Data Mining Applied to the Varicocele Condition

MIS

2019

INSTITUTO POLITÉCNICO DE COIMBRA

INSTITUTO SUPERIOR DE ENGENHARIA DE COIMBRA

Data Mining Applied to the Varicocele Condition

Trabalho projeto para a obtenção do grau de Mestre em

Informática e Sistemas

Autor

Judith Santos Pereira

Orientação

Prof. Doutor Jorge Bernardino

Dezembro de 2019

Mestrado em Engenharia Informática e Sistemas

Data Mining Applied to the Varicocele Condition

Trabalho Projeto
apresentado para a obtenção do grau de Mestre em
Informática e Sistemas
Especialização em Tecnologias da Informação e do Conhecimento

Autor
Judith Santos Pereira

Orientador
Prof. Doutor Jorge Bernardino
Departamento de Engenharia Informática e de Sistemas
Instituto Superior de Engenharia de Coimbra

Co-orientadores
Doutora Ana Paula Sousa
Centro de Neurociências e Biologia Celular
Centro Hospitalar e Universitário de Coimbra
Prof. Doutora Le Gruenwald
Universidade de Oklahoma

Supervisor
Prof. Doutor João Ramalho-Santos
Centro de Neurociências e Biologia Celular
Universidade de Coimbra

Coimbra, Dezembro, 2019

Acknowledgements

First, I would like to thank my supervisor, Professor Jorge Bernardino, from the Polytechnic of Coimbra - ISEC, for its encouragements and teachings on the scientific domain, as well as its support and understanding on my choices which enabled me to pursue my will of applying data mining techniques upon a real use case of the health care domain. However, its fulfillment would not have been possible if Professor João Ramalho-Santos, President of the Center for Neuroscience and Cell Biology, would not had accepted this collaboration and for it, I am sincerely grateful, as well as thankful to also letting me discover different domains and practices, such as biology and clinical investigation, by inviting me to their weekly working meetings of the Biology of Reproduction and Stem Cell group where I have learned a lot with other researchers in the male infertility domain.

Furthermore, I would like to leave a special thanks to my co-supervisor, Doctor Ana Paula Sousa, for having me in the laboratory of the Reproductive Medicine Unit of the CHUC - *Centro Hospitalar e Universitário de Coimbra* along with other researchers which helped me to dive into the health care domain; and hence, be more encouraged and taught. Moreover, I cannot forget to also thank the support, understanding and patience that was given throughout this project.

I would like to also thank my co-supervisor, Professor Le Gruenwald, from the University of Oklahoma, for its initial guidance, as well as Professor Carlos Pereira and Professor Viriato Marques, from the Polytechnic Institute of Coimbra, for their enlightenments on whether or not I was in the right path.

At last, I would like to leave a special thanks to my family: to my parents that have encouraged me to pursue my studies, to my sister, from the other side of the ocean, that still tells me to give my best and to my kind and understanding husband, that encourages me to chase my goals in a world where quitting is not an option…it might be a classic end, but it is what I truly feel…. without them, I wouldn´t be able to do this...Thank You.

Resumo

O sistema de saúde guarda cada vez mais informação dos seus utentes o que dificulta ou até impossibilita a descoberta de novos conhecimentos só com as técnicas usualmente utilizadas, i.e., as tradicionais técnicas estatísticas. De facto, os investigadores clínicos têm sentido uma crescente necessidade em extrair novos conhecimentos para continuadamente contribuir para o melhoramento dos serviços de saúde prestados. Essa necessidade tem vindo a ser colmatada com a aplicação de um processo, chamado “data mining”, que auxilia, através da aplicação de diversas técnicas (i.e., classificação, *clustering*, associação, etc.), a descoberta de padrões de dados vistos como interessantes, mas ocultados com as tradicionais técnicas estatísticas. A área da infertilidade masculina já começou a aplicar o data mining, por exemplo, através da aplicação da técnica de classificação para prever o sucesso de uma técnica de Procriação Medicamente Assistida. Contudo, o varicocelo - um síndrome anatómico de varizes escrotais caracterizado pela dilatação das veias que drenam o sangue da região dos testículos que em certos casos dá origem à infertilidade - não foi ainda explorado com uma técnica de data mining. A sua prevalência atinge 40% dos homens tratados por infertilidade, sendo que a infertilidade masculina abrange 50% das causas da infertilidade de um casal. A correção do varicocelo pode ser alcançada com um tratamento radiológico chamado embolização, que tem por objetivo desvitalizar as veias dilatadas através da introdução de substâncias terapêuticas na circulação sanguínea. Neste contexto, este trabalho teve os seguintes principais objetivos: i) averiguar o sucesso da correção do varicocelo com a técnica da embolização através da identificação de algum melhoramento na média dos valores dos parâmetros seminais ou das categorias seminais com recurso a técnicas estatísticas inferenciais (i.e. ANOVA e Chi-quadrado); ii) predizer o sucesso da embolização com técnicas de classificação através da aplicação do *decision tree* do RapidMiner e do algoritmo W-J48; iii) identificar padrões que caracterizam os pacientes embolizados com a técnica de clustering através do algoritmo K-Means e eleger as relações de atributos que ocorrem mais frequentemente através da técnica de associação com o algoritmo FP-Growth. Este processo de análise de dados seguiu a metodologia *Cross-Industry Standard Process for Data mining* (CRISP-DM) aplicando-a à análise de uma amostra de 293 homens inférteis descritos com 64 atributos que foram submetidos à embolização no Centro Hospitalar e Universitário de Coimbra (CHUC) entre Janeiro de 2007 e Abril de 2016. Os resultados obtidos indicam que a embolização melhora significativamente a média das concentrações de espermatozoides até 12 meses e de suas morfologias até 6 meses depois da embolização (ANOVA p<0.05) o que permite fundamentar o interesse em prever o sucesso desta técnica terapêutica. Sua previsão computarizada com a árvore de decisão do RapidMiner permitiu prever com uma *Accuracy* e *F-measure* de 70.59% e uma *AUC* de 0.750 que a probabilidade condicional de engravidar tendo um homem com uma severidade baixa ou média do varicocelo e uma parceira entre os 24 e 33 anos inclusive é de 70.83%. Também se viu que a frequência relativa, de pacientes com uma concentração de espermatozoides normal 3 meses depois da embolização e uma motilidade progressiva normal destes antes do tratamento, é mais alta em grupos de pacientes que raramente trabalham em ambientes tóxicos. Estes resultados permitem contribuir para as investigações em curso no domínio da infertilidade, assim como na identificação de medidas que permitem um maior auxílio na descoberta do conhecimento. Nomeadamente, vimos que a aplicação conjunta dos algoritmos de data mining com as técnicas estatísticas inferenciais, assim como a aplicação de diversas técnicas de data mining (i.e., classificação, clustering e associação), potencia a descoberta do conhecimento em dados clínicos.

Palavras Chaves: Data mining, Varicocelo, Embolização, Parâmetros Seminais

Abstract

Traditional statistic´s limitations took researchers to seek for advanced data analytic methods to better identify factors that may improve, for instance, treatment success. Since data mining techniques enhances the potential of knowledge discovery in data, by following the Cross-Industry Standard Process for Data mining (CRISP-DM) methodology, we have applied data mining techniques to leverage our descriptive statistical findings on the male infertility domain. More precisely, on the varicocele condition corrected with the embolization treatment. The aim of this study is to identify data patterns on patient's data with the varicocele condition and predict the success of the embolization treatment with data mining techniques in order to contribute to actual researches with an innovative data analysis approach that, to the best of our knowledge, have not been yet applied to the varicocele condition. In this context, after presenting our descriptive statistical results, we describe and apply the most commonly applied Data mining techniques in the healthcare domain: Classification with the RapidMiner´s Decision tree algorithm and the W-J48 java implementation of the C4.5 algorithm; Clustering, with the K-Means algorithm and Association, with the FP-Growth algorithm. Further on, we identify the most interesting obtained results, and at last, discuss the elected results and contributions for the studied domain.

Keywords: Data mining, Varicocele, Embolization, Seminal Parameters

Table of Contents

List of Figures

List of Tables

List of Formulas

Chapter 1 Introduction

The healthcare industry daily generates complex data from multiple sources, such as electronic patient records, medical reports, hospital devices, and billing systems (Cerquitelli, Baralis, Morra, & Chiusano, 2016), which makes knowledge discovery from this data harder or impossible to achieve using traditional statistics. In fact, several studies have suggested the advanced data analysis technique called data mining to overcome these data challenges (Cerquitelli et al., 2016) (Gonzalez, Tahsin, Goodale, Greene, & Greene, 2016).

Data mining is the process of discovering interesting data patterns (Han, Kamber, & Pei, 2012) where standard statistical exploratory data analysis procedures - traditional statistics - could not discover useful insights (Hand, Blunt, Kelly, & Adams, 2000). In our era, traditional statistics is viewed as the primary data analysis technique and data mining as the secondary technique due to its strengths and rapid developments (Hand, 1998). While the groundwork of both techniques is mathematics, data mining extends it with other subjects such as machine learning, database systems and visualization which brings important gains over the traditional statistics techniques (Tekieh & Raahemi, 2015). The main advantages of data mining over the traditional statistics techniques are its capability to analyze different types of data (i.e numbers, names, severity degrees etc.) as well as its ability to perform inductive analysis. This last advantage is fundamental in cases where researchers are trying to understand, for instance, the consequences of a treatment that are not fully known since many potential variables can difficult the formulation of a hypothesis to prove or reject.

The varicocele condition is characterized by the dilation of the veins of the spermatic cord (Arif et al., 2018). Studies estimates that the varicocele condition is present in more than 35% of infertile couples (Kirby, Wiener, Rajanahally, Crowell, & Coward, 2016). The McGraw-Hill Concise Dictionary of Modern Medicine (“varicocele Definition,” 2002) goes even further, by stating that the varicocele condition is linked to infertility in 40% of males treated for infertility. By having in mind that male infertility factors are responsible for 50% of infertility causes (Kirby et al., 2016), the importance of assessing data patients with a condition with such prevalence is clear, given that infertility affects an estimated 15% of couples globally (Agarwal, Mulgund, Hamada, & Chyatte, 2015). Varicocele correction can be achieved with the radiological embolization technique that introduces substances into the circulation to devitalize the enlarged veins (Lippincott, Williams, & Wilkins, 2012).

Several studies were carried out on the varicocele domain, however, to the best of our knowledge, none of them have applied an advanced data analysis process such as data mining. Therefore, data mining techniques were in this study applied on a data set of 293 infertile male patients with the varicocele condition that had undergone the embolization treatment in the *Centro Hospitalar e Universitário de Coimbra* (CHUC) with the aim of improving their chances of conceiving. This data set not only had general male patient´s information and external factors (i.e. if the patient drinks, smokes etc.) but also covered all semen analysis results carried out before and after the embolization treatment at 3, 6 and 12 months which were at the Reproductive Medicine Unit of the CHUC collected.

The aim of this work is to predict the varicocele embolization success and identify data patterns with data mining techniques. Accordingly, the main contributions of this study are the following:

- Contribute to the ongoing research on varicocele embolization;
- Leverage the findings in the global field of male infertility;
- Identify measures that can leverage knowledge discovery on similar data sets.

Its aim was achieved by following the Cross-Industry Standard Process for Data Mining (CRISP-DM) with the application of the commonly used data mining algorithms (i.e. Decision tree, K-Means and FP-Growth) which encompass the most commonly used data mining techniques in the health care industry (i.e. Classification, Clustering, and Association).

This master thesis is structured as follows. Chapter 1, introduces the carried-out study, as just discussed. Chapter 2, briefly describes the data mining technique and the varicocele condition with its correction. Chapter 3, discusses the surveys and researches that have been published on Data mining, as well as on the field of male infertility and varicocele. Chapter 4, presents the materials that this study has used and explains how this study was carried-out after describing each method applied. Chapter 5, presents the results of this real use case through several sections that reflects the CRISP-DM methodology. Finally, in Chapter 6, a conclusion is presented where future work is indicated. To better support the exposed study, a glossary is presented after Chapter 6 and most modelling results are documented in the Appendix C.

Chapter 2 Background

In order to better convey the carried-out work, this chapter presents a background on Data mining and on the varicocele condition as follows: in section 2.1, we provide a brief overview of data mining and in section 2.2, we disclose in more details, what the varicocele condition is all about and how this condition can be corrected.

2.1 Data mining

Data mining is the extraction of implicit, previously unknown, and potential useful information from data (Witten, Frank, & Hall, 2011). It is also defined by Han *et al*. (2012) as the process of discovering interesting patterns and knowledge from data; and by Ting, Shum, Kwok, Tsang, and Lee (2009), as the process of finding patterns, associations or relationships among data using different analytical techniques involving the creation of a model that will compute useful information or knowledge. This is why data mining is also popularly called "knowledge discovery from data", or KDD (Han et al., 2012).

As specified by Ting *et al*. (2009), data mining techniques are applied upon the data through the creation of data mining models. A data mining model, is a set of several steps (e.g. data reading, *feature selection*, algorithm application etc.) build with a data mining tool that aims to optimize the knowledge discovery process.

Data mining techniques extract knowledge from data with different methods that are implemented with various algorithms. These techniques can be divided into two main categories that reflects the main purposes of data mining (i.e. predict and describe data). Below, we describe these two main categories based on Tekieh & Raahemi (2015):

- **Predictive** – analysis that tries to generate predictive rules with the classification of *instances* on a specific *label attribute*. Since most of the predictive algorithms carry out predictive analysis through a *label attribute*, we also call them *supervised learning*. These algorithms enable to build predictive data mining models to predict, for instance, the success of a treatment. Classification is the most widely used predictive data mining technique.
- **Descriptive** – exploratory analysis that attempts to measure the similarity between *data values*, and discover data patterns and relationships. Hence, the aim of this category is to describe the data. The data mining algorithms that are applied to describe data are called *unsupervised learning* because they do not need a *label attribute* to compute as the predictive algorithms do. The mostly applied descriptive algorithms are the ones that perform Clustering or Association data mining techniques. However, Classification techniques can also be used to describe data in conjunction with Clustering techniques.

As previously said, data mining is an advanced data analytic technique that has several advantages over the traditional statistical techniques. These main gains are the following (Tekieh & Raahemi, 2015)(Dj Hand, 1999):

- **Heuristic technique** – data mining is a technique designed to extract knowledge from data in a better and quicker way than with traditional statistics and surpass its results.
- **Open approach** – the data mining technique is open to consider various approaches to mine the data and apply them in different orders. In opposition, traditional statistics is more conservative.
- **Inclusive** – the nature of statistics methods is to run analysis only on a sample of data. In contrast, data mining can consider the whole dataset for analysis.
- **Generalized** – most data mining methods can handle all types of data. In opposition, traditional statistics technique only analyzes numeric data.
- **Inductive** – in statistics, a hypothesis is first created and then the data gets analyzed to prove or reject the hypothesis (hypothetical-deductive analysis). On the other hand, data mining explores the data and tries to find knowledge out of the data (inductive analysis).

2.1.1 Data mining techniques

Through our study, we have identified that the most commonly used data mining techniques in the health care domain were: Classification, Regression, Clustering and Association (Tekieh & Raahemi, 2015) (Ahmad, Qamar, Qasim, & Rizvi, 2015) (Tomar & Agarwal, 2013). Therefore, in the following sub-sections we describe each of these techniques.

2.1.1.1 Classification

Classification is a data analysis technique constructing models that predict categorical *label attributes* (Han et al., 2012). This technique is used when the data is required to be classified into different groups (Tekieh & Raahemi, 2015), and/or predict the conditional probability of a *label attribute* outcome based on historical records. This method has been used in various healthcare applications: the classification technique was applied to better identify if a patient has dementia based on his/her neuropsychological test in Maroco *et al*. (2011). In Geman, Chiuchisan, & Covasa (2016), the Support Vector Machine and Artificial Neural Network algorithms were used to find correlations between a specific intestinal microbiota and the presence or absence of diabetes in order to predict metabolic diseases such as diabetes. In another work, the Support Vector Machine algorithm was also used along with the Particle Swarm Optimization, to predict seminal quality (Sahoo & Kumar, 2014). In Mirroshandel, Ghasemian, and Monji-Azad (2016), the K-Star algorithm was used to predict the outcome of individual sperm injection on humans in order to increase the implantation rate. In Kourou, Exarchos, Exarchos, Karamouzis, and Fotiadis (2015), a survey of works on data mining applications in the cancer prognosis and prediction field was presented. It turned out that all the presented works had applied algorithms that are classifiers. Just to name a few, Decision tree algorithms were used to predict breast cancer survival (Delen, Walker, & Kadam, 2005) and Bayesian Network to predict the recurrence of oral cancer considering several data types such as clinical imaging and genomic data from tissue and blood (Exarchos, Goletsis, & Fotiadis, 2012).

2.1.1.2 Regression

Regression is mainly used to demonstrate the correlation among different numerical attributes (Tomar & Agarwal, 2013) and predict a numerical value. Its limitation is that we already have to know the attribute that is independent and the attribute that is dependent in order to apply it (Ahmad et al., 2015). Since it is a technique that is closely related with what we do in statistics and the aim of this study is to explore data mining techniques by among other things, predict a nominal value, we did not delve further in this technique.

2.1.1.3 Clustering

Clustering is the process of partitioning a set of instances into subsets (Han et al., 2012) and by doing so, categorizing it. This technique is used when we do not have much information about the different types of instances (in our case, patients) involved in a population. As it is an unsupervised learning technique, it tries to find clusters of instances that are similar to each other without considering any specific target label (Tekieh & Raahemi, 2015). Since clustering is a technique specially used in the descriptive analysis stage, several works have applied clustering algorithms to usually categorize the handled data prior to classification. In fact, in Sharma, Singh, and Khatri (2016) a survey is presented on medical publications that have used Clustering techniques along with Classifications techniques where the following applications were pinpointed: the K-Means clustering algorithm was used to contribute to the diagnose of heart disease patients (Shouman, Turner, & Stocker, 2012) and to categorize colon tumors (Kumar & Wasan, 2010); Clustering methods were also used to categorize proteins into functional groups (Xu et al., 2012) to predict the likelihood of diseases (Paul & Hoque, 2010) and to detect disease-specific clusters within medical image data (Bruse et al., 2017).

2.1.1.4 Association

Association is the process of finding common sets of attributes, also called in the literature as “frequent item sets”, to generate rules that can describe the data set (Han et al., 2012) by identifying data patterns. These rules can also have the ability to predict events if the *consequent* of the rule is a label attribute or an attribute that can enable the prediction of an event, as studied in Azevedo and Jorge (2007) and Liu, Hsu, and Ma (1998). The association technique is applied with a “frequent item set mining algorithm” and the most commonly applied algorithms are in this context the APRIORI and the FP_Growth algorithms. In fact, the APRIORI algorithm was used to find associations between clinical data from diabetic patients (Stilou, Bamidis, Maglaveras, & Pappas, 2001) and in Hanirex, Kaliyamurthie, and Kerana (2015), the FP_Growth algorithm was used to infer disease association. Other association rule mining algorithms were proposed to find associations between time, place and patient´s infections on public health surveillance data (Brossette et al., 1998); between clinical data and therapeutic treatments (Ting, Wang, Kwok, Tsang, & Lee, 2010); between medical data and rhinitis conditions (Yang, Li, & Luo, 2016); and between patient´s data for coronary heart disease diagnosis (Orphanou et al., 2016).

2.2 Varicocele

Since the provided data set only covers male patients that have undergone varicocele correction, exploring the outlines of the varicocele condition with its corrections techniques was needed to better understand the data provided. Hence, in section 2.2.1, we present the definitions found on the varicocele condition and in section 2.2.2, we specify how this condition can be corrected.

2.2.1 Definition

As previously said, the varicocele condition is characterized by the dilation of the veins of the spermatic cord (Arif et al., 2018) that, depending of its severity grade, can be seen with naked eye on the patient´s scrotum. However, to better understand the provided data, a greater comprehensive and clinical definition was needed. Hence, complementary definitions were searched. Below, we present several clinical definitions of the varicocele condition that were retrieved from several medical dictionaries and related works.

The Farlex Partner Medical Dictionary (“varicocele Definition,” 2012), defines varicocele as a condition manifested by “abnormal dilation of the veins of the spermatic cord, caused by incompetent valves in the internal spermatic vein and resulting in impaired drainage of blood into the spermatic cord veins when the person assumes an upright position”. The McGraw-Hill Concise Dictionary (“varicocele Definition,” 2009) goes further on this definition by specifying that the incompetent valves that patients with varicocele have in the internal spermatic vein are, in fact, abnormal valves that “obstruct normal blood flow causing a backup of blood, resulting in venous dilatation”. This same definition goes further by characterizing the varicocele condition by specifying that it usually develops “slowly” and that it may be “asymptomatic”. They also specify that the incidence of the varicocele condition is higher in male than in female patients, since the varicocele condition can also affect female patients with the enlargement of the veins within the uterus. The McGraw-Hill Concise Dictionary also characterizes varicocele patients by specifying that it has a higher incidence on patients between 15 and 25 years old and that 40% of the males treated for infertility has the varicocele condition in a context where male factors contribute to 50% of s infertility causes. Finally, McGraw-Hill also specifies that an “abrupt appearance of a varicocele in older male patients may be caused by renal tumor [..] that alters the blood flow through the spermatic vein”. The Collins Dictionary of Medicine also adds to this definition (“varicocele Definition,” 2009) by relating it to “Varices” that surrounds testicles and forms “irregular swelling in the scrotum” that may cause a “dragging ache”. The Collins Dictionary of Medicine also specifies that the varicocele incidence is higher in the left testis and that it may affect fertility, and in these cases, correction is needed.

The varicocele condition has 3 severity grades (i.e. severity grade I, II and III) where the severity grade I is the mildest and the severity grade III is the severest.

Through the study of related works on varicocele, we have seen that the incidence of the varicocele condition in normal healthy males is estimated to be between 8 to 23% with the majority of cases affecting the left side (Makris et al., 2018). Furthermore, the left testicular varicoceles were associated with decreased testicular volumes in 73%, 53% and 43% in varicocele with grade III, II and I, respectively (Aza Mohammed & Frank Chinegwundoh, 2009).

2.2.2 Correction

The correction of a varicocele can be carried out through surgery or a radiological technique. Over the last decades, the radiological technique has become increasingly popular as a less invasive technique - initially used when surgery failed but now has a treatment on its own (Makris et al., 2018). The radiological technique that is being used is the embolization technique which is the treatment that all our patients have undergone. In the medical dictionary for the health professions and nursing, embolization is defined as a therapeutic introduction of various substances into the circulation [..] to devitalize a structure or organ by occluding its blood supply (Lippincott Williams & Wilkins, 2012). In the varicocele domain, the objective of the Embolization is to devitalize the enlarged veins in the man´s scrotum to divert the blood flow away from the varicocele.

The embolization technique uses mechanically occlusive solid or liquid embolic agents such as coils, sclerosants or glue to block blood circulation on the veins with varicocele. In the systematic review of 30 clinical studies on embolic agents carried out by Makris *et al*. (2018), 898 patients out of 3505 were treated with coils alone and the average technical success rate was of 92%. The most common complication with coils was epididymis-orchitis (i.e testicle inflammation), pampiniform plexus phlebitis (i.e. inflammation of the veins) or hydrocele (i.e accumulation of fluids around testicle) which was observed in 3.4% of the patients.

The glue was the least embolic material used with 251 patients in total and the average technical success rate was equal to the coils. The most common complication with this embolic material was the perforation of the internal spermatic vein with contrast extravasation which occurred in 5.8% of the cases. In terms of complications, the glue was associated with a significantly higher risk than the other embolic materials (p <0.05) and in terms of pain after the treatment, moderate post-embolization pain was also reported in 3,7% of these patients which shown to be significantly higher than the other materials ($p < 0.05$). Potential improvement in sperm characteristics were evaluated in 11 of the 30 studies assessed, all of which reported statistically significant improvement in sperm count and/or sperm motility. Furthermore, we have seen that 4 studies out of 30 have reported outcomes on the frequency of successful pregnancies.

Chapter 3 Related Work

Prior to the application of data mining in our project, several related works were studied to provide some guidance. The election of the data mining tool and algorithms to use, as well as the guidance on factors that influence male infertility or are clinically relevant to assess the varicocele condition, were some of the aspects researched in the literature and in this chapter summarized as follows: subsection 3.1 presents a summary of surveys on data mining tools; subsection 3.2 provides an overview of studies that have applied data mining techniques to male infertility by presenting the data mining algorithms that have been used; subsection 3.3 presents an overview of works that identify factors that influence male infertility and subsection 3.4 describes the latest statistical findings performed on data of patients with varicocele.

3.1 Data mining tools

Since the election of a suitable data mining tool is important to promote the success of a data mining project, we have looked up in several data mining tool surveys for guidance on which data mining tool we should use. Below, a summary of identified surveys is presented.

In Mikut and Reischl (2011), an overview on most existing data mining tools was presented. Its main contribution is the presentation of tool categorization criteria based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and license policies. The authors used tool categorization criteria to classify the data mining tools into nine types (e.g. data mining Suites, Business intelligence packages, Mathematical packages etc.). The benefit of the authors´ approach is that they listed most data mining tools by specifying their tool types. Moreover, they also identified which types of tool are suitable for which identified user groups which are business applications, applied research, algorithm development and education. The drawback of this work is that they did not describe data mining tools in depth and did not compare them.

In Begum (2013), the authors discussed the KDD process and various open source tools (R, Weka, Orange, RapidMiner, Tanagara). Its highlights are the identification of data mining current and future trends in most domains – cloud and distributed computing were identified as well as the heterogeneous and complex data characteristic. Another useful input is the identification of data mining methods to tackle the trends that are basically challenges of the KDD process. However, the discussed data mining tools were not compared or selected by any criteria.

Due to the increased popularity of data mining tools, a number of data mining tools surveys were conducted. In Rangra and Bansal (2014) a theoretical analysis of six open source data mining tools, Weka, Keel, R, KNIME, RapidMiner, and Orange, was given. The strongest points of this paper it is that each data mining tool was described through its technical specifications, features and specializations, as well as its advantages and limitations. Unfortunately, the authors did not say why they have considered these tools in their survey, nor specified the areas to which they are suited. Furthermore, for healthcare researchers who are beginners in the data mining field and want to identify the most suitable tools for their needs, all the technical aspects shown can be overwhelming.

In Jović, Brkić, and Bogunović (2014), a data mining tool survey was conducted that specifies the reason of its tools of choice – mostly on the results of the KDnuggets poll. Their work emphasizes the quality of RapidMiner, R, Weka, and KNIME platforms, identified in the 2013 KDNuggets poll, but also acknowledges the significant advancements made in other tools like Orange and Scikit-Learn. The strongest aspect of this paper it its technical aspects: selected tools were compared based on their general characteristics (i.e. programming language, license, etc.), applicability (i.e. Big Data, Text Mining, etc.) and data mining algorithms and procedures (Data visualization, Decision tree algorithms, etc.).

The performance of the classification algorithms (Naïve Bayes, Random Forest, Random Tree and Bagging) were evaluated through the use of three data mining tools (Weka, RapidMiner and Support Vector Machine) in Mishra and Thakur (2014). The contribution of this work is the identification of the best algorithm and tool for spam/junk mail classification. One of the benefits of this work is the performance evaluation of the tools in a specified field. Although tool performance was compared, tools were not described.

A similar work was carried out in Singh, Liu, Ding, and Li (2016) that presents an evaluation of RapidMiner and KNIME, on a customized predictive and descriptive model built with the VCF feature for telecom monitoring data. The strongest points of this paper are that it shows how it is possible to build a model in the KNIME and RapidMiner with the VCF feature, as well as its performance evaluation through their benchmarking. Although the tools were analyzed qualitatively and quantitively, the authors did not present a qualitative comparative table. We believe that this type of comparison would help in selecting a tool to adopt.

In Al-odan and Saud (2015), a comparative study of data mining tools suited for small to medium enterprises (SMEs) was presented. The reviewed data mining tools were KNIME, RapidMiner, Weka, RStudio and Orange. These tools were evaluated by 17 participants with more than 15 work years' experience in the IT field to assess their user experience though their intuitiveness, consistency, navigation, usability, installation manual, configuration guide, troubleshooting guide, and user tutorials. The weakest point of this paper is that the data mining tools are not clearly described in detail, in spite of their comparison.

In P. Aalam and T. Siddiqui (2016) seven data mining tools - Weka, ELKI, Orange, R, KNIME, Scikit-learn, and RapidMiner – were compared for clustering. The positive aspect of this paper is that they describe and compare qualitatively (programming language, interface type, covered clustering algorithm, etc.) the data mining tools. However, its focus is only on clustering.

In Almeida and Bernardino (2016) a survey on seven open source data mining tools for SMEs. KEEL, KNIME, Orange, RapidMiner, RProject, Tanagra and Weka were qualitatively compared (programming language, interface existence, data types supported, etc.). The strongest points of this paper are the good descriptions of the selected tools and the identification of the Cloud Services and Big Data (large amounts of data) support of the tools. Since those aspects are important to assess data mining tools for the healthcare domain as well as due to its good tools description, we have adopted some of its conclusions. The weakest point of this paper is that it does not evaluate the performance of the selected tools. In a subsequent paper (Almeida, Gruenwald, & Bernardino, 2016), the authors have addressed the interest of data mining for business and analyzed three popular Open Source data mining Tools – KNIME, Orange and RapidMiner. The strongest point of this paper is its tool evaluation that besides comparing the execution times of the tested algorithms, has also compared the results on seven other performance metrics – Precision, Recall, F-Measure, ROC, Accuracy, Specificity and Sensitivity. Although this work is related to the business domain, our work gains from it through its tool analysis and evaluation.

In Sharma *et al*. (2016) a survey on the application of the *Classification* and *Clustering* data mining methods to heart and Cancer diseases was provided. The paper briefly suggested the following data mining tools: RapidMiner, Weka, R-Programming, Orange, KNIME and NLTK. The strongest point of this work is its survey on its data mining applications in its project´s aim. However, the way that they have suggested the data mining tools is its weakest point - They are briefly described, neither compared nor suggested based on healthcare requirements.

In (Gui, Zheng, Ma, Fan, & Xu, 2016), a data management architecture for the healthcare industry, more precisely, for the personal health problem detection and real-time vital sign monitoring was proposed. The strongest point of this work is its data architecture suggestion that has in account the large amount of data characteristic with its specificities. However, the domain requirements are not deeply investigated, leaving behind other requirements such as the user experience of the proposed data management architecture. Their prototyped system was constructed with the Hadoop Database named HBase, the Hadoop Data Warehouse Hive and the data mining Libraries MLib and Spark Streaming. The proposed combination of tools makes sense since the Spark Libraries works perfectly with the Hadoop tools.

By analyzing all these works on data mining tools, we have seen that the Rapid Miner and the KNIME platform have been the most studied. Most of these studies elect the RapidMiner platform as the most intuitive and complete and by looking into the reports from the consulting company Gartner, we see that the RapidMiner platform has been placed for the fifth year in a row as a leader in the market of Data Science and Machine-learning platforms.

3.2 Data mining algorithms

To the best of our knowledge, no other work applies data mining techniques to a data set of patients with varicocele. However, we have found several works that use data mining techniques to study seminal parameters (i.e. sperm concentration, sperm progressive motility and sperm morphology), as well as external factors, in the male infertility context. Since seminal parameters and male external factors encompass a good part of the provided/collected data set, these studies were analyzed to guide us on the selection of the data mining algorithms for this study. Hence, Table 3.1 was built to give a glance on the data mining algorithms that have been applied in related works to better support the election of the data mining algorithms applied to this study.

Table 3.1 Studies that apply data mining algorithms to the infertility domain

Reference	Subject	Data Set	Preprocessing	Algorithms	Best Accuracy
(Sahoo & Kumar, 2014)	Predict seminal quality from environmental factors and life habits data.	100 instances, 9 attributes	Feature Selection	Multilayer perceptron (MLP), Decision Tree (DT), Naive Bayes (Kernel), Support vector machine + Particle swarm optimization (SVM+PSO) Support vector machine (SVM)	SVM + PSO -> 94%
(Bidgoli, Komleh, & Mousavirad, 2015)	Predict seminal quality from life habits and health status data.	100 instances, 9 attributes	Balance Data Set	Optimized MLP with genetic algorithm, SVM, DT and Naive Bayes (NB).	Optimized MLP -> 93.86%
(Gil, Girela, De Juan, Gomez-Torres, & Johnsson, 2012)	Predict seminal quality by associating environmental factors and lifestyle	100 instances, 9 attributes	Feature Selection	DT (C4.5) in binary method, MLP and SVM.	MLP -> 86% and SVM -> 86%;
(Guh, Wu, & Weng, 2011)	Predict IVF outcome from several patient´s descriptions (e.g., patient’s age, number of embryos transferred, number of frozen embryos, culture days of embryo, sperm parameters etc.)	5275 instances, 38 attributes	Balance Data Set & Feature Selection	DT (C4.5) with Genetic Algorithm (GA) for attribute selection	DT -> 73.2%
(Chen, Hsu, Cheng, & Li, 2009)	Predict IVF outcome from the patient’s physiology and the results of the stages of the IVF cycle	654 instances, 10 attributes	Feature Selection	PSO, Decision Tree J48, Naive Bayes, Bayes Net, MLP ANN	PSO -> 73.03%

By analyzing the built Table 3.1, we see that most identified studies have worked with small data sets and used the feature selection technique to select the attributes to mine. Furthermore, we see that only two of the five identified studies have balanced its data set, four of the five studies have applied the MLP algorithm and three of them, have also applied the SVM algorithm. Although several algorithms were used, the ones that gave the best accuracy were: Support Vector machine (SVM); Particle Swarm Optimization (PSO), Multilayer perceptron (MLP) and Decision Tree (DT). The study that has applied data mining algorithms upon sperm parameter values to predict a treatment outcome, similarly as we aim to do, is the work carried out by (Guh et al., 2011). Hence, we have chosen to go with the application of the decision tree algorithm (C4.5) since it is also a well-known algorithm that has been widely applied. In fact, in Table 3.1 we see that all identified works have applied it.

3.3 Risk Factors of Male Infertility

To establish guidelines on the type of information to collect/select for this study, we have selected several related works on male infertility to identify risk factors. Table 3.2 presents an overview of all identified studies.

Table 3.2 Risk Factors linked to male infertility

Reference	Risk Factors
(Keller, Chen, & Lin, 2012)	Erectile dysfunction
(May et al., 2006)	Weight, height and body mass index during childhood and adolescence (people with left varicocele were heavier and taller than an age-correlated normal population)
(Mohammadali Beigi, Mehrabi, & Javaherforooshzadeh, 2007)	Prevalence of varicocele in the patients’ brothers
(Niederberger, 2015)	Different infertility etiologies are genetically and clinically linked with other diseases
(Wang et al., 2016)	Environmental exposure to metals: tin, nickel, zinc and molybdenum may be associated decreased total testosterone or luteinizing hormone (total T/LH ratio). Manganese may induce spermatozoa apoptosis. Iron may be important for living spermatozoa.
(Williams & Alderman, 2001)	Woman Age
(Xu et al., 2012)	Proteins in the sperm
(Yan et al., 2014)	Nitric Oxide

By considering the identified risk factors presented in Table 3.2, we see that the male patient partner age, as well as the previous diseases of the male patient were attributes seen as risks factors for male infertility; and therefore, they have also been studied in this work.

3.4 Varicocele

The varicocele condition has been widely covered and assessed with statistical techniques. Since this study also applies statistical techniques to better understand the data prior the application of data mining techniques, we found that the study of those works was worthwhile. Hence, we have analyzed the varicocele-related works by identifying the type of information that has been studied in its results, as well as the statistical tests that have been applied to guide us through our work. By targeting these interests, this section presents an overview of some of the latest studies on the varicocele condition, and specifies how our work is different, and innovative.

In Delavar *et al*. (2014), it is seen that the percentage of varicocele is significantly higher in smokers compared with non-smokers. However, they have not found a significant difference between the varicocele condition and the occupation or the drinking alcohol habit of their patients. This study used the SPSS software version 16.0 for statistical analysis. Adjusted regression analysis was used to test associations between the nominal attributes and the significance level used was considered at $p<0.05$. This study relates to ours because our study also encompasses the external factors of the patients but for a different purpose: to assess if there is a relation between the embolization success, through pregnancy outcome, and external factors.

In DeWitt *et al*. (2018), it is shown that the laterality of the varicocele condition was not significantly associated with cancer diagnosis, nor vascular anomalies. Statistics were performed with the JMP Pro 12 of the SAS Institute. They used the mean and the standard deviation for the descriptive statistics and for the comparative statistics, they have used the Chi-square test with Fisher exact tests for categorical data. The significance level used in this study was set at p<0.05. This study relates to ours because we also have the information of the laterality of the varicocele condition.

In Bilreiro *et al*. (2017), it is seen that both materials improved the sperm parameters with similar success rates. This study applied the Student´s t test in numeric continuous attributes and Fisher´s exact test for categorical variables. The significant level used was p<0.05. This study used the Microsoft Office Excel 2010 and the Graphpad Prism 6 software. This study relates to ours since we also have had the interest to collect patient information regarding the embolic material used.

In Çayan and Akbay (2018), it is seen that patients that have redone the microsurgical sub inguinal procedure have significantly improved their postoperative sperm parameters, serum total testosterone level and spontaneous pregnancy rates. Statistical significance was assessed with the Student´s t test on sperm parameters and with the Chi-square test, for comparison of pregnancy results. These tests were applied with the SPSS 16.0 software package. This study relates to ours because, although covering the microsurgical sub inguinal varicocelectomy treatment rather than the embolization treatment, both studies assesses the success of their treatments with the increase of seminal parameters values, as well as pregnancy rates.

In Samplaski, Lo, Grober, Zini, and Jarvi (2017), it is shown that the correction of a varicocele could reduce the need of *in vitro* fertilization (IVF). This study used the Student´s t test to compare the changes in the semen parameters and used the Chi-square test to compare groups of patients. The level of significance used was p<0.05. This study relates to ours since we also use the sperm parameters attributes to assess the improvement of the embolization treatment.

In Kirby *et al*. (2016), we identify that the varicocele correction improves pregnancy and live birth rates. Since this paper is a review, the statistical tests used does not relate to our application. However, its main contribution does since our work also analysis the impact of the varicocele correction by assessing the success of the embolization treatment through the patient´s pregnancy outcome and analyses the evolution of semen categorizations (i.e. azoospermia, oligospermia) after the treatment.

In terms of attributes assessed in the varicocele domain, we see that sperm parameters, as well as semen categorization have been analyzed. Moreover, external factors such as previous diseases, occupation, drinking and smoking habits have also been assessed, as well as the laterality and embolic materials used during embolization. If we look up for the attributes that have been used to assess the impact of varicocele, we see that the pregnancy outcome and the rate of live babies has been used. Hence, these two attributes could also be used to predict the success of the embolization treatment.

Regarding the statistical tests used, we see that the Student´s t test (i.e. the ANOVA test only applied to two groups of data) and the Chi-square test were frequently used in varicocele studies. Furthermore, we have seen that in the WHO laboratory manual for the examination and processing of human sperm, the ANOVA test is also used to assess systematic differences among the sperm parameter values recorded by the technicians (WorldHealthOrganization, 2010). Moreover, all studies used a significance level of p<0.05. Hence, this study has also used these statistical tests and configuration to explore its data.

As we know, statistical "power" is achieved with a large amount of data. Hence, to see if our data set has an acceptable dimension in comparison to the data sets assessed in related works, we have also looked up for the number of patients (i.e. examples or instances in data mining) that these works had. To achieve that goal, we have also studied the work carried out by Makris *et al*. (2018) which studies 30 related works that statistically assesses patients treated with the varicocele embolization that, on average, entails data sets of 117 patients - the smallest data set had 16 and the largest, 468 patients, followed by 244 patients. The standard deviation of the number of patients assessed by all these 30 studies is of 102 patients which means that the dimension of the data sets that have been studied in this domain is very heterogeneous. Since our final and preprocessed data set has 293 patients, we consider that we have a very good data dimension since it is a number well above the average in varicocele-related works.

In contrast to all other studies noted above, this study applies *descriptive* and *predictive* data mining techniques while, to the best of our knowledge, only statistical techniques were previously used. Hence, this work has a high degree of innovation in terms of varicocele-based research.

Chapter 4 Materials and Methods

This study has used several materials and methods to achieve its data mining goals; and therefore, this section discloses in section 4.1, the materials that were used, and describes in section 4.2, the methods that were followed to better convey how, and for which purpose, they were used. The results of the application of all these methods are showcased in Chapter 5 with the CRISP-DM methodology.

4.1 Materials

In this section, in order to better convey the several assessed attributes disclosed in section 4.1.2, we previously expose in section 4.1.1, in which context these assessed attributes were obtained and how semen parameters were interpreted. At last, in section 4.1.3, we disclose the tools that were used to tackle the aim of this work.

4.1.1 Data Collection and Selection

The investigation team named “Biology of Reproduction and Stem Cells” (BRSC) of the Center for Neuroscience and Cell Biology (CNC) of Coimbra University has initially provided the data set for this study. The initial data set had 320 examples and 67 attributes, where 28 of them, were attributes generated by the BRSC investigators for statistical analysis purposes, 5, for data set management purposes - such as extra notes and example´s identifications - and 2, were empty or filled with the same value which down sized the data set width to 32 attributes that were selected with the BRSC team. These 32 attributes (i.e. 67-28-5-2=32) are in the Appendix A disclosed and are in this study identified as the “*initially provided and selected attributes*”.

After assessing the data quality of the attributes disclosed in the appendix A, we have seen that data preprocessing was needed to achieve the goals set. During the data preprocessing step, and based on the related works presented in the subsections 3.3 and 3.4, as well as our understanding of the varicocele condition and its correction described in subsection 0, we have reorganized and validated these 32 attributes. In the end, we have ended up with a total of 167 attributes. Out of these 167 attributes, 39 attributes mainly encompassed the 32 initially provided and selected attributes. Based on these 39 attributes, 25 were generated to better tackle the goals of this study. The remaining 103 attributes (167 preprocessed attributes minus 39 reorganized attributes minus 25 generated attributes) were mainly created to support the filling of missing values in the possible label attributes. For instance, to know if a patient´s partner got pregnant, we have recorded the pregnancy test results of all fertility procedures that the patient partner had undergone and at last, we have created an EXCEL condition clause that checked if the patient´s partner had at least one positive pregnancy result. If so, it would fill the missing “Gravidez” attribute with the value Yes. Since a patient partner can get also pregnant spontaneously (i.e. without an ART procedure) we have further on also checked the hospital´s information systems to look up for a possible child that could have born after the embolization treatment. These remaining 103 attributes were not considered for analysis since the goal of this study was to mainly analyze the information that the initially provided data set encompassed and mainly focus on the semen analysis since it is the most reliable and complete information that we could have. In fact, sperm parameter values could be validated with its related clinical test reports; and therefore, were all manually validated. Remaining patient data were validated or collected through the medical dossiers and hospital information systems where the physicians had recorded the medical appointment outcomes and scheduled the next medical appointments or tests. When some of the patient data were not found, the biologists of the Reproductive Medicine Unit of the CHUC that work for the BRSC research team have also contacted the patients by phone. Unfortunately, not all data could be to validated or collected, and we carried on with the data available. Why and how the *initially provided and select attributes* were preprocessed to achieve the 39 initially preprocessed attributes is disclosed in detail in the data Preparation section 5.4 and which attributes were selected for data mining purposes are at the end of this same section disclosed.

Since sperm parameter values were previously seen in the overview of related works as an important patient information, in the below subsection, we describe in more details how semen was collected and how this study has analyzed semen parameters.

4.1.1.1 Semen Collection and Analysis

To the patients of the provided data set it was asked to carry out a semen analysis in the Coimbra University Hospitals before the embolization treatment and 3, 6 and 12 months after the treatment. In order to carry out semen analyses, a sample of semen was obtained through masturbation after 3 to 5 days of sexual abstinence. Afterwards, sperm samples were analyzed and then elaborated a semen analysis report with the following information:

- the number of spermatozoa per milliliter of semen (i.e. sperm concentration);
- the percentage of spermatozoa as moving forward rapidly and slowly (i.e sperm progressive motility);
- the percentage of spermatozoa seen in the semen that have a normal shape (i.e. sperm morphology).

Further on, the recorded sperm parameter values were manually gathered in an EXCEL file that was then validated by comparing if the recorded values in the EXCEL file was the same with the ones recorded in the semen analysis paper report.

From then on, we have analyzed the validated sperm parameter values based on the thresholds defined by the last World Health Organization (WHO) Report published in 2010 (WorldHealthOrganization, 2010). These thresholds state that a male patient is considered with normozoospermia (i.e. with normal sperm parameter values) if he has the following semen characteristics:

- sperm concentration equal or above 15 million/ml;
- sperm progressive motility with at least 32%;
- sperm morphology with at least 4%.

Based on these thresholds, sperm parameter values were in this study assessed quantitatively and qualitatively as follows:

A. Quantitatively:

A.1) Sperm parameter values;

A.2) Number of altered sperm parameters (encompasses missing values).

B. Qualitatively:

B.1) Semen classification (see table 4.1 for definitions)

B.2) Sperm category (i.e. normality/abnormality of the sperm parameter values).

In the below Table 4.1, we present and describe the semen classifications. Hence, in the column called "Semen Classification", we present the name of the classifications used to tackle the assessment B.1, the column called "Semen characteristics", indicates the sperm parameter values that the semen analysis report must deliver to classify a semen into the corresponding semen classification, and in the column called "Number of altered sperm parameters", we specify the number of altered sperm parameter values in each semen classification.

Table 4.1 Description of Semen classifications

Semen Classification	Semen characteristics	Number of altered sperm parameters
Normozoospermia	Sperm concentration => 15 million/mL Sperm progressive motility => 32% Sperm morphology => 4%	0
Oligozoospermia	Sperm concentration < 15 million/mL Sperm progressive motility => 32% Sperm morphology => 4%	1
OligoAsthenozoospermia	Sperm concentration < 15 million/mL Sperm progressive motility < 32% Sperm morphology => 4%	2
OligoTeratozoospermia	Sperm concentration < 15 million/mL Sperm progressive motility => 32% Sperm morphology < 4%	2
Asthenozoospermia	Sperm concentration => 15 million/mL Sperm progressive motility < 32% Sperm morphology => 4%	1
AsthenoTeratozoospermia	Sperm concentration => 15 million/mL Sperm progressive motility < 32% Sperm morphology < 4%	2
Teratozoospermia	Sperm concentration => 15 million/mL Sperm progressive motility => 32% Sperm morphology < 4%	1
OligoAstenoTeratozoospermia	Sperm concentration 15 < million/mL Sperm progressive motility < 32% Sperm morphology < 4%	3
Azoospermia	Sperm concentration = 0 million/mL Sperm motility does not exist Sperm morphology does not exist	1

4.1.2 Data set

Data analysis was carried out upon a preprocessed data set of 293 heterosexual infertile couples (i.e. couples that were unable to get pregnant after 1 year of regular intercourse) - arising from the 320 provided instances - described throughout 64 preprocessed features (i.e. attributes), where the male partner had undergone an embolization treatment for varicocele correction between January 2007 and April 2016.

Regarding patient age, that male patients was between 23 and 54 years old, and their female partner, between 20 and 46 years old at the time of embolization treatment.

All male partners had undergone a semen analysis test before and at 3, 6, 12 months after the embolization treatment, with a previous sexual abstinence of 3 to 5 days. Furthermore, couples were followed in fertility appointments where some of the female partners have undergone ART procedures such as intrauterine insemination (IUI), *in vitro* fertilization (IVF), intracytoplasmic sperm injection (ICSI) or intracytoplasmic morphologically selected sperm injection (IMSI). However, some couples were able to achieve pregnancy spontaneously. Moreover, all follow up tests were performed in the CHUC and most patients were from Portuguese origin.

In the following subsection 4.1.2.1, we describe the assessed 64 attributes.

4.1.2.1 Attributes´ Description

The description of an attribute not only encompasses its definition but also specifies its attribute type. Attribute type is determined by the set of possible values that an attribute can have. Hence, we have qualitative (usually words) and quantitative (numbers) attributes (Barbara Ilowsky; Susan Dean, 2017): we can say that any kind of value from which a mean can be computed, is a quantitative attribute, otherwise, it is a qualitative attribute. Moreover, qualitative attributes can also be called ordinal attributes, if there is a meaningful order among its values; or binary, if the attribute can only have two nominal values (i.e. states) (Han et al., 2012). In the below Table 4.2 we describe these different types of attributes organized by its qualitative and quantitative category.

Table 4.2 Description of attribute types

Category	Type	Description	Example
Qualitative Attributes	Nominal (No)	Values that are symbols or names of things that do not have any meaningful order or a maximum of two diferent values. Note: Numeric labels are nominal values.	Marital status, Occupation, Medical test names.
	Binary (Bi)	Nominal values with only two states.	Gender, HIV medical test outcome.
	Ordinal (Or)	Nominal values that have a meaningful order or ranking among them.	Size, Professional Category, Disease severity
Quantitative Attributes	Numeric Discrete (ND)	A finite or countably infinite value.	Number of Babies
	Numeric Continuous (NC)	A measurable quantity typically represented as a floating-point value.	Spermatozoa concentration, dates (i.e. birth date, treatment dates etc.)

Regarding the dimensionality of the preprocessed data set, in Table 4.3, the 39 initially provided attributes are referenced with the ID 1 to 39 and the generated 25 attributes, referenced with the ID 40 to 64.

Table 4.3 can be interpreted as follows: the column named “ID”, presents the id of the attribute; the column named “Based on ID”, indicates the id of the attributes upon which the recorded value in the corresponding attribute was generated from; the column named “Attribute name”, presents the name of the attribute; the column named “Attribute code name”, indicates the code of the attribute used in the RapidMiner; the column named “Description”, describes the information that the attribute records; the column named “Type”, specifies the attribute type of the corresponding attribute with the several abbreviations specified in the previous Table 4.2 and at last, the column named “Attribute value”, discloses the range of values (for numeric continuous attributes), the values (for binominal or ordinal attributes) or some values (for nominal attributes) recorded in the corresponding attribute. For layout purposes, we have renamed the attribute code name called “ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos” and referenced with the ID 44, into the “ProfissãoComRiscoDeContacto” code name.

Table 4.3 Attribute ´s description

ID	Based on ID	Attribute name	Attribute code name	Description	Type	Attribute value
1		Man age	Idade_H	Age of the male patient at embolization time	NC	23-54
2		Woman age	Idade_M	Age of the patient´s partner at embolization	NC	20-46
3		Infertility time	Tempo_Infert	Months the couple have been trying to conceive	ND	4-192
4		Type of infertility	Prim_Sec	Patient´s partner first or second pregnancy	Bi	Primary, Secondary
5		Woman infertility factor	Factor_Infertilidade_Feminino	Patient´s partner diagnosed infertility cause	No	Anovulation
6		Man infertility factor	Factor_Infertilidade_Masculino	Male patient diagnosed infertility cause	No	Azoospermia, OAT
7		Smoking habit	HabitosTabagicos	Male patient smoking habits	No	4 cigarettes per day
8		Drinking habit	HabitosAlcoolicos	Male patient drinking habits	No	Socially, Rarely
9		Surgeries	Cirurgias	Male patient surgeries before treatment	No	Hernioplasty
10		Diseases	Doença	Male patient diseases before treatment	No	Left Epididymis cyst
11		Occupation	Profissao	Male patient occupation before treatment	No	Factory worker
12		Severity grade	Grau_Varicoc	varicocele severity grade before treatment	Or	I, II, III
13		Laterality	Lateralidade	Scrotum site of the varicocele condition	No	Left, Right, Both
14		Testis volume	Volume_Testiculo_Médico	Categorization of the patient´s testis volume	No	Above 20cc, Normal
15		Embolization date	Data_Embolização	Date of the embolization treatment	NC	01/17/2007-04/28/2016
16		Embolized laterality	TratamentoFeito_lateralidade	Treated scrotum laterality	No	Left, Right, Both
17		Material of Embolization	TratamentoFeito_material	Material used during the treatment	No	Coils, Glue
18		Complications	Complicações	Complications after the embolization treatment	No	None, Pain
19		Repeat embolization	Repetia_embolização	Whether the patient would repeat the treatment	No	Unknown, Yes, No
20		Reason to not repeat	Razão_não_repetir	Reason told for not repeating the treatment	No	Unknown, Pain
21		Concentration before treatment	Conc_Pre	Concentration of spermatozoa before	NC	0-220
22		Concentration at 3 months	Conc_3M	Concentration of spermatozoa at 3 months	NC	0-170
23		Concentration at 6 months	Conc_6M	Concentration of spermatozoa at 6 months	NC	0-160
24		Concentration at 12 months	Conc_1A	Concentration of spermatozoa at 12 months	NC	0-80
25		Progressive motility before treatment	A_B_pré	Percentage of fast/slow spermatozoa before	NC	0-89
26		Progressive motility at 3 months	A_B_3M	Percentage of fast/slow spermatozoa at 3 months	NC	0-94
27		Progressive motility at 6 months	A_B_6M	Percentage of fast/slow spermatozoa at 6 months	NC	0-83
28		Progressive motility at 12 months	A_B_1A	Percentage of fast/slow spermatozoa at 12 months	NC	0-83
29		Morphology before treatment	Formas_N_pré	Percentage of normal spermatozoa before	NC	0-38
30		Morphology at 3 months	Formas_N_3M	Percentage of normal spermatozoa at 3 months	NC	0-21
31		Morphology at 6 months	Formas_N_6M	Percentage of normal spermatozoa at 6 months	NC	0-21
32		Morphology at 12 months	Formas_N_1A	Percentage of normal spermatozoa at 12 months	NC	1-10
33		Pregnancy outcome	Gravidez	Couple got or not pregnant after embolization	Bi	No, Yes

ID	Based on ID	Attribute name	Attribute code name	Description	Type	Attribute value
34		Number of pregnancies	Num_Gravidezes	Number of pregnancies had after embolization	ND	0-3
35		Birth	Nascimento	Couple got or not a birth after embolization	Bi	No, Yes
36		Number of alive babies	Num_Bebés	Number of alive babies born after embolization	ND	0-3
37		Time took to conceive	Gravidez_pós_emb	Number of months after embolization	ND	0-79
38		ART	PMA	Patient´s partner got pregnant with ART	Bi	No, Yes
39		Spontaneous pregnancy	Gravidez_espontanea	Patient´s partner got pregnant spontaneously	Bi	No, Yes
40	7	Preprocessed smoking habit	HabitosTabagicos_Processado_Simplificado	Male patient smokes or not	Bi	No, Yes
41	8	Preprocessed drinking habit	HabitosAlcoolicos_Processado_Simplificado	Male patient drinks or not	Bi	No, Yes
42	9	Preprocessed surgeries	Cirurgias_Processado_Simplificado	Male patient got surgeries before treatment	Bi	No, Yes
43	10	Preprocessed diseases	DoençaSimplificada	Male patient got diseases before treatment	No	Epididymis
44	11	Hazardous occupation	ProfissãoComRiscoDeContacto	Male patient works or not in a toxic environment	Bi	No, Yes
45	21; 25; 29	Altered before	Numero_alterações_Pre	Number of altered sperm parameters before	ND	0, 1, 2, 3
46	22; 26; 30	Altered at 3 months	Numero_alterações_3M	Number of altered sperm parameters at 3 months	ND	0, 1, 2, 3
47	23; 27; 31	Altered at 6 months	Numero_alterações_6M	Number of altered sperm parameters at 6 months	ND	0, 1, 2, 3
48	24; 28; 32	Altered at 12 months	Numero_alterações_1A	Number of altered sperm parameters at 12 months	ND	0, 1, 2, 3
49	21; 25; 29	Semen classification before treatment	Qualificar_Espermograma_Pre	Semen classification before treatment	No	OAT
50	22; 26; 30	Semen classification at 3 months	Qualificar_Espermograma_3M	Semen classification 3 months after treatment	No	Normozoospermia
51	23; 27; 31	Semen classification at 6 months	Qualificar_Espermograma_6M	Semen classification 6 months after treatment	No	Azoospermia
52	24; 28; 32	Semen classification at 12 months	Qualificar_Espermograma_1A	Semen classification 12 months after treatment	No	Azoospermia
53	21	Concentration category before treatment	Conc_Pre_Qualificado	Normality of the concentration value before	Bi	Abnormal, Normal
54	22	Concentration category at 3 months	Conc_3M_Qualificado	Normality of the concentration value at 3 months	Bi	Abnormal, Normal
55	23	Concentration category at 6 months	Conc_6M_Qualificado	Normality of the concentration value at 6 months	Bi	Abnormal, Normal
56	24	Concentration category at 12 months	Conc_1A_Qualificado	Normality of the concentration value at 12 months	Bi	Abnormal, Normal
57	25	Progressive motility category before	A_B_Pre_Qualificado	Normality of the motility value before	Bi	Abnormal, Normal
58	26	Progressive motility category at 3 months	A_B_3M_Qualificado	Normality of the motility value at 3 months	Bi	Abnormal, Normal
59	27	Progressive motility category at 6 months	A_B_6M_Qualificado	Normality of the motility value at 6 months	Bi	Abnormal, Normal
60	28	Progressive motility category at 12 months	A_B_1A_Qualificado	Normality of the motility value at 12 months	Bi	Abnormal, Normal
61	29	Morphology category before treatment	Formas_N_Pre_Qualificado	Normality of the morphology value before	Bi	Abnormal, Normal
62	30	Morphology category at 3 months	Formas_N_3M_Qualificado	Normality of the morphology value at 3 months	Bi	Abnormal, Normal
63	31	Morphology category at 6 months	Formas_N_6M_Qualificado	Normality of the morphology value at 6 months	Bi	Abnormal, Normal
64	32	Morphology category at 12 months	Formas_N_1A_Qualificado	Normality of the morphology value at 12 months	Bi	Abnormal, Normal

If we analyze the generated attributes with the id 40 to 64, we see that their aim was to simplify the information of the original attributes. Regarding the hazardous occupation attribute, we have classified the man occupation recorded in the occupation attribute into whether its occupation is in contact with putative toxic environments or products, as stated by the international labour organization for hazardous occupations (international labour organization, n.d.). In fact, the international labour organization state that the most hazardous occupations are related to agriculture, construction, mining and ship-breaking, as well as any occupation exposed to chemical substances or radiation, to name a few.

By considering the attributes described in the above Table 4.3 we can say, based on the related works presented in 3.3 and 3.4, that on the matter of attribute selection, the preprocessed data set has the needed information to tackle the aims of this project. In fact, it has sperm parameter data before and after the treatment to see if the embolization treatment improves said parameters, it has several attribute candidates for the prediction of embolization success and it encompasses external factors such as surgeries, diseases, occupation, drinking, smoking habits and more, which can enable us to find interesting insights on eventual data patterns.

4.1.3 Tools

In terms of hardware, this study was carried out with one computer with the following characteristics:

- Operative system: Microsoft Windows 10 Home, 64 bits
- CPU: Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz
- RAM: 16 GB

In terms of software, this study used the software tools that we below specify for each purpose of use in the carried-out work:

- Data collection and preparation: Microsoft Excel 2016, Home and Student Edition installed in Portuguese.
- Data integration: Microsoft SQL Server Management Studio 2012.
- Data Analysis (Statistical & Mining): RapidMiner Studio Educational platform, version 8.1.001.

RapidMiner was the data mining tool that was selected since related works and consulting companies (as seen in subsection 3.1) has elected it as the best free/open source tool to apply data mining techniques. Since there is no open source (i.e. non-commercial and free) tools that fits all needs and the aim of this study is the application of data mining techniques, the need for a good data mining tool outweighed in this study the need of a tool that implements statistical tests. In fact, statistical tests are applied to better understand the preprocessed data and highlight possible data patterns that are further on explored with data mining techniques. Hence, we have only explored and applied statistical tests that the Rapid Miner had and that were applied in related works.

RapidMiner is a data science software platform developed by the company of the same name. It was formerly known as YALE (Yet Another Learning Environment) and was developed at the Artificial Intelligence Unit of the Technical University of Dortmund, Germany, that has its initial release in 2006. This software platform as a Free Edition that can be used on data sets up to 10 000 rows with a limit of 1 logical processor that is distributed under the AGPL license. AGPL is a license that can be attributed to open source software that can be run over a network.

RapidMiner is written in the Java programming language and provides a GUI to design and execute analytical processes. These processes are built with the mean of drag and drop components that can apply data transformation tasks, descriptive statistical tests and data mining algorithms to the data. These components are in RapidMiner called “operators” and the connection of several operators is called visual composition framework (VCF). Each operator has several parameters that can be configured and always has input and output ports to respectively receive the data from the previous operator and send it to the next operator.

4.2 Methods

This study used several methods that are described below. Hence, in section 4.2.1, we specify how the assessment of the quality of the initially provided and selected data - presented in the appendix A - was carried out, and in section 4.2.2, how this data was preprocessed to achieve the first 39 collected attributes described in Table 4.3. Next, in section 4.2.3, we specify how the 25 generates attributes were produced and in the following section 4.2.4, how and which label attribute was mainly selected to tackle our predictive goals. Furthermore, in section 4.2.5, we describe some of the statistical measures that this study has used and in the following section 4.2.6, we briefly specify how the selected statistical tests were applied. Further on, in section 4.2.7, we describe the data mining algorithms that were selected to tackle the aim of this work and afterwards, in section 4.2.8 we specify how these data mining algorithms were trained/tested and evaluated, and at last, in section 4.2.9, we provide a glance on how data modeling was carried out.

4.2.1 Data Quality Assessment

Before any data analysis, analyzing whether the provided data is trustworthy to respond to the data mining goals is fundamental (Thatipamula, 2013) (Maydanchik, 2007). Therefore, the data was checked through the assessment of a set of *key data dimensions* before analyzing it with statistical and/or data mining techniques. Hence, this study checked the initially provided and selected attributes disclosed in Appendix A with the *key data dimensions* described in Table 4.4 to know whether we could pursue our data analysis with or without preparing it (i.e. with or without data preparation). As shown in Table 4.4, after describing each *key data dimension* under the column named “Definition” based on the author Thatipamula (2013), we have specified how they were measured under the column named “Score formula” based on the author Maydanchik (2007). At last, under the column named “Assessment”, we specify how the *key data dimensions* were assessed.

By analyzing the “Score formula” column, we see that only the *Completeness* and the *Accuracy* dimension is specified since only these two dimensions were measured. In fact, the CRISP-DM methodology (Chapman et al., 2000) and one of the main authors in the data quality domain, Maydanchik (2007), supports this option. If we analyze the “Assessment” column, we see that the *Completeness*, the *Consistency* and the *Conformity* dimensions were assessed by *verifying* whether the provided data set complied with the data characteristics specified under the key dimension´s definition. However, the *Accuracy* and the *Integrity* dimensions were assessed by *validating* if the provided data was the same with the ones recorded in the patient´s medical dossiers and information technology systems gathered/installed at CHUC and if the data values were coherent with other data values of the dataset.

Table 4.4 Definition of the key data dimensions

Key data dimension	Definition	Score formula	Assessment
Completeness	Having all the attributes needed, in a usable sate (e.g. not having the male and female infertility risk factors mixed in one attribute) and filled (i.e. without missing values), to tackle the Data mining goals that were initially set.	((Number of total instances - Number of missing values) / Number of total instances)*100	Verification
Consistency	Showing data coherence (e.g. if a patient got an live baby, its related "Gravidez" attribute as to be set to yes "Sim"). Furthermore, the dataset must not have duplicated instances.		Verification
Conformity	Showing that data comply with a specific format in all instances (e.g. all embolization dates are with the formal dd/mm/yyyy)		Verification
Accuracy	Having the correct data	((Number of total instances - Number of erroneous values) / Number of total instances)*100	Validation
Integrity	Having the correct data linkage - the data can be traced and connected to other data correctly (e.g. the information of the patient´s partner is in fact related with its correct patient´s partner)		Validation

4.2.2 Data Preparation

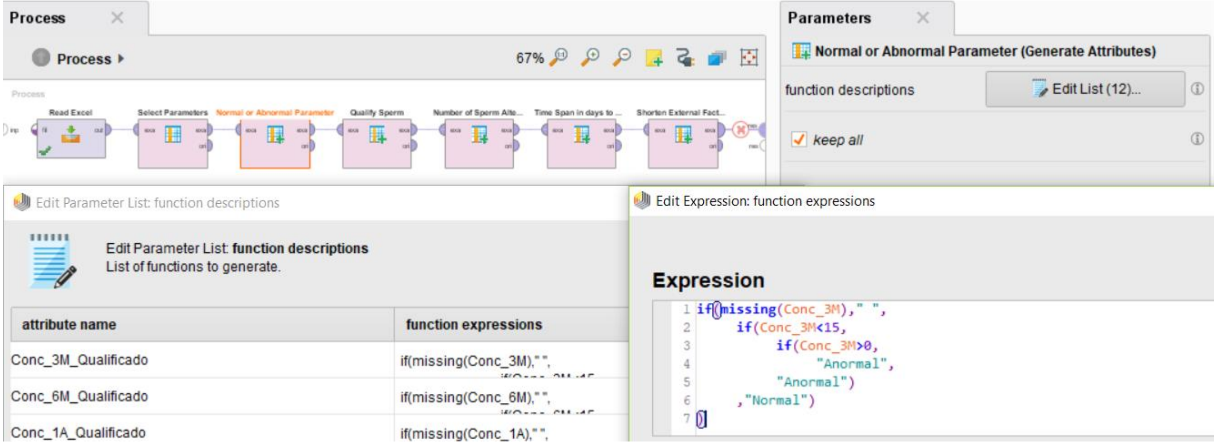
After assessing the initially provided and selected data set with the data quality dimensions described in Table 4.4, we have seen that the data needed to be preprocessed to achieve the Data mining goals that were set. Thus, this study carried out the data preparation tasks disclosed in Table 4.5 to produce the *initially preprocessed data set*. The order of their execution is roughly specified under the column named "Order". We say roughly, because after the *data integration* task more *data cleaning* and *reformatting* was also carried out. The *final preprocessed data set* was obtained after generating new attributes upon the *initially preprocessed data set*.

Table 4.5 Carried out data preparation tasks

Order	Data Preparation Tasks
1	Data Construction
2	Data Reorganization
3	Data Cleaning
4	Data Format
5	Data Integration
6	Attributes selection for Data mining purposes

4.2.3 Attribute Generation

To better detect some possible data patterns and optimize data mining results, sperm and semen analysis results were qualified, as previously indicated, and patient external factors were simplified (e.g. smokes 4 cigarettes per day, to only the word “Yes”). Most of these data transformations were carried out in the RapidMiner platform with the “Generate Attribute” operator. This operator has the ability to, based on a code wrote by the data analyst (called in RapidMiner “Expression”), generate a new attribute with the result of the devised expression. Figure 4.1 presents the RapidMiner´s interface at attribute generation time which can be interpreted as follows: at the top left, we have the built process to generate new attributes; at the top-right, the settings of the selected operator that is shown in the built process highlighted in orange and called “Normal or Abnormal parameter” (this operator implements the qualitative semen analysis identified in 4.1.1.1 as B.2); at the bottom left, some of the generated attributes from the selected operator; and at bottom right, one of the several expressions devised, which in this case is the expression devised to qualify sperm concentrations at 3 months. The expression presented in Figure 4.1 generates the information specified in the previous subsection 4.1.1.1 as B.2 which can be translated as follows: “if the sperm concentration at 3 months is missing, write nothing. Otherwise, look if the sperm concentration is below 15 million/ml, and if so, check if it is above 0 million/ml, if so, write *Abnormal* and otherwise, also write *Abnormal* for the cases that will be equal to 0 - the data set does not have negative values. If the sperm concentration is equal or above 15 million/ml, then write *Normal*”. All other generated attributes were in this process similarly coded, in exception to the attribute that simplifies the occupation attribute.



Normal or Abnormal Parameter (Generate Attributes)

function descriptions: Edit List (12)...

☑ keep all

Edit Parameter List: **function descriptions**
List of functions to generate.

attribute name	function expressions
Conc_3M_Qualificado	if(missing(Conc_3M)," ",
Conc_6M_Qualificado	if(missing(Conc_6M)," ",
Conc_1A_Qualificado	if(missing(Conc_1A)," ",

```
if(missing(Conc_3M)," ",
    if(Conc_3M<15,
        if(Conc_3M>0,
            "Anormal",
        "Anormal")
    ,"Normal")
)
```

Figure 4.1 Generate Attributes with the RapidMiner platform.

4.2.4 Label Selection

The final preprocessed data set has a set of attributes that could be used as labels for *Classification* tasks. However, not all these labels were seen good classifiers due to their frequency of possible and missing values. Therefore, this subsection presents the analysis upon these data characteristics, as well as specifies whether these label attributes provide a balanced data set based on the number of instances they can classify. For this purpose, Table 4.6 was built to better compare the characteristics of the possible label attributes referenced with its id. If we analyze this table, we see that the "Pregnancy outcome" attribute seems to be the best label candidate since it does not have a lot of missing values, in comparison to others, and it provides a quite balanced data set which enables to train the positive and the negative instances equally. Furthermore, this attribute was already seen used in related works to assess the impact of the embolization treatment, as well as the live births. If we check this last attribute, we can see that the "Number of alive babies" attribute has too many missing values in our data set to consider it as a label. Regarding the other possible label attributes, we see that the ones without missing values are far from being balanced. Therefore, the "Pregnancy outcome" was the elected label attribute. Please note that the "Pregnancy outcome" value that was set as the most important, was the value/class "Yes" since this study aims to predict the success of the embolization through the male patient´s partner ability to conceive.

Table 4.6 Analysis of label characteristics

ID	Possible Lables	Possible Values	Missing	Balanced Data Set?
33	Pregnancy outcome	Yes (123); No (107)	64	Quite YES
34	Number of pregnancies	{0 (187); 1 (84); 2 (19); 3 (4)}	0	NO
35	Birth	No (146); Yes (84)	64	NO
36	Number of alive babies	{0 (3); 1 (57); 2 (22); 3 (2)}	210	NO
37	Time to conceive	[0,79]	189	NO
38	ART	No (227); Yes (66)	0	NO
39	Spontaneous pregnancy	No (181); Yes (49)	64	NO

4.2.5 Statistical Measures

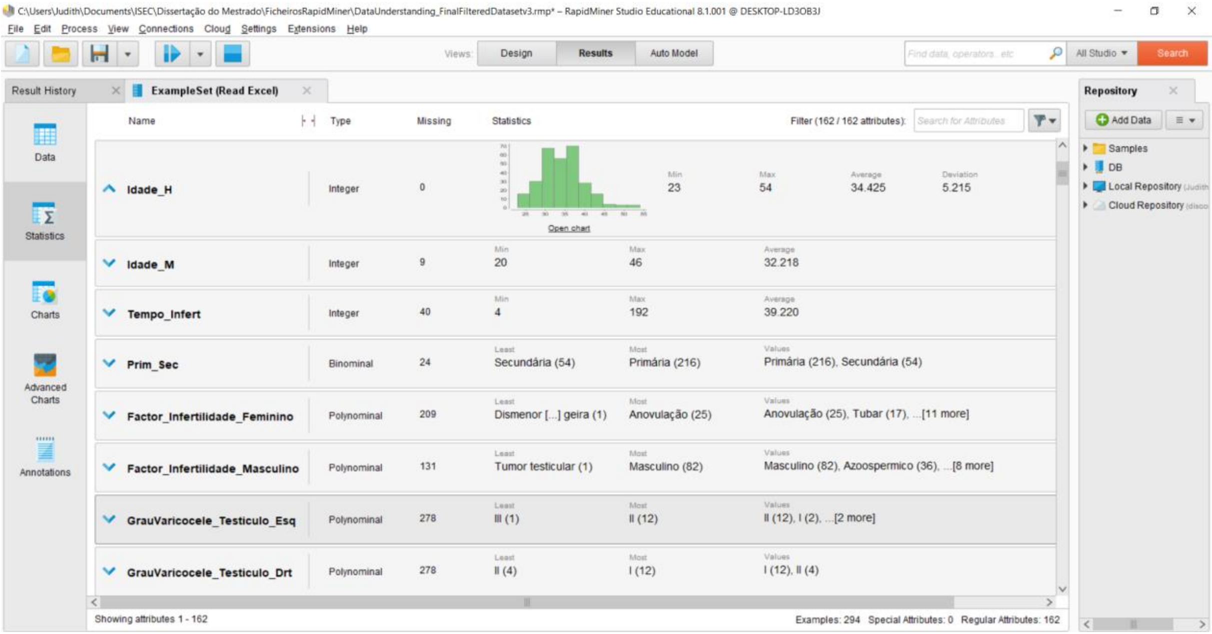
In DeWitt *et al*. (2018) the statistical measures used were the mean and the standard deviation. However, the data mining authors Han *et al*. (2012) asserts that to preprocess the data successfully, it is essential to have an overall picture of the data with a wider statistical description. The basic statistical description suggested in Han *et al*. (2012) encompasses the analysis of the central tendency of the data (i.e. the Mean, Median and Mode of each attribute), as well as its dispersion (i.e. the minimum value (Min), the value of the first quartile (Q_1), the value of the third quartile (Q_3), the maximum value (Max) and the standard deviation (SD) of each attribute). Hence, this study has computed all these statistical measures to statistically understand its quantitative data. The Min, Max, Mean (in RapidMiner called Average) and SD (in RapidMiner called Deviation) was with the RapidMiner platform computed. In Figure 4.2, we can see a print screen of the RapidMiner platform where in the center of the figure, part of the descriptive statistics computed by this tool for this study, can be seen. However, in some situations, such as generating graphs in EXCEL, we have also computed these measures with the Portuguese edition of the Microsoft Excel software. In Table 4.7 we present the definition of each statistical measure calculated in this study and specify the EXCEL formula used to compute them.

Table 4.7 Measures´ Definition

Measure Name	Definition (Barbara Ilowsky; Susan Dean, 2017)	Excel Formula Used
Mean	A number that measures the central tendency of data. This measure is obtained by the sum of all values in the sample divided by the number of values in the sample.	AVERAGE()
Median	A number that separates ordered data into halves and corresponds to the second quartile value (Q_2) of a sample.	MEDIAN()
Mode	The value that appears most frequent in a set of data.	MODE()
Min	The minimum value of a sample.	MIN()
Q_1	The Median of the values below the Median of the sample. This value indicates that 25% of the values of a sample is below Q_1.	QUARTILE(;1)
Q_3	The Median of the values above the Median of the sample. This value indicates that 25% of the values of a sample is above Q_3.	QUARTILE(;3)
Max	The maximum value of a sample.	MAX()
SD	A number that is equal to the square root of the sum of the squares of the distance between each value of a sample to their mean divided by the difference of the sample size and one. This measure tells how far the data values are in average from their mean.	STDEV()

As we know, qualitative attributes do not have statistical measures as do quantitative attributes. Thus, to have an overall picture of the values of a qualitative attribute we have identified the possible values, as well as the least and the most frequent values of each qualitative attribute. In fact, in the last 5 rows of the center table depicted in Figure 4.2, we can see that the RapidMiner platform presents this information including their frequency (in parentheses). Therefore, the possible values, as well as the least and the most frequent values of each qualitative attribute and its frequency were retrieved from RapidMiner.

The number of filled values per attribute (indicated in this study with the name "*Filled*") is a measure that is also presented. In fact, in the center of Figure 4.2, we can see the descriptive statistics computed by the RapidMiner platform where we see under the column name "Missing", the number of missing values for each attribute. The *Filled* measure was calculated by doing the number of total patients of the data set, which is 293 patients, minus the missing values indicated in the RapidMiner. Thereby, the *Filled* value for the attribute that records the age of each male patient partner, indicated with the attribute called "Idade_M", is 293 patients minus the 9 missing values indicated in the RapidMiner which gives 285 *Filled* records.


C:\Users\Judith\Documents\ISEC\Dissertação do Mestrado\FicheirosRapidMiner\DataUnderstanding_FinalFilteredDatasetv3.rmp* – RapidMiner Studio Educational 8.1.001 @ DESKTOP-LD3OB3J
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Auto Model
Find data, operators...etc
All Studio
Search
Result History
ExampleSet (Read Excel)
Data
Statistics
Charts
Advanced Charts
Annotations
Filter (162 / 162 attributes):
Search for Attributes
Open chart
Repository
Add Data
Samples
DB
Local Repository
Cloud Repository

Name	Type	Missing	Statistics			
Idade_H	Integer	0	Min 23	Max 54	Average 34.425	Deviation 5.215
Idade_M	Integer	9	Min 20	Max 46	Average 32.218	
Tempo_Infert	Integer	40	Min 4	Max 192	Average 39.220	
Prim_Sec	Binominal	24	Least Secundária (54)	Most Primária (216)	Values Primária (216), Secundária (54)	
Factor_Infertilidade_Feminino	Polynominal	209	Least Dismenor [...] geira (1)	Most Anovulação (25)	Values Anovulação (25), Tubar (17), ...[11 more]	
Factor_Infertilidade_Masculino	Polynominal	131	Least Tumor testicular (1)	Most Masculino (82)	Values Masculino (82), Azoospermico (36), ...[8 more]	
GrauVaricocele_Testiculo_Esq	Polynominal	278	Least III (1)	Most II (12)	Values II (12), I (2), ...[2 more]	
GrauVaricocele_Testiculo_Drt	Polynominal	278	Least II (4)	Most I (12)	Values I (12), II (4)	

Showing attributes 1 - 162

Examples: 294 Special Attributes: 0 Regular Attributes: 162

Figure 4.2 RapidMiner´s Descriptive Statistics

4.2.6 Statistical Tests

Most previously disclosed varicocele-related works applied the Chi-square and the ANOVA test upon sperm parameter values for *statistical inference*. Hence, these statistical tests were also applied in this work to identify possible data patterns to guide us through the construction of the data mining models.

How these statistical tests were applied is briefly discussed in the following subsections. Hence, subsection 4.2.6.1, discusses the Chi-square test and subsection 4.2.6.2, discusses the One-way ANOVA test. Further on, other statistical tests were also explored and tested in the RapidMiner platform. Hence, subsection 4.2.6.3, covers the Kolmogorov-Smirnov test and subsection 4.2.6.4, the Pearson Correlation test.

4.2.6.1 Chi-square

The Chi-square test, or $\boldsymbol{x^2}$ test, is a nonparametric statistical test applied to, among other applications, assess if two nominal attributes, *A* and *B*, are independent through the assessment of its relative frequencies. The null hypothesis for this statistical test is that *A* and *B* are independent and the level of significance used was *p*=0.05. However, in the context of the assessment of the computed association rules, we have raised the significance level to *p*=0.10 to complement the *lift* measure with the aim of ascertain with a greater precision the relation between the antecedent and the consequent attribute of a rule.

4.2.6.2 One-Way ANOVA

The purpose of the One-Way ANOVA parametric test, is to determine the existence of a statistically significant difference among several group means through variance calculation in order to assess if their mean differences are not random variations (Barbara Ilowsky; Susan Dean, 2017). The null hypothesis is that all group population means are equal and the level of significance used was *p*=0.05, as applied in varicocele-related works.

4.2.6.3 Kolmogorov-Smirnov

In this study, the Kolmogorov Smirnov Test operator was applied to sperm parameter values to assess if the sperm parameters values of patients that got pregnant had the same data distribution than the ones that did not got pregnant to identify data distribution differences.

The confidence level was in this study set to its default value which is 0.05 and its null hypothesis was defined as whether the mean of the population from which example sets are drawn are equal. This Kolmogorov Smirnov operator returns true if the null hypothesis can be rejected (i.e if the null hypothesis is rejected it means that both example sets are different in shape since their population´s mean is different; and therefore, might show a statistical significance when p is less than 0.05).

4.2.6.4 Pearson Correlation

The Pearson correlation, also called correlation coefficient, is a measure that indicates through a coefficient identified as *r*, how the linear correlation is between two attributes (Barbara Ilowsky; Susan Dean, 2017). This *r* coefficient can be interpreted as follows:

- *r* > 0.9 -> Very high correlation
- 0.7 < *r* < 0.9 -> High
- 0.5 < *r* < 0.7 -> Moderate
- 0.3 < r <0.5 -> Low
- 0.0 to 0.3 -> Despicable

This measure was used because it is widely suggested in the data mining field (Han et al., 2012) to identify the attributes to model. Hence, in the Rapid Miner platform, we have applied it to quantitative attributes.

4.2.7 Data mining algorithms

As we have seen in Table 3.1, several data mining algorithms have already been applied in the context of male infertility. Since this study covers similar patient information, we have at first applied the decision tree algorithm due to its acceptable outcome seen in Guh *et al*. (2011) for prediction purposes upon sperm parameter data, and then applied the K-means and Association rules algorithms due to their popularity in the identification of interesting data patterns in the healthcare domain. Hence, in this section, we briefly describe all these algorithms that we have applied upon the *final preprocessed dataset* by covering the following aspects: their main purposes; their type of results; how they work; their specificities and the RapidMiner operator we have used to apply them in the built models. Thus, this section is organized as follows: in section 4.2.7.1, we approach the decision tree algorithm; in section 4.2.7.2, we disclose the K-means algorithm and at last, in section 4.2.7.3, we cover the FP-Growth algorithm.

4.2.7.1 Decision tree

Decision trees are mostly applied to identify the most interesting attributes that one should use to mine and/or to predict the conditional probability of a *label attribute* outcome based on its historical records. The decision tree algorithm C4.5 is the most commonly applied for these main purposes and even in this field, it has already been applied (see section 3.2). In fact, as the authors Witten *et al*. (2011) state, the C4.5 algorithm is probably one of the most popular classifiers.

The final result of this type of algorithm is a tree that begins with a root, ramifies with decision nodes and at last, ends with leaf nodes – analogous to real trees. A decision node has two or more branches that reflect the attribute values of the ramified node/attribute and the leaf node represents a classification or decision based on the selected label attribute ("Decision Tree," n.d.). Hence, these leaf nodes present a label attribute value for each tree branch.

Classifiers as the decision tree algorithm C4.5, began with the attribute that promotes the highest gain of information by placing it at its root and its ramification is guided with the *entropy* measure. In fact, the C4.5 algorithm aims to decrease the *entropy* through the downward splitting of the nodes; and hence, choose as attribute nodes, the one that produces the purest daughter nodes (i.e. *entropy* equal to zero) to compute the smallest tree as soon as possible (Witten et al., 2011). Therefore, the C4.5 algorithm works as follows: after splitting a node and testing whether the entropy of the next node is lesser than the entropy before splitting and if this value is the least as compared to all possible test-cases for splitting, then the node is split into its purest constituents (i.e. attribute values). This assessment is recursively performed with the remaining attributes until all leaf nodes are pure (i.e. leaf nodes with instances belonging to one class, such as: "Pregnancy outcome"= "Yes" or "Pregnancy outcome"= "No"), or until it is not possible to further on split because the *entropy* is equal to 1. In other words, as Witten *et al*. (2011) states, this algorithm works top-down, seeking at each stage for an attribute that can better split the classes (i.e. yields the highest gain of information at each stage). The gain of information is in the C4.5 algorithm computed with the *Gain ratio* measurement which is an extension of the *information gain* measure used by the ID3 algorithm.

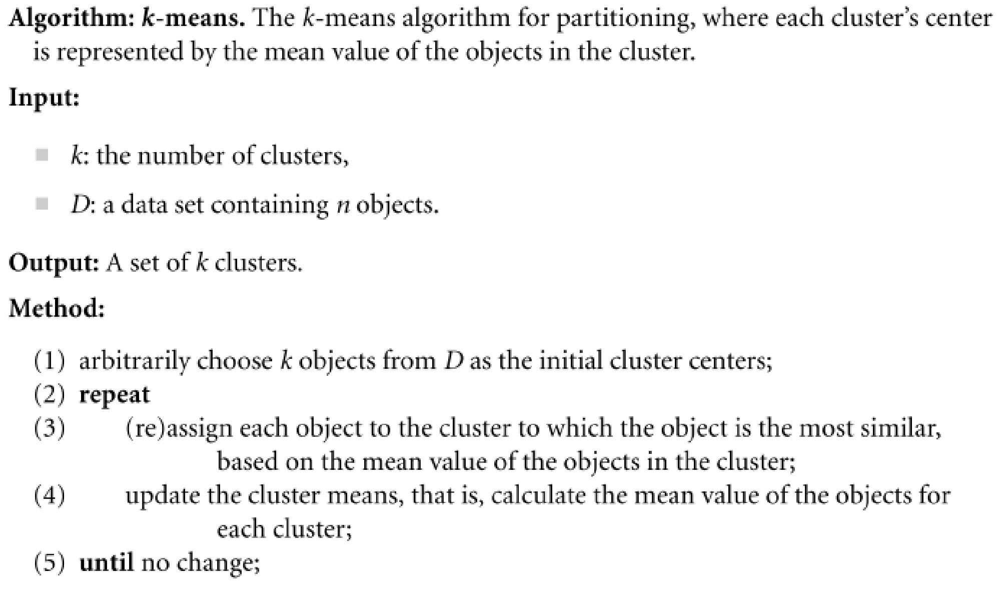
Since decision trees are supervised learners, most decision trees algorithms need a binomial attribute as a label and some implementations, require non-missing values in this special attribute. Hence, prior to the application of the algorithm we need to select a label attribute, transform it as binomial (if it is not yet binomial), filter the rows of the data set by non-missing value on the label attribute, apply some algorithm optimizations (e.g., other algorithms, attribute discretization, attribute normalization etc.), train the decision tree algorithm with different algorithm configurations (i.e. different parameter values) and test the algorithm.

Decision trees have the possibility to be pruned. Note that pruning refers to the removal of those branches in the decision tree which do not contribute significantly to our decision process. In the RapidMiner platform, pruning can be applied or not into the generated decision trees. In fact, pruning comes in the RapidMiner platform as a setting option. When we have a small data, such as is the case, it is important to activate that option; and hence, prune the computed decision trees because they tend to overfit (i.e. to adapt to the dataset). The C4.5 algorithm can perform the highest percentage of pruning which can go up to 10%.

Regarding the decision tree algorithms that are available in the RapidMiner platform, we have seen that the "Decision tree" operator was an own implementation and that the "W-J48" operator, which is an operator from the Weka data mining platform, is the free java implementation of the C4.5 algorithm. The senior community manager Scott Genzer at RapidMiner has stated in a post that the “Decision tree” operator was by far better; hence, both decision tree algorithms were applied in this study upon the *final preprocessed data set* to explore its statement. The W-J48 operator was get through the installation of the Weka extension 7.3.000 into our RapidMiner environment.

4.2.7.2 K-Means

K-means is a commonly used data mining algorithm for the application of *clustering* techniques; and therefore, its aim is to partition a data set into k groups of instances (called clusters). Its partitioning is performed with an agglomerative/partitional technique and not a hierarchical one, so its results are not expressed into a dendrogram. Hence, its results can be interpreted through its *centroid table* and its related plots. In the RapidMiner platform, the identified clusters are colored in different colors to enable us to conclude that the instances (i.e. patients) that are in the same cluster are similar to each other on the attributes assessed because they are all closer to the same cluster´s central point, called centroid. In the k-means algorithm this centroid is defined as the mean value of the points within the cluster so the centroid is not a value from the data set. Hence, the k-means algorithm mines numeric attributes and calculates the distance between the instance values and the centroid to decide in which cluster the instance should be part of. In Figure 4.3, we present the K-mean algorithm shown in Han *et al*. (2012) to provide a clearer understanding on the way this algorithm works.



Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

```
(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) repeat
(3)     (re)assign each object to the cluster to which the object is the most similar,
            based on the mean value of the objects in the cluster;
(4)     update the cluster means, that is, calculate the mean value of the objects for
            each cluster;
(5) until no change;
```

Figure 4.3 K-means algorithm (Han et al., 2012)

The K-mean algorithm works well for finding spherical-shaped clusters in small to medium-size data sets and are good at handling low-dimensional data such as data sets involving only two or three attributes (Han et al., 2012). Since our data set is small, as is the number of selected attributes, we have applied the K-means algorithm to, in the first place, identify data patterns and, more importantly, define how the selected attributes could be *discretized* in the built predictive models to optimize the *classifiers* applied. In fact, we could *discretize,* for instance, the sperm parameters values by WHO thresholds but the question is: “does that discretization really distribute the patients according to success in the treatment?" and the answer is no, because we have patients in the *final preprocessed data set* that had low sperm parameters values, yet was able to conceive. Hence, we ran the K-means algorithm upon each single selected attribute with the "Pregnancy outcome" attribute to see if we could find an interesting partition of the data based on the main aim of this study. In the RapidMiner platform, this algorithm was applied with the operator called "K-Means".

4.2.7.3 FP-Growth

Through our data understanding, we have found that we were more successful at identifying the most correlated attributes with the label attribute with the *Chi-square* test than the *Pearson correlation* test. Hence, our idea to apply a data mining algorithm that would identify common frequent item sets has flourished and made us apply an association rule algorithm not only to find data patterns, but also to find rules that could predict the success of the embolization treatment. In fact, as previously seen, decision trees are "greedy" algorithms that do not provide interesting results in small data sets. For this reason, we have sought which association rule algorithm we could apply to our data set, and have seen that the RapidMiner platform had the association rule algorithm called FP-Growth.

The FP-Growth algorithm aims to find frequent patterns and interesting relationships among the data set attributes. This algorithm is an optimization of the APRIORI algorithm since it has the ability to only perform two scans of the data set to identify the most frequent item sets (i.e. the first scan is to detect the frequency of each attribute and the other one, is to build the FP-Tree). Furthermore, the FP-Growth algorithm can better convey its results due to its several ways of disclosing them (i.e. tabularly, textually and graphically).

The FP_Growth algorithm performs the following ordered steps to generate its rules (Han et al., 2012):

1. Scan the dataset to find the *frequent* single items (i.e. the algorithm begins to read the data set and returns, in this study, the number of instances set to "TRUE" for each attribute of the dataset with at least the set *min_support*).
2. Sort the *frequent* items computed in the previous step by its frequency and in descending order (this list is called the f-list). Then, for each instance, we will have an associated list of frequent ordered attributes based on the computed f-list.
3. Scan the data set again and construct a tree that presents the association between the frequent attributes with the indication of their *support* in each node (this tree is called FP-tree). This FP-tree is constructed based on the *frequent* ordered attributes computed in the previous step for each instance and the *support* indicated along each tree node tells the number of times each path occurs in the data set until the assessed node.
4. Mine the FP-tree to generate *conditional pattern-base*. This mining is performed by partitioning the FP-tree by each attribute that is presented as nodes in the FP-tree. This is why this step is called "divide and conquer".
5. Mine each *conditional pattern-base* recursively to identify *frequent* patterns and at last, formulate the association rules based on the *min-support* and *min-confidence* that

was initially set. In the RapidMiner platform, the formulation of the association rules are not performed by the FP_Growth algorithm but by another algorithm/operator.

The FP_Growth algorithm requires that all input attributes have to be binomial, and to better interpret the result it is a good practice to map the attribute values. In contrast to the K_Means algorithm that cannot accept data sets with empty values, this algorithm can. However, the algorithm does not consider the missing attributes values since this algorithm seeks to count frequencies (i.e. count attribute values set to TRUE). Nevertheless, this point is useful for our type of data set since it has a low number of instances with all attributes filled. Furthermore, this technique is widely used in the bioinformatics field which reinforced its selection to tackle the aim of this work. In the RapidMiner platform, this algorithm was applied with the operator called “FP-Growth”.

4.2.8 Model´s Training and Assessment

As suggested by the CRISP-DM methodology, before building the data mining models, it is recommended to define how the models will be trained, tested and assessed because the built VCFs implement the decisions made on these matters. Hence, before disclosing how the modeling phase was carried out, we here disclose the decisions made on these matters.

As we have previously seen, to tackle the aim set we have applied several data mining algorithms that have their own specificities. Hence, the application of these different algorithms entailed the training of different parameters (i.e. RapidMiner´s operator settings) and the assessment of different performance metrics. Therefore, to better convey how these algorithms were tested, trained and assessed we disclose them grouped by each applied data mining technique as follows: in section 4.2.8.1, we specify how the Decision Tree algorithm was tested, trained and assessed - the Decision Tree algorithm is a supervised learner so we have also tested it through the selected label attribute defined in section 4.2.4; in section 4.2.8.2, we specify how the K-means algorithm was trained and assessed and in section 4.2.8.3, we specify how the FP-Growth algorithm was trained and assessed.

4.2.8.1 Classification

The followed test designs were:

A. Split the data set into 3 parts – 80% for training/testing and 20% for validation, where in the 80% part, 70% is taken to train and the remaining 30%, to test the data set.
B. Split the data set into 2 parts – 70% for training and 30% for testing.

In Guh *et al*. (2011), the followed testing design was the one noted in B since they have divided their data set of 5275 instances into two parts: They have used 80% for training and 20% for testing; and hence, have not validated the model. However, the CRISP-DM methodology (Chapman et al., 2000), the founder of the RapidMiner platform (Mierswa, 2012), as well as the consulting company SimaFore (Deshpande, 2012), suggest the test design disclosed in A to avoid test overfitting which especially occurs with small data sets. Therefore, we have chosen test design A, but to further on compare our results with related works, we have recomputed the best decision tree model by following test design B. This enabled us to assess the generated decision tree upon all preprocessed data set (advantage of the test design B), as well as check the stability of the model throughout the comparison of the different test results.

The main data set splitting performed within the test design A was executed with the “Split data” operator. This split of data was performed with the sampling type called “Stratified” to ensure that we have in each sub-dataset the same number of instances classified as “Sim” and “Não”, although we have a quite balanced data set (107 instances classified as “Sim” and 123 instances classified as “Não”). Regarding the splitting ratio used, we have gone for the ratios that the RapidMiner platform uses in its tutorials which is the exposed 80% for training/testing and 20% for validation. Afterwards, the training/testing data set was further on divided into the training and testing dataset with the following operators, which were also used in the test design B:

A. Split Validation – operator that splits the data set with a single iteration.
B. Cross Validation – operator that performs several split validations.

The “Split Validation” operator split the data set into 70% for training and 30% for testing. Since these sub-datasets can be generated with several types of samplings, the ones that were tested were: *Linear*, *Shuffle* or *Stratified*.

The “Cross Validation” operator internally performs several “Split Validations”. In fact, the “Cross Validation” operator splits the data set into k sub-datasets and keeps one sub-dataset for testing and the remaining ones for training. Next, it recursively selects another sub-dataset for testing and considers the remaining ones for training. This test is done k times (i.e. until all sub-datasets were at least 1 time a testing dataset) and several k values can be tested. In this study, we have tested the model with k=2 to k=4.

The error measures delivered by the cross-validation operator are an average of all computed error measures since this operator generates k models during its testing. This is why we present with each error measure its standard deviations. The cross validation delivers the model applied to all trained/tested data set; and hence, with the test design A, it has returned the model applied with the 80% training/testing part.

Since we have applied 2 decision tree algorithms (i.e. the decision tree from the RapidMiner platform and the W-J48 algorithm), all decision tree models built have executed the 4 testing steps disclosed in Table 4.8:

Table 4.8 Testing steps of Decision tree´s algorithms

Testing Step Number	Task
1	Test the RapidMiner´s Decision tree algorithm within a Split Validation operator.
2	Test the RapidMiner´s Decision tree algorithm within a Cross Validation operator.
3	Test the W-J48 algorithm within a Split Validation operator.
4	Test the W-J48 algorithm within a Cross Validation operator.

The training of the Decision tree algorithm entailed its application on several groups of selected attributes with the variation of the parameters disclosed in Table 4.9 within each testing step disclosed in Table 4.8. Therefore, the Decision tree models were exhaustively trained/tested in order to choose the optimal parameter values; and hence, the best model. The parameters were selected based on the guidelines explained by the founder and principal of the SimaFore company in Deshpande (2012). The variation of the model parameters entailed the execution of 8664 tests per modeling step of the decision tree algorithm - in each modeling step we have carried out 2160 tests for the decision tree algorithm ran within a simple validation, 6480 tests for the decision tree algorithm ran within a cross validation, 6 tests for the J-W48 algorithm ran within a simple validation and 18 tests for the J-W48 algorithm ran within a cross validation. The best computed results are in the Appendix C.1 disclosed.

Table 4.9 Parameters varied through decision tree training

Related Operator	Parameter Name	Tested Values	Parameter Description
Decision tree	Criterion	*Information_gain*; *Gain_ratio*; *Gini_index*; *Accuracy*.	Selects the criterion on which Attributes will be selected for splitting
Decision tree	Minimal size for split	4; 5; 6.	The size of a node is the number of Examples in its subset. Only those nodes whose size is greater than or equal to the *minimal size for split* parameter are split. If we set a minimal number too high, we can end up with leaves that are not exclusive to one class.

Decision tree	Minimal gain	0.100; 0.140;
		0.180; 0.220;
		0.260; 0.300.
			The node is split if its gain is greater than the *minimal gain*. A higher value of *minimal gain* results in fewer splits and thus a smaller tree.

Decision tree	Minimal leaf size	2;3;4;5; 6.	The size of a leaf is the number of Examples in its subset. The tree is generated in such a way that every leaf has at least the *minimal leaf size* number of Examples. Hence, a high minimal leaf size also reduces the tree size.
Decision tree	Maximal depth	20	The tree stops to grow when it achieves 20 levels. Since it never reaches that value, this value was set as a dummy value.
Decision tree; W-J48	Apply pruning	Yes; No.	If the *Apply pruning* parameter is set to “Yes”, some branches are replaced by leaves according to the default error calculation of pruning set by the RapidMiner platform which is a *confidence* value equal to 0.25.

Split Validation;
Cross Validation
	Sampling Type	Linear sampling;
		Shuffled sampling;
		Stratified sampling.
			Types of sampling for building subsets.

Cross Validation	Number of folds	2;3;4.	The number of folds is the number of subsets the dataset should be divided into. Each subset has equal number of Examples. Furthermore, the number of iterations that will take place is equal to the *number of folds* set.

The range of values that were tested in each parameter was based on the numeric default values of the RapidMiner platform since we have placed these default values as the maximum values in the corresponding range of values. The minimum value was based on the minimum value that the computed decision tree could manage. For instance, we know that the "decision tree" operator splits numerical values into two branches and to prevent leaves with one instance we have set the "minimal size for split" parameter starting with the value 4 and the "minimal leaf size" parameter, with the value 2. Furthermore, regarding the "minimal gain" parameter, we have computed the *gain ration* of all attributes and have identified that the minimal gain could start with the value 0.018 since it was the gain ratio for the "Severity grade" attribute on the "Pregnancy outcome" attribute. Moreover, in the first runs, we have seen that the computed decision tree only had a leaf as an output so we have also tested the "apply_prunning" set to *True* and *False* to see if we could see a subjectively interesting decision tree. Therefore, the maximal depth was not a concern; and hence, we have left the default value as the only value to test with (i.e. the value 20). Additionally, several splitting criteria were also tested (i.e. *gain_ratio*, *information_gain*, *gini_index*, *accuracy*). All these parameter values were tested with the "Optimize Parameters" operator as suggested in Deshpande (2012) and the varied parameters were selected based on the guidelines explained in Deshpande (2012).

Please note that this training was carried-out upon each selected groups of attributes disclosed in section 5.4.4, as well as upon the groups of preprocessed attributes that delivered the most interesting results during the application of the *Clustering* technique. By doing so, we were able to train the decision tree model upon the attributes preprocessed differently (i.e. upon the original preprocessed attribute values; afterwards, upon its categorized values, as well as, upon its binomial, numerical, normalized and discretized values).

At the end of the decision tree training/testing, evaluation measures called performance metrics or error rates, were retrieved to elect the right and best model to tackle the goals of this study. Bellow, in Table 4.10, we disclose under the column named "Performance Metric", some of the performance metrics that could be used to assess the several trained/tested data mining models with classification techniques. In order to better convey which ones were in this study considered as determinant to choose the best model and why, under the column named "Definition", we present their definitions, and under the column named "Formula", how they are computed in the Rapid Miner platform. Note that the first 4 *performance metrics* are counts; and therefore, they do not have a formula specified in the RapidMiner´s documentation. The formula for the AUC performance metric is not also specified in the RapidMiner´s documentation.

Table 4.10 Performance Metrics Used (a.k.a error measures)

Performance Metric	Definition	Formula
True Positive (TP)	The number of instances classified correctly as Positive since the *label attribute* was set to Positive (i.e. Gravidez = "Sim").	
False Positive (FP)	The number of instances classified wrongly as Positive since the *label attribute* was in fact Negative (i.e. Gravidez = "Não").	
True Negative (TN)	The number of instances classified correctly as Negative since the *label attribute* was set to Negative (i.e. Gravidez = "Não").	

Performance Metric	Definition	Formula
False Negative (FN)	The number of instances classified wrongly as Negative since the *label attribute* was in fact Positive (i.e. Gravidez = “Sim”).	
Accuracy	The proportion of instances classified correctly among the total number of instances.	(TP + TN) / (TP + FP + FN + TN)
Classification Error	The proportion of instances classified wrongly among the total number of instances.	(FP + FN) / (TP + FP + FN + TN)
Precision a.k.a. Positive Predicted Value	The proportion of instances classified correctly as Positive among the number of instances classified/predicted Positively (i.e. predicted that Gravidez = “Sim”).	(TP)/(TP+FP)
Recall a.k.a. Sensitivity or True Positive Rate	The proportion of instances classified correctly as Positive among the number of instances with the *label attribute* set to Positive (i.e. Gravidez = “Sim”).	(TP)/(TP+FN)
Specificity a.k.a. True Negative Rate	The proportion of instances classified correctly as Negative among the number of instances with the *label attribute* set to Negative (i.e. Gravidez = “Não”).	(TN) / (TN + FP)
F-Measure	The harmonic mean of *Precision* and *Recall*.	2 (Precision * Recall) / (Precision + Recall) = 2TP / (2TP + FP + FN)
False Positive Rate	The proportion of instances classified wrongly as Positive among the number of instances with the *label attribute* set to Negative (i.e. Gravidez = “Não”).	(FP)/(FP+TN) = 1 - *Specificity*
False Negative Rate	The proportion of instances classified wrongly as Negative among the number of instances with the *label attribute* set to Positive (i.e. Gravidez = “Sim”).	(FN)/(FN+TP)
AUC a.k.a. Area Under the ROC Curve	The probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Note that the *ROC curve* is drawn by ordering in ascending order the obtained *True Positive* and *False Positive Rates* of all possible classification thresholds of a generated model and by plotting the *False Positive Rate* in the x axis and the *True Positive Rate,* in the y axis.	

The metrics that were determinant in the choice of the right data mining model for the prediction of the embolization success were, in the following order: *F-Measure*, *AUC*, *Recall* and the *Accuracy* metric that are highlighted in Table 4.10 with orange lettering. In fact, in the context of this study, it is important to have a *Classifier* that can correctly classify its *instances* (i.e. have a good *Accuracy* rate) and specially, as Positive (i.e. have a good *Recall* rate) since this study prizes more the instances with the “Pregnancy outcome” label attribute set to “Yes” than the other instances. Note that if we maximize the number of *True Positive*, we diminish the number of *False Negative* (see the *Recall* formula), which is what we want since we do not want to see our algorithm classifying instances as *Negative* when they are in fact *Positive*. However, there is a trade-off: if we only optimize the *Recall* metric, the *Precision* tends to diminish which means that if the model classifies all the instances correctly as Positive, the classifier will tend to classify wrongly an instance as Positive when in fact they had the "Pregnancy outcome" label attribute set to "No". This is why it is important to consider, along with the *Recall* metric, the *Precision* metric to have a broader knowledge of how the model classifies positive instances; and therefore, the *F-Measure* was in this context considered as the most important metric to select the best model because, as we can see in Table 4.10, it considers both metrics. However, the AUC metric was also checked because through this metric we can assess the cost-sensitive learning (i.e. see the False Positives generated when True Positives are classified) and learning in the presence of unbalanced classes (Fawcett, n.d.), as the select label attribute is slightly.

These metrics were computed by the RapidMiner platform with the operator called "Performance Binomial Classification" added at the end of the built *VCF*s. The best computed metrics were recorded into a table with the specifications of the ran models to ease the election of the best model. During the training of the models we have considered the *Accuracy* measure to select the parameter values that we could use during validation of the model. The best model was at last considered as the one with the highest *F-Measure* value, although we also checked if the *Recall* and *Accuracy* values were also high. However, the *ROC Curve* was at a starting point assessed to see if the built *Classifier* was better than a random one by checking if the plotted graph of the *True Positive Rate* vs the *False Positive Rate* generated by the Rapid Miner platform (i.e. the ROC curve), had its plotted curve close to the top-left corner of its graph ("Visualizing the Confusion Matrix - Sanyam Kapoor," n.d.). In fact, if this graph shows a straight line at a 45 degree angle it indicates that for every *False Positive* he detects, he has a corresponding *True Positive* so the algorithm performs a random performance that shows an AUC=0.5; and therefore, the *AUC* metrics were also looked for to be high (Fawcett, n.d.) - note that a model whose predictions are 100% correct has an AUC of 1.0.

Since all related works that have applied data mining techniques to sperm parameters have obtain Accuracies above 73% (see subsection 3.2), this study considers that an Accuracy above this value is also acceptable during the training/testing of the classification model.

Concerning previous work (Guh et al., 2011), these authors have used the following measures to assess its Decision tree model: sensitivity, specificity and accuracy. Hence, the AUC measure was not used, which we see as a drawback, since the AUC measure allows to assess if the model is worth applying.

Regarding the AUC measure, we have followed the following interpretation (Tape, n.d.):

- 0.90-1 = excellent (A)
- 0.80-0.90 = good (B)
- 0.70-0.80 = fair (C)
- 0.60-0.70 = poor (D)
- 0.50-0.60 = fail (F)

4.2.8.2 Clustering

The training of the K-means algorithm entailed its application on several groups of selected attributes with the variation of the parameters disclosed in the below Table 4.11. This Clustering training entailed the execution of at least 114 runs during the variation of its parameters upon different groups of attributes. These computed results are in the Appendix C.2 disclosed.

Table 4.11 Varied parameters through K-means training

Parameter Name	Tested Range of Values	Description
Number of clusters	2 to 4 clusters	Number of clusters that the algorithm aims to form. We have not tested more clusters because we would end up with small groups of patients due to the small data set we have.
Numerical measures	Euclidean and Manhattan	Distance between a point to assign to a cluster and the centroids

Clustering results were assessed externally and internally as follows: externally, by analyzing generated plots (e.g. scatter plot) and internally, by assessing its centroid table that presents the means of each cluster per attribute. For the most interesting models, we have applied the ANOVA statistical test upon these generated centroid means, inspired by Zancanaro, Kuflik, Boger, Goren-Bar, and Goldwasser (2007). Furthermore, clusters were also internally assessed with the distance similarity index called *Davies Bouldin* which is indicated for crisp/hard clusters (i.e. clusters where each instance only falls within one cluster). The RapidMiner platform computes this index and specifies that the closer the absolute index is to 0, the better, since it tells that the identified clusters have a low intra-cluster distance and a high inter-cluster distance (i.e. through a scatter plot, we would see that the identified clusters have a high density of data points and that clusters are apart from each other). Please note that in cases where we are assessing more than 3 attributes at the time, it is not possible to represent the generated clusters with a scatter plot; and hence, the *centroid table* is assessed with its corresponding series plot to identify interesting patient data patterns.

4.2.8.3 Association

The training of the FP-Growth algorithm entailed its application upon the selected groups of attributes, as well as the variation of the parameters below disclosed. Hence, this FP-Growth algorithm was at least 19 times applied upon the preprocessed data set through 6 modeling steps that we disclose in the next section which are mainly disclosed in the Appendix C.3.

Table 4.12 Varied parameters through FP-Growth training

Parameter Name	Tested Range of Values	Parameter Description
Support	0 to 0.1	Defines the support threshold.
Confidence	0 to 0.8	Defines the confidence threshold.

As performed in Yildirim (2015), association rules were evaluated objectively (i.e. through the computed rules´ measures) and subjectively (i.e. through the evaluation of the clinical sense and interest of the generated rule).

Objectively, the rules were assessed through their computed *support*, *confidence*, *lift* and *conviction* measure since a high *support* indicates that the rule occurs frequently; if it has a high *confidence*, it tells that its conditional probability is high; if the *lift* measure presents a different measure than 1, it means that the attributes covered in the rule are related with each other - which means that the generated rule can be considered as interesting - and if the *conviction* measure is different than 1, it means that the rule direction has an implication; and hence, it also contributes for its interestingness.

The initial *support* value was set to 0.1 and the *confidence* to 0.8 to find the objectively most interesting rules. Which means that all generated rules that were bellow these metric values were excluded; and hence, the setting of these measures pruned the generated results. This practice is recommended in the mining of health care data since this type of data tend to generate a large number of rules with low support. In fact, in Shukla, Patel, and Sen (2014) – a study that performs a review on the application of data mining techniques in the health care domain – the authors state that in the health care domain, we tend to have a significant fraction of association rules that are irrelevant and that the most relevant rules, often appear with high quality metrics but with a low support. We believe that this is the reason why in Yildirim (2015), the support was set to 1% and the confidence to 40% since its work was also performed in the health care domain. Furthermore, after our first application of the FP-Growth algorithm, we have seen that a *support*=0.1 and a *confidence*=0.8 would not generate subjectively interesting rules so we have lowered this threshold to 0.0 and after that, have seen that a support equal to 0.1 and a confidence equal to 0.4 was in our case also enough to identify the most objectively and subjectively interesting rules as in Yildirim (2015).

But what are subjectively interesting rules in this study? Well, we could say that since all assessed attributes are related with pregnancy outcome, all generated rules could be seen in the clinical perspective as interesting. However, an association rule that provides a predictive information is more interesting in this case (i.e. rules where an attribute filled **before** the treatment implies another one that occurs **after** the embolization treatment, can be seen as a predictive rule). Furthermore, since one of the main goals of this study is to predict the success of the embolization treatment, looking for rules that have the “Pregnancy outcome” attribute as conclusions is in line with our goals. In fact, even if the *association* data mining technique is usually applied to discover patterns/relations between sets of attributes (i.e. descriptive purposes), we can also use this technique for predictive purposes as performed by Azevedo and Jorge (2007). Hence, we have further on filtered our data set by non-missing values in the “Pregnancy outcome” attribute prior the application of the *association* algorithm, with the aim of generating rules upon patients which we know were able to get their partner pregnant or not. Hence, the interestingness factor (IF) of the lift and conviction measure value was interpreted as follows (Yildirim, 2015):

- IF(X, Y) =1.0, X and Y are independent,
- IF(X, Y) >1.0, X and Y are positively correlated,
- IF(X, Y) <1.0, X and Y are negatively correlated.

However, in terms of *lift* and *conviction*, we have focused, as suggested in Rapid Miner´s tutorials, on seeking rules higher than 1.0 because it conveys that the rule is more interesting than below 1.0. In fact, if we have a *lift* higher than 1.0, it tells us that the probability of the attribute occurring *consequent* is lower than its conditional probability - based on the *lift* formula.

The choice of all these measures was based on Yildirim (2015). In fact, in Yildirim (2015), the author states that the *support* measures the statistical significance of the computed rule; the *confidence*, the strength of the rule, and the *lift* and *conviction*, its interestingness through the assessment of the correlation between the antecedent and the consequent attributes. Based on that statement, we have ranked the results by statistical significance (i.e. the *support* measure) and if the other measures, especially the *confidence* measure, had good values, we have considered that first ranked rule as objectively the most interesting one. In fact, the authors (Han et al., 2012) state that association rules are considered interesting if they both satisfy a minimum *support* threshold and a minimum *confidence* threshold; and therefore, a rule that has these two measurements above a set threshold should be assessed with the *lift* and *conviction* measures to seek for objectively interesting rules. Furthermore, in Azevedo and Jorge (2007) the *conviction* measure proved to be effective for predictive tasks, which also reinforces the selection of these metrics.

The initial *support* value was set to 0.1 and the *confidence* to 0.8 to find the most objectively interesting rules. Further on, to find some subjectively interesting rules, we have lowered the *support* and *confidence* to 0.0 and have seen through the first tests that we could adjust the threshold values to 0.1, for the *support*, and 0.4, for the *confidence*, as in Yildirim (2015). Hence, the selection of the most objectively and subjectively interesting association rules entailed the assessment of each computed association rule by the following conditions which can be seen as the pruning conditions:

- Objectively interesting:
    - *support* > = 0.1
    - *confidence* > = 0.4
    - *lift* and *conviction* > = 1.1
- Subjectively interesting:
    - The *antecedent* occurred before, or at the same time of, the *consequent*.
    - *support* > = 0.15 which means that the rule encompasses at least 35 patients.

In order to better convey the computed results, we have presented them by the *consequent* attributes that were seen in each test as objectively interesting where the objectively interesting rules were marked with a check mark and the subjectively interesting ones, with an exclamation mark in the generated result tables.

The selection of the most interesting rules was performed by checking if the previously selected/prunned association rules had a statistically significant dependence between their *antecedent* and *consequent* (i.e. if the relation between the antecedent and consequent, presented by the lift and conviction measure, is statistically significant). This assessment was carried out by applying the Chi-square test as firstly proposed in Brin, Motwanit, and Silverstein, (1997) for a significance level of p=0.10, p=0.05 and p=0.01 to elect a greater number of interesting association rules. Rule selection was carried out by inserting into an excel she*et al*l previously selected association rules and by calculating, for each of them, the

Chi-square test in terms of confidence, support and lift measures as performed in Yildirim (2015). In fact, it has also been shown (Alvarez, 2003) that the x^2 value could be directly calculated in terms of these standard measures (i.e. the Chi-square statistical formula presented in Formula 4.1 satisfies the equality defined by the author Sergio A. Alvarez). This formula is defined as:

$$x^2 = n(lift - 1)^2 \frac{Support * Confidence}{(Confidence - Support)(Lift - Confidence)}$$

Formula 4.1 Chi-square (Alvarez, 2003)

Hence, in our study the above formula is applied with an *n* equal to 230 - since it is the number of instances without missing values in the “Pregnancy outcome” attribute - to elect the best computed association rules.

We have considered that an association rule with an antecedent and a consequent is statistically significant if the computed x^2 value, calculated with the Formula 4.1**Erro! A origem da referência não foi encontrada.**, is equal or above the following values:

For p= 0.10	2.706
For p= 0.05	3.841
For p= 0.01	6.635

These values were retrieved from the x^2 distribution for a degree of freedom of 1 since all attribute values are binomial; and therefore, the contingency table would always have 2 rows and 2 columns, called a two-dimensional table. In fact, (Alvarez, 2003) aggregates the items of the antecedent and, separately, the items of the consequent by performing a Boolean product over each of these item sets to end up with a two-dimensional table with the aim of increasing the possibility of achieving the minimum cell counts required for the validity of the Chi-square analysis - which is 5. Hence, the above x^2 threshold values were always used for this purpose.

On the set of the most interesting rules (i.e. pruned rules with a statistically significant dependence between the *antecedent* and *consequent*) we have considered that the best rule was the one with the highest *support* and/or *confidence* value since it indicates, respectively, the statistical significance and strength of the rule. Most works only use the *confidence* measure to elect the best rules but since we have a small data set and the healthcare domain tend to generate rules with low support, it was in the context of this study important to firstly check for their *support*.

4.2.9 Modeling

The CRISP-DM modeling phase was carried-out by following a set of sub-phases (see Table 4.13) which in turn, entailed the execution of a set of modeling steps that are in the Appendix B described along with the built models.

Table 4.13 Modeling Sub-Phases

Sub-phase N^a	Carried out modeling steps	General Tasks
1	Decision tree modeling step 1 and 2 (see step details in Table 6.1)	Apply a Decision tree model upon the attributes that have a good data quality, and afterwards, only upon the first group of selected attributes.
2	K-Means modeling step 1 to 5 (see step details in Table 6.3)	Find interesting data patterns with several K-Means models.
3	All FP-Growth modelling steps (see step details in Table 6.6)	Find interesting data patterns and predictive information with several FP-Growth models.
4	Decision tree modeling step 3 to 9 (see step details in Table 6.1)	Continue the Decision tree application by applying its models upon the remaining groups of selected attributes and afterwards, upon the data set transformed into numerical and nominal values and at last, try to optimize one of the best computed result with the *Bagging* ensemble meta-algorithm.
5	K-Means modeling step 6 (see step detail in Table 6.3)	Apply the best decision tree model upon the most interesting clustered instances to better describe its findings.

Chapter 5 Study of the Varicocele Condition

This section presents the results of the application of the previously described methods by following the CRISP-DM project methodology. Hence, this section begins by describing in section 5.1, the CRISP-DM project methodology, and then presents the outcomes of each CRISP-DM phase as follows: in section 5.2, we present the requirements of the “Biology of Reproduction and Stem Cells” (BRSC) research team, what the project resources and constraints/project risks were – with the indication of the ways to tackle them -, the data mining Goals set, as well as the data mining tools and techniques used. In section 5.3, we present the statistical results that were obtained with the *final preprocessed data set*. In section 5.4, we specify de quality of the *initially provided data set,* as well as describe the data preparation that was performed to achieve the *final preprocessed data set*. In section 5.5, we disclose, identify and interpret the best data mining results generated for each applied data mining technique and in section 5.6, we present a summary of the best results and discuss them by evaluating their contributions.

5.1 Project Methodology

This study follows the CRISP-DM methodology since it is the most popular for data mining projects (Piatetsky, n.d.). After the data preparation phase, the data mining techniques are applied to the preprocessed data set to build data mining models and evaluate them. Finally, the selected models are deployed/disclosed to the client. In our case, the discloser of the data mining models built was performed through the presentation of the acquired results presented in this chapter. The schema of the CRISP-DM methodology is shown in Figure 5.1.

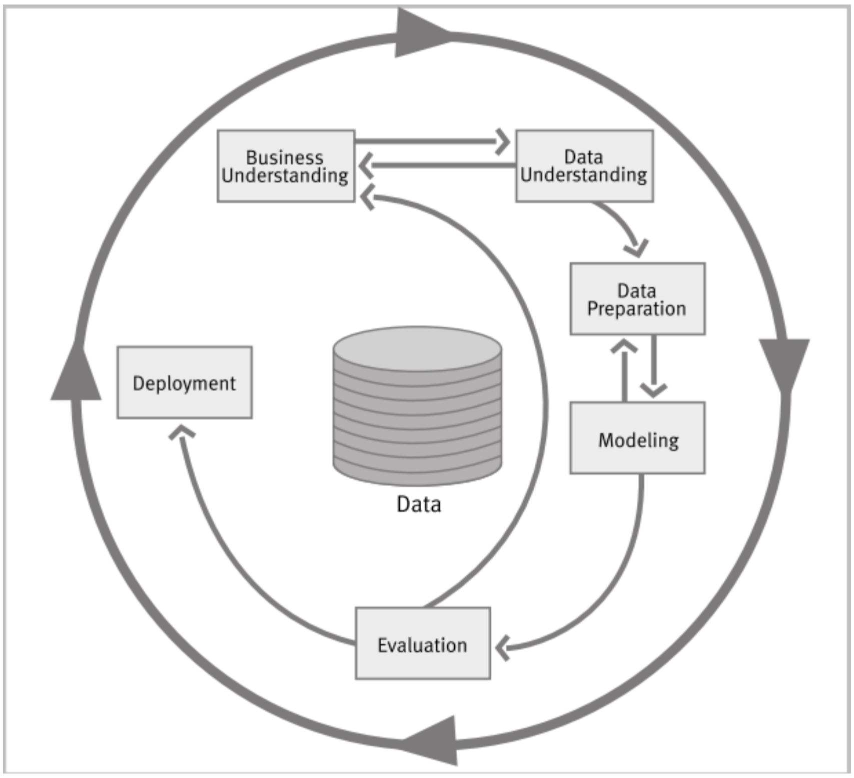


Figure 5.1 Phases of the CRISP-DM project methodology (Chapman *et al*., 2000)

As we can see in Figure 5.1, CRISP- DM is a project methodology that encompasses a set of phases that are executed in iterative ways. In 2014, CRISP-DM was still the mostly used data mining project methodology (Piatetsky, n.d.) and CRISP-DM attributes its success to its practical approach on planning projects since it was built upon real-world experiences of how people conduct data mining projects (Chapman et al., 2000).

5.2 Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into data mining Goals and preliminary plan the project to achieve the previously defined data mining Goals (Chapman et al., 2000). Thus, in this section we present in subsection 5.2.1 the business background and objectives of the research team with whom this work was carried out. In 5.2.2, an inventory of the resources available to the project, as well as the constraints and problems/risks that this project encountered and tackled. In 5.2.3, the data mining goals of our project. In 5.2.4, the selected data mining tool, and in 5.2.5, the data mining techniques/algorithms that were used.

5.2.1 Background, Objective and Requirements

This work was carried out with the BRSC research team of the Center for Neuroscience and Cell Biology (CNC) that works with the Medicine of Reproduction Unit of the *Centro Hospitalar e Uiversitário de Coimbra* (CHUC) located in Portugal.

The main research domain of the BRSC team is to study male infertility and its main objectives are to study the mechanisms, conditions (e.g varicocele) and external aspects (e.g environmental aspects, patient lifestyle and fiscal characteristics etc.) that might be associated with male infertility, to increase treatment success.

One of the ongoing research areas of the BRSC team is the study of the varicocele condition where its actual main needs/requirements are to predict the success of the varicocele treatment, performed in CHUC, as well as its prevalence.

5.2.2 Resources, Constraints and Project´s Risks

This project had several resources available shown in Table 5.1, that indicates these resources by purpose and when applicable, specifies the preprocessed attributes (attributes presented in Table 4.3) that used the corresponding resource to validate and collect its data (indicated under the column named “Attributes”). If we analyze the attributes specified under this column, we see that some attributes used several information systems to validate and collect its data. It is the case of the male patient external factors (i.e. Man infertility factor, Smoking habits, Drinking habits, Surgeries, Diseases), that used the information technology system called “*SMR*” to validate and retrieve that information, and when the “*SMR*” system had empty fields for these attributes, we have looked up into the patient medical dossiers. The pregnancy outcome attribute was also checked in another information system due to the lack of information that was found in the initially available information systems for some of the patients.

Table 5.1 Project´s Resources

Purpose	Resource	Attributes
Business expertise	BRSC team members and CHUC personnel	
Data Validation and Collection	CHUC´s information technology system called “*Doentes*”	Man age; Woman age; Occupation; Embolization date; Birth; Number of alive babies; Spontaneous pregnancy; Pregnancy outcome
	SHR´s information technology system called “*SMR*”	Type of infertility; Woman infertility factor; Male infertility factor; Smoking habit; Drinking habit; Surgeries; Diseases; Pregnancy outcome; Number of pregnancies; ART
	CHUC information technology system called “*Anatomia Patológica*”	Pregnancy outcome (before 2012, pregnancy test results were in this system recorded)
	Patient medical dossiers	Type of infertility; Woman infertility factor; Man infertility factor; Smoking habit; Drinking habit; Surgeries; Diseases; Severity grade; Laterality; Testis volume; Embolized laterality; Material of embolization; Complications

	Original Semen Analysis Reports	Concentration before treatment;
		Concentration at 3 months;
		Concentration at 6 months;
		Concentration at 12 months;
		Progressive motility before treatment;
		Progressive motility at 3 months;
		Progressive motility at 6 months;
		Progressive motility at 12 months;
		Morphology before treatment;
		Morphology at 3 months;
		Morphology at 6 months;
		Morphology at 12 months

The initially identified constraints were:

- The missing values in the data set, as well as in the final and preprocessed data set;
- The small data set that we ended up with (i.e. 293 instances) for the data mining universe.

However, these data constraints are expectable for the health care domain and the size of the final and preprocessed data set is good for the varicocele domain as we have seen through the review of related works presented in section 3.4.

For the data preparation step, another main constraint was raised:

- The impossibility of exporting patient information to a digital format.

However, we were able to export the patient´s information gathered in the "*SMR*" information system to an EXCEL file, and afterwards, import that data to a temporary Database, built for this purpose in the Microsoft SQL Server 2012, to at last, retrieve the needed information with SQL queries. All other attributes, that were not validated with the "*SMR*" information system, were validated and filled manually.

As in any project, this study also encountered some problems/risks. The main ones were:

- The time line of the project was short to acquaint with knowledge needed to further understand and analyze the data, to look up and retrieve the needed information in the available information systems and to preprocess and generate/build new attributes;
- The generation of an interesting data mining model due to the missing values and small correlations seen with the initial data set;
- The risk of not being able to formulate some conclusions for each data mining goals set due to the outcomes achieved.

To tackle these problems and risks, I dedicated a big part of my time to fill in the missing values, correct the provided ones and generate new attributes to increase our chances of having higher correlations. Moreover, we have decided to carry out an *inductive statistical analysis* with the ANOVA and Chi-square statistical tests during the *Data understanding* step to further on serve as an input for the data mining algorithms, and, finally, we have identified and exhaustively applied the data mining algorithm that showed the best performance on sperm parameter data (i.e. Decision tree seen in (Guh et al., 2011)) to guide and increase our chances of formulating interesting conclusions for data mining by also using the knowledge acquire with the other applied data mining techniques.

5.2.3 Data mining goals

As seen in section 5.2.1, the BRSC team has been studying the varicocele condition, as well as its treatment outcomes with embolization. By considering their main requirements (i.e. predict the prevalence of the varicocele condition, as well as its treatment success), the following data mining goals were identified with a different degree of importance which made us highlight them with different colors to better convey its priority; i.e., goals with orange lettering are the most important to achieve, goals with yellow lettering are secondary and goals with black lettering could not be achieved:

With *predictive* data mining techniques:

Goal 1) Predict the varicocele condition.

Goal 2) Predict the success of the varicocele embolization carried out at CHUC.

With *descriptive* data mining techniques assess:

Goal 3) Semen classification vs the condition laterality.

Goal 4) The varicocele condition vs the patients that do not have the condition.

Goal 5) Other male infertility patterns (e.g. sperm evolution through time with statistical tests and semen classification vs patient´s external factors with data mining techniques).

After understanding and assessing the *completeness* of the provided dataset, we have seen that this study could achieve all these data mining goals, except the Goal 1 and 4 (see section 5.4.1 for the rationale). Furthermore, we have identified that the goals with the highest interest for the BRSC team was: the *prediction* of the embolization success (i.e. Goal 2), as well as the *description* of the relation between the patient semen classification and its external factors (i.e. occupation, drinking, alcohol, disease) and the statistical description of the evolution of the patient sperm parameter values through time (i.e. Goal 5); and therefore, this study focused on Goal 2 and 5, and afterwards, on Goal 3.

5.2.4 Data mining Tool Selection

Through previous work, we have analyzed the requirements of the healthcare industry and have identified that the RapidMiner platform was one of the mostly suited tools to perform knowledge discovery (KDD) on clinical data sets. Hence, in this work we have choose to use the RapidMiner Educational tool, version 8.1.001.

5.2.5 Techniques Used

By considering the studies presented in section 3.2, as well as the characteristics of the data mining algorithms described in section 4.2.7, we have selected the data mining algorithms for this study. Since data mining algorithms can be categorized as predictive or descriptive, we specify below the selected data mining algorithms grouped by these categories. Hence, in section 5.2.5.1, we specify the predictive algorithms that were used in this study, and in section 5.2.5.2, the descriptive ones.

5.2.5.1 Predictive Algorithms

Due to their good performance shown in related works, this study applies the following algorithms:

- RapidMiner´s Decision tree; W-J48 (Classification data mining technique);
- FP-Growth (Association data mining technique).

5.2.5.2 Descriptive Algorithms

Due to its popularity, this study also applies the following algorithms:

- K-means (Clustering data mining technique);
- FP-Growth (Association data mining technique).

5.3 Data Understanding

The data understanding phase starts with initial data collection and proceeds with activities that enable to become familiar with the provided data (Chapman et al., 2000). In fact, studying the varicocele condition and its treatments, as shown in section 2.2, as well as its related works, as show in sections 3.3 and 3.4, acquainted us with the provided data to further on statistically understand it.

Statistical analysis is an important asset to a data mining study because its findings can give us clues on data patterns than can be further on investigated with data mining techniques. In fact, we cannot forget that the data mining technique also encompasses, along with other fields, statistical concepts. Hence, this section aims to present our statistical findings on the attributes described in the section 4.1.2.1 since they have guided us through our data mining application.

In order to better convey our statistical findings, the statistical results are presented as follows: in subsection 5.3.1, we disclose the basic statistical descriptions of each attribute disclosed in Table 4.3, as specified in section 4.2.5, and in the subsection 5.3.2, we statistically explore the previously disclosed attributes into different angles by going further with the application of the statistical tests described in section 4.2.6.

5.3.1 Data Description

In order to know where do most values fall, we have computed the measures of central tendency (i.e. the Mean, Median and Mode of each attribute). Table 5.2 presents these results, as well as the number of filled values that each attribute has by specifying it under the column named “n”.

Table 5.2 Statistical description of quantitative attributes - Measures of central tendency

ID	Attribute code name	Mean	Median	Mode	*n*
1	Idade_H	34.43	34	36	293
2	Idade_M	32.22	32	33	284
3	Tempo_Infert	39.22	34,5	24	254
14	Data_Embolização	Dec 29, 2012	Oct 12, 2013		293
21	Conc_Pre	13.93	4.2	0	281
22	Conc_3M	19.07	11	0	245
23	Conc_6M	17.21	7.6	0	131
24	Conc_1A	15.78	8	0	137
25	A_B_pré	26.90	24	0	251
26	A_B_3M	31.12	29	0	217
27	A_B_6M	28.91	23	0	116
28	A_B_1A	28.44	26	0	121
29	Formas_N_pré	4.06	2	1	210
30	Formas_N_3M	4.67	3	2	180
31	Formas_N_6M	4.79	3	3	47
32	Formas_N_1A	3.13	3	1	23
34	Num_Gravidezes	0.46	0	0	293
36	Num_Bebés	1.27	1	1	84
37	Gravidez_pós_emb	17.70	13	9	105

In order to have an idea of how the quantitative data is spread out (Han et al., 2012), we have also identified for each quantitative attribute the following values: the minimum value (Min), the value of the first quartile (Q_1), the value of the third quartile (Q_3), the maximum value (Max) and the standard deviation (SD). Table 5.3 presents these results.

Table 5.3 Statistical description of quantitative attributes - Measures of data dispersion

ID	Attribute code name	Min	Q_1	Q_3	Max	SD
1	Idade_H	23	31	37,75	54	5.22
2	Idade_M	20	30	35	46	4.40
3	Tempo_Infert	4	24	48	192	28.87
14	Data_Embolização	Jan 17, 2007			Apr 28, 2016	

ID	Attribute code name	Min	Q_1	Q_3	Max	SD
21	Conc_Pre	0	0.8	14	220	23.64
22	Conc_3M	0	1	27.1	170	24.94
23	Conc_6M	0	1.1	27	160	24.07
24	Conc_1A	0	1.5	26	80	18.66
25	A_B_pré	0	7	40.5	89	22.46
26	A_B_3M	0	12	46	94	22.21
27	A_B_6M	0	5	50	83	25.05
28	A_B_1A	0	11	43	83	23.14
29	Formas_N_pré	0	1	5	38	4.826
30	Formas_N_3M	0	2	6	21	4.443
31	Formas_N_6M	0	2	6	21	4.515
32	Formas_N_1A	0	1.5	3.5	10	2.510
34	Num_Gravidezes	0	0	1	3	0.679
36	Num_Bebés	0	1	2	3	0.567
37	Gravidez_pós_emb	0	9	21	79	15.09

To have an overall picture of each qualitative attribute, we have identified some of the least and most frequent values of each qualitative attributes and specified between parentheses its frequency. We have also specified the number of filled values has in the previous tables. This information is presented in Table 5.4.

Table 5.4 Statistical description of quantitative attributes - Basic statistical description

ID	Attribute code name	Least Frequent	Most Frequent	n
4	Prim_Sec	Secundária (54)	Primária (216)	270
5	Factor_Infertilidade_Feminino	Anovulação + Tubar (1) Baixa reserva (1) Baixa reserva + Patologia Uterina (1) Dismenorreia I Ligeira (1)	Anovulação (25)	85
6	Factor_Infertilidade_Masculino	Hiperprolactinemia (1) Oligoespermia Severa (1) Tumor testicular (1)	Masculino (82)	163
7	HabitosTabagicos	Não – exfumador há 18 anos (1) Não – exfumador há 2 anos (1) Sim – 1 cigarro por dia (1) Sim – 13 cigarros por dia (1) Sim – esporadicamente (1) Sim – ocasional (1)	Não (88)	204
8	HabitosAlcoolicos	Não – 50oils50ca50c (1) Sim – 1 copo ao jantar (1) Sim – 50oils50ca50c (1) Sim – 50oils50 (1) Sim – à refeição (1)	Não (80)	116
9	Cirurgias	Sim – Amigdalectomia aos 11 anos (1) Sim – Amigdalectomia. Adenoidectomia (1) Sim – Apendicectomia aos 21 anos (1) Sim – Bypass Gástrico (1) Sim – Circuncisão (1) Sim – Criptorquidia (1)	Não (53)	138
10	Doença	Apendicite aguda gangrenosa (1) Ardor ejaculatório (1) Artrite reumatoide (1) Doença celíaca (1) Epididimite pré-embolização (1) Excesso de peso – IMC 30,9 (1)	Parotidite (12)	92

ID	Attribute code name	Least Frequent	Most Frequent	n
11	Profissao	Afinador maquinas (1) Agente da PSP (1) Assistente comercial (1) Engenheiro Zootécnico (1) Engenheiro agrónomo (1) Fabricante de capacete (1) Ferroviário (1) Fiel de armazém (1) GNR (1)	Empresário (6)	202
12	Grau_Varicoc	III (33)	II (111)	211
13	Lateralidade	Direito (1)	Esquerdo (178)	218
14	Volume_Testiculo_Médico	Reduzido (1)	Bom (14)	44
16	TratamentoFeito_lateralidade	Embolização apenas Direita (1)	Embolização apenas Esquerda (200)	206
17	TratamentoFeito_material	Cola + Lipiodol (2)	Coils (15)	23
18	Complicações	Abcesso (1) Dor intensa (1) Lombargia e dor (1) Muita dor. Fez reacção 51oils51ca ao contraste iodado (1) Sim – aquando a cateterização da artéria(1)	Não (271)	293
19	Repetia_embolização	Não (10)	Desconhecido (144)	293
20	Razão_não_repetir	Infeção (1) Recuperação difícil (1) Técnica (1)	NA (283)	293
33	Gravidez	Sim (107)	Não (123)	230
35	Nascimento	Sim (84)	Não (146)	230
38	PMA	Sim (66)	Não (228)	293
39	Gravidez_espontanea	Sim (49)	Não (181)	230

5.3.2 Data Exploration

In this subsection, we present the results of the data exploration carried out with *descriptive* and *inferencial statistics* upon the attributes described contextually in Table 4.3, and statistically, in Table 5.2, Table 5.3 and Table 5.4. Please note that important results disclosed in tables are usually highlighted in orange color.

To better convey our findings, this section presents our statistical results through several subsections that analyze the attributes that are clinically important. Clinical importance was related with the understanding of some of the aspects of male infertility learned from the BRSC team and complemented with the study of related works on the varicocele condition that were previously summarized. As seen in related works, sperm parameters are one of the most common attributes that are clinically assessed in the male infertility and varicocele domains; and therefore, this study has assessed them from different angles (subsection 5.3.2.2, 0and 5.3.2.4) after analyzing the descriptive statistics of the 39 initially preprocessed attributes (subsection 5.3.2.1). Later, we analyze if there is a pattern on the month on which the patient´s partner conceives (subsection 5.3.2.5) and then identified the attributes that were more correlated with the pregnancy attribute (subsection 5.3.2.6). Finally, through several subsections, we have also analyzed other data relationships to guide us on the identification of other data patterns (in subsection 5.3.2.7, between laterality and severity grade, and in the subsections 5.3.2.8, 5.3.2.9 and 5.3.2.10, between the semen classification and varicocele laterality, as well as drinking and smoking habits, respectively). At last, in section 5.3.2.11, we assess the relation between the patient´s age and the pregnancy outcome.

5.3.2.1 Analysis of the descriptive statistics

In this subsection we aim to, through data visualization (i.e. graphs), analyze the attributes statistically described in section 5.3.1 and draw the first statistical conclusions or give insights of the attributes that might not be considered for data mining analysis due to its poor provided data knowledge. To better convey our first conclusions, we have grouped our results by several subjects and have disclosed them through several subsections.

5.3.2.1.1 Patient age vs date of embolization treatment

By analyzing the values of the attribute "*Data_embolization*" we see that the embolization treatment was performed between January 2007 and April 2016. If we analyze the dispersion of the male patients´ ages though time, we see that half of them (51.70%) were treated between July 2013 and July 2016 (152 on 293 instances) which is clearly seen in the scatter graph depicted in Figure 5.2 through a higher dot concentration in that time span. Moreover, if we analyze the evolution of the male patient age through time, we see that they have been treated at a similar age. In fact, if we divide the patients into halves by median date of treatment (i.e. *Data_Embolização*); and hence, compare the male patient age treated between December 10th of 2008 and October 10th of 2013 inclusively, that have an age mean of 34.46 ±5.18 years old on a sample of 142 patients, with the ones treated between October 14th of 2013 and February 24th of 2016 inclusively, that have an mean age of 34.36 ±5.10 years old, for a sample with the same size, we see that the age mean is similar, as well as its standard deviation. The same happens for their corresponding partners: 32.25 ±4.36 for the time span between December 10th of 2008 and October 10th of 2013, and 32.23 ±4.40, for the time span between October 14th of 2013 and February 24th of 2016.

If we analyze the male patient age with the age of the corresponding partner, we see that women are on average younger than males, and with a lower standard deviation which means that the women´s age are not as spread out as the men are. Figure 5.3 depicts a scatter graph of the women ages by the date of the embolization to see how the data is spread out through time, as well as the corresponding age of their male partner. As expected, most women have male partners with similar ages and between the first and third quartile of the male patient´s age which is from 31 to 38 years old since most dots in the scatter graph are colored in turquoise to green, representing this range of years.

Regarding the success of pregnancy after varicocele treatment, we see in Figure 5.2 that the provided data set is quite homogeneous in that matter, where if the male patient´s partner got pregnant (blue dot), did not got pregnant (green dot) and we do not know (red dot), dispersed through time and with homogeneity. We have calculated the number of patients that did not got and got pregnant and have seen that of the 293 patients assessed, we were able to categorize 230 male patients in terms of if they were able to conceive their partner, and from these 230 male patients, 46.52% (107/230) were successful.

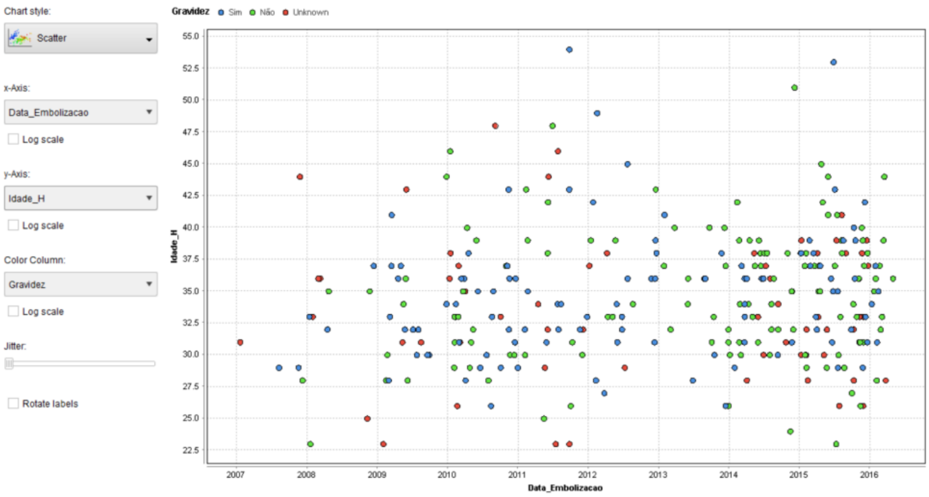


Chart style: Scatter

x-Axis: Data_Embolizacao

□ Log scale

y-Axis: Idade_H

□ Log scale

Color Column: Gravidez

□ Log scale

Jitter:

□ Rotate labels

Figure 5.2 Scatter graph of male patient´s age by treatment date

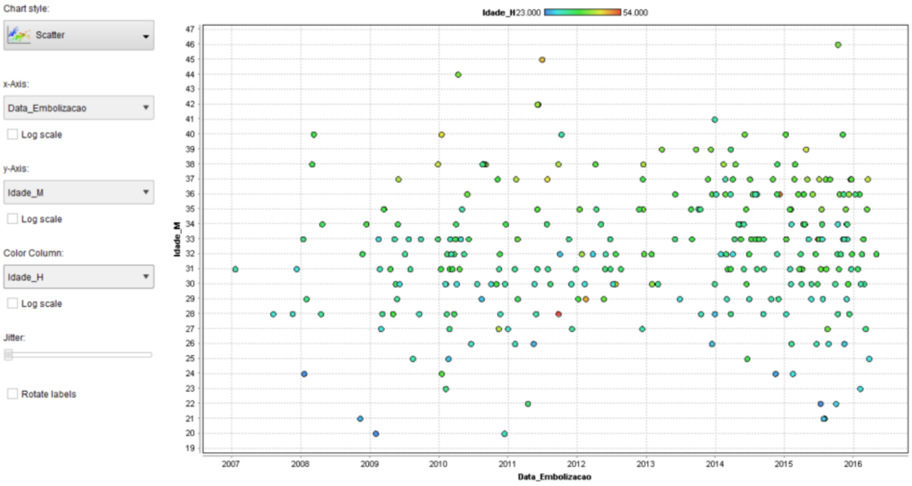


Chart style: Scatter

x-Axis: Data_Embolizacao

□ Log scale

y-Axis: Idade_M

□ Log scale

Color Column: Idade_H

□ Log scale

Jitter:

□ Rotate labels

Figure 5.3 Scatter graph of male patient´s partner age by treatment date

5.3.2.1.2 Infertility time vs male patient´s age and outcome

During the exploration of the values of the "Infertility time" attribute, we have seen that the most common time span for a couple to seek medical help is 24 months (25.59% 65/254), followed by 36 months (16.14% 41/254). By generating a histogram, we have seen that these times were positively skewed as suspected since its mean, median and mode are different (see Figure 5.4).

If we analyze the "Man age" attribute with the "Infertility time" attribute, we see that the data seems to create a cluster of 198 patients that encompasses 77.95% of this sample (198/254). This cluster seems to go from 25 to 40 years old and from 5 to 60 months of infertility. Out of these 198 patients, 81 patients achieved pregnancy at most 77 months after the embolization treatment, 63 out of these 81 patients, achieved live births and 61 out of these 63 pregnancies resulted in 1 or 2 live babies whereas 37 out of these 61 patients, conceived with an ART procedure.

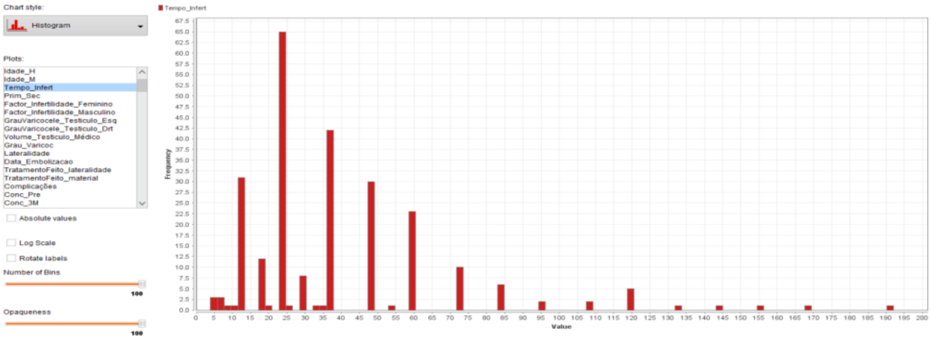


Chart style: Histogram

Plots:
Idade_H
Idade_M
Tempo_Infert
Prim_Sec
Factor_Infertilidade_Feminino
Factor_Infertilidade_Masculino
GrauVaricocele_Testiculo_Esq
GrauVaricocele_Testiculo_Drt
Volume_Testiculo_Médico
Grau_Varicoc
Lateralidade
Data_Embolizacao
TratamentoFeito_lateralidade
TratamentoFeito_material
Complicações
Conc_Pre
Conc_3M

□ Absolute values

□ Log Scale

□ Rotate labels

Number of Bins: 100

Opaqueness: 100

Figure 5.4 Histogram of the infertility time

5.3.2.1.3 Dispersion of the sperm parameters values through time

The exploration of the sperm parameter values was carried out through the analysis of its five-number summary (i.e. minimum, first quartile, median, third quartile and maximum value), as well as its mean, standard deviation and sample dimension (*n*). Since the range of values of each sperm parameter differs substantially – specially sperm concentration – each sperm parameter was separately analyzed through the elaboration of summary tables presented in Table 5.5, Table 5.6 and Table 5.7, and box plots presented in Figure 5.5, Figure 5.6 and Figure 5.7. The built box plots can be interpreted as follows: the value at the end of the top whisker denote the maximum value of the sperm parameter values indicated in Table 5.5, Table 5.6 and Table 5.7 with the id (*Max)*; the length of the top whiskers represents 25% of the sperm parameter values and goes from the maximum sperm parameter value to the third quartile (*Q3*); the gray block depicts the difference between the third quartile (*Q3*) and the median (*Median*); the orange bloc depicts the difference between the median and the first quartile (*Q1*) and the bottom whisker, the difference between the first quartile (*Q1*) and the minimum value (*Min*). Hence, the median is represented in the box plot by de separation between the gray and the orange block. Furthermore, as the top whisker, the length of the bottom whisker represents 25 % of the lowest sperm parameter values. Hence, the data between the third and the first quartile, depicted with the gray and orange box, represent 50% of the data that is close to the median of the sperm parameter values.

Table 5.5 Sperm Concentrations through time – Main statistical results

Attribute code name	Com_Pre	Com_3M	Com_6M	Com_1A
Min	0	0	0	0
Q1	0.8	1	1.1	1.5
Median	4.2	11	7.6	8
Q3	14	27.1	27	26
Max	220	170	160	80
Mean	13.93	19.07	17.21	15.78
Mean-Median	9.73	8.071	9.61	7.78
Standard Deviation	23.64	24.94	24.07	18.66
n	281	245	131	137

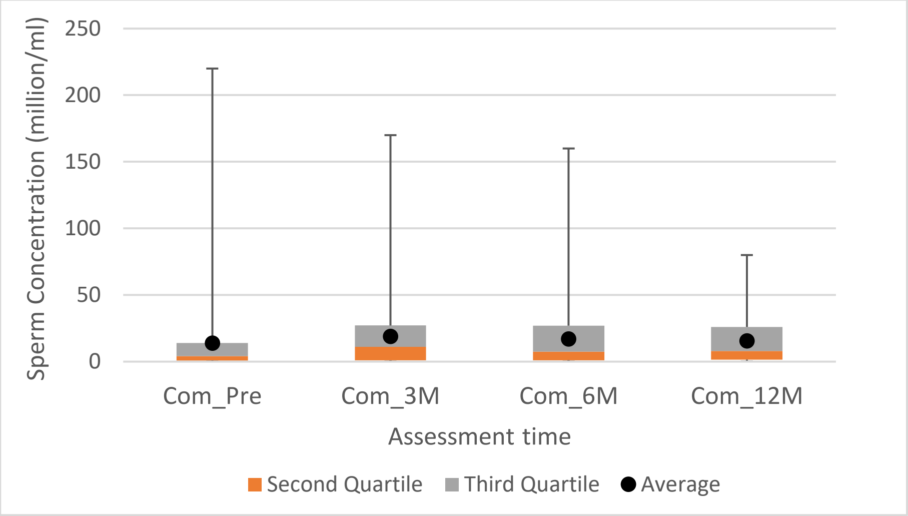


Figure 5.5 Sperm Concentrations through time - Box Plot with mean values

Table 5.6 Sperm Progressive Motilities through time – Main statistical results

Attribute code name	A_B_Pre	A_B_3M	A_B_6M	A_B_1A
Min	0	0	0	0
Q1	7	12	5	11
Median	24	29	23	26
Q3	40.5	46	50	43
Max	89	94	83	83
Mean	26.90	31.12	28.91	28.44
Mean-Median	2.90	2.12	5.91	2.44
Standard Deviation	22.46	22.21	25.05	23.15
n	251	217	116	121

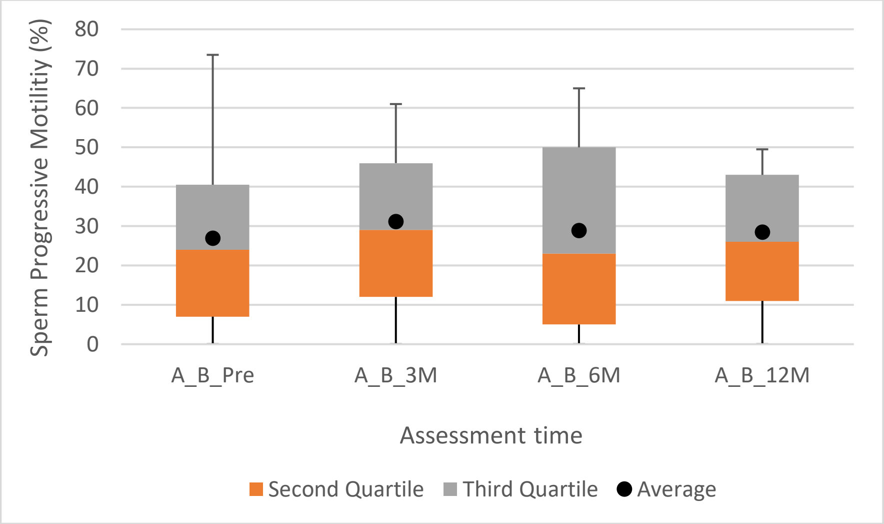


Figure 5.6 Sperm Progressive Motilities through time – Box Plot with mean values

Table 5.7 Sperm Morphologies through time – Main statistical results

Attribute code name	Forma_N_Pre	Forma_N_3M	Forma_N_6M	Forma_N_1A
Min	0	0	0	0
Q1	1	2	2	1.5
Median	2	3	3	3
Q3	5	6	6	3.5
Max	38	21	21	10
Mean	4.062	4.67	4.79	3.13
Mean-Median	2.06	1.67	1.79	0.13
Standard Deviation	4.83	4.44	4.51	2.51
n	210	180	47	23

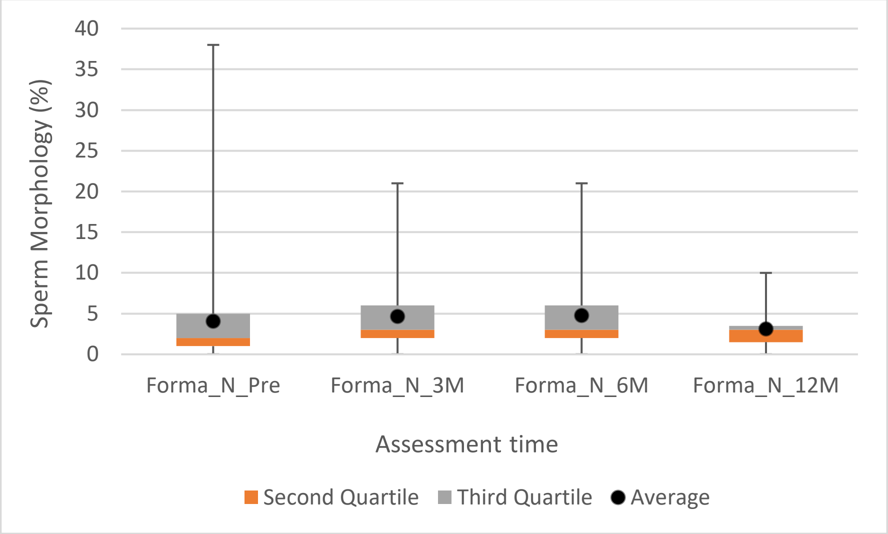


Figure 5.7 Sperm Morphologies through time – Box Plot with mean values

If we analyze the *Min* and *Max* values presented in the summary tables, we see that the sperm concentration and morphology have its widest data dispersion before the embolization treatment in contrast to the progressive motility, where this occurs at 3 months after treatment. Furthermore, mean and median values are different for all sperm parameters. In fact, in the box plots above, we see that the black dots, that represents the mean value, are all above the median value.

Since the mean value for all patient´s follow-up times and sperm parameters is larger than the median value with a difference between both values always above 0, this indicates that sperm parameter values are skewed (i.e. not normally distributed). To validate this idea, histograms were generated with the RapidMiner Platform and shown in Figure 5.8, Figure 5.9 and Figure 5.10, where we can indeed see that the data is right-skewed, also called positive skewness, since right-skewed distributions have a larger mean value than the median value ("Skewed Distribution: Definition, Examples - Statistics How To," n.d.). That happens because the mean value is influenced by extreme scores; and therefore, pulls towards, in right-skewed distributions, to the right long tail of the f(x) function. In fact, this explains why the sperm concentration, before the embolization treatment, has the larger "Mean – Median" value since the sperm concentration before the embolization treatment has also the biggest range of values (*Max-Min*) which tends to further pull the mean from the median.

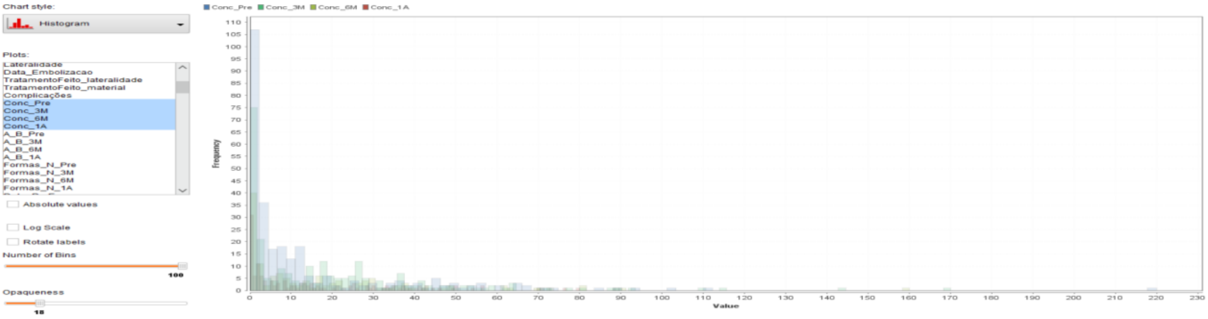


Chart style: Histogram

☐ Absolute values

☐ Log Scale

☐ Rotate labels

Number of Bins 100

Opaqueness 18

Figure 5.8 Histogram of sperm concentrations

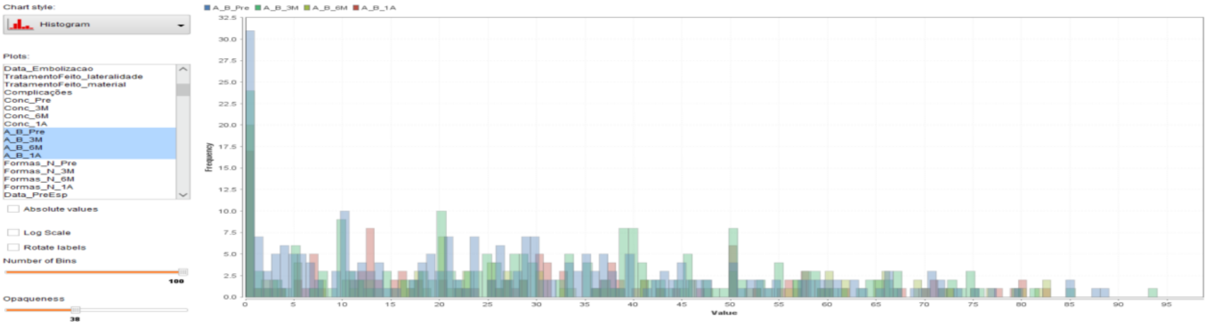


Chart style: Histogram

☐ Absolute values

☐ Log Scale

☐ Rotate labels

Number of Bins 100

Opaqueness 38

Figure 5.9 Histogram of sperm progressive motilities

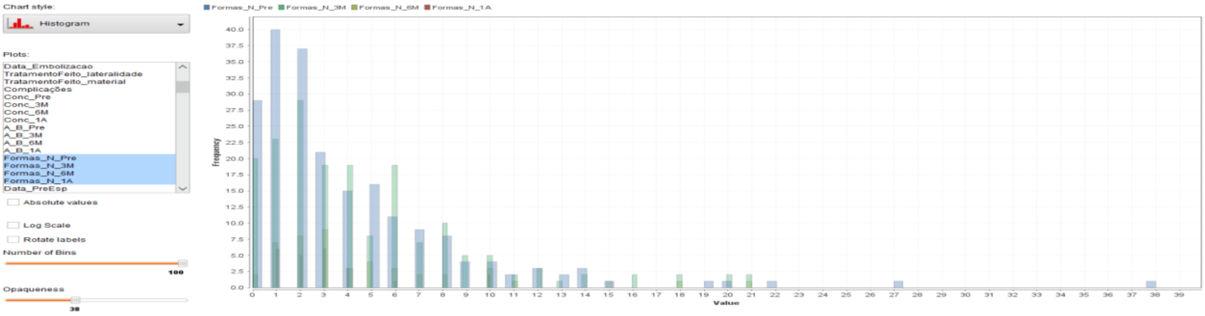


Chart style: Histogram

☐ Absolute values

☐ Log Scale

☐ Rotate labels

Number of Bins 100

Opaqueness 38

Figure 5.10 Histogram of sperm morphologies

5.3.2.1.4 Pregnancy related attributes

Out of the 230 women that we were able to determine the pregnancy outcome, 107 (i.e. 46.52%) got pregnant and mostly, for the first time. In fact, we have 85 women with the "*Prim_Sec"* attribute set to "*Primária*" out of these 107 pregnant women leading to a rate of 79.44%.

If we explore how these pregnant women conceived, we see that 54.21% (58/107) of them have conceived with an ART procedure, 38.31% (41/107) spontaneously and 7.48% (8/107) with both methods, through at least 2 different pregnancies. Out of these 107 pregnant women, most of them had 1 valid pregnancy: valid pregnancies are in this study pregnancies that have occurred after the embolization treatment (this condition was guaranteed/validated during data preparation). In fact, we have seen that 78.50% (84/107) had 1 pregnancy, 17.75% (19/107) had 2 pregnancies and 3.74% (4/107) had 3 pregnancies after the embolization treatment.

In terms of births, we see that 78.50% of these patients (84/107) were able to give birth to a child and 96.43% (81/84) of them, to an alive baby. 67.86% (57/84) of these couples had one child. Hence, we can say that the alive baby rate of the dataset is of 35.22% (i.e. 84 births minus 3 stillborn babies divided by the 230 patients of the pregnancy sample).

5.3.2.1.5 Infertility factors

Out of the 85 women with infertility factor information, 29.41% had anovulation (25/85), 23.53% had diseases related with tubal ((17 tubal + 1 endometrioses + 1 endometrioses with tubal + 1 anovulation with tubal) /85) and 17.65% had previous abortions (15/85).

Regarding male infertility factors, we have seen that half of them (50.31% 82/163) had the indication that the infertility cause was only from themselves. To look up for more details, we checked their corresponding semen classification before treatment and have seen that out of these 82 men, 20 had OligoAsthenoTeratozoospermia. Please note that the semen classifications indicated under the man infertility factor attribute were not considered for analysis since through the years the World Health Organization (WHO) has updated its threshold.

5.3.2.1.6 External factors

By analyzing the frequency of the values specified on all external factor attributes statistically described in section 5.3.1, we can characterize the male patients of the data set as follows:

- 48.04% of the patients (98/204) smoke;
- 30.17% of the patients (35/116) drink alcohol;
- 61.59% of the patients (85/138) undergone surgery before the embolization treatment;
- 21.98% of the patients (20/91) had parotitis disease before the embolization treatment, followed by the epididymis cysts 8.79% (8/91) and overweight condition 6.59% (6/91).
- 36.14% of the patients (73/202) have an occupation in an environment that has a certain level of toxicity, and the textile industry was the most frequent and toxic occupation (8/202) of this sample since most of these patients alleged that they work in an environment with high temperatures. Since the occupation attribute is highly fragmented (i.e has several different attributes with low frequencies), it lead us to generate a new attribute upon it to categorize the recorded occupations into whether they encompass toxic products or environments.

5.3.2.1.7 Severity grade vs varicocele´s laterality

To explore whether the severity grade of varicocele is related with the site on which the condition appears (i.e. if the varicocele appears on the right, left or both testes), several graphs were generated in RapidMiner, with the values of the attribute "*Grau_Varicoc*" and "*Lateralidade".*

If we analyze the *frequency* of the values of the "*Grau_Varicoc*" attribute, from the 211 patients that had as established severity grade 111 patients had severity grade II (52.61%), 67 patients had severity grade I (31.75%) and 33 patients had severity grade III (15.64%). Hence, most patients of the dataset have the severity grade I or II (84.36%).

If we analyze the *frequency* of the values of the "*Lateralidade"* attribute, we see that from the 218 patients where data was available, 178 had the condition on its left testicle (i.e. 81.65% 178/218).

By crossing that data with a *scatter graph*, we see that 176 out of 207 patients had the varicocele condition on the left testicle and with a grade I or II which includes 85.02% of the sample. In Figure 5.11, we present the *scatter graph* where we can see a higher dot concentration for the left testicle, indicated in the scatter graph with the name "*Esquerdo*", and for the severity grade II and I.

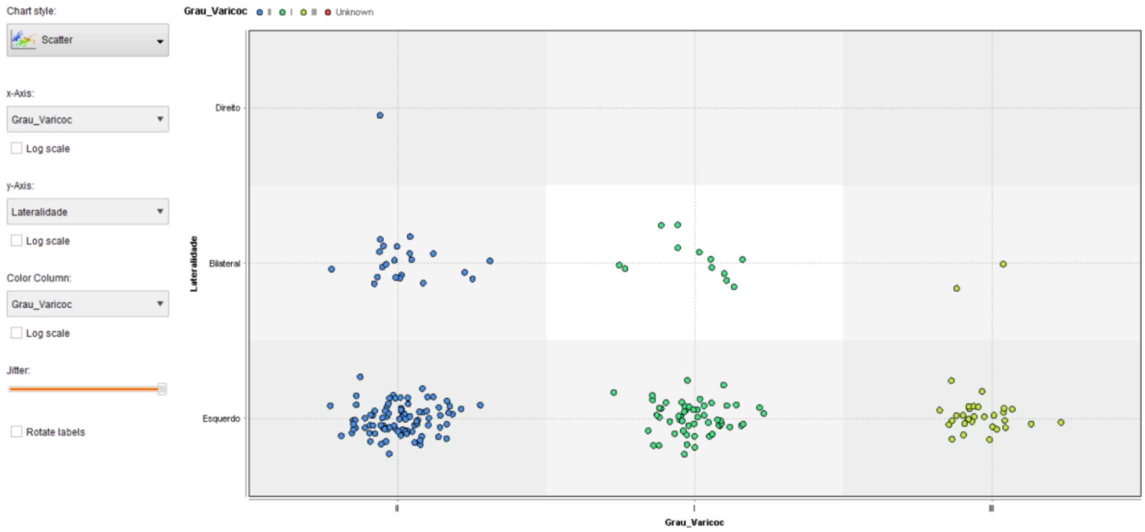


Chart style: Scatter

x-Axis: Grau_Varicoc

□ Log scale

y-Axis: Lateralidade

□ Log scale

Color Column: Grau_Varicoc

□ Log scale

Jitter:

□ Rotate labels

Figure 5.11 varicocele´s laterality by severity grade

5.3.2.1.8 Testicle volume vs varicocele´s severity grade

Through previous research, we have seen that patients with an advanced and untreated varicocele condition are prone to testis reduction since the veins on the scrotum do not irrigate well the testis; and therefore, cause testis atrophy (Aza Mohammed & Frank Chinegwundoh, 2009). Due to that information, we analyze the information of testis volume of some of the patients. Thus, of the 294 provided patients, we could retrieve information on testis volume for 44 patients and analyzed that information by generating the *scatter graph* depicted in Figure 5.12. By analyzing this *scatter graph*, we see that, according to diagnosis, 75% of the assessed patients (33 out of 44 patients), had normal to good testis volume (i.e. a testis volume equal or bigger than 20cc) and only 6 out of these 33 patients (18.18%), had a varicocele severity grade of III which tells us that most patients of this sample have treated the varicocele condition in its initial stage since most of them have lower severity grades and from normal to good testis volumes.

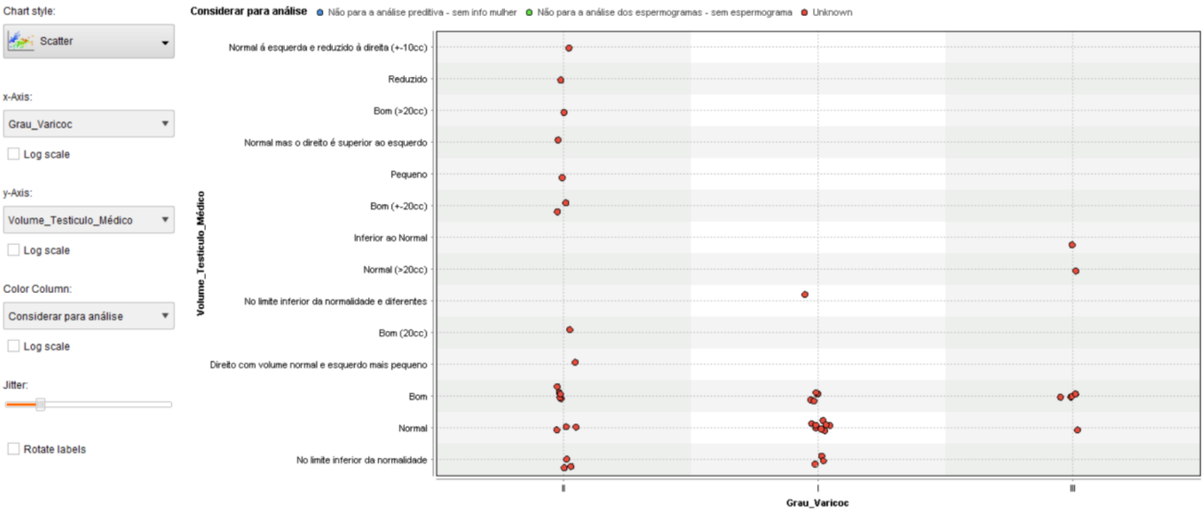


Chart style: Scatter

x-Axis: Grau_Varicoc

□ Log scale

y-Axis: Volume_Testiculo_Médico

□ Log scale

Color Column: Considerar para análise

□ Log scale

Jitter:

□ Rotate labels

Considerar para análise
● Não para a análise preditiva - sem info mulher
● Não para a análise dos espermogramas - sem espermograma
● Unknown
Volume_Testiculo_Médico
Normal á esquerda e reduzido à direita (+-10cc)
Reduzido
Bom (>20cc)
Normal mas o direito é superior ao esquerdo
Pequeno
Bom (+-20cc)
Inferior ao Normal
Normal (>20cc)
No limite inferior da normalidade e diferentes
Bom (20cc)
Direito com volume normal e esquerdo mais pequeno
Bom
Normal
No limite inferior da normalidade
II
I
III
Grau_Varicoc

Figure 5.12 Scatter graph of the testicle volume by the severity grade

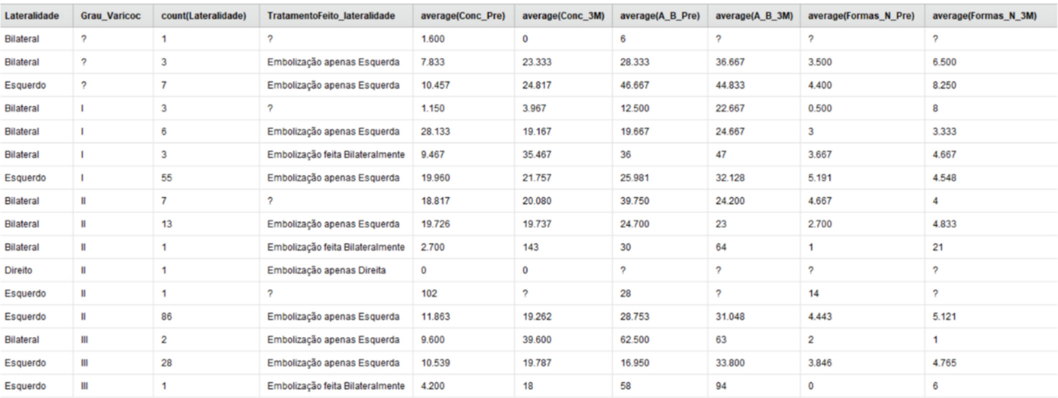
5.3.2.1.9 Diagnosed and Treated laterality vs sperm improvement

From the 293 patients assessed, we knew on which testicle the varicocele correction was carried out in 206. We have seen that 97.9% (200/206) of these 206 patients, underwent embolization on the left testicle. From these 200 patients, 12% (24/200) had the varicocele condition on both testis. The remaining 6 patients out of the 206 patients, underwent embolization on both testis (5 patients) or on the right testicle (1 patient). If we analyze initial diagnoses, the ones that had the procedure carried out on both testis, 4 out of the 5 patients treated, had initially the varicocele condition also diagnosed in both testis, and the remaining patient, had it diagnosed on the left testicle. Hence, the data set has 28 patients (24+4) with the varicocele condition diagnosed in both testis but 85.71% of them (24/28), only underwent embolization on the left testicle.

Due to the high percentage of varicocele correction on the left testicle, this attribute does not provide much information. However, despite the small sample size on patients with varicocele on both testicles (*n*=28), it was seen of clinical importance to explore if the ones that underwent embolization on both testicles (4 patients) had better sperm parameter values than the other patients (24 patients). To assess that situation, we have grouped sperm parameter values in the RapidMiner platform by varicocele laterality, severity grade and treated laterality (Table 5.8).

If we only focus on the patients that were embolized on both testicles, indicated with the word "*Bilateral*", and consider only the grouped sperm parameter mean values that have laterality information,", the severity grade, and its treated laterality, we see that the patients with a severity grade I or II and embolized in both testis have improved all sperm parameters at 3 months after the treatment. However, if we analyze the ones that were only embolized at their left testis, we see that for the ones with a severity grade of I, there is a mean sperm concentration decrease of 8.97 millions/ml at 3 months after the treatment although other sperm parameters improving with treatment; and for patients with severity grade II, we see that the progressive motility suffered a slight decrease of 1.7%. Therefore, these results suggest that embolization should be carried out in both testis when the patient also has varicocele in both testis. However, the small number of patients that we have as diagnosed and treated in both testis is to small (4 patients) to formulate a founded conclusion, and require further investigations.

Table 5.8 Sperm parameters values before and 3 months after the treatment grouped by the diagnosed laterality, its severity grade and its treated laterality



Lateralidade	Grau_Varicoc	count(Lateralidade)	TratamentoFeito_lateralidade	average(Conc_Pre)	average(Conc_3M)	average(A_B_Pre)	average(A_B_3M)	average(Formas_N_Pre)	average(Formas_N_3M)
Bilateral	?	1	?	1.600	0	6	?	?	?
Bilateral	?	3	Embolização apenas Esquerda	7.833	23.333	28.333	36.667	3.500	6.500
Esquerdo	?	7	Embolização apenas Esquerda	10.457	24.817	46.667	44.833	4.400	8.250
Bilateral	I	3	?	1.150	3.967	12.500	22.667	0.500	8
Bilateral	I	6	Embolização apenas Esquerda	28.133	19.167	19.667	24.667	3	3.333
Bilateral	I	3	Embolização feita Bilateralmente	9.467	35.467	36	47	3.667	4.667
Esquerdo	I	55	Embolização apenas Esquerda	19.960	21.757	25.981	32.128	5.191	4.548
Bilateral	II	7	?	18.817	20.080	39.750	24.200	4.667	4
Bilateral	II	13	Embolização apenas Esquerda	19.726	19.737	24.700	23	2.700	4.833
Bilateral	II	1	Embolização feita Bilateralmente	2.700	143	30	64	1	21
Direito	II	1	Embolização apenas Direita	0	0	?	?	?	?
Esquerdo	II	1	?	102	?	28	?	14	?
Esquerdo	II	86	Embolização apenas Esquerda	11.863	19.262	28.753	31.048	4.443	5.121
Bilateral	III	2	Embolização apenas Esquerda	9.600	39.600	62.500	63	2	1
Esquerdo	III	28	Embolização apenas Esquerda	10.539	19.787	16.950	33.800	3.846	4.765
Esquerdo	III	1	Embolização feita Bilateralmente	4.200	18	58	94	0	6

5.3.2.2 Evolution of semen and sperm categorizations through time

As seen previously, semen can be qualitatively and quantitatively analyzed. Hence, some of the first questions raised in this context was: which is the largest semen and sperm qualification group before and after the embolization treatment? Does it change during follow up? And how about the number of normal sperm parameters, does it change through time/with treatment? To address all these questions, several *crosstabs* were built and are presented and interpreted through the application of the *Chi-square* teste.

To identify the largest semen classification throughout the patient follow up times, a *crosstab* was built (Table 5.9) and a bar chart was generated based on its computed relative frequencies (Figure 5.13). As we can appreciate in Table 5.9, before the embolization treatment, the biggest semen classification is the OligoAsthenoTeratozoospermia (OAT) with 26.89% (64/238); 3 months after embolization, the biggest group is Normozoospermia with 19.90% (41/206) and 6 months later, Azoospermia with 24.19% (15/62) which continued to lead 12 months after the treatment with 39.47% (15/38) – all these values are highlighted in orange in Table 5.9 . Since there is a difference between the *relative frequencies* through time, we have analyzed if these differences occurred by chance or if they were influenced by the time at which the semen analysis was carried out, to assess if the semen classifications have improved with time.

With the *Chi-square test*, we have computed a *p* value less than 0.05 (i.e. 0.0000001582755) which tells us that there is enough evidence to conclude that there is a statistically significant relationship between the semen classification and the time when the semen analysis was carried out. Since the biggest *relative frequency* at 3 months after the embolization treatment is from patients with normal sperm parameters, we can say that the embolization treatment improved sperm parameters in 14.02% of cases (i.e. 19,90% with normozoospermia 3 months after the treatment minus 5.99% with normospermia before the treatment).

Table 5.9 Crosstab of semen classifications by patient´s follow up time

	Before treatment		3 Months		6 Months		12 Months		Total
	Count	% of total	Count	% of total	Count	% of total	Count	% of total	
Normozoospermia	14	5.88	41	19.90	7	11.29	3		65
Oligozoospermia	21	8.82	15	7.28	4	6.45	0	0	40
OligoAsthenozoospermia	27	11.34	11	5.34	2	3.23	3	7.89	43
OligoTeratozoospermia	31	13.03	16	7.77	2	3.23	3	7.89	52
Asthenoozoospermia	21	8.82	22	10.68	8	12.90	0	0	51
AsthenooTeratozoospermia	18	7.56	24	11.65	9	14.52	4	10.53	55
Teratozoospermia	14	5.88	22	10.68	3	4.84	3	7.89	42
OligoAstenoTeratozoospermia	64	26.89	28	13.59	12	19.35	7	18.42	111
Azoospermia	28	11.76	27	13.11	15	24.19	15	39.47	85
Total	238	100	206	100	62	100	38	100	544

Statistically significant p<0.05

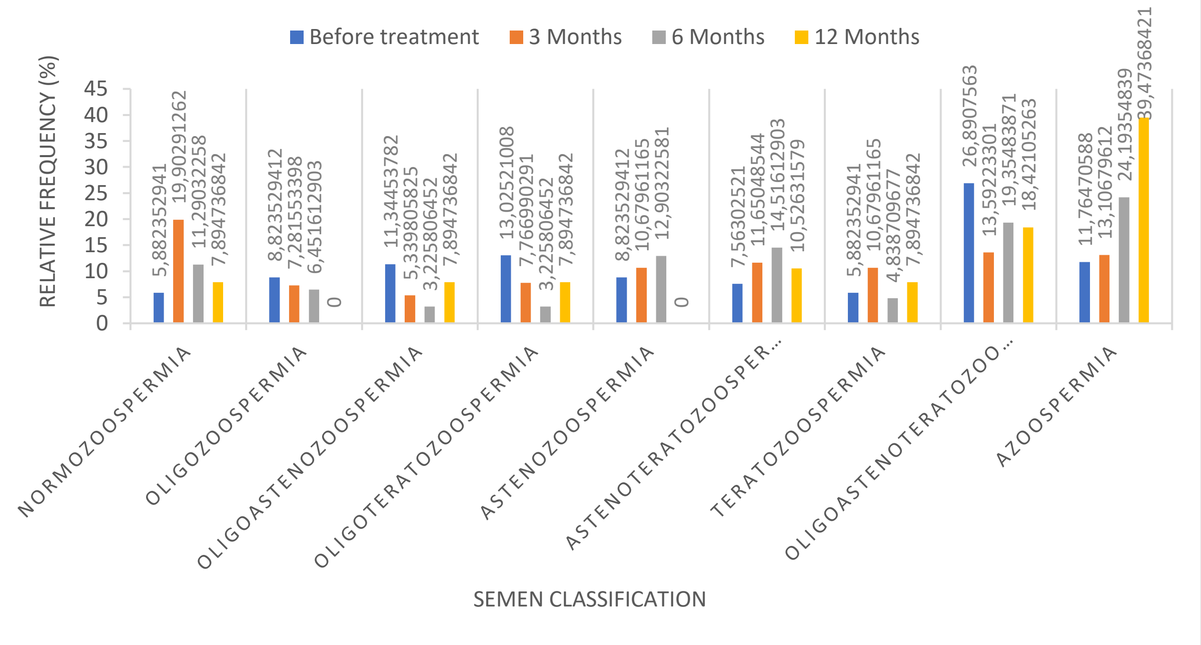


Figure 5.13 Evolution of the relative frequencies of the semen classifications through time

When we analyze the normality of the sperm parameter values through time, we see that there is a statistically significant relationship between the semen normality and when they were clinically assessed. In fact, after applying the *Chi-square* test to the *crosstab* presented in Table 5.10, the null hypothesis was rejected for all three sperm parameters with a p value under 0.001.

Table 5.10 Crosstab of sperm parameter normality according to follow up time

	Before treatment		3 Months		6 Months		12 Months		
	Count	% of Total	Count	% of Total	Count	% of Total	Count	% of Total	Total
Normal Concentration	67	28.15	109	52,91	27	43,55	10	26,32	213
Abnormal Concentration	171	71.85	97	47,09	35	56,45	28	73,68	331
Total	238	100.00	206	100.00	62	100.00	38	100.00	544
Normal Progressive Motility	80	33.61	94	45.63	16	25.81	9	23.68	199
Abnormal Progressive Motility	130	54.62	85	41.26	31	50.00	14	36.84	260
Missing Progressive Motility	28	11.76	27	13.11	15	24.19	15	39.47	85
Total	238	100.00	206	100.00	62	100.00	38	100.00	544
Normal Morphology	83	34.87	89	43.20	21	33.87	6	15.79	199
Abnormal Morphology	127	53.36	90	43.69	26	41.94	17	44.74	260
Missing Morphology	28	11.76	27	13.11	15	24.19	15	39.47	85
Total	238	100.00	206	100.00	62	100.00	38	100.00	544

Statistically significant p<0.05

In Table 5.10, we can see that for sperm progressive motility, as well as for the sperm morphology, their corresponding missing values were also covered by the *Chi-square* test. In fact, these missing values were not completely obtained by chance since they express the patients with azoospermia; and therefore, must also be computed by the *Chi-square* test.

By analyzing the values presented in Table 5.10, we see that 3 months after embolization the proportion of normal sperm concentrations tend to be 24.76% higher than before the treatment (i.e. 52.91% 3 months after the treatment minus 28.15% before the treatment) for the target population. The same trend occurs in the two other sperm parameters but with a lower improvement: 12.02% higher for sperm progressive motility and 8.33% higher for sperm morphology. Unfortunately, these improvements decrease at 6 months and further on at 12 months after the treatment as seen in Figure 5.14. Moreover, at 12 months, the sample proportion of normal sperm parameters is lower than before the treatment: -1.84% for sperm concentration (i.e. 26.32% at 12 months minus the corresponding 28.15% before the treatment), -9.93% for sperm progressive motility and -19,08 for sperm morphology.

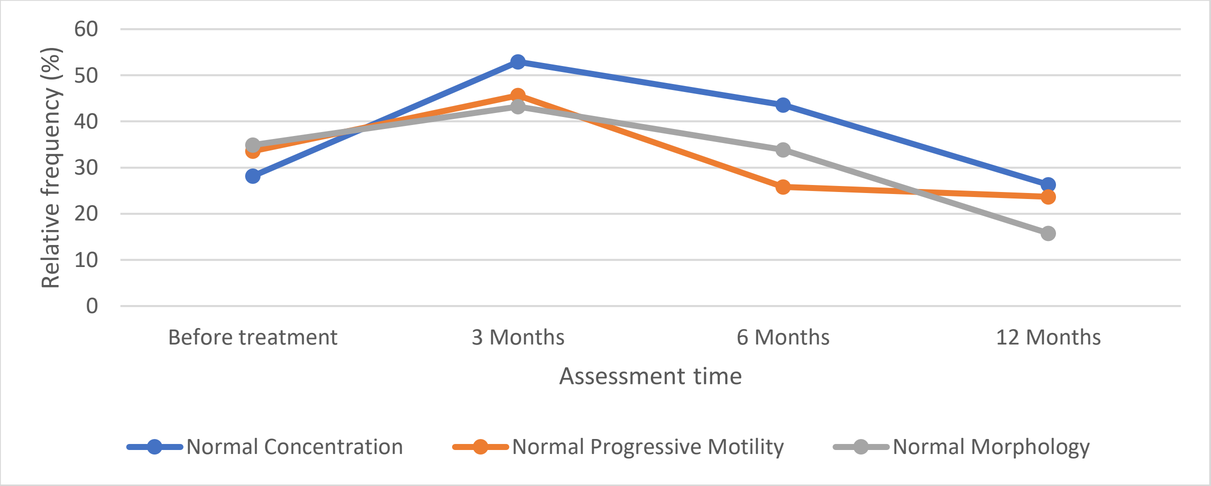


Figure 5.14 Evolution of the frequencies of the sperm parameter´s normality through time

To assess the evolution of the abnormality of sperm parameter values through time, we have analyzed the number of abnormal (i.e. altered) sperm parameters that were identified in each semen analysis report (described in Table 4.1). When we assess if there is an association between the number of altered sperm parameters with when the semen analysis report was carried out (Table 5.11), we see that there is a statistical significant relationship between both attributes (p<0.001). By analyzing the higher values highlighted in orange in Table 5.11, we see that at each patient´s follow up time, the biggest sample proportion is always held by samples with one altered sperm parameter, which increases throughout the year but sees a slight depletion after 12 months. Figure 5.15 shows that trend by presenting a time series graph of the sample proportions presented in Table 5.11. By analyzing this graph, we can also see that the proportion of patients without altered sperm parameter values (i.e. patients with *Normozoospermia*) increases at 3 months and decreases gradually until 12 months but at the end, it is still 2.01% higher than before the treatment which is good since this sample represent the patients with normal semen. Furthermore, we see that the sample proportion of patients with OAT decreases throughout the follow up year in -8.47% which is also good (i.e. 18.42% at 12 Months for OAT, minus 26.89% which is the corresponding sample proportion that exists before the treatment).

Table 5.11 Number of abnormal sperm parameters through time

	Before treatment		2 Months		6 Months		12 Months		Total
	Count	% of Total	Count	% of Total	Count	% of Total	Count	% of Total	
Normozoospermia	14	5.88	41	19.90	7	11.29	3	7.89	65
1 altered sperm parameter	84	35.29	86	41.75	30	48.39	18	47.37	218
2 altered sperm parameters	76	31.93	51	24.76	13	20.97	10	26.32	150
OAT	64	26.89	28	13.59	12	19.35	7	18.42	111
Total	238	100	206	100	62	100	38	100	544

Statistically significant p<0.05; OAT stands for OligoAsthenoTeratozoospermia

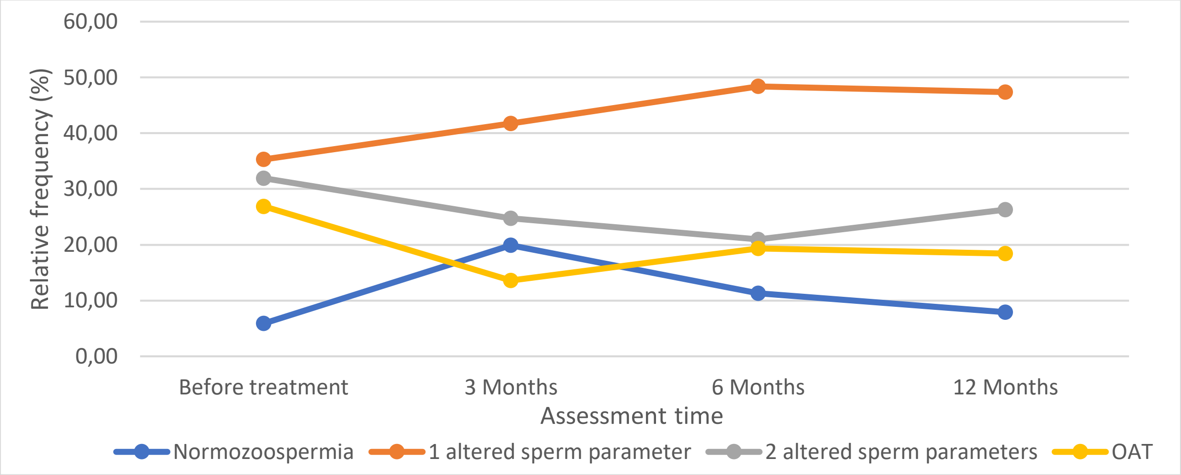


Figure 5.15 Evolution of the relative frequencies of the number of abnormal sperm parameters through time

5.3.2.3 Evolution of sperm parameters values through time

If we analyze sperm mean values (Figure 5.16) before and at 12 months after the treatment, we see that the sperm concentration goes from being under the 15 million/ml threshold (i.e. 13.93 million/ml) to slightly above this threshold (i.e. 15.78 million/ml) suggesting improvement with treatment. To assess statistical significance, the ANOVA test was applied upon the mean of the preprocessed sperm parameter values at each follow up time (i.e. before the treatment, 3,6 and 12 after the treatment) and the following results were generated:

- For sperm concentration -> Statistically significant difference since $p < 0.05$ (p=0.017);
- For sperm progressive motility -> No statistical significance since $p > 0.05$ (p=0.376);
- For sperm morphology -> Statistically significant difference since $p < 0.05$ (p=0.001).

By analyzing the time series graph in Figure 5.16, as well as the bar graph depicted in Figure 5.17 of the sperm parameter mean values according to patient follow up time, we can say that the embolization treatment only significantly improves sperm concentrations (p=0.017) and morphologies (p=0.001) which still confirms the importance of the procedure.

Furthermore, if we analyze the dimension of the population at each patient follow up time (see Table 5.12, Table 5.13 and Table 5.14 at row "*n"*), we see that the information regarding sperm morphologies at 6 and 12 months, is based on a reduced population (i.e. *n* at 6 months is 47 patients and at 12 months, is 23 patients) in comparison with sperm concentrations and sperm progressive motilities. This situation occurs because in some cases, sperm morphology cannot be assessed or was not evaluated. Please note that for any ART procedure a semen analysis is always carried out, but only concentration and progressive motility are usually assessed. Nevertheless, we have also considered these ART semen analysis results although we do not have the corresponding sperm morphologies.

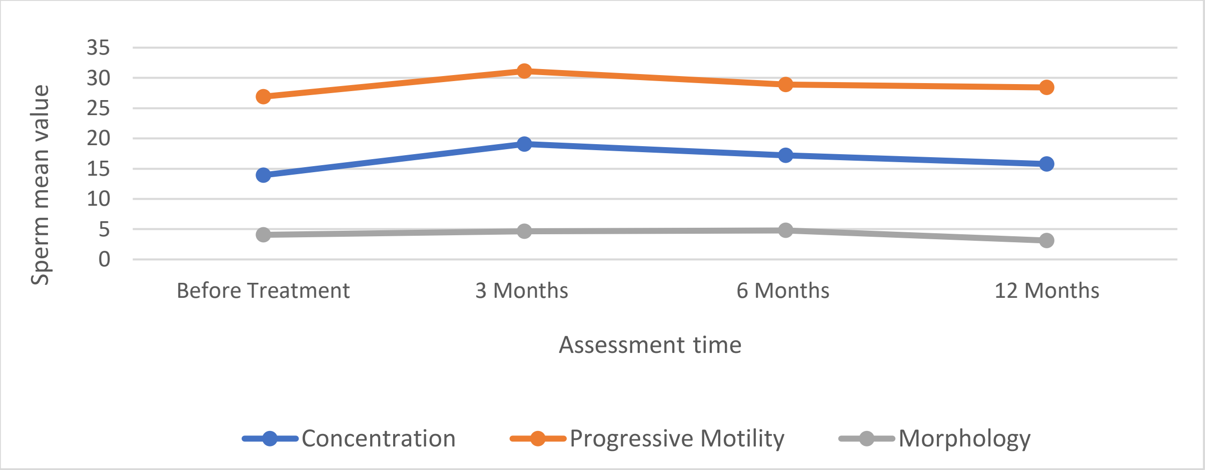


Figure 5.16 Time series of sperm parameters´ evolution – analysis of mean values

Table 5.12 Main statistical results for sperm concentrations

Concentration	Before treatment	3 months	6 months	12 months
Mean	13.93	19.07	17.21	15.78
SD	23.64	24.94	24.07	18.66
n	281	245	131	137

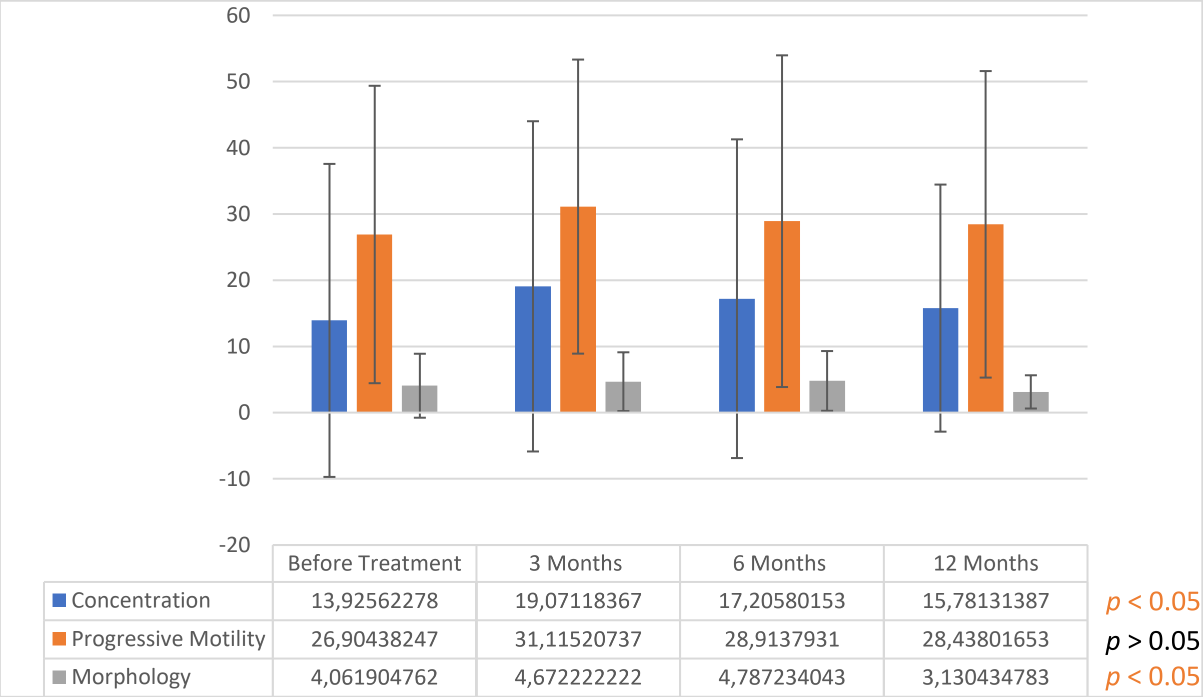
Table 5.13 Main statistical results for progressive motilities

Progressive Motility	Before treatment	3 months	6 months	12 months
Mean	26.90	31.12	28.91	28.44
SD	22.46	22.21	25.05	23.15
n	251	217	116	121

Table 5.14 Main statistical results for morphologies

Morphology	Before treatment	3 months	6 months	12 months
Mean	4.062	4.672	4.79	3.13
SD	4.83	4.44	4.51	2.51
n	210	180	47	23

By analyzing the above standard deviations, we can say that the standard deviations are roughly equal for each sperm parameter through the follow up times. In fact, for the ANOVA test, it is good enough if the largest standard deviation is less than double the smallest standard deviation (Sullivan, 2011), which is in our case true.



	Before Treatment	3 Months	6 Months	12 Months	
🟦 Concentration	13,92562278	19,07118367	17,20580153	15,78131387	$p < 0.05$
🟧 Progressive Motility	26,90438247	31,11520737	28,9137931	28,43801653	$p > 0.05$
■ Morphology	4,061904762	4,672222222	4,787234043	3,130434783	$p < 0.05$

Figure 5.17 Bar Graph of the mean values of the sperm parameters with its standard deviations by the patient ´s follow up times

5.3.2.4 Sperm parameter values vs pregnancy outcome

Since this study aims to predict the success of the embolization treatment, the assessment of sperm parameter mean values as related to pregnancy outcomes is very important. Hence, this subsection aims to present the analysis with a partitioned data set.

To assess if sperm parameter values are different in these two groups, the two-sample K-S test of the Kolmogorov Smirnov statistical test was in the RapidMiner Platform applied with the operator “Kolmogorov Smirnov test”. This statistical test was selected since sperm parameters values are continuous and without a well know underlying data distribution – even though our sample seems to be positively skewed. The results are presented in Table 5.15 and can be interpreted as follows:

- *attribute* -> Name of the attribute tested;
- *p_value* -> Statistical results of the two-sample K-S test that indicates that there is a statistical difference if p_value is below 0.05;
- *null_hypothesis_rejected* -> Boolean value that indicates whether the null hypothesis of the two samples K-S test is rejected.

Table 5.15 Kolmogorov Smirnov test results on the comparison of seminal parameters per pregnancy outcome

Attribute	p_value	Null_hypothesis_rejected
Concentration before treatment	0.549	false
Concentration at 3 months	0.040	true
Concentration at 6 months	0.000	true
Concentration at 12 months	0.002	true
Progressive motility before treatment	0.362	false
Progressive motility at 3 months	0.002	true
Progressive motility at 6 months	0.000	true
Progressive motility at 12 months	0.000	true
Morphology before treatment	0.002	true
Morphology at 3 months	0.002	true
Morphology at 6 months	0.000	true
Morphology at 12 months	0.000	true

As previously seen, the K-S operator returns true if the null hypothesis can be rejected, which means that both samples are different in shape and their population´s mean values are different. The sperm parameters with a different data distribution are highlighted in orange. With this in mind, and by analyzing the results presented in Table 5.15, we see that, notwithstanding the sperm concentration (*p*=0.549) and progressive motility (*p*=0.362) before embolization, all other sperm parameter values have a different data distribution at any patient follow up time. These results allow us to say that, before the treatment, the values of sperm morphology are statistically different in terms of data distribution for patients that we could achieve a pregnancy, in comparison to the ones that did not.

If we analyze the values disclosed in Table 5.16, we see that patients that got their partner pregnant have higher mean values in all three sperm parameters and at all patient follow-up times after the embolization treatment. To assess whether these improvements are significant, the ANOVA test was applied. The computed results are presented in Table 5.17 where statistically significant differences are highlighted in orange. Please note that the attribute Morphology at 12 months is not highlighted since it is not regarded as interesting due to the small sample size.

Table 5.16 Sperm parameter mean values per pregnancy outcome and assessment time

Sperm Parameter ** ANOVA *p*<0.05 *** ANOVA *p*<0.01		Pregnancy Outcome	*n*	mean	mean difference	SD
Concentration	before treatment	Yes	107	14.5	-0.4	21.6
		No	119	14.9		27.2
	3 months	Yes	94	22.9	4.8	24.2
		No	107	18.1		27.5
	6 months**	Yes	50	22.9	8.2	29.5
		No	65	14.7		20.6
	12 months	Yes	60	18.8	4.3	19.5
		No	69	14.5		18.2
Progressive Motility	before treatment**	Yes	102	29.9	6.9	23.3
		No	107	23.0		20.8
	3 months	Yes	92	33.2	3.3	21.6
		No	92	29.9		21.1
	6 months	Yes	49	33.5	6.2	25.1
		No	57	27.3		25.6
	12 months	Yes	58	30.8	4.1	23.6
		No	58	26.7		22.8
Morphology	before treatment	Yes	89	4.0	0.4	5.0
		No	87	3.6		3.5
	3 months***	Yes	78	5.5	1.6	5.0
		No	74	3.9		3.4
	6 months	Yes	22	5.0	0.5	3.2
		No	19	4.5		5.2
	12 months***	Yes	10	4.0	1.1	3.3
		No	12	2.9		1.4

Table 5.17 ANOVA results for sperm parameter differences between pregnancy results

Sperm Parameter	ANOVA p value	Difference statistically significant?
Concentration before treatment	0.903	No
Concentration at 3 months	0.165	No
Concentration at 6 months	0.015	Yes
Concentration at 12 months	0.081	No
Progressive motility before treatment	0.018	Yes
Progressive motility at 3 months	0.236	No
Progressive motility at 6 months	0.064	No
Progressive motility at 12 months	0.171	No
Morphology before treatment	0.488	No
Morphology at 3 months	0.004	Yes
Morphology at 6 months	0.327	No
Morphology at 12 months	0.000	Yes

By analyzing the ANOVA test results (Table 5.17), we see that even though sperm progressive motility does not significantly improve after the embolization treatment, its value highlights a data pattern which indicates that sperm progressive motility before the treatment of patients that were able to get their partner pregnant is statistically different (ANOVA test, *p*=0.018), despite their data distribution being the same (two sample KS test, *p*=0.362). Furthermore, we see that sperm morphology at 3 months is significantly different in its data distribution (two sample KS test, *p*=0.002) and values (ANOVA test, *p*=0.004). Since our sample size for the *Formas_N_1A* is of 22 patients, we only consider the significance given to the sperm morphology at 3 months that includes 152 patients which is a more acceptable sample dimension to formulate a valid conclusion. For sperm morphology before the treatment that showed a significant difference in its data distribution with the two sample KS test, their values are not significantly different (ANOVA test gave *p*=0.488). Despite not having a significant difference on all sperm parameter values, the mean values of all sperm parameters are all higher in group b (*Gravidez=Sim*) after the treatment which suggests that the patients that got their partner pregnant had a greater response to the treatment with a statistical significance on the sperm morphology at 3 months (*p*=0.004) and at 6 months on the sperm concentration (*p*=0.015).

To analyze the differences between the mean values presented under the column name "*mean*" of Table 5.16, graphs were generated using the RapidMiner platform. Each graph below depicts the mean values of each sperm parameter: in Figure 5.18, we can see the mean sperm concentration values, in Figure 5.19, mean sperm progressive motility values and in Figure 5.20, mean sperm morphology values.

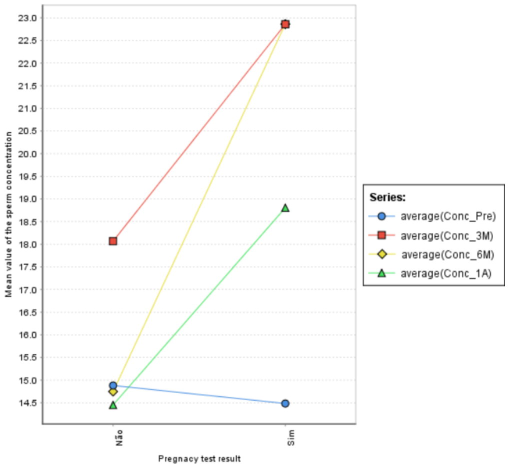


Figure 5.18 Means of sperm concentrations

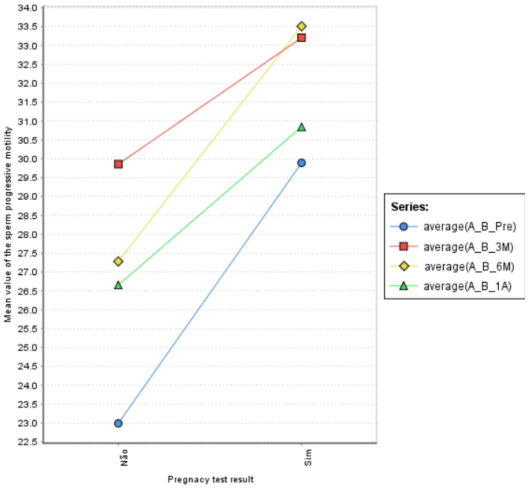


Figure 5.19 Means of sperm progressive motility

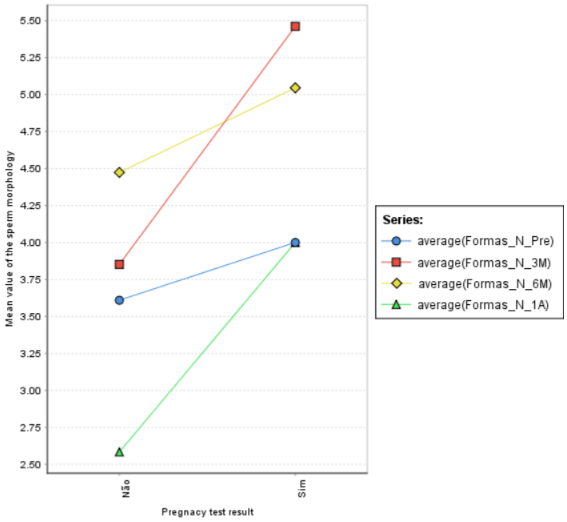


Figure 5.20 Means of sperm morphology

5.3.2.5 Time took to conceive vs type of conception

If we analyze the values of the "Gravidez_após_emb" attribute, we see that the dataset has 105 instances filled with the number of months a couple took to conceive after the embolization treatment. Moreover, we also see, based on quartiles values, that 50% of the male patients get their partner pregnant between the 9th and the 21th month after the embolization treatment. However, on average, couples get pregnant around the 13th month and with a high standard deviation of ±15.09 months.

To identify a pattern on when the pregnancies are achieved after the embolization, we have built a *bar graph* to analyze the number of pregnancies per the number of months the couple took to conceive. If we analyze the resultant *bar graph* depicted in Figure 5.21, we see that in some months we have much more pregnancies than others (e.g. at 3, 4, 6, 8, 9, 12, 17, 24,26, 28 and 48 months after the embolization we have much more pregnancies than in the other months) encompassing 54 out of the 105 patients assessed (51.43%). To see if there is a statistically significant difference, we have applied an ANOVA and have seen that the number of pregnancies occurring at 3, 4, 6 etc. months, is statistically significantly different (p=0.016) than the number of pregnancies occurring at 0, 1, 2, 5 etc. months where we clearly see less pregnancies happening.

To have some insights on the causes of this difference, we have analyzed how the patient partners have conceived (i.e. if they have conceived spontaneously or/and with an ART procedure after the treatment). The group of patients that stands out in Figure 5.21 (i.e. group of patients that have conceived at the month 3, 4, 6 etc) are presented in Table 5.18 with the name "Stand out" and the other patients (i.e. group of patients that have conceived at the month 0, 1, 2, 5 etc.), with the name "Does not stand out". If we analyze the *frequencies* (count of patients by type of conception) presented in Table 5.18 we see that the ART procedures are the most frequent among both groups. However, the number of ART procedures performed in the "*Stand out*" group is lower than the "Does not stand out" group which in turn has more spontaneous pregnancies. Moreover, the *relative frequency* of spontaneous pregnancies tends to be 15% higher in the "*Stand out*" group in comparison with the other group. However, the highest *relative frequency* gap between both groups are for the patients that after the embolization treatment had more than one pregnancy, one with an ART procedure, the other spontaneously (that information can be analyzed under the column name “Both”). In fact, the “*Stand out*” group tends to have 25% more pregnancies and with both conceptions’ methods. In terms of the relationship between the two subjects, we see that there is no significant relationship between these groups of patients and the type of conception they used (*Chi-square* test *p*=0.42)

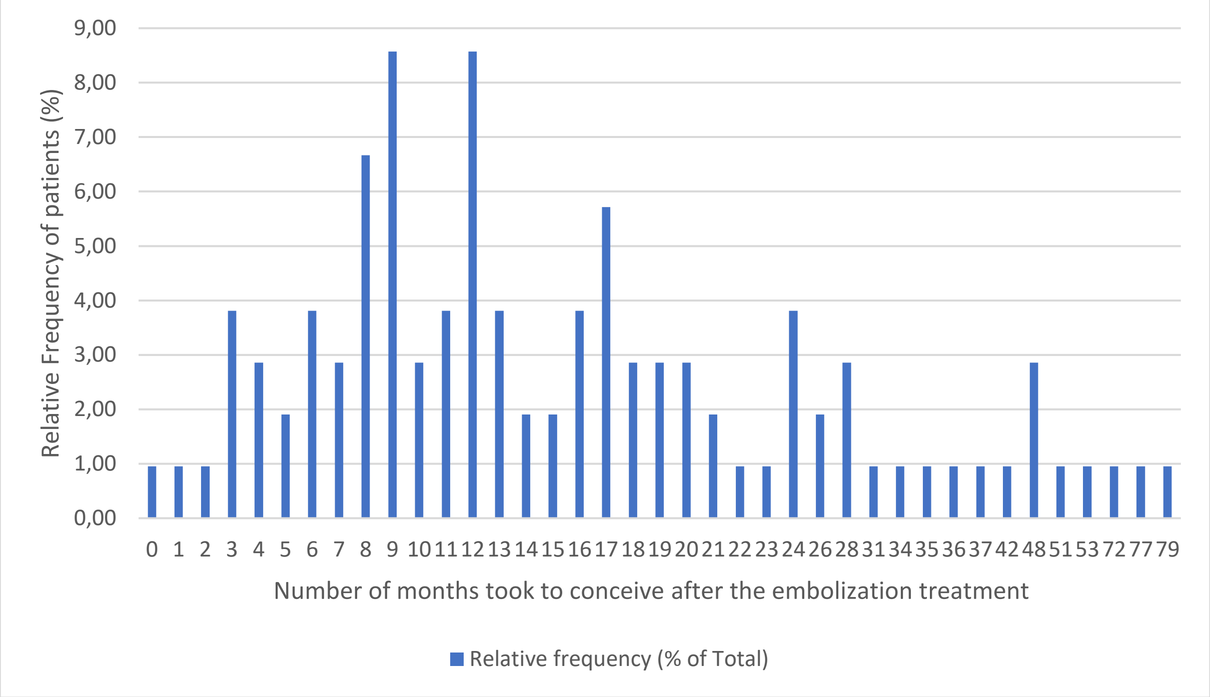


Figure 5.21 Relative frequency of patients by the number of months they took to conceive

Table 5.18 Crosstab of groups of patients by the type of conception

	Type of conception						Total
Groups of patients	ART		Spontaneously		Both		
	Count	% of Total	Count	% of Total	Count	% of Total	
Stand out	26	45.61	23	57.5	5	62.5	54
Does not stand out	31	54.39	17	42.5	3	37.5	51
Total	57	100	40	100	8	100	105

Statistically not significant p>0.05

If we consider the first 12 months after the embolization treatment – which is when the male patient follow up time was carried out – we see that 48.57% (51/105*100) of the 105 couples were able to conceive until the end of the year. In Figure 5.22 we present a *cumulative relative frequency bar graph* that depicts that reality since the height of the bar of the graph at 12 months is quite half the height of the last cumulative bar which represents all 105 instances

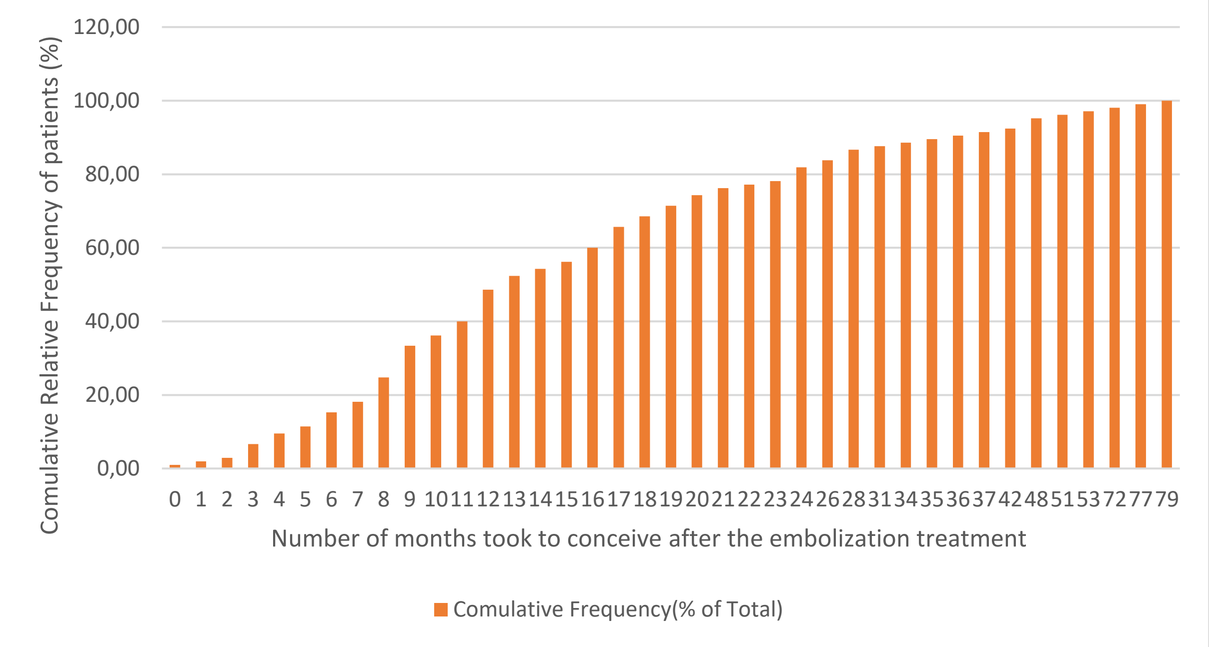


Figure 5.22 Cumulative frequencies of patients by the number of months they took to conceive

Since we have previously seen that the sperm parameter values improve after the embolization treatment, we have applied an ANOVA test, to further on inspect the difference in sperm parameter values between both patients' groups.

If we analyze the obtained results presented in Table 5.19, we see that the mean values of the sperm concentrations are significantly different at 12 months after embolization between the "*Stand out*" and the "Does not stand out" patient group (*Concentration at 12 months* has *p*=0.011) which corresponds to the 12th month when the highest percentage of patients achieved pregnancy (8.57%) along with the 9th month. In terms of the other sperm parameters, we see that only the sperm morphology at 6 months is significantly different (*Morphology at 6 months* has *p*=0.013).

Table 5.19 Statistical parameters per sperm parameter and patient group

Sperm Parameter	Patients group	n	Mean	SD	ANOVA (*p*)
Concentration before treatment	Does not Stand out	51	13.89	21.36	0.754
	Stand out	54	15.24	22.35	
Concentration at 3 months	Does not Stand out	45	27.39	28.92	0.086
	Stand out	47	19.21	18.51	
Concentration at 6 months	Does not Stand out	17	21.29	26.60	0.728
	Stand out	32	23.29	31.63	
Concentration at 12 months	Does not Stand out	33	23.11	20.98	0.011
	Stand out	26	13.46	16.83	
Progressive motility before treatment	Does not Stand out	49	32.12	23.91	0.389
	Stand out	51	28.16	23.04	
Progressive motility at 3 months	Does not Stand out	45	35.20	22.20	0.393
	Stand out	45	31.53	21.60	
Progressive motility at 6 months	Does not Stand out	17	32.77	23.15	0.746
	Stand out	31	34.36	26.76	
Progressive motility at 12 months	Does not Stand out	31	31.90	22.35	0.635
	Stand out	26	29.65	25.77	
Morphology before treatment	Does not Stand out	45	3.13	2.69	0.090
	Stand out	43	4.83	6.60	

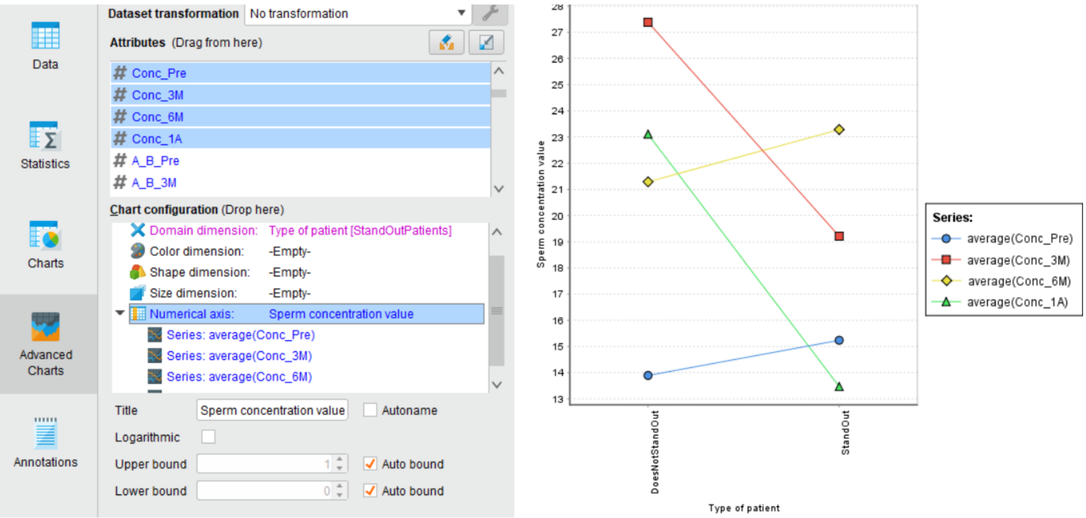
Sperm Parameter	Patients group	n	Mean	SD	ANOVA (*p*)
Morphology at 3 months	Does not Stand out	39	4.82	5.05	0.146
	Stand out	37	6.24	4.91	
Morphology at 6 months	Does not Stand out	8	5.75	3.11	0.013
	Stand out	13	4.23	3.03	
Morphology at 12 months	Does not Stand out	6	3.67	3.39	0.232
	Stand out	4	4.50	3.70	

In Figure 5.23, on the right, we have the graph for sperm concentrations and on the left, the configurations set in the *advanced charts* option of the RapidMiner platform to generate that type of graph (since other similar graphs were analogously configurated, we will from now on only present the generated graph). If we focus on the higher sperm concentration mean values that the “Stand out” group has, we see that they had a higher sperm concentration before and 6 months after the treatment but they are not significantly higher statistically (*Conc_Pre* has *p*=0.754 and *Conc_Pre* has *p*=0.728). However, we see that the mean value of the sperm concentration at 12 months is significantly lower statistically (*Conc_1A* has *p*=0.011) than the “Does not Stand out group” with a mean value of 13.46 millions/ml in contrast to the good concentration average that the “Does not Stand out group” has (23.11 millions/ml). If we check how these 26 “Stand out” patients conceived at the 12th month, we see that 73.08% (19/26) have conceived with an ART procedure: 19 with ART, 3 spontaneously and 4 with both methods since as of the 12th month they had 2 pregnancies, one with ART and the other one spontaneously.

In Figure 5.24, we see that the patients of the “Stand Out” group have a higher sperm progressive motility mean value at 6 months but it is not significant (*A_B_6M* has *p*=0.746).

If we analyze the sperm morphologies depicted in Figure 5.25, we see that the “Stand Out” group does not have significantly better sperm morphology before and 3 months after the treatment (Formas_N_Pre *p*=0.090 and Formas_N_3M *p*=0.146) but has a significantly lower sperm morphology at 6 months after the treatment (Formas_N_6M *p*=0.013). If we analyze the 13 patients of that sample we see that 6 of them conceived with ART procedures, 5 spontaneously and 2 with both methods.

Contrarily to what we had initially expected, the “Stand out” group distinguished itself from the “Does not stand out group” from its statistically significantly lower mean values in one of its sperm parameters: at 6 months, with a lower mean value of its sperm morphology and at 12 months, with a lower mean value of its sperm concentration. Since the sample size of the sperm morphology at 6 months is of 8 patients, for the “Does not stand out”, and 13 patients, for the “Stand out” group, we will not consider the significance obtained for sperm morphology in our final conclusions, due to the small sample size at 6 months.



Dataset transformation: No transformation

Attributes (Drag from here)

- Conc_Pre
- Conc_3M
- Conc_6M
- Conc_1A
- A_B_Pre
- A_B_3M

Chart configuration (Drop here)

- Domain dimension: Type of patient [StandOutPatients]
- Color dimension: -Empty-
- Shape dimension: -Empty-
- Size dimension: -Empty-
- Numerical axis: Sperm concentration value
  - Series: average(Conc_Pre)
  - Series: average(Conc_3M)
  - Series: average(Conc_6M)

Title: Sperm concentration value □ Autoname

Logarithmic □

Upper bound 1 ☑ Auto bound

Lower bound 0 ☑ Auto bound

Figure 5.23 Mean values of the sperm concentrations by the patient group on RapidMiner

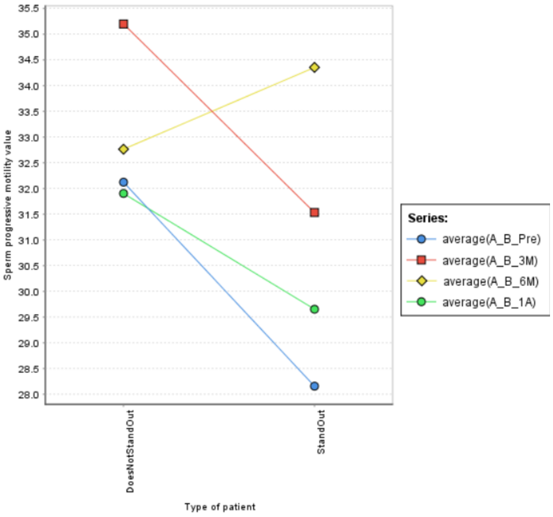


Figure 5.24 Mean values of the sperm progressive motilities by the patient group

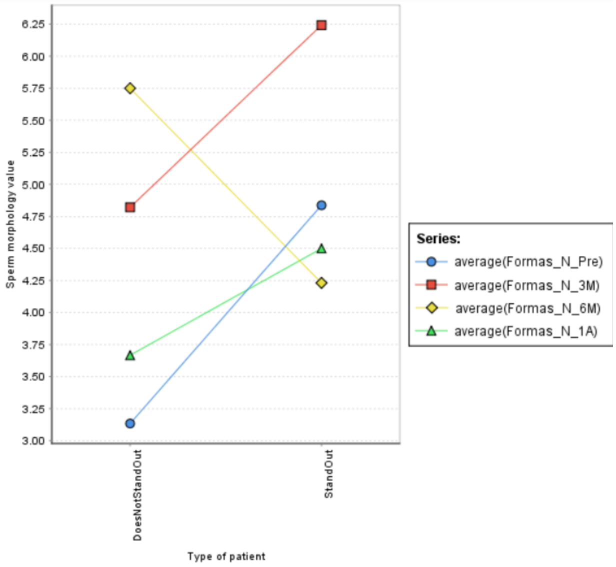
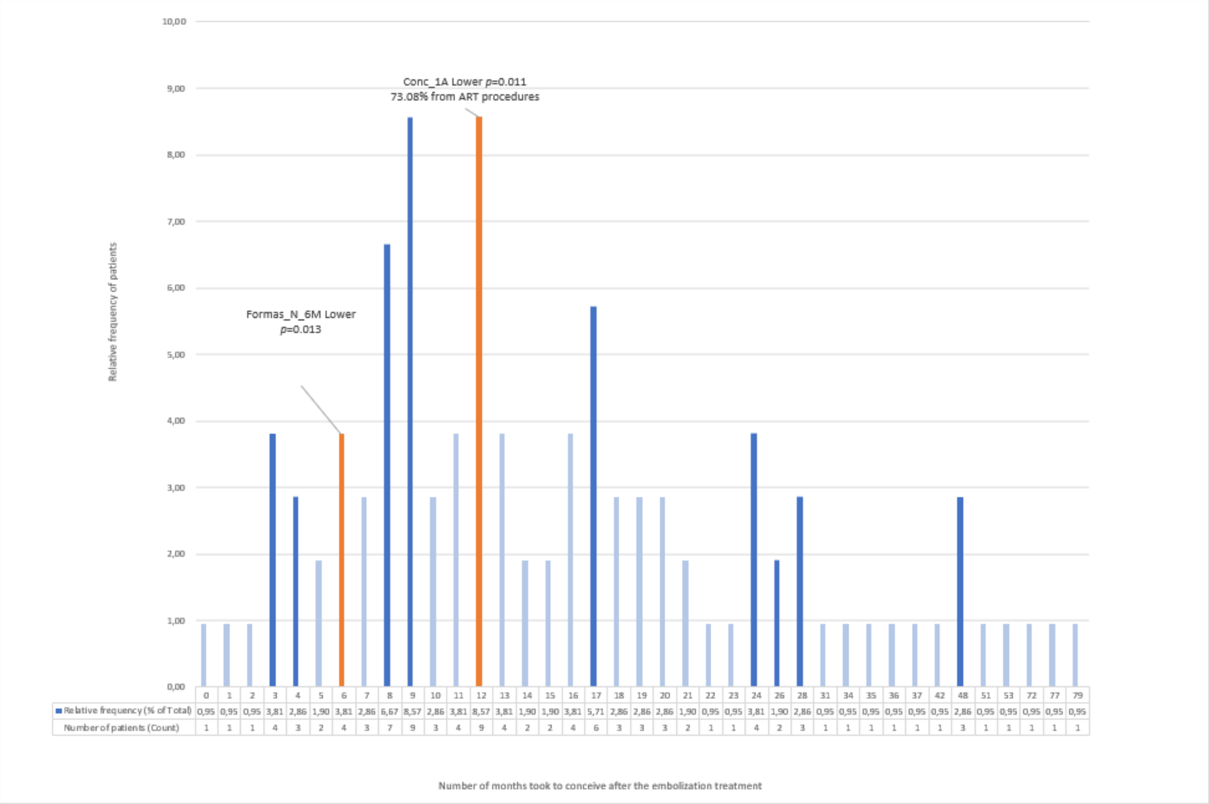


Figure 5.25 Mean values of the sperm morphologies by the patient group

To sum up the findings on this matter, we have built the *bar graph* presented in Figure 5.26 that depicts the statistically significant findings on the perspective of this “Stand out” patient group. This sum up *bar graph* as the following legend:

- Light blue bar = month with a normal number of patients that achieved pregnancy (i.e. the “Does not stand out” patients’ group). It encompasses 51 patients;
- Orange and strong blue bar = month that has a significant number of patients that achieved pregnancy (*p*=0.016) (i.e. the “Stand out” patients’ group) in comparison to the other months. These bars encompass 54 patients;
- Orange bar = month that have a statistically significant number of patients that achieved pregnancy and has a significant difference in one of its sperm parameter values;
- Bar´s notes = indication of the statistically significant difference on sperm parameter values of the patients in the “Stand out” group and other important findings.

Under the *bar graph* we can see the relative frequency of patients that achieved pregnancy by month the pregnancy, and below it, we have the number of patients encompassed in each of these months which in total covers the 105 patients assessed in this analysis.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
■ Relative frequency (% of Total)	0,95	0,95	0,95	3,81	2,86	1,90	3,81	2,86	6,67	8,57	2,86	3,81	8,57	3,81	1,90	1,90	3,81	5,71	2,86
Number of patients (Count)	1	1	1	4	3	2	4	3	7	9	3	4	9	4	2	2	4	6	3

	19	20	21	22	23	24	26	28	31	34	35	36	37	42	48	51	53	72	77	79
■ Relative frequency (% of Total)	2,86	2,86	1,90	0,95	0,95	3,81	1,90	2,86	0,95	0,95	0,95	0,95	0,95	0,95	2,86	0,95	0,95	0,95	0,95	0,95
Number of patients (Count)	3	3	2	1	1	4	2	3	1	1	1	1	1	1	3	1	1	1	1	1

Figure 5.26 Sum up bar graph with the significant sperm parameter results

5.3.2.6 Most correlated attributes

As suggested by Han *et al*. (2012) correlation analysis can be performed with the Pearson correlation test, between numeric attributes, or by the Chi-square test, between nominal or discrete attributes (Wujek, Hall, & Güneș, n.d.). Since the aim of this study is to predict the success of the embolization treatment through pregnancy outcomes, we have assessed the relation of the attributes with the selected label attribute to identify the ones that are more related with the pregnancy outcome.

Pearson correlation measures showed that sperm morphology at 12 months "*Formas_N_1A*" (*r*=-0.286) and the time of the treatment "*Data_Embolização*" (*r*=0.204) were the attributes that presented the highest correlation with the pregnancy; and hence, all correlations were seen as minor for the "*Gravidez*" *label attribute* (see Table 5.20).

Table 5.20 Pearson Correlation results

Attribute Name	Pearson Correlation with the pregnancy outcome
Idade_H	0.021
Idade_M	0.156
Tempo_Infert	0.154
Data_Embolizacao	0.204
Conc_Pre	0.008
Conc_3M	-0.092
Conc_6M	-0.161
Conc_1A	-0.115
A_B_Pre	-0.155
A_B_3M	-0.079
A_B_6M	-0.123
A_B_1A	-0.091
Formas_N_Pre	-0.045
Formas_N_3M	-0.186
Formas_N_6M	-0.068
Formas_N_1A	-0.286
Numero_alterações_Pre	0.007
Numero_alterações_3M	0.143
Numero_alterações_6M	0.089
Numero_alterações_1A	0.033

In terms of interesting correlations, we have seen that sperm parameters are mainly moderately dependent between themselves during the follow up times (all computed results with the Pearson correlation test are in Appendix A):

- Sperm concentration before the treatment "Conc_Pre" with the sperm concentration at 6 months after the treatment "Conc_6M" (*r*=0.633);
- Sperm concentration 3 months after the treatment "Conc_3M" with the sperm concentration at 6 months after the treatment "Conc_6M" (*r*=0.685);
- Sperm concentration 6 months after the treatment "Conc_6M" with the sperm concentration at 12 months after the treatment "Conc_1A" (*r*=0.738);
- Sperm progressive motility before the treatment "A_B_Pre" with the sperm progressive motility at 6 months "A_B_6M" (*r*=0.588);
- Sperm progressive motility 3 months after the treatment "A_B_3M" with the sperm progressive motility at 6 months "A_B_6M" (*r*=0.597);
- Sperm progressive motility 12 months after the treatment "A_B_1A" with the sperm progressive motility at 3 months "A_B_3M" (*r*=0.427);
- Sperm morphology at 3 months after the treatment "Formas_N_3M" with the sperm morphology at 12 months (*r*=0.544).

Furthermore, we also have seen that the age of the male patient is moderately correlated with the age of his female partner (*r*=0.544).

By analyzing the results computed with the *Chi-square* test, we see that the attributes that are more related with pregnancy are the following:

- Severity grade (*p*=0.049);
- Occupational hazard due to possible toxic exposure (*p*=0.023);
- Categorization of the sperm concentration at 3 months (*p*=0.017);
- Categorization of the sperm progressive motility before (*p*=0.027) and 3 months (*p*=0.022) after the treatment;
- Categorization of the semen analysis reports performed before (*p*=0.017), 3 (*p*=0.018) and 6 (*p*=0.036) months after the treatment.

In Table 5.21, we present the results of the computed Chi-square test for each qualitative attribute described in Table 4.3.

Table 5.21 Chi square results

Attribute Name	Chi square value	*p* value
Prim_Sec	0.080	0.961
Factor_Infertilidade_Feminino	17.551	0.093
Factor_Infertilidade_Masculino	17.828	0.058
HabitosTabagicos	25.692	0.691
HabitosAlcoolicos	12.124	0.277
Cirurgias	62.988	0.512
Doença	68.543	0.527
Profissao	119.292	0.445
Grau_Varicoc	7.869	0.049
Lateralidade	1.503	0.472
Volume_Testiculo_Médico	10.468	0.655
TratamentoFeito_lateralidade	2.653	0.265
TratamentoFeito_material	3.917	0.271
Complicações	7.045	0.532
Repetia_embolização	1.077	0.584
Razão_não_repetir	6.610	0.251
HabitosTabagicos_Processado_Simplificado	1.007	0.604
HabitosAlcoolicos_Processado_Simplificado	2.409	0.300
Cirurgias_Processado_Simplificado	1.045	0.593
DoençaSimplificada	45.400	0.413
ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	7.549	0.023
Qualificar_Espermograma_Pre	20.145	0.017
Qualificar_Espermograma_3M	20.007	0.018
Qualificar_Espermograma_6M	17.776	0.038
Qualificar_Espermograma_1A	8.588	0.284
Conc_Pre_Qualificado	3.760	0.153
Conc_3M_Qualificado	8.135	0.017

Attribute Name	Chi square value	*p* value
Conc_6M_Qualificado	3.186	0.203
Conc_1A_Qualificado	0.677	0.713
A_B_Pre_Qualificado	7.221	0.027
A_B_3M_Qualificado	7.625	0.022
A_B_6M_Qualificado	0.405	0.817
A_B_1A_Qualificado	1.175	0.556
Formas_N_Pre_Qualificado	5.359	0.069
Formas_N_3M_Qualificado	5.446	0.066
Formas_N_6M_Qualificado	4.123	0.127
Formas_N_1A_Qualificado	1.503	0.472

If we analyze the *Crosstab* in Table 5.22, we see that the most common severity grade for the patients that were and were not able to conceive is the severity grade II which tends to be 13.77% more prevalent on patients that were able to conceive. In contrast, the severity grade III, that is seen as the less prevalent severity grade among both types of patients, tends to be 13.06% more prevalent on patients that were not able to conceive. Most important values are in Table 5.22, and similarly in other below crosstabs, highlighted in orange.

Table 5.22 Crosstab of the severity grade by the pregnancy test result (*p*=0.049)

Severity Grade	Pregnancy test result				Total
	Yes		No		
	Count	% of Total	Count	% of Total	
I	29	34.12	31	34.83	60
II	48	56.47	38	42.70	86
III	8	9.41	20	22.47	28
Total	85	100	89	100	174

Statistically significant p<0.05

If we analyze the *Crosstab* depicted in Table 5.23, we see that the putative toxicity of the male patient´s occupation (a.k.a. toxic occupation) is related with achieving pregnancy (*p*=0.023). Even if most male patients have a non-toxic occupation, we see that the male patients that were able to achieve pregnancy tend to have 2.46% less chance of having a toxic occupation.

Table 5.23 Crosstab of the toxicity of the patient´s occupation by the pregnancy test result

Toxic Occupation	Pregnancy test result				Total
	Yes		No		
	Count	% of Total	Count	% of Total	
Yes	22	31.88	34	34.34	56
No	47	68.12	65	65.66	112
Total	69	100	99	100	168

Statistically significant p<0.05

Furthermore, we have seen that it is not by chance that the qualification of the sperm concentration values at 3 months are related with the fact of being able to achieve pregnancy (*p*=0.017). In fact, we see (Table 5.24) that the male patients that were able to achieve pregnancy tend to have at 3 months 20.07% more normal sperm concentration values than other patients since the male patients that were not able to achieve pregnancy tend to have at 3 months 20.07% more abnormal sperm concentration values.

Table 5.24 Crosstab of the qualification of the sperm concentration at 3 months by pregnancy test result

	Pregnancy test result				Total
Conc_3M_Qualificado	Yes		No		
	Count	% of Total	Count	% of Total	
Normal	54	57.45	40	37.38	94
Abnormal	40	42.55	67	62.62	107
Total	94	100	107	100	201

Statistically significant p<0.05

If we analyze the qualification of sperm progressive motility before the treatment, we see (Table 5.25) that the male patients that could achieve pregnancy after the embolization treatment tend to have already 10.38% more normal values on sperm progressive motility before the treatment. However, even if the relative frequency of the abnormality of the sperm parameter values overcomes its normality in both patient groups, the patients that were not able to achieve pregnancy tend to have 10.38% more abnormal values on its sperm progressive motility than other patients.

Table 5.25 Crosstab of the qualification of the sperm progressive motility before the treatment by the pregnancy test result

	Pregnancy test result				Total
A_B_Pre_Qualificado	Yes		No		
	Count	% of Total	Count	% of Total	
Normal	43	42.16	34	31.78	77
Abnormal	59	57.84	73	68.22	132
Total	102	100	107	100	209

Statistically significant p<0.05

Similarly with what happened with the sperm concentration values at 3 months, the values of sperm progressive motility at 3 months tends to have 13.05% more normal values on patients that were able to achieve pregnancy than on other patients. In contrast, the patients that were not able to achieve pregnancy tend to have 13.05% more abnormal sperm progressive motility values (see Table 5.26).

Table 5.26 Crosstab of the qualification of the sperm progressive motility at 3 months after the treatment by pregnancy test result

	Pregnancy test result				Total
A_B_3M_Qualificado	Yes		No		
	Count	% of Total	Count	% of Total	
Normal	50	54.35	38	41.30	88
Abnormal	42	45.65	54	58.70	96
Total	92	100	92	100	184

Statistically significant p<0.05

Moreover, if we analyze the qualification of the semen analysis report before and after the treatment (at 3 and 6 months), we see that they also have a statistically significant relationship with the pregnancy test result. If we analyze the Crosstab depicted in Table 5.27, we see that the male patients that were not able to conceive tend to be 6.92% more Azoospermic, 7.03% more Teratozoospermic and 8.94% more OligoAsthenoTeratozoospermic than the male patients that were able to conceive. Furthermore, this last semen qualification is more prevalent on patients that were not able to conceive, despite being the most frequent semen qualification in both patient groups before treatment. The semen classification that has the biggest relative frequency discrepancy between the two patient groups is the OligoTeratozoospermia which is 15.33% more prevalent in patients that were able to conceive.

Table 5.27 Crosstab of the qualification of the semen analysis report before the treatment by the pregnancy test result

	Pregnancy test result				Total
Qualificar_Espermograma_Pre	Yes		No		
	Count	% of Total	Count	% of Total	
Normozoospermia	6	6.45	5	5.10	11
Oligozoospermia	10	10.75	9	9.18	19
OligoAsthenozoospermia	12	12.90	10	10.20	22
OligoTeratozoospermia	19	20.43	5	5.10	24
Asthenozoospermia	9	9.68	8	8.16	17
AsthenoTeratozoospermia	8	8.60	8	8.16	16
Teratozoospermia	2	2.15	9	9.18	11
OligoAsthenoTeratozoospermia	23	24.73	33	33.67	56
Azoospermia	4	4.30	11	11.22	15
Total	93	100	98	100	191

Statistically significant p<0.05

In terms of the sample proportions of the semen classification, we see that 3 months after the embolization treatment, patients that were able to conceive tend to be 18.76% more normozoospermic and that normozoospermia is the most prevalent semen classification 3 months after the embolization treatment in contrast to before the treatment, where the main category was OligoAsthenoTeratozoospermia. If we analyze the male patients that were not able to conceive, we see that the most prevalent semen classification is Azoospermia that tends to be 14.35% greater, followed by the OligoAsthenoTeratozoospermia that tends to be 4.48% greater than for the patients that were able to conceive (see Table 5.28).

Table 5.28 Crosstab of the qualification of the semen analysis report at 3 months after the treatment by the pregnancy test result

	Pregnancy test result				Total
Qualificar_Espermograma_3M	Yes		No		
	Count	% of Total	Count	% of Total	
Normozoospermia	24	30.00	10	11.24	34
Oligozoospermia	4	5.00	9	10.11	13
OligoAstenozoospermia	5	6.25	4	4.49	9
OligoTeratozoospermia	8	10.00	7	7.87	15
Asthenozoospermia	9	11.25	10	11.24	19
AsthenoTeratozoospermia	8	10.00	12	13.48	20
Teratozoospermia	11	13.75	8	8.99	19
OligoAstenoTeratozoospermia	9	11.25	14	15.73	23
Azoospermia	2	2.50	15	16.85	17
Total	80	100	89	100	169

Statistically significant p<0.05

If we analyze the evolution of the sample proportions of the semen classifications at 6 months (Table 5.29), we see that Asthenozoospermia overtakes Normozoospermia in patients that have conceived by increasing by 10.49% (21.74% - 11.25%). However, the most prevalent semen classifications are still the same for the patients that were not able to conceive and a greater discrepancy of the relative frequencies between each patiens groups is seen: 25.28% (29.63%-4.35%) for Azoospermia and 16.59% (29.63%-13.04%) for OligoAsthenoTeratozoospermia. Furthermore, in the group of patients that were not able to conceive, we see that at 6 months after treatment there is no patient with Oligozoospermia, OligoAsthenozoospermia and OligoTeratozoospermia.

Table 5.29 Crosstab of the qualification of the semen analysis report at 6 months after the treatment by pregnancy test result

	Pregnancy test result				Total
Qualificar_Espermograma_6M	Yes		No		
	Count	% of Total	Count	% of Total	
Normozoospermia	2	8.70	4	14.81	6
Oligozoospermia	4	17.39	0	0.00	4
OligoAstenozoospermia	2	8.70	0	0.00	2
OligoTeratozoospermia	2	8.70	0	0.00	2
Asthenozoospermia	5	21.74	2	7.41	7
AsthenoTeratozoospermia	3	13.04	3	11.11	6
Teratozoospermia	1	4.35	2	7.41	3
OligoAstenoTeratozoospermia	3	13.04	8	29.63	11
Azoospermia	1	4.35	8	29.63	9
Total	23	100	27	100	50

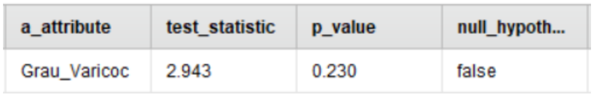
Statistically significant p<0.05

5.3.2.7 Varicocele laterality vs severity grade

When have analyzed the patients with the varicocele condition on the left or on both testes along with its severity (subsection 5.3.2.1.7), we have further on wondered if there was a relationship between these two attributes.

Table 5.30 presents the computed result. If we analyze the result shown, we see that the laterality of the varicocele condition is not related with its severity grade (p>0.05).

Table 5.30 Chi Square result for the varicocele´s severity grade vs its laterality

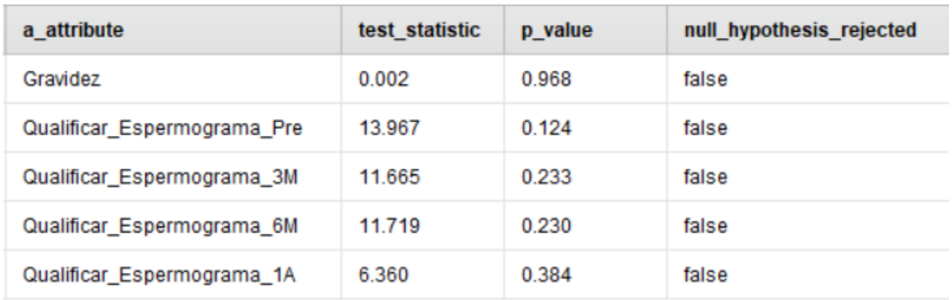


a_attribute	test_statistic	p_value	null_hypoth...
Grau_Varicoc	2.943	0.230	false

5.3.2.8 Semen classification vs Laterality

Of the 230 male patients that were able to conceive, we have assessed the statistical relationship between the semen classification and the laterality of the varicocele condition and have only considered the patients that had the condition on the left and both testicles since only 1 patient has it on its right testicle. We have seen that there is no statistical significant relationship between the patients that have the varicocele condition on the left side or on both sides in terms of semen classification or pregnancy (p>0.05; Table 5.31).

Table 5.31 Chi Square results for semen classification vs laterality

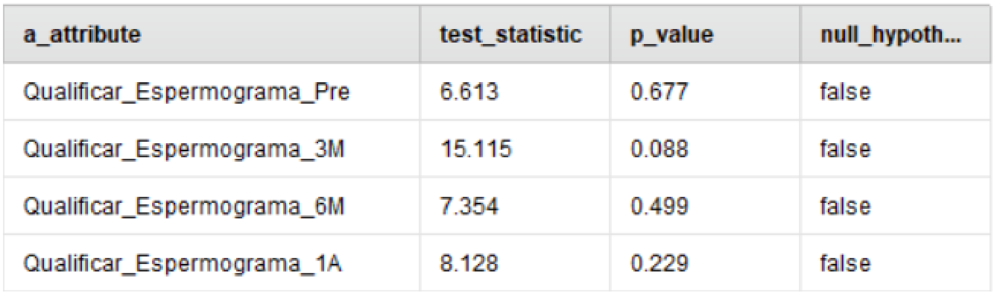


a_attribute	test_statistic	p_value	null_hypothesis_rejected
Gravidez	0.002	0.968	false
Qualificar_Espermograma_Pre	13.967	0.124	false
Qualificar_Espermograma_3M	11.665	0.233	false
Qualificar_Espermograma_6M	11.719	0.230	false
Qualificar_Espermograma_1A	6.360	0.384	false

5.3.2.9 Semen classification vs Drinking habit

We have assessed the statistical relationship between the semen classification and the simplified drinking habits of the patients and have seen that there is no statistical significant relationship between the semen classification and the fact that the patients drink or do not drink (Table 5.32)

Table 5.32 Chi Square results for semen classification vs drinking habit

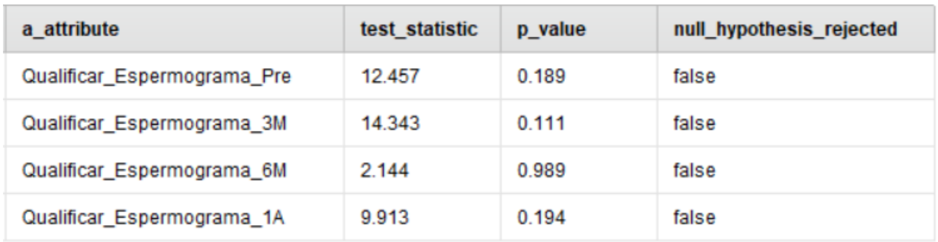


a_attribute	test_statistic	p_value	null_hypoth...
Qualificar_Espermograma_Pre	6.613	0.677	false
Qualificar_Espermograma_3M	15.115	0.088	false
Qualificar_Espermograma_6M	7.354	0.499	false
Qualificar_Espermograma_1A	8.128	0.229	false

5.3.2.10 Semen classification vs Smoking habit

We have also assessed the statistical relationship between the semen classification and the simplified smoking habits of the patients and have seen that there is no statistically significant relationship between the semen classifications and the smoking habits. (Table 5.33)

Table 5.33 Chi Square results for Semen classification vs Smoking habit



a_attribute	test_statistic	p_value	null_hypothesis_rejected
Qualificar_Espermograma_Pre	12.457	0.189	false
Qualificar_Espermograma_3M	14.343	0.111	false
Qualificar_Espermograma_6M	2.144	0.989	false
Qualificar_Espermograma_1A	9.913	0.194	false

5.3.2.11 Patient age vs Pregnancy outcome

By applying the ANOVA test, we have seen that the age of the male patient does not significantly differ between the ones that were able to conceive and the ones that were not (*p*=0.752). However, the age of the male patient´s partner (i.e. the woman patient) significantly differs (*p*=0.018). Figure 5.27, depicts this difference for the 230 patients that had no-missing values under the "Gravidez" attribute and shows that the patient´s partners that did not conceived were in average 1.318 (±1.228) year older than the ones that were able to conceive.

In contrast, we see that the mean of the male patient age is quite the same (± 34 years old). The values of the means and standard deviations depicted in Figure 5.27 can be seen in Table 5.34.

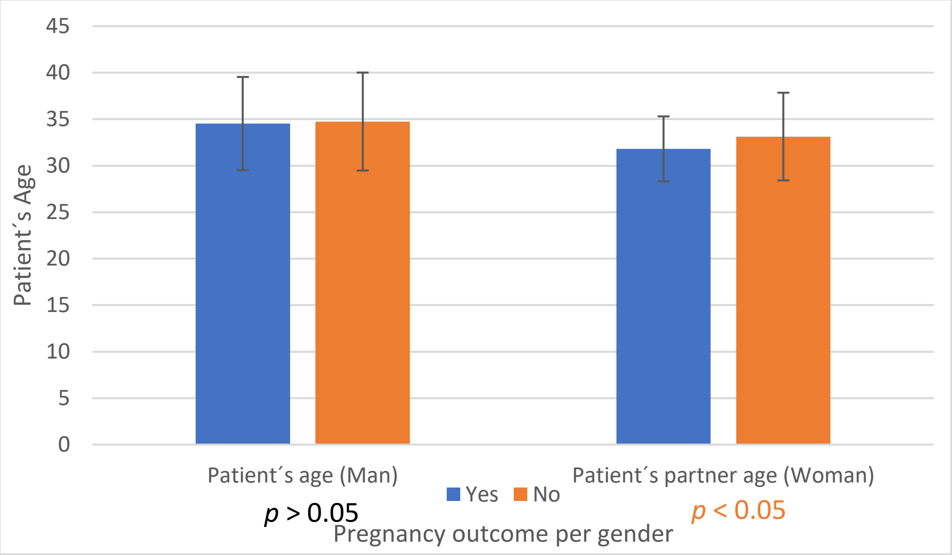


Figure 5.27 Bar Graph of the patient´s mean age with its standard deviations by the patient´s pregnancy outcome

Table 5.34 Means and Standard Deviations of the patient´s ages vs the pregnancy outcomes

Mean	Patient´s age (Man)	Patient´s partner age (Woman)
Pregnancy Outcome = Yes	34.533	31.813
Pregnancy Outcome = No	34.748	33.131
Standard Deviation	Patient´s age (Man)	Patient´s partner age (Woman)
Pregnancy Outcome = Yes	5.008	3.489
Pregnancy Outcome = No	5.268	4.717

5.4 Data Preparation

In this section, we disclose how the data preparation was carried out. Hence, in section 5.4.1, we present the results of data quality assessment; in section 5.4.2, we specify how new attributes were constructed and how the *initially provided and selected attributes* were reorganized, cleaned and formatted; in section 5.4.3, we specify how we have merged the imported data that we have used to fill some of the attributes that were seen missing in the *final preprocessed data set*, and finally, in section 5.4.4., we showcase the attributes that were selected to mine.

5.4.1 Data Quality Assessment

The quality of the data of the attributes disclosed in the Appendix A was carried out, as suggested by the CRISP-DM methodology, by addressing the following questions: if the data covered all the cases required to achieve the data mining goals set; if the data was correct; if it had errors, and if so, their frequency; if there are missing values (i.e. seen in the initially provided data set with blanks or with the *unknown* values) and if so, where do they occur and how common they were.

To check if the provided data covered all the cases required to tackle all data mining goals set, we have assessed if the initially provided and selected attributes encompassed the information needed to tackle them. For it, we have built the below Table 5.35 were we have identified the attributes of the Appendix A that could be analyzed to tackle each data mining goal based on the type of information they provide. Hence, this table has the following structure: the column named "*Original Attributes"* specifies all attributes in Appendix A, the column named "Type of information" specifies the type of information that the *Original Attribute* provides, and the remaining columns, specifies the data mining goals that were initially set out – these Goals are grouped by the *type of analysis* that one can perform in the data mining domain. This table was filled based on the information that was studied in the related works exposed in section 3.3 and 3.4 and acquired from the *BRSC* team. Further on, this table helped us to identify if all Goals were achievable, if more information (i.e. Attributes) was needed to fulfill the aims set or if the reorganization of the information gathered in some of the attributes was needed. The attributes that could be used to tackle each data mining goal is identified in the next Table 5.35 with a green check mark. When an attribute needed to be reorganized into new attributes a black check mark was written, otherwise nothing was specified.

If we analyze Table 5.35, we can see that the data mining goals 1 and 4 do not have any attributes that could be used to tackle them. In fact, these goals aim to predict how a male patient can develop a varicocele condition (Goal 1) and identify some patterns (Goal 4). Since the initially provided data set only has 2 patients (i.e. instances) that did not have the varicocele condition, these goals were not possible to achieve due to the low number of instances for these cases for the data mining algorithms to use. Hence, the incompleteness of the provided data set on this matter, incapacitated us to achieve the Goals 1 and 4 since it was not even possible to retrieve more data in the CHUC databases on that matter. For this reason, this study only focused on men with the varicocele condition and deleted the remaining and duplicated instances which made the data set end up with 293 instances.

For the other data mining goals, we see that the "*FR*" attribute, as well as the "*Factor_F*" and "*Notas*", needed to be reorganized. In fact, each of them gathers data from different entities/subjects. For example, the "*FR*" attribute gathers external factors such as occupation, diseases and smoking/drinking habits in the same attribute, which makes the assessment of each of these subjects harder. Hence, new attributes were created.

If we analyze the attributes that the initially provided data set to tackle Goal 2, we see that all attributes could be used, in terms of the type of information they provide. However, we could think that the attributes "*Idade_M*", "*Tempo_Infert*" and "*Prim_Sec*" would not provide at first sight useful information to predict the success of the male´s treatment due to their relation with the patient female partner, but since the success of the embolization is measured with pregnancy outcome, these attributes might also contribute for the aim set in Goal 2.

Concerning sperm parameter values, we see that to achieve the data mining Goal 3, we need to generate several attributes that categorizes the sperm parameter values to further on assess seminal quality (i.e. assess if they have *Azoospermia* etc.). At last, for the Goal 5, we see that any attribute can be interesting to assess since it aims to identify any kind of data patterns; and therefore, the creation of new attributes is welcomed. However, some patterns were initially suggested to assess which led us to focus on these specific subjects: the patient´s seminal quality vs smoking and drinking habits, as well as laterality. Furthermore, since we have seen that the materials used during the embolization procedure could influence its success (Bilreiro et al., 2017), and the testis size, indicate the advanced state of the varicocele condition (Aza Mohammed & Frank Chinegwundoh, 2009), we have also created new attributes to gather those data. Unfortunately, we were not able to retrieve those data for all patients that we had in our dataset, but it at least provided us with some clues.

Table 5.35 Data Set Completeness Assessment

		Predictive		Descriptive		
		1) Varicocele	2) Embolization Success	3) Semen Classification Vs Laterality	4) Varicocele vs Male´s Infertility	5) Find other male´s infertility patterns
Original Attribute	Type of information					
Idade_H	Patient´s info		✓			✓
Idade_M	Patient´s Partner info		✓			✓
Tempo_Infert	Patient´s Partner info		✓			✓
Prim_Sec	Patient´s Partner info		✓			✓
FR	Male Patient´s risk factors		✓			✓
Fator_F	Couple´s infertility factor		✓			✓
Grau_Varicoc	varicocele´s info		✓			✓
Lateralidade	varicocele´s info		✓	✓		✓
Data	Embolization Data		✓			✓
Notas	Embolization´s feedback		✓			✓
Complicações	Embolization´s feedback		✓			✓
Conc_Pre	Sperm parameter value		✓	✓		✓
Conc_3M	Sperm parameter value		✓	✓		✓
Conc_6M	Sperm parameter value		✓	✓		✓
Conc_1A	Sperm parameter value		✓	✓		✓
A_B_pré	Sperm parameter value		✓	✓		✓
A_B_3M	Sperm parameter value		✓	✓		✓
A_B_6M	Sperm parameter value		✓	✓		✓
A_B_1A	Sperm parameter value		✓	✓		✓
Formas_N_pré	Sperm parameter value		✓	✓		✓
Formas_N_3M	Sperm parameter value		✓	✓		✓
Formas_N_6M	Sperm parameter value		✓	✓		✓
Formas_N_1A	Sperm parameter value		✓	✓		✓
Gravidez	Possible Label attribute		✓			✓
Num_Gravidezes	Possible Label attribute		✓			✓
Nascimento	Possible Label attribute		✓			✓
Num_Bébés	Possible Label attribute		✓			✓
Gravidez_pós_emb	Possible Label attribute		✓			✓
PMA	Possible Label attribute		✓			✓
Gravidez_espontanea	Possible Label attribute		✓			✓
Repetia_embolização	Embolization´s feedback		✓			✓
Razão_não_repetir	Embolization´s feedback		✓			✓

In the following Table 5.36 we present the results of the overall data quality assessment performed upon the data attributes initially provided and selected. Hence, in this table we have specified for each key dimension the *type of information* that did not comply with the data characteristics disclosed in the *key data dimension´* specified in section 4.2.1. Note that the *type of information* in this table was categorized in the previous table and the *Requirements* in the key dimension definitions was disclosed in Table 4.4

Table 5.36 Data Quality Assessment of the initially provided data set

Key dimension	Requirements	Type of information that did not comply
Completeness	Directly usable (without data preparation needed) and Filled	All attributes
Consistency	Coherent and without duplicated instances.	Some *Possible Label attributes* showed in the same instance incoherence between them and the provided data set had duplicated instances.
Conformity	Rightly Formatted	The *Embolization data* had the format mm/dd/yyyy so it was then formatted to dd/mm/yyyy. *The Male Patient risk factors*, the *Couple´s Infertility factors* and the *Embolization´s feed backs* had different values for the same meaning; and therefore, the values were standardized on meaning and format before generating new attributes upon these attributes. The *Sperm parameter values* were provided rounded in a wrong way; and therefore, it was corrected: for the sperm concentrations, to a decimal value with only one decimal point, and the other sperm parameters, to an integer value.
Accuracy	Correct Data	All data needed to be validated with the medical dossier or the information technology systems in the CHUC
Integrity	Correct Data linkage	One instance had the wrong *Patient ´s Partner info,* so its related *Possible Label attributes* and *Couple´s infertility factors* information, was recollected and fixed.

If we analyze the type of information that did not comply with the requirements of each key dimension, we see that all attributes needed to be preprocessed since none of them fully complied with all *requirements* (see Table 5.36). The effort that was needed to preprocess that data can be seen in Table 5.37. This table presents, as specified in section 4.2.1, the data quality scoreboard for each initially provided and selected attribute for the final 293 filtered instances. The original attributes that were not validated with information technology systems or medical dossiers, or replaced by new attributes, have its related meta-data colored in gray and the "*Validation rate*" column set to 0% or NA respectively. In fact, if we analyze the *Validation rate* of the "*Temp_Infert*", "*Repetia_embolização*" and "*Razão_não_repetir*" attribute, we see that these attributes were not validated. The reason behind it is that they are attributes that gather approximated values told by the patient that cannot be validated with accuracy; and therefore, were not considered for further data analysis. However, these attributes were initially explored to better understand the data – as previously disclosed in section 5.3.2.1. Despite not having validated the "Repetia_embolização" and "Razão_não_repetir" attribute, we have verified them with the other initially provided and selected attributes. This data verification enabled us to reorganize and produce a coherent data set that through the process enabled the identification of erroneous data.

The other computed *Validation rate* that can be seen in this table, is the 27.3 validation rate. This rate represents the 80 male patient medical dossiers that it was possible to look at in the CHUC (i.e. 80 patients were validated / 293 filtered attributes = 27.3%). These s medical dossiers were used to validate the data values other than the *sperm parameter* and *possible label attribute* values.

By analyzing Table 5.37, we can see that the sperm parameter values required the highest validation effort due to its possibility to be accurately validated with clinical test reports and be considered important by related works (see related works studied in section 3.3 and 3.4 on this matter), as well as the *BRSC* team. Furthermore, since the prediction of the embolization success was since the beginning seen as the most important from the BRSC team, we have also dedicated our time to validate and fill the attributes that could be possible labels to further apply supervised predictive algorithms, such as decision trees, to the preprocessed data set. Therefore, the validation rate of these related attributes, as well as the patient age attributes and embolization date, were our focus.

If we analyze the numbers of the "Final Missing Values", we see that some them are even higher than before. This situation is seen in some of the sperm parameters values because some of the Azoospermic patients had the value 0 in motility and morphology when they should have had blank values instead, and some of the added sperm parameters values, only had sperm concentration and motility specified in the medical dossiers since they were from sperm analysis reports conducted for ART procedures. Moreover, the filling of the *possible label attributes* had a different mindset, and some of them were incoherent, which also led to a higher number of missing values in the final preprocessed data set.

Table 5.37 Data Quality Scoreboard for the original attributes

Original Attribute	Initial Missing Values	Final Missing Values	Validation Rate	Identified Erroneous Values	Initial Completeness Score	Final Completeness Score	Initial Accuracy Score
Idade_H	15	0	100%	44	94.88	100.00	84,98
Idade_M	27	9	100%	112	90.78	96.93	61.77
Tempo_Infert	46	39	0%		84.30	86.69	
Prim_Sec	35	23	100%	25	88.05	92.15	91.47
FR	176		NA		39.93		
Fator_F	266		NA		9.22		
Grau_Varicoc	155	82	27.30%		47.10	72.01	
Lateralidade	150	75	27.30%		48.81	74.40	
Data	2	0	100%	0	99.32	100.00	100.00
Notas	279		NA		4.78		
Complicações	282	0	27.30%		3.75	100.00	
Conc_Pre	23	12	100%	33	92.15	95.90	88.74
Conc_3M	49	48	100%	35	83.28	83.62	88.05
Conc_6M	218	162	100%	46	25.60	44.71	84.30
Conc_1A	241	156	100%	39	17.75	46.76	86.69
A_B_pré	53	42	100%	23	81.91	85.67	92.15
A_B_3M	73	76	100%	34	75.09	74.06	88.40
A_B_6M	230	177	100%	39	21.50	39.59	86.69
A_B_1A	247	172	100%	28	15.70	41.30	90.44
Formas_N_pré	73	83	100%	32	75.09	71.67	89.08
Formas_N_3M	97	113	100%	22	66.89	61.43	92.49
Formas_N_6M	268	246	100%	29	8.53	16.04	90.10

Original Attribute	Initial Missing Values	Final Missing Values	Validation Rate	Identified Erroneous Values	Initial Completeness Score	Final Completeness Score	Initial Accuracy Score
Formas_N_1A	287	270	100%	18	2.05	7.85	93.86
Gravidez	83	63	100%	27	71.67	78.50	90.78
Num_Gravidezes	78	0	100%	33	73.38	100.00	88.74
Nascimento	87	63	100%	35	70.31	78.50	88.05
Num_Bébés	87	209	100%	19	70.31	28.67	93.52
Gravidez_pós_emb	144	188	100%	43	50.85	35.84	85.32
PMA	80	0	100%	56	72.70	100.00	80.89
Gravidez_espontanea	142	63	100%	14	51.54	78.50	95.22
Repetia_embolização	145	144	0%		50.51	50.85	
Razão_não_repetir	1	0	0%		99.66	100.00	
Average					55.86	70.40	88.34

If we analyze the average of filled attributes in the initially provided data set indicated in the above table row named “Average”, we see that nearly half the attributes values were filled (55.86%). However, after the data preparation step, we could increase this data set on average by 14.54% (i.e. 70.40-55.86). In terms of the correctness of the provided data, we have seen that the values were on average 11.66% (100-88.34) incorrect. However, all these erroneous values were corrected throughout the data preparation step. Please note that the number of erroneous values was only counted when all the values of the corresponding attribute was validated since this value was only used to compute the “Initial Accuracy Score”.

As we can see in Table 5.37 above, some of the attributes were not validated since new attributes were created in replacement (attributes with the *Validation rate* set to NA). Hence, in Table 5.38 below we present the number of missing values that these newly created attributes had. These attributes were already described in Table 4.3. Their values came from the information gathered in the *original attributes*, specified under the column “Original Attribute Based On”, as well as the information technology systems and medical dossiers looked up in the CHUC. The gray cells under the column named “Original Attribute Based On” means that the attribute was created from scratch. Despite being newly created attributes, we also had missing values since patients do not have all the same data filled in their respective information technology systems or medical dossiers. Furthermore, some of these attributes, such as “Volume_Testicular_Médico” and “TratamentoFeito_material”, which has one of the highest number of missing values, were created during the analysis of the patient medical dossiers and identified as a possible interesting information; and therefore, only encompasses the patients that we could check.

Table 5.38 Data Quality Scoreboard for the newly created attributes

Original Attribute Based On	Attribute Name	Final Missing Values
Factor_F	Factor_Infertilidade_Feminina	208
Factor_F	Factor_Infertilidade_Masculina	130
FR	Hábitos Tabágicos	89
FR	Hábitos Alcoolicos	177
FR	Cirurgias	155
FR	Doença	201
FR	Profissão	91
	Volume_Testicular_Médico	249
Notas	TratamentoFeito_lateralidade	87

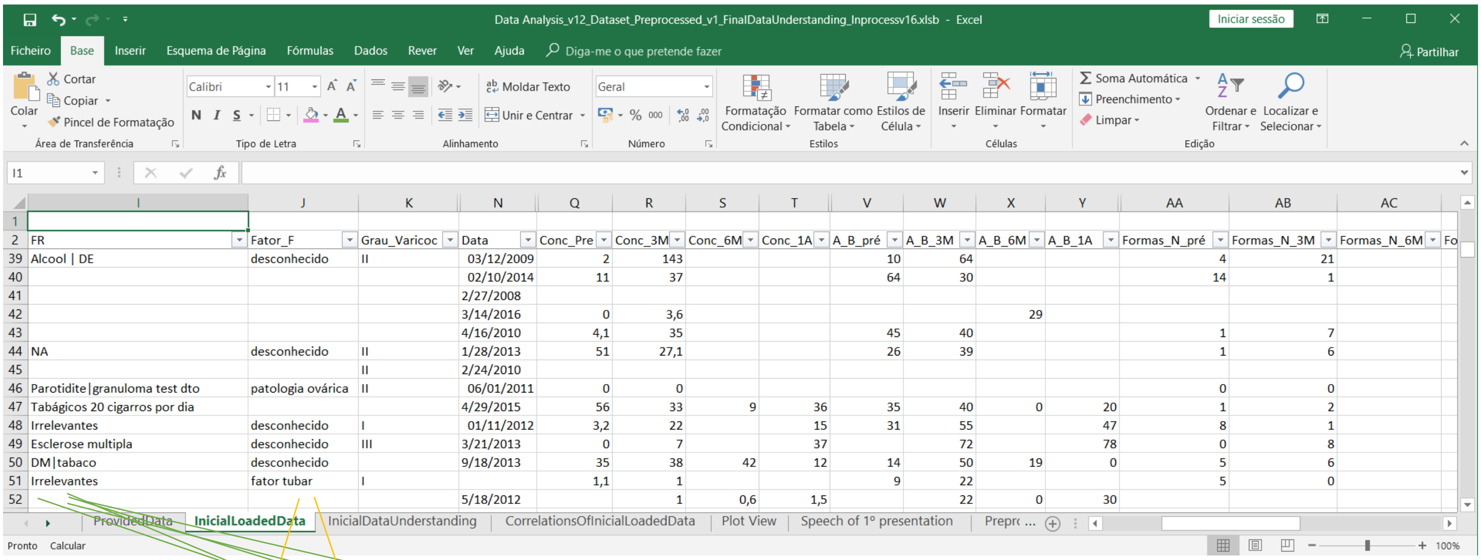
Original Attribute Based On	Attribute Name	Final Missing Values
	TratamentoFeito_material	270

5.4.2 Data Construction, Reorganization, Cleaning and Format

Before carrying out the data integration task, we constructed new attributes to reorganize the information recorded in the *original attributes Factor_F*, *FR* and *Notas*, as seen in Table 5.38, and then, cleaned the data set by retrieving the instances that were duplicated or out of the aim of this study. At last, we have formatted the embolization data. To better convey what was done during these data preparation tasks, we will partly present how the *initially provided data set* was, and how this *initially provided data set* was reorganized and preprocessed. Hence, in Figure 5.28 we see a print screen of part of the *initially provided data set* and in the Figure 5.29, the same *instances* reorganized with the where we can see (noted with green arrows) the newly created attributes for the *FR* attributes, and with the yellow arrows, the ones created for the *Factor_F* attributes. In this context, we can relate each instance of Figure 5.28 with the preprocessed instance shown in Figure 5.29 through the embolization date specified under the column name “Data” in Figure 5.28, and “Data Embolização” in the Figure 5.29.

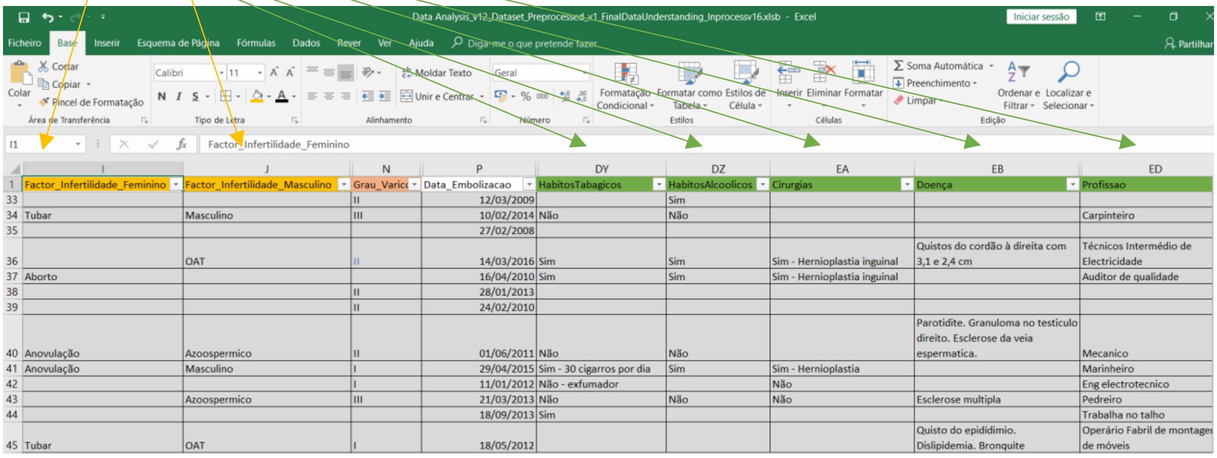
In Figure 5.28 the *initially provided data set* can be seen in the center-bottom. In the second row of this EXCEL table we see the names of some of the *original attributes* described in Appendix A, and below it, its values. If we analyze the values of the attribute *FR*, we see, as previously indicated, that this attribute gathers information from different subjects. In fact, the EXCEL row 39 indicates for this attribute that the patient drinks (specified with the word “Alcool”), in the row 46, that the patient had Parotitis disease (specified with the word “Parotidite”), in the row 47, that the patient smokes (specified with the word “Tabágicos …”) and in the row 50, that the patient also smokes but specified differently (with the word “Tabaco”). Hence, this small portion of data can already disclose why we have reorganized that information and uniformized it under the green attributes named seen in Figure 5.29. For the attributes names colored in yellow in Figure 5.29, we can see that we have already filled these attributes with more information than those provided in its original attribute “Factor_F”. Furthermore, we also see that the patient embolization date recorded in the *initially preprocessed dataset* under the attribute named “Data_Embolização” is formatted into dd/mm/yyyy. Moreover, the initially preprocessed dataset shown in Figure 5.29 has one less instance because the 2 last instances shown in Figure 5.28 were duplicates.

At the end of the data integration task, we had to uniformize the data and one of the most important uniformizations were the sperm parameter values. In fact, as we can see in the EXCEL row 49 of Figure 5.28 and under the column named “Formas_N_pré”, this patient has Azoospermia and has the value 0 for its morphology which is wrong. Hence, in these situations, motility and morphology were replaced with blank values.



	I	J	K	N	Q	R	S	T	V	W	X	Y	AA	AB	AC
1															
2	FR	Fator_F	Grau_Varicoc	Data	Conc_Pre	Conc_3M	Conc_6M	Conc_1A	A_B_pré	A_B_3M	A_B_6M	A_B_1A	Formas_N_pré	Formas_N_3M	Formas_N_6M
39	Alcool | DE	desconhecido	II	03/12/2009	2	143			10	64			4	21	
40				02/10/2014	11	37			64	30			14	1	
41				2/27/2008											
42				3/14/2016	0	3,6					29				
43				4/16/2010	4,1	35			45	40			1	7	
44	NA	desconhecido	II	1/28/2013	51	27,1			26	39			1	6	
45			II	2/24/2010											
46	Parotidite|granuloma test dto	patologia ovárica	II	06/01/2011	0	0							0	0	
47	Tabágicos 20 cigarros por dia			4/29/2015	56	33	9	36	35	40	0	20	1	2	
48	Irrelevantes	desconhecido	I	01/11/2012	3,2	22		15	31	55		47	8	1	
49	Esclerose multipla	desconhecido	III	3/21/2013	0	7		37		72		78	0	8	
50	DM|tabaco	desconhecido		9/18/2013	35	38	42	12	14	50	19	0	5	6	
51	Irrelevantes	fator tubar	I		1,1	1			9	22			5	0	
52				5/18/2012		1	0,6	1,5		22	0	30			

Figure 5.28 Part of the initially provided data set



	I	J	N	P	DY	DZ	EA	EB	ED
1	Factor_Infertilidade_Feminino	Factor_Infertilidade_Masculino	Grau_Varicc	Data_Embolizacao	HabitosTabagicos	HabitosAlcoolicos	Cirurgias	Doença	Profissao
33			II	12/03/2009		Sim			
34	Tubar	Masculino	III	10/02/2014	Não	Não			Carpinteiro
35				27/02/2008					
36		OAT	II	14/03/2016	Sim	Sim	Sim - Hernioplastia inguinal	Quistos do cordão à direita com 3,1 e 2,4 cm	Técnicos Intermédio de Electricidade
37	Aborto			16/04/2010	Sim	Sim	Sim - Hernioplastia inguinal		Auditor de qualidade
38			II	28/01/2013					
39			II	24/02/2010					
40	Anovulação	Azoospermico	II	01/06/2011	Não	Não		Parotidite. Granuloma no testiculo direito. Esclerose da veia espermatica.	Mecanico
41	Anovulação	Masculino	I	29/04/2015	Sim - 30 cigarros por dia	Sim	Sim - Hernioplastia		Marinheiro
42			I	11/01/2012	Não - exfumador		Não		Eng electrotecnico
43		Azoospermico	III	21/03/2013	Não	Não	Não	Esclerose multipla	Pedreiro
44				18/09/2013	Sim				Trabalha no talho
45	Tubar	OAT	I	18/05/2012				Quisto do epididimio. Dislipidemia. Bronquite	Operário Fabril de montagem de móveis

Figure 5.29 Part of the initially preprocessed data set

5.4.3 Data Integration

Data Integration contributed for the improvement of some of the *key dimensions qualities* exposed in Table 5.36. In fact, some of the provided data was validated, corrected and filled using a temporary Database created with the Microsoft SQL server and EXCEL scripting/clauses. How this process was exactly performed is specified here by disclosing the steps that were followed to preprocess the provided data set and produce the *initial preprocessed data set*. These steps were executed in the following order:

1) Retrieved from the "*Doentes*" information system the identification of the patient´s partner with their birth date, the patient´ birth date with occupation and the number of babies the patient´s partner had after her partner´s embolization. Through this process we have also validated the embolization dates recorded in the *initially provided data set* and checked if there was an indication of whether the patient´s partner got pregnant after the embolization and gave birth to a child. All this information was recorded in new temporary attributes created in the *reorganized initially provided data set* produced in the previous subsection 5.4.2. Note that the "SMR" information system only has the

information of patients that undergone procedures after 2012; and therefore, data collection entailed to looked up into the several resources specified in Table 5.1 that also indicates which attribute was checked from where;

2) Created a database in the Microsoft SQL server without any tables, called “varicoceleBD”;
3) Imported to the database previously created, the data set produced in the step 1) – during data import we have edited the mapping of the source data with the destination data and verified if the autogenerated type of data was correct;
4) Exported from the “*SMR*” information technology system the data to fill the following attributes: Prim_Sec, Factor_Infertilidade_Feminino, Factor_Infertilidade_Masculino, HabitosTabagicos, HabitosAlcoolicos, Cirurgias, Doença, Gravidez, Num_Gravidezes, PMA. Note that the “*SMR*” information technology system only exports the data by its type of information and when it was recorded in the system; and therefore, the data to fill the first 7 attributes, were exported into one Excel file, and the data to fill the other 3 attributes, needed to be exported into several Excel files. For instance, the first *IVF*, *ICSI* or *ISMI* procedure that the patient´s partner had undergone in CHUC was exported into one excel file, the second one, into another excel file etc. At the end, we had exported 30 Excel files to fill the aimed attributes: 7 from the undergone *IVF*, *ICSI* or *ISMI* procedures, 7 from the undergone *IVF*, *ICSI* or *ISMI* procedures carried out with cryopreservation, 8 from the undergone *IUI*, 7 from the fertilized embryos and 1 from the couple’s general information;
5) Imported to the database created in 2) the data exported in the previous step – as done in 3);
6) Built in the Microsoft SQL server a SQL query to join all the needed information from the imported Excel files;
7) Exported to EXCEL the result of the SQL query computed in the previous step;
8) Created temporary attributes, in the data set produced in 7), to gather the possible values of the *possible label attributes* – these values were achieved by building EXCEL scripts/clauses upon the dataset generated in 7);
9) Validated, Corrected and Filled the *Possible label attributes* with the created attributes in 8) if the values of the *Possible label attributes* were missing or if the patient was not previously questioned. Note that in the *initially provided data set*, the *Possible label attributes* were filled by questioning patients by phone. Unfortunately, not all patients answered the call;
10) Validated, Corrected and Filled the remaining attributes – the *Patient Partner info*, the *Male Patient risk factors* and the *Couple´s infertility factor* – with the data set produced in 9);
11) Validated, Corrected and Filled the remaining instances that were not preprocessed with the exported information from the “*SMR*” and “*Doentes*” information systems with the 80 medical dossiers that we had received. In parallel, sperm parameter values were also validated, corrected and filled with the original semen analysis reports;
12) Uniformized the format of the attribute values of the data set produced in 11);

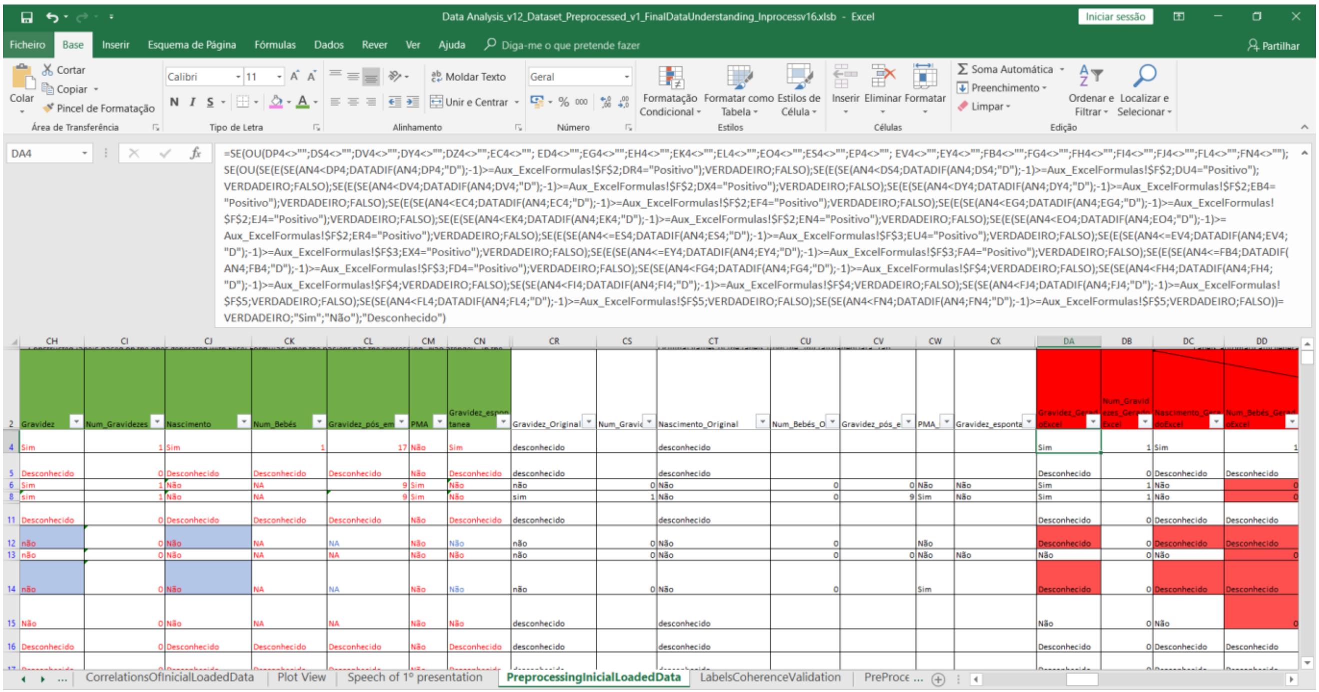
13) Loaded the data set produced in 12) into the RapidMiner platform. This loaded data set is the so called *initially preprocessed data set* that has the first 39 attributes described in Table 4.3 upon which other attributes were generated with the RapidMiner.

In Table 5.39, we present the rationale used in the creation of the EXCEL scripts/clauses indicated in step 8 for the validation, correction and filling of the *Possible label attributes*.

Table 5.39 Rationale of the Excel clauses

Attribute Name	Mind Set
Gravidez	if the woman of the male patient has a hospital event such has an ART procedure ("PMA" or "TEC") 3 days after the embolization with a positive pregnancy test, or an ultrasound, 10 days after the embolization, or a birth, 168 days after the embolization, we say that the partner had a pregnancy after the embolization. Otherwise, we write the value no "Não". If we do not have any information of the patient, we write nothing (i.e. *blank* value).
Num_Gravidezes	The number of pregnancies is equal to the number of ART procedures performed after the embolization date with a positive pregnancy test result "*Num_PMAs_Positivos_após_Embolização*" + (the total number of ultrasounds "*Num_Total_Ecos*" – the number of ART procedures with an ultrasound "*Num_PMAs_com_Ecos*") + (the total number of births "*Num_Total_Partos*" – the number of births with an ultrasound "*Num_Partos_com_Ecos*")
Nascimento	if the patient´s partner has at least one birth 168 days after her partner´s embolization treatment, we say that the partner had a birth.
Num_Bébés	The number of babies is equal to the sum of babies recorded in the "Doentes" system under the patient´s partner id with birth dates that have occurred at least 168 days after her partner´s embolization date.
Gravidez_pós_emb	We consider the date of the hospital event (ART procedure, ultrasound or birth) that has a shorter time span with the embolization date and records its difference in days
PMA	if the patient´s partner has a hospital event such has an ART procedure ("PMA" or "TEC") 3 days, or an insemination ("IIU") 0 days, after the embolization treatment with a positive pregnancy test, we put the value Yes "Sim". Otherwise, we put the value No "Não".
Gravidez_espontanea	if the patient´s partner has more natural conceptions than valid pregnancies from ART procedures, we put the value Yes "Sim". Otherwise, we put the value No "Não" if we have at least an information about the patient´s partner such as the ART dates and techniques performed or a ultrasound or birth date.

Below, in Figure 5.30, we can see a print screen of the data set during the preparation of the possible label attributes where in the bottom of the figure the preprocessed possible label attributes can be seen with names colored in green, the original ones in white, and the temporary ones created with the excel clauses, in red. Above it, the Excel Clause built for the "Gravidez" attribute can be seen. By analyzing the data preparation carried out on these attributes, we see that the values under the green colored attributes, have less missing/unknown values. Note that all unknown values that are in this figure seen, were at the end replaced with blank values to be easier/cleaner to manage in the RapidMiner platform.


Data Analysis_v12_Dataset_Preprocessed_v1_FinalDataUnderstanding_Inprocessv16.xlsx - Excel
Iniciar sessão
Ficheiro
Base
Inserir
Esquema de Página
Fórmulas
Dados
Rever
Ver
Ajuda
Diga-me o que pretende fazer
Partilhar
Área de Transferência
Tipo de Letra
Alinhamento
Número
Estilos
Células
Edição
DA4
CorrelationsOfInicialLoadedData
Plot View
Speech of 1º presentation
PreprocessingInicialLoadedData
LabelsCoherenceValidation
PreProce ...

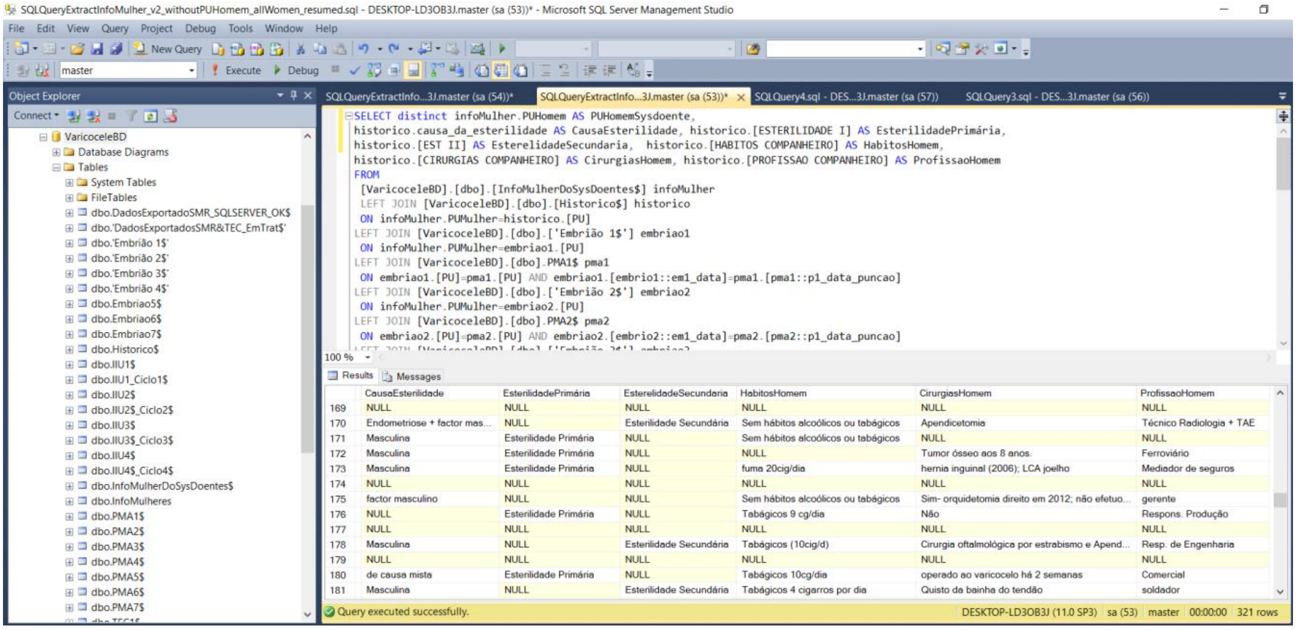
=SE(OU(DP4<>"";DS4<>"";DV4<>"";DY4<>"";DZ4<>"";EC4<>""; ED4<>"";EG4<>"";EH4<>"";EK4<>"";EL4<>"";EO4<>"";ES4<>"";EP4<>""; EV4<>"";EY4<>"";FB4<>"";FG4<>"";FH4<>"";FI4<>"";FJ4<>"";FL4<>"";FN4<>""); SE(OU(SE(E(SE(AN4<DP4;DATADIF(AN4;DP4;"D");-1)>=Aux_ExcelFormulas!$F$2;DR4="Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<DS4;DATADIF(AN4;DS4;"D");-1)>=Aux_ExcelFormulas!$F$2;DU4="Positivo"); VERDADEIRO;FALSO);SE(E(SE(AN4<DV4;DATADIF(AN4;DV4;"D");-1)>=Aux_ExcelFormulas!$F$2;DX4="Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<DY4;DATADIF(AN4;DY4;"D");-1)>=Aux_ExcelFormulas!$F$2;EB4= "Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<EC4;DATADIF(AN4;EC4;"D");-1)>=Aux_ExcelFormulas!$F$2;EF4="Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<EG4;DATADIF(AN4;EG4;"D");-1)>=Aux_ExcelFormulas! $F$2;EJ4="Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<EK4;DATADIF(AN4;EK4;"D");-1)>=Aux_ExcelFormulas!$F$2;EN4="Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<EO4;DATADIF(AN4;EO4;"D");-1)>= Aux_ExcelFormulas!$F$2;ER4="Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<=ES4;DATADIF(AN4;ES4;"D");-1)>=Aux_ExcelFormulas!$F$3;EU4="Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<=EV4;DATADIF(AN4;EV4; "D");-1)>=Aux_ExcelFormulas!$F$3;EX4="Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<=EY4;DATADIF(AN4;EY4;"D");-1)>=Aux_ExcelFormulas!$F$3;FA4="Positivo");VERDADEIRO;FALSO);SE(E(SE(AN4<=FB4;DATADIF(AN4;FB4;"D");-1)>=Aux_ExcelFormulas!$F$3;FD4="Positivo");VERDADEIRO;FALSO);SE(SE(AN4<FG4;DATADIF(AN4;FG4;"D");-1)>=Aux_ExcelFormulas!$F$4;VERDADEIRO;FALSO);SE(SE(AN4<FH4;DATADIF(AN4;FH4; "D");-1)>=Aux_ExcelFormulas!$F$4;VERDADEIRO;FALSO);SE(SE(AN4<FI4;DATADIF(AN4;FI4;"D");-1)>=Aux_ExcelFormulas!$F$4;VERDADEIRO;FALSO);SE(SE(AN4<FJ4;DATADIF(AN4;FJ4;"D");-1)>=Aux_ExcelFormulas! $F$5;VERDADEIRO;FALSO);SE(SE(AN4<FL4;DATADIF(AN4;FL4;"D");-1)>=Aux_ExcelFormulas!$F$5;VERDADEIRO;FALSO);SE(SE(AN4<FN4;DATADIF(AN4;FN4;"D");-1)>=Aux_ExcelFormulas!$F$5;VERDADEIRO;FALSO))= VERDADEIRO;"Sim";"Não");"Desconhecido")

	CH	CI	CJ	CK	CL	CM	CN	CR	CS	CT	CU	CV	CW	CX	DA	DB	DC	DD
2	Gravidez	Num_Gravidezes	Nascimento	Num_Bebés	Gravidez_pós_em	PMA	Gravidez_espontanea	Gravidez_Original	Num_Gravic	Nascimento_Original	Num_Bebés_O	Gravidez_pós_e	PMA_	Gravidez_esponta	Gravidez_GeradoExcel	Num_Gravidezes_GeradoExcel	Nascimento_GeradoExcel	Num_Bebés_GeradoExcel
4	Sim	1	Sim	1	17	Não	Sim	desconhecido		desconhecido					Sim	1	Sim	1
5	Desconhecido	0	Desconhecido	Desconhecido	Desconhecido	Não	Desconhecido	desconhecido		desconhecido					Desconhecido	0	Desconhecido	Desconhecido
6	Sim	1	Não	NA	9	Sim	Não	não	0	Não	0	0	Não	Não	Sim	1	Não	0
8	sim	1	Não	NA	9	Sim	Não	sim	1	Não	0	9	Sim	Não	Sim	1	Não	0
11	Desconhecido	0	Desconhecido	Desconhecido	Desconhecido	Não	Desconhecido	desconhecido		desconhecido					Desconhecido	0	Desconhecido	Desconhecido
12	não	0	Não	NA	NA	Não	Não	não	0	Não	0		Não		Desconhecido	0	Desconhecido	Desconhecido
13	não	0	Não	NA	NA	Não	Não	não	0	Não	0	0	Não	Não	Não	0	Não	0
14	não	0	Não	NA	NA	Não	Não	não	0	Não	0		Sim		Desconhecido	0	Desconhecido	Desconhecido
15	Não	0	Não	NA	NA	Não	Não	desconhecido		desconhecido					Não	0	Não	0
16	Desconhecido	0	Desconhecido	Desconhecido	Desconhecido	Não	Desconhecido	desconhecido		desconhecido					Desconhecido	0	Desconhecido	Desconhecido

Figure 5.30 Validation, Correction and Filling of the Possible label attributes

Below, we present some print screens of our working environment during the data integration task performed with the Microsoft SQL server 2012.

In Figure 5.31, on the left, the database created c "varicoceleBD" can be seen with part of its imported tables underneath; on the top right, we can see part of the SQL query that was built to join the information needed to complement the one initially provided in the attributes that gathered the *Couple´s infertility factors* (with the attribute seen in the first column), the *Patient´s Partner info* (with the attribute seen in the second and third column) and the *Male Patient´s risk factors* (with the attributes seen in the remaining columns). To the reorganized *initially provided data set* produced in the cleaning and formatted process, we have added this joined information to the newly constructed attributes exposed in Table 5.38 and afterwards, checked if in the patient medical dossier, we could find more information to fill the remaining missing values.


SQLQueryExtractInfoMulher_v2_withoutPUHomem_allWomen_resumed.sql - DESKTOP-LD3OB3J.master (sa (53))* - Microsoft SQL Server Management Studio
File Edit View Query Project Debug Tools Window Help
New Query
master
Execute
Debug
Object Explorer
Connect
VaricoceleBD
Database Diagrams
Tables
System Tables
FileTables
dbo.DadosExportadosSMR_SQLSERVER_OK$
dbo.'DadosExportadosSMR&TEC_EmTrat$'
dbo.'Embrião 1$'
dbo.'Embrião 2$'
dbo.'Embrião 3$'
dbo.'Embrião 4$'
dbo.Embriao5$
dbo.Embriao6$
dbo.Embriao7$
dbo.Historico$
dbo.IIU1$
dbo.IIU1_Ciclo1$
dbo.IIU2$
dbo.IIU2$_Ciclo2$
dbo.IIU3$
dbo.IIU3$_Ciclo3$
dbo.IIU4$
dbo.IIU4$_Ciclo4$
dbo.InfoMulherDoSysDoentes$
dbo.InfoMulheres
dbo.PMA1$
dbo.PMA2$
dbo.PMA3$
dbo.PMA4$
dbo.PMA5$
dbo.PMA6$
dbo.PMA7$
SQLQueryExtractInfo...3J.master (sa (54))*
SQLQueryExtractInfo...3J.master (sa (53))*
SQLQuery4.sql - DES...3J.master (sa (57))
SQLQuery3.sql - DES...3J.master (sa (56))
100 %
Results
Messages

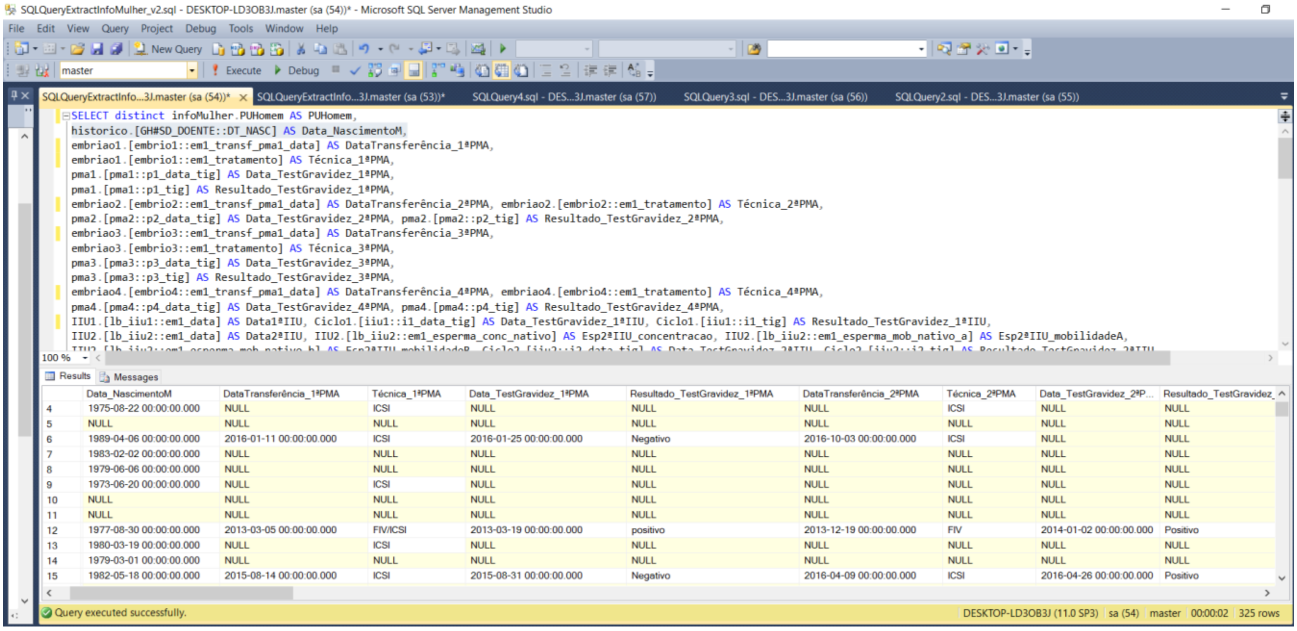
```sql
SELECT distinct infoMulher.PUHomem AS PUHomemSysdoente,
historico.causa_da_esterilidade AS CausaEsterilidade, historico.[ESTERILIDADE I] AS EsterilidadePrimária,
historico.[EST II] AS EsterelidadeSecundaria,  historico.[HABITOS COMPANHEIRO] AS HabitosHomem,
historico.[CIRURGIAS COMPANHEIRO] AS CirurgiasHomem, historico.[PROFISSAO COMPANHEIRO] AS ProfissaoHomem
FROM
 [VaricoceleBD].[dbo].[InfoMulherDoSysDoentes$] infoMulher
 LEFT JOIN [VaricoceleBD].[dbo].[Historico$] historico
 ON infoMulher.PUMulher=historico.[PU]
LEFT JOIN [VaricoceleBD].[dbo].['Embrião 1$'] embriao1
 ON infoMulher.PUMulher=embriao1.[PU]
LEFT JOIN [VaricoceleBD].[dbo].PMA1$ pma1
 ON embriao1.[PU]=pma1.[PU] AND embriao1.[embrio1::em1_data]=pma1.[pma1::p1_data_puncao]
LEFT JOIN [VaricoceleBD].[dbo].['Embrião 2$'] embriao2
 ON infoMulher.PUMulher=embriao2.[PU]
LEFT JOIN [VaricoceleBD].[dbo].PMA2$ pma2
 ON embriao2.[PU]=pma2.[PU] AND embriao2.[embrio2::em1_data]=pma2.[pma2::p1_data_puncao]
```

	CausaEsterilidade	EsterilidadePrimária	EsterelidadeSecundaria	HabitosHomem	CirurgiasHomem	ProfissaoHomem
169	NULL	NULL	NULL	NULL	NULL	NULL
170	Endometriose + factor mas...	NULL	Esterilidade Secundária	Sem hábitos alcoólicos ou tabágicos	Apendicetomia	Técnico Radiologia + TAE
171	Masculina	Esterilidade Primária	NULL	Sem hábitos alcoólicos ou tabágicos	NULL	NULL
172	Masculina	Esterilidade Primária	NULL	NULL	Tumor ósseo aos 8 anos.	Ferroviário
173	Masculina	Esterilidade Primária	NULL	fuma 20cig/dia	hernia inguinal (2006); LCA joelho	Mediador de seguros
174	NULL	NULL	NULL	NULL	NULL	NULL
175	factor masculino	NULL	NULL	Sem hábitos alcoólicos ou tabágicos	Sim- orquidetomia direito em 2012; não efetuo...	gerente
176	NULL	Esterilidade Primária	NULL	Tabágicos 9 cg/dia	Não	Respons. Produção
177	NULL	NULL	NULL	NULL	NULL	NULL
178	Masculina	NULL	Esterilidade Secundária	Tabágicos (10cig/d)	Cirurgia oftalmológica por estrabismo e Apend...	Resp. de Engenharia
179	NULL	NULL	NULL	NULL	NULL	NULL
180	de causa mista	Esterilidade Primária	NULL	Tabágicos 10cg/dia	operado ao varicocelo há 2 semanas	Comercial
181	Masculina	NULL	Esterilidade Secundária	Tabágicos 4 cigarros por dia	Quisto da bainha do tendão	soldador

```
Query executed successfully.    DESKTOP-LD3OB3J (11.0 SP3) | sa (53) | master | 00:00:00 | 321 rows
```

Figure 5.31 Query result of the *Couple infertility factor*, the *Patient Partner info* and the *Male Patient risk factors*

At the top of Figure 5.32, we can see part of the SQL query that was built to join the information needed to validate, correct or fill the possible label attributes and at its bottom, we can see part of the result generated. This result was by the built EXCEL clauses, exposed in Table 5.39, transformed to further on validate, correct and fill the values of the *possible label attributes*.


SQLQueryExtractInfoMulher_v2.sql - DESKTOP-LD3OB3J.master (sa (54))* - Microsoft SQL Server Management Studio
File Edit View Query Project Debug Tools Window Help
New Query
master
Execute
Debug
SQLQueryExtractInfo...3J.master (sa (54))*
SQLQueryExtractInfo...3J.master (sa (53))*
SQLQuery4.sql - DES...3J.master (sa (57))
SQLQuery3.sql - DES...3J.master (sa (56))
SQLQuery2.sql - DES...3J.master (sa (55))
100 %
Results
Messages

```sql
SELECT distinct infoMulher.PUHomem AS PUHomem,
historico.[GH#SD_DOENTE::DT_NASC] AS Data_NascimentoM,
embriao1.[embrio1::em1_transf_pma1_data] AS DataTransferência_1ªPMA,
embriao1.[embrio1::em1_tratamento] AS Técnica_1ªPMA,
pma1.[pma1::p1_data_tig] AS Data_TestGravidez_1ªPMA,
pma1.[pma1::p1_tig] AS Resultado_TestGravidez_1ªPMA,
embriao2.[embrio2::em1_transf_pma1_data] AS DataTransferência_2ªPMA, embriao2.[embrio2::em1_tratamento] AS Técnica_2ªPMA,
pma2.[pma2::p2_data_tig] AS Data_TestGravidez_2ªPMA, pma2.[pma2::p2_tig] AS Resultado_TestGravidez_2ªPMA,
embriao3.[embrio3::em1_transf_pma1_data] AS DataTransferência_3ªPMA,
embriao3.[embrio3::em1_tratamento] AS Técnica_3ªPMA,
pma3.[pma3::p3_data_tig] AS Data_TestGravidez_3ªPMA,
pma3.[pma3::p3_tig] AS Resultado_TestGravidez_3ªPMA,
embriao4.[embrio4::em1_transf_pma1_data] AS DataTransferência_4ªPMA, embriao4.[embrio4::em1_tratamento] AS Técnica_4ªPMA,
pma4.[pma4::p4_data_tig] AS Data_TestGravidez_4ªPMA, pma4.[pma4::p4_tig] AS Resultado_TestGravidez_4ªPMA,
IIU1.[lb_iiu1::em1_data] AS Data1ªIIU, Ciclo1.[iiu1::i1_data_tig] AS Data_TestGravidez_1ªIIU, Ciclo1.[iiu1::i1_tig] AS Resultado_TestGravidez_1ªIIU,
IIU2.[lb_iiu2::em1_data] AS Data2ªIIU, IIU2.[lb_iiu2::em1_esperma_conc_nativo] AS Esp2ªIIU_concentracao, IIU2.[lb_iiu2::em1_esperma_mob_nativo_a] AS Esp2ªIIU_mobilidadeA,
```

	Data_NascimentoM	DataTransferência_1ªPMA	Técnica_1ªPMA	Data_TestGravidez_1ªPMA	Resultado_TestGravidez_1ªPMA	DataTransferência_2ªPMA	Técnica_2ªPMA	Data_TestGravidez_2ªP...	Resultado_TestGravidez_
4	1975-08-22 00:00:00.000	NULL	ICSI	NULL	NULL	NULL	ICSI	NULL	NULL
5	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
6	1989-04-06 00:00:00.000	2016-01-11 00:00:00.000	ICSI	2016-01-25 00:00:00.000	Negativo	2016-10-03 00:00:00.000	ICSI	NULL	NULL
7	1983-02-02 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
8	1979-06-06 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
9	1973-06-20 00:00:00.000	NULL	ICSI	NULL	NULL	NULL	NULL	NULL	NULL
10	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
11	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
12	1977-08-30 00:00:00.000	2013-03-05 00:00:00.000	FIV/ICSI	2013-03-19 00:00:00.000	positivo	2013-12-19 00:00:00.000	FIV	2014-01-02 00:00:00.000	Positivo
13	1980-03-19 00:00:00.000	NULL	ICSI	NULL	NULL	NULL	NULL	NULL	NULL
14	1979-03-01 00:00:00.000	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
15	1982-05-18 00:00:00.000	2015-08-14 00:00:00.000	ICSI	2015-08-31 00:00:00.000	Negativo	2016-04-09 00:00:00.000	ICSI	2016-04-26 00:00:00.000	Positivo

```
Query executed successfully.    DESKTOP-LD3OB3J (11.0 SP3) | sa (54) | master | 00:00:02 | 325 rows
```

Figure 5.32 Query result of the undergone ART procedures

5.4.4 Attribute Selection

The selection of the attributes to mine was the culmination of all the work performed until the modelling step. Hence, based on all the acquired data understanding and data quality assessment, we have built a summary table to better identify the attributes that were valuable to achieve the data mining goals set in section 5.2.3 and finally planned how the selected attributes could be mined. Table 5.40 is the summary table that can be interpreted as follows:

The column named "*ID*" and "*Attribute Name*", specifies the name of the *final preprocessed attributes*; the column named "*Significance*", indicates if the corresponding attribute values present a statistical significant difference between the patients that got and did not conceive – probabilities computed with the *Chi-square* test came from section 5.3.2.6 and the ones computed with the *ANOVA* test, came from section 5.3.2.4; the column named "*Correlation*", indicates if the corresponding attribute is correlated with the "*Gravidez*" *label attribute* – the presented coefficients were computed with the *Pearson Correlation* measure and came from section 5.3.2.6; the column named "*Stability*", indicates how stable/constant the corresponding attribute values are – s*tabilities* were retrieved from the RapidMiner platform and are equal to the number of rows with the most frequent non-missing value divided by the total number of data rows with non-missing values; the column named "*Missing*", indicates the percentage of missing values of the corresponding attribute – these values were also retrieved from the RapidMiner platform and are equal to the number of missing values divided by the total number of instances; the column named "*Selected*", presents a check mark at the attributes that were considered for data mining and discloses if the RapidMiner suggested it or not for prediction tasks through its cell color (i.e. green cells mean that RapidMiner suggests the corresponding attribute for prediction, yellow cells mean that RapidMiner does not suggest it because of its low or high correlation and red cells mean that RapidMiner does not suggest it because of poor data quality); Furthermore, the column named "*Goal*", specifies if the corresponding attribute will be used to predict the success of the embolization (Goal 2) – by indicating the number "2" – or describe the relation between the semen classification and the varicocele laterality (Goal 3) – by indicating the number "3" – or the relation between the semen classification and the patient external factors (Goal 5) - by indicating the number "5" – or several – by indicating the id of the goals that the corresponding attribute can be mined for, for example, the combination "2 & 5" means that the corresponding attribute can be used to achieve Goals 2 and 5.

Rows colored in gray indicate that the corresponding attribute has a direct cause-and-effect relationship with the "*Gravidez*" *label attribute* (e.g. the attributes that were initially set as *possible label attributes*), or has no relation at all with the "*Gravidez*" *label attribute* (e.g. the date of the varicocele treatment called "Data_Embolização") or has a validation rate equal to 0% (computed and showcased in the data quality scoreboard presented in Table 5.37); and therefore, they will not be considered for analysis.

Attribute names colored in blue indicate that they are the *initially preprocessed attributes* described in Table 4.3. The other attributes, are the ones that were later on generated.

Good statistical "*Significance*" or "*Correlation*" values are highlighted with orange lettering.

The attributes were selected for mining purposes based on the following criteria:

- Low values for the "*Significance*" probability (i.e. <0.05);
- High values for the "*Correlation*" criteria (or at least higher than 0.01 which is the minimum considered by the RapidMiner platform);

- Low rates for the "*Stability*" and "*Missing*" criteria (i.e. lower than 90% for "*Stability*" and lower than 70% for "*Missing*" values – these thresholds were also suggested by the RapidMiner platform).

The reason behind the criteria noted above is the way data mining algorithms work: Algorithms used for prediction tasks tend to choose the attributes that are more correlated/significant with the label attribute since it indicates that they are the dependent factors (i.e. attributes) of the label attribute. Decision trees are often used for prediction tasks and they compute for their root, the attribute that promotes the highest gain of information. The attribute that promotes the highest gain of information is the one that can better separate its instances; and therefore, has a greater capability to reduce *Entropy*. In fact, classifiers such as decision trees, began with the attribute that promotes the highest gain of information and ramifies with the *Entropy*. The *Entropy* measures the state of confusion of a set of *instances* based on its *label attribute*; and therefore, if in a set of instances the label attribute has the same value for all instances, for example, all "Gravidez"="Sim", we then can say that this set of instances is pure (i.e. not confused) since they all have the same label value on that set of instances; and therefore, its *Entropy* is equal to 0. In contrast, if we have as many instances with the "Gravidez" attribute set to "Sim" as set to "Não", we will then have a set of instances that is totally mixed/confused; and therefore, the *Entropy* is equal to 1. If we consider the first set of instances that gave an *Entropy* equal to 0, we can say that all these patients had a pregnancy, and this conclusion is the biggest gain of information that the algorithm has at that moment. On the other hand, if we consider the case that ended up with an *Entropy* equal to 1, we see that the algorithm cannot conclude much on its instances so we say that the algorithm did not gain any information. Furthermore, when we have an attribute with a high "*Stability*", we often end up with an attribute that also generates a high *Entropy* (i.e. that provides a low gain of information). For instance, our initially preprocessed data set has half "Gravidez"="Sim" and "Gravidez"="Não" which is a good thing but if we have, lets say, an infertility time attribute equal to 6 months for all instances of the data set, the algorithm will not be capable to identify an information that differentiates the patients that conceived from the ones that did not, in terms of the infertility time attribute; and therefore, the algorithm will not select the infertility time attribute for its decision tree since it could not gain any information from it. This is why we look for attributes with low stability (i.e. with variability on values). Moreover, since data mining algorithms have to first train their models, having a high number of missing values will decrease its capacity to learn; and therefore, the probability of generating a model with good quality. Please note that for fitting purposes of the Table 5.40 longer attribute names where truncated.

Table 5.40 Summary table for attributes selection upon 293 instances

		Selection criteria					
ID	Attribute Name	Significance (*p*)	Correlation (*r*)	Stability	Missing	Select	Goal
1	Man age	ANOVA 0.752	0.021	9.22%	0.00%	🟩	
2	Woman age	ANOVA 0.018	0.156	11.27%	3.07%	✓	2 & 5
3	Infertility time						
4	Type of infertility	Chi-square 0.961		80.00%	7.85%	🟨	
5	Woman infertility factor	Chi-square 0.093		29.41%	70.99%	🟥	
6	Man infertility factor	Chi-square 0.058		50.31%	44.37%	🟩	
7	Smoking habit	Chi-square 0.691		43.14%	30.38%	🟩	

		Selection criteria					
ID	Attribute Name	Significance (*p*)	Correlation (*r*)	Stability	Missing	Select	Goal
8	Drinking habit	Chi-square 0.277		68.97%	60.41%	🟩	
9	Surgeries	Chi-square 0.512		38.41%	52.90%	🟩	
10	Diseases	Chi-square 0.527		13.04%	68.60%	🟩	
11	Occupation	Chi-square 0.445		3.96%	31.06%	🟩	
12	Severity grade	Chi-square 0.049		52.61%	27.99%	🟩 ✓	2 & 5
13	Laterality	Chi-square 0.472		81.65%	25.60%	🟩	
14	Testis volume	Chi-square 0.655		31.82%	84.98%	🟥	
15	Embolization date						
16	Embolized laterality	Chi-square 0.265		97.09%	29.69%	🟥	
17	Material of Embolization	Chi-square 0.271		65.22%	92.15%	🟥	
18	Complications	Chi-square 0.532		92.47%	0.00%	🟥	
19	Repeat embolization						
20	Reason to not repeat						
21	Concentration before treatment	ANOVA 0.903	0.008	9.96%	4.10%	🟨	
22	Concentration at 3 months	ANOVA 0.165	-0.092	11.02%	16.38%	🟩	
23	Concentration at 6 months	ANOVA 0.015	-0.161	11.45%	55.29%	🟩 ✓	2 & 5
24	Concentration at 12 months	ANOVA 0.081	-0.115	10.95%	53.24%	🟩	
25	Progressive motility before treatment	ANOVA 0.018	-0.155	12.35%	14.33%	🟩 ✓	2 & 5
26	Progressive motility at 3 months	ANOVA 0.236	-0.079	11.06%	25.94%	🟩	
27	Progressive motility at 6 months	ANOVA 0.064	-0.123	17.24%	60.41%	🟩	
28	Progressive motility at 12 months	ANOVA 0.171	-0.091	14.05%	58.70%	🟩	
29	Morphology before treatment	ANOVA 0.488	-0.045	19.05%	28.33%	🟩	
30	Morphology at 3 months	ANOVA 0.004	-0.186	16.11%	38.57%	🟩 ✓	2 & 5
31	Morphology at 6 months	ANOVA 0.327	-0.068	19.15%	83.96%	🟥	
32	Morphology at 12 months	ANOVA 0.000	-0.286	26.09%	92.15%	🟥	
33	Pregnancy outcome						
34	Number of pregnancies						
35	Birth						
36	Number of alive babies						
37	Time took to conceive						
38	ART						
39	Spontaneous pregnancy						
40	Preprocessed smoking habit	Chi-square 0.604		51.96%	30.38%	🟩	
41	Preprocessed drinking habit	Chi-square 0.300		69.83%	60.41%	🟩	
42	Preprocessed surgeries	Chi-square 0.593		61.59%	52.90%	🟩	
43	Preprocessed diseases	Chi-square 0.413		21.98%	68.94%	🟩	
44	Hazardous Occupation	Chi-square 0.023		63.86%	31.06%	🟩 ✓	2 & 5
45	Number of altered sperm parameters before		0.007	36.65%	4.10%	🟨	
46	Number of altered sperm parameters at 3		0.143	37.96%	16.38%	🟩	
47	Number of altered sperm parameters at 6		0.089	41.98%	55.29%	🟩	
48	Number of altered sperm parameters at 12		0.033	40.88%	53.24%	🟩	
49	Semen classification before treatment	Chi-square 0.017		26.89%	18.77%	🟩 ✓	2 & 5
50	Semen classification at 3 months	Chi-square 0.018		19.90%	29.69%	🟩 ✓	2 & 5
51	Semen classification at 6 months	Chi-square 0.038		24.19%	78.84%	🟥	
52	Semen classification at 12 months	Chi-square 0.284		39.47%	87.03%	🟥	
53	Concentration category before treatment	Chi-square 0.153		75.80%	4.10%	🟩	
54	Concentration category at 3 months	Chi-square 0.017		54.69%	16.38%	🟩 ✓	2 & 5
55	Concentration category at 6 months	Chi-square 0.203		61.07%	55.29%	🟩	
56	Concentration category at 12 months	Chi-square 0.713		62.04%	53.24%	🟩	
57	Progressive Motility category before	Chi-square 0.027		62.55%	14.33%	🟩 ✓	2 & 5
58	Progressive Motility category at 3 months	Chi-square 0.022		53.46%	25.94%	🟩 ✓	2 & 5
59	Progressive Motility category at 6 months	Chi-square 0.817		62.93%	60.41%	🟩	
60	Progressive Motility category at 12 months	Chi-square 0.556		61.98%	58.70%	🟩	
61	Morphology category before treatment	Chi-square 0.069		60.48%	28.33%	🟩	
62	Morphology category at 3 months	Chi-square 0.066		50.56%	38.57%	🟩	
63	Morphology category at 6 months	Chi-square 0.127		55.32%	83.96%	🟥	
64	Morphology category at 12 months	Chi-square 0.472		73.91%	92.15%	🟥	

By analyzing Table 5.40, we see that the Goal 3 could not be carried out since, as we have seen in section 5.3.2.8, patient semen classification is not related with varicocele laterality. However, the main data mining goals set were possible to carry out (i.e. the prediction of the embolization success (Goal 2) and the identification of patterns between semen classification and external factors (Goal 5)).

If we analyze the attributes that were selected in Table 5.40, we see that some of them will not be possible to mine together since they are related to each other (e.g. sperm parameter values are related with their corresponding sperm and semen categorization attributes). Hence, we have applied a feature selection technique to filter the *final preprocessed attributes* by having this aspect in consideration. We specify below in which order and with which attribute combination we have mined the *final preprocessed data set*:

1. Grau_Varicoc, Conc_6M, A_B_pré, Formas_N_3M, ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos, Gravidez.
2. Grau_Varicoc, Conc_3M_Qualificado, A_B_Pre_Qualificado, A_B_3M_Qualificado, ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos, Gravidez.
3. Grau_Varicoc, Qualificar_Espermograma_Pre, Qualificar_Espermograma_3M, ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos, Gravidez.
4. Idade_M, Grau_Varicoc, Conc_6M, A_B_pré, Formas_N_3M, ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos, Gravidez.
5. Idade_M, Grau_Varicoc, Conc_3M_Qualificado, A_B_Pre_Qualificado, A_B_3M_Qualificado, ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos, Gravidez.
6. Idade_M, Grau_Varicoc, Qualificar_Espermograma_Pre, Qualificar_Espermograma_3M, ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos, Gravidez.

The attribute combinations presented in the above step 1, 2 and 3 only focus on male patient information. In contrast, the other attributes combinations (i.e. step 4, 5 and 6) adds the woman patient partner info to see if a better and more interesting model can be generated. These attribute combinations were mined to predict the success of the embolization treatment (Goal 2) and identify data patterns (Goal 5).

To these groups of attributes, we have at last added the label attribute “Gravidez” since it is the main attribute that through this study was used to predict the success of the embolization treatment.

5.5 Modeling

As previously seen in section 2.1.1, the most commonly applied data mining techniques in the health care domain are the *Classification, Clustering* and *Association* techniques; and therefore, this study has also applied these techniques upon the *final preprocessed data set* described in section 4.1.2.1 to tackle the goals set and disclosed in section 5.2.3. To achieve the best possible results, we have applied these data mining techniques with well-known data mining algorithms upon the groups of attributes disclosed in the previous section 5.4.4 since they are the ones that are more related with the pregnancy outcome.

This study has applied the following algorithms: the Decision Tree algorithm, for the *Classification* technique; the K-MEANS algorithm, for the *Clustering* technique and the FP-Growth algorithm, for the *Association* technique. These algorithms were applied with the RapidMiner platform and most of the generated results, are in the Appendix C documented. The most interesting ones are in this section disclosed grouped by each applied data mining technique initially described in section 2.1.1. In order to better identify the most interesting findings, we have highlighted them with orange lettering.

5.5.1 Classification

All results generated during the first 8 decision tree modeling steps described in the Appendix B.1, can be seen in the Appendix C.1. From all these generated and disclosed results, we have identified that the model that outperformed during its training/testing was the model implemented with the RapidMiner Decision tree algorithm ran within a simple validation operator that achieved an F-measure of 75%. This model was executed upon a set of 85 instances with non-missing and parsed to numerical values that were at first preprocessed at the time of the application of the K-means algorithm. Hence, this model was applied upon the following attributes - that corresponds to the 5th group of assessed attributes - with the decision tree Model 2 that was built during the decision tree modeling step 6 depicted in Figure 6.5:

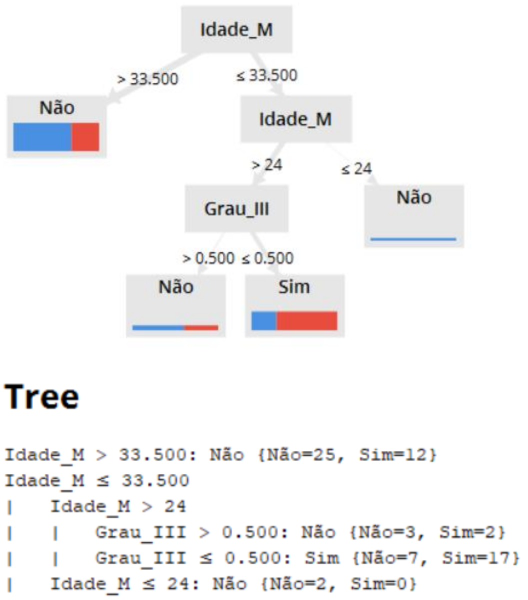
- Idade_M – not transformed;
- Grau_Varicoc - manually dichotomized (i.e. mapped and generated new attributes);
- ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos – mapped “Não” to 0 and “Sim” to 1, as well as parse to number;
- Conc_3M_Qualificado - mapped “Anomal” to 0 and “Normal” to 1, as well as parse to number;
- A_B_Pre_Qualificado - mapped “Anomal” to 0 and “Normal” to 1, as well as parse to number;
- A_B_3M_Qualificado - mapped “Anomal” to 0 and “Normal” to 1, as well as parse to number;
- Gravidez – not transformed (i.e. remained the nominal values “Sim” and “Não”) and informed the algorithm, with the “Remap binomials” operator, which was the positive and negative value.

The performance of the best model upon the above attributes is shown in the below Table 5.41.

Table 5.41 Best decision tree model of the first 8 modeling steps – Model of step 6

<table>
<tr><th>Step Nº |<i>n</i></th><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>AUC</th><th>Output</th></tr>
<tr><td>6.1
Training/ Testing

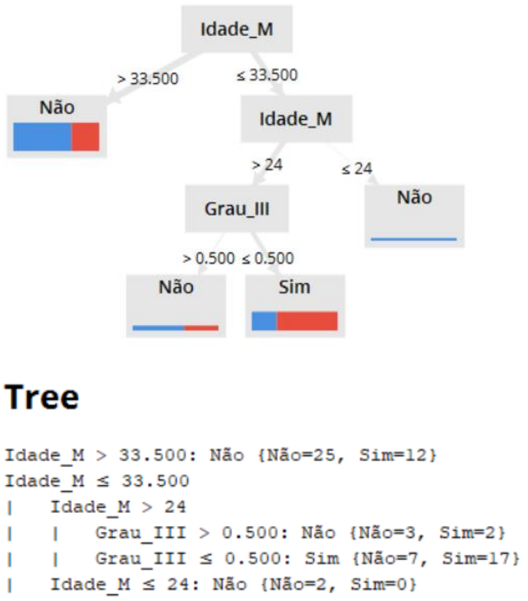
48 instances for training

20 instances for testing</td><td>80.00%</td><td>85.71%</td><td>66.67%</td><td>75.00%</td><td>0.717</td><td>

Idade_M
> 33.500
≤ 33.500
Não
Idade_M
> 24
≤ 24
Grau_III
Não
> 0.500 ≤ 0.500
Não
Sim

Tree
Idade_M > 33.500: Não {Não=25, Sim=12}
Idade_M ≤ 33.500
| Idade_M > 24
| | Grau_III > 0.500: Não {Não=3, Sim=2}
| | Grau_III ≤ 0.500: Sim {Não=7, Sim=17}
| Idade_M ≤ 24: Não {Não=2, Sim=0}

Model´s characteristics:
1-Validation.sampling_type = linear sampling
1-Decision Tree.criterion = accuracy
1-Decision Tree.apply_pruning = true
1-Decision Tree.minimal_size_for_split = 4
1-Decision Tree.minimal_gain = 0.1
1-Decision Tree.minimal_leaf_size = 2
1-Decision Tree.maximal_depth = 20</td></tr>
<tr><td>Validation

17 instances to validate the model 6.1.1</td><td>70.59%</td><td>66.67%</td><td>75.00%</td><td>70.59%</td><td>0.750</td><td>

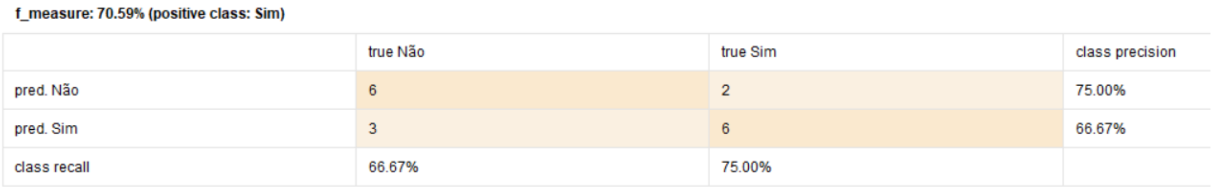
Idade_M
> 33.500
≤ 33.500
Não
Idade_M
> 24
≤ 24
Grau_III
Não
> 0.500 ≤ 0.500
Não
Sim

Tree
Idade_M > 33.500: Não {Não=25, Sim=12}
Idade_M ≤ 33.500
| Idade_M > 24
| | Grau_III > 0.500: Não {Não=3, Sim=2}
| | Grau_III ≤ 0.500: Sim {Não=7, Sim=17}
| Idade_M ≤ 24: Não {Não=2, Sim=0}

Model´s characteristics:
1-Validation.sampling_type = linear sampling
1-Decision Tree.criterion = accuracy
1-Decision Tree.apply_pruning = true
1-Decision Tree.minimal_size_for_split = 4
1-Decision Tree.minimal_gain = 0.1
1-Decision Tree.minimal_leaf_size = 2
1-Decision Tree.maximal_depth = 20</td></tr>
</table>

As we can see, the F-Measure and the AUC of the training/testing of the above model are close to the ones computed during the validation which tells us that the model is quite stable. Furthermore, despite most of its performance measures being slightly lower in the validation test, it is still the best validation measures that we have found during the first 8 modeling steps. Moreover, all other model validations have roughly failed the test (i.e. the other models have generated AUC measures between 0.5 and 0.613).

The confusion matrix and the ROC plot of the validation of the model ran in step 6.1 and disclosed in Table 5.41, can be seen below.



f_measure: 70.59% (positive class: Sim)

	true Não	true Sim	class precision
pred. Não	6	2	75.00%
pred. Sim	3	6	66.67%
class recall	66.67%	75.00%	

Figure 5.33 Confusion Matrix of the Validation of the Model 6.1

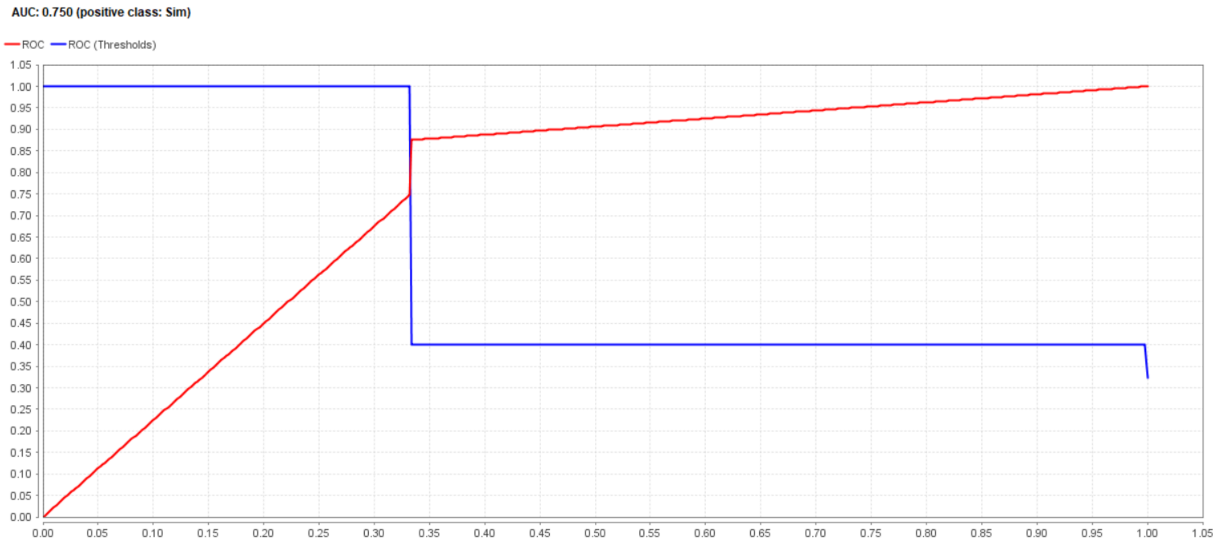


Figure 5.34 ROC plot of the validation of the model 6.1

If we analyze the confusion matrix depicted in Figure 5.33, we can see that it encompasses the 17 instances split for validation on the model 6.1, as well as its reported metrics. If we look at the ROC plot presented next, we see that the model 6.1 is fair due to its computed value (0.750).

If we interpret its decision tree model 6.1 we can say with 70.59% of accuracy through the model validation that:

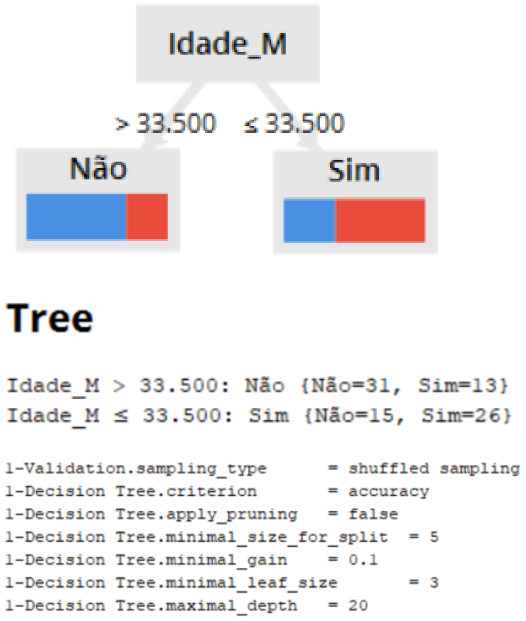
- The probability of a woman getting pregnant above 33 years old is lower (i.e. 32.43% (12/(12+25))) than for the ones below 33 years old (i.e. 61.29% (19/(19+12))).
- All women patients with 24 years old or younger were not able to get pregnant; and hence, the decision tree also tells us that woman below 24 years old does not get pregnant.

- Most of the women patients between 24 and 33 years old, that have a male partner that does not have a varicocele with a high severity grade (i.e. has a severity grade equal to I or II), are able to get pregnant (i.e. 70.83% (17/(17+7)).

Since some related works have not divided its dataset into training/testing and validation, we have also trained/tested this best generated model (i.e. model 6.1) without validating it to further on better compare our results with the ones obtained in related works (e.g. (Guh et al., 2011)). To do so, we have considered the decision tree Model 2 and mainly have deleted the "Split Data" operator from it. The computed results can be seen below in Table 5.42.

Table 5.42 Model of step 6 without validation

Step Nº |*n*	Accuracy	Precision	Recall	F-Measure	AUC	Output

6.1.1
Training/
Testing
59 instances for training
26 instances for testing
	80.77%	70.00%	77.78%	73.68%	0.801	
						
						Tree
						`Idade_M > 33.500: Não {Não=31, Sim=13}`
						`Idade_M ≤ 33.500: Sim {Não=15, Sim=26}`
						`1-Validation.sampling_type = shuffled sampling`
						`1-Decision Tree.criterion = accuracy`
						`1-Decision Tree.apply_pruning = false`
						`1-Decision Tree.minimal_size_for_split = 5`
						`1-Decision Tree.minimal_gain = 0.1`
						`1-Decision Tree.minimal_leaf_size = 3`
						`1-Decision Tree.maximal_depth = 20`

As we can see, the performance measures of model 6.1.1 surpassed the ones obtained in the model 6.1 during its validation. However, the computed decision tree is not very interesting: It only says that woman patients 33 years old or younger achieve pregnancy and that woman patients older than 33 years old do not. However, this result was useful to assess if the tree root of the model 6.1 and 6.1.1 remained the same (i.e. with the same splitting attribute and value).

If we analyze the number of instances obtained in each tree leaves of the model 6.1, we see that we have a small number of instances in most of them, specially favored by the restricted number of assessed instances (i.e. 85 instances). Hence, we have also looked up for the right ran decision tree model upon the 230 instances that we have also assessed and have seen that the decision tree modeling step 3.1 has generated the model with the second highest performance measure during the training of the 8 modeling steps with an Accuracy of 74.55%, an F-measure of 73.08% and a AUC measure of 0.747 during the training/testing of the model. However, its issue is its validation measures that indicates that it is a worthless test since its AUC measure is equal to 0.554 and its Accuracy and F-Measure measure is equal to 47.83%. However, due to its acceptable performance during its training upon the 230 instances, we have tried to improve its classification by applying the *Bagging* ensemble method with the aim of also minimizing the clear test overfit that we have got in the training/testing of the model. To do this, the model depicted in Figure 6.12 was built to run the Rapid Miner´s decision tree algorithm within the *Bagging* RapidMiner operator with the parameter values found to be the best during the execution of the model 3.1. The right parameter values of model 3.1 are:

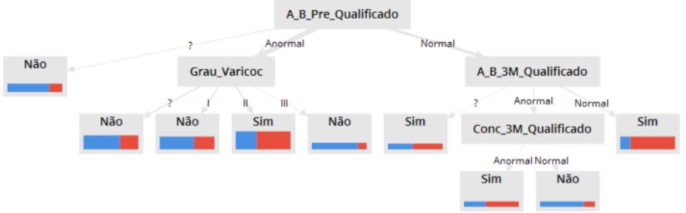
- Validation sampling type during training/testing: stratified;
- Decision tree splitting criterion: accuracy;
- Decision tree pruning: True;
- Decision tree minimal size for split: 4;
- Decision tree minimal gain: 0.1;
- Decision tree minimal leaf size: 4;
- Decision tree maximal depth: 20.

Regarding the choice of the *Bagging* ensemble method, we have used it because the related work carried out in (Guh et al., 2011) - the one that had its best decision tree model computed with an accuracy of 73.2% - also applied the *Bagging* ensemble method during the training/testing of their decision tree model to achieve its 73.2% of accuracy. However, we have seen that their 73.2% of accuracy was delivered during the training/testing of their model and not during its validation which gave us confidence to optimize the model found in step 3.1 since during the training/testing of our model and without the *bagging* ensemble method we have surpassed their accuracy by 1.35% (i.e. Accuracy of model 3.1 = 74.55%, see first row of Table 5.43). In the table below, we show the results with the application of the *Bagging* ensemble method on model 3.1, as described in the corresponding methods section, and at last, show the results of applying the model 3.1 on all the dataset.

Table 5.43 Bagging Application – Improvement of Model 3.1

<table>
<tr><th>Step Nº |<i>n</i></th><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>AUC</th><th>Output</th></tr>
<tr><td>3.1 Training/ Testing

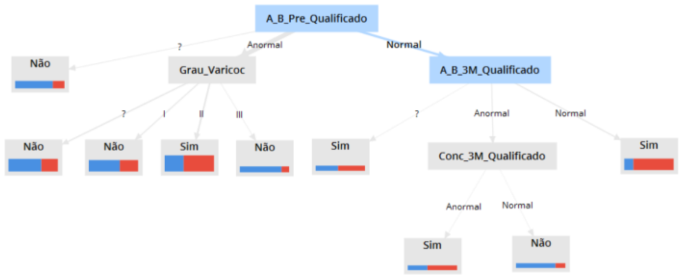
129 instances to train

55 instances to test</td><td>74.55%</td><td>73.08%</td><td>73.08%</td><td>73.08%</td><td>0.747</td><td>

A_B_Pre_Qualificado
?
Anormal
Normal
Não
Grau_Varicoc
A_B_3M_Qualificado
?
I
II
III
?
Anormal
Normal
Não
Não
Sim
Não
Sim
Conc_3M_Qualificado
Sim
Anormal Normal
Sim
Não

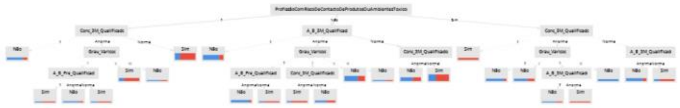
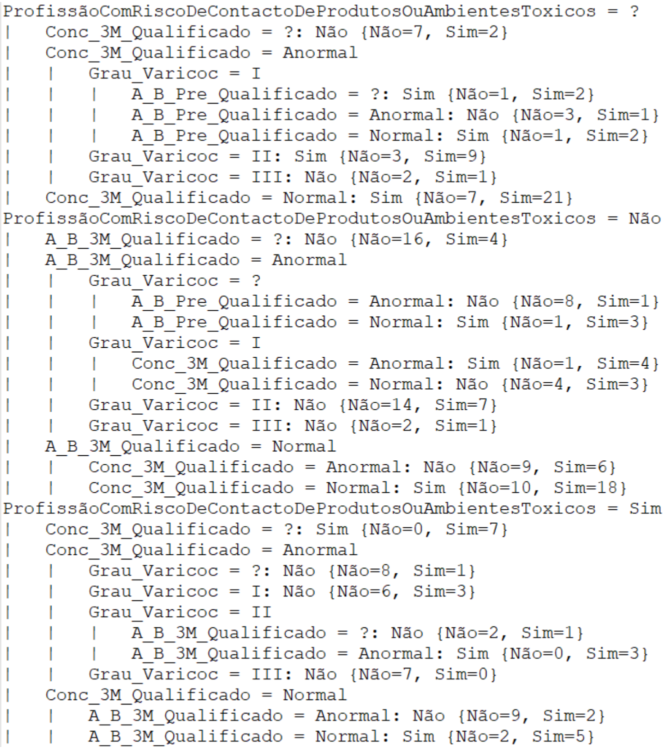
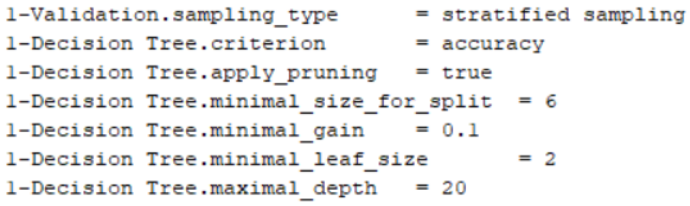
<b>Tree</b>
A_B_Pre_Qualificado = ?: Não {Não=13, Sim=4}
A_B_Pre_Qualificado = Anormal
| Grau_Varicoc = ?: Não {Não=20, Sim=10}
| Grau_Varicoc = I: Não {Não=17, Sim=10}
| Grau_Varicoc = II: Sim {Não=15, Sim=24}
| Grau_Varicoc = III: Não {Não=11, Sim=2}
A_B_Pre_Qualificado = Normal
| A_B_3M_Qualificado = ?: Sim {Não=5, Sim=6}
| A_B_3M_Qualificado = Anormal
| | Conc_3M_Qualificado = Anormal: Sim {Não=4, Sim=6}
| | Conc_3M_Qualificado = Normal: Não {Não=8, Sim=2}
| A_B_3M_Qualificado = Normal: Sim {Não=5, Sim=22}

1-Validation.sampling_type   = stratified sampling
1-Decision Tree.criterion   = accuracy
1-Decision Tree.apply_pruning   = true
1-Decision Tree.minimal_size_for_split = 4
1-Decision Tree.minimal_gain   = 0.1
1-Decision Tree.minimal_leaf_size   = 4
1-Decision Tree.maximal_depth   = 20</td></tr>
</table>

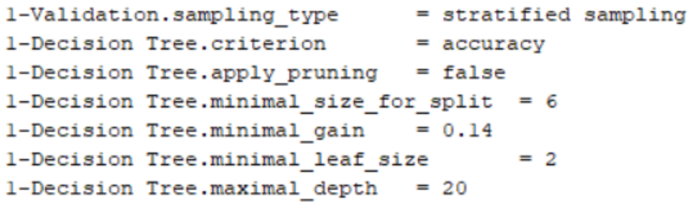
Validation of model 3.1 46 instances to validate the model 3.1	47.83%	44.00%	52.38%	47.83%	0.554	 A_B_Pre_Qualificado ? Anormal Normal Não Grau_Varicoc A_B_3M_Qualificado ? I II III ? Anormal Normal Não Não Sim Não Sim Conc_3M_Qualificado Sim Anormal Normal Sim Não **Tree** A_B_Pre_Qualificado = ?: Não {Não=13, Sim=4} A_B_Pre_Qualificado = Anormal | Grau_Varicoc = ?: Não {Não=20, Sim=10} | Grau_Varicoc = I: Não {Não=17, Sim=10} | Grau_Varicoc = II: Sim {Não=15, Sim=24} | Grau_Varicoc = III: Não {Não=11, Sim=2} A_B_Pre_Qualificado = Normal | A_B_3M_Qualificado = ?: Sim {Não=5, Sim=6} | A_B_3M_Qualificado = Anormal | | Conc_3M_Qualificado = Anormal: Sim {Não=4, Sim=6} | | Conc_3M_Qualificado = Normal: Não {Não=8, Sim=2} | A_B_3M_Qualificado = Normal: Sim {Não=5, Sim=22}
3.1.1 Trainning/ Testing of model 3.1 with Bagging and with the optimized parameters of model 3.1 129 instances to train 55 instances to test	70.91%	77.78%	53.85%	63.64%	0.668	(Generated 10 decision trees with the baggining ensemble method)
Validation of model 3.1.1 with Bagging and with the optimized parameters of model 3.1 46 instances to validate the model 3.1	58.70%	54.55%	57.14%	55.81%	0.644	(Generated 10 decision trees with the baggining ensemble method)

<table>
<tr><th></th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td>3.1.1.1 Training/ Testing of model 3.1 applied to all instances (i.e. without Validation) and without Bagging

161 instances to train

69 instances to test</td><td>68.12%</td><td>63.16%</td><td>75.00%</td><td>68.57%</td><td>0.652</td><td>

ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?
| Conc_3M_Qualificado = ?: Não {Não=7, Sim=2}
| Conc_3M_Qualificado = Anormal
| | Grau_Varicoc = I
| | | A_B_Pre_Qualificado = ?: Sim {Não=1, Sim=2}
| | | A_B_Pre_Qualificado = Anormal: Não {Não=3, Sim=1}
| | | A_B_Pre_Qualificado = Normal: Sim {Não=1, Sim=2}
| | Grau_Varicoc = II: Sim {Não=3, Sim=9}
| | Grau_Varicoc = III: Não {Não=2, Sim=1}
| Conc_3M_Qualificado = Normal: Sim {Não=7, Sim=21}
ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não
| A_B_3M_Qualificado = ?: Não {Não=16, Sim=4}
| A_B_3M_Qualificado = Anormal
| | Grau_Varicoc = ?
| | | A_B_Pre_Qualificado = Anormal: Não {Não=8, Sim=1}
| | | A_B_Pre_Qualificado = Normal: Sim {Não=1, Sim=3}
| | Grau_Varicoc = I
| | | Conc_3M_Qualificado = Anormal: Sim {Não=1, Sim=4}
| | | Conc_3M_Qualificado = Normal: Não {Não=4, Sim=3}
| | Grau_Varicoc = II: Não {Não=14, Sim=7}
| | Grau_Varicoc = III: Não {Não=2, Sim=1}
| A_B_3M_Qualificado = Normal
| | Conc_3M_Qualificado = Anormal: Não {Não=9, Sim=6}
| | Conc_3M_Qualificado = Normal: Sim {Não=10, Sim=18}
ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim
| Conc_3M_Qualificado = ?: Sim {Não=0, Sim=7}
| Conc_3M_Qualificado = Anormal
| | Grau_Varicoc = ?: Não {Não=8, Sim=1}
| | Grau_Varicoc = I: Não {Não=6, Sim=3}
| | Grau_Varicoc = II
| | | A_B_3M_Qualificado = ?: Não {Não=2, Sim=1}
| | | A_B_3M_Qualificado = Anormal: Sim {Não=0, Sim=3}
| | Grau_Varicoc = III: Não {Não=7, Sim=0}
| Conc_3M_Qualificado = Normal
| | A_B_3M_Qualificado = Anormal: Não {Não=9, Sim=2}
| | A_B_3M_Qualificado = Normal: Sim {Não=2, Sim=5}

1-Validation.sampling_type   = stratified sampling
1-Decision Tree.criterion   = accuracy
1-Decision Tree.apply_pruning   = true
1-Decision Tree.minimal_size_for_split = 6
1-Decision Tree.minimal_gain   = 0.1
1-Decision Tree.minimal_leaf_size   = 2
1-Decision Tree.maximal_depth   = 20</td></tr>
<tr><td>3.1.1.1.1 Training/ Testing of model 3.1.1.1 with Bagging

161 instances to train

69 instances to test</td><td>72.46%</td><td>67.57%</td><td>78.12%</td><td>72.46%</td><td>0.758</td><td>(Generated 10 decision trees with the bagging ensemble method)

1-Validation.sampling_type   = stratified sampling
1-Decision Tree.criterion   = accuracy
1-Decision Tree.apply_pruning   = false
1-Decision Tree.minimal_size_for_split = 6
1-Decision Tree.minimal_gain   = 0.14
1-Decision Tree.minimal_leaf_size   = 2
1-Decision Tree.maximal_depth   = 20</td></tr>
</table>

As we can appreciate, the results obtained with the bagging ensemble increased the accuracy of the validation of model 3.1 (i.e the accuracy of the validation of model 3.1=47.83% and the accuracy of the validation of model 3.1.1=58.70%) but its AUC, despite also increasing from 0.554, in model 3.1, to 0.644, in model 3.1.1, remains poor.

Regarding the training/testing of the model 3.1, we can see that the values have decreased after the application of Bagging (i.e. the accuracy of the training of the model 3.1=74.55% and the accuracy of the training of the model 3.1.1=70.91%) which depicts the minimization of the overfit that this ensemble method promises.

If we look at the two last rows of the table above, we see the results of applying model 3.1 to all 230 instances without *bagging* (i.e. Model 3.1.1.1) and with *bagging* (i.e. Model 3.1.1.1.1). If we compare these two results, we see that with the *bagging*, the model has increased its accuracy (i.e the accuracy of the model 3.1.1.1=68.12% and the accuracy of the model 3.1.1.1.1=72.46%) and the AUC has reached a fair performance measure (AUC of model 3.1.1.1 = 0.652 and AUC of model 3.1.1.1.1=0.758). However, the F-measure of model 3.1.1.1.1 is slightly under the one obtained in model 3.1 (i.e. F-measure of the training model 3.1 =73.08% and the F-measure of the training model 3.1.1.1.1 =72.0846%); and hence, model 3.1 remained the right model during step 3 for the assessment of all 230 instances. If we compare the training of the model 3.1 to the model 6.1, we conclude that 6.1 is a better model due to its higher f-measure (i.e. the f-measure of the training of the model 3.1=73.08% and the f-measure of the training of the model 6.1=75%). Furthermore, model 6.1 shows a lower difference between the performance measures obtained during the training of the model vs the ones obtained during its validation.

5.5.2 Clustering

The aim of identifying data patterns and better understanding the selected numerical attributes in the context of the label attribute “Gravidez”, as well as the need to discretize the “Idade_M” attribute for further data mining applications, made us apply the well-known K-means algorithm upon the *final preprocessed data set*.

From the groups of attributes tested discussed in section 5.4.4, we have seen that the fifth group (i.e. the group with the attributes: Idade_M, A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_Varicoc, ProfissãoComRiscoDeContactoDeProdutosOuAmbientes, Gravidez) was the one that provided, with the model shown in Figure 6.14, the most interesting clusters with the following configurations: Number of clusters = 4 and Numerical measure for the distance calculation = Euclidean Distance (see in the Appendix C.2 the results of all 114 performed tests).

One of the things that made this model stand out from the rest was its higher number of covered patients/instances (85 instances for the fifth group of attributes vs 28 instances for the first and forth group and 76 instances, for the third and sixth group of attributes). In fact, due to the K-Means specificity that requires all attribute values to be filled, the computed results ended up with a fewer number of instances in each generated cluster due to the reduced number of filtered instances. However, since it was the one that covered the highest number of instances, we have decided to further on seek for an interesting data pattern based on that group of attributes that could maintain the same number of instances. With that in mind, we have decided to initially add to the fifth group of attributes, the age of the male patient recorded in the attribute “Idade_H” since it does not have missing values. In fact, despite the attribute “Idade_H” not being a statistically significant attribute, clinically, it is an interesting one that could support a more domain wise data pattern. Furthermore, we have also tested the K-Means algorithm with several different data transformations upon the attribute “Grau_Varicoc” and “Idade_M”: The “Grau_Varicoc” attribute was manually dichotomized as shown in Figure 6.21 and mapped as shown in Figure 6.24; and at last, women ages were discretized as shown in Figure 6.25 and applied to the model depicted in Figure 6.21 with the fine-tuned model disclosed in Figure 6.26. Therefore, in the following subsections, we present and interpret the obtained results from each of these built data mining models described in the Appendix B.2.

5.5.2.1 Severity grade dichotomized (Model 2)

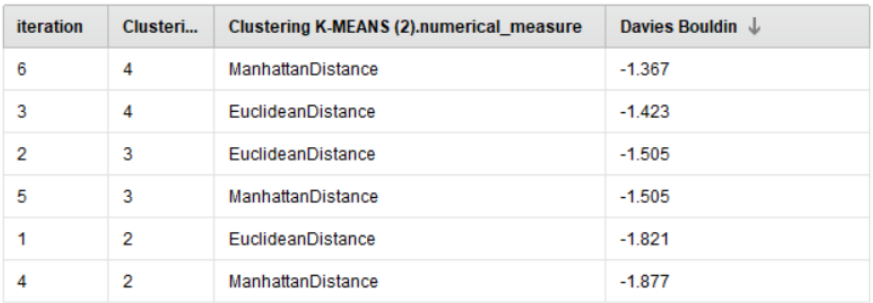
In this subsection, we present the results obtained with the model depicted in Figure 6.21 where we have manually dichotomized the severity grade attribute “Grau_Varicoc” as specified in Table 6.5.

This model was executed with the following conditions:

- Selected attributes: Idade_M, Idade_H, A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_I, Grau_II, Grau_III, ProfissãoComRiscoDeContactoDeProdutosOuAmbientes, Gravidez;
- Filtered the attributes listed above by non-missing values; and therefore, iterations were performed upon 85 instances;
- All attributes were transformed into numerical values (i.e. 0 and 1) and the woman´s age attribute “Idade_M” was normalized; and hence, the assessed data set only contained numerical attributes going from 0 to 1;
- Number of clusters: 2 to 4;
- Distance measure: Euclidean and Manhattan Distance.

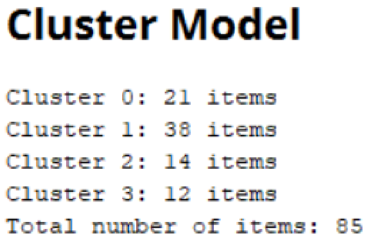
After testing the data mining model depicted in Figure 6.21 with different settings (i.e. a number of clusters going from 2 to 4 and a numerical measure of Manhattan or Euclidean), we have concluded that the best iteration was the one with the following K-Means setting: Number of clusters = 4 and Numerical measure for the distance calculation = Manhattan Distance. In fact, this iteration can be seen in Table 5.44, identified as the iteration 6, with the lowest Davies Bouldin index (Davies Bouldin index = -1.367). Hence, we further on showcase its results.

Table 5.44 Davies Bouldin results for the model depicted in Figure 6.21



iteration	Clusteri...	Clustering K-MEANS (2).numerical_measure	Davies Bouldin ↓
6	4	ManhattanDistance	-1.367
3	4	EuclideanDistance	-1.423
2	3	EuclideanDistance	-1.505
5	3	ManhattanDistance	-1.505
1	2	EuclideanDistance	-1.821
4	2	ManhattanDistance	-1.877

This iteration has separated the 85 patients into 4 groups of similar patient’s characteristics and each cluster covered between 12 to 38 patients, as we can see in Figure 5.35.



Cluster Model

```
Cluster 0: 21 items
Cluster 1: 38 items
Cluster 2: 14 items
Cluster 3: 12 items
Total number of items: 85
```

Figure 5.35 Cluster´s distribution in the iteration nº 6 – Best result

The *centroid table* computed in this iteration clearly shows that what mainly differentiates each generated cluster is varicocele severity grade. In fact, if we analyze this centroid table depicted in Table 5.45, along with its complementary ones disclosed in Table 5.46 and Table 5.47 we can formulate the following conclusions – please note that for the attributes that can only be assigned the value 0 or 1, we can consider the centroid means in Table 5.46 as relative frequencies for the value 1:

- Cluster 0 only has patients with a low varicocele severity grade (i.e. "Grau_Varicoc"=I);
- Cluster 1 only has patients with a moderate varicocele severity grade (i.e. "Grau_Varicoc"=II);
- Cluster 2 has a mix of low and high varicocele severity grades (i.e. "Grau_Varicoc"=I or III). However, most of them (78.6%, which is 11 patients), have the severity grade I and the remaining ones (21.4%, which is 3 patients), have the severity grade III;
- Cluster 3 only has patients with a high varicocele severity grade (i.e. "Grau_Varicoc"=III).

Regarding the other attributes shown in the centroid table we can also conclude that:

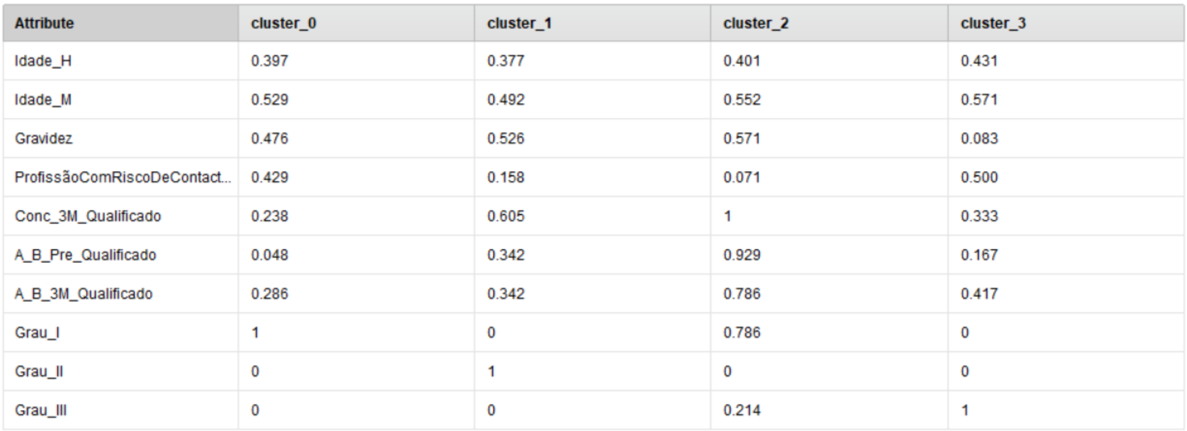
- The means of the patients´ ages do not vary much across the several clusters (see centroid´s means in the Idade_H and Idade_M attribute) which means that the patients of each data cluster have similar ages. In fact, there is no statistical significance between them ($p>0.05$);
- Despite cluster 0, 1 and 2 having a similar relative frequency of patients that conceived in the cluster 3, only 8.3% (i.e. 1 patient out of the 12) conceived and all of them had a high varicocele severity grade;
- Despite cluster 0 and 3 having a similar relative frequency of patients that work in toxic environments or with toxic products (respectively 42.9% and 50%), cluster 1 and 2 have a very low number of patients in these conditions. In fact, cluster 1 indicates that only 15.8% of its patients do work in a toxic environment and in cluster 2, only 7.1% also do so. Further on, we see that patients that did not work in toxic environments mainly had a severity grade equal to I or II. In fact, the joining of cluster 1 and 2 includes 52 patients (38 patients for the cluster 1 and 14 patients for cluster 2) and only 21.4% of the patients in cluster 2 (i.e. 3 patients) had a severity grade III, which means that 94.23% of the patients in these 2 clusters ((52 patients – 3 patients with the high severity grade in the cluster 2)/52) have a severity grade that is low to moderate when the patient works in a none toxic environment;

- Cluster 2 has also the highest relative frequency of patients with a normal sperm parameter value (i.e. 100% with a “Con_3M_Qualificado” set to “Normal”, 92.9% for the “A_B_Pre_Qualificado” and 78.6% for the A_B_3M_Qualificado). Furthermore, this cluster has the lowest relative frequency of patients that work in a toxic environment (7.1%).

The above-mentioned conclusions can be visualized through the series plot depicted in Figure 5.36 where we can see that each colored line represents each cluster described in the centroid table disclosed in Table 5.45. These colored lines can be interpreted as:

- Blue line – presents the mean value for each attribute within cluster 0;
- Green line – presents the mean value for each attribute within cluster 1;
- Yellow line – presents the mean value for each attribute within cluster 2;
- Red line – presents the mean value for each attribute within cluster 3.

Table 5.45 Centroid Table for the iteration nº6 – Best result



Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Idade_H	0.397	0.377	0.401	0.431
Idade_M	0.529	0.492	0.552	0.571
Gravidez	0.476	0.526	0.571	0.083
ProfissãoComRiscoDeContact...	0.429	0.158	0.071	0.500
Conc_3M_Qualificado	0.238	0.605	1	0.333
A_B_Pre_Qualificado	0.048	0.342	0.929	0.167
A_B_3M_Qualificado	0.286	0.342	0.786	0.417
Grau_I	1	0	0.786	0
Grau_II	0	1	0	0
Grau_III	0	0	0.214	1

Table 5.46 Complementary Centroid Table for the iteration nº6 – Mean, Standard-deviation and ANOVA result – Best result

	Cluster 0 n=21	Cluster 1 n=38	Cluster 2 n=14	Cluster 3 n=12	*P*
Idade_H	36.524 (±5.006)	35.947 (±5.266)	36.643 (±4.236)	37.500 (±5.248)	0.778
Idade_M	33.762 (±4.049)	32.789 (±4.400)	34.357 (±4.517)	34.833 (±3.927)	0.336
Gravidez	0.476 (±0.512)	0.526 (±0.506)	0.571 (±0.514)	0.083 (±0.289)	0.030
ProfissãoComRiscoDeContacto	0.429 (±0.507)	0.158 (±0.370)	0.071 (±0.267)	0.500 (±0.522)	0.007
Conc_3M_Qualificado	0.238 (±0.436)	0.605 (±0.495)	1 (±0)	0.333 (±0.492)	0.001
A_B_Pre_Qualificado	0.048 (±0.218)	0.342 (±0.481)	0.929 (±0.267)	0.167 (±0.389)	0.001
A_B_3M_Qualificado	0.286 (±0.463)	0.342 (±0.481)	0.786 (±0.426)	0.417 (±0.515)	0.011
Grau_I	1 (±0)	0 (±0)	0.786 (±0.426)	0 (±0)	0.001
Grau_II	0 (±0)	1 (±0)	0 (±0)	0 (±0)	0.001
Grau_III	0 (±0)	0 (±0)	0.214 (±0.426)	1 (±0)	0.001

Table 5.47 Complementary Centroid Table for the iteration nº6 – Frequency – Best result

	Value	Cluster 0 n=21	Cluster 1 n=38	Cluster 2 n=14	Cluster 3 n=12
Idade_H					
Idade_M					
Gravidez	0 (Não)	11	18	6	11
	1 (Sim)	10	20	8	1
ProfissãoComRiscoDe Contacto	0 (Não)	12	32	13	6
	1 (Sim)	9	6	1	6
Conc_3M_Qualificado	0 (Anormal)	16	15	0	8
	1 (Normal)	5	23	14	4
A_B_Pre_Qualificado	0 (Anormal)	20	25	1	10
	1 (Normal)	1	13	13	2
A_B_3M_Qualificado	0 (Anormal)	15	25	3	7
	1 (Normal)	6	13	11	5
Grau_I	0	0	38	3	12
	1 (Grau_Varicoc=1)	21	0	11	0
Grau_II	0	21	0	14	12
	1 (Grau_Varicoc=2)	0	38	0	0
Grau_III	0	21	38	11	0
	1 (Grau_Varicoc=3)	0	0	3	12

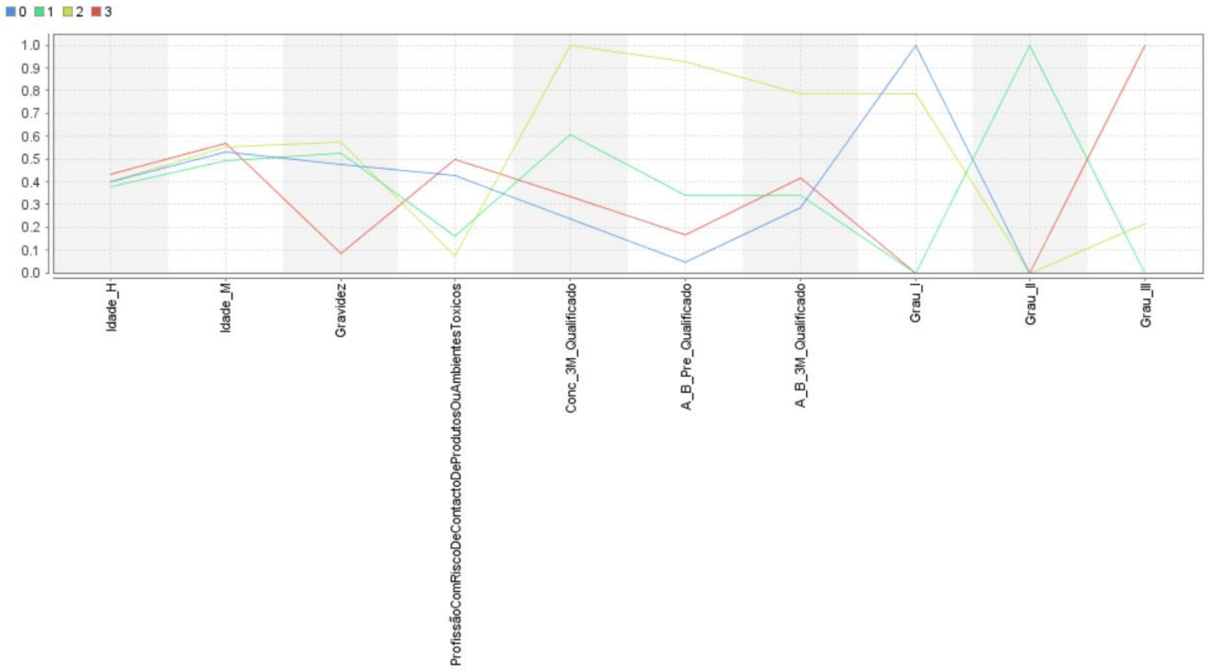


Figure 5.36 Series plot of the centroid means presented in Table 5.45 – Best result

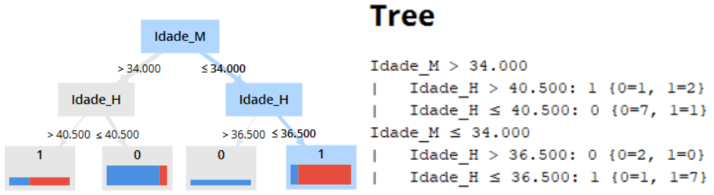
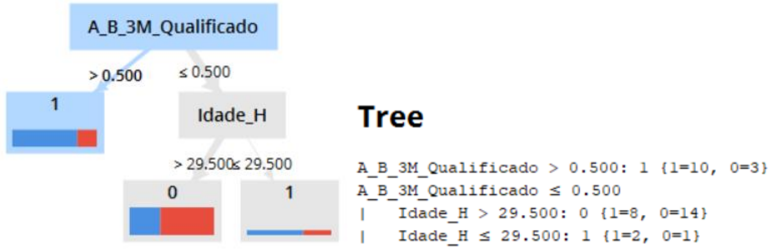
Due to the interesting results obtained, we have applied the Decision tree model showcased in Figure 6.22 (i.e. Model 2.1) on the clustered instances generated. Since the first goal of this study is to predict the success of the embolization treatment, we have generated trees by considering the "Gravidez" attribute as a label. These decision trees are shown in Table 5.48 and were generated with the following model parameter values that were found to be the best during the application of the decision tree algorithm (i.e. modeling step 6.1 depicted in Table 5.41):

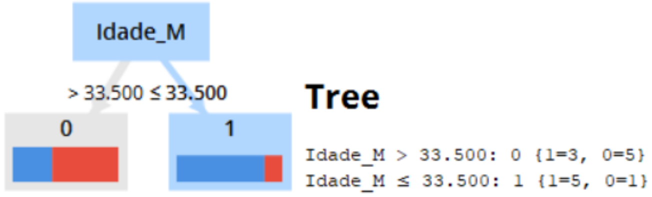
- Criterion = Accuracy;

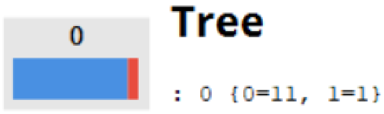
- Apply pruning = True;
- Minimal Size For split = 4;
- Minimal gain = 0.1;
- Minimal Leaf Size = 2;
- Maximal depth = 20;

The decision trees in Table 5.48 have “age” denormalized to better interpret the computed results. The decision tree leaf nodes have the pregnancy outcome normalized; and hence, the value 0 can be interpreted as *not pregnant* and the value 1, as *pregnant*. The decision tree paths that cover the highest frequency of pregnancies are highlighted in blue.

Table 5.48 Decision Tree results upon the clusters generated in the last iteration 6

Cluster Nº | n	Decision tree
Cluster 0 n=21	 Idade_M > 34.000 ≤ 34.000 Idade_H Idade_H > 40.500 ≤ 40.500 > 36.500 ≤ 36.500 1 0 0 1 Tree Idade_M > 34.000 | Idade_H > 40.500: 1 {0=1, 1=2} | Idade_H ≤ 40.500: 0 {0=7, 1=1} Idade_M ≤ 34.000 | Idade_H > 36.500: 0 {0=2, 1=0} | Idade_H ≤ 36.500: 1 {0=1, 1=7}
Cluster 1 n=38	 A_B_3M_Qualificado > 0.500 ≤ 0.500 1 Idade_H > 29.500 ≤ 29.500 0 1 Tree A_B_3M_Qualificado > 0.500: 1 {1=10, 0=3} A_B_3M_Qualificado ≤ 0.500 | Idade_H > 29.500: 0 {1=8, 0=14} | Idade_H ≤ 29.500: 1 {1=2, 0=1}

Cluster 2 n=14	
	Tree
	Idade_M > 33.500: 0 {1=3, 0=5}
	Idade_M ≤ 33.500: 1 {1=5, 0=1}

Cluster 3 n=12	 Tree : 0 {0=11, 1=1}

If we analyze the decision trees above, we can appreciate that the cluster that delivers the most interesting information is the cluster 1, which has the highest number of patients. The decision tree of cluster 1 tells us that a normal sperm progressive motility at 3 months after embolization generates more pregnancies (i.e. 10/(10+3)=76.9%) and that the highest frequency of unsuccessful couples (14/(8+14)=63.63%) is held by male patients above 29 years old with abnormal sperm progressive motility at 3 months after treatment. Regarding Cluster 0, it tells us that most pregnancies occur when couples are of a male patient 36 years old or younger and a woman 34 years old or younger (7(/1+7)=87.5%). This cluster also tells us that these couples are less successful when the male patient is 40 years old or younger and the woman older than

34. The decision tree generated from the cluster 3 complies with what was said previously: most patients in this cluster are unsuccessful (i.e. 11/(11+1)=91.67%) . Cluster 2 indicates that woman patients 33 years old or younger are more prone to get pregnant (5/(5+1)=83.33%). This result makes us recall what we have seen during the Decision tree modelling step 6 depicted in Table 5.42.

Further on, we have rerun this k-means model without the "Idade_H" attribute to assess if the Davies Bouldin index would decrease and it has slightly lowered to -1.349. However, since all other computed results remained the same and our focus is on the male infertility, we have continued with model 2 as the best model.

5.5.2.2 Severity grade mapped (Model 3)

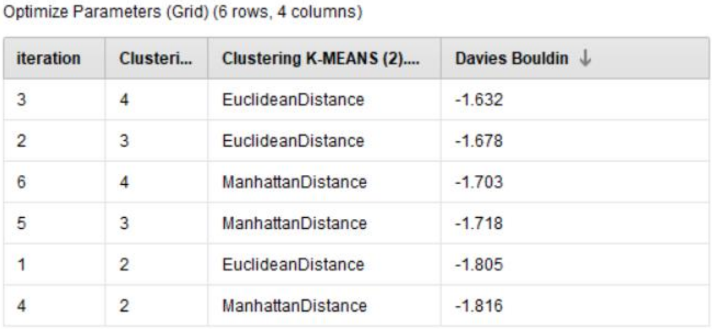
In this subsection, we present the results obtained with the model depicted in Figure 6.24. This model had the severity grade mapped, instead of dichotomized as previously disclosed.

This model was executed with the following conditions:

- Selected attributes: Idade_M, Idade_H, A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_Varicoc, ProfissãoComRiscoDeContactoDeProdutosOuAmbientes, Gravidez.
- Filtered the above listed attributes by non-missing values; and therefore, all these iterations were ran on the same 85 instances.
- Number of clusters: 2 to 4.
- Distance measure: Euclidean and Manhattan Distance

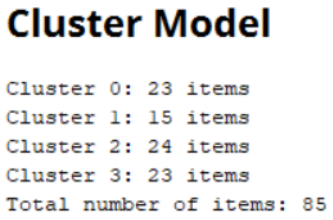
This model generated the Davies Bouldin results presented in Table 5.49 where we can see that the best setting for this model is: Distance measure = Euclidean Distance and Number Clusters=4 (Davies Bouldin index=-1.632). The generated clusters have the distribution presented in the Figure 5.37.

Table 5.49 Davies Bouldin results for the model depicted in Figure 6.24



Optimize Parameters (Grid) (6 rows, 4 columns)

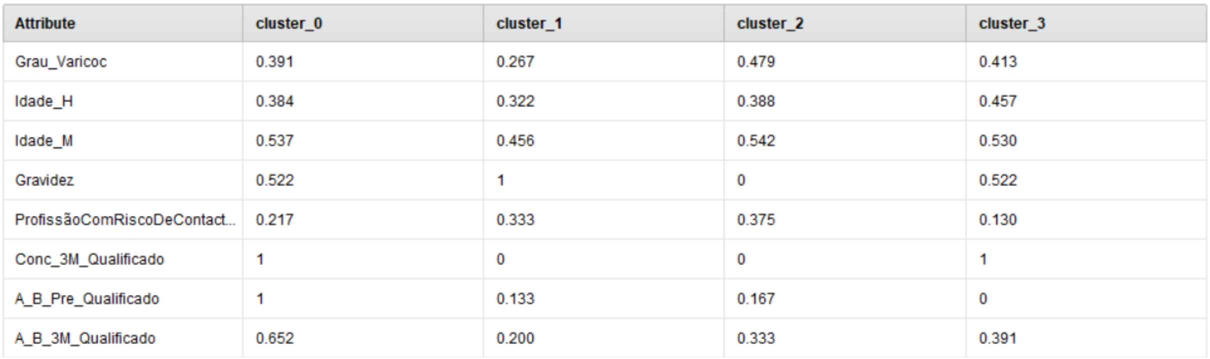
iteration	Clusteri...	Clustering K-MEANS (2)....	Davies Bouldin ↓
3	4	EuclideanDistance	-1.632
2	3	EuclideanDistance	-1.678
6	4	ManhattanDistance	-1.703
5	3	ManhattanDistance	-1.718
1	2	EuclideanDistance	-1.805
4	2	ManhattanDistance	-1.816



```
Cluster Model

Cluster 0: 23 items
Cluster 1: 15 items
Cluster 2: 24 items
Cluster 3: 23 items
Total number of items: 85
```

Figure 5.37 Cluster´s distribution in the iteration nº 3

Table 5.50 Centroid Table for the iteration nº3



Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Grau_Varicoc	0.391	0.267	0.479	0.413
Idade_H	0.384	0.322	0.388	0.457
Idade_M	0.537	0.456	0.542	0.530
Gravidez	0.522	1	0	0.522
ProfissãoComRiscoDeContact...	0.217	0.333	0.375	0.130
Conc_3M_Qualificado	1	0	0	1
A_B_Pre_Qualificado	1	0.133	0.167	0
A_B_3M_Qualificado	0.652	0.200	0.333	0.391

To better support our analysis of the series plot depicted in Figure 5.38, we have built complementary centroid tables that are depicted in the Table 5.51 and Table 5.52, below. The first complementary centroid table presents: de-normalized means (these means have its cells colored in beige); ± the standard deviations; and the statistical significance (*p*). The next complementary centroid table presents the frequencies for each attribute value specified under the column "Value".

Table 5.51 Complementary Centroid Table for the iteration nº3 – Mean, Standard-deviation and ANOVA result

	Cluster 0 n=23	Cluster 1 n=15	Cluster 2 n=24	Cluster 3 n=23	*p*
Grau_Varicoc	1.783 (±0.795)	1.533 (±0.640)	1.958 (±0.806)	1.826 (±0.576)	0.345
Idade_H	36.13 (±4.605)	34.333 (±3.940)	36.25 (±4.674)	38.261 (±5.856)	0.111
Idade_M	33.957 (±4.084)	31.867 (±3.543)	34.083 (±4.872)	33.783 (±4.199)	0.384
Gravidez	0.522 (±0.511)	1 (±0)	0 (±0)	0.522 (±0.511)	0.000
ProfissãoComRiscoDeContacto	0.217 (±0.422)	0.333 (±0.488)	0.375 (±0.495)	0.130 (±0.344)	0.236
Conc_3M_Qualificado	1 (±0)	0 (±0)	0 (±0)	1 (±0)	(not significant)
A_B_Pre_Qualificado	1 (±0)	0.133 (±0.352)	0.167 (±0.381)	0 (±0)	0.000
A_B_3M_Qualificado	0.652 (±0.487)	0.2 (±0.414)	0.333 (±0.482)	0.391 (±0.499)	0.028

Table 5.52 Complementary Centroid Table for the iteration nº3 – Frequency

	Value	Cluster 0 n=23	Cluster 1 n=15	Cluster 2 n=24	Cluster 3 n=23
Grau_Varicoc	1 (Grau_Varicoc=1)	10	8	8	6
	2 (Grau_Varicoc=2)	8	6	9	15
	3 (Grau_Varicoc=3)	5	1	7	2
Idade_H					
Idade_M					
Gravidez	0 (Não)	11	0	24	11
	1 (Sim)	12	15	0	12
ProfissãoComRiscoDeContacto	0 (Não)	18	10	15	20
	1 (Sim)	5	5	9	3
Conc_3M_Qualificado	0 (Anormal)	0	15	24	0
	1 (Normal)	23	0	0	23
A_B_Pre_Qualificado	0 (Anormal)	0	13	20	23
	1 (Normal)	23	2	4	0
A_B_3M_Qualificado	0 (Anormal)	8	12	18	13
	1 (Normal)	15	3	8	9

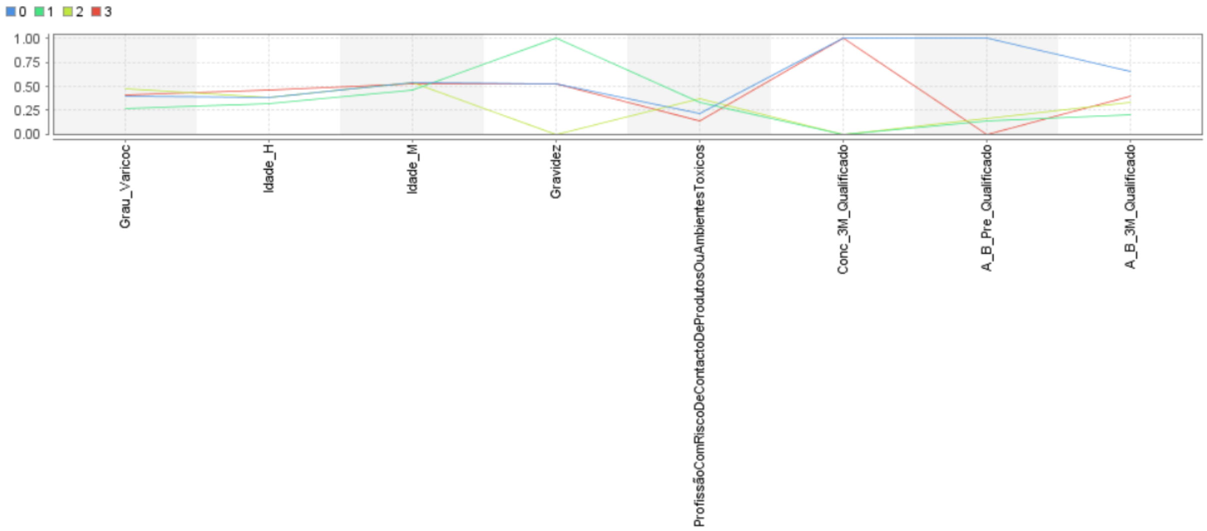


Figure 5.38 Series plot of the centroid means presented in Table 5.50

If we analyze the plot in Figure 5.38, we see that in all clusters, the varicocele severity grade, as well as the patient ages are quite the same (*p*>0.05). Moreover, the cluster represented in green (cluster 1) is quite the same as the one represented in yellow (cluster 2). In fact, we can see that most attributes have almost the same mean in these two clusters which tells us that the patients that were able to conceive do not have a different pattern from the ones that did not got pregnant. Similarly, we have the other two clusters – blue (Cluster 0) and red (Cluster 3) – that only differ in terms of sperm progressive motility. In fact, we see that the group of patients with normal sperm progressive motility before the treatment (Cluster 0) is still the largest one 3 months after the varicocele embolization in spite of its decline; i.e., from the group of 23 patients with normal sperm progressive motility before the treatment (Cluster 0), 8 of them show a decline in their sperm progressive motility 3 months after the treatment. Furthermore, if we compare cluster 0 and 3 with clusters 1 and 2, we see that we have a lower percentage of patients in cluster 0 and 3 that are exposed to toxic products or environments at their job than the patients in cluster 1 and 2. However, this gap is not statistically significant (p>0.05); and therefore, it did not add useful information.

5.5.2.3 Idade_M discretization (Model 4)

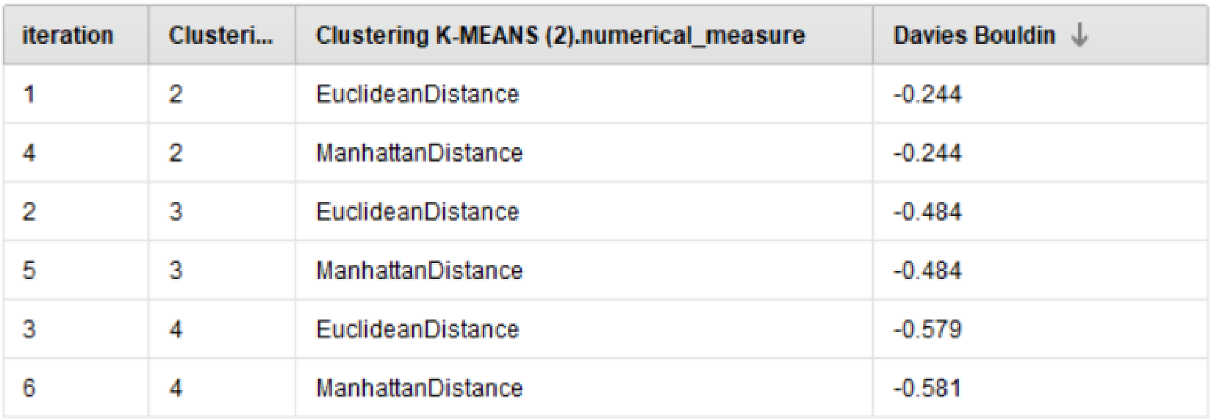
In this subsection, we present the results computed with the model depicted in Figure 6.25. This model aims to visually show how the woman age varies with the "Gravidez" attribute through a scatter plot since the ANOVA statistical results, disclosed in section 5.3.2.11, indicated that the mean of the woman age significantly varied (*p*=0.018) with the label attribute.

This model was executed with the following conditions:

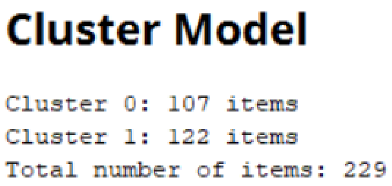
- Selected attributes: Idade_M, Gravidez;
- Filtered the above listed attributes by non-missing values; and therefore, all these iterations were upon the same 229 instances ran (Note that we have 230 instances with the "Gravidez" attribute filled but we have 1 missing value in the "Idade_M" attribute which leads us to 229 instances);
- Number of clusters: 2 to 4;
- Distance measure: Euclidean and Manhattan Distance.

This model generated the Davies Bouldin results presented in Table 5.53 where we can see that the best setting for this model is: Distance measure = Euclidean Distance and Number Clusters=2 (Davies Bouldin index=-0.244). The generated clusters have the distribution presented in Figure 5.39 where Cluster 0, represents the patients that conceived (i.e. "Gravidez"= "Sim") and Cluster 1, represents the patients that did not conceive (i.e. "Gravidez"= "Não"). Through the clustering distribution shown below, we can see that we have quite 7% less patients that conceived, as we have statistically previously seen (i.e. (122-107)/229=0.065. In the scatter plots depicted in Figure 5.40 and Figure 5.41, we can see the distribution of the women ages within these 2 clusters.

Table 5.53 Davies Bouldin results for the model depicted in Figure 6.25



iteration	Clusteri...	Clustering K-MEANS (2).numerical_measure	Davies Bouldin ↓
1	2	EuclideanDistance	-0.244
4	2	ManhattanDistance	-0.244
2	3	EuclideanDistance	-0.484
5	3	ManhattanDistance	-0.484
3	4	EuclideanDistance	-0.579
6	4	ManhattanDistance	-0.581



Cluster Model

```
Cluster 0: 107 items
Cluster 1: 122 items
Total number of items: 229
```

Figure 5.39 Cluster´s distribution in the iteration nº1

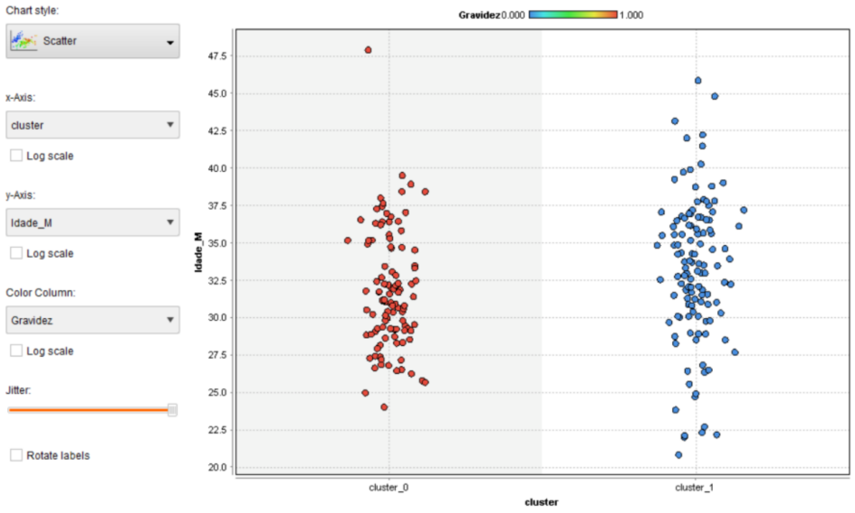


Chart style: Scatter

x-Axis: cluster

□ Log scale

y-Axis: Idade_M

□ Log scale

Color Column: Gravidez

□ Log scale

Jitter:

□ Rotate labels

Figure 5.40 Scatter plot between Idade_M vs Clusters – max jitter

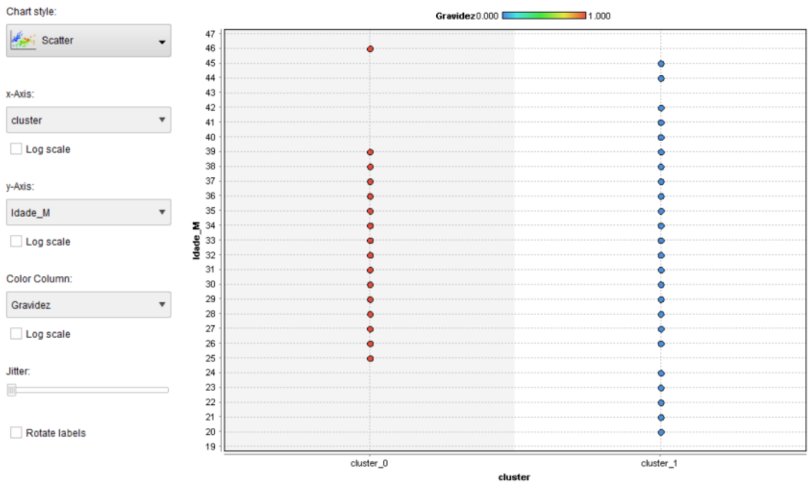


Chart style: Scatter

x-Axis: cluster

□ Log scale

y-Axis: Idade_M

□ Log scale

Color Column: Gravidez

□ Log scale

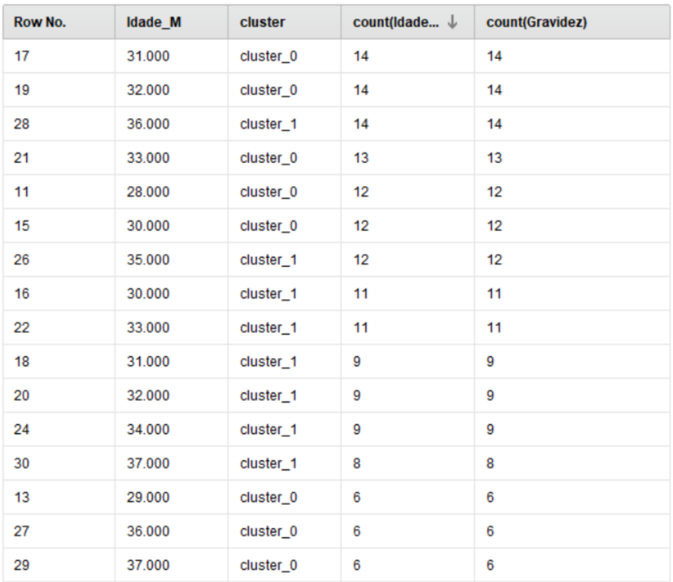
Jitter:

□ Rotate labels

Figure 5.41 Scatter plot between Idade_M vs Clusters – min jitter

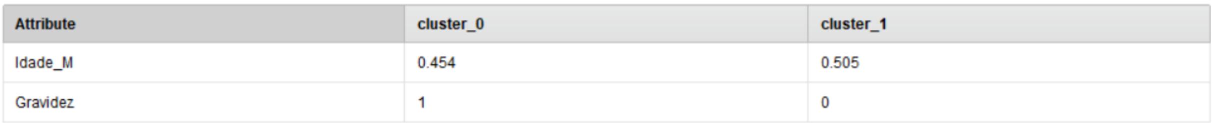
If we assess the frequencies of the women age depicted in Table 5.54, we see that the most common ages in the group of patients that conceived are: 31 (14 patients) and 32 (14 patients) years old and in the group of patient that did not conceive are: 36 (14 patients) and 35 (12 patients) years old. Hence, just by the most common ages of each cluster, we see that there is in fact a difference since the most frequent ages of the group of patients that got pregnant is lower than the other group which contributes to a lower age mean, as we can see in the centroid tables depicted in Figure 5.43.

Table 5.54 Highest frequencies of the filtered woman´s ages grouped by clusters



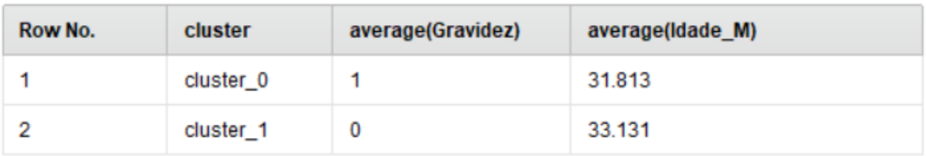
Row No.	Idade_M	cluster	count(Idade... ↓	count(Gravidez)
17	31.000	cluster_0	14	14
19	32.000	cluster_0	14	14
28	36.000	cluster_1	14	14
21	33.000	cluster_0	13	13
11	28.000	cluster_0	12	12
15	30.000	cluster_0	12	12
26	35.000	cluster_1	12	12
16	30.000	cluster_1	11	11
22	33.000	cluster_1	11	11
18	31.000	cluster_1	9	9
20	32.000	cluster_1	9	9
24	34.000	cluster_1	9	9
30	37.000	cluster_1	8	8
13	29.000	cluster_0	6	6
27	36.000	cluster_0	6	6
29	37.000	cluster_0	6	6

Table 5.55 Normalized Centroid Table for the iteration nº1



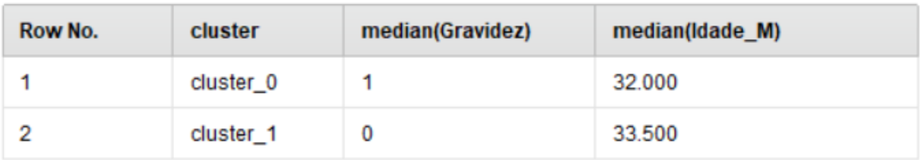
Attribute	cluster_0	cluster_1
Idade_M	0.454	0.505
Gravidez	1	0

Table 5.56 Denormalized Centroid Table for the iteration nº1 – mean



Row No.	cluster	average(Gravidez)	average(Idade_M)
1	cluster_0	1	31.813
2	cluster_1	0	33.131

Table 5.57 Denormalized Centroid Table for the iteration nº1 – median



Row No.	cluster	median(Gravidez)	median(Idade_M)
1	cluster_0	1	32.000
2	cluster_1	0	33.500

By assessing the scatter plots depicted in Figure 5.40 and Figure 5.41 we can say that there is a different age distribution between these two clusters that can be further explored. In fact, we can see ranges of age values with only the “Gravidez” attribute set to “Não” which leads us to discretize the “Idade_M” attribute by its entropy to potentiate the discovery of interesting information within these ranges of age values. Hence, if we focus on the scatter plot depicted in Figure 5.41 and in cluster 1, we can formulate the following discretization for the “Idade_M” attribute (the number of patients that falls in each range is specified between parentheses):

Idade_M<=24 (8 patients); 25<= Idade_M <= 39 (211 patients); Idade_M>= 40 (10 patients)

After looking up for the values of the label attribute “Gravidez” in each range of values, we have seen that in the middle range of values that goes from 25 to 39 years old, we have 106 patients with the “Gravidez” attribute set to “Sim” and 105 patients with the “Gravidez” attribute set to “Não”; and therefore, the *entropy* would be in this range of values close to 1. On the other hand, the other range of values would be close to 0: all patients below 24 years old have the “Gravidez” attribute set to “Não” and for the patients above 40 years old, only one patient out of the 10 patients identified, has the “Gravidez” attribute set to “Sim”. The problem with this discretization is that we have a small number of patients in the range of value with low entropy and the one close to 1 has the main patients; and therefore, we have sought for another type of discretization, but before that, we have applied the “Discretize by Entropy” operator that is available in the RapidMiner platform – this operator implements the entropy discretization proposed in (Fayyad & Irani, 1993). This operator has computed a range of values going from negative to positive infinity. This operator/discretization method was not able to propose a discretization for the “Idade_M” attribute which made us move to another more suitable method for our dataset. Hence, since this study also applies the association rule algorithm, which is based on frequencies of values, we have decided to explore the discretization per frequency method. Hence, we had the idea of building a box plot for the woman age, since each part of the box plot (i.e. Top Wisker, Upper Box, Lower Box and Bottom Wisker) cover 25% of the data; and therefore, we have formulated the discretization of the “Idade_M” attribute based on these parts of the box plot depicted in Figure 5.42. This box plot was built in EXCEL with the statistical values disclosed in Table 5.58 that were computed as specified in section 4.2.5.

Table 5.58 Main Statistical values of the “Idade_M” attribute

ID	Idade_M
Min	20
Q1	30
Median	32
Q3	36
Max	46
Mean	32.52
Mean-Median	0.515
Standard Deviation	4,23
n	229

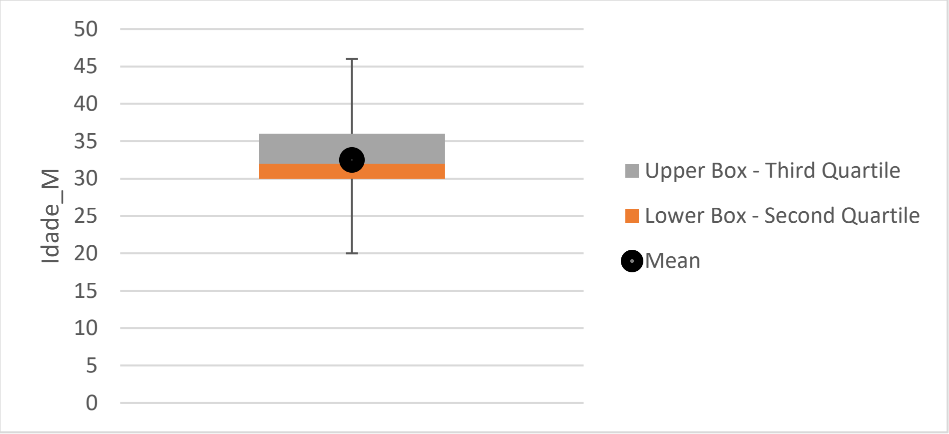


Figure 5.42 Woman´s age distribution – Box Plot of “Idade_M”

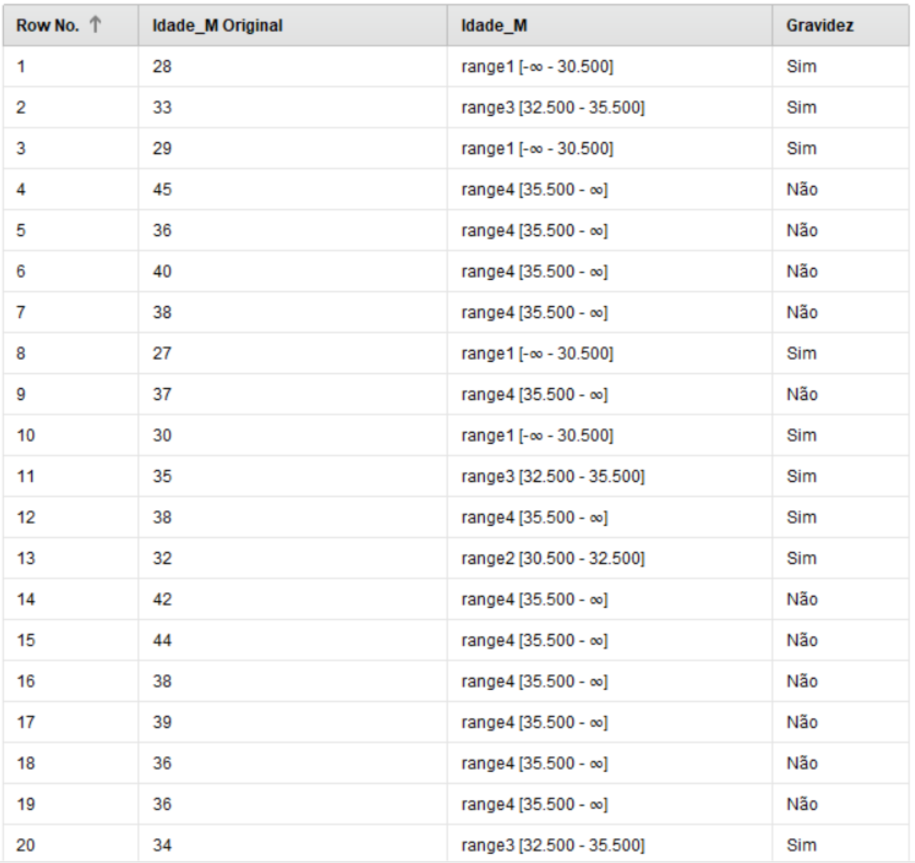
Thus, based on the statistical measures disclosed in Table 5.58Table 5.58, we have formulated and fine-tuned the following discretization for the “Idade_M” attribute:

Idade_M<=30; 30< Idade_M <=32; 32< Idade_M <36; Idade_M>=36

The above discretization was fine-tuned on the matter of its equalities with the RapidMiner operator called “Discretize By Frequency” for 4 bins. The “Discretize By Frequency” operator creates bins in such a way that the number of unique values in all bins are (almost) equal. This operator has computed the following ranges of values which enabled us to not only fine-tune the discretization previously set, but also go with the decision of using the “Discretize by frequency” operator to discretize the “Idade_M” attribute:

- Range 1: infinity to 30.500;
- Range 2: 30.500 to 32.500;
- Range 3: 32.500 to 35.500;
- Range 4: 35.500 to infinity.

In the below Figure 5.43, we can see a print screen of part of the computed results by the “Discretize by frequency” operator where we can see the original value under the column named “Idade_M Original”, the discretized value under the column named “Idade_M” and the label outcome under the column named “Gravidez”. The original value was computed with the RapidMiner´s “Generate Attribute” operator to assess if the assigned discretization was correct.



Row No. ↑	Idade_M Original	Idade_M	Gravidez
1	28	range1 [-∞ - 30.500]	Sim
2	33	range3 [32.500 - 35.500]	Sim
3	29	range1 [-∞ - 30.500]	Sim
4	45	range4 [35.500 - ∞]	Não
5	36	range4 [35.500 - ∞]	Não
6	40	range4 [35.500 - ∞]	Não
7	38	range4 [35.500 - ∞]	Não
8	27	range1 [-∞ - 30.500]	Sim
9	37	range4 [35.500 - ∞]	Não
10	30	range1 [-∞ - 30.500]	Sim
11	35	range3 [32.500 - 35.500]	Sim
12	38	range4 [35.500 - ∞]	Sim
13	32	range2 [30.500 - 32.500]	Sim
14	42	range4 [35.500 - ∞]	Não
15	44	range4 [35.500 - ∞]	Não
16	38	range4 [35.500 - ∞]	Não
17	39	range4 [35.500 - ∞]	Não
18	36	range4 [35.500 - ∞]	Não
19	36	range4 [35.500 - ∞]	Não
20	34	range3 [32.500 - 35.500]	Sim

Figure 5.43 Discretization of the “Idade_M” attribute by frequency

To better understand the data in each range of age values, the below Table 5.59 shows the number of patients covered in each range of values under the column named “n”, the frequencies of the “Gravidez” attribute values under the columns with the prefix “Gravidez”, as well as the means and standard deviations of the ages in each range of values under the column named “Mean(±SD)”.

Table 5.59 Description of the instances in each range of values

Range	n	“Gravidez” = “Sim”	“Gravidez” = “Não”	Mean (±SD)
Range 1 infinity to 30.5	71	39	32	27.831 (±2.426)
Range 2 30.5 to 32.5	46	28	18	31.500 (±0.506)
Range 3 32.5 to 35.5	54	22	32	33.852 (±0.856)
Range 4 35.5 to infinity	58	18	40	37.810 (±2.259)

By analyzing the information disclosed in Table 5.59, we see that the first range of values encompasses the highest number of patients (i.e. 71 patients – we have 23 women with 30 years old) and that the frequencies of patients with the “Gravidez” attribute equal to “Sim”, is almost the same with the ones that have the “Gravidez” attribute equal to “Não” (i.e. 39 vs 32). In contrast, the other ranges of values have a reasonable and similar number of patients, as well as an interesting difference between the frequencies of the “Gravidez” attribute values. An interesting aspect that we can see in this table is that in the first two ranges (i.e. ages under 32.5) we have more patients that have conceived, and in the last two ranges (i.e. ages above 32.5), the inverse is seen (i.e. less patients have conceived). Hence, we can say that women have conceived more until they were 32 years old.

To assess the interest of this discretization, we have applied the ANOVA test upon the discretized ages (i.e. set “Idade_M Original” grouped by “Idade _M” in the ANOVA operator upon the data that was partly shown in Figure 5.43). The ANOVA test showed that the means of the ages between the generated ranges were significantly different (p<0.001) which makes an interesting discretization to test since we could draw the hypothesis that if there is a difference in the age means, some attributes might influence this difference. Hence, this possibility made us go with the discretization by frequency presented in Table 5.59.

5.5.2.4 K-means application of the patient´s age discretization (Model 5)

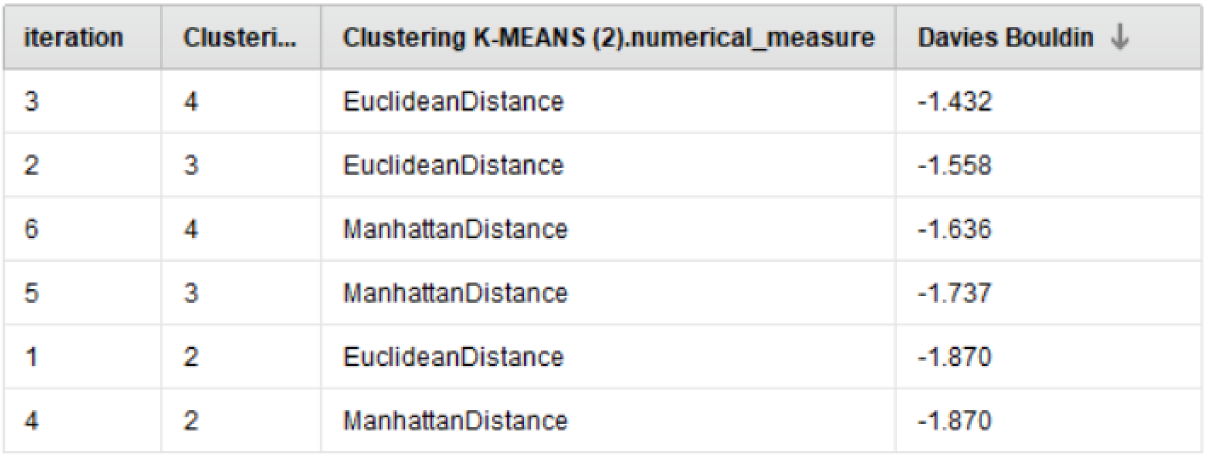
In this subsection, we present the results obtained with the model depicted in Figure 6.26 (Model 5). This model aimed to apply the previously defined discretization of the “Idade_M” attribute upon the most interesting clustering result (i.e. the model with the “Grau_Varicoc” attribute manually dichotomized depicted in Figure 6.21).

This model was executed with the following conditions:

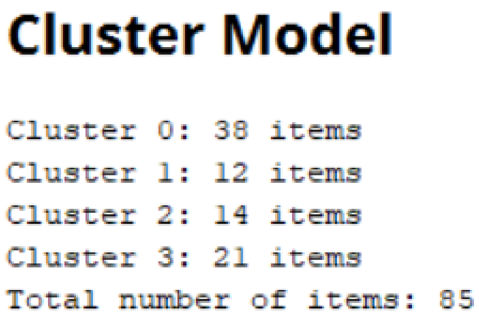
- Selected attributes: Idade_M, A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_I, Grau_II, Grau_III, ProfissãoComRiscoDeContactoDeProdutosOuAmbientes, Gravidez;
- Filtered the above listed attributes by non-missing values; and therefore, all these iterations were running on the same 85 instances;
- Number of clusters: 2 to 4;
- Distance measure: Euclidean and Manhattan Distance.

This model generated the Davies Bouldin results presented in Table 5.60 where we can see that the best setting for this model is: Distance measure = Euclidean Distance and Number Clusters = 4 (Davies Bouldin index= -1.432). The generated clusters have the distribution presented in Figure 5.44 and its means in Table 5.61 and its related series plot, Figure 5.45.

Table 5.60 Davies Bouldin results for the model depicted in Figure 6.26



iteration	Clusteri...	Clustering K-MEANS (2).numerical_measure	Davies Bouldin ↓
3	4	EuclideanDistance	-1.432
2	3	EuclideanDistance	-1.558
6	4	ManhattanDistance	-1.636
5	3	ManhattanDistance	-1.737
1	2	EuclideanDistance	-1.870
4	2	ManhattanDistance	-1.870

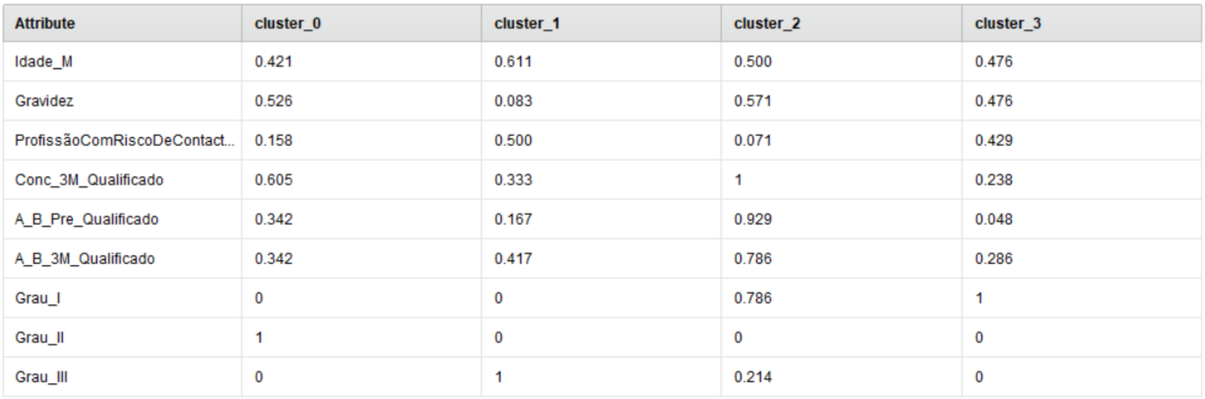


Cluster Model

```
Cluster 0: 38 items
Cluster 1: 12 items
Cluster 2: 14 items
Cluster 3: 21 items
Total number of items: 85
```

Figure 5.44 Cluster´s distribution in the iteration nº3

Table 5.61 Centroid Table for the iteration nº3



Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Idade_M	0.421	0.611	0.500	0.476
Gravidez	0.526	0.083	0.571	0.476
ProfissãoComRiscoDeContact...	0.158	0.500	0.071	0.429
Conc_3M_Qualificado	0.605	0.333	1	0.238
A_B_Pre_Qualificado	0.342	0.167	0.929	0.048
A_B_3M_Qualificado	0.342	0.417	0.786	0.286
Grau_I	0	0	0.786	1
Grau_II	1	0	0	0
Grau_III	0	1	0.214	0

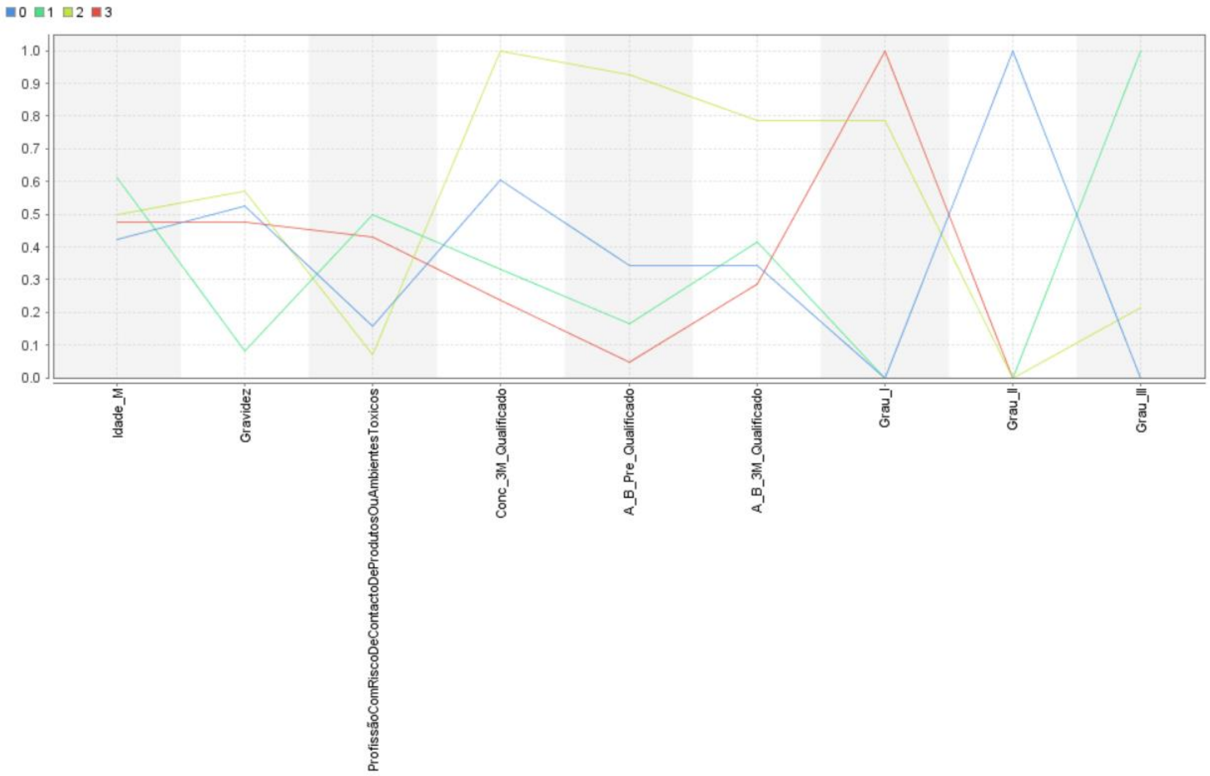


Figure 5.45 Series plot of the centroid means presented in Table 5.61

As we can see, the results generated are exactly the same as previously disclosed in section 5.5.2.1 (all but the mean value of the "Idade_M" attribute – Note that the clusters were renamed in this test), which means that even if the mean is statistically different between the defined age ranges, when partitioned, the clusters generated end up with a mix of woman age ranges that makes it still statistically insignificant (p=0.441). Hence, we can say that the discretization of the "Idade_M" attribute did not improve this model. Moreover, the best Davies Bouldin result is even worse (Model 2: -1.367 vs Model 5: -1.432). However, we have continued to explore its application with the FP-Growth algorithm and its results can be seen in the next subsection.

5.5.3 Association

The association technique was applied with the *FP-Growth* algorithm to identify attribute relations that were interesting and statistically significant with the aim of tackling the goals 2 and 5 described in section 5.2.3. As documented in Appendix C.3, several tests were carried out with the *FP-Growth* algorithm through six modeling steps described in Table 6.6 that aim to identify the most interesting rules.

In order to select the most interesting rules we have at first selected the most objectively and subjectively interesting ones by identifying the rules that complied with the pruning conditions defined in section 4.2.8.3 and at last, computed the C*hi-square* test based on the standard measures of *confidence*, *support* and *lift* of each pruned rule with the formula disclosed in Formula 4.1. In Table 5.62, we disclose all selected/pruned association rules during the first 5 modeling steps of the application of the FP-Growth algorithm, as well as identify the ones that have a statistically significant dependence between the *antecedent* and *consequent* attributes of the rule. The corresponding x^2 value of these rules highlighted below in bold and with one asterisk (*) if the dependency between the attributes was statistically significant for a p<=0.10 (i.e. x^2 >= 2.706), or with two asterisks (**), for a p<=0.05 (i.e. x^2 >= 3.841), or with three asterisks (***), for a p<=0.01 (i.e. x^2 >= 6.635). However, despite all these significance levels, the rules that were in fact elected are the ones that have a p<=0.05. Elected rules are highlighted in yellow in the following result tables.

To specify how the generated results - results disclosed under the column names colored in a stronger gray color - were computed during the first 5 modeling steps by the RapidMiner platform, other information was added to Table 5.62 that indicates the following information:

- "Rule ID" - Rule identification number.
- "Test Nº" - Test number related to the outputs shown in Appendix C.3, that has the following format: [Modeling step Number].[Time Ran]. Hence, for example, the test 1.2, references the results obtained during the second run of the association modelling step 1. Note that the tests 1.x to 2.x were computed with the model disclosed in Figure 6.27 FP-Growth model 1 (used in step 1 and 2) and the tests 3.x to 5.x, were computed with the model disclosed in Figure 6.28 FP-Growth model 2 (used in step 3 to 5).
- "Gravidez Filtered" - Indicates whether the "Gravidez" attribute was filtered by non-missing values.
- "Support threshold" and "Confidence threshold" - Indicates the *support* and *confidence* thresholds set in the corresponding test.
- "x^2" - Chi-square value computed with the EXCEL software.
- "n" - Number of instances encompassed in the corresponding test.

Table 5.62 Selected results in the first 5 modeling steps

Rule ID	Test Nº	"Gravidez" Filtered	Support threshold	Confidence threshold	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2	n
1	1.1	No	0.1	0.8	Grau_Varicoc=I	A_B_Pre	0.198	0.866	1.153	1.855	**6.13444398 ****	293
2	1.2	No	0.0	0.0	A_B_Pre	Formas_N_3M	0.468	0.623	1.14	1.203	**20.894662 *****	293
3	1.2	No	0.0	0.0	A_B_Pre	Gravidez	0.317	0.423	1.158	1.1	**12.58891263 *****	293
4	1.2	No	0.0	0.0	Formas_N_3M	Gravidez	0.253	0.463	1.266	1.181	**14.40116299 *****	293
5	1.2	No	0.0	0.0	A_B_Pre, Formas_N_3M	Gravidez	0.229	0.489	1.339	1.242	**17.06157362 *****	293
6	2.1	No	0.0	0.0	A_B_Pre	Formas_N_3M, Gravidez	0.229	0.305	1.206	1.075	**12.68232854 *****	293
7	2.1	No	0.0	0.0	Conc_6M	Gravidez	0.167	0.422	1.157	1.099	**2.715615806 ***	293
8	2.1	No	0.0	0.0	Grau_Varicoc=II	Gravidez	0.164	0.432	1.184	1.118	**3.487211005 ***	293
9	3.2	Yes	0.0	0.0	A_B_Pre	Formas_N_3M	0.517	0.654	1.106	1.181	**14.11070966 *****	230
10	3.2	Yes	0.0	0.0	Formas_N_3M	Gravidez	0.322	0.544	1.17	1.173	**8.378246668 *****	230
11	3.2	Yes	0.0	0.0	A_B_Pre	Formas_N_3M, Gravidez	0.291	0.368	infinity	1.583	**Infinity *****	230
12	3.2	Yes	0.0	0.0	A_B_Pre, Formas_N_3M	Gravidez	0.291	0.563	1.21	1.224	**9.442665918 *****	230
13	3.3	Yes	0.1	0.4	Grau_Varicoc=II	Formas_N_3M	0.222	0.593	1.003	1.004	0.001791514	230
14	3.3	Yes	0.1	0.4	Grau_Varicoc=II	A_B_Pre,Formas_N_3M	0.196	0.523	1.011	1.012	0.01787736	230
15	3.3	Yes	0.1	0.4	A_B_Pre,Grau_Varicoc=II	Formas_N_3M	0.196	0.662	1.119	1.208	1.984420491	230
16	3.3	Yes	0.1	0.4	Grau_Varicoc=I	Formas_N_3M	0.178	0.683	1.156	1.291	**2.848821379 ***	230
17	3.3	Yes	0.1	0.4	Grau_Varicoc=I	A_B_Pre, Formas_N_3M	0.152	0.583	1.127	1.158	1.402075017	230
18	3.3	Yes	0.1	0.4	A_B_Pre,Grau_Varicoc=I	Formas_N_3M	0.152	0.686	1.161	1.303	2.450825696	230
19	3.4	Yes	0.1	0.4	Grau_Varicoc=II	Gravidez	0.209	0.558	1.2	1.21	**4.78859224 ****	230
20	3.4	Yes	0.1	0.4	Grau_Varicoc=II	A_B_Pre, Gravidez	0.174	0.465	1.15	1.114	2.100530514	230
21	3.4	Yes	0.1	0.4	A_B_Pre, Grau_Varicoc=II	Gravidez	0.174	0.588	1.264	1.299	**5.860237633 ****	230
22	4.1	Yes	0.1	0.4	Conc_3M_Qualificado	Gravidez	0.235	0.574	1.235	1.257	**7.646138449 *****	230
23	4.1	Yes	0.1	0.4	A_B_3M_Qualificado	Gravidez	0.217	0.568	1.221	1.238	**6.040879256 ****	230
24	4.1	Yes	0.1	0.4	A_B_Pre_Qualificado	Gravidez	0.187	0.558	1.2	1.211	**4.03046074 ****	230

Rule ID	Test Nº	"Gravidez" Filtered	Support threshold	Confidence threshold	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2	n
25	4.1	Yes	0.1	0.4	Conc_3M_Qualificado, A_B_3M_Qualificado	Gravidez	0.157	0.667	1.433	1.604	**11.55926365 *****	230
26	4.2	Yes	0.1	0.4	A_B_3M_Qualificado	Conc_3M_Qualificado	0.235	0.614	1.501	1.53	**24.7785958 *****	230
27	4.2	Yes	0.1	0.4	A_B_Pre_Qualificado	Conc_3M_Qualificado	0.178	0.532	1.303	1.265	**7.326340305 *****	230
28	4.2	Yes	0.1	0.4	A_B_3M_Qualificado	Gravidez, Conc_3M_Qualificado	0.157	0.409	1.742	1.295	**24.20627243 *****	230
29	4.3	Yes	0.1	0.4	Conc_3M_Qualificado	A_B_3M_Qualificado	0.235	0.574	1.501	1.451	**24.78013169 *****	230
30	4.3	Yes	0.1	0.4	A_B_Pre_Qualificado	A_B_3M_Qualificado	0.165	0.494	1.29	1.219	**6.020405714 ****	230
31	5.1	Yes	0.1	0.4	Qualificar_Espermograma_3M = Normozoospérmico	Gravidez	0.104	0.706	1.517	1.818	**9.24548327 *****	230

As previously seen, the FP-Growth algorithm considers the attribute values set to "True" to compute the most frequent item sets; and therefore, to interpret the above generated association rules we must keep in mind the attribute values that were considered as "True" in the above ran tests. Table 5.63 presents these values: the attributes specified with the value *All*, under the column named "True Value", indicates that all non-missing values of the corresponding attribute are considered as "True" by the algorithm since these attributes are polynominal attributes that were dichotomized with the "Nominal to Binomial" RapidMiner´s operator - as seen in the model depicted in Figure 6.27 FP-Growth model 1 and Figure 6.28 FP-Growth model 2.

Table 5.63 Attribute´s True Values in the first 5 modeling steps

Attribute Name	True Value
Grau_Varicoc	*All*
Conc_6M	Conc_6M>0
A_B_pré	A_B_pré>0
Formas_N_3M	Formas_N_3M>0
Conc_3M_Qualificado	Normal
A_B_Pre_Qualificado	Normal
A_B_3M_Qualificado	Normal
Qualificar_Espermograma_Pre	*All*
Qualificar_Espermograma_3M	*All*
ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	Sim
Gravidez	Sim

Upon the set of the most interesting rules (i.e. the pruned rules with a statistically significant dependence between the *antecedent* and *consequent* attributes that are in shown in Table 5.62), we have elected the best rules. Those best rules are highlighted in yellow in Table 5.62 and can be interpreted as follows in the context of its main measures:

- The rule with the highest statistical significance (i.e. *support*) is:

  A_B_Pre -> Formas_N_3M (for n=230)

  This rule tells us that the conditional probability (i.e. confidence) of observing 3 months after the embolization treatment a normal sperm morphology given a normal sperm motility before the embolization treatment is of 65% (i.e. 119/182). In the dataset, we have 119 male patients out of 230 that have before the treatment a normal sperm motility and 3 months after the embolization treatment, a normal sperm morphology, which gives a *support* of 0.198. This rule is interesting since there is a strong relation between these two sperm parameters (p<0.01).

- The rule that showed the best strength (i.e. *confidence*) was:

  Grau_Varicoc=1 -> A_B_Pre (for n=293)

  This rule tells us that the conditional probability of observing normal sperm motility before the embolization treatment given a low severity grade of the varicocele condition is of 87% (i.e. 58/67). In the dataset, we have 58 male patients out of 293 that have before the treatment a low severity grade of the varicocele condition and a sperm motility above 0, which gives a *support* of 0.198. This rule is interesting since there is a relation between the severity grade and the sperm motility before the treatment ($p<0.05$).

- The rule that computed the highest dependency between its *antecedent* and *consequent* attributes (i.e. x^2) was:
  A_B_Pre -> Formas_N_3M, Gravidez (for n=230)

  This rule tells us that the conditional probability of observing a normal sperm morphology 3 months after the embolization treatment and getting pregnant (i.e. “Gravidez”=”Sim”) given a normal sperm motility before the treatment is 37% (i.e. 67/182). In the dataset, we have 67 male patients out of 230 that were able to conceive with a normal sperm motility before the treatment and a normal sperm morphology 3 months after the treatment which gives a *support* of 0.291. This rule is interesting since there is a very strong relation between the antecedent and consequent attributes of the rule ($p<0.01$).

- The rule that showed the highest statistical significance (i.e. *support*) for the *consequent* attribute “Gravidez” was:
  Formas_N_3M -> Gravidez (for n=230)

  This rule tells us that the conditional probability of a woman conceiving (i.e. “Gravidez”=”Sim”) given a partner with normal sperm morphology 3 months after the treatment is of 54% (i.e. 74/136). In the dataset, we have 74 male patients out of 230 that could impregnate their partner with normal sperm morphology 3 months after the treatment above 0 which gives a *support* of 0.322. This rule is interesting since there is a strong relation between the antecedent and consequent attributes of the rule ($p<0.01$) and through all the FP-Growth application, it was seen the rule with the highest support towards the “Gravidez” consequent.

- The top 3 rules that showed the best strength (i.e. *confidence*) for the *consequent* attribute “Gravidez” was, in the order of appearance:
  Qualificar_Espermograma_3M = Normozoospérmico -> Gravidez (for n=230)

  This rule tells us that the conditional probability of a woman conceiving (i.e. “Gravidez”=”Sim”) given a Normozoospermic partner at 3 months after the treatment is of 71% (i.e.24/34). In the dataset, we have 24 male patients out of 230 that impregnated their partner by being Normozoospermic at 3 months after the treatment which gives a *support* of 0.104. This rule is interesting since there is a strong relation between the antecedent and consequent attributes of the rule ($p<0.01$).

  Conc_3M_Qualificado, A_B_3M_Qualificado -> Gravidez (for n=230)

  This rule tells us that the conditional probability of a woman conceiving (i.e. “Gravidez”=”Sim”) given a partner with a Normal sperm concentration and sperm motility at 3 months after the treatment is of 67% (i.e. 36/54). In the dataset, we have 36 male patients out of 230 that impregnated their partner by having these sperm parameters above or equal to the WHO thresholds. This rule is interesting since there is a strong relation between the antecedent and consequent attributes of the rule ($p<0.01$).

> **A_B_Pre, Grau_Varicoc=II -> Gravidez** (for n=230)
>
> This rule tells us that the conditional probability of a woman conceiving (i.e. “Gravidez”= “Sim”) given a partner with sperm motility before the treatment above 0 and a moderate varicocele´severity grade is of 59% (40/68). In the dataset, we have 40 male patients out of 230, that impregnated their partner with normal sperm motility and a moderate severity grade of varicocele before the treatment. These last attributes are related with the pregnancy outcome since it has computed a p<0.05.

If we analyze the assessed “True” attribute values disclosed in Table 5.63 above, we see that some attributes values were not mined: the abnormal sperm parameter values and other ranges of sperm parameters values. Since one of the aims of this study is to find data patterns, essentially between sperm related attributes and external factors (i.e. Goal 5), we have applied the FP-Growth algorithm on the sperm parameters covered in the fourth, fifth and sixth group of attributes disclosed in section 5.4.4 with other attributes related with patient external factors to seek for interesting association rules. Furthermore, we have also looked for rules with the consequent attribute “Gravidez” to tackle the prediction of embolization success (i.e. Goal 2). Hence, in this last six modeling steps we have discretized the sperm parameters and have added other external factors (i.e. smoking and drinking habits) that were not seen in this study as statistically significant with the “Gravidez” attribute, nor with a good data quality (see section 5.4.4), but yet has been studied in related works (Delavar et al., 2014). Nevertheless, we have assessed the statistical significance between these external factors and the semen classification but not with the sperm parameters, so we have also aimed to do it through this sixth modeling step. Furthermore, we have added the “PMA” and “Gravidez_espontanea” attribute to see if there were attribute values related with these types of conceptions and have added the age of the woman discretized as defined in section 5.5.2.3, but this time, with the “Discretized by User Specification” operator to ensure that the previously defined ranges were retested.

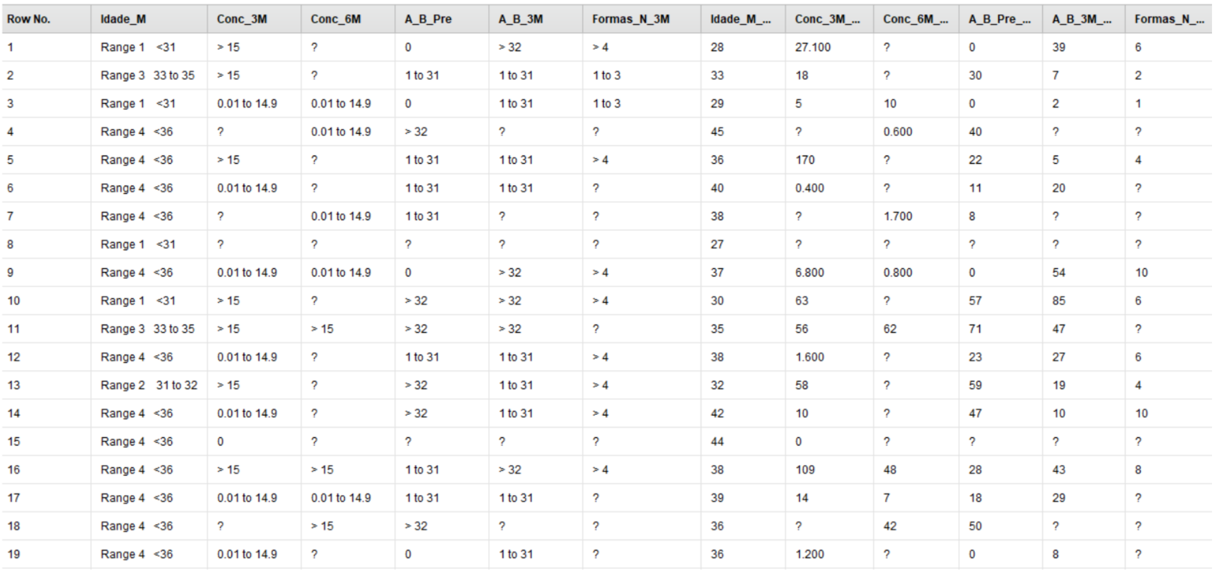
Sperm parameter discretization was carried out in the sixth modeling step similarly as with the “Idade_M” attribute that was discussed in section 5.5.2.3. Hence, we have at first tested the RapidMiner´s operator called “Discretization by Entropy” where we have seen that the proposed ranges of values, for all assessed sperm parameters (i.e. Conc_3M , Conc_6M, A_B_pré, A_B_3M, Formas_N_3M), were from negative infinity to positive infinity and then have tested the “Discretization by frequency” operator for 4 bins. When we analyzed the proposed ranges of values of this last operator, we have realized that the best way to discretize these values was to use a user specified discretization with the operator called “Discretized by User Specification” to better interpret the patients that have normal or abnormal sperm parameter values. In fact, with the “Discetization by Frequency” operator, for, for example the “Conc_3M” attribute, the second proposed range went from 1.75 to 18.5 million sperm per milliliter; and therefore, this range covered patients that had normal and abnormal sperm concentration values since the WHO threshold is set to 15 million/milliliter. To facilitate the interpretation of the results, we have discretized the sperm parameters by the WHO thresholds described in section 4.1.1.1. Therefore, we have ended up with the discretization disclosed in Table 5.64 that can be interpreted as follows: under the column named “Attribute Name”, we have the name of the numeric attributes that were discretized for this test; under the column named “Discretization”, we have the proposed ranges of values; under the column named “n”, the number of instances that falls in each of these specified ranges of values with non-missing values in the “Gravidez” attribute and in the following two columns, we have the number of instances that have the “Gravidez” attribute filled with the value Yes (“Sim”) or No (“Não”) for the corresponding attribute and range of values.

This proposed discretization encompasses patients with the value 0, as well as patients with abnormal sperm parameter values, which makes us consider information that was not yet assessed in this study with the FP-Growth algorithm. If we analyze the number of instances that falls into abnormal sperm parameter values, we see that this data set as more patients with abnormal sperm parameters than normal sperm parameters in the selected attributes which supports our will to also assess that population in this last modeling step.

Table 5.64 Attribute discretization - step 6 – test 6.1

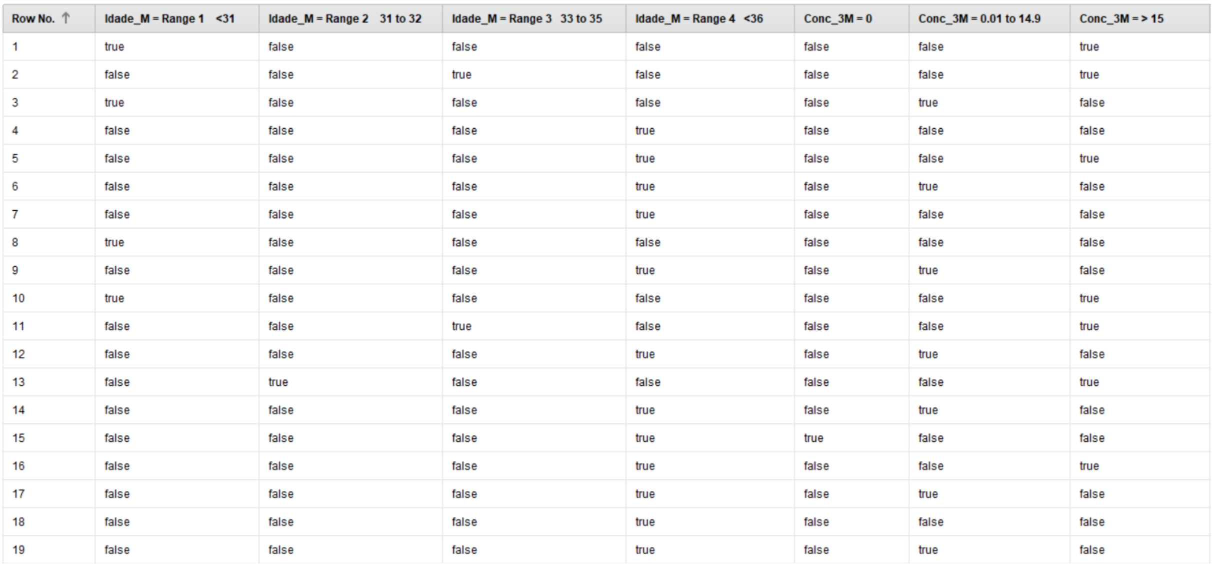
Attribute Name	Discretization	n	Gravidez (Sim)	Gravidez (Não)
Conc_3M	0	17	2	15
	0.01 to 14.9	90	38	52
	15 to positive infinity	94	54	40
Conc_6M	0	9	1	8
	0.01 to 14.9	60	25	35
	15 to positive infinity	46	24	22
A_B_pré	0	27	18	9
	1 to 31	105	50	55
	32 to positive infinity	77	43	34
A_B_3M	0	18	7	11
	1 to 31	78	35	43
	32 to positive infinity	88	50	38
Formas_N_3M	0	16	4	12
	1 to 3	61	32	29
	4 to positive infinity	75	42	33
Idade_M	Infinity to 30.500	71	39	32
	30.500 to 32.500	46	28	18
	32.500 to 35.500	54	22	32
	35.500 to infinity	58	18	40

In Figure 5.46, we can see the discretization performed by the RapidMiner platform with the “Discretized by User Specification” operator. As we can see, at left, we have all discretized attributes and at right, its original values that enabled the validation of the performed discretization. In the following Figure 5.47, we can see the “Idade_M” and the “Conc_3_M” attribute that were subsequently dichotomized with the “Nominal to Binomial” operator. The FP_Growth algorithm has then run upon this type of dichotomized data.



Row No.	Idade_M	Conc_3M	Conc_6M	A_B_Pre	A_B_3M	Formas_N_3M	Idade_M_...	Conc_3M_...	Conc_6M_...	A_B_Pre_...	A_B_3M_...	Formas_N_...
1	Range 1 <31	> 15	?	0	> 32	> 4	28	27.100	?	0	39	6
2	Range 3 33 to 35	> 15	?	1 to 31	1 to 31	1 to 3	33	18	?	30	7	2
3	Range 1 <31	0.01 to 14.9	0.01 to 14.9	0	1 to 31	1 to 3	29	5	10	0	2	1
4	Range 4 <36	?	0.01 to 14.9	> 32	?	?	45	?	0.600	40	?	?
5	Range 4 <36	> 15	?	1 to 31	1 to 31	> 4	36	170	?	22	5	4
6	Range 4 <36	0.01 to 14.9	?	1 to 31	1 to 31	?	40	0.400	?	11	20	?
7	Range 4 <36	?	0.01 to 14.9	1 to 31	?	?	38	?	1.700	8	?	?
8	Range 1 <31	?	?	?	?	?	27	?	?	?	?	?
9	Range 4 <36	0.01 to 14.9	0.01 to 14.9	0	> 32	> 4	37	6.800	0.800	0	54	10
10	Range 1 <31	> 15	?	> 32	> 32	> 4	30	63	?	57	85	6
11	Range 3 33 to 35	> 15	> 15	> 32	> 32	?	35	56	62	71	47	?
12	Range 4 <36	0.01 to 14.9	?	1 to 31	1 to 31	> 4	38	1.600	?	23	27	6
13	Range 2 31 to 32	> 15	?	> 32	1 to 31	> 4	32	58	?	59	19	4
14	Range 4 <36	0.01 to 14.9	?	> 32	1 to 31	> 4	42	10	?	47	10	10
15	Range 4 <36	0	?	?	?	?	44	0	?	?	?	?
16	Range 4 <36	> 15	> 15	1 to 31	> 32	> 4	38	109	48	28	43	8
17	Range 4 <36	0.01 to 14.9	0.01 to 14.9	1 to 31	1 to 31	?	39	14	7	18	29	?
18	Range 4 <36	?	> 15	> 32	?	?	36	?	42	50	?	?
19	Range 4 <36	0.01 to 14.9	?	0	1 to 31	?	36	1.200	?	0	8	?

Figure 5.46 Discretized Attributes – step 6 – test 6.1



Row No. ↑	Idade_M = Range 1 <31	Idade_M = Range 2 31 to 32	Idade_M = Range 3 33 to 35	Idade_M = Range 4 <36	Conc_3M = 0	Conc_3M = 0.01 to 14.9	Conc_3M = > 15
1	true	false	false	false	false	false	true
2	false	false	true	false	false	false	true
3	true	false	false	false	false	true	false
4	false	false	false	true	false	false	false
5	false	false	false	true	false	false	true
6	false	false	false	true	false	true	false
7	false	false	false	true	false	false	false
8	true	false	false	false	false	false	false
9	false	false	false	true	false	true	false
10	true	false	false	false	false	false	true
11	false	false	true	false	false	false	true
12	false	false	false	true	false	true	false
13	false	true	false	false	false	false	true
14	false	false	false	true	false	true	false
15	false	false	false	true	true	false	false
16	false	false	false	true	false	false	true
17	false	false	false	true	false	true	false
18	false	false	false	true	false	false	false
19	false	false	false	true	false	true	false

Figure 5.47 Dichotomized Attributes - step 6 – test 6.1

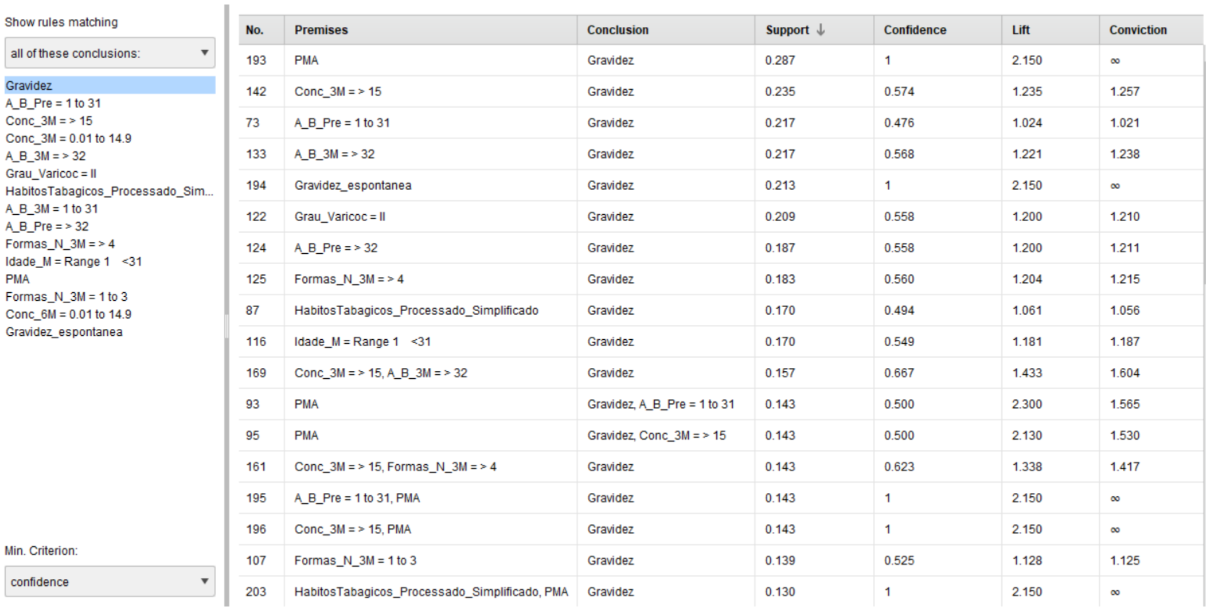
In orther to better convey the generated results with these discretized and dichotomized attributes, in the following subsections we present the computed results during the sixth modeling step of the FP-Growth algorithm grouped by its performed tests.

5.5.3.1 Generated results – Step 6 – Test 6.1 (Model 3)

In the first test of the sixth modeling step of the FP-Growth algorithm, we have applied the model depicted in Figure 6.29 with the following selected attributes:

Idade_M, Grau_Varicoc, Conc_3M , Conc_6M, A_B_pré, A_B_3M, Formas_N_3M,
ProfissãoComRiscoDeContactoDeProdutosOuAmbientes,
HabitosTabagicos_Processado_Simplificado,
HabitosAlcoolicos_Processado_Simplificado,
Gravidez, PMA, Gravidez_espontanea.

A print screen of part of the generated association rules for the consequent attribute “Gravidez” is shown in Figure 5.48 - all generated rules are in Appendix C.3.6.



Show rules matching

all of these conclusions: ▼

Gravidez
A_B_Pre = 1 to 31
Conc_3M = > 15
Conc_3M = 0.01 to 14.9
A_B_3M = > 32
Grau_Varicoc = II
HabitosTabagicos_Processado_Sim...
A_B_3M = 1 to 31
A_B_Pre = > 32
Formas_N_3M = > 4
Idade_M = Range 1 <31
PMA
Formas_N_3M = 1 to 3
Conc_6M = 0.01 to 14.9
Gravidez_espontanea

Min. Criterion:

confidence ▼

No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction
193	PMA	Gravidez	0.287	1	2.150	∞
142	Conc_3M = > 15	Gravidez	0.235	0.574	1.235	1.257
73	A_B_Pre = 1 to 31	Gravidez	0.217	0.476	1.024	1.021
133	A_B_3M = > 32	Gravidez	0.217	0.568	1.221	1.238
194	Gravidez_espontanea	Gravidez	0.213	1	2.150	∞
122	Grau_Varicoc = II	Gravidez	0.209	0.558	1.200	1.210
124	A_B_Pre = > 32	Gravidez	0.187	0.558	1.200	1.211
125	Formas_N_3M = > 4	Gravidez	0.183	0.560	1.204	1.215
87	HabitosTabagicos_Processado_Simplificado	Gravidez	0.170	0.494	1.061	1.056
116	Idade_M = Range 1 <31	Gravidez	0.170	0.549	1.181	1.187
169	Conc_3M = > 15, A_B_3M = > 32	Gravidez	0.157	0.667	1.433	1.604
93	PMA	Gravidez, A_B_Pre = 1 to 31	0.143	0.500	2.300	1.565
95	PMA	Gravidez, Conc_3M = > 15	0.143	0.500	2.130	1.530
161	Conc_3M = > 15, Formas_N_3M = > 4	Gravidez	0.143	0.623	1.338	1.417
195	A_B_Pre = 1 to 31, PMA	Gravidez	0.143	1	2.150	∞
196	Conc_3M = > 15, PMA	Gravidez	0.143	1	2.150	∞
107	Formas_N_3M = 1 to 3	Gravidez	0.139	0.525	1.128	1.125
203	HabitosTabagicos_Processado_Simplificado, PMA	Gravidez	0.130	1	2.150	∞

Figure 5.48 Association rules for the 4th and 5th group of attributes – step 6 - test 6.1

As we can see, this step has generated a wide variety of association rules (208 association rules) mainly lead by the several dichotomized attributes that we have ended up with (i.e. 28 attributes). Hence, we have focused on the identification of the association rules with the highest statistical significance (i.e. *support*) and/or strength (i.e. *confidence*) for the several patient aspects we aimed to assess through this test: pregnancy outcome, type of conception, abnormal sperm parameters, external factors and woman age range. Thus, Table 5.65 presents these rules with its corresponding x^2 value. Note that all these association rules were generated upon the data set filtered by non-missing values in the “Gravidez” attribute (i.e. n=230) and for the following thresholds: *support*=0.1 and *confidence*=0.4.

Table 5.65 Selected Results – step 6 - test 6.1

Rule ID	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
1	PMA	Gravidez	0.287	1	2.15	∞	**106.4677419** ***
2	Gravidez_espontanea	Gravidez	0.213	1	2.15	∞	**71.5864** ***
3	A_B_Pre = 1 to 31	A_B_3M = 1 to 31	0.209	0.457	1.348	1.217	**12.03980423** ***
4	A_B_3M = 1 to 31	Conc_3M = 0.01 to 14.9	0.178	0.526	1.343	1.283	**8.9108785** ***
5	Formas_N_3M = 1 to 3	Conc_3M = > 15	0.170	0.639	1.564	1.64	**18.31982** ***
6	Idade_M = Range 1 <31	Gravidez	0.170	0.549	1.181	1.187	**2.935959** *
7	Conc_3M = 0.01 to 14.9	Conc_6M = 0.01 to 14.9	0.157	0.4	1.533	1.232	**14.90409858** ***
8	HabitosTabagicos_Processado_Simplificado	A_B_3M = > 32	0.152	0.443	1.158	1.108	1.858191
9	Idade_M = Range 1 <31	HabitosTabagicos_Processado_Simplificado	0.143	0.465	1.353	1.227	**6.664957734** ***
10	A_B_Pre = 1 to 31, PMA	Gravidez	0.143	1	2.15	∞	**44.13477246** ***
11	Conc_3M = > 15, PMA	Gravidez	0.143	1	2.15	∞	**44.13477246** ***

Rule ID	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
12	HabitosTabagicos_Processado_Simplificado, PMA	Gravidez	0.130	1	2.15	∞	**39.52298851 *****
13	A_B_Pre = 1 to 31, Conc_3M = > 15	Gravidez	0.122	0.596	1.281	1.323	**4.067032595 ****
14	Conc_3M = > 15, Gravidez_espontanea	Gravidez	0.122	1	2.15	∞	**36.75284738 *****
15	Grau_Varicoc = II, Gravidez_espontanea	Gravidez	0.117	1	2.15	∞	**35.04699887 *****
16	Idade_M = Range 1 <31, PMA	Gravidez	0.117	1	2.15	∞	**35.04699887 *****
17	Conc_3M = 0.01 to 14.9, PMA	Gravidez	0.113	1	2.15	∞	**33.69616685 *****
18	Grau_Varicoc = II, PMA	Gravidez	0.113	1	2.15	∞	**33.69616685 *****
19	Formas_N_3M = 1 to 3	HabitosTabagicos_Processado_Simplificado	0.113	0.426	1.241	1.144	2.520855881
20	A_B_3M = 1 to 31, PMA	Gravidez	0.109	1	2.15	∞	**32.35746352 *****
21	PMA, Formas_N_3M = 1 to 3	Gravidez	0.109	1	2.15	∞	**32.35746352 *****
22	Formas_N_3M = > 4, Gravidez_espontanea	Gravidez	0.104	1	2.15	∞	**30.70089286 *****
23	A_B_Pre = 1 to 31, Grau_Varicoc = II	Gravidez	0.100	0.622	1.336	1.413	**4.333392833 ****
24	A_B_3M = > 32, PMA	Gravidez	0,122	1	2.15	∞	**36,75284738 *****
25	A_B_Pre = > 32, PMA	Gravidez	0,113	1	2.15	∞	**33,69616685 *****
26	A_B_3M = > 32, Gravidez_espontanea	Gravidez	0,122	1	2.15	∞	**36,75284738 *****

The first association rule that is presented in the Table 5.65 (i.e. PMA -> Gravidez), as well as the second one, are not subjectively interesting rules despite being the ones with the highest *support* and/or *confidence* value in this test. In fact, if we consider the first rule, the "PMA" attribute is highly related (x^2 = 106,468**)** with the "Gravidez" attribute because it specifies the pregnancy´s type of conception (i.e. whether the recorded pregnancy was achieved with an ART procedure or not). Hence, the confidence value equal to 1 does not tell that all performed ART procedures were successful. Thus, the only information that we can extract from these types of rules is that 66 patients out of 230 (i.e. 28%) conceived with an ART procedure carried out at CHUC and that 49 patients conceived spontaneously which tells us that **107 couples out of 230 were able to conceive** and that we have more patients in the data set that conceived with an ART procedures than spontaneously (i.e. we have 17 more patients that conceived with an ART procedure). Note that these 107 couples had a total of 115 pregnancies because 8 out of them got pregnant twice: 5 couples got pregnant twice and at last, had 2 live babies; 2 other couples, got pregnant twice, but only completed one of the pregnancies with a live baby and the last couple had also two pregnancies but we do not have information on outcome.

The following rules are the most subjectively and objectively interesting rules that were identified in this test (i.e. rule 3,4,5 and 7 that are highlighted in yellow in Table 5.65) and can be interpreted as follows:

- **A_B_Pre = 1 to 31 -> A_B_3M = 1 to 31** (Rule ID =3)

The conditional probability of having an abnormal sperm motility going from 1% to 31% three months after the embolization treatment having had the same result before the treatment is of 45.7% (i.e. *confidence*=0.457). In fact, we have 48 patients out of 230 that have abnormal sperm motility before and after the treatment over 105 patients that have abnormal sperm motility before the treatment (i.e. 48/105=0.457).

- **A_B_3M = 1 to 31 -> Conc_3M = 0.01 to 14.9** (Rule ID =4)
  The conditional probability of having an abnormal sperm concentration 3 months after the embolization treatment that goes from 0.01 to 14.9 million/milliliter given an abnormal sperm motility that goes from 1% to 31% at the same time is of 52.6% (i.e. *confidence* = 0.526). In fact, we have 41 patients that had abnormal sperm motility and an abnormal sperm concentration 3 months after the treatment in these ranges of values over 78 patients that had an abnormal sperm motility at the same time (i.e. 41/78=0.526).
- **Formas_N_3M = 1 to 3 -> Conc_3M = > 15** (Rule ID =5)
  The conditional probability of having a normal sperm concentration 3 months after the treatment given at the same time an abnormal sperm morphology that goes from 1% to 3% is of 63.9% (i.e. *confidence* = 0.639 which is 39/61). In fact, we have 39 patients that had an abnormal sperm morphology 3 months after the treatment that goes from 1% to 3% and a normal sperm concentration over 61 patients that had an abnormal sperm morphology from 1% to 3% at the same time (i.e. 39/61=0.639).
- **Conc_3M = 0.01 to 14.9 -> Conc_6M = 0.01 to 14.9** (Rule ID =7)
  The conditional probability of having an abnormal sperm concentration going from 0.01 to 14.9 million/milliliter 6 months after the treatment given an abnormal sperm concentration in that same range of values 3 months earlier is of 40% (i.e. 36 patients with abnormal sperm concentration at 3 and 6 months from 0.01 to 14.9 million per milliliter over 90 patients with abnormal sperm concentrations in that same range of values 3 months earlier).

Through the identification of the rules listed above, we have seen that the rules with the highest *support* were mostly related with normal sperm parameter values and/or rules that were already showcased in Table 5.62; and therefore, these rules were not redisclosed in Table 5.65. However, the maximum *support* found through this test was the value 0.287 - related with the first association rule disclosed in the above Table 5.65 – which tells us that this test did not generated rules that encompassed a wide number of patients.

Even though the rules 9 to 26 were only objectively interesting, due to low *support* value (i.e. below 0.15), we have decided to include them anyway in Table 5.65 because they encompassed the subjects that we aim to assess. In fact, these last rules mainly characterize the pregnancies carried-out with ART procedures and/or spontaneously among the 230 patients that were able to conceive or not. Hence, through their *support* value, we can see the percentage of patients that encompasses each of these displayed “if then” conditions which lead us to the following statements:

- 14.3% of the woman that got pregnant with an ART procedure had a partner that had abnormal sperm motility before the embolization treatment that went from 1% to 31% (i.e. we have 33 instances that has the “Gravidez” attribute set to TRUE, the “PMA” attribute set to TRUE and the "A_B_Pre = 1 to 31" set to TRUE among the 230 instances assessed in the test 1.1 which gives 33/230=14.3%. Analogously, the following rules must be interpreted the same way) (Rule ID = 10);

- 14.3% of woman that conceived with an ART procedure had a partner that had a normal sperm concentration 3 months after the treatment (Rule ID = 11);
- 13% of the women that conceived with an ART procedure had a partner that smoked (Rule ID = 12);
- 12.2% of the woman that conceived spontaneously had a partner with a normal sperm concentration 3 months after the embolization treatment (Rule ID = 14);
- 11.7% of the woman that conceived spontaneously had a partner with a moderate varicocele (severity grade II) (Rule ID = 15);
- 11.3% of the woman that conceived with an ART procedure had a partner with an abnormal sperm concentration that went from 0.01 to 14.9 million/milliliter 3 months after the embolization treatment (Rule ID = 17);
- 11.3% of the woman that conceived with an ART procedure had a partner with a moderate varicocele (severity grade II) (Rule ID = 18);
- 10.9% of the woman that conceived with an ART procedure had a partner with an abnormal sperm motility that went from 1% to 31% 3 months after the embolization treatment (Rule ID = 20);
- 10.9% of the woman that conceived with an ART procedure had a partner with an abnormal sperm morphology three months after the embolization treatment that went from 1% to 3% (Rule ID = 21);
- 10.4% of the woman that conceived spontaneously had a normal sperm morphology 3 months after the embolization treatment (Rule ID =22).

Regarding the above interpreted association rules, we can see that there is a pattern between them and the rules identified with the id 24, 25 and 26 in the above Table 5.65. This pattern is disclosed below by its related rules and context:

- Rule 10 vs 25 – ART procedure ("PMA") vs Sperm motility **before treatment**
  A slightly higher percentage of patients with an abnormal sperm motility value before the treatment has conceived with an ART procedure. In fact, 14.3% had an abnormal sperm motility value different than 0 (rule 10) vs 11,3%, had a normal sperm motility value (rule 25);
- Rule 24 vs 20 – ART procedure ("PMA") vs Sperm motility **3 months** after the treatment
  A slightly higher percentage of patients with a normal sperm motility value 3 months after the embolization treatment conceived with an ART procedure. In fact, 12.2% had a normal sperm motility value (rule 24) vs 10,9%, had an abnormal sperm motility value different than 0 (rule 20);
- Rule 14, 15, 22 and 26 – Spontaneous conception ("Gravidez_espontanea")
  The generated objectively interesting rules have shown that the biggest groups of couples identified (i.e. support > 0.104) that conceived spontaneously all had at least

one of the sperm parameter values categorized as normal 3 months after the embolization treatment (i.e. Rule 14, 22 and 26 encompasses sperm parameter values that are normal by the WHO thresholds), as well as a moderate varicocele condition (Rule 15).

Since we could not assess the association measures for the rules that covered the types of conceptions (with exception of the *support* measure), in Table 5.66 below, we present the generated results for the "PMA" and "Gravidez_espontanea" attribute set as consequent during test 6.1. As we can see, the *support* values of these rules are the same as disclosed in Table 5.65 but we have here unbiased *confidence*, *lift*, *conviction* and x^2 values; and hence, these values were the ones that were considered for further analyses on the matter of the types of conceptions among the 230 couples that achieved pregnancy or not. In the Table 5.67, we have these same rules but only generated on the couples that were successful (i.e. only for instances with "Gravidez"="Sim"). In Appendix C.3.6, we disclose all these results but in Table 5.66 and in Table 5.67 we only present the objectively interesting rules.

Table 5.66 Selected Results – step 6 - test 6.1 - Types of Conceptions for n=230

Related Rule ID	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x2
1	Gravidez	PMA	0.287	0.617	2.15	1.861	**106.4718*****
2	Gravidez	Gravidez_espontanea	0.213	0.458	2.15	1.452	**71.58172*****
10	Gravidez, A_B_Pre = 1 to 31	PMA	0.143	0.66	2.3	2.097	**43.26733*****
11	Gravidez, Conc_3M = > 15	PMA	0.143	0.611	2.13	1.834	**36.09594*****
12	Gravidez, HabitosTabagicos _Simplificado	PMA	0.13	0.769	2.681	3.09	**53.1795*****
14	Gravidez, Conc_3M = > 15	Gravidez_espontanea	0.122	0.519	2.434	1.634	**39.39073*****
15	Gravidez, Grau_Varicoc = II	Gravidez_espontanea	0.117	0.562	2.64	1.799	**43.98779*****
16	Gravidez, Idade_M = Range 1 <31	PMA	0.117	0.692	2.413	2.317	**37.57123*****
17	Gravidez, Conc_3M = 0.01 to 14.9	PMA	0.113	0.684	2.384	2.258	**35.07919*****
18	Gravidez, Grau_Varicoc = II	PMA	0.113	0.542	1.888	1.556	**19.23664*****
20	Gravidez, A_B_3M = 1 to 31	PMA	0.109	0.714	2.489	2.496	**36.95628*****
21	Formas_N_3M = 1 to 3	PMA	0.109	0.41	1.428	1.208	**6.144852****
21	Gravidez, Formas_N_3M = 1 to 3	PMA	0.109	0.781	2.723	3.26	**44.54074*****
22	Gravidez, Formas_N_3M = > 4	Gravidez_espontanea	0.104	0.571	2.682	1.836	**39.19622*****
24	Gravidez, A_B_3M = > 32	PMA	0.122	0.56	1.952	1.621	**23.35803*****
25	Gravidez, A_B_Pre = > 32	PMA	0.113	0.605	2.107	1.804	**26.07485*****
26	Gravidez, A_B_3M = > 32	Gravidez_espontanea	0.122	0.56	2.629	1.789	**46.01329*****

Table 5.67 Selected Results – step 6 - test 6.1.1 - Types of Conceptions for n=170

Related Rule id	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x2
10	A_B_Pre = 1 to 31	PMA	0.308	0.66	1.07	1.127	0.738495732
11	Conc_3M = > 15	PMA	0.308	0.611	0.991	0.985	0.014165584
12	HabitosTabagicos_Simplificado	PMA	0.28	0.769	1.247	1.66	**6.01347209 ****
14	Conc_3M = > 15	Gravidez_espontanea	0.262	0.519	1.132	1.126	1.609187652

Related Rule id	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x2
15	Grau_Varicoc = II	Gravidez_espontanea	0.252	0.562	1.228	1.239	**3.815525892 ***
16	Idade_M = Range 1 <31	PMA	0.252	0.692	1.122	1.245	1.467874555
17	Conc_3M = 0.01 to 14.9	PMA	0.243	0.684	1.109	1.213	1.127383383
18	Grau_Varicoc = II	PMA	0.243	0.542	0.878	0.836	2.087846344
20	A_B_3M = 1 to 31	PMA	0.234	0.714	1.158	1.341	2.094053694
21	Formas_N_3M = 1 to 3	PMA	0.234	0.781	1.267	1.752	**5.243841981 ****
22	Formas_N_3M = > 4	Gravidez_espontanea	0.224	0.571	1.248	1.265	**3.583052094 ***
24	A_B_3M = > 32	PMA	0.262	0.56	0.908	0.871	1.28130713
25	A_B_Pre = > 32	PMA	0.243	0.605	0.98	0.969	0.046351691
26	A_B_3M = > 32	Gravidez_espontanea	0.262	0.56	1.223	1.232	**3.951418507 ****

The rules in highlighted inTable 5.66 and in Table 5.67 enables us to formulate the following conclusions regarding the types of conceptions:

- 61.7% of the pregnancies were from an ART procedure carried out in the CHUC. In fact, we have 66 patients that conceived with an ART procedure out of the 107 patients achieved pregnancy (*confidence*= 66/107=0.617).
- 45.8% of the pregnancies were spontaneous. In fact, we have 49 patients that conceived spontaneously out of the 107 pregnant patients (*confidence*= 49/107=0.458).
- 76% of the woman that conceived did so with an ART procedure and having a partner that smokes. In fact, we have 30 women that have a smoker partner and conceived via ART over the 39 male smokers that we have among the 107 couples that conceived (i.e. *confidence* = 30/39=0.769). Hence, among the 173 patients with defined smoking habits, 45.66% (79/173) smoked and half of them (49.37% (39/79)) were able to conceive mainly with the help of an ART procedure with a confidence of 76% (30/39) since the ones that conceived spontaneously in these conditions were only of 33.33% (13/39). Hence, we have a *support* of 28% (i.e. *support* = 30/107 = 0.280). Note that we have not identified an interesting rule that relates smoking habits with spontaneously getting pregnant.
- 78.10% of the woman that conceived, did so via ART by having a partner with an abnormal sperm morphology 3 months after the treatment that went from 1% to 3%. In fact, we have 25 women that have a partner with an abnormal sperm morphology that ranges from 1% to 3% and that conceived with an ART over the 32 male partners that have an abnormal sperm morphology in that range of values (i.e. *confidence* = 25/32=0.781), among the 107 couples that conceived. Hence, we have a *support* of 23.4% (i.e. *support* = 25/107 = 0.234).
- 56% of the woman that conceived did so spontaneously by having a partner with normal sperm morphology 3 months after the embolization treatment. In fact, we have 28 women with a partner with normal sperm motility and conceived spontaneously over the 50 male partners that we have with a normal sperm motility 3 months after the embolization treatment (i.e. *confidence* = 28/50=0.56), among the 107 couples that conceived. Hence, the *support* is equal to 26.2% (i.e. *support* = 28/107 = 0.262).

Regarding the age ranges, we have only identified rules related with the woman patient ages under 31 years old but none of them were seen subjectively and objectively interesting and with a good Chi-square value. However, the one that showed to be only subjectively and objectively interesting was:

- **Idade_M = Range 1 <31 -> Gravidez** (Rule ID =6)
  The conditional probability of getting pregnant before 31 years old is of 54.9%. In fact, we have 39 women that got pregnant before 31 years old, over the 71 women that we have in that range of age values (i.e. 39/71=0.549). However, there is not a statistically significant relation between being under 31 years old and getting pregnant since the computed Chi-square value disclosed in Table 5.65 was only above the significance level of p=0.010 since its x^2 is equal to 2.9.

Furthermore, in that range of women ages we have seen with the 9th rule of Table 5.65 that 14.3% of these couples had a male patient that smoked and these two aspects were shown to be related (i.e. $x^2 = 6.6649$ which is a $p<0.01$). Moreover, with the 16th rule of Table 5.66 we have seen that 11.7% of ART procedures in the CHUC were performed on women under the age of 31 and the relation of this aspect was seen statistically significant (i.e. x^2=37,57123 which is a $p<0.01$).

Regarding the attribute association between the drinking or smoking habit with other attribute values, we have seen that the drinking habit does not even appear in the 208 generated rules which means that the drinking habit is not a relevant attribute. However, regarding the smoking habit, we have seen some rules with the “HabitosTabagicos_Processado_Simplificado” attribute as an antecedent or a consequent in the identified rules (i.e. Rule ID 8, 9,12, 19). The most subjectively interesting association rule in the context of sperm parameters is the 8th rule since it conveys that 15.2% of the patients that smoke have normal sperm motility. However, the computed x^2 value did not indicate that there was a relation between these two attributes. In the same context, the 19th rule was objectively interesting and also relevant for this context since it tells that 11.3% of the male patients that had abnormal sperm parameters 3 months after the treatment that went from 1% to 3% was a smoker. Unfortunately, even though the computed x^2 value is close to be statistically significant for $p<0.01$, it is not (it has a $x^2 = 2.5208$ and the x^2 for a p=0.10 is 2.706). Hence, the smoking habit of the male patient is not related with the abnormal sperm morphology at 3 months, neither with other attributes; and therefore, we can say that the *final preprocessed data set* did not compute statistically significant association rules regarding patient smoking or drinking habit with the assessed attributes. Note that for rule 19th we have even retested the model with only two ranges of values for the sperm morphology by appending the range 0% to the following range 1% to 3%, and no rule was generated for the sperm morphology attribute, so it showed to be an even worse discretization.

Regarding the sperm parameters and the pregnancy outcome, only two rules were seen objectively (i.e. support < 0.15) interesting (Rule 13 and 23) because the other rules were biased - since they had a type of conception as an antecedent and the “Gravidez” attribute, as a consequent or they were related with the woman age. These two rules can be interpreted as follows:

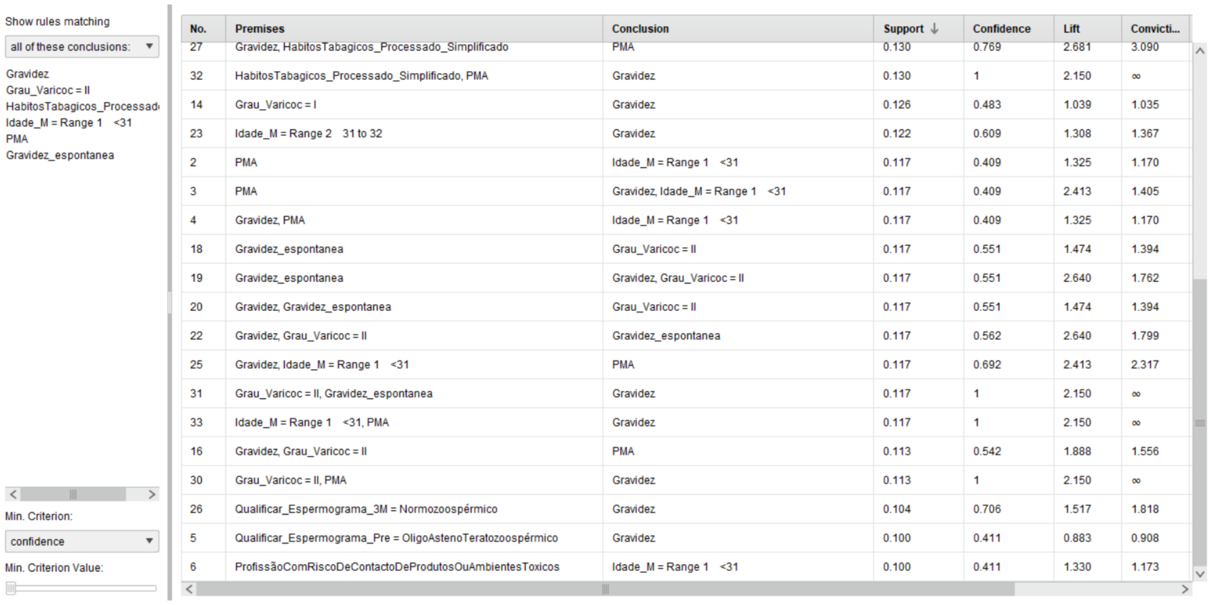
- **A_B_Pre = 1 to 31, Conc_3M = > 15 -> Gravidez** (RULE 13)
  The conditional probability of getting pregnant given an abnormal sperm motility before the treatment that goes from 1% to 31% and a normal sperm concentration 3 months after the embolization treatment is of 59.6% (i.e. 28/47).
- **A_B_Pre = 1 to 31, Grau_Varicoc = II -> Gravidez** (RULE 23)
  The conditional probability of getting pregnant given an abnormal sperm motility before the treatment that goes from 1% to 31% with a moderate varicocele is of 62.2% (i.e. 23/37).

5.5.3.2 Generated results – Step 6 – Test 6.2 (Model 4)

To look up into the 230 instances for what the semen categorization can tell us, we have adapted and rerun the model depicted in Figure 6.29 by getting rid of all sperm parameter values, as well as its related discretization, and have added the semen categorization attributes related with before and 3 months after the treatment (i.e. the attribute "Qualificar_Espermograma_Pre" and "Qualificar_Espermograma_3M" that were seen statistically related with the "Gravidez" attribute). Finally, we have ended up with the model depicted in Figure 6.30.

In this test, all subjectively and objectively interesting rules were seen as the same as previously identified and the ones that appeared related with the semen categorization, had very low support (i.e. *support* = 0.10) (see Figure 5.49); and therefore, they were not considered for further analyses. However, the semen categorization at 3 months after the embolization treatment presented a high Chi-square above 0.01 (x^2 = 9,245483) but we unfortunately had a small sample (see Rule No. 26 in Table 5.68): 24 patients out of the 230 patients assessed had a pregnancy by having 3 months after the treatment a normal semen (24/230=0,104) and the conditional probability of conceiving having a normal semen is of 70.6% (24/34), which is a pretty high probability. Hence, it is for sure an objectively interesting rule but not subjectively interesting since it encompasses a small number of patients (i.e. has a support lower than 0.15).

The results can be partly seen below, where we can see these only two generated rules coming at last in the table depicted in Figure 5.49 that presents a print screen of the generated results in the RapidMiner platform. The following Table 5.68 and Table 5.69, disclose the generated resulted for the semen classifications where the only conclusions that were seen generated for these attributes were the type of conception.



Show rules matching

all of these conclusions:

Gravidez
Grau_Varicoc = II
HabitosTabagicos_Processad
Idade_M = Range 1 <31
PMA
Gravidez_espontanea

Min. Criterion: confidence

Min. Criterion Value:

No.	Premises	Conclusion	Support	Confidence	Lift	Convicti...
27	Gravidez, HabitosTabagicos_Processado_Simplificado	PMA	0.130	0.769	2.681	3.090
32	HabitosTabagicos_Processado_Simplificado, PMA	Gravidez	0.130	1	2.150	∞
14	Grau_Varicoc = I	Gravidez	0.126	0.483	1.039	1.035
23	Idade_M = Range 2 31 to 32	Gravidez	0.122	0.609	1.308	1.367
2	PMA	Idade_M = Range 1 <31	0.117	0.409	1.325	1.170
3	PMA	Gravidez, Idade_M = Range 1 <31	0.117	0.409	2.413	1.405
4	Gravidez, PMA	Idade_M = Range 1 <31	0.117	0.409	1.325	1.170
18	Gravidez_espontanea	Grau_Varicoc = II	0.117	0.551	1.474	1.394
19	Gravidez_espontanea	Gravidez, Grau_Varicoc = II	0.117	0.551	2.640	1.762
20	Gravidez, Gravidez_espontanea	Grau_Varicoc = II	0.117	0.551	1.474	1.394
22	Gravidez, Grau_Varicoc = II	Gravidez_espontanea	0.117	0.562	2.640	1.799
25	Gravidez, Idade_M = Range 1 <31	PMA	0.117	0.692	2.413	2.317
31	Grau_Varicoc = II, Gravidez_espontanea	Gravidez	0.117	1	2.150	∞
33	Idade_M = Range 1 <31, PMA	Gravidez	0.117	1	2.150	∞
16	Gravidez, Grau_Varicoc = II	PMA	0.113	0.542	1.888	1.556
30	Grau_Varicoc = II, PMA	Gravidez	0.113	1	2.150	∞
26	Qualificar_Espermograma_3M = Normozoospérmico	Gravidez	0.104	0.706	1.517	1.818
5	Qualificar_Espermograma_Pre = OligoAstenoTeratozoospérmico	Gravidez	0.100	0.411	0.883	0.908
6	ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	Idade_M = Range 1 <31	0.100	0.411	1.330	1.173

Figure 5.49 Association rules for the 6th group of attributes – Step 6 - test 6.2

Table 5.68 Selected Results – Step 6 - test 6.2 – Semen Classifications for n=230

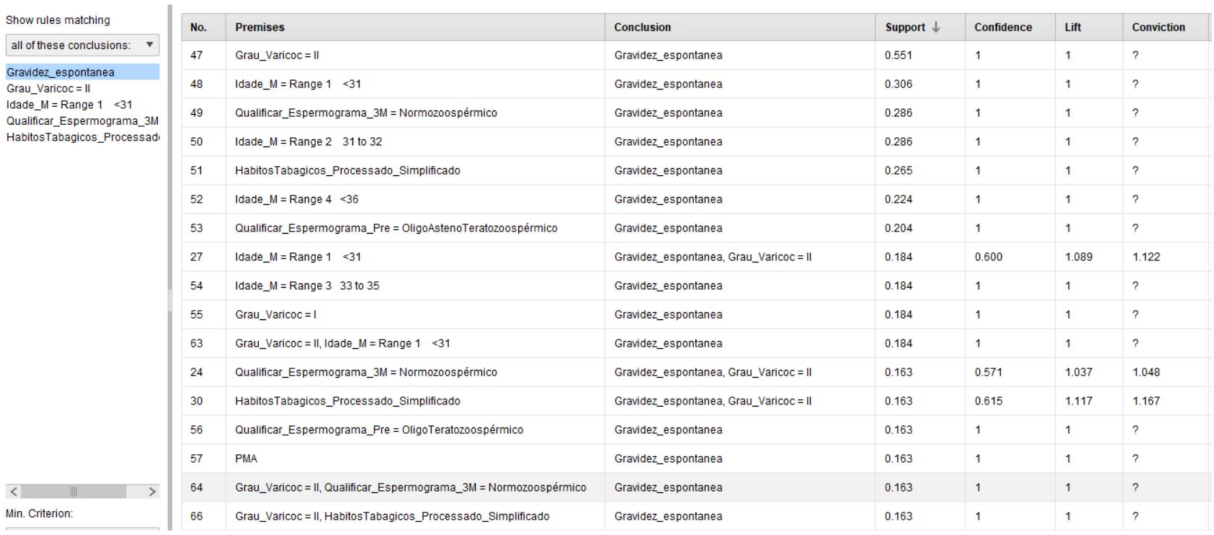
No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
5	Qualificar_Espermograma_Pre = OligoAstenoTeratozoospérmico	Gravidez	0.1	0.411	0.883	0.908	0.881534
26	Qualificar_Espermograma_3M = Normozoospérmico	Gravidez	0.104	0.706	1.517	1.818	**9.245483*****

Table 5.69 Selected Results – Step 6 - test 6.2.1 – Semen Classifications for n=170

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
24	Qualificar_Espermograma_Pre = OligoAstenoTeratozoospérmico	PMA	0.14	0.652	1.057	1.102	0.15303266
20	Qualificar_Espermograma_3M = Normozoospérmico	Gravidez_espontanea	0.131	0.583	1.274	1.301	1.96430206
17	Qualificar_Espermograma_3M = Normozoospérmico	PMA	0.121	0.542	0.878	0.836	0.73835755
25	Qualificar_Espermograma_Pre = OligoTeratozoospérmico	PMA	0.121	0.684	1.109	1.213	0.43972476
9	Qualificar_Espermograma_3M = Normozoospérmico	Grau_Varicoc = II	0.103	0.458	1.022	1.018	0.01220181
12	Qualificar_Espermograma_Pre = OligoAstenoTeratozoospérmico	Grau_Varicoc = II	0.103	0.478	1.066	1.057	0.10407063

To define a data pattern for the types of conceptions, we have further assessed the 49 patients that conceived spontaneously, and then the 66 patients that did so via ART pby adapting the model depicted in Figure 6.30 with the resetting of the “Select Attribute” operator accordingly to this aim.

Regarding the 49 patients that conceived spontaneously, we have looked up for patients with normal semen since we have previously identified an objectively interesting rule on that matter (Rule No 26 in Table 5.68). Hence, into the dataset we have looked up for the semen categorization set as “Normozoospermia” before and 3 months after the embolization treatment for the couples that spontaneously conceived and we have seen that out of the 49 couples, 3 had Normozoospermia before the treatment and 3 months later, 14 had Normozoospermia. Hence, the probability of getting pregnant spontaneously by having normozoospermia at 3 months after the treatment is of 28.6% (14/49) and by having a moderate varicocele condition is 55.1% (27/49). Regarding the patients that conceived spontaneously with a mild varicocele (severity grade=I), we have 18.4% (9/49). We have also seen that the greater group of patients that conceived spontaneously had less than 31 years old (15/49 = 30.6%) and that the male patient was a smoker 26.5% (13/49) of the time. All these probabilities were extracted through the support presented in the Figure 5:61. This figure depicts the most subjectively and objectively interesting results obtained on the data set filtered by the couples that conceived spontaneously t. Hence, these rules were generated on the 49 instances that had the “Gravidez_espontanea” attribute set to TRUE. This is why confidence values are seen all equal to 1 for the conclusions with the “Gravidez_espontanea” attribute alone. Hence, this analysis was only descriptive.



Show rules matching

all of these conclusions: ▼

Gravidez_espontanea
Grau_Varicoc = II
Idade_M = Range 1 <31
Qualificar_Espermograma_3M
HabitosTabagicos_Processad

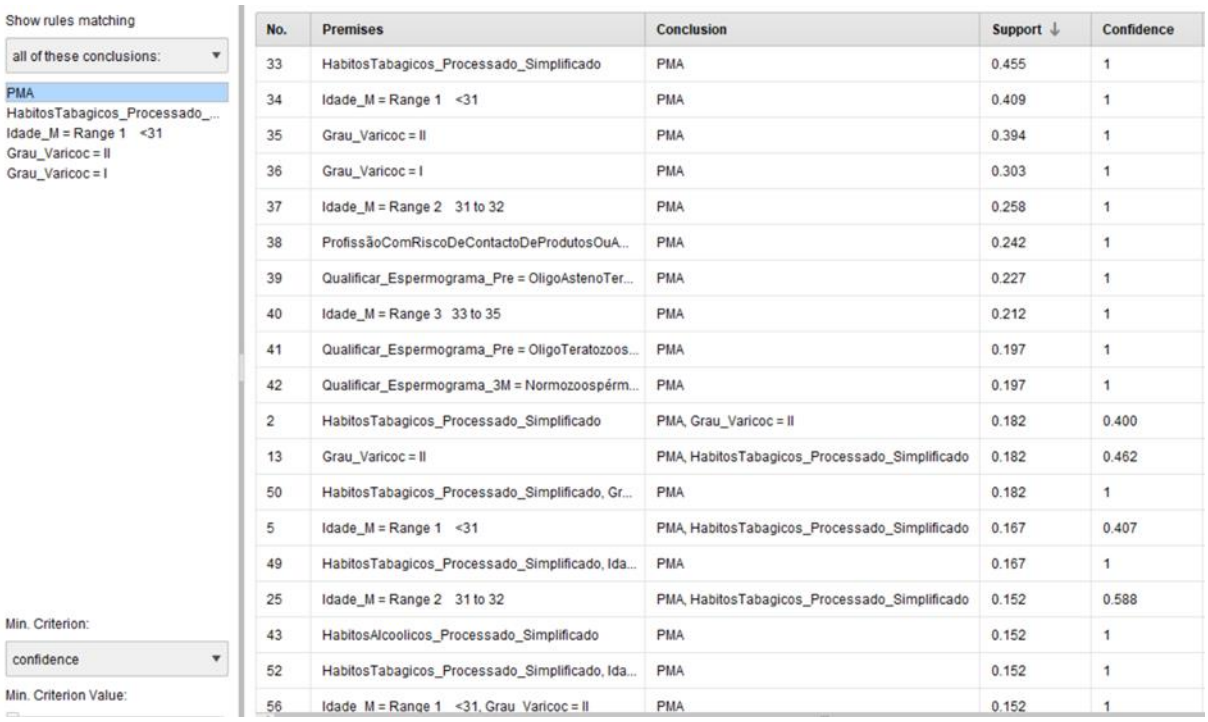
Min. Criterion:

No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction
47	Grau_Varicoc = II	Gravidez_espontanea	0.551	1	1	?
48	Idade_M = Range 1 <31	Gravidez_espontanea	0.306	1	1	?
49	Qualificar_Espermograma_3M = Normozoospérmico	Gravidez_espontanea	0.286	1	1	?
50	Idade_M = Range 2 31 to 32	Gravidez_espontanea	0.286	1	1	?
51	HabitosTabagicos_Processado_Simplificado	Gravidez_espontanea	0.265	1	1	?
52	Idade_M = Range 4 <36	Gravidez_espontanea	0.224	1	1	?
53	Qualificar_Espermograma_Pre = OligoAstenoTeratozoospérmico	Gravidez_espontanea	0.204	1	1	?
27	Idade_M = Range 1 <31	Gravidez_espontanea, Grau_Varicoc = II	0.184	0.600	1.089	1.122
54	Idade_M = Range 3 33 to 35	Gravidez_espontanea	0.184	1	1	?
55	Grau_Varicoc = I	Gravidez_espontanea	0.184	1	1	?
63	Grau_Varicoc = II, Idade_M = Range 1 <31	Gravidez_espontanea	0.184	1	1	?
24	Qualificar_Espermograma_3M = Normozoospérmico	Gravidez_espontanea, Grau_Varicoc = II	0.163	0.571	1.037	1.048
30	HabitosTabagicos_Processado_Simplificado	Gravidez_espontanea, Grau_Varicoc = II	0.163	0.615	1.117	1.167
56	Qualificar_Espermograma_Pre = OligoTeratozoospérmico	Gravidez_espontanea	0.163	1	1	?
57	PMA	Gravidez_espontanea	0.163	1	1	?
64	Grau_Varicoc = II, Qualificar_Espermograma_3M = Normozoospérmico	Gravidez_espontanea	0.163	1	1	?
66	Grau_Varicoc = II, HabitosTabagicos_Processado_Simplificado	Gravidez_espontanea	0.163	1	1	?

Figure 5:61 Association Rules for the spontaneous pregnancies – step 6 – test 6.2.2

Regarding the ART procedure, in the next figure, we present the most subjectively and objectively interesting results obtained on the data set filtered by the patients that conceived via ART p (i.e. these rules were generated on the 66 instances that had the “PMA” attribute set to TRUE analogously as disclosed in Figure 5:61).

If we focus on the rules with the highest support, we see that 45.5% (30/66) of the couples that have conceived with an ART procedure had the male patient as a smoker and 40.9% (27/66) of the woman were younger than 31 years. Regarding the severity grade of the varicocele condition, while 55.1% (27/49) of the spontaneous pregnancies had a moderate varicocele condition, with the ART procedure, only 39.4% (26/66) had that same severity grade followed by 30.3% (20/66) for the low severity grade (severity grade=I). On the matter of semen categorization, we see that while the spontaneous pregnancies had 28.9% (14/49) of its male patient with normozoospermia 3 months after the treatment, we here only have 19.7% (13/66) of the patients with normal sperm parameter values for that same time. Hence, male patients that were able to conceive through an ART procedure did not improve in semen categorization as much as the ones that were able to get pregnant spontaneously but we have to say that the panorama of the semen categorization before the treatment was also worse for patients that got pregnant with an ART procedure (i.e. OAT->PMA (support=0.227) vs OAT->Gravidez_espontanea (support=0.204)).



Show rules matching

PMA
HabitosTabagicos_Processado_...
Idade_M = Range 1 <31
Grau_Varicoc = II
Grau_Varicoc = I

Min. Criterion: confidence

Min. Criterion Value:

No.	Premises	Conclusion	Support ↓	Confidence
33	HabitosTabagicos_Processado_Simplificado	PMA	0.455	1
34	Idade_M = Range 1 <31	PMA	0.409	1
35	Grau_Varicoc = II	PMA	0.394	1
36	Grau_Varicoc = I	PMA	0.303	1
37	Idade_M = Range 2 31 to 32	PMA	0.258	1
38	ProfissãoComRiscoDeContactoDeProdutosOuA...	PMA	0.242	1
39	Qualificar_Espermograma_Pre = OligoAstenoTer...	PMA	0.227	1
40	Idade_M = Range 3 33 to 35	PMA	0.212	1
41	Qualificar_Espermograma_Pre = OligoTeratozoos...	PMA	0.197	1
42	Qualificar_Espermograma_3M = Normozoospérm...	PMA	0.197	1
2	HabitosTabagicos_Processado_Simplificado	PMA, Grau_Varicoc = II	0.182	0.400
13	Grau_Varicoc = II	PMA, HabitosTabagicos_Processado_Simplificado	0.182	0.462
50	HabitosTabagicos_Processado_Simplificado, Gr...	PMA	0.182	1
5	Idade_M = Range 1 <31	PMA, HabitosTabagicos_Processado_Simplificado	0.167	0.407
49	HabitosTabagicos_Processado_Simplificado, Ida...	PMA	0.167	1
25	Idade_M = Range 2 31 to 32	PMA, HabitosTabagicos_Processado_Simplificado	0.152	0.588
43	HabitosAlcoolicos_Processado_Simplificado	PMA	0.152	1
52	HabitosTabagicos_Processado_Simplificado, Ida...	PMA	0.152	1
56	Idade_M = Range 1 <31, Grau_Varicoc = II	PMA	0.152	1

Figure 5:62 Association Rules for the pregnancies from ART – step 6 – test 6.2.3

5.6 Evaluation and Discussion

Regarding the data understanding and the modeling phases, they were carried out with the RapidMiner platform version 8.1.001. In fact, from the last poll (Piatetsky, 2018), that has inquired 2025 participants on which Data Science tools they were using, the RapidMiner platform was seen the mostly used rising from 33% in 2017 to 52.7% in 2018 which reassured our choice after choosing it in 2016 through a comprehensive study in that matter. However, after its use, we have determined that it has some limitations in the statistics field (e.g. there is only one and unstable non-parametric inferential statistical test) which is understandable due its focus on data mining algorithms. Nevertheless, its wide range of operators and extensions made it possible to meet the goals set.

Regarding data volume, the initial data set had 320 instances and 32 attributes which was at first sight seen as small on the matter of its number of instances. However, after analyzing the literature review in Makris *et al*. (2018) that studies 30 clinical investigations on the varicocele embolization domain, we have seen that the provided data set had an interesting volume of data since related works were in average of 117 patients (± 102 patients). Hence, even the 230 preprocessed instances with non-missing values under the pregnancy outcome attribute - that were the instances mostly analyzed during this study - remained a good volume of data. Regarding the application of data mining techniques, we have seen that it was possible since the reviewed related works on the infertility domain mainly deal with similar volume of data (as seen in Table 3.1).

Data quality was assessed with key data quality dimensions to know if the attributes provided were directly usable to tackle the data mining goals set (i.e. completeness), as well as coherent (i.e. consistency), rightly formatted (i.e. conformity), correct (i.e. accurate with the available information systems) and correctly linked (i.e. integrity) as suggested in Maydanchik (2007). To have a glimpse of this assessment, we have built a data quality score board after preprocessing (see Table 5.37). As we have seen, most attributes were validated/filled/corrected with the available information systems of the CHUC which enabled us to increase it completeness by going in average from a 55.86% to a 70.40% filled dataset. However, the severity grade, the varicocele´s laterality and the embolization ´s complications were not fully validated (i.e. validated 27.30% of the 293 not duplicated and provided instances). The reason behind this validation rate was that we have decided to review the medical dossiers that we were able to receive in a stipulated time span and then go with the information that we had to minimize the overhead of retrieving data that could latter on reveal itself to be not statistically significant with the goals set.

After preprocessing the provided data set, our main concern was to statistically analyze the preprocessed dataset to not only better understand it, but also, check if the embolization treatment was in the first place a successful treatment since one of the goals of the BRSC team was to predict its success. To do so, we have first determined criteria of success, i.e. improvement of sperm parameters as carried out in Kirby *et al*. (2016), and then, have applied *inference statistics* such as the ANOVA statistical test, upon numerical attributes, and the Chi-square statistical test, upon nominal attributes, to test our hypothesis. The choice of these statistical tests was based on the literature since the ANOVA and the Chi-square statistical tests were mostly used in the varicocele domain and specially, upon sperm parameter data (see section 3.4).

Statistical results helped us to overcome some encountered difficulties. In fact, it enabled us to elect a balanced label attribute, as well as, select the attributes that were more related with it since the Pearson correlations were all seen as low with the selected label attribute.

Regarding the label attribute, we have selected the pregnancy outcome attribute as in Guh *et al*. (2011) since it delivered the most balanced data set (i.e. 123 instances have the pregnancy outcome set to No and 107 have it set to Yes). Concerning the identification of the most statistically significant attributes, we have seen with the ANOVA and the Chi-square test that the following attributes were seen related with the pregnancy outcome (Table 5.40): Woman age (ANOVA p=0.018); Severity grade (Chi-square p=0.049); Concentration at 6 months (ANOVA p=0.015); Progressive motility before treatment (ANOVA p=0.018); Morphology at 3 months (ANOVA p=0.004); Concentration category at 3 months (Chi-square p=0.017); Progressive Motility category before treatment (Chi-square p=0.027); Progressive Motility category at 3 months (Chi-square p=0.022); Semen classification before treatment (Chi-square p=0.017); Semen classification at 3 months (Chi-square p=0.018) and Hazardous Occupation (Chi-square p=0.023). As these attributes reveal, several data transformations were carried out upon the provided and preprocessed data set which showed to potentiate knowledge discovery. These data transformations were: dichotomization of the severity grade, normalization of the numeric attributes and transformation of the numerical attributes into different nominal attributes.

To maximize knowledge discovery, we have selected the most commonly applied data mining techniques in the healthcare industry (i.e. classification, clustering and association) with their well tested algorithm based on Tekieh and Raahemi (2015), Ahmad *et al*. (2015) and Tomar and Agarwal (2013). Thereby, these data mining techniques were applied with the following algorithms: classification, with the RapidMiner´s Decision tree algorithm and the W-J48 java implementation of the C4.5 algorithm; clustering, with the K-means algorithm and association rule, with the FP-Growth algorithm.

All these algorithms were mainly trained upon the identified attributes that were related with pregnancy outcome by varying its main parameters as specified in section 4.2.8. This task was achieved with the "optimized parameter" operator that helped us to automatically loop the several model parameters in order to select the best based on performance measures (i.e. mainly the *F-measure* along with the *Accuracy* and the *AUC* measure). This "optimized parameter" operator was very useful since during our first sub-modeling phase (i.e. decision tree modeling step 1 and 2) we struggled to find a decision tree with even 1 level. Hence, when we have sought a solution that could optimize the training process by exhaustively train/test the algorithms, we have found this operator which enabled us to also maximize knowledge discovery.

Since knowledge discovery was difficult with the decision tree algorithm, we have applied the clustering and the association rule technique in an early modeling stage to bring another understanding of the data that could help us on our search for the predictive model. This is the reason behind the order of the sub-modeling phases that this study has followed and disclosed in Table 4.13, where we see that right after applying 2 out of the 9 decision tree modeling steps during the sub-modelling phase 1, we have gone for the other data mining techniques in sub-modeling phase 2 and 3 to further on continue with the training of the decision tree algorithm in sub-modeling phase 4.

This modeling strategy was seen successful since the most interesting knowledge discovery was achieved during the K-Means application (disclosed in Table 5.70 as result id 14) which gave us the idea to train/test the decision tree algorithm on this same preprocessed dataset during the sub-modeling phase 4 and in turn, also gave us the predictive model (disclosed in Table 5.70 as result id 5). Through this data mining experiment we have seen that due to the small and missing data that we had, it is understandable that it is more achievable to extract interesting knowledge with a K-means algorithm, that is less influenced by missing data since it seeks to group the data through similarities between data points, than with a decision tree algorithm, that tries to train/test upon missing values; and hence, struggles to select the attribute that promotes the highest gain of information for its decision tree. Hence, due to this experience, we can say that it is important to first identify the most commonly applied data mining techniques in a research domain and apply them as a whole, since the different techniques can complement each other and potentiate interesting knowledge discovery.

Regarding the association rule technique, we have found that it is a good technique to identify attributes or relations that are interesting (i.e. mainly with the highest *support* and/or *confidence*). This technique clearly depicts one of the advantages of data mining, which is to be an inductive technique and not a hypothetic-deductive technique as statistical analysis is. Therefore, it is, in our point of view, an interesting technique to begin with during the data understanding modeling phase, even before *inferential statistics*, to identify relations or attributes that we might want to assess statistically later on - when we are not able to formulate a hypothesis – or detect interesting data patterns (e.g. result id 11). Since all clinical research on varicocele have used hypothetic-deductive techniques, such as the ANOVA and the Chi-square test (seen in section 3.4), we can say, based on what we have experienced, that the joining of data mining techniques with inferential statistical techniques, is very useful because they complete each other in spite of being rarely joined together in related work. In contrast, this study has joined these data analysis techniques as follows: along with the clustering technique, we have applied the ANOVA statistical test upon the computed k-means centroid means to assess if their was a statistical significant difference between the cluster means as in Zancanaro *et al*. (2007), Furthermore, along with the association rule technique, we have applied the Chi-square test upon the FP-Growth performance measure to assess if there was a statistical significant relationship between the antecedent and the consequent of the rule as in Brin *et al*. (1997) to complement the *lift* measure that only tells us how far from independent the events are. We believe that the reason why the joining of these techniques is rarely applied is that most works are not performed in a multidisciplinary investigation team where there is a sharing of practices between fields as we have experienced in this project.

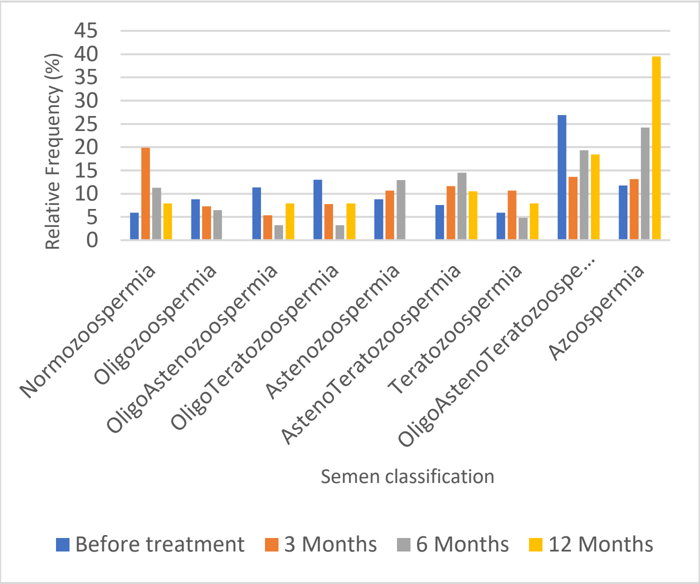
Regarding the testing of the decision tree algorithms, we have followed two test designs to unveil eventual test overfits; i.e, we have firstly tested the data set by splitting it into 3 parts (i.e. 80% for training/testing and 20% for validation, where in the 80% part, 70% was taken for training and the remaining 30%, for testing the data set) and the best models were retested by splitting into 2 parts (i.e. 70% for training and 30% for testing). The 3 parts test design was suggested by the CRISP-DM methodology (Chapman et al., 2000), the founder of the RapidMiner platform (Mierswa, 2012), as well as the consulting company SimaFore (Deshpande, 2012) to overcome test overfitting. However, most related works as (Guh et al., 2011) only implement the 2-part test design. Since we have applied both test designs, we were able to discuss its benefit: Through the several generated decision tree models depicted in Appendix C.1, we have seen that the only model that has computed an acceptable AUC during the validation was the one computed during the step 6 (i.e. AUC = 0.750). Nevertheless, we believe that the non-missing values requirement trained/tested in this step 6 has also contributed to its acceptable result. In contrast, all other models have failed with an AUC from 0.500 to 0.604 during the validation of their best model (i.e. model that computes the highest F-measure during a modeling step). Hence, if we would not have tested the models with the 3-part test design, we would not be able to say that the elected model was stable, but more importantly, that we were not misled by the performance measures obtained during its training/testing. In fact, in Appendix C.1 are models that had computed acceptable AUC and F-measures during training, but failed during validation (e.g. model of step 3 has a training/testing AUC=0.747 and a validation AUC=0.554). Hence, based on this experience, we believe that the 3-part test design is more suitable for health care data sets that usually have several missing values as stated in Tekieh and Raahemi (2015). Please note that the elected model 6 is identified as result id 5 in Table 5.70.

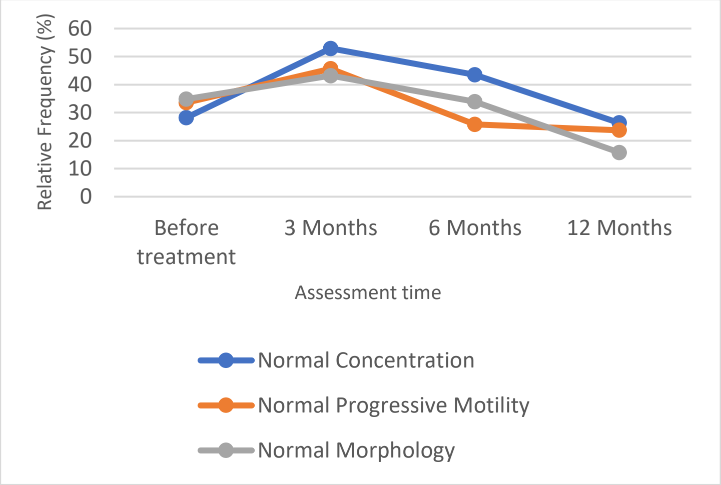
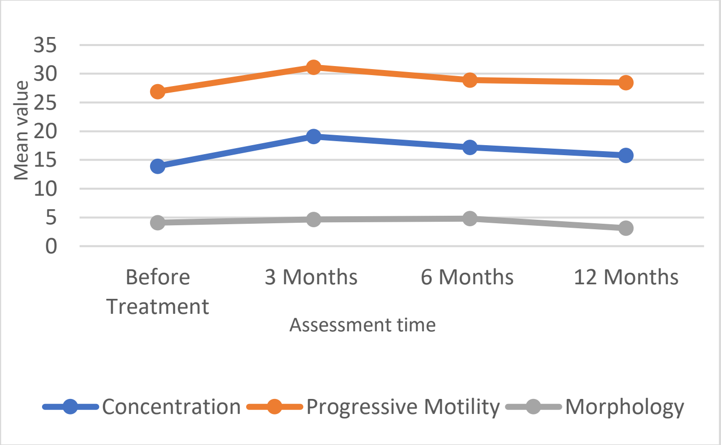
Results were firstly elected based on their performance metrics, as specified in section 4.2.8, and afterwards, on the fulfillment of the defined data mining goals. Finally, the BRSC team has approved these models and pinpointed the most interesting ones for their research. By doing so the evaluation Crisp-DM phase was achieved.

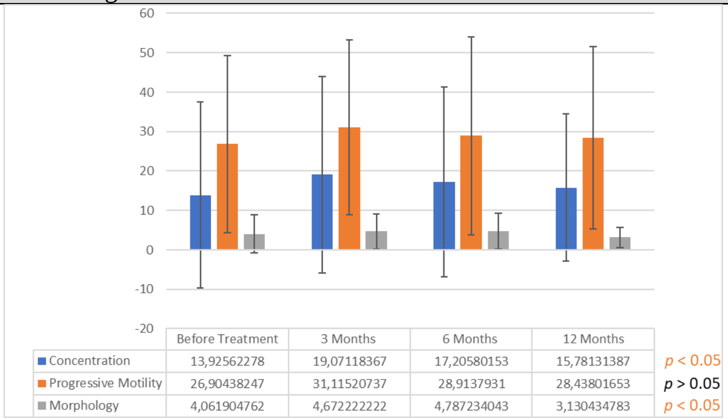
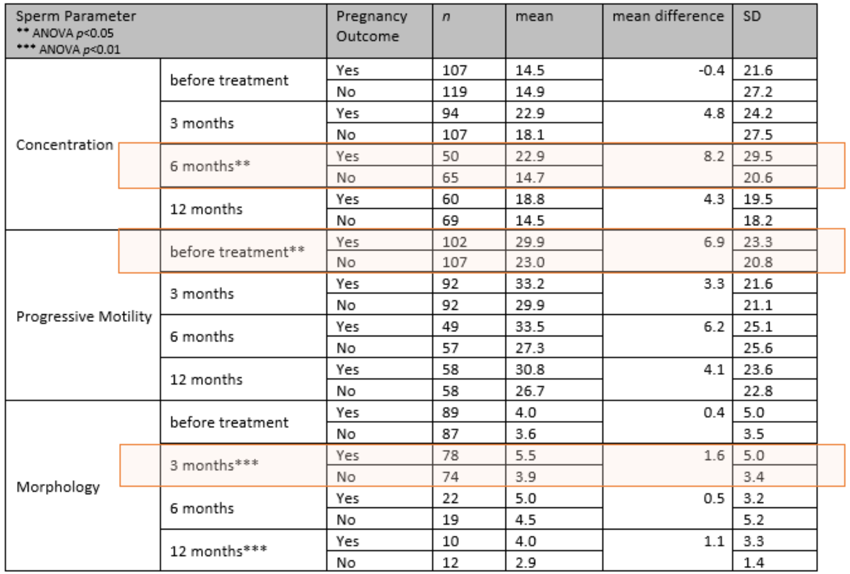
The outcome of the evaluation phase is depicted in the below Table 5.70 where we can see the mapping of the best elected results with the defined data mining goals, as well as the pinpointing of the most interesting results for the varicocele clinical research that are highlighted in light orange in the corresponding “Result ID” cell. Thereby, we can see that this study has achieved its possible data mining goals disclosed in section 5.2.3, as well as tackled with success the project risks initially identified in section 5.2.2. During the construction of Table 5.70, we have also rerun (i.e. tested) the built models and reviewed its modeling process, as the CRISP-DM suggests, to technically validate these clinically approved models. Some of the model results disclosed in Table 5.70 were adjusted to better convey them to the BRSC team.

Next, we discuss the highlighted results of Table 5.70 by relating them with the initially computed statistical results (i.e. section 5.3.2); the statements of medical dictionaries (i.e. Chapter 2) and the findings of related works (i.e. Chapter 3). At last, conclusions are defined based on the generalizations validated by the BRSC team.

Table 5.70 Summary of highlighted results and models

Result ID	Goal Nº	Task Description	Best elected Result	Interesting Related outcome	Data mining Step
1	5	Understanding the evolution of semen categorization and sperm normality through time	**There is a statistically significant relationship between the semen classification and the time when the semen analysis was carried out (x^2 p<0.05)**. We have seen that before the treatment, the biggest semen classification was the OligoAsthenoTeratozoospermia (OAT) with 26.89% (64/238) and 3 months later, it was Normozoospermia with 19.90% (41/206) by increasing in 14%. Further on, we have also seen that **there is a statistically significant relationship between the relative frequency of patients with normal sperm parameter values and the time when the semen analysis was carried out (x^2 p<0.05).** In this matter, we have seen that the relative frequency of normal sperm parameter values has increased 3 months after the embolization treatment for all sperm parameters. However, 12 months after the treatment, the relative frequency of normal sperm parameter values is lower than before the treatment and the largest semen categorization group is of Azoospermia.		Data Understanding (Bar graph from Figure 5.13 and series plot from Figure 5.14)

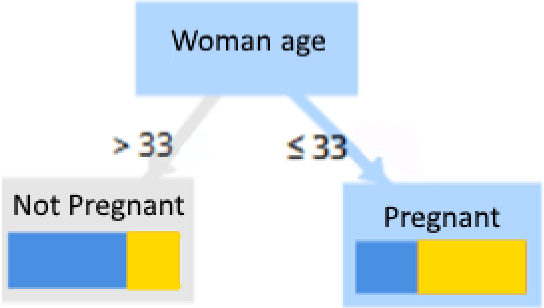
Result ID	Goal Nº	Task Description	Best elected Result	Interesting Related outcome	Data mining Step
					
2	5	Understanding the evolution of sperm parameter values through time	**The embolization treatment statistically significantly improves the mean value of sperm concentration (ANOVA *p*=0.017) and sperm morphology (ANOVA *p*=0.001) until 6 months after the treatment.** However, the sperm parameter that benefits more from it is the sperm concentration. In fact, 12 months after the embolization, it is still higher than before. Furthermore, **patients that got their partner pregnant had in average a greater response to the treatment since they had a higher mean value in its sperm morphology at 3 months (ANOVA *p*=0.004) and sperm concentration at 6 months (ANOVA *p*=0.015)**. Moreover, successful patients also had a higher mean value in its sperm progressive		Data Understanding (series plot from Figure 5.16, bar graph from Figure 5.17 and pregnancies results from Table 5.16)

Result ID	Goal Nº	Task Description	Best elected Result	Interesting Related outcome	Data mining Step
			motility before the treatment (ANOVA *p*=0.018).	 	
3	5	Assessment of the relation between semen	There is **no statistically significant relationship between the semen classification and the smoking or**		Data Understanding

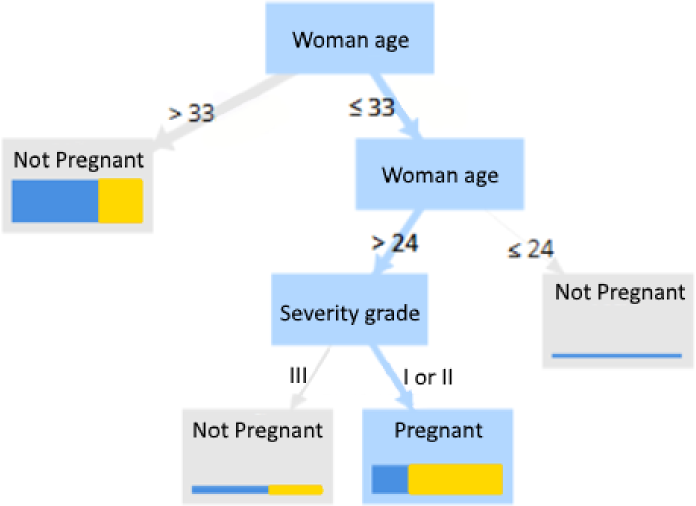
60
50
40
30
20
10
0
-10
-20

	Before Treatment	3 Months	6 Months	12 Months	
■ Concentration	13,92562278	19,07118367	17,20580153	15,78131387	$p < 0.05$
■ Progressive Motility	26,90438247	31,11520737	28,9137931	28,43801653	$p > 0.05$
■ Morphology	4,061904762	4,672222222	4,787234043	3,130434783	$p < 0.05$

Sperm Parameter ** ANOVA *p*<0.05 *** ANOVA *p*<0.01		Pregnancy Outcome	*n*	mean	mean difference	SD
Concentration	before treatment	Yes	107	14.5	-0.4	21.6
		No	119	14.9		27.2
	3 months	Yes	94	22.9	4.8	24.2
		No	107	18.1		27.5
	6 months**	Yes	50	22.9	8.2	29.5
		No	65	14.7		20.6
	12 months	Yes	60	18.8	4.3	19.5
		No	69	14.5		18.2
Progressive Motility	before treatment**	Yes	102	29.9	6.9	23.3
		No	107	23.0		20.8
	3 months	Yes	92	33.2	3.3	21.6
		No	92	29.9		21.1
	6 months	Yes	49	33.5	6.2	25.1
		No	57	27.3		25.6
	12 months	Yes	58	30.8	4.1	23.6
		No	58	26.7		22.8
Morphology	before treatment	Yes	89	4.0	0.4	5.0
		No	87	3.6		3.5
	3 months***	Yes	78	5.5	1.6	5.0
		No	74	3.9		3.4
	6 months	Yes	22	5.0	0.5	3.2
		No	19	4.5		5.2
	12 months***	Yes	10	4.0	1.1	3.3
		No	12	2.9		1.4

Result ID	Goal Nº	Task Description	Best elected Result	Interesting Related outcome	Data mining Step
		classification and patient´s external factors	**drinking habit**. The only statistical significant relation identified with the x^2 statistical test regarding smoking habits was related with the woman age (p<0.01) where we have seen that the conditional probability of having a male partner that smoked given a women under 31 years old is of 46.5% and both situations occur 14.3% of the times.		(x^2 statistical results are in Table 5.32 and in Table 5.33)
4	2	Prediction of the embolization success	We can say with 80.77% of *Accuracy*, 73.68% of *F-Measure* and 0.801 of *AUC* that most **women below and equal to 33 years old are able to get pregnant** (i.e. 63.41% (26/26+15)) in contrast to 29.55% (13/13+31) for women above 33 years old).	Model ran with the following parameter values: Sampling type during training/testing: shuffled; Decision tree splitting criterion: accuracy; Decision tree pruning: false; Decision tree minimal size for split: 5; Decision tree minimal gain: 0.1; Decision tree minimal leaf size: 3; Decision tree maximal depth: 20. Computed decision tree:  Tree description: Woman age > 33: Not Pregnant { No=31, Yes=13} Woman age <= 33: Pregnant { No=15, Yes=26}	Decision tree Modeling Step 6 without validation (result disclosed in Table 5.42)

<table>
<tr><th>Result ID</th><th>Goal Nº</th><th>Task Description</th><th>Best elected Result</th><th>Interesting Related outcome</th><th>Data mining Step</th></tr>
<tr><td>5</td><td>2</td><td>Prediction of the embolization success</td><td>We can say with 70.59% of *Accuracy* and *F-measure*, as well as with 0.750 of *AUC*, that **a man patient without a high varicocele severity grade is more prone to conceive if his partner is above 24 years old and below 33 years old inclusively**. In fact, 70.83% (17/(7+17)) of the assessed couples with these characteristics got pregnant (Please see the decision tree path highlighted in blue in the related model). Note that most female patients above 33 years old did not got pregnant (i.e. 67.57% (25/(25+12)), as well as none of the female patients below 24 years old inclusively.</td><td>Model ran with the following parameter values:
Sampling type during training/testing: linear;
Decision tree splitting criterion: accuracy;
Decision tree pruning: True;
Decision tree minimal size for split: 4;
Decision tree minimal gain: 0.1;
Decision tree minimal leaf size: 2;
Decision tree maximal depth: 20.

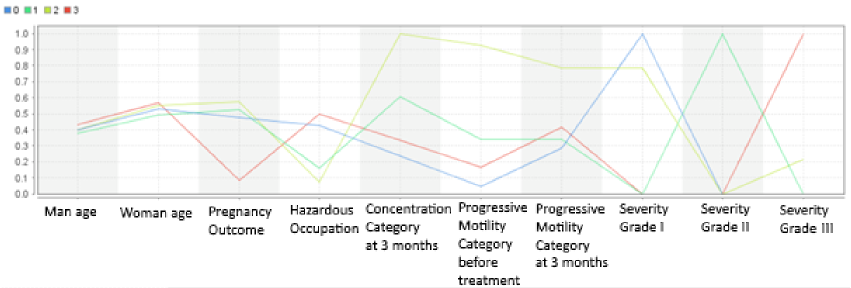
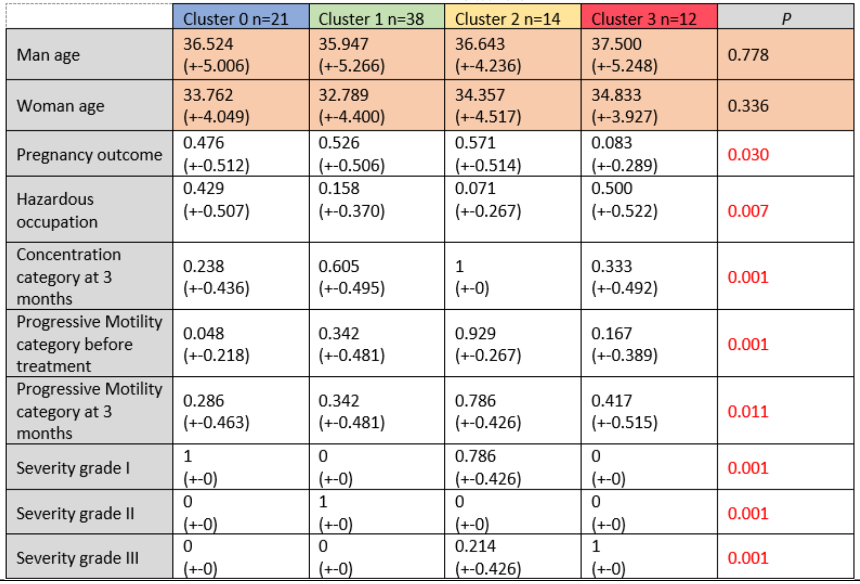
Computed decision tree:


Tree description:
Woman age > 33: Not Pregnant { No=25, Yes=12}
Woman age <= 33
| Woman age > 24
| | Severity grade = III : Not Pregnant {No=3, Yes=2}
| | Severity grade = I or II: Pregnant {No=7, Yes=17}
| Woman age <= 24: Not Pregnant {No=2, Yes=0}</td><td>Decision tree Modeling Step 6 with validation (result disclosed in Table 5.41)</td></tr>
</table>

Result ID	Goal Nº	Task Description	Best elected Result	Interesting Related outcome	Data mining Step
6	2	Prediction of the embolization success	**The conditional probability of a woman getting pregnant given a partner with a sperm morphology 3 months after the treatment greater than 0% is of 54.4%** (i.e. 74/136) and both situations occur 32.2% of the times. This association rule was the one that has computed the highest *support* value for the pregnancy outcome consequent attribute.	**Morphology at 3 months > 0% -> Pregnancy outcome=Yes** (p<=0.01 for n=230) *support*=0.322, *confidence*=0.544, *lift*=1.170, *Conviction*=1.173, x^2=8.38	Association Rule modeling step 3 (result summarized in Table 5.62)
7	5	Data pattern for Sperm parameters	**The conditional probability of observing 3 months after the embolization treatment a sperm morphology greater than 0% given a sperm progressive motility before the embolization treatment also greater than 0% is of 65.4%** (i.e. 119/182) and both situations occur 51.7% of the times. This association rule was the one that has computed the highest *support* value among the objectively and subjectively interesting rules generated in this study.	**Progressive Motility before treatment>0% -> Morphology at 3 months > 0%** (p<=0.01 for n=230) *support*=0.517, *confidence*=0.654, *lift*=1.106, *Conviction*=1.181, x^2=14.11	Association Rule modeling step 3 (result summarized in Table 5.62)
8	5	Data pattern for Sperm parameters	**The conditional probability of observing a sperm progressive motility before the embolization treatment greater than 0% given a low severity grade of the varicocele condition is of 86.6%** (i.e. 58/67) and both situations occur 19.8% of the times. This association rule was the one that has computed the highest *confidence* value among the objectively and subjectively interesting rules generated in this study.	**Severity grade=I -> Progressive Motility before treatment>0%** (p<=0.05 for n=293) *support*=0.198, *confidence*=0.866, *lift*=1.153, *conviction*=1.855, x^2=6.13	Association Rule modeling step 1 (result summarized in Table 5.62)
9	5	Data pattern for Sperm parameters	**The conditional probability of having a normal sperm concentration 3 months after the treatment given at the same**	**Morphology at 3 months = 1 to 3 -> Concentration at 3 months = > 15** (p<=0.01 for n=230) *support*=0.170, *confidence*=0.639, *lift*=1.564, *conviction*=1.64, x^2=18.32	Association Rule modeling step 6

Result ID	Goal Nº	Task Description	Best elected Result	Interesting Related outcome	Data mining Step
			time an abnormal sperm morphology is of 63.9% (i.e. 39/61) and both situations occur 17% of the times. This association rule was the one that presented the highest *confidence* for rules related with discretized sperm parameter values.		(result summarized in Table 5.65)
10	5	Data pattern per conception types	**61.7% of the pregnancies were carried out with an ART procedure** (*confidence*= 66/107=0.617) and **45.8% were achieved spontaneously** (*confidence*= 49/107=0.458).	**Pregancy outcome=Yes -> ART=Yes** (p<=0.01 for n=230) *support*=0.287, *confidence*=0.617, *lift*=2.15, *conviction*=1.861, x^2=106.47 **Pregancy outcome=Yes -> Spontaneous pregnancy =Yes** (p<=0.01 for n=230) *support*=0.213, *confidence*=0.458, *lift*=2.15, *conviction*=1.452, x^2=71.58	Association Rule modeling step 6 (result summarized in Table 5.66)
11	5	Data pattern for Spontaneous conceptions	**The biggest identified groups of couples (i.e. support > 0.104) that conceived spontaneously all had 3 months after the embolization treatment at least one of the sperm parameter values categorized as normal or a moderate varicocele condition**. However, the concentration at 3 months was seen to not be statistically significant. In the next column, all generated association rules for the consequent attribute Spontaneous pregnancy with a p<=0.10 are disclosed.	**Severity grade = II -> Spontaneous pregnancy =Yes** (p<=0.10 for n=107) *support*=0.252, *confidence*=0.562, *lift*=1.228, *conviction*=1.239, x^2=3.82 **Morphology at 3 months = > 4 -> Spontaneous pregnancy =Yes** (p<=0.10 for n=107) *support*=0.224, *confidence*=0.517, *lift*=1.248, *conviction*=1.265, x^2=3.58 **Progressive Motility at 3 months = > 32 -> Spontaneous pregnancy =Yes** (p<=0.05 for n=107) *support*=0.262, *confidence*=0.560, *lift*=1.223, *conviction*=1.232, x^2=3.95	Association Rule modeling step 6 for 107 instances (result summarized in Table 5.67).
12	5	Data pattern for ART conceptions	Among the 173 patients that we were able to know their smoking habits, 45.66% (79/173) smoked and quite half of them, (i.e. 49.37% (39/79)) were able to get pregnant. However, the conditional probability of a woman getting pregnant with an ART procedure given a male	**Preprocessed smoking habit =Yes -> ART=Yes** (p<=0.05 for n=107) *support*=0.28, *confidence*=0.769, *lift*=1.247, *conviction*=1.66, x^2=6.01	Association Rule modeling step 6 for 107 instances (result summarized in Table 5.67).

Result ID	Goal Nº	Task Description	Best elected Result	Interesting Related outcome	Data mining Step
			partner smoker is of 76.9% (30/39) and both situations occur 28% of the times. In other words, **man patients that smoke are more successful to conceive with the help of an ART procedure than spontaneously** (i.e. 76% (30/39) vs 33.33 (13/39)).		
13	5	Data pattern for ART conceptions	**The conditional probability of a woman getting pregnant with an ART procedure given a partner with an abnormal sperm morphology is of 78.1%** (25/32) and both situations occur 23.4% of the times.	**Morphology at 3 months =1 to 3 -> ART=Yes** (p<=0.05 for n=107) *support*=0.234, *confidence*=0.781, *lift*=1.267, *conviction*=1.752, x^2=5.24	Association Rule modeling step 6 for 107 instances (summarized in Table 5.67).

Result ID	Goal Nº	Task Description	Best elected Result	Interesting Related outcome	Data mining Step
14	5	Data pattern for varicocele´s patients	When we have non-missing values under the patient´s ages, the pregnancy outcome, the putative hazardous occupations of the male patient, the sperm concentration at 3 months, the sperm progressive motility before and 3 months after the embolization treatment, as well as under the varicocele´s severity grade, we are able to conclude the following: **Infertile male patients with a high varicocele severity grade rarely conceive and that the relative frequency of patients with normal sperm concentrations 3 months after the varicocele embolization and normal sperm progressive motility before the treatment is higher in clusters where fewer male patients work in putative hazardous occupations**. In the next column, the series plot with its complementary centroid table can be seen.	 	K-means modeling step 2 with ANOVA (series plot from Figure 5.36 and complementary centroid table from Table 5.46)

	Cluster 0 n=21	Cluster 1 n=38	Cluster 2 n=14	Cluster 3 n=12	*P*
Man age	36.524 (+-5.006)	35.947 (+-5.266)	36.643 (+-4.236)	37.500 (+-5.248)	0.778
Woman age	33.762 (+-4.049)	32.789 (+-4.400)	34.357 (+-4.517)	34.833 (+-3.927)	0.336
Pregnancy outcome	0.476 (+-0.512)	0.526 (+-0.506)	0.571 (+-0.514)	0.083 (+-0.289)	0.030
Hazardous occupation	0.429 (+-0.507)	0.158 (+-0.370)	0.071 (+-0.267)	0.500 (+-0.522)	0.007
Concentration category at 3 months	0.238 (+-0.436)	0.605 (+-0.495)	1 (+-0)	0.333 (+-0.492)	0.001
Progressive Motility category before treatment	0.048 (+-0.218)	0.342 (+-0.481)	0.929 (+-0.267)	0.167 (+-0.389)	0.001
Progressive Motility category at 3 months	0.286 (+-0.463)	0.342 (+-0.481)	0.786 (+-0.426)	0.417 (+-0.515)	0.011
Severity grade I	1 (+-0)	0 (+-0)	0.786 (+-0.426)	0 (+-0)	0.001
Severity grade II	0 (+-0)	1 (+-0)	0 (+-0)	0 (+-0)	0.001
Severity grade III	0 (+-0)	0 (+-0)	0.214 (+-0.426)	1 (+-0)	0.001

During data exploration of sperm parameter values, we have seen that sperm concentration and sperm morphology have their widest data dispersion before varicocele embolization (i.e. biggest max-min difference seen in Table 5.5 and Table 5.7, respectively) and that for sperm progressive motility it occurs 3 months after the varicocele embolization (seen in Table 5.6). These data dispersions have raised the hypothesis that the varicocele embolization might influence sperm parameter values through time; and hence, we have then assessed its evolution through the patient´s follow up times. During this assessment we have seen, with the series plot depicted as result id 2 in Table 5.70, that mean values have improved and reached the highest value at 3 months for sperm concentration (ANOVA $p<0.05$) and at 6 months for sperm morphology (ANOVA $p<0.05$). These improvements were coherent with the ones obtained in the result id 1, where the biggest semen classification at 3 months following treatment is normozoospermia. In fact, all sperm parameter mean values at 3 months were close to normality. However, in the following follow up times azoospermia becomes the leading semen classification which goes with the visual depletion seen in the result id 2 after its peak values at 3 and 6 months after treatment. Nevertheless, if we analyze the sperm parameter mean values at 12 months, we see that the panorama is not as pessimistic. In fact, considering the result id 2, we see that sperm concentration and sperm progressive motility mean values remain even higher at the 12 months follow up time than before varicocele embolization. Moreover, the sperm concentration mean value even went, on average, from an abnormal to a normal state based on the WHO thresholds (WHO, 2010); i.e., Concentration before treatment=13.93 vs Concentration at 12 months=15.78. However, the sperm morphology mean value showed a worse value than before the treatment at 12 months, which can be justified by the small number of filled values under this attribute (i.e. Morphology at 12 months has an n=23). A particularity of the sperm morphology mean values is that, even before the treatment, its values were always seen in average normal based on the WHO thresholds (i.e. Morphology before treatment=4.06). Hence, these results show that these embolized infertile men only have on average abnormal sperm concentrations and sperm progressive motility before the treatment and that the varicocele embolization improves all sperm parameter mean values with a statistical significance for sperm concentration and morphology. If we consider the varicocele embolization review carried out in Makris *et al*. (2018), we see that all 11 studies showed sperm parameter improvements, have reported a statistically significant improvement in sperm concentration and/or motility which goes along with our findings. Regarding the wide group of patients seen with azoospermia in the result id 1 after the 3 months follow up time, it can be explained by the fact that it only considers patients with all its sperm parameters filled and that most patients that are asked to repeat its semen analysis have azoospermia, which shows a version of the story that is not close to the reality expressed through the result id 2.

If we assess the sperm concentration as related to pregnancy outcome (section 5.3.2.4), we see that sperm concentration at 6 months after treatment is, on average, significantly greater for patients that were able to conceive; i.e., the mean value of the concentrations at 6 months for pregnancy outcome equal to *Yes* is of 22.86 million/ml vs 14.75 million/ml for pregnancy outcome equal to No. Therefore, this attribute was considered to mine (see selected attributes in Table 5.40), as well as the sperm morphology at 3 months after treatment which was seen in the result id 6 and 7 related with the association rules that have computed the highest *support* measure. In another perspective, 17.14% of the successful couples conceived at 9 (8.57%) or 12 (8.57%) months after the embolization treatment with a statistically significant difference in the sperm concentration mean value at 12 months; i.e. ANOVA p<0.05 for the concentration at 12 months with a mean value of 13.46 million/ml (Table 5.19). This sperm concentration mean value might be considered low but it enabled a successful ART procedure. In fact, 73.08% of the pregnancies carried out at 12 months after treatment were achieved with an ART procedure (see Figure 5.26).

Regarding the laterality of varicocele, most patients had the condition on the left testicle (178/218=81.65%) which is coherent with the statement in Makris *et al*. (2018) and similar to the 80.2% of left sided varicoceles identified in DeWitt *et al*. (2018). In the varicocele overview performed by Aza Mohammed and Frank Chinegwundoh (2009), the high incidence of the condition on the left testicle is due to its spermatic vein configuration since blood flow is drained at a right angle - rather than obliquely as on the right side - which causes a higher blood pressure. Concerning its relationship with other patient features, we have seen that it was not related with the semen classification (Table 5.31), or severity grade (Table 5.30) since they both have computed a *Chi-square* p>0.05.

Concerning varicocele severity grade, we have identified that it was related with the pregnancy outcome (*Chi-square p* = 0.049) and that severity grade III could even be used to classify patients where embolization was unsuccessful, since the K-means algorithm shows that they rarely conceive (cluster 3). In fact, only 3 patients out of the 15 patients that had a high severity grade among the 85 clustered patients have conceived - 2 out of these 3 patients were included in cluster 2 and the remaining patient in cluster 3 due to low pregnancy success. If we validate this last affirmation by assessing all our 230 preprocessed instances, we see that the conditional probability of conceiving given a severity grade III is of 28.57% (8/28), given a severity grade II, is 55.81% (48/86), and given a severity grade I, is 48.33% (29/60). Hence, patients with a severity grade III have a lower *support* and *confidence* than the other severity grades to conceive and these proportions are similar to the clustered data set. However, for patients with a high severity grade, the *confidence* in the clustered data set was slightly lower (i.e. 3/15=20.00% vs 8/28=28.57%). Nevertheless, this aspect was not an impediment to find interesting results since we have seen that cluster 3, that only has patients with the severity grade III, has the highest relative frequency of patients that work in hazardous occupations (50%), as well as the lowest relative frequency of pregnancies (8.3%) which was statistically significant (ANOVA *p*<0.05) in comparison to the other relative frequencies computed in each cluster and attribute. We should stress that the clustered data set only encompasses filled attributes; and hence, the conclusions related with the K-Means and the Decision Tree best results, disclosed in Table 5.70 as result id 4, 5 and 14, are related with non azoospermic patients.

Furthermore, during the application of the RapidMiner´s decision tree, we have also seen that the most interesting model has used the severity grade attribute to predict the success of varicocele embolization (result id 5 in Table 5.70) with an *Accuracy* and *F-measure* of 70.59%, as well as an *AUC* of 0.750 during the model validation. In fact, from this decision tree we could estimate the following conditional probability: the probability of conceiving given a couple with a man without a high varicocele severity grade and a woman older than 24 and younger 33 years old, inclusively, is of 70.83% (17/17+7). The interpretation of result id 5 enable us to say that woman age is overridden by a high varicocele severity grade when the woman is in a more fertile age. In contrast, for older women, the severity grade is no longer important to predict the pregnancy outcome which is mainly negative, likely due to the prevalence of a female factor. The cases under or equal to 24 years old can be discarded due to the the fact that there are only 2 patients in this age range.

Moreover, the severity grade also appeared in the association rule that has computed the highest *confidence* value (result id 8), as well as among the rules that characterizes spontaneous pregnancies (result id 11). Regarding the negative events caused by a high severity grade (i.e. grade III), other authors (Aza Mohammed & Frank Chinegwundoh, 2009) show that higher the severity grade are directly correlated with higher testicular volume reduction in adolescents. This situation can be seen as a negative clinical event since the lost of testicular volume in adolescents with varicocele was associated with a decrease on sperm concentration (Haans, Laven, Mali, te Velde, & Wensing, 1991); and hence, our idea that an high severity can be related to a negative events is also indirectly confirmed by previous literature.

Another attribute in the elected predictive models (result id 4 and 5) was woman age. During the 2 test designs, the attribute that delivered the highest gain of information was woman age with a splitting value at 33 years old in both test designs. In fact, woman age is the tree root in both approved results. If we cross this splitting value with the women age in Table 5.59, we see that the age range 1 and 2 - that includes women until 32 years old - clearly has a higher frequency of patients that were able to conceive. In contrast, this panorama completely changes at older female ages. Despite the splitting value of the decision tree being one year above of what it is depicted in Table 5.59, it enabled us to validate/support the decision tree results; and hence, say that the decision tree generated is coherent with what we have previously seen. Please note that the one-year difference can be justified with the fact that the result id 4 and 5 is computed in a smaller sample of data (i.e. 85 instances vs 229 assessed in Table 5.59). Clinically, this splitting age seemed acceptable since the provided and preprocessed data set only encompasses infertile couples. Furthermore, this splitting age is not far from the 35 years old indicated where women are have an accentuated drop in oocyte quality, which negatively influences pregnancy success (Williams & Alderman, 2001). Since the focus of this study is male infertility, we have contextualized this information and have seen that there is a statistically significant difference between male patient partner age means for whether they conceive or not (ANOVA *p*=0.018). Regarding male patient age, we have not seen a relationship with the success of the treatment (i.e. the pregnancy outcome attribute; ANOVA *p*=0.752). This result was expected since Figure 5.2 shows that pregnancies appeared homogenously through the scatter plot which was confirmed in Table 5.34, where male patient age mean was roughly at 34 years old for both pregnancy outcomes. Regarding the male age values, the age range went from 23 to 54 years old in the final preprocessed data set (Table 5.3) which is a different age range than the one specified in the McGraw-Hill Concise Dictionary that states that the higher incidence of varicocele is between 15 and 25 years old (disclosed in section 2.2.1). Nevertheless, this difference can be explained by the fact that our male patients are from a population that aim to achieve parenthood, in contrast to the medical dictionary that aims to depict the condition on the overall population. During the study of related works in male infertility (Williams & Alderman, 2001) greater pregnancy rates in patients with younger female partners. In fact, these authors (Williams & Alderman, 2001) state that the pregnancy rate for woman under 30 years old is of 15.8%, and above 40 years old is 0% in the context of couples undergoing donor sperm insemination. Furthermore, patient partner age was also identified as the first modeling attribute in other study (Guh et al., 2011). However, this work did not assess male patient age since its focus is the construction of a decision tree that predicts IVF fertilization with donor sperm.

Through our analysis of the varicocele severity grade, another interesting attribute arose during K-means application: the putative hazardous occupation of the male patient. In fact, the relative frequency of patients with normal sperm concentrations 3 months after varicocele embolization and normal sperm progressive motility before the treatment is higher in clusters where fewer male patients work in putative hazardous occupations (result id 14). If we take a closer look into this result, we see that cluster 1 and 2 have a lower relative frequency of patients that work in hazardous occupations (i.e. cluster 1 = 15.8% and cluster 2 = 7.1% vs cluster 0 = 42.9% and cluster 3 = 50% with an ANOVA $p = 0.007$) and that these same clusters are related with the highest relative frequencies of patients with normal sperm concentrations 3 months after varicocele embolization (i.e. cluster 1=60.5% and cluster 2= 100% vs cluster 0=23.8% and cluster 3=33.33% with an ANOVA $p = 0.001$), as well as the highest pregnancy rates, above 50% (i.e. cluster 1=52.6% and cluster 2= 57.1% vs cluster 0=47.6% and cluster 3=8.3% with an ANOVA $p = 0.030$). Furthermore, cluster 2 has the lowest relative frequency of patients with a putative hazardous occupation (i.e. 7.1%); and only has patients with normal sperm concentrations 3 months after the treatment. Similarly, the same pattern is seen for sperm progressive motility before the treatment (cluster 1=34.2% and cluster 2= 92.9% vs cluster 0=4.8% and cluster 3=16.7% with an ANOVA $p = 0.001$). If we assess the relation between the pregnancy outcome and the hazardous occupation attribute, we see that there is in fact a statistically significant relationship between these two attributes (Chi-square p=0.023) which supports the results obtained. This relation has indirectly already raised interest among the scientific community since other authors have studied the relation of some toxic components, such as the exposure to metals, with reproductive hormone levels (Wang et al., 2016).

When considering other factors, such as drinking alcohol or smoking, there was no statistically significant relationship with pregnancy outcome (depicted in Table 5.40), as well as between the drinking/smoking habit and semen classifications (section 5.3.2.9 and 5.3.2.10 respectively). Concerning patient occupation, we have considered it predictable due to the low data stability seen under the occupation attribute. In fact, this was the reason behind the transformation into an attribute that could record putative hazardous occupations to promote a higher information gain from the pregnancy outcome attribute. As expected, this transformation has increased data stability (i.e. occupation stability=3.96% vs hazardous occupation stability=63.86%) which enabled the identification of an interesting knowledge discovery (e.g. result id 14). In the result id 14 putative hazardous occupations are related with normal sperm concentration, which was only shown via data mining. Other authors (Delavar et al., 2014) h identified a higher prevalence of varicocele in smokers, however, they did not find a significant difference regarding the occupation or the alcohol drinking habits. We recall that in (Delavar et al., 2014), they compare patients with and without varicocele , which is different from this study that only assesses infertile patients that have undergone varicocele embolization. Hence, although relevant for the discussion, the results of (Delavar et al., 2014) cannot be directly compared with ours.

In summary, the results obtained are in line with the literature and made sense to the clinical experts (i.e. the BRSC team). However, as far as we know, this study cannot be totally compared with other studies. In fact, even though the varicocele condition is widely researched with *inferential statistics*, it was never studied with data mining techniques, which, as we have shown, potentiate knowledge discovery; and hence, enabled us to contribute with different and interesting findings as those in result id 5, 11 and 14. Therefore, this study not only contributes to the ongoing research of the BRSC team, but also serves as a basis to research other infertility-related matters. Moreover, the issues raised in this project that made us seek for a comprehensive data analysis to meet the data mining goals set such as joining the most popular data mining techniques applied in healthcare to the commonly used statistical tests. In fact, this enabled us to potentiate knowledge discovery along with the use of RapidMiner operators (e.g. the generate attribute operator and the optimize parameter operator). Therefore, we believe that this is what differentiates our work from most data mining works (see some in section 3.2) that usually only apply one technique, mostly the classification technique, and focus model choice on the performance measures which should, in our point of view, complement the seeking of interesting results since knowledge discovery is the essence of data mining.

Furthermore, since 68 features were needed to validate/generate the possible label attributes, but not directly assessed due to the domain of this study, we could use these male patient partner attributes to, for instance, research female infertility. Hence, our final data set could be used in other clinical projects, which usually struggle with data collection. In fact, data is a very valuable asset in this field, since healthcare data is usually incomplete (Tomar & Agarwal, 2013) and time-consuming to collect, as indeed we have experienced.

Chapter 6 Conclusions and Future Work

Using statistical and data mining techniques this study has analyzed a data set of 293 embolized varicocele infertile male patients in the *Centro Hospitalar e Universitário de Coimbra* described with 64 patients features (e.g., man partner age, varicocele severity grade, hazardous occupation and sperm parameter values collected before and after the treatment). More precisely, it has ascertained the success of the varicocele embolization with the Chi-square and ANOVA inferential statistical tests; predicted its success through the pregnancy outcome with the RapidMiner decision tree algorithm and the W-J48 java implementation of the C4.5 algorithm; and then identified interesting data patterns with the K-Means and FP-Growth algorithm which enabled the formulation of the following data analysis conclusions:

- Varicocele embolization improves sperm concentration mean values up to one year after the treatment and sperm morphology up to 6 months, which positively influences pregnancy success (mostly related with result id 1 and 2);
- A severity grade from low to moderate is mainly related with positive events; i.e., having a sperm progressive motility above 0% before the treatment (result id 8) and spontaneously conceiving (result id 11). In contrast, a high severity grade is related with negative events; i.e., rarely conceives (result id 14);
- Although varicocele embolization success is not related with the male patient age, it is however related with male patient partner age, decreasing when the partner is 33 years old or older (result id 4 and 5);
- On the context of non azoospermic patients, the varicocele embolization is more efficient on patients that do not work in putative hazardous occupations and sperm concentration values are more prone to normalize 3 months after the treatment for these patients (result id 14).

These findings were seen relevant to clinical experts and contributed to on-going research not only in the male infertility domain, but also in the knowledge discovery domain since it enabled us to identify measures that can potentiate the discovery of interesting results in similar data sets.

In fact, regarding the collection of data, recording it through an information system technology is the best method. However, these systems must also be thought for data analysis purposes to fully enhance knowledge discovery. Actually, through access of some of the CHUC information systems, we have formulated the following recommendations to improve knowledge discovery:

- The most relevant attributes for research purposes must have its corresponding fields set as mandatory and when possible, filled with the aid of an item selection box.
- Information systems should minimize the risk of inserting incoherent values by, for instance, firstly asking for the semen classification, and, if azoospermia is selected, automatically lock the remaining sperm parameter fields by automatically include the value 0 in all of them.
- Incorrect value formats must require correction to proceed with the form.

Based on the implementation of these recommendations, we believe that we would improve data quality; and hence, the possibility of achieving even more interesting results with clinical data.

Concerning the data mining application, the application of inferential statistical tests was seen useful to identify attributes that are more related with the label attribute when the Pearson correlations were low, as well as ascertain whether the predictive goal makes sense. Furthermore, we have seen that the application of several data mining techniques such as classification, clustering an association, provides greater knowledge discovery, even if the aim is to only predict a specific outcome, since it not only helps to guide the modeling process, but also enables to validate the knowledge that is found by comparing its different perspectives. Additionally, joining K-Means with the ANOVA statistical test, as well as the FP-Growth algorithm with the Chi-square statistical test, helps to identify results that are also objectively interesting. Moreover, the application of different test designs where models are also assessed with different performance measures such as the F-measure, Accuracy and AUC, minimizes the risk of selecting misleading models. In this context, we have identified that the following measures improves knowledge discovery:

- Follow the CRISP-DM methodology;
- Fill/validate the provided data;
- Collect more data;
- Use feature selection techniques that are mostly use in the studied domain;
- Apply the most commonly applied data mining techniques in the studied domain even if the aim is to only predict an outcome;
- Optimize the training of the models when possible;
- Follow a 3-parts test design;
- Not only focus on the performance of the models but also on its interestingness.

All these measures are in our point of view important contributions for further data mining projects in the healthcare field, since healthcare data sets are commonly known to be difficult to mine due to their characteristics.

As initially said, data mining techniques were not until know applied to the varicocele condition in spite of its high prevalence in the male infertility domain; and therefore, thanks to the data mining´s inductive capabilities, we were able to identify overlooked couple´s features that potentiate the success of the varicocele embolization and hence, the infertility treatment. In fact, our findings enabled us to not only validate with literature that the varicocele embolization can significantly increase sperm concentration (p<0.05) and morphology (p<0.05) but also, alert the clinical community that the varicocele severity grade, as well as the occupation of the male patient should require further investigations by the biologists and attention by the physicians since they were not until know very studied in related works. Therefore, we can say that this work contributes and leverage the findings in the global field of male infertility due to its innovative approach.

Regarding the obtained model performances, we have seen that the Decision tree model has surpassed the 73% of Accuracy seen in the similar related work of Guh et al. (2011) by reaching an accuracy of 80.77% within a similar test design that does not validate the model. However, within a test design that validates it, we have obtained an Accuracy and F-Measure of 70.59% with an AUC of 0.750 which was seen acceptable for a model validation. Since the domain-interestingness of the identified knowledge discoveries are usually not the central point in data mining-related works where model´s performance evaluations take the lead, this work shows the opposite approach raised from the need of achieving the data mining goals set. Since similar works would require the same approach, we have also sought to identify the tasks/measures that could potentiate the discovery of interesting findings. Since we have seen during the validation of the elected models by the BRSC teams that we had successfully reached the data mining goals set, we have seen that our identified measures could be useful to other data scientists.

This novel application enabled us to publish and present our most interesting results in the 53rd Annual Scientific Meeting of the European Society for Clinical Investigation that was published in the European Journal of Clinical Investigation (see Appendix D). Furthermore, we are actually in the submission process of the final paper of this work that can be seen in the Appendix E.

As future work, we would like to further explore the identified knowledge discovery measures with more data mining algorithms, as well as on other healthcare data sets, to formulate a good data mining practice guide for this field, since they contribute towards research that could improve our lives.

References

Agarwal, A., Mulgund, A., Hamada, A., & Chyatte, M. R. (2015). A unique view on male infertility around the globe. *Reproductive Biology and Endocrinology : RB&E*, *13*, 37. https://doi.org/10.1186/s12958-015-0032-1

Ahmad, P., Qamar, S., Qasim, S., & Rizvi, A. (2015). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications*, *120*(15), 975–8887. Retrieved from https://pdfs.semanticscholar.org/8228/9448146b86c6160bf5225bd5e3cea35a8c57.pdf

Al-odan, H. A., & Saud, A. A. A. K. (2015). Open Source Data Mining Tools. In *1st International Conference on Electrical and Information Technologies ICEIT'2015 Open* (pp. 369–374). https://doi.org/10.1109/EITech.2015.7162956

Almeida, P., & Bernardino, J. (2016). A survey on open source data mining tools for SMEs. *Advances in Intelligent Systems and Computing*, *444*, 253–262. https://doi.org/10.1007/978-3-319-31232-3_24

Almeida, P., Gruenwald, L., & Bernardino, J. (2016). Evaluating Open Source Data Mining Tools for Business. *Proceedings of the 5th International Conference on Data Management Technologies and Applications*, (Data), 87–94. https://doi.org/10.5220/0005939900870094

Alvarez, S. A. (2003). *Chi-squared computation for association rules: preliminary results*. Retrieved from http://www.cs.bc.edu/~alvarez/ChiSquare/chi2tr.pdf

Arif, C., Kotoulas, K., Georgellis, C., Frigkas, K., Bantis, A., & Patris, E. (2018). Two Case Reports of Varicocele Rupture during Sexual Intercourse and Review of the Literature. *Case Reports in Urology*, *2018*, 1–6. https://doi.org/10.1155/2018/4068174

Aza Mohammed, & Frank Chinegwundoh. (2009). Testicular Varicocele: An Overview. *Urologia Internationalis*. https://doi.org/10.1159/000218523

Azevedo, P. J., & Jorge, A. M. (2007). Comparing Rule Measures for Predictive Association Rules. In *European Conference on Machine Learning*. Retrieved from https://link.springer.com/chapter/10.1007/978-3-540-74958-5_47

Barbara Ilowsky; Susan Dean. (2017). *Introductory statistics*. OpenStax, Rice University. Retrieved from https://d3bxy9euw4e147.cloudfront.net/oscms-prodcms/media/documents/Statistics-LR.pdf

Begum, H. (2013). Data Mining Tools and Trends – An Overview. *International Journal of Emerging Research in Management &Technology*, *ISSN*, 2278–9359.

Bidgoli, A. A., Komleh, H. E., & Mousavirad, S. (2015). Seminal Quality Prediction using Optimized Artificial Neural Network with Genetic Algorithm. In *Conference: 9th International Conference on Electrical and Electronics Engineering(ELECO 2015)At: Bursa, Turkey*. https://doi.org/10.1109/ELECO.2015.7394596

Bilreiro, C., Donato, P., Costa, J. F., Agostinho, A., Carvalheiro, V., & Caseiro-Alves, F. (2017). Varicocele embolization with glue and coils: A single center experience. *Diagnostic and Interventional Imaging*, *98*(7–8), 529–534. https://doi.org/10.1016/j.diii.2017.01.006

Brin, S., Motwanit, R., & Silverstein, C. (1997). *Beyond Market Baskets: Generalizing Association Rules to Correlations*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.9473&rep=rep1&type=pdf

Brossette, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T., & Moser, S. A. (1998).

Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. *Journal of the American Medical Informatics Association*, *5*(4), 373–381. https://doi.org/10.1136/jamia.1998.0050373

Bruse, J. L., Zuluaga, M. A., Khushnood, A., McLeod, K., Ntsinjana, H. N., Hsia, T.-Y., … Schievano, S. (2017). Detecting Clinically Meaningful Shape Clusters in Medical Image Data: Metrics Analysis for Hierarchical Clustering applied to Healthy and Pathological Aortic Arches. *IEEE Transactions on Biomedical Engineering*, 1–1. https://doi.org/10.1109/TBME.2017.2655364

Çayan, S., & Akbay, E. (2018). Fate of Recurrent or Persistent Varicocele in the Era of Assisted Reproduction Technology: Microsurgical Subinguinal Redo Varicocelectomy Versus Observation. *Urology*, *0*(0). https://doi.org/10.1016/j.urology.2018.03.046

Cerquitelli, T., Baralis, E., Morra, L., & Chiusano, S. (2016). Data mining for better healthcare: A path towards automated data analysis? *2016 IEEE 32nd International Conference on Data Engineering Workshops, ICDEW 2016*, 60–63. https://doi.org/10.1109/ICDEW.2016.7495617

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.

Chen, C.-C., Hsu, C.-C., Cheng, Y.-C., & Li, S.-T. (2009). Knowledge Discovery on In Vitro Fertilization Clinical Data Using Particle Swarm Optimization. *Bioinformatics and BioEngineering, 2009. BIBE &#039;09. Ninth IEEE International Conference On*, 278–283. https://doi.org/10.1109/BIBE.2009.36

Decision Tree. (n.d.). Retrieved October 5, 2018, from https://www.saedsayad.com/decision_tree.htm

Delavar, M., Haydari, F., Mahdinejad, N., Abedi, S., Shafi, H., & Esmaeilzadeh, S. (2014). Prevalence of varicocele among primary and secondary infertile men: Association with occupation, smoking and drinking alcohol. *North American Journal of Medical Sciences*, *6*(10), 532. https://doi.org/10.4103/1947-2714.143285

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002

Deshpande, B. (2012). How to choose optimal decision tree model parameters in Rapidminer. Retrieved November 22, 2018, from http://www.simafore.com/blog/bid/107076/How-to-choose-optimal-decision-tree-model-parameters-in-Rapidminer

DeWitt, M. E., Greene, D. J., Gill, B., Nyame, Y., Haywood, S., & Sabanegh, E. (2018). Isolated Right Varicocele and Incidence of Associated Cancer. *Urology*, *0*(0). https://doi.org/10.1016/j.urology.2018.03.047

Dougherty, J., Kohavi, R., & Sahami, M. (1995). *Supervised and Unsupervised Discretization of Continuous Features*. Morgan Kaufmann Publishers. Retrieved from http://www.math.unipd.it/~dulli/corso04/disc.pdf

Exarchos, K. P., Goletsis, Y., & Fotiadis, D. I. (2012). Multiparametric Decision Support System for the Prediction of Oral Cancer Reoccurrence. *IEEE Transactions on Information Technology in Biomedicine*, *16*(6), 1127–1134. https://doi.org/10.1109/TITB.2011.2165076

Fawcett, T. (n.d.). How to evaluate classification models for business analytics - Part 2. Retrieved August 12, 2018, from http://www.simafore.com/blog/bid/57470/How-to-evaluate-classification-models-for-business-analytics-Part-2

Fayyad, U. M., & Irani, K. B. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In M. Kaufmann (Ed.), *Proceedings of the 13th International Joint*

Conference on Artificial Intelligence (pp. 1022–1027). Retrieved from https://www.ijcai.org/Proceedings/93-2/Papers/022.pdf

Geman, O., Chiuchisan, I., & Covasa, M. (2016). Data mining and knowledge discovery tools for human microbiome big data. *2016 6th International Conference on Computers Communications and Control, ICCCC 2016*, (Icccc), 91–96. https://doi.org/10.1109/ICCCC.2016.7496744

Gil, D., Girela, J. L., De Juan, J., Gomez-Torres, M. J., & Johnsson, M. (2012). Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*, *39*(16), 12564–12573. https://doi.org/10.1016/j.eswa.2012.05.028

Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., & Greene, C. S. (2016). Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics*, *17*(1), 33–42. https://doi.org/10.1093/bib/bbv087

Guh, R. S., Wu, T. C. J., & Weng, S. P. (2011). Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes. *Expert Systems with Applications*, *38*(4), 4437–4449. https://doi.org/10.1016/j.eswa.2010.09.112

Gui, H., Zheng, R., Ma, C., Fan, H., & Xu, L. (2016). An Architecture for Healthcare Big Data Management and Analysis (pp. 154–160). Springer, Cham. https://doi.org/10.1007/978-3-319-48335-1_17

Haans, L. C., Laven, J. S., Mali, W. P., te Velde, E. R., & Wensing, C. J. (1991). Testis volumes, semen quality, and hormonal patterns in adolescents with and without a varicocele. *Fertility and Sterility*, *56*(4), 731–736. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1915950

Han, J., Kamber, M., & Pei, J. (Computer scientist). (2012). *Data mining : concepts and techniques*. Elsevier/Morgan Kaufmann.

Hand, D. J. (1998). Data Mining: Statistics and More? *The American Statistician*, *52*(2), 112–118. https://doi.org/10.1080/00031305.1998.10480549

Hand, David, Blunt, G., Kelly, M., & Adams, N. (2000). Data Mining for Fun and Profit. *Statistical Science*, *15*(2), 111–131. https://doi.org/10.1214/ss/1009212753

Hand, Dj. (1999). Statistics and data mining: intersecting disciplines. *ACM SIGKDD Explorations Newsletter*, *1*(1), 16–19. https://doi.org/10.1145/846170.846171

IBM. (2010). IBM Knowledge Center - What is a data set? Retrieved August 4, 2018, from https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zconcepts/zconc_datasetintro.htm

international labour organization. (n.d.). Hazardous Work. Retrieved February 25, 2019, from https://www.ilo.org/safework/areasofwork/hazardous-work/lang--en/index.htm

Jović, A., Brkić, K., & Bogunović, N. (2014). An overview of free software tools for general data mining. *37th International Convention MIPRO …*, (May), 26–30. https://doi.org/10.1109/MIPRO.2014.6859735

Keller, J. J., Chen, Y.-K., & Lin, H.-C. (2012). Varicocele is associated with erectile dysfunction: a population-based case-control study. *The Journal of Sexual Medicine*, *9*(7), 1745–1752. https://doi.org/10.1111/j.1743-6109.2012.02736.x

Kerana Hanirex, D., Kaliyamurthie, D. K. P., & Kerana, D. (2015). AN ADAPTIVE TRANSACTION REDUCTION APPROACH FOR MINING FREQUENT ITEMSETS: A COMPARATIVE STUDY ON DENGUE VIRUS TYPE1. *Int J Pharm Bio Sci*, *6*(2), 336–340. Retrieved from www.ijpbs.net

Kirby, E. W., Wiener, L. E., Rajanahally, S., Crowell, K., & Coward, R. M. (2016). Undergoing varicocele repair before assisted reproduction improves pregnancy rate and live birth rate in azoospermic and oligospermic men with a varicocele: a systematic review and meta-analysis. *Fertility and Sterility*, (August), 3–8. https://doi.org/10.1016/j.fertnstert.2016.07.1093

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V, & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *CSBJ*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Kovács, F., Legány, C., & Babos, A. (n.d.). *Cluster Validity Measurement Techniques*. Retrieved from https://pdfs.semanticscholar.org/c4f9/df3c66105382d05e58ec35faa8d435f55c91.pdf

Kumar, P., & Wasan, S. K. (2010). Analysis of X-means and global k-means USING TUMOR classification. In *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)* (pp. 832–835). IEEE. https://doi.org/10.1109/ICCAE.2010.5451883

Lippincott Williams & Wilkins (Ed.). (2012). *Medical Dictionary for the Health Professions and Nursing*. Julie K. Stegman. Retrieved from https://medical-dictionary.thefreedictionary.com/_/cite.aspx?url=https%3A%2F%2Fmedical-dictionary.thefreedictionary.com%2Fembolization&word=embolization&sources=MillerKeane,wkMed,dorland,MGH_Med,wkHP,gem,evPod,vet,iMedix

Liu, B., Hsu, W., & Ma, Y. (1998). *Integrating Classification and Association Rule Mining*. Retrieved from www.aaai.org

Makris, G. C., Efthymiou, E., Little, M., Boardman, P., Anthony, S., Uberoi, R., & Tapping, C. (2018). Safety and effectiveness of the different types of embolic materials for the treatment of testicular varicoceles: a systematic review. *The British Journal of Radiology*, 20170445. https://doi.org/10.1259/bjr.20170445

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., de Mendonça, A., … Duin, R. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, *4*(1), 299. https://doi.org/10.1186/1756-0500-4-299

Math´s Fun. (2017). Normal distribution. Retrieved April 29, 2018, from https://www.mathsisfun.com/data/standard-normal-distribution.html

May, M., Taymoorian, K., Beutner, S., Helke, C., Braun, K. P., Lein, M., … Hoschke, B. (2006). Body size and weight as predisposing factors in varicocele. *Scandinavian Journal of Urology and Nephrology*, *40*(1), 45–48. https://doi.org/10.1080/00365590500407795

Maydanchik, A. (2007). *Data Quality Assessment*. Technics Publications, LLC.

Mierswa, I. (2012). Optimize selection, how to get the resulting best model? Retrieved November 23, 2018, from https://community.rapidminer.com/discussion/16851/solved-optimize-selection-how-to-get-the-resulting-best-model

Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(5), 431–443. https://doi.org/10.1002/widm.24

Minitab 18. (2017). What are independent samples? Retrieved April 28, 2018, from https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/tests-of-means/what-are-independent-samples/

Mirroshandel, S. A., Ghasemian, F., & Monji-Azad, S. (2016). Applying data mining techniques for increasing implantation rate by selecting best sperms for intra-cytoplasmic sperm injection treatment. *Computer Methods and Programs in Biomedicine*. https://doi.org/10.1016/j.cmpb.2016.09.013

Mishra, R., & Thakur, R. S. (2014). An Efficient Approach for Supervised Learning Algorithms Using Different Data Mining Tools for Spam Categorization. In *2014 Fourth International Conference on Communication Systems and Network Technologies* (pp. 472–477). https://doi.org/10.1109/CSNT.2014.100

Mohammadali Beigi, F., Mehrabi, S., & Javaherforooshzadeh, A. (2007). Varicocele in brothers of patients with varicocele. *Urology Journal*, *4*(1), 33–35.

Murray R. Spiegel; Larry J. Stephens. (n.d.). *Statistics*. (McGraw-Hill, Ed.) (Fourth). Schaum´s Outline Series. Retrieved from http://www.buders.com/UNIVERSITE/Universite_Dersleri/istatistik/statistics.pdf

Niederberger, C. (2015). Re: Infertility etiologies are genetically and clinically linked with other diseases in single meta-diseases. *Journal of Urology*, *194*(6), 1712. https://doi.org/10.1016/j.juro.2015.09.045

Orphanou, K., Dagliati, A., Sacchi, L., Stassopoulou, A., Keravnou, E., & Bellazzi, R. (2016). Combining Naive Bayes Classifiers with Temporal Association Rules for Coronary Heart Disease Diagnosis. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 81–92). IEEE. https://doi.org/10.1109/ICHI.2016.15

P. Aalam and T. Siddiqui. (2016). Comparative Study of Data Mining Tools used for Clustering. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 3971–3975). New Delhi.

Paul, R., & Hoque, A. S. M. L. (2010). Clustering medical data to predict the likelihood of diseases. In *2010 Fifth International Conference on Digital Information Management (ICDIM)* (pp. 44–49). IEEE. https://doi.org/10.1109/ICDIM.2010.5664638

Piatetsky, G. (n.d.). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved September 26, 2017, from http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved July 31, 2018, from https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Piatetsky, G. (2018). Data Science tools poll. Retrieved February 25, 2019, from https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

Rangra, K., & Bansal, K. L. (2014). Comparative Study of Data Mining Tools. *International Journal of Advanced Research in Computer Science and Software Engineering*, *4*(6), 2277–128.

RapidMiner. (2016). *RapidMiner Radoop 7 Operator Reference Manual*. Retrieved from https://docs.rapidminer.com/downloads/rapidminer-radoop-operator-reference.pdf

Sahoo, A., & Kumar, Y. (2014). Seminal quality prediction using data mining methods. *Technology and Health Care*, *22*(4), 531–545. https://doi.org/10.3233/THC-140816

Samplaski, M. K., Lo, K. C., Grober, E. D., Zini, A., & Jarvi, K. A. (2017). Varicocelectomy to upgrade semen quality to allow couples to use less invasive forms of assisted reproductive technology. *Fertility and Sterility*, *108*(4), 609–612. https://doi.org/10.1016/j.fertnstert.2017.07.017

Sharma, R., Singh, S., & Khatri, S. (2016). Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey. In *Second International Conference on Computational Intelligence & Communication Technology*. https://doi.org/10.1109/CICT.2016.142

Shouman, M., Turner, T., & Stocker, R. (2012). Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients. *School of Engineering and Information Technology, University of New South Wales at the Australian Defence Force Academy, Northcott Drive, Canberra ACT 2600*, (August 2014), 1–7. Retrieved from http://weblidi.info.unlp.edu.ar/worldcomp2012-mirror/p2012/DMI9007.pdf

Shukla, D. P., Patel, S., & Sen, A. (2014). *A Literature Review in Health Informatics Using Data Mining Techniques Keywords Data mining, frequent patterns, data mining techniques, medical data mining*. Retrieved from http://cmapspublic2.ihmc.us/rid=1P0PJQH3F-1WTS7CH-273X/Shukla et al. - 2014 - A literature review in health informatics using da.pdf

Singh, S., Liu, Y., Ding, W., & Li, Z. (2016). Evaluation of Data Mining Tools for Telecommunication Monitoring Data Using Design of Experiment. In *2016 IEEE International Congress on Big Data Evaluation* (pp. 283–290). https://doi.org/10.1109/BigDataCongress.2016.43

Skewed Distribution: Definition, Examples - Statistics How To. (n.d.). Retrieved April 25, 2018, from http://www.statisticshowto.com/probability-and-statistics/skewed-distribution/

Statistics Wikibooks.org. (2012). Retrieved from http://de.wikibooks.org/wiki/Benutzer:Dirk_Huenniger/wb2pdf.

Stilou, S., Bamidis, P. D., Maglaveras, N., & Pappas, C. (2001). Mining Association Rules from Clinical Databases: An Intelligent Diagnostic Process in Healthcare. *MEDINFO*.

Sullivan, M. (2011). *Fundamentals of statistics*. Prentice Hall.

Tape, T. G. (n.d.). The Area Under an ROC Curve. Retrieved December 10, 2018, from http://gim.unmc.edu/dxtests/roc3.htm

Ted Wrigley. (2016). What are the differences between the Kolmogorov-Smirnov test and the Mann-Whitney U test? - Quora. Retrieved April 30, 2018, from https://www.quora.com/What-are-the-differences-between-the-Kolmogorov-Smirnov-test-and-the-Mann-Whitney-U-test

Tekieh, M. H., & Raahemi, B. (2015). Importance of Data Mining in Healthcare. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, 1057–1062. https://doi.org/10.1145/2808797.2809367

Thatipamula, S. (2013). Data Done Right: 6 Dimensions of Data Quality (Part 1) - Smartbridge. Retrieved July 22, 2018, from https://smartbridge.com/data-done-right-6-dimensions-of-data-quality-part-1/

The Sample Proportion. (n.d.). Retrieved May 21, 2018, from http://www.stat.wmich.edu/s216/book/node68.html

Ting, S. L., Shum, C. C., Kwok, S. K., Tsang, A. H. C., & Lee, W. B. (2009). Data Mining in Biomedicine: Current Applications and Further Directions for Research. *J. Software Engineering & Applications*, *2*, 150–159. https://doi.org/10.4236/jsea.2009.23022

Ting, S. L., Wang, W. M., Kwok, S. K., Tsang, A. H. C., & Lee, W. B. (2010). RACER: Rule-Associated CasE-based Reasoning for supporting General Practitioners in prescription making. *Expert Systems with Applications*, *37*(12), 8079–8089. https://doi.org/10.1016/j.eswa.2010.05.080

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241–266. https://doi.org/10.14257/ijbsbt.2013.5.5.25

Varicocele Definition. (2002). Retrieved from https://medical-dictionary.thefreedictionary.com/varicocele

Varicocele Definition. (2009). In *Mosby's Medical Dictionary* (9th ed.). Retrieved from https://medical-dictionary.thefreedictionary.com/varicocele

Varicocele Definition. (2012). In *Farlex Partner Medical Dictionary*. Retrieved from https://medical-dictionary.thefreedictionary.com/varicocele

Visualizing the Confusion Matrix - Sanyam Kapoor. (n.d.). Retrieved August 12, 2018, from https://www.sanyamkapoor.com/machine-learning/confusion-matrix-visualization/

Wang, Y. X., Sun, Y., Huang, Z., Wang, P., Feng, W., Li, J., … Lu, W. Q. (2016). Associations of urinary metal levels with serum hormones, spermatozoa apoptosis and sperm DNA damage in a Chinese population. *Environment International*, *94*, 177–188. https://doi.org/10.1016/j.envint.2016.05.022

Williams, R. S., & Alderman, J. (2001). Predictors of success with the use of donor sperm. *American Journal of Obstetrics and Gynecology*, *185*(2), 332–337. https://doi.org/10.1067/mob.2001.116733

Witten, I. H. (Ian H. ., Frank, E., & Hall, M. A. (Mark A. (2011). *Data mining : practical machine learning tools and techniques*. Morgan Kaufmann.

WorldHealthOrganization. (2010). *WHO laboratory manual for the examination and processing of human semen*. *World Health Organization* (Fifth edit). World Health Organization. https://doi.org/10.1038/aja.2008.57

Wujek, B., Hall, P., & Güneș, F. (n.d.). Best Practices for Machine Learning Applications. Retrieved from https://support.sas.com/resources/papers/proceedings16/SAS2360-2016.pdf

Xu, W., Hu, H., Wang, Z., Chen, X., Yang, F., Zhu, Z., … Qiao, Z. (2012). Proteomic characteristics of spermatozoa in normozoospermic patients with infertility. *Journal of Proteomics*, *75*(17), 5426–5436. https://doi.org/10.1016/j.jprot.2012.06.021

Yan, L., Guo, W., Wu, S., Liu, J., Zhang, S., Shi, L., … Gu, A. (2014). Genetic variants in nitric oxide synthase genes and the risk of male infertility in a Chinese population: A case-control study. *PLoS ONE*, *9*(12), 1–14. https://doi.org/10.1371/journal.pone.0115190

Yang, L., Li, Z., & Luo, G. (2016). MH-ARM: A Multi-Mode and High-Value Association Rule Mining Technique for Healthcare Data Analysis. In *2016 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 122–127). IEEE. https://doi.org/10.1109/CSCI.2016.0030

Yeager, K. (n.d.). *Crosstabs*. Retrieved from https://libguides.library.kent.edu/SPSS/Crosstabs

Yildirim, P. (2015). Association Patterns in Open Data to Explore Ciprofloxacin Adverse Events. *Applied Clinical Informatics*, *6*(4), 728–747. https://doi.org/10.4338/ACI-2015-06-RA-0076

Zancanaro, M., Kuflik, T., Boger, Z., Goren-Bar, D., & Goldwasser, D. (2007). Analyzing Museum Visitors' Behavior Patterns. In *User Modeling 2007* (pp. 238–246). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73078-1_27[i]

Glossary

Since this study uses some knowledges of biology, statistics and computer science, this section aims to present the definition of some of the technical terms used during the discloser of the carried-out work from these domain knowledges.

To better present these definitions, we have grouped the technical terms by domain of knowledge and presented them in the following order: Biology, Statistics and Computer Science. For the Computer Science domain, we have described the technical terms used in the context of data mining and applied to the carried-out work to better explain them. Hence, in this section we also present some expressions that we have also used.

Biology

ART: Refers to any Assisted Reproductive Technology used to get a woman pregnant. These technologies can be: *IUI, IVF, ICSI* or *IMSI*.

Concentration: the percentage of spermatozoa in the semen.

ICSI: Intracytoplasmic Sperm Injection.

IMSI: Intracytoplasmic Morphologically Selected Sperm Injection.

IUI: Intrauterine Insemination.

IVF: In Vitro Fertilization.

Laterality: Testis with the varicocele condition (e.g. left, right or both testis)

Morphology: the percentage of spermatozoa with a normal shape in the semen.

Motility: the percentage of motile spermatozoa.

Semen Analysis Report: Report carried-out by a biologist to record the concentration, the morphology and the motility of the spermatozoa.

Semen: organic fluid that may contain spermatozoa.

Spermatozoa: sperm cells that aims to join an ovum to form a zygote that normally develops into an embryo.

Computer Science

In this section, we describe the main terms used to describe the carried-out work in the domain of Computer science:

Absolute Support: The number of instances in the data set (i.e. frequency) that contains an attribute or a set of attributes filled, in this study, with the value “TRUE” (RapidMiner´s definition).

Accuracy: It is a performance metric that also gave its name to the splitting criterion tested with the variation of the Decision tree´s “criterion” parameter that as the same name. By the RapidMiner platform, this criterion selects an attribute for splitting if it maximizes the performance metric accuracy of the whole tree. We here recall that the performance metric Accuracy is the proportion of instances classified correctly among the total number of instances.

Antecedent or Premises: In the context of *association*, the antecedent is the attribute that appears at the left side of the implication. In RapidMiner, it is called "Premises". For example, in a rule as (X -> Y), X is the antecedent or premise of the rule.

Association rule: Association found between attributes, read X implies Y and written (X->Y) that expresses that "if X occurs then Y occurs". The probability of this association is assessed through several metrics as *Support*, *Confidence*, *Lift* and *Conviction*, that are below described.

Attribute: column of the data set that describes, in our case, patients in one subject (e.g. their age, treatment date, sperm concentration etc. are all attributes).

Bagging: Machine learning ensemble meta-algorithm, also called "Bootstrap aggregating", that aims to improve classification in terms of stability and classification accuracy. Model´s stability is achieved through the reduction of model´s variance which helps to avoid data overfitting. Bagging works as follows: it repeatedly sub-samples the dataset (i.e. it sub-samples the data set the number of times specified under the iteration field of the Bagging operator) and applies the nested classifier (e.g. decision tree) upon the data ratio defined for training in the Bagging operator. In practice, very different trees are often grown from the different sub-samples which illustrates instability of models. The single generated prediction measures are obtained through simple voting. This voting elects as the final classification the one that is most often predicted by the different generated trees (RapidMiner´s definition).

Blank Value: Data value that appears empty in the EXCEL file but when exported to the RapidMiner platform, presents a *Null* value shown with a question mark "?".

Centroid table: Table generated after running the K-Means algorithm that presents the mean attribute values for each identified cluster. This artefact is important because through the mean of each attribute within each cluster, we can identify what differentiates each cluster in terms of patient's characteristics to further on identify a patient´s data pattern.

Classifier: A data mining algorithm that implements classification.

Conditional pattern-bases: Conditional pattern-bases are what the FP_Growth algorithm first generates when it mines the computed FP-tree. These conditional pattern-bases are all the paths that can lead to a specific attribute in the FP-tree. Usually it begins to identify the paths that leads to one of the FP-tree´s leaves and then it recursively identifies the paths that leads to the other nodes of the FP-tree by going up in the tree. These paths do not include the attribute that is being assessed, so if the algorithm is seeking for all the paths that leads to an attribute that is for instance a leaf, this leaf attribute will not be part of the identified path. This is why we call them "conditional".

Confidence: Confidence it is the **conditional probability,** written *Pr*, of observing Y given X (RapidMiner´s definition). This measure is defined as:

$$\text{confidence (X implies Y)} = \frac{support\ \text{(X U Y)}}{support\ \text{(X)}} = Pr(Y|X)$$

Formula Glossary 1 Association measure - Confidence (RapidMiner´s definition)

The "*support* (X U Y)" is the proportion of instances where the values of the attributes X and Y both appear, in this study, both with the value "TRUE". Note that the FP_Growth algorithm only manages binominal attributes; and therefore, the attributes´ values were transformed in this study with the RapidMiner´s operator called "Numerical to Binomial" and "Nominal to Binomial" to transform the filled attributes´ values into "TRUE" or "FALSE".

Since confidence is a conditional probability, the higher the confidence of a rule is, the better. In fact, a high confidence, let's say of 100% - in RapidMiner expressed by 1 - would mean in this study that each time a patient has both attributes filled with the value "TRUE", the attribute that is *antecedent* in the rule only occurs when both attributes are filled with the value "TRUE", so the rule has an high statistical strength.

The confidence measure ranges from 0 to 1.

Consequent or Conclusion: In the context of *association*, the consequent is the attribute that appears at the right side of the implication. In RapidMiner, it is called "Conclusion". For example, in a rule as (X -> Y), Y is the consequent or conclusion of the rule.

Conviction: Conviction attempts to measure the **degree of implication** of a rule since the generated rules are sensitive to the rule direction (RapidMiner´s definition). Conviction is defined as:

$$\text{conviction (X implies Y)} = \frac{(1\ -\ supp\text{ort (Y))}}{(1\ -\ confidence\ \text{(X implies Y))}}$$

Formula Glossary 2 Association measure - Conviction (RapidMiner´s definition)

As the *lift* measure, conviction values different than 1.0 indicates interesting rules. If we analyze the conviction formula we can say that the conviction is equal to the ratio of the probability of **not** occurring the consequent attribute over its conditional probability of not also occurring; and therefore, if we lower the conditional probability of **not** occurring Y given X, the conviction value gets higher which expresses a more interesting rule.

The conviction measure goes from 0.5 to positive infinity and a value equal to positive infinity indicates that the implication is logical (Azevedo & Jorge, 2007).

Correlation Similarity: Correlation between the Attribute vectors of the two Examples (RapidMiner´s definition).

Data Quality: The degree to which a set of data characteristics fulfills the domain requirements. For instance, data completeness is one of the data characteristics that shows the quality of a provided data.

Data set: A file that contains one or more records. A record is a basic unit of information that can be used by a program (IBM, 2010). A data set is organized in rows and columns, as our provided data set is, and records are the rows of the data set that are in the data mining universe named *Instances* or *Examples*.

Data validation: Checking the accuracy of the source/provided data prior data analyzes.

Data verification: Checking the consistency of the source/provided data prior data analyzes to assess the coherence between the attribute´s values.

Davies Bouldin: Index used for the assessment of the quality of identified clusters by clustering algorithms. This index measures the average of similarity between each cluster and its most similar one (Kovács, Legány, & Babos, n.d.). Since the aim of clustering is to partition a data set into similar data, we will try to minimize this index by testing the built models with several parameters and inputs to see if we can find a clustering model that can minimize this index and achieve the aims of the study. Hence, lower values are the ones that we seek.

Discretize: Convert a numeric continuous attribute into a numeric discreate attribute by creating several ranges of values (i.e. bins).

Entropy: Measure that indicates the state of confusion of a set of *instances* based on its *label attribute*. This measure calculates the proportion of values that are positive minus the ones that are negative and can be interpreted as the **expected information needed to classify an instance** in the partition *D* is given by the below formula where p_i is the nonzero probability that an arbitrary instance in *D* belongs to class C_i (Han et al., 2012).

$$Entropy\ (D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Formula Glossary 3 Entropy - based on (Han et al., 2012)

In our case, positive instances have the corresponding “Gravidez” attribute value set to “Sim” and negative instances, set to “Não”. Hence, when the entropy is equal to 0, it means that the data set is pure in the assessed attribute because all its instances are from the same class (i.e. all instances have the same value in the “Gravidez” attribute; and therefore, they all are equal to “Sim” or equal to “Não”). If the entropy is equal to 1, it means that half of the instances are positive, and the other half, negative, so the data set is totally “confused”.

Euclidean Distance: Square root of the sum of quadratic differences over all attributes (RapidMiner´s definition).

Example set*:* collection of rows of a *data set*.

Feature selection: Selection of attributes during the modeling process built in the RapidMiner platform.

Final preprocessed attributes*: attributes* of the *Final preprocessed data* set.

Final preprocessed data set*:* The data set that is already treated with data preparation tasks that includes the *initially preprocessed attributes* and the ones that were upon them generated which encompasses 293 instances and 64 attributes (39 *initially preprocessed attributes*, and 25, upon them generated).

Frequent item/attribute: In the context of the APRIORI and the FP_Growth algorithm, we say that a frequent item/attribute is an attribute that appears in the data base, in this study, with its value set to "TRUE" more than the defined *min_support*.

Gain ratio: Measure used by the C4.5 algorithm to compute the gain of information of each attribute to build its decision tree. This measure is an extension of the *information gain* measure and attempts to overcome its issue: biased toward attributes with many values. This measure applies a kind of normalization to the *information gain* previously presented with the "split information" value that can be seen at the denominator of the below gain ratio formula - the split information formula is in the next formula disclosed.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A\ (D)}$$

Formula Glossary 4 Gain Ratio - based on (Han et al., 2012)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * \log_2 \left(\frac{|D_j|}{|D|}\right)$$

Formula Glossary 5 Split information - based on (Han et al., 2012)

The split information value represents the potential information generated by splitting the training data set, D, into v partitions, corresponding to the v outcomes of a test on attribute A. Note that, for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D which is different than the information gain which measures the information with respect to classification that is acquired based on the same partitioning (Han et al., 2012).

Gini_Index: In this context, it is a splitting criterion tested with the variation of the Decision tree´s "criterion" parameter. By the RapidMiner platform, this criterion is a measure of inequality between the distributions of label characteristics.

Information gain: Measure used by the decision tree algorithm ID3 (predecessor of the C4.5 algorithm) to compute the nodes of its decision tree. The information gain measures the reduction of the entropy when it ramifies in each node (i.e. attribute). Note that a node with an entropy=0 tells us that the value of the information gain measure is high since we can say that all instances are positive or negative; and therefore, it generates "information" - this is why decision trees algorithms build its trees by reducing the entropy. The information gain is calculated with the below formula:

$$Gain(A) = Entropy(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Entropy(D_j)$$

Formula Glossary 6 Information Gain - based on (Han et al., 2012)

The term $\frac{|D_j|}{|D|}$ acts as the weight of the *j*th partition which is the number of instances that goes down at each branch of the assessed node and $Entropy(D_j)$ calculates the *entropy* of the attribute node value at the *j*th partition. Hence, the Information gain is defined as the difference between the original information requirement calculated with the entropy for the root/node (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A) (Han et al., 2012). Hence, what the algorithm does at each node before splitting is to see if the split in one attribute generates a gretter gain of information than splitting in another attribute values.

Initially preprocessed attributes: Attributes of the *initially preprocessed data set* which is the data set with the 39 attributes.

Initially preprocessed data set: The data set that is already treated with data preparation tasks but does not yet include the generated attributes.

Initially provided and selected attributes: Attributes selected from the *initially provided data set* to carry out this study. These attributes were with the BRSC team selected and without statistical tests or data mining algorithms. These attributes are also called in this study “Original attribute”.

Initially provided data set: The first data set provided by the BRSC team.

Instance/example: row of the data set that records, in our case, the information of one patient.

Key Data Dimensions: Key characteristics of the *data value*s of a data set in the matter of data quality.

Label Attribute: Special attribute role that can be set in the RapidMiner platform to an attribute to act as a target attribute for learning tasks, as classification. Label attributes are also often called “target variable” or “class” (RapidMiner, 2016).

Lift: Lift measures **how far from independence** are X and Y. Values equal to 1 imply that X and Y are independent; and hence, the rule is not interesting (RapidMiner´s definition). In fact, a lift equal to 1 tells that the conditional probability of Y given X is equal to Y occurring randomly so it is not an interesting rule. We seek rules with lifts higher than 1.0, as suggested in RapidMiner´s tutorials. The lift measure can be simplified as (RapidMiner´s definition):

$$\text{lift (X implies Y)} = \frac{confidence\ \text{(X implies Y)}}{support\ \text{(Y)}}$$

Formula Glossary 7 Association measure - Lift (RapidMiner´s definition)

By observing the above lift formula and by considering the previous measure´s definitions, we can in fact say that the lift is equal to the ratio of the conditional probability of the event Y given X over the probability of occurrence of the event Y.

By RapidMiner, the lift measure ranges within 0 to positive infinity where a value close to 1.0 implies that the attributes are independent so a value equal to positive infinity indicates that the attributes are totally dependent.

Manhattan Distance: Sum of the absolute distances of the Attribute values (RapidMiner´s definition).

Min_confidence: confidence initially set by the programmer to prune the results computed by the association algorithm.

Min_support: support initially set by the programmer to prune the results computed by the association algorithm. It therefore works as a threshold where we say that any attribute that is frequent, has a support higher than the defined *min_support*.

Regular Attribute: Attribute that does not have a special attribute role and is used as an input for learning tasks (RapidMiner, 2016).

Relative Support or Support : The *relative support* of a rule, or simply called *support*, is the proportion of instances in the data set that contains an attribute or a set of attributes filled, in this study, with the value "TRUE" (RapidMiner´s definition). In (Han et al., 2012), the authors describes the support as the probability of having a transaction with both X and Y events and in some algorithms it is called "coverage".

The support is indicated as "*support* ()" and can be interpreted in this study as the **probability** of a patient having an attribute set to "TRUE". For example, if we have a total of 100 patients and we have the information of if they smoke (recorded in the data set as "Sim") or not (recorded in the data set as "Não") in an attribute called X and for 30 of them, we have the information that they smoke since we have 30 "Sim" values under that attribute X, then the support of , written support(X), is equal to 30/100=0.30. Which means that in this dataset, the probability of having a patient that smokes is of 30%. The absolute support would be in this case 30 patients. Hence, this measure is defined as:

$$\text{support}\,(X, Y) = \frac{P\ (\mathrm{X\ U\ Y})}{N}$$

Formula Glossary 8 Association measure - Support , based on (Yildirim, 2015)

Since it is a probability, the higher the value of the support is, the better. The support measure ranges from 0 to 1.

Supervised learning: Data mining task that needs a *label attribute* to train to further on build a function that will classify new *instances*.

Unsupervised learning: Data mining task that describes the structure of a "unlabeled" data set. The drawback of these algorithms is that they cannot be assessed with performance metrics.

Value: In our case, it is the patient´s information gathered in the rows and columns of the data set.

VCF: work flow built in the RapidMiner platform that implements a data mining model. The VCF name stands for "visual composition framework".

Statistics

In this section, we describe all the statistical terms that were in this study used to describe the performed work.

Bar Graph: A Bar Graph consists 2 attributes, one dependent arranged in the y axis of a graph and the other one, independent arranged in the x axis ("Statistics Wikibooks.org," 2012). This graph is a set of horizontal or vertical bars that aims to show the frequencies of each value of an attribute through the length, for horizontal bar graphs, or the height, for vertical bar graphs, of its bars (Barbara Ilowsky; Susan Dean, 2017).

Box plot: A box plot is a type of graph that allows to visually see how data is distributed. This graph presents the following main points: the minimum value, the first quartile, the median, the third quartile and the maximum value of the data set (Barbara Ilowsky; Susan Dean, 2017). To assess the data distribution of sperm parameters, this plot was performed.

Crosstab: A Crosstab (also called *contingency table* (Yeager, n.d.)) is a *frequency table* that describes the relationship between two nominal attributes - let us say attribute *A* and *B* – to assess their statistical relation (some authors also calls it correlation but we have kept the correlation term for Pearson correlations only). In a crosstab, the nominal or discreate values of the attribute *A* determine the rows of the table (*r*) and the nominal or discreate values of the attribute *B*, determine the columns (*c*). *B* is the independent attribute and *A* is the dependent attribute. The cells of the table contain the number of times that a particular combination of nominal values occurred ("Count"). The dimension of a crosstab is reported as *r* × c (Han et al., 2012). This table also presents the row sums and column sums specified in the crosstab table under the name "Total". These totals are called marginal frequencies (Yeager, n.d.). As suggested by (Yeager, n.d.), in some cases we also refer in these tables the *sample proportions* of the counts of the attribute *A* against the corresponding total of the independent attribute *B* to directly compare, across the independent attribute, the percentages that the values of the attribute *A* has in each category of the attribute *B*. In this study, the *B* attribute is usually the patient´s follow-up time.

Cumulative Relative Frequency: It is the total relative frequency up to a given data value.

Descriptive Statistics: A facet of statistics that deals with describing the observed data with basic statistical measures (i.e. mean, median, mode etc.), without considering its population.

Expected frequencies: Expected frequencies are events that according to probability rules are expected to occur with frequencies e_1, e_2, e_3, ..., e_k (Murray R. Spiegel; Larry J. Stephens, n.d.). These expected frequencies are calculated based on the *Crosstab* generated with the *observed frequencies*. These expected frequencies are the number of times an event occurs if the two join attributes are independent (i.e. can occur by chance). These values are computed when we want to analyse the discrepancy between the *observed* and *expected frequencies* to assess if two attributes are correlated (i.e dependent). Since these expected frequencies are with *Crosstabs* presented, expected frequencies are presented as e_{ij} since it represents the expected join event that the attribute *A* takes on a_i and attribute *B* takes on b_j. Expected frequencies are calculated as follows (Han et al., 2012):

$$e_{\mathrm{ij}} = \frac{count(A = a_{\mathrm{i}}) * count(B = b\mathrm{j})}{n}$$

Formula Glossary 9 Expected frequencies

where n is the number of data tuples, count (A = a_{i}) is the number of tuples having value a_{i} for A, and count (B = b_{j}) is the number of tuples having value b_{j} for B.

Frequency table: A Frequency table is used to describe a single nominal or discreat attribute (Yeager, n.d.). This table is a data representation in which a grouped data, or not, is displayed along with the corresponding *frequencies* (Barbara Ilowsky; Susan Dean, 2017).

Frequency: The number of times an attribute value appears in a data set. In this study, frequencies are in tables specified under the name “Count”.

Histogram: An histogram is a graphic version of frequency distribution that allows to, not only see data frequencies of continuous quantitative data (Barbara Ilowsky; Susan Dean, 2017), but also, have an idea of the shape of the distribution of the data to see if the data is normally distributed or not (i.e. if data distribution forms a bell shape or not). Histograms can also tell if the data has a uniform range of values, which indicates that the data might have a high entropy equal to 1, or if there is one or two peaks and a lot of valleys of values, which indicates that the data might have a low entropy close to 0. That last information is interesting since some algorithms, such as decision trees uses the low entropy to identify the nodes of its tree through the information gain which is a measure based on the entropy.

This graph consists a set of bars of equal width drawn adjacent to each other where the horizontal scale represents classes/ranges of quantitative data values and the vertical scale, frequencies. The ranges for the horizontal scale are selected by the data analyst (Barbara Ilowsky; Susan Dean, 2017).

In the RapidMiner platform, histograms can be built and the ranges for the horizontal scale can be configured in the generated graph setting named *bins*. *Bins* can also be 0 which mean that in this case RapidMiner plots one bar for each unique data value that the example set has to indicate the number of times a specific value appears in the data set (i.e. frequency distribution). In the RapidMiner platform, histograms were in this study generated to better assess if sperm parameters values are normally distributed.

Independent samples: *Samples* that are selected randomly so that its observations do not depend on values of other observations (Minitab 18, 2017). Moreover, independent samples are also set of values that are not exactly covering the same population. For instance: if some of the patients that have performed a stereogram before the treatment does not carry out stereograms in all patient´s follow ups performed at 3, 6 and 12 months after the treatment, these different samples have a different number of patients and observations; and therefore, independent.

Inferential/Inductive Statistics: A facet of statistics that deals with estimating a population parameter based on a *sample* statistic (Barbara Ilowsky; Susan Dean, 2017) or inferring properties of a population to a *sample* to test statistical hypothesis.

<u>Nonparametric test</u>: Statistical tests applied to data samples that are not drawn from a *normal distribution.* These statistical tests are independent of population distributions and associated *parameters* (i.e mean and standard deviation of population). They are also valuable in dealing with nonnumerical data such as ordinal data (e.g severity grade of a condition) (Murray R. Spiegel; Larry J. Stephens, n.d.).

<u>Normal distribution</u>: After plotting a graphic that shows the frequency distribution of a data population, with for instance a *histogram*, the analyses of how the data is distributed is performed in order to assess if the data tends to be around a central value or not. An approximation of the plotted frequencies called probability distribution function f(x) is usually drawn. This function presents the undercurve probability of occurring a value in a range of values (Barbara Ilowsky; Susan Dean, 2017).

In many situations, f(x) depicts a bell curve shape. For instance, the heights of people, the size of things produced by machines, errors in measurements, blood pressure and marks on a test are some examples where the approximation of data frequencies forms a bell shape (Math´s Fun, 2017). When f(x) depicts a bell curve shape, we statistically say that the data is normally distributed. Figure Glossary 1 presents an example of a f(x) function normally distributed were the shaded area in the figure depicts the probability of a value occurring between the value 1 and 2.

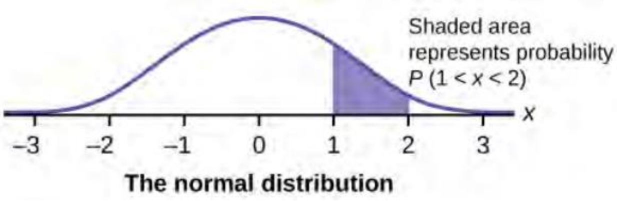


Figure Glossary 1 Example of a normal distribution (Barbara Ilowsky; Susan Dean, 2017)

Data displayed in a normal distribution have the following property (Math´s Fun, 2017):

- The mean of the data is equal to the median and the mode.
- The histogram presents a symmetry at its center
- 50% of the values are less than the mean and 50% are greater

Figure Glossary 2 depicts these properties.

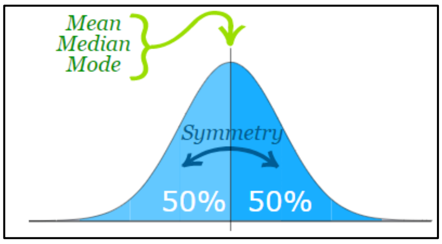


Figure Glossary 2 Properties of normal distribution (Math´s Fun, 2017)

Knowing that everything tends towards the mean in a normal distribution, that data in a normal distribution has a specific shape or that the standard deviation has an unambiguous relationship to the probability of an outcome, means we can do very powerful statistical tests (Ted Wrigley, 2016). Based on these statistical concepts, several statistical tests were form statisticians devised to, for instance, assess if there is a statistical significance difference between samples based on their f(x). We are saying shape since f(x) can have or not a normal distribution. To encompass this situation, statistical tests were devised for each type of data distribution. Nerveless, even if we apply a statistical test for normal distributed data to data that is not normally distributed, it can also work, but in the case of assessing statistical significance differences between samples, some differences might not show through. For this reason, it is always good to also apply to data a statistical test that is suitable to its distribution; and thereby, initially assess the distribution of the sample by analyzing, from other aspects, if its mean is equal to its median to further on select the most suitable statistical tests that one should use.

Observed frequencies: Observed frequencies are events that in a particular sample were observed to occur with frequencies o_1, o_2, o_3, ... , o_k (Murray R. Spiegel; Larry J. Stephens, n.d.). These observed frequencies are presented in *Crosstabs* when we want to assess the relationship between two attributes and are in these cases presented as o_{ij} which presents the join event that the attribute *A* takes on value a_i and attribute *B* takes on value b_j.

Parameter: A number that is a property of a population such as the median, mean etc (Barbara Ilowsky; Susan Dean, 2017).

Pie Graph: A Pie Graph is a circular statistical graphic which is divided into slices. Each slice is proportional to the quantity of data values it represents. This study has used the pie graph to have an overall picture of how the data values of an attribute qualitative or quantitative is distributed in terms of its value´s frequency since the RapidMiner presents by default, near each slice, the number of times each attribute value occurs (i.e. frequency) in contrast to the bar charts were the frequencies can only be seen in the RapidMiner through its y axis, which can bring some confusion. Therefore, in this study we have more used this type of graph than the most common *Bar Graph*.

Sample proportion a.k.a Relative Frequency: if *X* is the number of successes out of a sample of n observations and each observation is considered has a success, we can say that a sample proportion represented by $\hat{p}$ is equal to ("The Sample Proportion," n.d.):

$$\hat{p} = \frac{X}{n} * 100$$

Formula Glossary 10 Sample Proportion

In this study, sample proportions are in tables specified under the name "% of total". A sample proportion is in statistics also called relative frequency.

Sample: A portion of a larger population represented with the size *n*. In this study, the provided data is a sample since it is a portion of all patients that frequented the Coimbra´s Hospital.

Scatter plot: A scatter plot is a graph of data points where each pair of attributes´ values is treated as a pair of coordinates and plotted as points in a plane (Han et al., 2012). Scatter plots were in this study used to provide a first look at bivariate data (i.e. where each value of one attribute is paired with a value of the other attribute) to see clusters of points and outilers, or to explore the possibility of correlation relationships.

Time series: A time series is a graph of data points plotted by time. Hence, the independent variable (i.e. the x axis) on these graphs is time and the dependent variable (i.e. the y axis), is the value of the data that can be plotted by time. The line or curve plotted by joining the data points are often called a trend line or trend curve (Murray R. Spiegel; Larry J. Stephens, n.d.). Hence, time series were in this study built to visually assess the trend of sperm parameters through time (i.e. before the treatment and 3,6,12 months after the treatment).

Appendix A: Data understanding

The table below describes the attributes that were initially provided and selected.

Table A. 1 Description of the initially provided and selected attributes

ID	Attribute code name	Description	Attribute Category	Attribute Type
1	Idade_H	Age of the male patient at the time of the embolization	Quantitative	Numeric Continuous
2	Idade_M	Age of the woman´s male patient at the time of the partner´s embolization	Quantitative	Numeric Continuous
3	Tempo_Infert	Infertility time (i.e Time told in months by the patient´s partner regarding the time that they have been trying to conceive until the first fertility appointment)	Quantitative	Numeric Continuous
4	Prim_Sec	Woman´s Infertility type (i.e told by couple - if it is an infertility condition related to the first, indicated with the value "Primária", or second, indicated with the value "Secundária", pregnancy). Note: if the male patient has a child from a previous relationship but his current partner does not have a child yet, it is "Primária".	Qualitative	Binary
5	FR	Risk factors that the male patient has that might contribute to his infertility. These factors were identified through male´s interrogation and clinical report analysis. (i.e The patient did not told or their are no information about it so it is unknown (desconhecido), ….).	Qualitative	Nominal
6	Fator_F	Infertility factors indicated by the couple and from clinical avaliation for the male patient and its female partner (i.e unknow (desconhecido), Clinically studied but infertility cause were not identified (Não tem)....).	Qualitative	Nominal
7	Grau_Varicoc	Degree of varicocele´s severity (i.e lowest severity (1), medium severity (2) and highest severity (3))	Qualitative	Ordinal
8	Lateralidade	Scrotum site of the varicocele condition (i.e if varicocele is in the right (2), left (1) or on both (3) scrutum)	Qualitative	Nominal
9	Data	Scheduled date for the embolization treatment	Quantitative	Numeric Continuous
10	Notas	Notes taken by the physician after the embolization treatment. It can specify in which testicle the embolization was carried out or the type of complication the male patient felt after it.	Qualitative	Nominal
11	Complicações	Complications from Embolization treatment (i.e none (0), orquiepididimite (1), pain (2) or other (3))	Qualitative	Nominal

ID	Attribute code name	Description	Attribute Category	Attribute Type
12	Conc_Pre	Quantity of spermatozoa in millions per milliliters (concentration) of semen before the Embolization treatment. The minimum quantificated value for concentration is 0.1 so even if someone as only 3 spermatozoa e put 0.1.	Quantitative	Numeric Discrete
13	Conc_3M	Quantity of spermatozoa in millions per milliliters (concentration) of semen 3 months after the Embolization treatment. The minimum quantificated value for concentration is 0.1 so even if someone as only 3 spermatozoa e put 0.1.	Quantitative	Numeric Discrete
14	Conc_6M	Quantity of spermatozoa in millions per milliliters (concentration) of semen 6 months after the Embolization treatment. The minimum quantificated value for concentration is 0.1 so even if someone as only 3 spermatozoa e put 0.1.	Quantitative	Numeric Discrete
15	Conc_1A	Quantity of spermatozoa in millions per milliliters (concentration) of semen 1 year after the Embolization treatment. The minimum quantificated value for concentration is 0.1 so even if someone as only 3 spermatozoa e put 0.1.	Quantitative	Numeric Discrete
16	A_B_pré	Percentage of spermatozoa with progressive motility in the semen before the Embolization treatment.	Quantitative	Numeric Discrete
17	A_B_3M	Percentage of spermatozoa with progressive motility 3 months after the Embolization treatment.	Quantitative	Numeric Discrete
18	A_B_6M	Percentage of spermatozoa with progressive motility 6 months after the Embolization treatment.	Quantitative	Numeric Discrete
19	A_B_1A	Percentage of spermatozoa with progressive motility 12 months after the Embolization treatment.	Quantitative	Numeric Discrete
20	Formas_N_pré	Percentage of spermatozoa that have a normal shape in the semen before the Embolization treatment.	Quantitative	Numeric Discrete
21	Formas_N_3M	Percentage of spermatozoa that have a normal shape in the semen 3 Months after the Embolization treatment.	Quantitative	Numeric Discrete
22	Formas_N_6M	Percentage of spermatozoa that have a normal shape in the semen 6 Months after the Embolization treatment.	Quantitative	Numeric Discrete
23	Formas_N_1A	Percentage of spermatozoa that have a normal shape in the semen 1 Year after the Embolization treatment.	Quantitative	Numeric Discrete
24	Gravidez	Indicates if the female partner got pregnant after her partner ´s emolization treatment (i.e. Got pregnant (1), Did not got pregnant (0), Do not know if she got or not pregnant (2))	Qualitative	Nominal

ID	Attribute code name	Description	Attribute Category	Attribute Type
25	Num_Gravidezes	How many pregnancies the female partner had after her partner´s embolization treatment.	Quantitative	Numeric Discrete
26	Nascimento	If their was a birth after the embolization (i.e Yes (1), No (0), do not know if their was or not a birth (2))	Qualitative	Nominal
27	Num_Bébés	Number of alive babies born after her partner´s embolization treatment.	Quantitative	Numeric Discrete
28	Gravidez_pós_emb	Months that took the female partner to conceive after her partner´s embolization treatment (Told by the patient or through clinical report analysis).	Quantitative	Numeric Continuous
29	PMA	Inicial definition: if the female partner had a fertility treatment (ART). Actual definition: if the female partner had a fertility treatment (ART) to conceive after the embolization in the SMR service (i.e Yes (1), No (0)). (PMA stands for "Procriação Medicamente Assitida" which can be: insemination "inseminação artificial intrauterina" (IIU), in vitro fertilization "fertilização in vitro" (FIV), ICSI "microinjecção intracitoplasmática de espermatozóides", etc.)	Qualitative	Binary
30	Gravidez_espontanea	if the pregnancy was spontaneous after the embolization (i.e Yes (1), No (0)).	Qualitative	Binary
31	Repetia_embolização	If the male pacient would repeate the Embolization Treatment (i.e Yes (1), No (0), we could not get the opinion of the pacient (2))	Qualitative	Nominal
32	Razão_não_repetir	Reason to not repeat the Embolization treatment (i.e Pain (1), Infection (2), Technique (3), Do not have a specifique reason why he would not repeat the Embolization treatment (4)).	Qualitative	Nominal

Table A. 2 and Table A. 3 depict all Pearson correlations that were by the RapidMiner platform computed. These correlations were computed with the “Gravidez” attribute mapped to “Sim”=1 and “Não”=2 and the data set filtered by non-missing values in the “Gravidez” attribute for the numerical attributes of the final data set. Hence, these correlations were performed upon the 230 filtered and *finally preprocessed instances* of the provided data set.

Table A. 2 Pearson Correlations of related sperm parameter attributes

Attributes	Conc_Pre	Conc_3M	Conc_6M	Conc_1A	A_B_Pre	A_B_3M	A_B_6M	A_B_1A	Formas_N_Pre	Formas_N_3M	Formas_N_6M	Formas_N_1A	Numero_alterações_Pre	Numero_alterações_3M	Numero_alterações_6M	Numero_alterações_1A
Conc_Pre	1	0.491	0.633	0.508	0.138	-0.001	0.055	0.022	0.104	0.034	0.418	0.013	-0.383	-0.211	-0.233	-0.221
Conc_3M	0.491	1	0.685	0.661	0.177	0.178	0.345	0.088	0.217	0.271	0.040	0.212	-0.230	-0.429	-0.447	-0.311
Conc_6M	0.633	0.685	1	0.738	0.202	0.213	0.223	0.276	0.188	-0.069	0.389	-0.053	-0.345	-0.205	-0.463	-0.471
Conc_1A	0.508	0.661	0.738	1	0.312	0.170	0.284	0.248	0.195	0.131	0.322	0.180	-0.322	-0.317	-0.496	-0.547
A_B_Pre	0.138	0.177	0.202	0.312	1	0.444	0.588	0.409	0.123	0.259	0.329	0.224	-0.581	-0.304	-0.427	-0.386
A_B_3M	-0.001	0.178	0.213	0.170	0.444	1	0.597	0.427	0.123	0.377	0.440	0.246	-0.173	-0.582	-0.560	-0.396
A_B_6M	0.055	0.345	0.223	0.284	0.588	0.597	1	0.398	0.241	0.305	0.392	0.287	-0.246	-0.421	-0.701	-0.438
A_B_1A	0.022	0.088	0.276	0.248	0.409	0.427	0.398	1	-0.087	0.290	0.202	0.247	-0.161	-0.150	-0.259	-0.678
Formas_N_Pre	0.104	0.217	0.188	0.195	0.123	0.123	0.241	-0.087	1	0.392	0.206	0.089	-0.482	-0.173	-0.173	0.012
Formas_N_3M	0.034	0.271	-0.069	0.131	0.259	0.377	0.305	0.290	0.392	1	0.342	0.544	-0.240	-0.636	-0.143	-0.233
Formas_N_6M	0.418	0.040	0.389	0.322	0.329	0.440	0.392	0.202	0.206	0.342	1	0.964	-0.337	-0.326	-0.598	-0.346
Formas_N_1A	0.013	0.212	-0.053	0.180	0.224	0.246	0.287	0.247	0.089	0.544	0.964	1	-0.029	-0.386	-0.114	-0.487
Numero_alterações_Pre	-0.383	-0.230	-0.345	-0.322	-0.581	-0.173	-0.246	-0.161	-0.482	-0.240	-0.337	-0.029	1	0.284	0.340	0.243
Numero_alterações_3M	-0.211	-0.429	-0.205	-0.317	-0.304	-0.582	-0.421	-0.150	-0.173	-0.636	-0.326	-0.386	0.284	1	0.428	0.275
Numero_alterações_6M	-0.233	-0.447	-0.463	-0.496	-0.427	-0.560	-0.701	-0.259	-0.173	-0.143	-0.598	-0.114	0.340	0.428	1	0.452
Numero_alterações_1A	-0.221	-0.311	-0.471	-0.547	-0.386	-0.396	-0.438	-0.678	0.012	-0.233	-0.346	-0.487	0.243	0.275	0.452	1
Gravidez	0.008	-0.092	-0.161	-0.115	-0.155	-0.079	-0.123	-0.091	-0.045	-0.186	-0.068	-0.286	0.007	0.143	0.089	0.033
Idade_H	-0.009	0.186	-0.023	0.186	-0.007	-0.086	0.082	0.069	0.079	-0.007	-0.062	0.156	-0.050	-0.099	-0.021	-0.177

Table A. 3 Pearson Correlations of patient´s information

Attributes	Gravidez	Idade_H	Idade_M	Tempo_Infert	Data_Embolizacao
Gravidez	1	0.021	0.156	0.154	0.204
Idade_H	0.021	1	0.526	0.126	0.164
Idade_M	0.156	0.526	1	0.088	0.139
Tempo_Infert	0.154	0.126	0.088	1	0.094
Data_Embolizacao	0.204	0.164	0.139	0.094	1

Appendix B: Data Mining Models

This section discloses the modeling steps followed during the application of each data mining algorithm, as well as describes the models that were ran during this process. Hence, this section is organized as follows: in section B.1, we present the modeling steps and models built to apply the classification technique; in section B.2, we disclose the modeling steps and models built to apply the clustering technique, and finally, in section B.3, we showcase the modeling steps and models built to apply the FP-Growth algorithm.

B.1 Classification with Decision tree

As we have seen, we have at first applied the Decision tree algorithm upon the originally provided and preprocessed attributes that the RapidMiner platform determined to have good data quality (i.e. attributes disclosed in Table 5.40 with the cells under the column named "Attribute Name" colored in blue and the cells under the column named "Selected", colored in green); afterwards, we have applied the Decision tree algorithm to the groups of selected attributes identified in section 5.4.4; and finally we have reapplied it upon these same selected attributes but with the numerical attributes discretized as specified in Table 5.64. After recording the parameters of the models that had the best *Accuracy* during the training/testing of the RapidMiner´s Decision Tree and J-W48 algorithm (i.e. C4.5 algorithm) for the simple and cross validations (see Appendix B.1 for these results), we have selected the best model. The best model had operator parameter values that enabled a better performance during model training/testing (i.e. output of the "optimize parameter operator"), as well as during its validation. Hence, to predict the success of the embolization treatment we have followed several modeling steps that we summarize in Table 6.1.

Table 6.1 Modeling steps of the application of the Decision tree Algorithm

Step Number	Tested Attributes	Task performed
1	Idade_H, Idade_M, Cirurgias, Doença, Factor_Infertilidade_Masculino, Grau_Varicoc, HabitosAlcoolicos, HabitosTabagicos, Lateralidade, Profissao, Conc_3M, Conc_6M, Conc_1A, A_B_Pre, A_B_3M, A_B_6M, A_B_1A, Formas_N_Pre, Formas_N_3M, Gravidez	Applied the Decision Tree algorithm with the model depicted in Figure 6.4 (Model 1) upon the attributes with good data quality by the RapidMiner´s assessment.
2	A_B_Pre Conc_6M Formas_N_3M Grau_Varicoc Gravidez ProfissãoComRiscoDeContacto…	Applied the Decision Tree algorithm with the model depicted in Figure 6.4 (Model 1) on the first group of selected attributes.

Step Number	Tested Attributes	Task performed
3	A_B_Pre_Qualificado A_B_3M_Qualificado Conc_3M_Qualificado Grau_Varicoc Gravidez ProfissãoComRiscoDeContacto…	Applied the Decision Tree algorithm with the model depicted in Figure 6.4 (Model 1) on the second group of selected attributes.
4	Grau_Varicoc Gravidez ProfissãoComRiscoDeContacto... Qualificar_Espermograma_3M Qualificar_Espermograma_Pre	Applied the Decision Tree algorithm with the model depicted in Figure 6.4 (Model 1) on the third group of selected attributes.
5		Retested the steps 2 to 4 by adding the Idade_M attribute to check if we could surpass the f-measure found in phase 3. Hence, this step also ran the model depicted in Figure 6.4 (Model 1).
6	A_B_3M_Qualificado A_B_Pre_Qualificado Conc_3M_Qualificado Grau_Varicoc Gravidez Idade_M ProfissãoComRiscoDeContacto…	Filtered the data set by non-missing values under the fifth group of selected attributes, transformed its nominal values into numerical values through mapping and dichotomization and normalized all values to test the decision tree algorithm on a normalized and numerical data set. The choice of this set of attributes and transformations was guided by the most interesting result identified with the clustering technique (i.e. K-means modeling step 2). This task was performed by running the VCF depicted in Figure 6.5 (Model 2).
7	A_B_3M_Qualificado A_B_Pre_Qualificado Conc_3M_Qualificado Grau_Varicoc Gravidez Idade_M ProfissãoComRiscoDeContacto…	Reran the model 2 without filtering the data set by non-missing values under the fifth group of selected attributes. This task was executed with the VCF depicted in Figure 6.8 (Model 3).
8	A_B_3M A_B_Pre Conc_3M Conc_6M Formas_N_3M Grau_Varicoc Gravidez HabitosAlcoolicos_Processado… HabitosTabagicos_Processado… Idade_M ProfissãoComRiscoDeContacto…	Transformed the numerical attributes into nominal with a user defined discretization that reflects the WHO thresholds for the sperm parameters attributes, the woman´s age quartiles for the Idade_M attribute and the severity grade was automatically dichotomized; Hence, all the data set was transformed into nominal attribute values. The choice of this set of attributes and transformations was raised after the application of the association technique. This task was executed with the VCF depicted in Figure 6.9 (Model 4).
9	A_B_Pre_Qualificado A_B_3M_Qualificado Conc_3M_Qualificado Grau_Varicoc Gravidez ProfissãoComRiscoDeContacto…	Optimized the best decision tree model obtained upon the 230 preprocessed instances (i.e. model obtained in step 3) with the Bagging ensemble method. This task was executed with the VCF depicted in Figure 6.12 (Model 5). Please note that the best model was obtained from the step 6.

As we can see from the table above, the Decision tree algorithm was mainly applied with the model depicted in Figure 6.4. This model is described in the following figures: Figure 6.1, depicts a close-up of the upper half of the decision tree model shown in Figure 6.4 where we can see the 4 testing steps depicted within the 4 colored rectangles; Figure 6.2, shows the nested built model for the training and testing of the decision tree algorithm and Figure 6.3, depicts the application of the decision tree model upon the validation data set – this figure is a close-up of the lower half of the model disclosed in Figure 6.4.

The only things that varied between the several runs that were carried out with this model (i.e. modeling steps) were the selection of attributes to mine - specified within the “Select Attributes” operator here entitled “Select 19 Original” - and the alteration of the model parameters to validate at the end of each application – carried out within the Decision tree operator that is highlighted in Figure 6.3 with the name “Optimized Decision Tree”.

All the tests performed with the model depicted in Figure 6.4 were carried out on 230 labeled instances (i.e. 129 instances to train + 55 instances to test + 46 instances to test).

If we look closer into the partial view of the decision tree model in Figure 6.1 we see that it reflects the validation process previously disclosed in section 4.2.8.1. In fact, the operator that we can see entitled “Split Data” at the bottom-left of Figure 6.1 splits the data set into training and validation, and the training data set, is served as an input to the “Optimize Parameters” operator which is a nested operator that contains the “Split Validation” operator or the “Cross Validation” operator, which in turn, has the decision tree algorithm nested with its application and computed performance measures that are depicted in Figure 6.2.

If we analyze the model´s validation depicted in Figure 6.3, we can see that the model´s validation is performed with the same operators than those shown in Figure 6.2 since it performs the same task (i.e. test/validate the application of a model). The only difference here, is that the apply model operator, here called “Apply Optimized Mo..” receives as an input the partitioned data set split for validation purposes instead of testing purposes, as depicted in Figure 6.2.

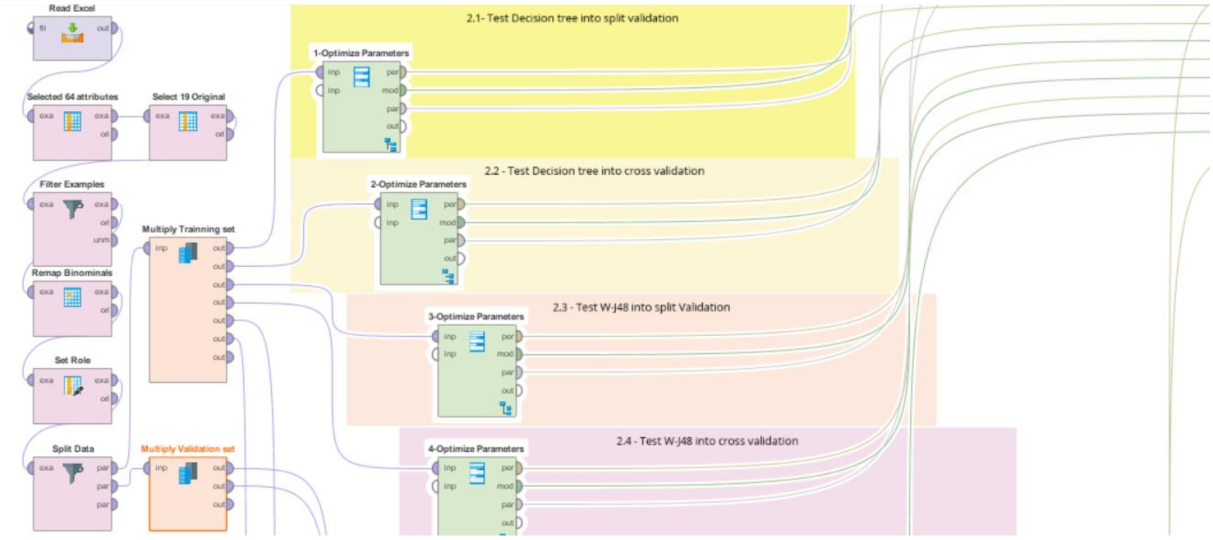


Figure 6.1Main Decision tree model - partial view

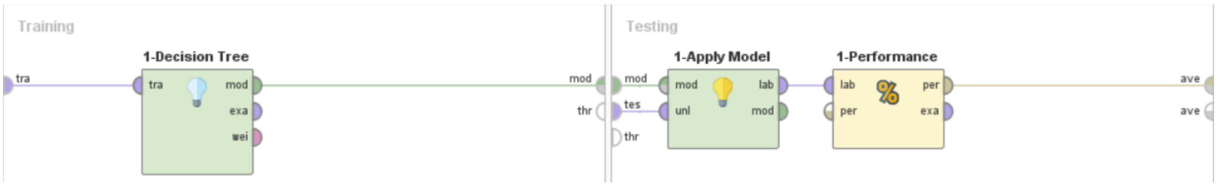


Figure 6.2 Decision Tree´s training and testing

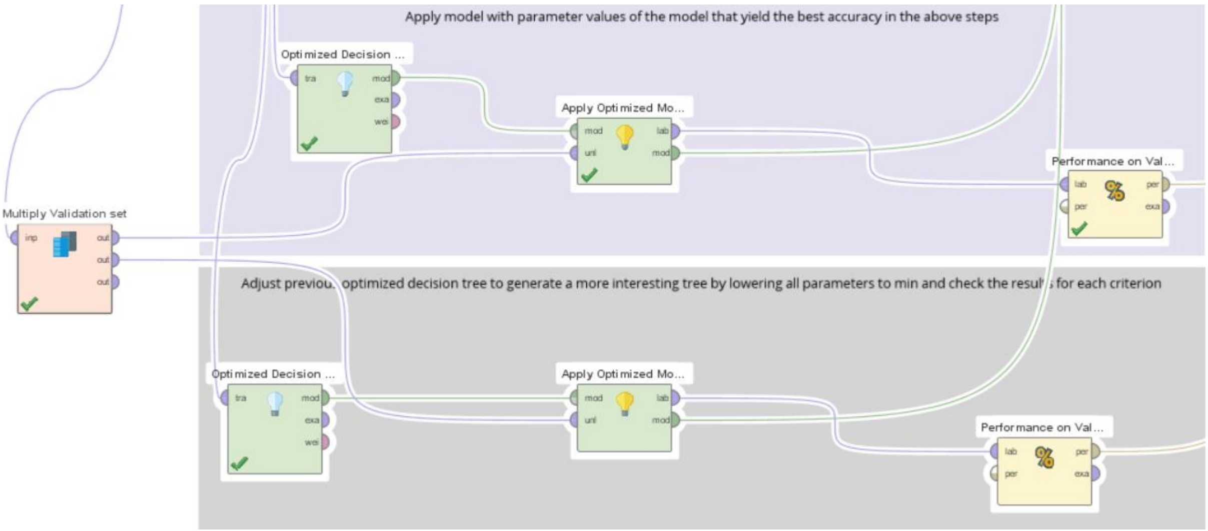

Apply model with parameter values of the model that yield the best accuracy in the above steps
Optimized Decision ...
Apply Optimized Mo...
Performance on Val...
Multiply Validation set
Adjust previou optimized decision tree to generate a more interesting tree by lowering all parameters to min and check the resul for each criterion
Optimized Decision ...
Apply Optimized Mo...
Performance on Val...

Figure 6.3 Model´s validation

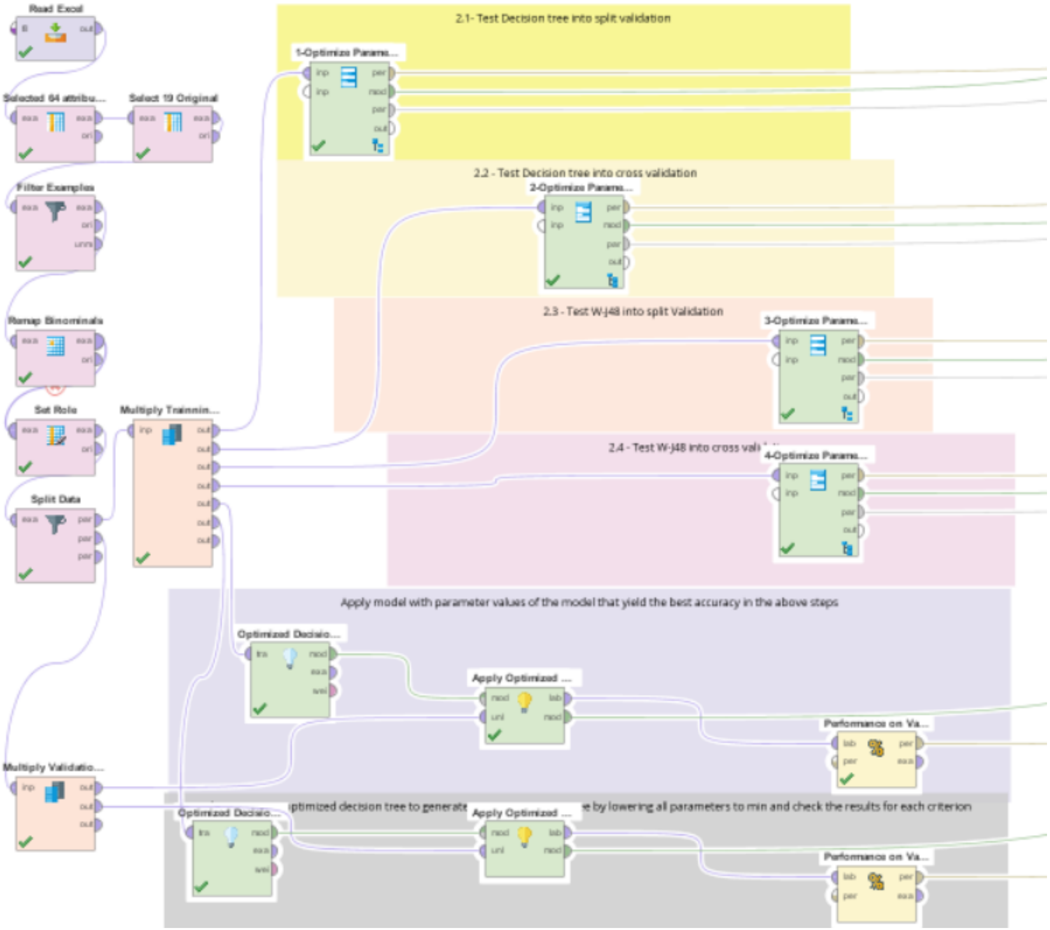

Read Excel
Selected 64 attribu...
Select 19 Original
Filter Examples
Remap Binominals
Set Role
Multiply Trainnin...
Split Data
Multiply Validatio...
2.1- Test Decision tree into split validation
1-Optimize Parame...
2.2 - Test Decision tree into cross validation
2-Optimize Parame...
2.3 - Test W-J48 into split Validation
3-Optimize Parame...
2.4 - Test W-J48 into cross vali
4-Optimize Parame...
Apply model with parameter values of the model that yield the best accuracy in the above steps
Optimized Decisio...
Apply Optimized ...
Performance on Va...
ptimized decision tree to generate
e by lowering all parameters to min and check the results for each criterion

Figure 6.4 Main Decision Tree model - complete view (Model 1 – used in step 1 to 5)

In order to establish a model that computes the highest f-measure, we have fine-tuned the model depicted in Figure 6.4 by adding a set of RapidMiner´s operators to transform the 5th group of selected attribute values into numerical and dichotomized values in order to test the model on a numerical data set (step 6 and 7). A close-up of the beginning of this fine-tuned model can be seen in Figure 6.5 based on the model in Figure 6.4. During the construction of this model we have also tested the transformation of the "Gravidez" attribute into 1 and 0 and the normalization of the woman´s age but its performance metrics and results remained the same; and therefore, we have decided not to go with these transformations since they only made the interpretation of the computed models more difficult. Further on, we have also tested the Decision tree model on all sperm parameters and woman´s age attributes, as well as on Varicocele severity grade manually dichotomized to also test the model only on nominal values (step 8). A close-up of the beginning of this model can also be seen in Figure 6.9.

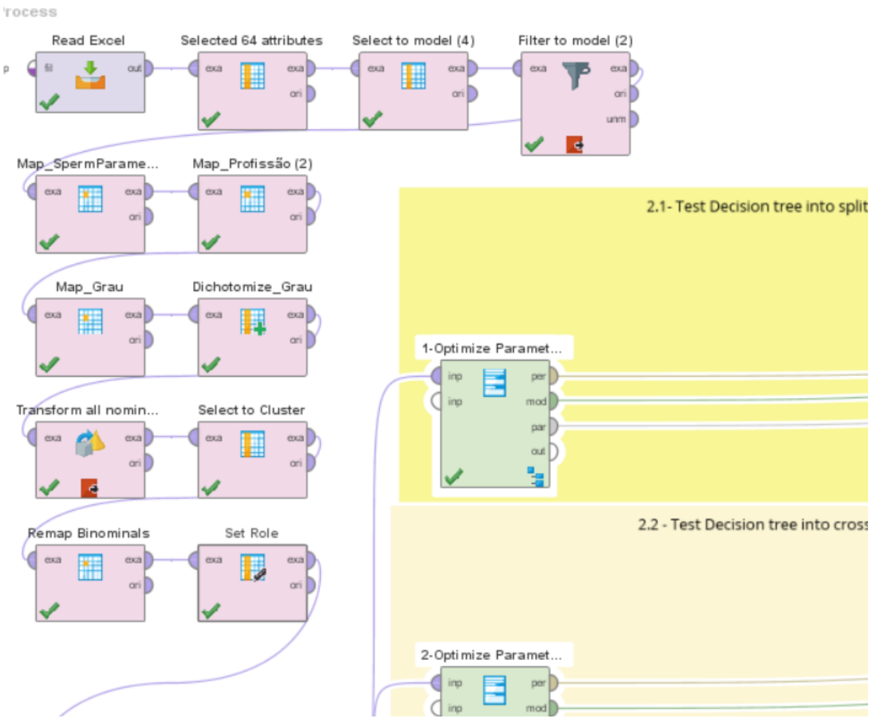


Figure 6.5 First Fine-Tuned model - partial view (Model 2 - used in step 6)

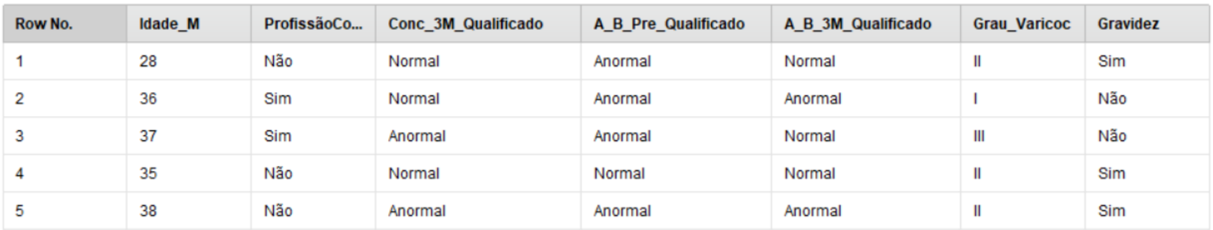
The pink operators seen in the above model have its purposes disclosed in the below Table 6.2.

Table 6.2 Attribute transformations of the Decision tree Model 2

Operator name	Purpose
Selected 64 attributes	Selects the 64 attributes that were selected to assess.
Select to model (4)	Selects the attributes encompassed in the 5th group of attributes for this model to mine: A_B_3M_Qualificado A_B_Pre_Qualificado Conc_3M_Qualificado Grau_Varicoc Gravidez Idade_M ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos

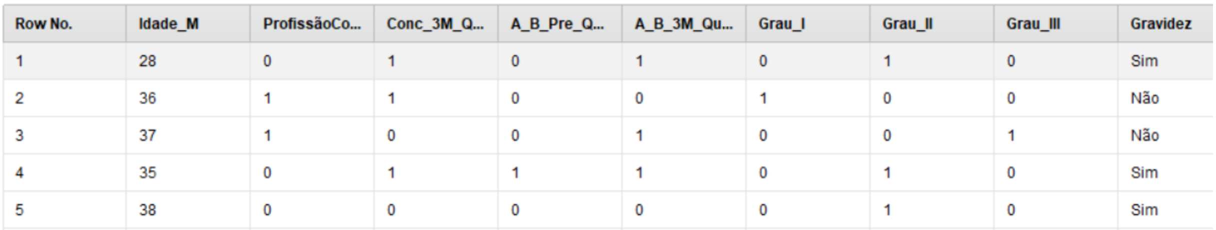
Operator name	Purpose
Filter to model (2)	Filters the data set by non-missing values in the previously selected attributes.
Map_SpermParamet…	Maps the sperm parameter values into: Anormal -> 0 Normal -> 1
Map_Profissão (2)	Maps the values of the “ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos” atribute into: Não -> 0 Sim -> 1
Map_Grau	Maps the values of the “Grau_Varicoc” attribute into: I -> 1 II -> 2 III -> 3
Dichotomize_Grau	Dichotomizes the mapped “Grau_Varicoc” attribute by generating 3 new attributes to manage separately the severity grades. For this purpose the following attributes were generated: Grau_I Grau_II Grau_III
Transform all nomin…	Transforms all previously mapped attributes into numerical attributes.
Select to Cluster	Reselects all transformed attributes to mine, except the “Grau_Varicoc” attribute since its dichotomized attributes are the ones that we aim to mine.
Remap Binomials	Indicates that the “Gravidez” attribute value “Sim” must be considered as a positive value and the “Não” value, as a negative value.
Set Role	Indicates that the “Gravidez” attribute is a lable attribute.

These attribute transformations enabled us to normalize the data and end-up with a numerical data set. We disclose this transformation by showing, in Figure 6.6 the first filtered data set rows by the decision tree Model 2, and in Figure 6.7 these same rows transformed after the application of the operators described in Table 6.2.



Row No.	Idade_M	ProfissãoCo...	Conc_3M_Qualificado	A_B_Pre_Qualificado	A_B_3M_Qualificado	Grau_Varicoc	Gravidez
1	28	Não	Normal	Anormal	Normal	II	Sim
2	36	Sim	Normal	Anormal	Anormal	I	Não
3	37	Sim	Anormal	Anormal	Normal	III	Não
4	35	Não	Normal	Normal	Normal	II	Sim
5	38	Não	Anormal	Anormal	Anormal	II	Sim

Figure 6.6 First filtered data set rows by the Decision tree Model 2



Row No.	Idade_M	ProfissãoCo...	Conc_3M_Q...	A_B_Pre_Q...	A_B_3M_Qu...	Grau_I	Grau_II	Grau_III	Gravidez
1	28	0	1	0	1	0	1	0	Sim
2	36	1	1	0	0	1	0	0	Não
3	37	1	0	0	1	0	0	1	Não
4	35	0	1	1	1	0	1	0	Sim
5	38	0	0	0	0	0	1	0	Sim

Figure 6.7 Transformation of the first filtered data set rows by the Decision tree Model 2

Since the Decision tree Model 2 filters its data set by non-missing values, we have ended up with 85 instances to mine (i.e. 48 instances to train + 20 instances to test + 17 instances to validate); and hence, the resulting decision tree shown leaves with a small number of instances. To tackle this situation, we have rerun the decision tree Model 2 on all 230 instances which produced Model 3 that we below partly disclose in Figure 6.8. As we can see, this model does not have the “Filter to model (2)” operator that was previously seen, which enabled us to mine all 230 instances with selected attributes transformed as shown in Table 6.2. The remainder of this model is the same as the one depicted in Figure 6.4.

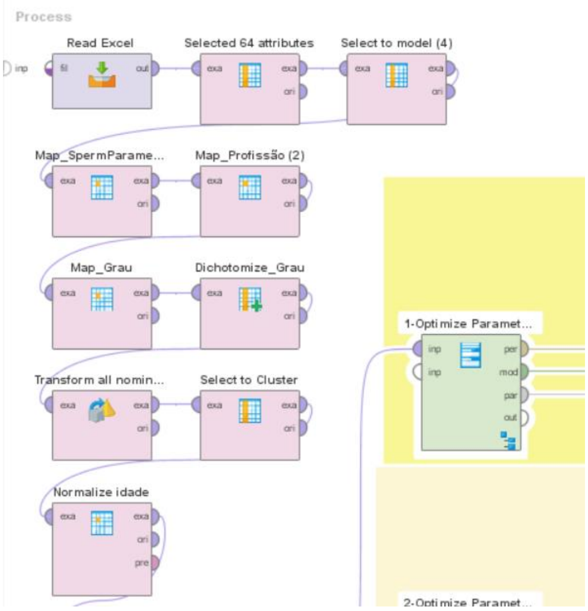


Figure 6.8 Second Fine-Tuned model - partial view (Model 3 - used in step 7)

If we look closer into the operators used to implement the 8th decision tree modeling step depicted in Figure 6.9, we see that we have several operators named after sperm parameters and woman´s age attributes in the center-left of the presented VCF. These operators implement the discretization of their corresponding attribute that was carried-out with the “Discretize by User Specification” operator. Afterwards, the operator called “Remap Binomials (2)”, indicates that the attribute value “Não” must be considered as a negative value and the attribute value “Sim”, as a positive attribute value. At last, the “Nominal to Binomial” operator dichotomizes the “Grau_Varicoc” attribute. Hence, after all these attribute transformations we have ended-up with a binomial data set. This transformed data set can be partly seen in Figure 6.10 and Figure 6.11 where in Figure 6.10, we show the first 5 rows of the data set before attribute transformation (at the time of the “Set Role (2)” operator) and in Figure 6.11, we showcase part of these same rows transformed by model operators. The remainder of the Model 4 was built as the Model 1 in Figure 6.4.

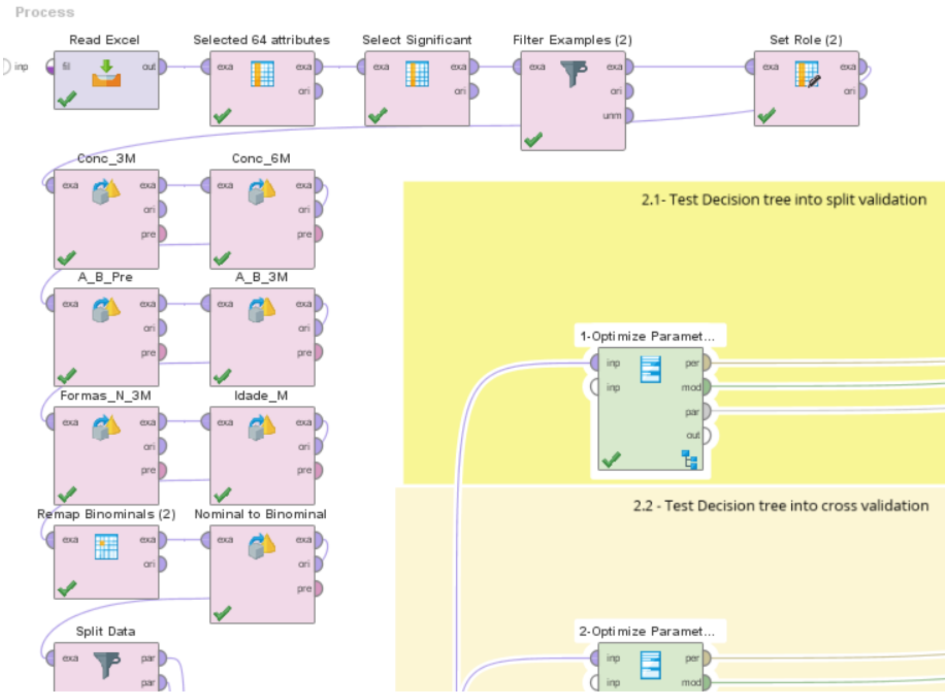
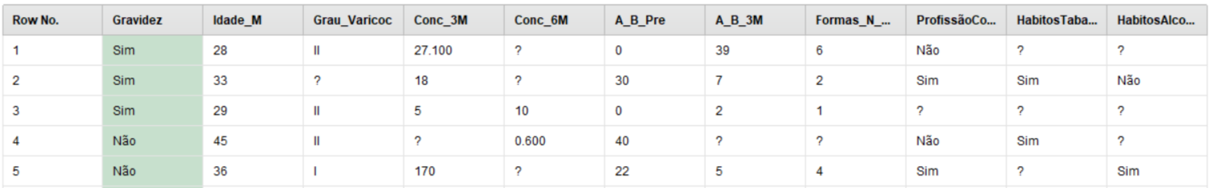
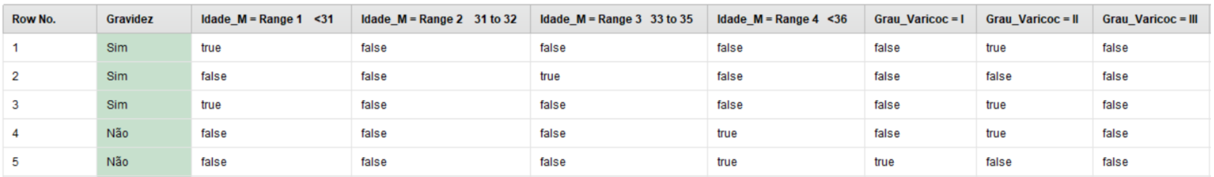


Figure 6.9 Third Fine-Tuned model - partial view (Model 4 - used in step 8)



Row No.	Gravidez	Idade_M	Grau_Varicoc	Conc_3M	Conc_6M	A_B_Pre	A_B_3M	Formas_N_...	ProfissãoCo...	HabitosTaba...	HabitosAlco...
1	Sim	28	II	27.100	?	0	39	6	Não	?	?
2	Sim	33	?	18	?	30	7	2	Sim	Sim	Não
3	Sim	29	II	5	10	0	2	1	?	?	?
4	Não	45	II	?	0.600	40	?	?	Não	Sim	?
5	Não	36	I	170	?	22	5	4	Sim	?	Sim

Figure 6.10 Filtered rows by the “Filter Examples (2)” operator of the Decision tree Model 4



Row No.	Gravidez	Idade_M = Range 1 <31	Idade_M = Range 2 31 to 32	Idade_M = Range 3 33 to 35	Idade_M = Range 4 <36	Grau_Varicoc = I	Grau_Varicoc = II	Grau_Varicoc = III
1	Sim	true	false	false	false	false	true	false
2	Sim	false	false	true	false	false	false	false
3	Sim	true	false	false	false	false	true	false
4	Não	false	false	false	true	false	true	false
5	Não	false	false	false	true	true	false	false

Figure 6.11 Transformation of the filtered data set rows by the Decision tree Model 4

To the best trained/tested model on the 230 assessed instances (i.e. model obtained during the decision tree modeling step 3 at the first testing step), we have finally applied the *Bagging* ensemble method to increase the performance measures and minimize overfit.

The *Bagging* ensemble method was applied with the “Bagging” RapidMiner operator which is a nested operator that ran the decision tree algorithm with the best parameter values identified through the several modeling steps. The “Bagging” RapidMiner´s operator can be seen highlighted in orange in the model depicted in Figure 6.12 and next, in Figure 6.13, we can see the built process executed within the “1-Validation (2)” operator of Model 5.

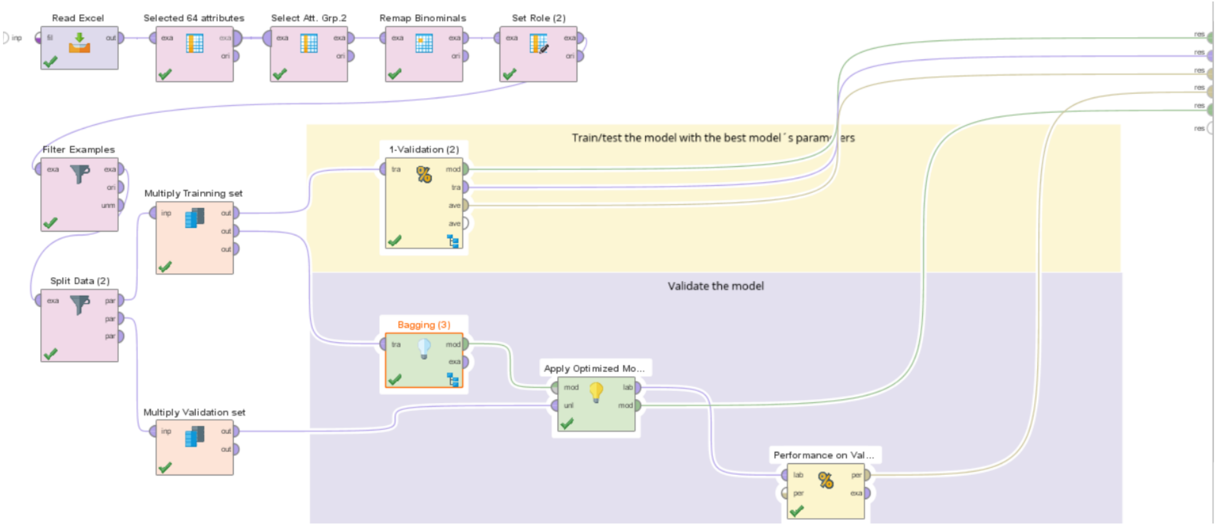


Figure 6.12 Bagging Application – (Model 5 – used in step 9)

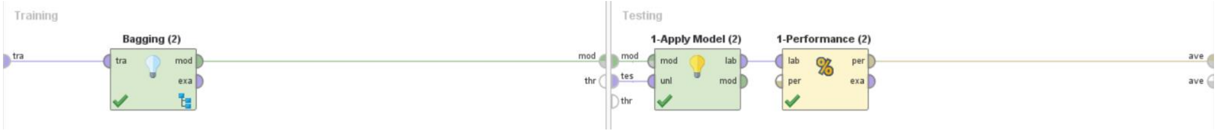


Figure 6.13 Bagging application within the “1-Validation (2)” operator of Model 5

The “Bagging” operator was executed with its default parameter value in the iteration field; (i.e. a number of iterations equal to 10) and the sample ratio was set to 0.7.

B.2 Clustering with K-means

The K-Means partitioning clustering algorithm was also applied through several modeling steps. These steps are summarized in Table 6.3.

Table 6.3 Modeling steps of the application of the K-Means Algorithm

Step Number	Tested Attributes	Task performed
1		Applied the K-Means algorithm upon all selected groups of attributes. This task was performed with model 1 shown in Figure 6.14.
2	Idade_M, Idade_H, A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_I, Grau_II, Grau_III, ProfissãoComRiscoDeContacto,, Gravidez.	Filtered the tested attributes - that were based on the fifth group of selected attributes with more filled instances - by non-missing values. Afterwards, manually dichotomized the severity grade attribute named "Grau_Varicoc", normalized all attribute values by previously transforming them into numerical values and at last, applied the K-Means algorithm. These tasks were performed with model 2 depicted in Figure 6.21. In this step, we have also performed another sub-step where we have retested model 2 without the "Idade_H" attribute which improved its performance.
3	Idade_M, Idade_H, A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_Varicoc, ProfissãoComRiscoDeContacto…, Gravidez.	Reran the step 2 with the severity grade mapped instead of dichotomized. This step was executed with model 3 depicted in Figure 6.24.
4	Idade_M, Gravidez	Defined the best discretization for the idade_M attribute to be used by also other data mining algorithms. This step was executed with model 4 depicted in Figure 6.25.
5	Idade_M, A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_I, Grau_II, Grau_III, ProfissãoComRiscoDeContacto, Gravidez.	Applied the previously defined discretization of the "Idade_M" attribute upon the most interesting clustering result until then found (i.e. model generated in step 2 without the "Idade_H" attribute). This task was executed with the model 5 that can be seen in Figure 6.26.
6	Idade_M, A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_I, Grau_II, Grau_III, ProfissãoComRiscoDeContacto,, Gravidez.	Applied the best decision tree model (i.e. decision tree obtained in the decision tree modeling step 6) upon each clustered data set generated from the step 2. This task was executed with the model 6 that is depicted in Figure 6.22.

As we have previously seen, this technique was applied with the RapidMiner´s "K-mean" operator. Furthermore, distances were calculated with the Euclidean and Manhattan measure since all nominal values were converted into numeric values. Moreover, prior to the "k-Means" application, the data selection operators called "Select Attributes" and "Filter Examples" were firstly applied to the data set, as well as the "Nominal to Numerical" operator to prepare the nominal attributes for the k-Means specificities. Afterwards we also tested the K-means algorithm by converting the nominal attributes that were not binomial, to binomial with the

"Nominal to Binomial" operator before converting them to numeric, and have also normalized the sperm parameter values with the "Normalize" operator to have all sperm parameters values between 0.0 and 1. Hence, we ended up with a data set with all its attribute values in the same range of values to identify clusters of data in the same scale of values. To identify clusters of data, we have tested the K-means algorithm upon the *final preprocessed data set* for 2 to 4 clusters and for the Euclidean and Manhattan distances. Figure 6.14 presents the first model that was built to apply the K-means algorithm on the final and preprocessed data set where all operators mentioned can be seen.

The K-means was tested with several selected and filtered attributes that were manually changed in the operator within the yellow rectangle that can be seen in the center-left of the figure below. The operator in green called "Optimize Parameter…" performs a loop on all parameters that were tested within the K-means algorithm. This loop is presented in the following Figure 6.15 where we can see the nested operators of the "Optimize Parameter…" operator. As we can see, the K-means is linked to a performance operator to retrive the *Davies Bouldin* index; and hence, internally evaluate the generated clusters – all the tests performed with this model are reported in the Appendix B.2.

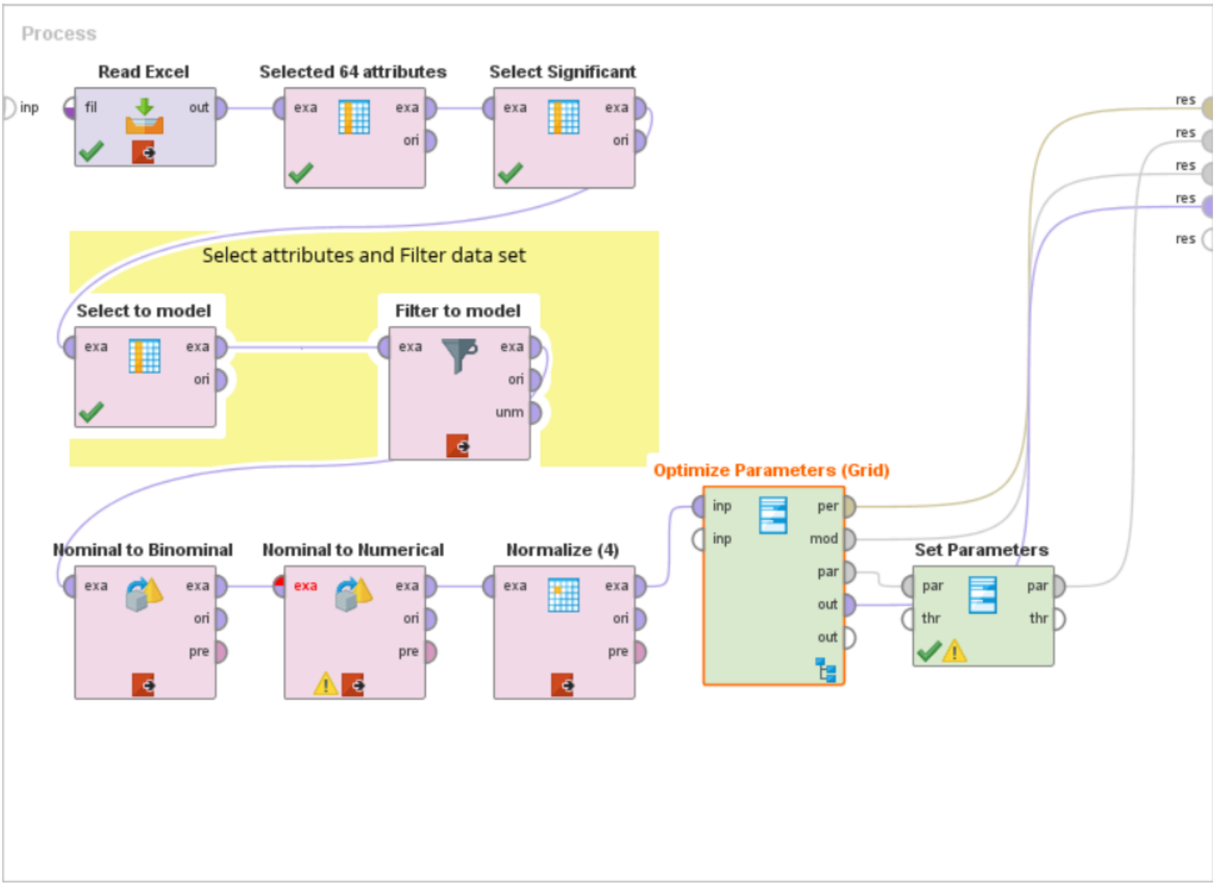


Figure 6.14 General Clustering Model (Model 1)

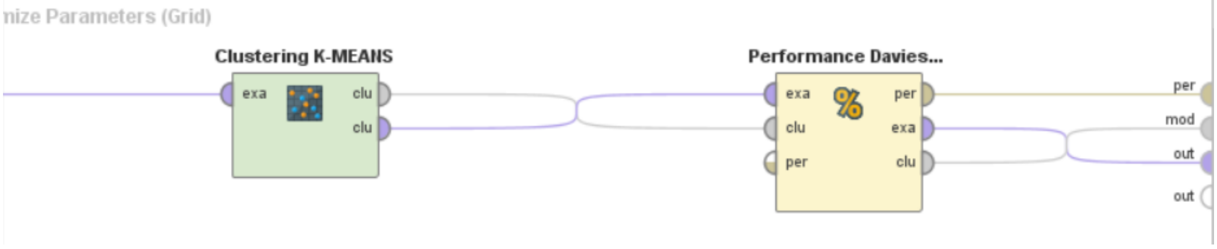


Figure 6.15 K-Means application

After analyzing the results that were generated from model in Figure 6.14 we have selected the most interesting data pattern in terms of the number of instances it covers, and from there, fine-tuned this model to not only seek for more interesting data patterns, based on the elected attributes, but mainly, enhance our capability of interpreting the results generated. Therefore, we have altered the model depicted in Figure 6.14 by personalizing the nominal to numerical mapping – with the “Map”, “Nominal to numerical” and “Parse numbers” operators – and de-normalizing the numerical attributes; and afterwards, by applying statistical analysis upon the clustered data. In summary, the operators between the operator named in Figure 6.14 as “Filter to model” and “Normalize” were altered/added and several operators were then, after the “Optimize Parameters (Grids)” operator, added to better interpret the generated clusters of data.

Figure 6.16 presents the first fine-tuned clustering data mining model, Figure 6.17, partly discloses the process that was built to interpret its results and Figure 6.24, presents an overview of the best built K-means´ data mining model that mainly joins the two last *VCFs*.

Regarding the model depicted in Figure 6.16, we can see that its main differences in comparison with the one in Figure 6.14, are the operators within the beige rectangle called “Map attribute values”. In fact, these are the ones that were added to the model depicted in Figure 6.14: the first two operators with the prefix “Map..” used the operator called “Map”; the third operator, with the prefix “Grau_Varicoc..”, used the “Nominal to Numerical” operator with the “coding_type” setting set to the option “dummy coding” – this option transforms the “Grau_Varicoc” attribute, that is nominal, into 3 dichotomized bi-numerical attributes – and the last operator, with the prefix “Transform..”, used the “Parse numbers” operator, to transforms all nominal values into numerical ones. The operator named “Normalize idade”, that appears in the following pink rectangle, only normalizes the original numerical attributes (i.e. “Idade_M” and “Idade_H”). Hence, it only transforms the filtered patient´s ages into a value within 0 and 1, were 0, represents the lowest age that appears in the filtered data set, and 1, the highest one.

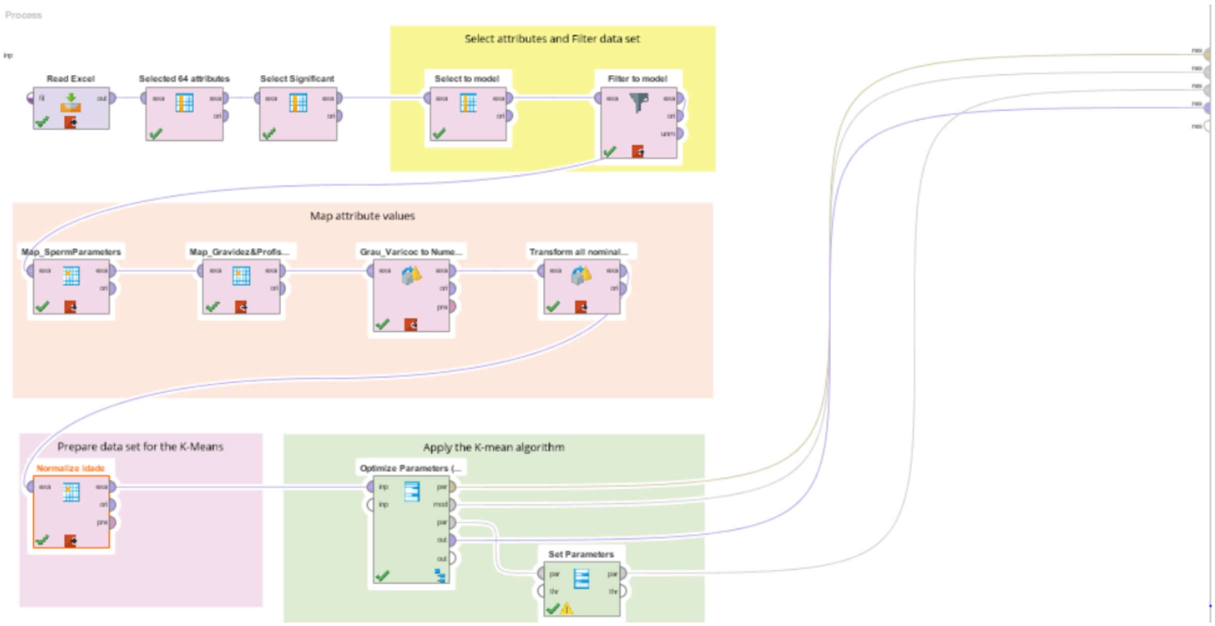


Figure 6.16 Fine-tuned Clustering Model

In Table 6.4 we show how the nominal attributes tested with the model depicted above in Figure 6.16 were mapped with the “Map” operator. Note that, as previously stated in Figure 6.16, the “Grau_Varicoc” attribute was automatically dichotomized with the “Nominal to Numerical” operator; and therefore, it is not covered in the table below.

Table 6.4 Mapping values of the fine-tuned clustering model depicted in Figure 6.16

Attribute Name	Original Value	New mapped value
Conc_3M_Qualificado A_B_Pre_Qualificado A_B_3M_Qualificado	Anormal	0
	Normal	1
Gravidez ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	Não	0
	Sim	1

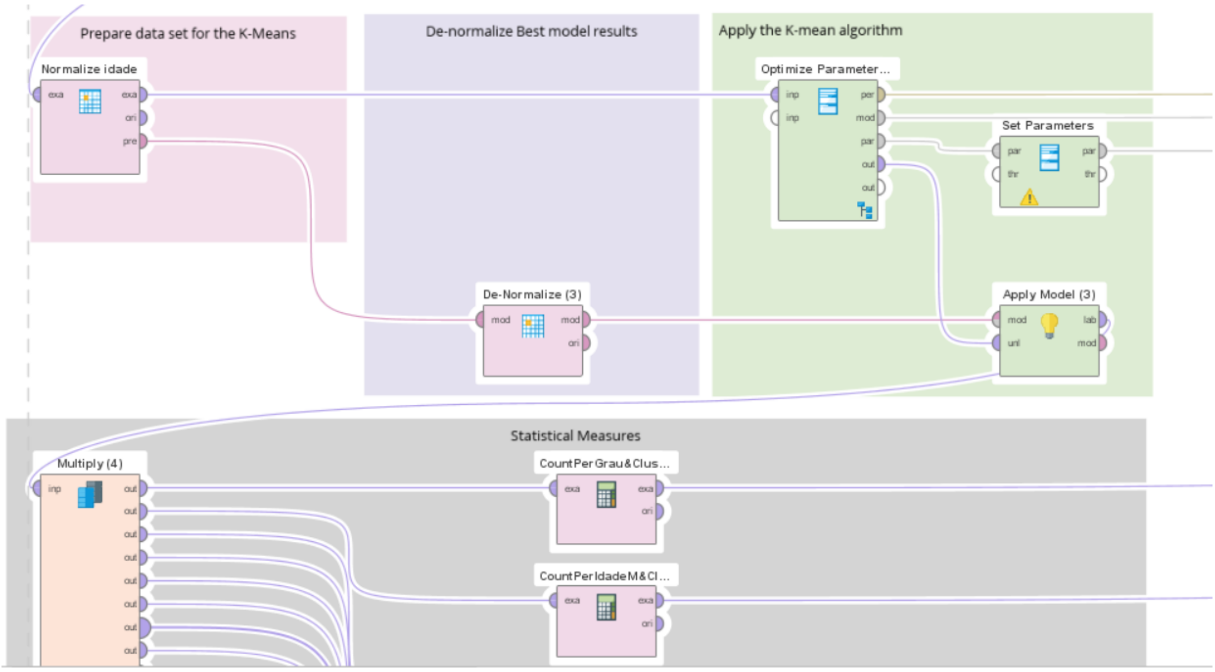


Figure 6.17 De-normalize K-Means results (Partly)

To put in context the de-normalization process shown in Figure 6.17, the operator called “Normalize idade”, seen in the previous Figure 6.16, was connected to the operator “De-Normalize (3)”. This last operator was then connected to the operator “Apply Model (3)” that also receives the best K-means´ result computed by the operator “Optimize Parameter”. Hence, based on the information that the operator “Apply Model” has (i.e. the patient´s ages de-normalized and the best clustered data set generated by the K-means algorithm), the model computes all main statistical measures (i.e. count, mean, standard deviation, maximum value, minimum value and median), for each attribute within each generated cluster (i.e. performs a group by cluster and attribute) – these statistical measures were calculated with the “Aggregate” operator which are the last pink operators seen with a calculator icon inside them. Furthermore, we have also applied the statistical significance test ANOVA with the “Group ANOVA” operator to each group of clustered data and attribute.

During the testing of the RapidMiner´s statistical operators, we have seen that the dichotomized “Grau_Varicoc” attribute by the “Nominal to Numerical” operator - in Figure 6.16 named as “Grau_Varicoc to nume…” - was not being handled by the RapidMiner´s statistical operators; and therefore, we have manually dichotomized this attribute with the RapidMiner´s operators that can be seen highlighted in orange in Figure 6.18 below and described as follows:

- “Map_Grau” – uses the “Map” operator to map the “Grau_Varicoc” values.
- “Dichotomize_Grau” – uses the “Generate Attributes” operator to create 3 new attributes called “Grau_I”, “Grau_II” and Grau_III” filled with the results of the function expressions seen in Figure 6.19 below - the values of each of these newly created attributes are described in Table 6.5.
- “Select to Cluster” – uses the “Select Attributes” operator to discard the former “Grau_Varicoc” attribute of the model.

All attributes modeled by the VCF depicted in the below Figure 6.18 are depicted in Table 6.5 where all data transformations performed are disclosed.

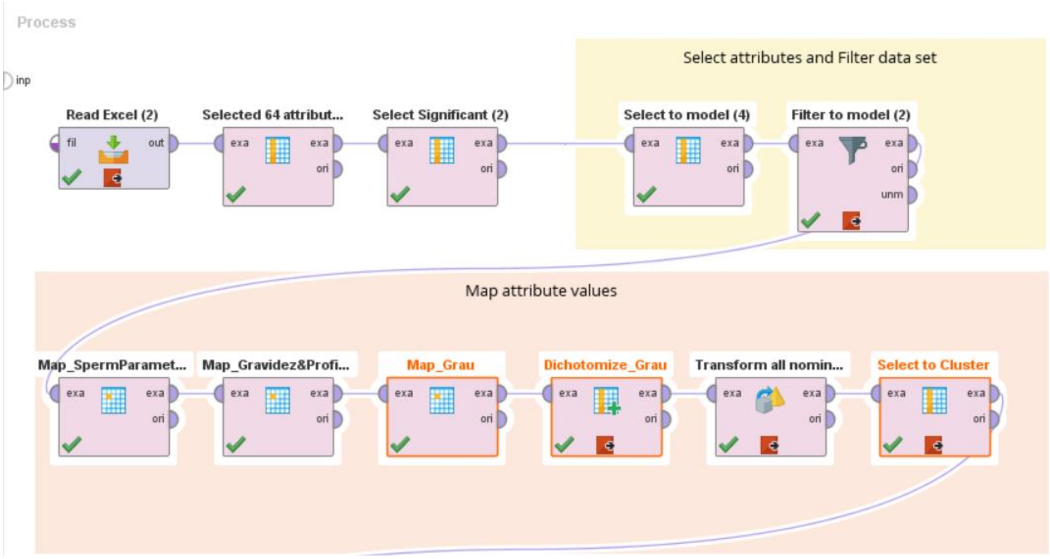
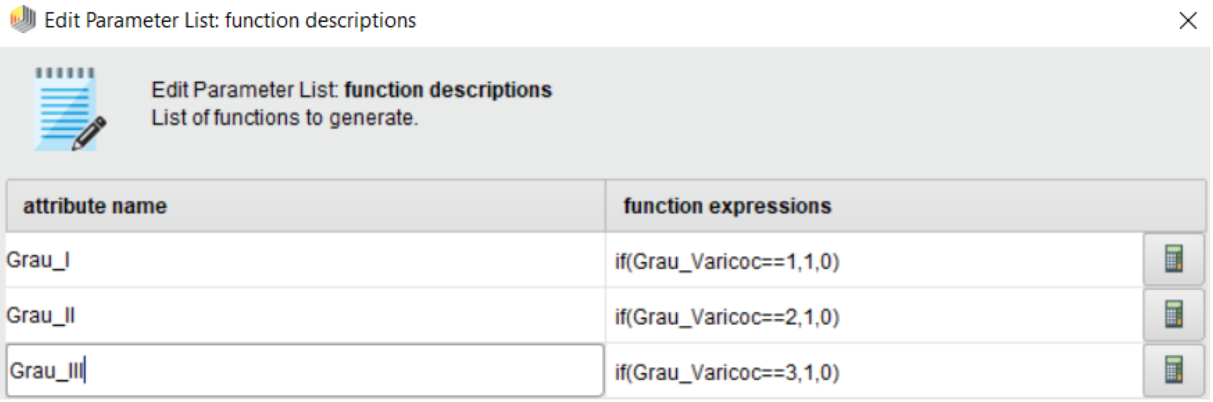


Figure 6.18 Manual dichotomization of the “Grau_Varicoc” attribute – VCF

Table 6.5 Data transformations of the model depicted in Figure 6.18

Type of data transformation	Attribute Name	Original Value	New mapped value
Mapping	Conc_3M_Qualificado A_B_Pre_Qualificado A_B_3M_Qualificado	Anormal	0
		Normal	1
	Gravidez ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	Não	0
		Sim	1
	Grau_Varicoc	I	1
		II	2
		III	3
Dichotomization	Grau_I	Dichotomized attribute that has the value 1 when the corresponding value of the attribute “Grau_Varicoc” has the value 1.	
	Grau_II	Dichotomized attribute that has the value 1 when the corresponding value of the attribute “Grau_Varicoc” has the value 2.	
	Grau_III	Dichotomized attribute that has the value 1 when the corresponding value of the attribute “Grau_Varicoc” has the value 3.	

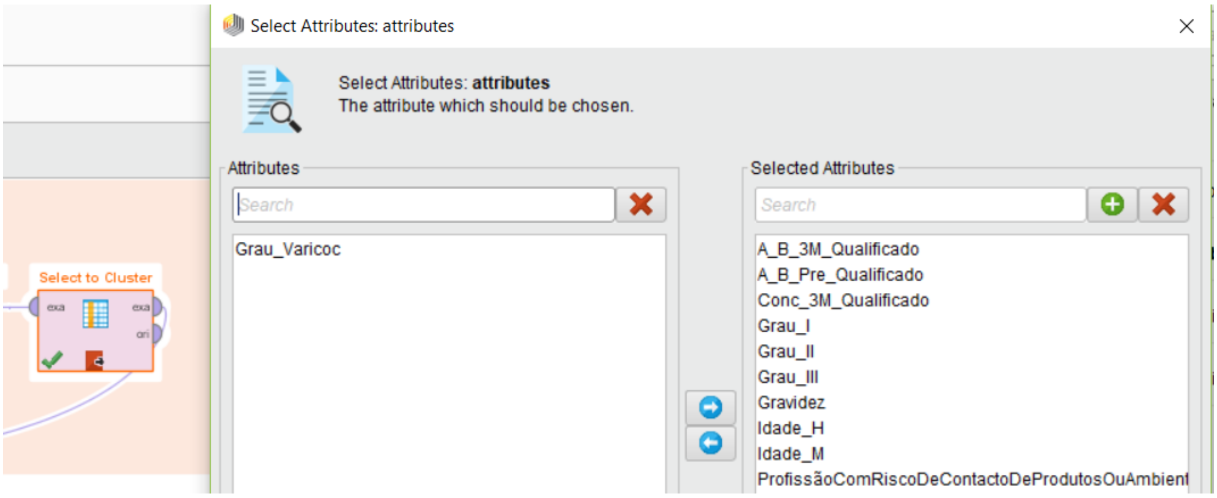


Edit Parameter List: function descriptions

Edit Parameter List: **function descriptions**
List of functions to generate.

attribute name	function expressions
Grau_I	if(Grau_Varicoc==1,1,0)
Grau_II	if(Grau_Varicoc==2,1,0)
Grau_III	if(Grau_Varicoc==3,1,0)

Figure 6.19 Manual dichotomization of the “Grau_Varicoc” attribute – Implementation



Select Attributes: attributes

Select Attributes: **attributes**
The attribute which should be chosen.

Attributes

Grau_Varicoc

Selected Attributes

A_B_3M_Qualificado
A_B_Pre_Qualificado
Conc_3M_Qualificado
Grau_I
Grau_II
Grau_III
Gravidez
Idade_H
Idade_M
ProfissãoComRiscoDeContactoDeProdutosOuAmbient

Figure 6.20 Selected attributes to model in the *VCF* depicted in Figure 6.18

An overview of the model partly shown in Figure 6.18 can be seen in Figure 6.21 belowm where we show in the upper/first half of the model the *VCF* depicted in Figure 6.18 and in the lower/second half, part of the *VCF* presented in Figure 6.17. Furthermore, the application of the statistical significance test ANOVA is shown in the bright yellow rectangle.

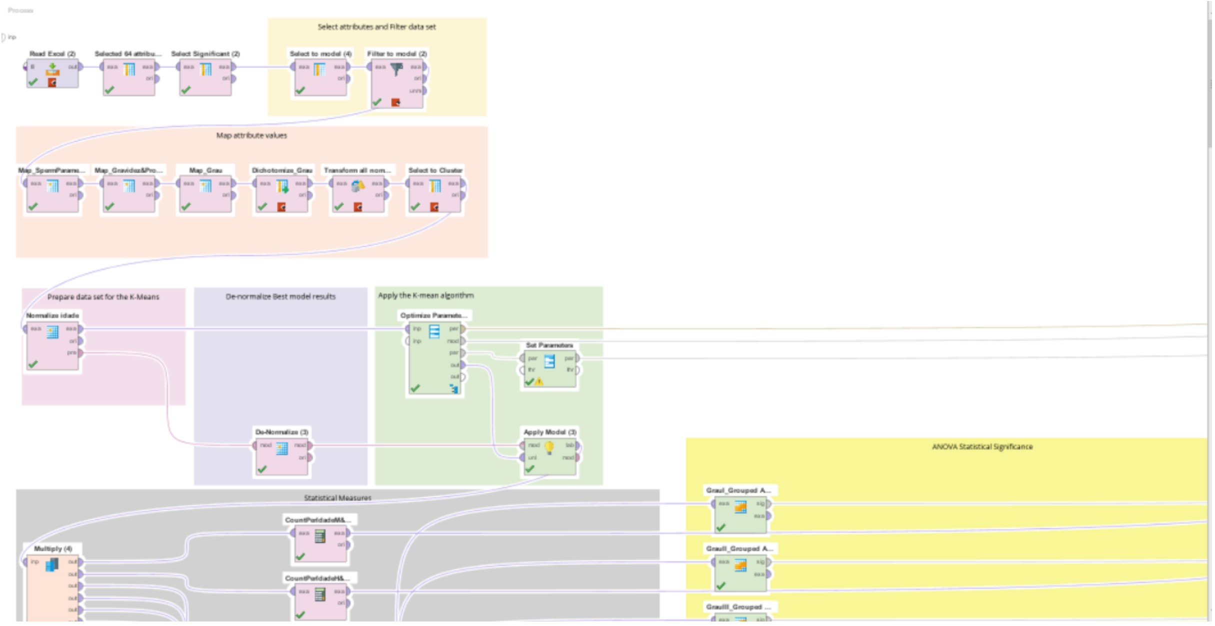


Figure 6.21 Overview of the model partly shown in Figure 6.18 – “Grau_Varicoc” manually dichotomized (Model 2)

Due to the interestingness of the results obtained in Model 2, we have at last applied the best decision tree model that we have obtained during the decision tree application step 6.1.1 upon the delivered partitioned data set by Model 2. This task was achieved by:

0 Filtering the data set by one of the attribute values of the “Cluster” attribute (the attribute “Cluster” is generated by the K-means algorithm to classify the instances by its clusters).

1 Transforms the “Gravidez” attribute into Binomial, but by remaining the 0 and 1 value instead of the “Não” and “Sim” value because the decision tree modeling step 6.1.1 gave better results than the step 6.1. Please note that this transformation was needed to use the decision tree´s splitting criteria that we have tested (i.e. accuracy, gain ratio etc.)

2 Set the “Gravidez” attribute as a label “attribute”.

3 Apply upon the filtered instances in step 1 the Decision tree model with the RapidMiner´s operator called “Decision Tree”, with its parameters set with the characteristics found for the model generated in the decision tree aplication step 6.1.1 – since the aim is to describe the clusters, we have not trained/tested the data set within each cluster because we have a small number of instances.

4 Record the generated decision tree and redo all the steps until all clusters are modelled by the decision tree operator.

The built model that implements the above steps is shown in Figure 6.22 where we can see that the results delivered by the K-means operator within the “optimize parameters” operator is connected to the “Multiply (4)” operator that creates several copies of the partitioned data set to serve as inputs to the following operators that implement the steps disclosed previously. These operators append the previous Model 2.

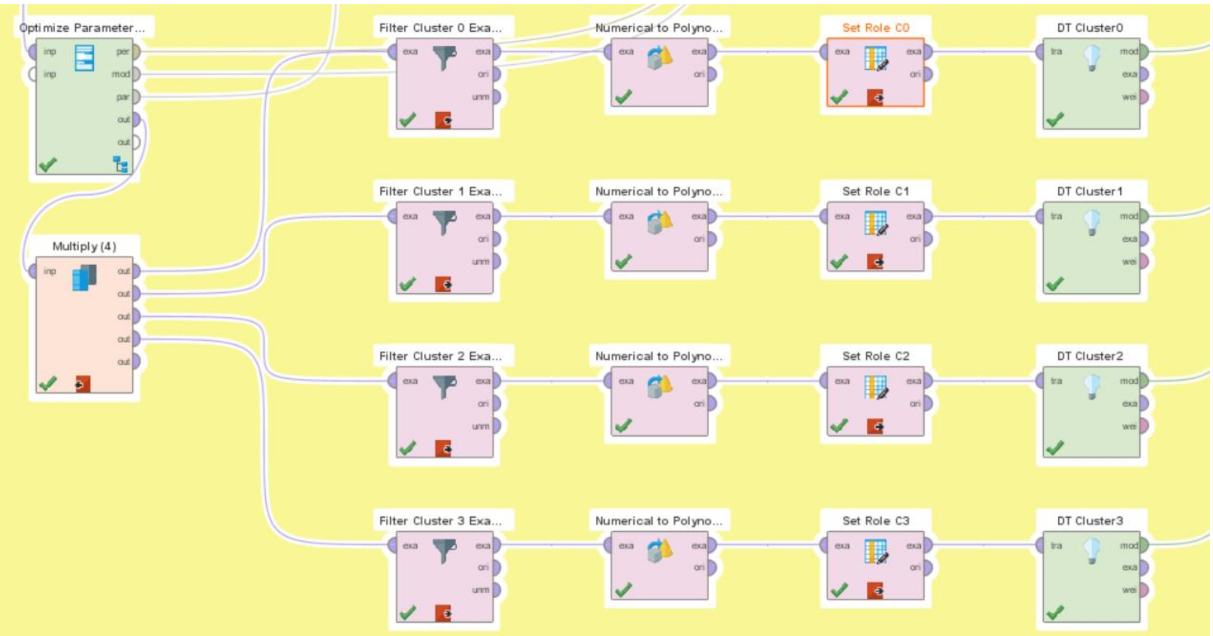


Figure 6.22 Decision Tree applied upon Clustered Data (Model 6)

Model 2 gave us the idea to also test the K-means algorithm with only the “Grau_Varicoc” attribute mapped, as specified in Table 6.5. Hence, we have deleted the operators called “Dichotomize_Grau” and “Select to Cluster” of the model depicted in Figure 6.18, and in the operator with the prefix “Normalize..”, also normalized the “Grau_Varicoc” attribute since in this context the severity grade was ranging between 1 and 3. This altered *VCF* can be see in Figure 6.23 below, and the overview of this model is depicted in Figure 6.24. This last overview can also serve as a general view of the fine-tuned K-means´ models that were ran since the only difference between all these K-means models tested was the way the attributes “Grau_Varicoc”, “Idade_M” and “Idade_H” were transformed/preprocessed within the beige rectangle, after electing the best starting model (model depicted in Figure 6.16).

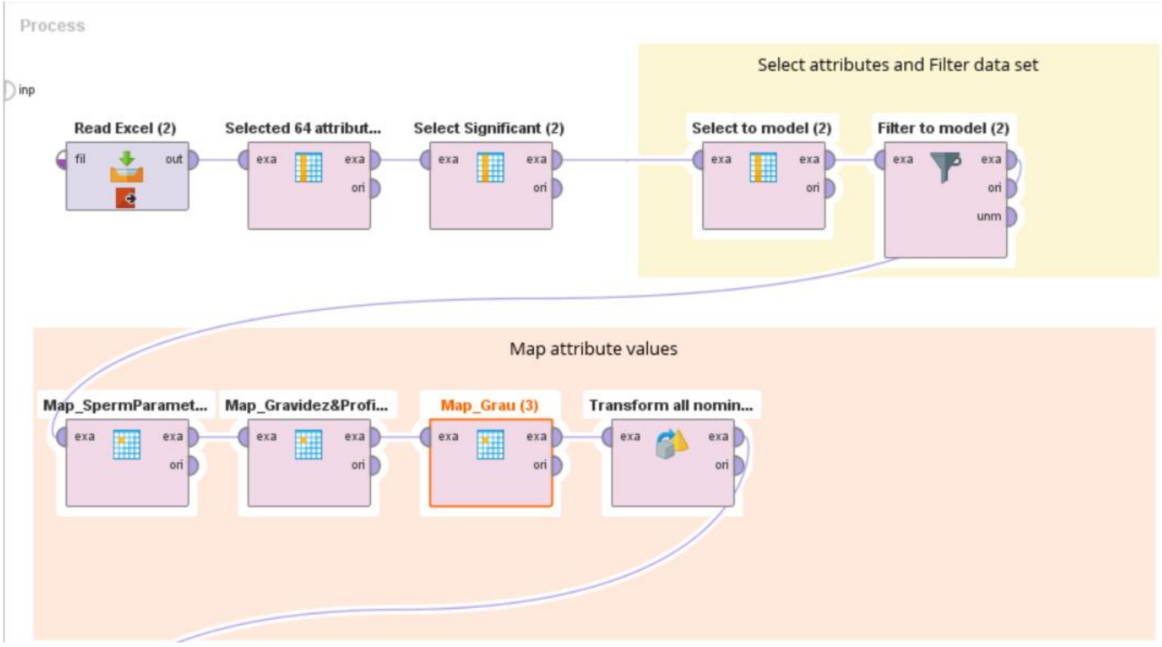


Figure 6.23 Mapping of the “Grau_Varicoc” attribute – VCF

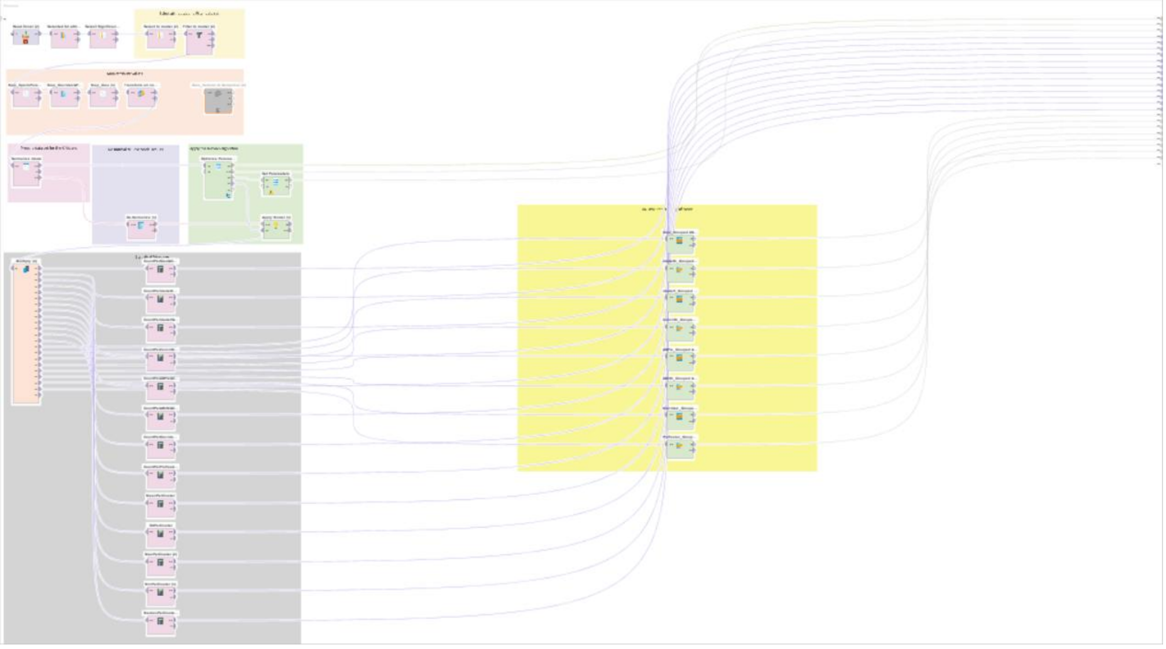


Figure 6.24 Overview of the model partly shown in Figure 6.23 – “Grau_Varicoc” Mapped (Model 3)

We also had the idea to discretize woman ´s age “Idade_M” to at last, reapply the K-Means model depicted in Figure 6.21 – where the “Grau_Varicoc” attribute is manually dichotomized – with the defined discretization. Note that we have only applied the woman´s age discretization to the model where the “Grau_Varicoc” attribute was manually dichotomized because this model gave more interesting results. The application of the K-Means algorithm served in this case to better understand the pattern of the patient´s partner age prior the formulation of its discretization to guide us through the best way to potentiate information discovery.

In Dougherty, Kohavi, and Sahami (1995), an interesting study on how we can discretize continuous attributes is presented. They expose the “Equal Interval Width” method, that merely divides a range of observed values into k equal sized bins, where k is a user supplied parameter, to end up with bins with the same range width (i.e. if the width is 10, all boundaries of the ranges will be multiples of 10); the “Equal Frequency Intervals” method, that divides a continuous variable into k bins to end up with a similar number of instances in each bin (if all values were unique we would end up with a quite same number of instances in each bin); and the “Discretize by Entropy” method, that considers the label attribute to define ranges of values that minimized the *Entropy*. This last option has caught our interest since the woman´s ages varies with the label attribute values; and therefore, we have explored that option by previously assessing with the K-means algorithm if there were ranges of woman´s ages that were only part of one label class to seek for ranges with low *Entropy*. To do so, we have altered the model depicted in Figure 6.21 to only apply the K-means algorithm upon the “Idade_M” and Gravidez” attribute. This altered model can be seen in Figure 6.25 (Model 4).

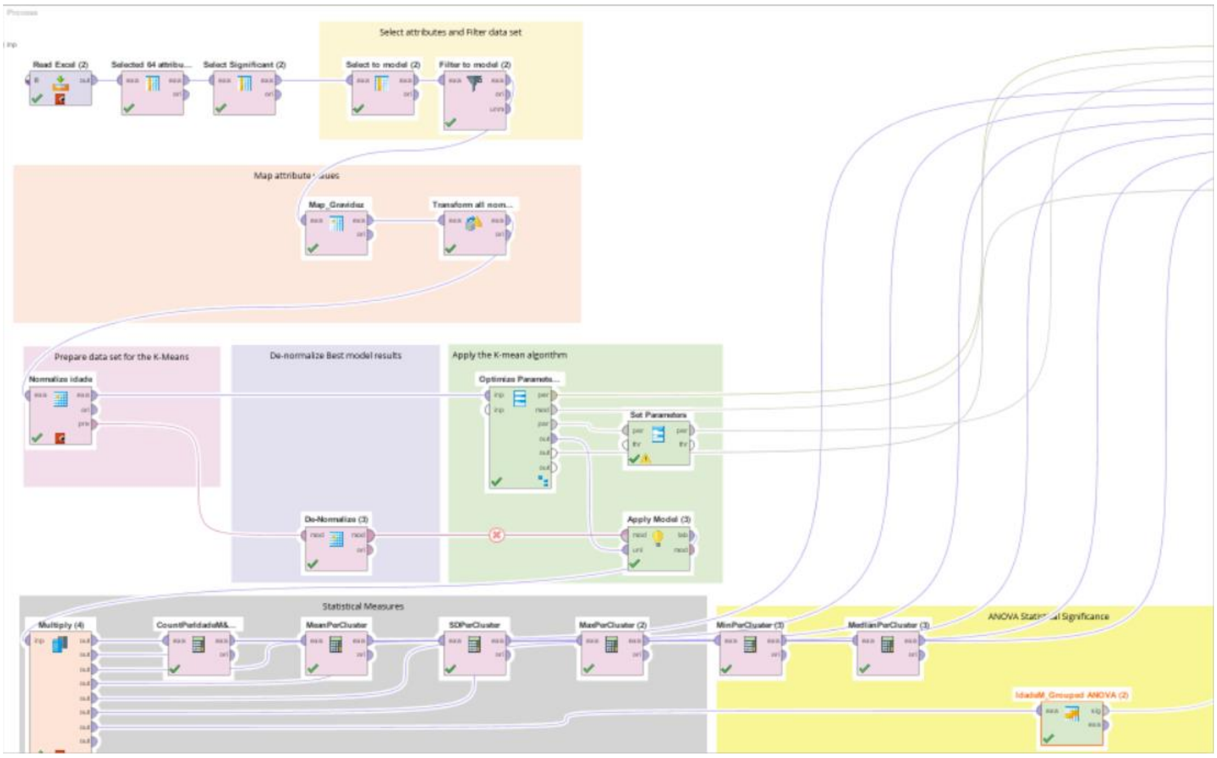


Figure 6.25 Assessment to the “Idade_M” attribute vs the “Gravidez” attribute (Model 4)

In Figure 6.26 below we disclose the model depicted in Figure 6.21 with the selected discretization of the “Idade_M” attribute. This discretization was carried out with RapidMiner operators that can be seen in the bright yellow rectangle called “Discretize Idade_M”. The selected discretization method was the “Equal Frequency Intervals” method and was implemented with the “Discretized by Frequency” operator that can be seen in Figure 6.26 renamed as “Discretize” – the reason behind the election of this discretization is in section 5.5.2.3 disclosed. The two following operators with the prefix “Transform” handle the applied discretization: we have changed the display of the discretization values gathered in the “Idade_M” attribute to a numerical value by applying the “Generate Attribute” operator with the script “cut(Idade_M,5,1)” and transformed the resulting value into a numerical value with the “Parse Numbers” operator (i.e. has transformed “range1 [-∞-30.500]” to “1” with the “Generate Attribute” operator and the resulting value “1” was transformed into the number 1 with the “Parse Numbers” operator).

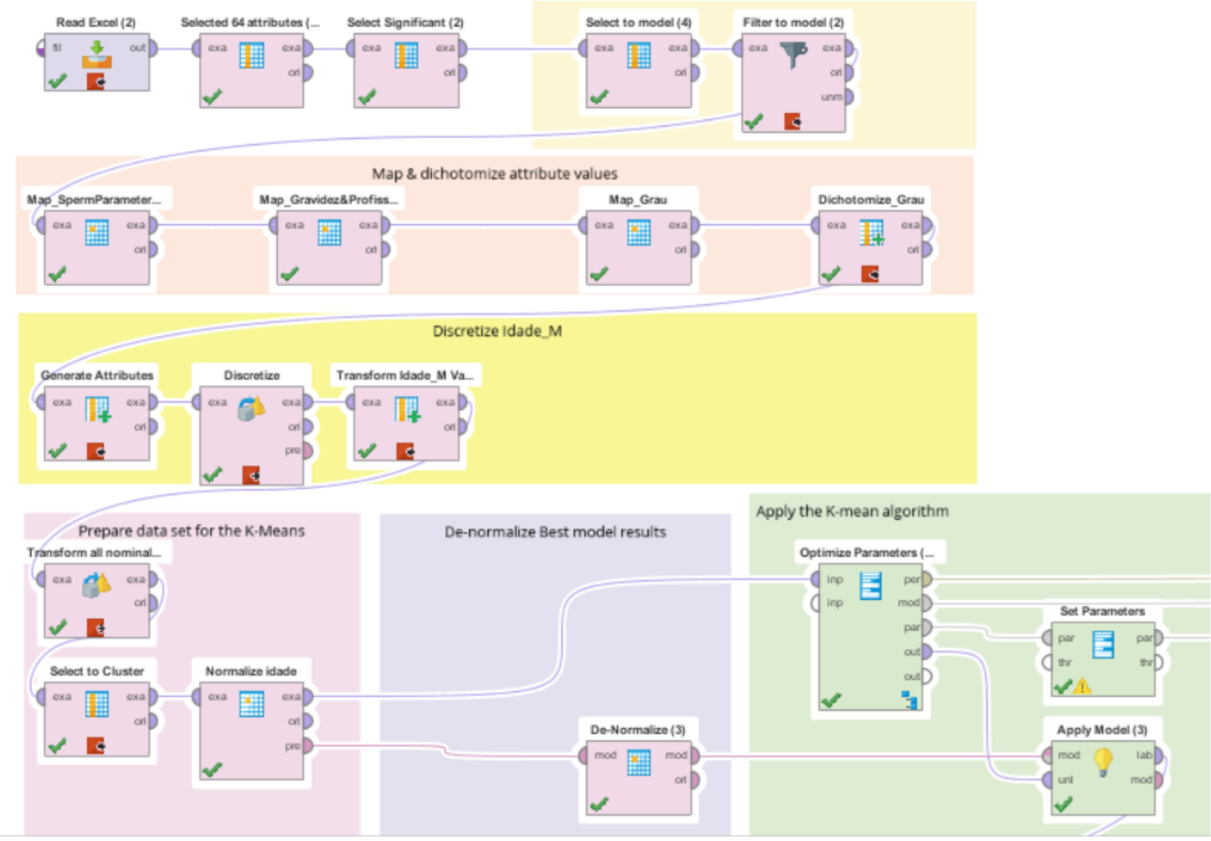


Figure 6.26 “Grau_Varicoc” manually dichotomized and “Idade_M” discretized (Model 5)

B.3 Association with FP-Growth

The FP_Growth algorithm was applied in 6 modeling steps that are summarized in Table 6.6.

Note: the names of the following attributes were for formatting reasons summarized as: “ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos” to “ProfissãoComRisco”, “HabitosTabagicos_Processado_Simplificado” to “HabitosTabagicos_Simplificado”, and “HabitosAlcoolicos_Processado_Simplificado” to “HabitosAlcoolicos_Simplificado).

Table 6.6 Modeling steps of the application of the FP-Growth Algorithm

Step Number	Tested Attributes	Settings	Task performed

1	Grau_Varicoc, Conc_6M, A_B_pré, Formas_N_3M, ProfissãoComRisco, Gravidez	Support=0.1
		Confidence=0.8
			Applied the *FP_Growth* algorithm with the model 1 depicted in Figure 6.27 FP-Growth model. This model applied the association algorithm upon the first group of selected attributes with a *support* set to 0.1 and a *confidence* set to 0.8 to identify objectively interesting rules ordered by its *support*.

2	Grau_Varicoc, Conc_6M, A_B_pré, Formas_N_3M, ProfissãoComRisco, Gravidez	Support=0.0
		Confidence=0.0
			Applied the *FP_Growth* algorithm with model 1 presented in Figure 6.27. This step also applies the association algorithm upon the first group of selected attributes but sets the *support* to 0.0 and *confidence* set to 0.0 to identify subjectively interesting rules.

3	Grau_Varicoc, Conc_6M, A_B_pré, Formas_N_3M, ProfissãoComRisco, Gravidez	Filtered by non-missing values in the “Gravidez” Attribute
		and
		Support=0.0
		Confidence=0.0
		and at last, adjusted to:
		Support=0.1
		Confidence=0.4
			Filtered the data set by the instances that have non-missing values in the “Gravidez” attribute and reapplied the *FP_Growth* algorithm with the model depicted in Figure 6.28 (model 2). This step applies the model upon the first group of selected attributes to seek for interesting rules with a *support* and *confidence* maintained to 0.0 and afterwards, fine-tunes the thresholds to support=0.1 and confidence=0.4 to seek for objectively and subjectively interesting rules based on the conditions set (i.e. conditions disclosed in section 4.2.8.3).

4	Grau_Varicoc, Conc_3M_Qualificado, A_B_Pre_Qualificado, A_B_3M_Qualificado, ProfissãoComRisco, Gravidez.	Filtered by non-missing values in the “Gravidez” Attribute
		and
		Support=0.1
		Confidence=0.4
			Applied the *FP_Growth* algorithm with the model presented in Figure 6.28 (model 2), upon the second group of selected attributes to seek for objectively and subjectively interesting rules based on the conditions set (i.e. conditions disclosed in section 4.2.8.3).

5	Grau_Varicoc, Qualificar_Espermograma_Pre, Qualificar_Espermograma_3M, ProfissãoComRisco, Gravidez.	Filtered by non-missing values in the “Gravidez” Attribute
		and
		Support=0.1
		Confidence=0.4
			Applied the *FP_Growth* algorithm with the model presented in Figure 6.28 (model 2), upon the third group of selected attributes to seek for objectively and subjectively interesting rules based on the conditions set (i.e. conditions disclosed in section 4.2.8.3).

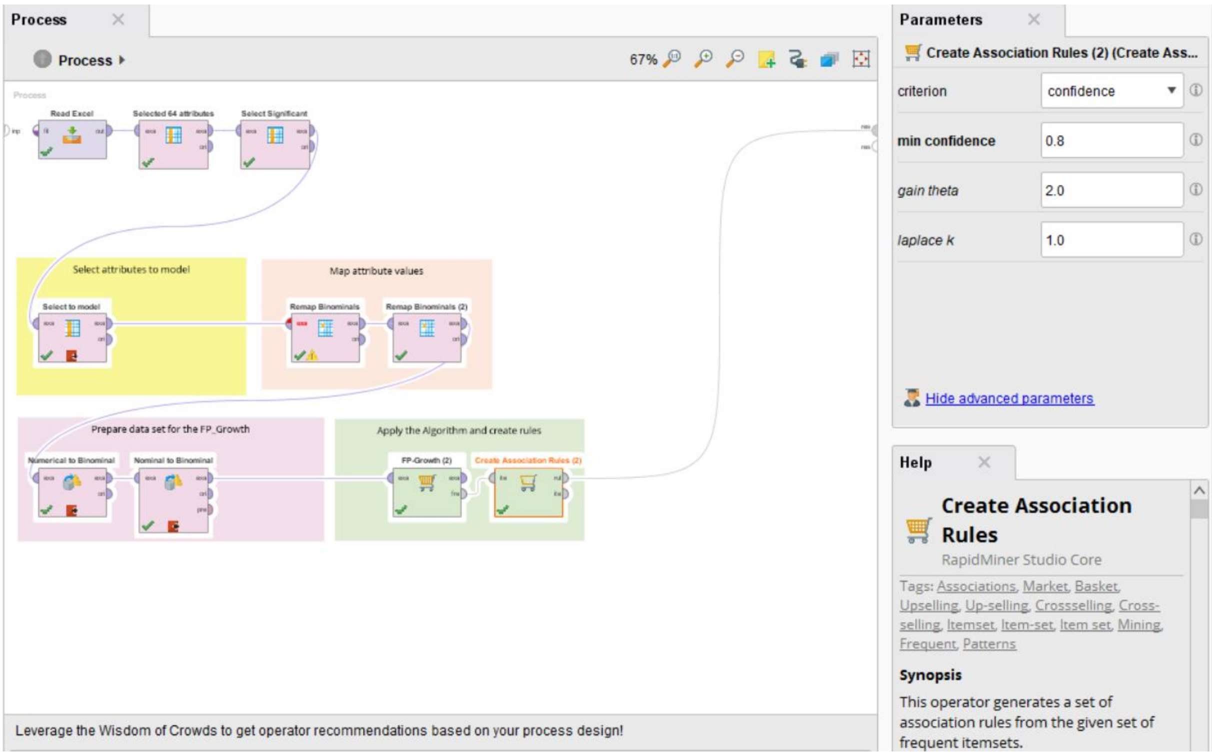
Step Number	Tested Attributes	Settings	Task performed

6	Idade_M, Grau_Varicoc, Conc_3M , Conc_6M, A_B_pré, A_B_3M, Formas_N_3M, Qualificar_Espermograma_Pre, Qualificar_Espermograma_3M, ProfissãoComRisco, HabitosTabagicos_Simplificado, HabitosAlcoolicos_Simplificado, Gravidez, PMA, Gravidez_espontanea	Filtered by non-missing values in the "Gravidez" Attribute
		and
		Support=0.1
		Confidence=0.4
			Applied the FP_Growth algorithm with the models presented in Figure 6.29 (model 3) (to mainly test the sperm parameter values discretized) and Figure 6.30 (model 4) (to mainly test the semen classifications) to further explore the groups of selected attributes already tested to seek for other objectively and subjectively interesting rules based on the conditions set (i.e. conditions disclosed in section 4.2.8.3). This step entailed the addition of more attributes, as well as the discretization of all numerical attributes.

All results are presented in Appendix B., and the identification of the most interesting rules of these first 5 modeling steps are shown in section 5.5.3, as well as the results of the final step 6 since it is the most interesting test that we have carried out with the FP-Growth algorithm in terms of aspects assessed. The reason behind the selection of the tested attributes in this sixth step of the application of the FP-Growth algorithm is also shown in section 5.5.3.

In Figure 6.27, we present the model that was built to implement the first and second step of the application of the FP_Growth algorithm upon all 293 instances; afterwards, in Figure 6.28, we present the model that was built for the third to fifth step upon the 230 filtered instances; in Figure 6.29, we disclose the model that was built for the last sixth step to test all numerical attributes discretized in these 230 filtered instances and the final model depicted in Figure 6.30, depicts the same previous model but adapted to test the attributes related with semen classifications.

If we analyze Figure 6.27 we see that on the left we have the model that began to filter the dataset using the selected attributes, mainly with the operators named "Select Significant" and "Select to model" with the RapidMiner operator called "Select Attributes". Afterwards, the model prepares the data set for the FP_Growth algorithm by mapping the attribute values to "True" and "False" (e.g. "Normal" and "Sim" attribute values were set as positive values; and hence, the algorithm as interpreted them as "True") and by transforming the selected attributes to binomial attribute (e.g. The sperm categorizations were transformed to binomial; and therefore, several new columns were generated to say if a patient was normozoospermic before the treatment or not, or azoospermic or not, etc. to end up with a binomial attribute); then, the FP_Growth algorithm is applied, generating the most frequent item set based on the *support* and *confidence* values set in the above right corner of the figure in the field called respectively *min_support* and *min_confidence*; and at last, it creates the association rules with the last operator called "Create Association Rules" – in this figure appearing with the name "Create Association…" – based on the frequent item sets previously generated.



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Create Association Rules (2) (Create Ass...

criterion: confidence

min confidence: 0.8

gain theta: 2.0

laplace k: 1.0

Create Association Rules

RapidMiner Studio Core

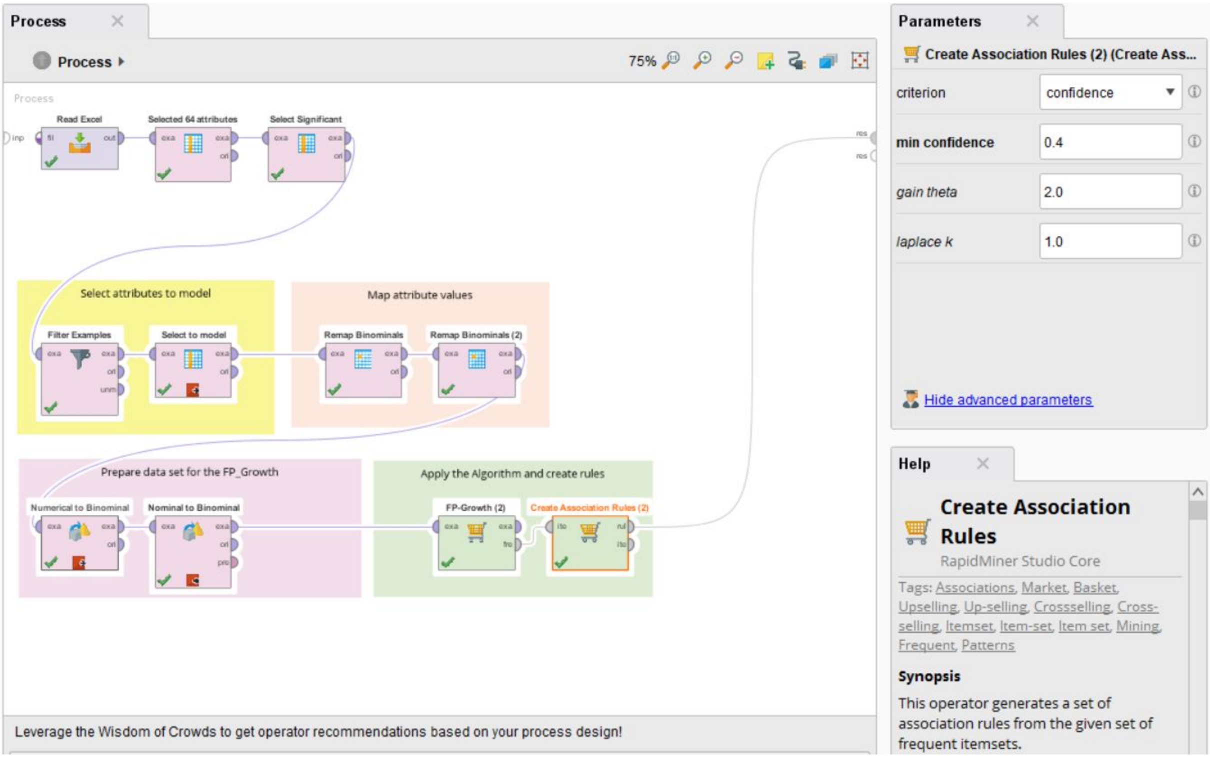
Tags: Associations, Market, Basket, Upselling, Up-selling, Crossselling, Cross-selling, Itemset, Item-set, Item set, Mining, Frequent, Patterns

Synopsis

This operator generates a set of association rules from the given set of frequent itemsets.

Figure 6.27 FP-Growth model 1 (used in step 1 and 2)

If we look at Figure 6.28, we see that the unique difference from the previous figure is that we have a filter operator called “Filter examples” to filter the instances by the non-missing values of the “Gravidez” attribute prior the “select to model” operator. Furthermore, we can see that in the above right corner of the figure, the *min confidence* value is set at the end of this step to 0.4.



Create Association Rules (2) (Create Ass...

criterion: confidence

min confidence: 0.4

gain theta: 2.0

laplace k: 1.0

Create Association Rules

RapidMiner Studio Core

Tags: Associations, Market, Basket, Upselling, Up-selling, Crossselling, Cross-selling, Itemset, Item-set, Item set, Mining, Frequent, Patterns

Synopsis

This operator generates a set of association rules from the given set of frequent itemsets.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 6.28 FP-Growth model 2 (used in step 3 to 5)

If we analyze the model depicted in Figure 6.29, we see that this model mainly adds a set of new operators to the previous model, that can be seen in the center of the model within a gray rectangle. These new operators, named by sperm parameters, discretizes each sperm parameter with the “Discretized by User Specification” operator but before that, uses the “Generate Attributes” operator, here named with the prefix “Copy Sperm Param…”, to create a copy of the discretized attributes to further on validate the computed discretization.

The next model depicted in Figure 6.30, is exactly the same as the model depicted in Figure 6.29, but only has the discretization of woman age.

In spite of not being seen, the main difference between these last two models is also the configuration of the “Select Attributes” and “Filter Examples” operators that can be seen within the yellow rectangle at the top of Figure 6.29 and Figure 6.30. In fact, the “Select Attributes” operator, here called “Select Significant”, was used to select the attributes and the “Filter example” operator, was used to filter the data set by: non-missing values in the “Gravidez” attribute (test 6.1 and 6.2 of the 6th step); the “Gravidez” attribute set to “Sim” (test 6.1.1 and 6.2.1 of the 6th step); the “Gravidez_espontanea” attribute set to “Sim” (test 6.2.2 of the 6th step) and the “PMA” attribute set to “Sim” (test 6.2.3 of the 6th step).

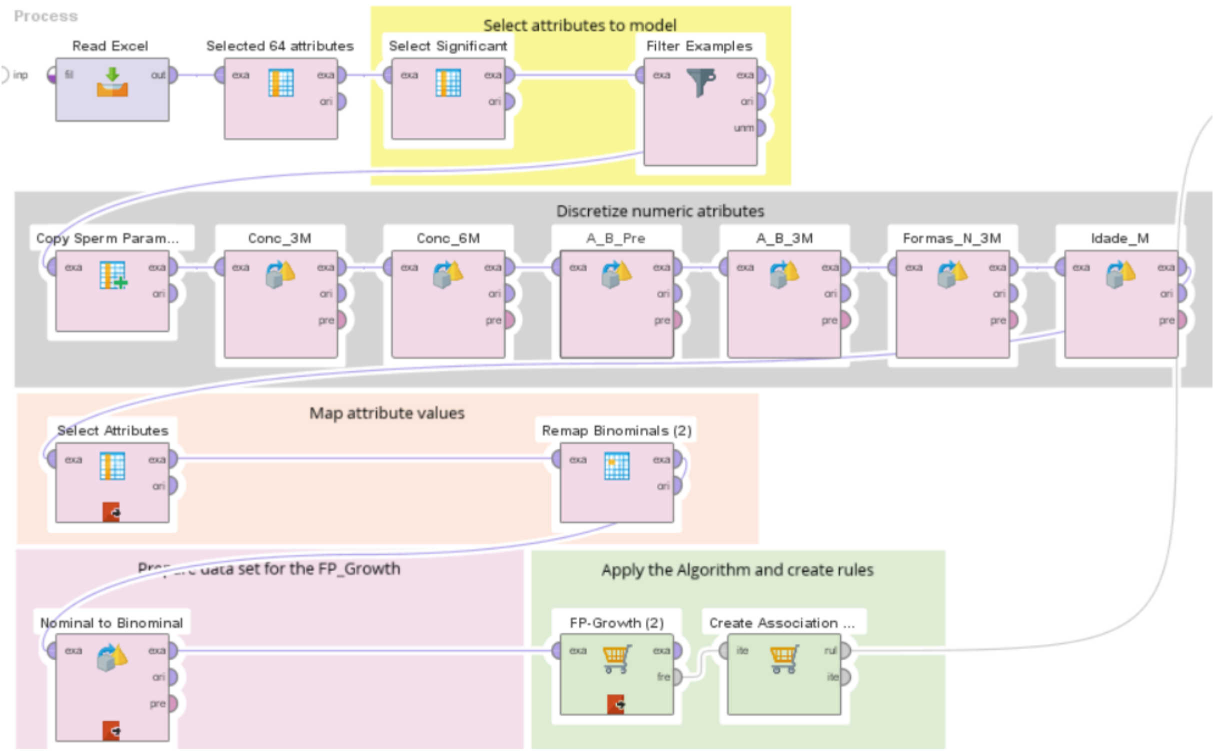


Figure 6.29 FP-Growth model 3 (used in step 6 - test 6.1)

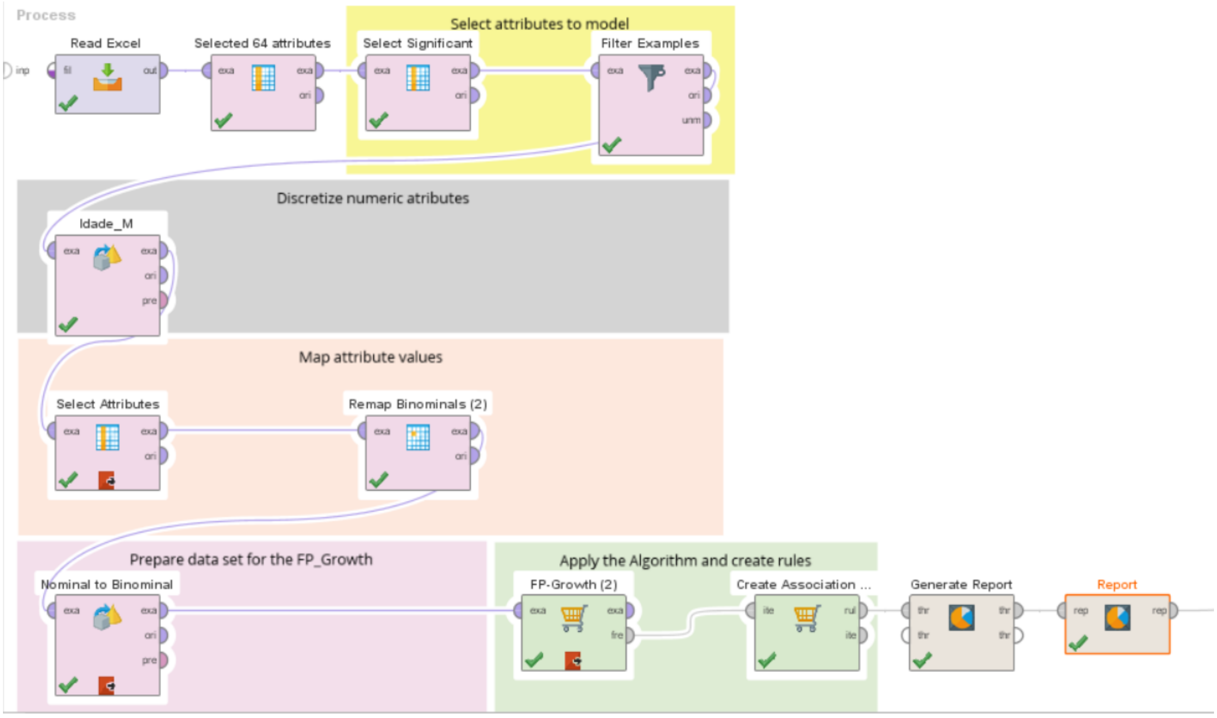


Figure 6.30 FP-Growth model 4 (used in step 6 - test 6.2)

Appendix C: Models´ results

In this section, we present the results that were obtained through the application of the several Data mining techniques that this study has applied upon the *final preprocessed dataset*. Hence, section C.1, presents the results obtained during the application of the *Classification* Data mining technique; section C.2, presents the results obtained with the K-means algorithm during the application of the *Clustering* Data mining technique and section C.3, presents the ones obtained with the FP-Growth algorithm during the application of the *Association* Data mining technique.

C.1 Classification

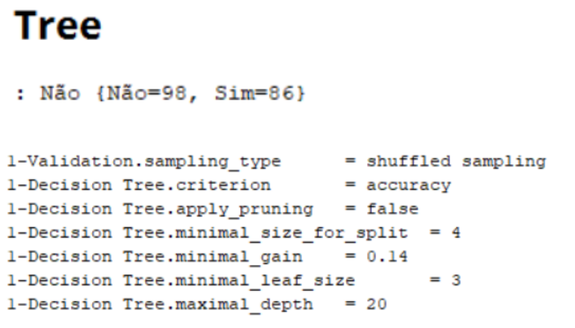
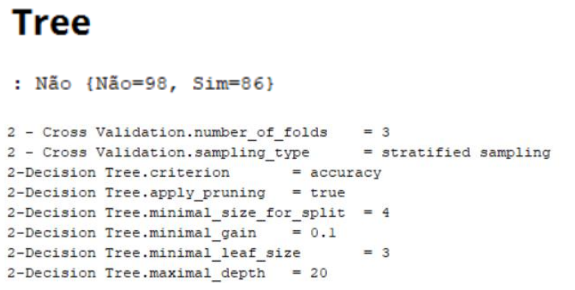
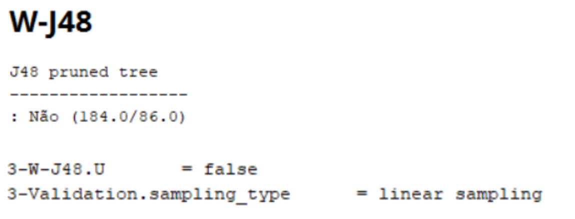
In this section, we present in the below Table C.1 1 the best results achieved with the Decision tree´s modeling steps disclosed in Table 6.1.

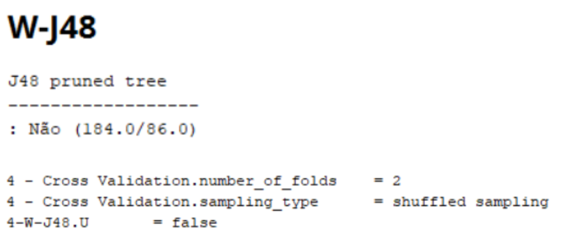
The step numbers specified under the column named "Step Nº" of Table C.1 1 Best Decision tree model´s results

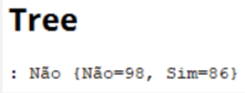
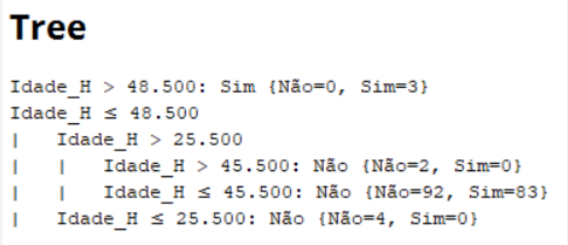
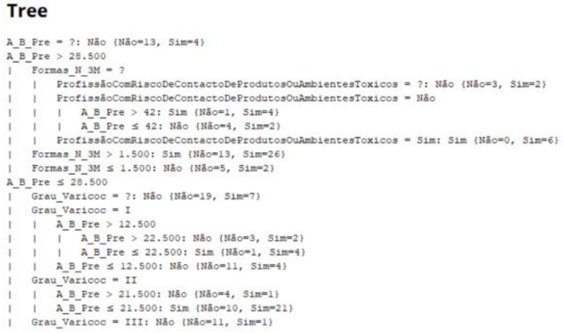
specify the modeling step along with the testing step number (e.g. the row identified with the step 2.4, showcases the best result obtained during the fourth testing step of the decision tree model during the second decision tree modeling step). The best F-Measures per modeling step were highlighted in orange, as well as the best validation test, which corresponds to the elected model described in the modelling section 5.5.1.

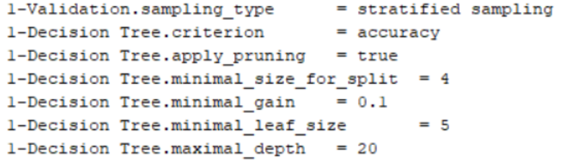
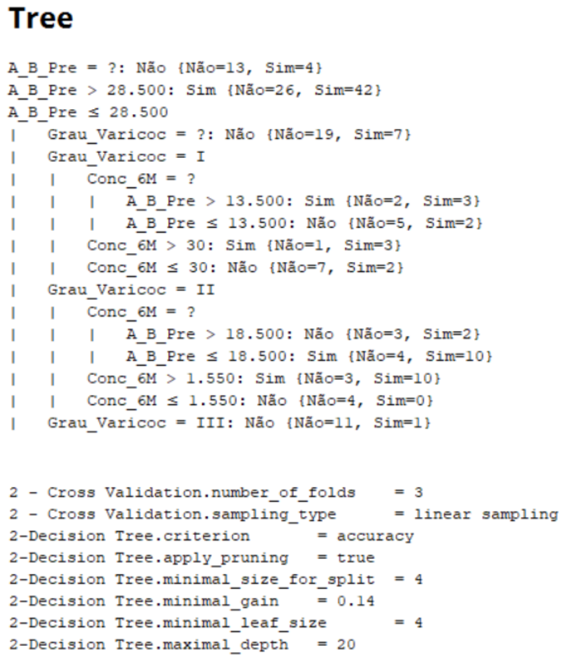
We recall that the results disclosed in step 1 to 5, inclusively, were computed with the decision tree model depicted in Figure 6.4; in step 6 and 7, with the fine-tuned decision tree model depicted in Figure 6.5 and Figure 6.8, respectively; and in step 8, with the fine-tuned decision tree model depicted in Figure 6.9.

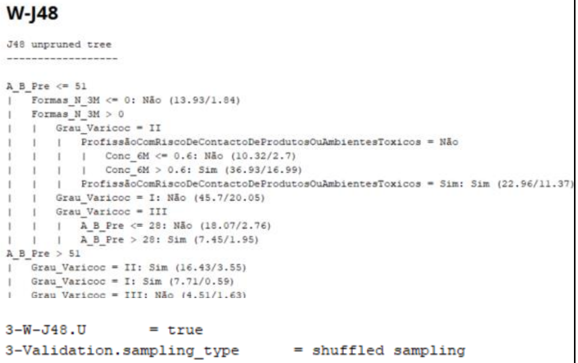
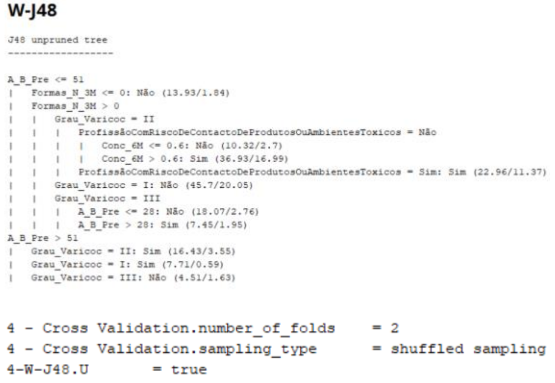
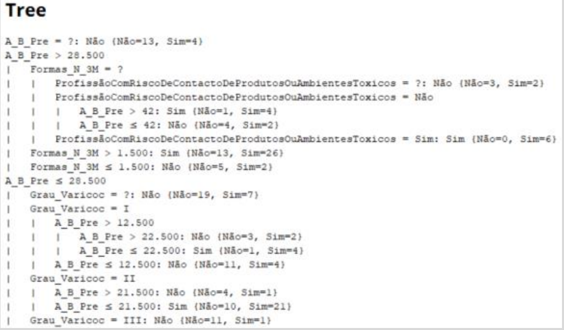
Table C.1 1 Best Decision tree model´s results

Step Nº	Tested Attributes	Attribute´s Transformations	Algorithm	Model´s Testing	Accuracy	Precision	Recall	F-Measure	AUC	Output (Decision Tree and related Model´s Parameters)
1.1	Idade_H, Idade_M, Cirurgias, Doença, Factor_Infertilidade_Masculino, Grau_Varicoc, HabitosAlcoolicos, HabitosTabagicos, Lateralidade, Profissao, Conc_3M, Conc_6M, Conc_1A, A_B_Pre, A_B_3M, A_B_6M, A_B_1A, Formas_N_Pre, Formas_N_3M, Gravidez	No	RapidMiner´s Decision tree	Simple Validation	67.27%	60.71%	70.83%	65.38%	0.680	 **Tree** : Não {Não=98, Sim=86} 1-Validation.sampling_type = shuffled sampling 1-Decision Tree.criterion = accuracy 1-Decision Tree.apply_pruning = false 1-Decision Tree.minimal_size_for_split = 4 1-Decision Tree.minimal_gain = 0.14 1-Decision Tree.minimal_leaf_size = 3 1-Decision Tree.maximal_depth = 20
1.2				Cross Validation	58.14% +/- 3.00%	57.38%	40.23% +/- 28.48%	47.62%	0.575 +/- 0.054	 **Tree** : Não {Não=98, Sim=86} 2 - Cross Validation.number_of_folds = 3 2 - Cross Validation.sampling_type = stratified sampling 2-Decision Tree.criterion = accuracy 2-Decision Tree.apply_pruning = true 2-Decision Tree.minimal_size_for_split = 4 2-Decision Tree.minimal_gain = 0.1 2-Decision Tree.minimal_leaf_size = 3 2-Decision Tree.maximal_depth = 20
1.3			W-J48	Simple Validation	60.00%	unkown	0.00%	unknown	0.500	 **W-J48** J48 pruned tree ------------------ : Não (184.0/86.0) 3-W-J48.U = false 3-Validation.sampling_type = linear sampling

<table>
<tr><td>1.4</td><td></td><td></td><td></td><td>Cross Validation</td><td>54.35% +/- 0.00%</td><td>58.33%</td><td>7.95% +/- 7-95%</td><td>14.29%</td><td>0.498 +/- 0.002</td><td>
W-J48
J48 pruned tree

: Não (184.0/86.0)
4 - Cross Validation.number_of_folds = 2
4 - Cross Validation.sampling_type = shuffled sampling
4-W-J48.U = false</td></tr>
<tr><td colspan="3">Validation of the right model found in step 1 with the best parameter´s settings found for the operator called “1-Decision tree” which are related to the highest f-measure found in step 1.1 that is above highlighted in orange color.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>54.35%</td><td>unknown</td><td>0.00%</td><td>unknown</td><td>0.500</td><td>
Tree
: Não {Não=98, Sim=86}</td></tr>
<tr><td colspan="3">Validation of the model in step 1 with the lowest parameter´s settings (i.e. minimal size for split=4; minimal gain=0.018; minimal leaf size=2; maximal depth=20 and without prunning) and the splitting criterion gain ratio.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>52.17%</td><td>0.00%</td><td>0.00%</td><td>unknown</td><td>0.500</td><td>
Tree
Idade_H > 48.500: Sim {Não=0, Sim=3}
Idade_H ≤ 48.500
| Idade_H > 25.500
| | Idade_H > 45.500: Não {Não=2, Sim=0}
| | Idade_H ≤ 45.500: Não {Não=92, Sim=83}
| Idade_H ≤ 25.500: Não {Não=4, Sim=0}</td></tr>
<tr><td>2.1</td><td>A_B_Pre
Conc_6M
Formas_N_3M
Grau_Varicoc
Gravidez
ProfissãoComRiscoDeContacto…</td><td>No</td><td>RapidMiner´s Decision tree</td><td>Simple Validation</td><td>67.27%</td><td>68.18%</td><td>57.69%</td><td>62.50%</td><td>0.701</td><td>
Tree
A_B_Pre = ?: Não {Não=13, Sim=4}
A_B_Pre > 28.500
| Formas_N_3M = ?
| | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?: Não {Não=3, Sim=2}
| | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não
| | | A_B_Pre > 42: Sim {Não=1, Sim=4}
| | | A_B_Pre ≤ 42: Não {Não=4, Sim=2}
| | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Sim {Não=0, Sim=6}
| Formas_N_3M > 1.500: Sim {Não=13, Sim=26}
| Formas_N_3M ≤ 1.500: Não {Não=5, Sim=2}
A_B_Pre ≤ 28.500
| Grau_Varicoc = ?: Não {Não=19, Sim=7}
| Grau_Varicoc = I
| | A_B_Pre > 12.500
| | | A_B_Pre > 22.500: Não {Não=3, Sim=2}
| | | A_B_Pre ≤ 22.500: Sim {Não=1, Sim=4}
| | A_B_Pre ≤ 12.500: Não {Não=11, Sim=4}
| Grau_Varicoc = II
| | A_B_Pre > 21.500: Não {Não=4, Sim=1}
| | A_B_Pre ≤ 21.500: Sim {Não=10, Sim=21}
| Grau_Varicoc = III: Não {Não=11, Sim=1}</td></tr>
</table>

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>
1-Validation.sampling_type   = stratified sampling
1-Decision Tree.criterion   = accuracy
1-Decision Tree.apply_pruning   = true
1-Decision Tree.minimal_size_for_split = 4
1-Decision Tree.minimal_gain   = 0.1
1-Decision Tree.minimal_leaf_size   = 5
1-Decision Tree.maximal_depth   = 20</td></tr>
<tr><td>2.2</td><td></td><td></td><td></td><td>Cross Validation</td><td>60.33% +/- 2.72%</td><td>65.84% +/- 7.41%</td><td>38.57% +/- 15.19%</td><td>45.88% +/- 8.39%</td><td>0.638 +/- 0.029</td><td>
Tree
A_B_Pre = ?: Não {Não=13, Sim=4}
A_B_Pre > 28.500: Sim {Não=26, Sim=42}
A_B_Pre ≤ 28.500
| Grau_Varicoc = ?: Não {Não=19, Sim=7}
| Grau_Varicoc = I
| | Conc_6M = ?
| | | A_B_Pre > 13.500: Sim {Não=2, Sim=3}
| | | A_B_Pre ≤ 13.500: Não {Não=5, Sim=2}
| | Conc_6M > 30: Sim {Não=1, Sim=3}
| | Conc_6M ≤ 30: Não {Não=7, Sim=2}
| Grau_Varicoc = II
| | Conc_6M = ?
| | | A_B_Pre > 18.500: Não {Não=3, Sim=2}
| | | A_B_Pre ≤ 18.500: Sim {Não=4, Sim=10}
| | Conc_6M > 1.550: Sim {Não=3, Sim=10}
| | Conc_6M ≤ 1.550: Não {Não=4, Sim=0}
| Grau_Varicoc = III: Não {Não=11, Sim=1}
2 - Cross Validation.number_of_folds   = 3
2 - Cross Validation.sampling_type   = linear sampling
2-Decision Tree.criterion   = accuracy
2-Decision Tree.apply_pruning   = true
2-Decision Tree.minimal_size_for_split = 4
2-Decision Tree.minimal_gain   = 0.14
2-Decision Tree.minimal_leaf_size   = 4
2-Decision Tree.maximal_depth   = 20</td></tr>
</table>

2.3			W-J48	Simple Validation	61.82%	58.82%	41.67%	48.78%	0.603	
2.4				Cross Validation	58.70% +/- 2.17%	57.88% +/- 2.12%	48.16% +/- 20.89%	49.60% +/- 12.10%	0.570 +/- 0.051	
Validation of the right model found in step 2 with the best parameter´s settings found for the operator called “1-Decision tree” which are related to the highest f-measure found in step 2.1 that is above highlighted in orange color.			RapidMiner´s Decision tree	-	56.52%	52.38%	52.38%	52.38%	0.509	

W-J48

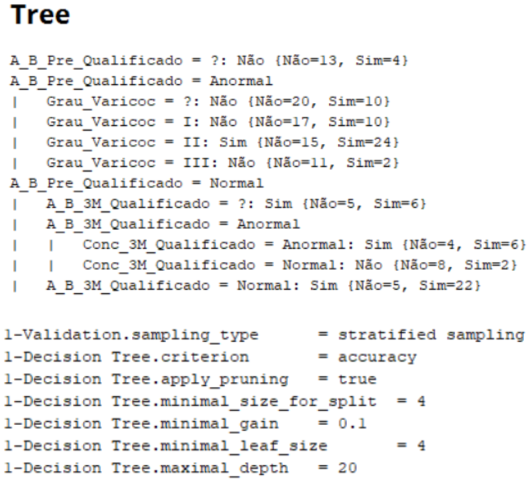
```
J48 unpruned tree
------------------

A_B_Pre <= 51
|   Formas_N_3M <= 0: Não (13.93/1.84)
|   Formas_N_3M > 0
|   |   Grau_Varicoc = II
|   |   |   ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não
|   |   |   |   Conc_6M <= 0.6: Não (10.32/2.7)
|   |   |   |   Conc_6M > 0.6: Sim (36.93/16.99)
|   |   |   ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Sim (22.96/11.37)
|   |   Grau_Varicoc = I: Não (45.7/20.05)
|   |   Grau_Varicoc = III
|   |   |   A_B_Pre <= 28: Não (18.07/2.76)
|   |   |   A_B_Pre > 28: Sim (7.45/1.95)
A_B_Pre > 51
|   Grau_Varicoc = II: Sim (16.43/3.55)
|   Grau_Varicoc = I: Sim (7.71/0.59)
|   Grau_Varicoc = III: Não (4.51/1.63)

3-W-J48.U        = true
3-Validation.sampling_type      = shuffled sampling
```

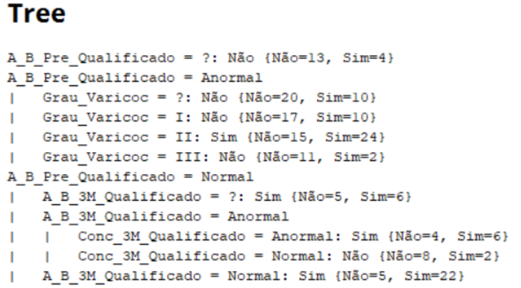
W-J48

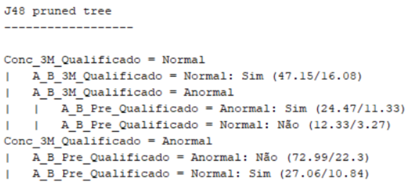
```
J48 unpruned tree
------------------

A_B_Pre <= 51
|   Formas_N_3M <= 0: Não (13.93/1.84)
|   Formas_N_3M > 0
|   |   Grau_Varicoc = II
|   |   |   ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não
|   |   |   |   Conc_6M <= 0.6: Não (10.32/2.7)
|   |   |   |   Conc_6M > 0.6: Sim (36.93/16.99)
|   |   |   ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Sim (22.96/11.37)
|   |   Grau_Varicoc = I: Não (45.7/20.05)
|   |   Grau_Varicoc = III
|   |   |   A_B_Pre <= 28: Não (18.07/2.76)
|   |   |   A_B_Pre > 28: Sim (7.45/1.95)
A_B_Pre > 51
|   Grau_Varicoc = II: Sim (16.43/3.55)
|   Grau_Varicoc = I: Sim (7.71/0.59)
|   Grau_Varicoc = III: Não (4.51/1.63)


4 - Cross Validation.number_of_folds    = 2
4 - Cross Validation.sampling_type      = shuffled sampling
4-W-J48.U       = true
```

Tree

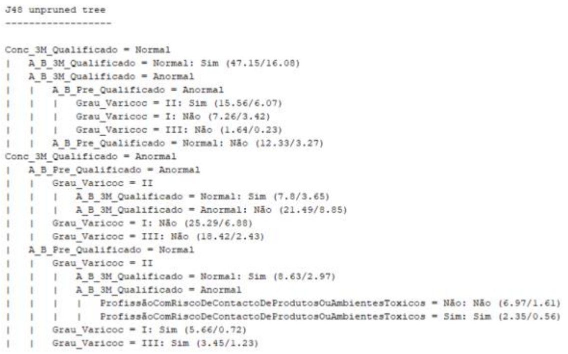
```
A_B_Pre = ?: Não {Não=13, Sim=4}
A_B_Pre > 28.500
|   Formas_N_3M = ?
|   |   ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?: Não {Não=3, Sim=2}
|   |   ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não
|   |   |   A_B_Pre > 42: Sim {Não=1, Sim=4}
|   |   |   A_B_Pre ≤ 42: Não {Não=4, Sim=2}
|   |   ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Sim {Não=0, Sim=6}
|   Formas_N_3M > 1.500: Sim {Não=13, Sim=26}
|   Formas_N_3M ≤ 1.500: Não {Não=5, Sim=2}
A_B_Pre ≤ 28.500
|   Grau_Varicoc = ?: Não {Não=19, Sim=7}
|   Grau_Varicoc = I
|   |   A_B_Pre > 12.500
|   |   |   A_B_Pre > 22.500: Não {Não=3, Sim=2}
|   |   |   A_B_Pre ≤ 22.500: Sim {Não=1, Sim=4}
|   |   A_B_Pre ≤ 12.500: Não {Não=11, Sim=4}
|   Grau_Varicoc = II
|   |   A_B_Pre > 21.500: Não {Não=4, Sim=1}
|   |   A_B_Pre ≤ 21.500: Sim {Não=10, Sim=21}
|   Grau_Varicoc = III: Não {Não=11, Sim=1}
```

<table>
<tr><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td colspan="3">Validation of the model in step 2 with the lowest parameter´s settings (i.e. minimal size for split=4; minimal gain=0.018; minimal leaf size=2; maximal depth=20 and without prunning) and the splitting criterion gain ratio.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>60.87%</td><td>60.00%</td><td>42.86%</td><td>50.00%</td><td>0.572</td><td>(generated a too long decision tree)</td></tr>
<tr><td>3.1</td><td>A_B_Pre_Qualificado
A_B_3M_Qualificado
Conc_3M_Qualificado
Grau_Varicoc
Gravidez
ProfissãoComRiscoDeContacto…</td><td>No</td><td>RapidMiner´s Decision tree</td><td>Simple Validation</td><td>74.55%</td><td>73.08%</td><td>73.08%</td><td>73.08%</td><td>0.747</td><td>
Tree
A_B_Pre_Qualificado = ?: Não {Não=13, Sim=4}
A_B_Pre_Qualificado = Anormal
| Grau_Varicoc = ?: Não {Não=20, Sim=10}
| Grau_Varicoc = I: Não {Não=17, Sim=10}
| Grau_Varicoc = II: Sim {Não=15, Sim=24}
| Grau_Varicoc = III: Não {Não=11, Sim=2}
A_B_Pre_Qualificado = Normal
| A_B_3M_Qualificado = ?: Sim {Não=5, Sim=6}
| A_B_3M_Qualificado = Anormal
| | Conc_3M_Qualificado = Anormal: Sim {Não=4, Sim=6}
| | Conc_3M_Qualificado = Normal: Não {Não=8, Sim=2}
| A_B_3M_Qualificado = Normal: Sim {Não=5, Sim=22}

1-Validation.sampling_type   = stratified sampling
1-Decision Tree.criterion   = accuracy
1-Decision Tree.apply_pruning   = true
1-Decision Tree.minimal_size_for_split = 4
1-Decision Tree.minimal_gain   = 0.1
1-Decision Tree.minimal_leaf_size   = 4
1-Decision Tree.maximal_depth   = 20</td></tr>
<tr><td>3.2</td><td></td><td></td><td></td><td>Cross Validation</td><td>62.50% +/- 1.63%</td><td>60.28% +/- 3.14%</td><td>58.12% +/- 0.97%</td><td>59.16% +/- 2.02%</td><td>0.613 +/- 0.053</td><td>
Tree
A_B_Pre_Qualificado = ?: Não {Não=13, Sim=4}
A_B_Pre_Qualificado = Anormal
| Grau_Varicoc = ?: Não {Não=20, Sim=10}
| Grau_Varicoc = I: Não {Não=17, Sim=10}
| Grau_Varicoc = II: Sim {Não=15, Sim=24}
| Grau_Varicoc = III: Não {Não=11, Sim=2}
A_B_Pre_Qualificado = Normal
| A_B_3M_Qualificado = ?: Sim {Não=5, Sim=6}
| A_B_3M_Qualificado = Anormal
| | Conc_3M_Qualificado = Anormal: Sim {Não=4, Sim=6}
| | Conc_3M_Qualificado = Normal: Não {Não=8, Sim=2}
| A_B_3M_Qualificado = Normal: Sim {Não=5, Sim=22}</td></tr>
</table>

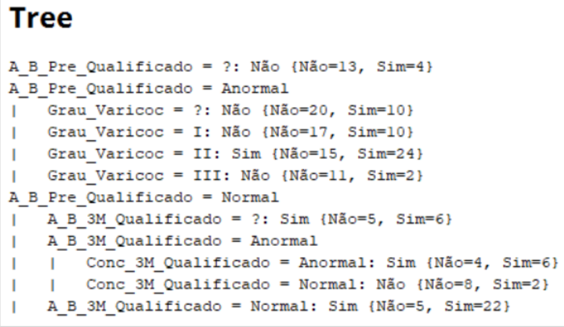
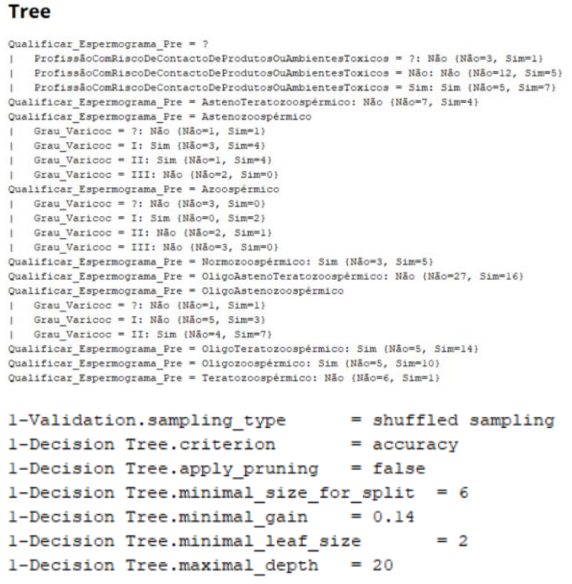
<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>2 - Cross Validation.number_of_folds = 2
2 - Cross Validation.sampling_type = shuffled sampling
2-Decision Tree.criterion = accuracy
2-Decision Tree.apply_pruning = true
2-Decision Tree.minimal_size_for_split = 4
2-Decision Tree.minimal_gain = 0.1
2-Decision Tree.minimal_leaf_size = 4
2-Decision Tree.maximal_depth = 20</td></tr>
<tr><td>3.3</td><td></td><td></td><td>W-J48</td><td>Simple Validation</td><td>63.64%</td><td>57.14%</td><td>36.36%</td><td>44.44%</td><td>0.606</td><td><b>W-J48</b>

J48 pruned tree

Conc_3M_Qualificado = Normal
| A_B_3M_Qualificado = Normal: Sim (47.15/16.08)
| A_B_3M_Qualificado = Anormal
| | A_B_Pre_Qualificado = Anormal: Sim (24.47/11.33)
| | A_B_Pre_Qualificado = Normal: Não (12.33/3.27)
Conc_3M_Qualificado = Anormal
| A_B_Pre_Qualificado = Anormal: Não (72.99/22.3)
| A_B_Pre_Qualificado = Normal: Sim (27.06/10.84)

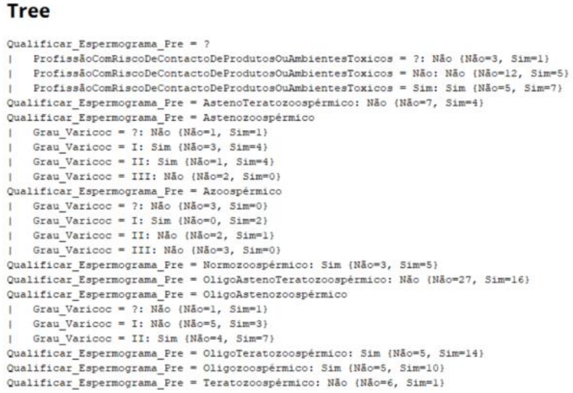
3-W-J48.U = false
3-Validation.sampling_type = linear sampling</td></tr>
<tr><td>3.4</td><td></td><td></td><td></td><td>Cross Validation</td><td>63.59% +/- 3.80%</td><td>61.63% +/- 2.66%</td><td>58.28% +/- 6.01%</td><td>59.85% +/- 4.43%</td><td>0.621 +/- 0.001</td><td><b>W-J48</b>

J48 unpruned tree

Conc_3M_Qualificado = Normal
| A_B_3M_Qualificado = Normal: Sim (47.15/16.08)
| A_B_3M_Qualificado = Anormal
| | A_B_Pre_Qualificado = Anormal
| | | Grau_Varicoc = II: Sim (15.56/6.07)
| | | Grau_Varicoc = I: Não (7.26/3.42)
| | | Grau_Varicoc = III: Não (1.64/0.23)
| | A_B_Pre_Qualificado = Normal: Não (12.33/3.27)
Conc_3M_Qualificado = Anormal
| A_B_Pre_Qualificado = Anormal
| | Grau_Varicoc = II
| | | A_B_3M_Qualificado = Normal: Sim (7.8/3.65)
| | | A_B_3M_Qualificado = Anormal: Não (21.49/8.85)
| | Grau_Varicoc = I: Não (25.29/6.88)
| | Grau_Varicoc = III: Não (18.42/2.43)
| A_B_Pre_Qualificado = Normal
| | Grau_Varicoc = II
| | | A_B_3M_Qualificado = Normal: Sim (8.63/2.97)
| | | A_B_3M_Qualificado = Anormal
| | | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não: Não (6.97/1.61)
| | | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Sim (2.35/0.56)
| | Grau_Varicoc = I: Sim (5.66/0.72)
| | Grau_Varicoc = III: Sim (3.45/1.23)

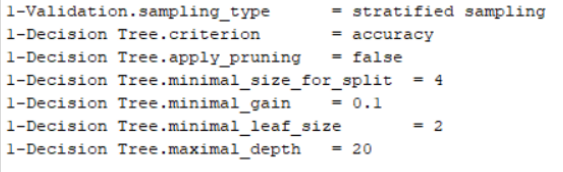
4 - Cross Validation.number_of_folds = 2
4 - Cross Validation.sampling_type = shuffled sampling
4-W-J48.U = true</td></tr>
</table>

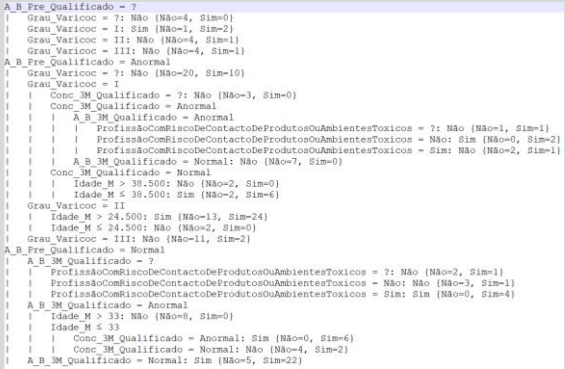
<table>
<tr><td colspan="3">Validation of the right model found in step 3 with the best parameter´s settings found for the operator called “1-Decision tree” which are related to the highest f-measure found in step 3.1 that is above highlighted in orange color.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>47.83%</td><td>44.00%</td><td>52.38%</td><td>47.83%</td><td>0.554</td><td>
Tree
<pre>
A_B_Pre_Qualificado = ?: Não {Não=13, Sim=4}
A_B_Pre_Qualificado = Anormal
| Grau_Varicoc = ?: Não {Não=20, Sim=10}
| Grau_Varicoc = I: Não {Não=17, Sim=10}
| Grau_Varicoc = II: Sim {Não=15, Sim=24}
| Grau_Varicoc = III: Não {Não=11, Sim=2}
A_B_Pre_Qualificado = Normal
| A_B_3M_Qualificado = ?: Sim {Não=5, Sim=6}
| A_B_3M_Qualificado = Anormal
| | Conc_3M_Qualificado = Anormal: Sim {Não=4, Sim=6}
| | Conc_3M_Qualificado = Normal: Não {Não=8, Sim=2}
| A_B_3M_Qualificado = Normal: Sim {Não=5, Sim=22}
</pre></td></tr>
<tr><td colspan="3">Validation of the model in step 3 with the lowest parameter´s settings (i.e. minimal size for split=4; minimal gain=0.018; minimal leaf size=2; maximal depth=20 and without prunning) and the splitting criterion gain ratio.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>60.87%</td><td>55.56%</td><td>71.43%</td><td>62.50%</td><td>0.595</td><td>(generated a too long decision tree)</td></tr>
<tr><td>4.1</td><td>Grau_Varicoc
Gravidez
ProfissãoComRiscoDeContacto...
Qualificar_Espermograma_3M
Qualificar_Espermograma_Pre</td><td>No</td><td>RapidMiner´s Decision tree</td><td>Simple Validation</td><td>67.27%</td><td>57.89%</td><td>52.38%</td><td>55.00%</td><td>0.616</td><td>
Tree
<pre>
Qualificar_Espermograma_Pre = ?
| ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?: Não {Não=3, Sim=1}
| ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não: Não {Não=12, Sim=5}
| ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Sim {Não=5, Sim=7}
Qualificar_Espermograma_Pre = AstenoTeratozoospérmico: Não {Não=7, Sim=4}
Qualificar_Espermograma_Pre = Astenozoospérmico
| Grau_Varicoc = ?: Não {Não=1, Sim=1}
| Grau_Varicoc = I: Sim {Não=3, Sim=4}
| Grau_Varicoc = II: Sim {Não=1, Sim=4}
| Grau_Varicoc = III: Não {Não=2, Sim=0}
Qualificar_Espermograma_Pre = Azoospérmico
| Grau_Varicoc = ?: Não {Não=3, Sim=0}
| Grau_Varicoc = I: Sim {Não=0, Sim=2}
| Grau_Varicoc = II: Não {Não=2, Sim=1}
| Grau_Varicoc = III: Não {Não=3, Sim=0}
Qualificar_Espermograma_Pre = Normozoospérmico: Sim {Não=3, Sim=5}
Qualificar_Espermograma_Pre = OligoAstenoTeratozoospérmico: Não {Não=27, Sim=16}
Qualificar_Espermograma_Pre = OligoAstenozoospérmico
| Grau_Varicoc = ?: Não {Não=1, Sim=1}
| Grau_Varicoc = I: Não {Não=5, Sim=3}
| Grau_Varicoc = II: Sim {Não=4, Sim=7}
Qualificar_Espermograma_Pre = OligoTeratozoospérmico: Sim {Não=5, Sim=14}
Qualificar_Espermograma_Pre = Oligozoospérmico: Sim {Não=5, Sim=10}
Qualificar_Espermograma_Pre = Teratozoospérmico: Não {Não=6, Sim=1}

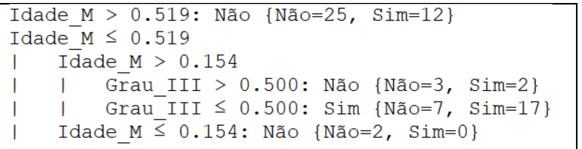
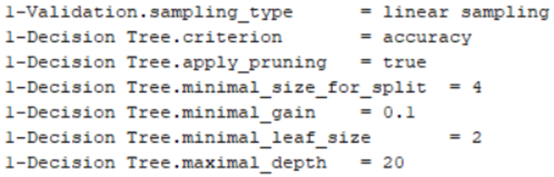
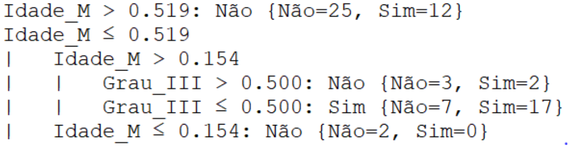
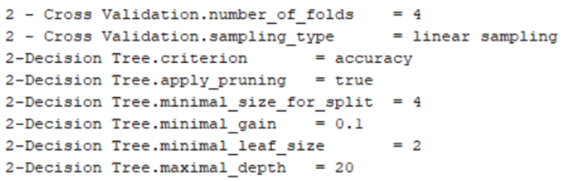
1-Validation.sampling_type   = shuffled sampling
1-Decision Tree.criterion   = accuracy
1-Decision Tree.apply_pruning   = false
1-Decision Tree.minimal_size_for_split = 6
1-Decision Tree.minimal_gain   = 0.14
1-Decision Tree.minimal_leaf_size   = 2
1-Decision Tree.maximal_depth   = 20
</pre></td></tr>
</table>

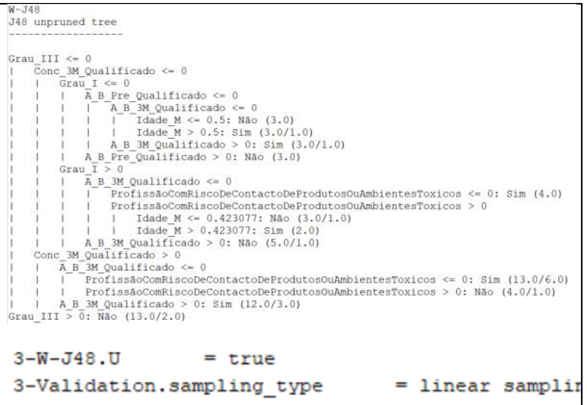
<table>
<tr><td>4.2</td><td></td><td></td><td></td><td>Cross Validation</td><td>61.92% +/- 4.91%</td><td>59.72% +/- 5.20%</td><td>54.61% +/- 8.97%</td><td>56.88% +/- 6.98%</td><td>0.569 +/- 0.076</td><td>Tree
Qualificar_Espermograma_Pre = ?
| ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?: Não {Não=3, Sim=1}
| ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não: Não {Não=12, Sim=5}
| ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Sim {Não=5, Sim=7}
Qualificar_Espermograma_Pre = AstenoTeratozoospérmico: Não {Não=7, Sim=4}
Qualificar_Espermograma_Pre = Astenozoospérmico
| Grau_Varicoc = ?: Não {Não=1, Sim=1}
| Grau_Varicoc = I: Sim {Não=3, Sim=4}
| Grau_Varicoc = II: Sim {Não=1, Sim=4}
| Grau_Varicoc = III: Não {Não=2, Sim=0}
Qualificar_Espermograma_Pre = Azoospérmico
| Grau_Varicoc = ?: Não {Não=3, Sim=0}
| Grau_Varicoc = I: Sim {Não=0, Sim=2}
| Grau_Varicoc = II: Não {Não=2, Sim=1}
| Grau_Varicoc = III: Não {Não=3, Sim=0}
Qualificar_Espermograma_Pre = Normozoospérmico: Sim {Não=3, Sim=5}
Qualificar_Espermograma_Pre = OligoAstenoTeratozoospérmico: Não {Não=27, Sim=16}
Qualificar_Espermograma_Pre = OligoAstenozoospérmico
| Grau_Varicoc = ?: Não {Não=1, Sim=1}
| Grau_Varicoc = I: Não {Não=5, Sim=3}
| Grau_Varicoc = II: Sim {Não=4, Sim=7}
Qualificar_Espermograma_Pre = OligoTeratozoospérmico: Sim {Não=5, Sim=14}
Qualificar_Espermograma_Pre = Oligozoospérmico: Sim {Não=5, Sim=10}
Qualificar_Espermograma_Pre = Teratozoospérmico: Não {Não=6, Sim=1}

2 - Cross Validation.number_of_folds   = 3
2 - Cross Validation.sampling_type   = linear sampling
2-Decision Tree.criterion   = accuracy
2-Decision Tree.apply_pruning   = true
2-Decision Tree.minimal_size_for_split   = 4
2-Decision Tree.minimal_gain   = 0.1
2-Decision Tree.minimal_leaf_size   = 2
2-Decision Tree.maximal_depth   = 20</td></tr>
<tr><td>4.3</td><td></td><td></td><td>W-J48</td><td>Simple Validation</td><td>70.91%</td><td>70.00%</td><td>58.33%</td><td>63.64%</td><td>0.739</td><td>(generated an unpruned tree that was too long)
3-W-J48.U   = true
3-Validation.sampling_type   = shuffled sampling</td></tr>
<tr><td>4.4</td><td></td><td></td><td></td><td>Cross Validation</td><td>66.30% +/- 2.43%</td><td>66.11% +/- 2.20%</td><td>58.12% +/- 12.89%</td><td>60.95% +/- 6.88%</td><td>0.684 +/- 0.053</td><td>(generated an unpruned tree that was too long)
4 - Cross Validation.number_of_folds   = 4
4 - Cross Validation.sampling_type   = stratified sampling
4-W-J48.U   = true</td></tr>
<tr><td colspan="3">Validation results of the right model found in step 4 with the best parameter´s settings found for the operator called “3-W-J48” which is related to the highest f-measure found in step 4.3 that is above highlighted in orange color.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>56.52%</td><td>53.33%</td><td>38.10%</td><td>44.44%</td><td>0.604</td><td>(generated a too long tree)</td></tr>
</table>

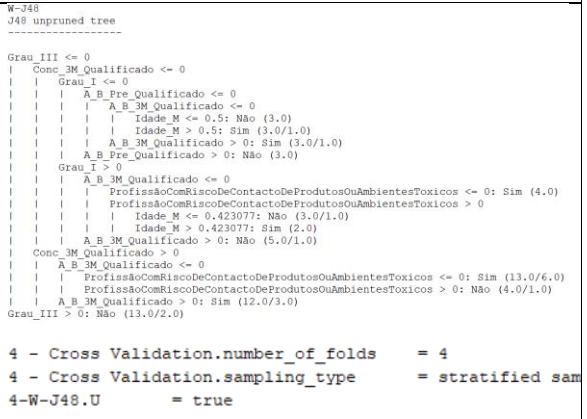
<table>
<tr><td colspan="2">Validation of the model in step 4 with the lowest parameter´s settings (i.e. minimal size for split=4; minimal gain=0.018; minimal leaf size=2; maximal depth=20 and without prunning) and the splitting criterion gain ratio.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>58.70%</td><td>56.25%</td><td>42.86%</td><td>48.65%</td><td>0.613</td><td>(generated a too long tree)</td></tr>
<tr><td>5</td><td colspan="9">Retested the step 2 to 4 by adding the Idade_M attribute to check if we could surpass the f-measure of 73.08% found in step 3 but the max f-measure generated was 70.18%. The best obtained measures were during this step all related with the operator called “1-Decision Tree” and the obtained results were for: Attribute group 4 -> 66.67%; Attribute group 5 -> 70.18% and Attribute group 6 -> 65.22% so the group of attributes that has generated the best result is still the group of attributes related with the categorized sperm parameters values (e.g. A_B_Pre_Qualificado etc).Since this group of attributes also has generated interesting results with the clustering algorithm, we have further on tested the 5th group of attributes in the next 6th and 7th step by transforming its attribute values into numerical and normalized values.</td></tr>
<tr><td colspan="2">Validation results of the right model found in step 5 with the best parameter´s settings found for the operator called “1-Decision tree” which computed the following performance measures during its training/testing: accuracy=69.09%, precision=64.52%, recall=76.92%, f-measure = 70.18%, AUC=0.711.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>58.70%</td><td>53.57%</td><td>71.43%</td><td>61.22%</td><td>0.577</td><td>Model´s parameter values that have computed the best performance measures during the step 5:

1-Validation.sampling_type   = stratified sampling
1-Decision Tree.criterion   = accuracy
1-Decision Tree.apply_pruning   = false
1-Decision Tree.minimal_size_for_split = 4
1-Decision Tree.minimal_gain   = 0.1
1-Decision Tree.minimal_leaf_size   = 2
1-Decision Tree.maximal_depth   = 20

Tree generated by the model applied to the validation data set with the model´s parameter values disclosed above:

A_B_Pre_Qualificado = ?
| Grau_Varicoc = ?: Não {Não=4, Sim=0}
| Grau_Varicoc = I: Sim {Não=1, Sim=2}
| Grau_Varicoc = II: Não {Não=4, Sim=1}
| Grau_Varicoc = III: Não {Não=4, Sim=1}
A_B_Pre_Qualificado = Anormal
| Grau_Varicoc = ?: Não {Não=20, Sim=10}
| Grau_Varicoc = I
| | Conc_3M_Qualificado = ?: Não {Não=3, Sim=0}
| | Conc_3M_Qualificado = Anormal
| | | A_B_3M_Qualificado = Anormal
| | | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?: Não {Não=1, Sim=1}
| | | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não: Sim {Não=0, Sim=2}
| | | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Não {Não=2, Sim=1}
| | | A_B_3M_Qualificado = Normal: Não {Não=7, Sim=0}
| | Conc_3M_Qualificado = Normal
| | | Idade_M > 38.500: Não {Não=2, Sim=0}
| | | Idade_M ≤ 38.500: Sim {Não=2, Sim=6}
| Grau_Varicoc = II
| | Idade_M > 24.500: Sim {Não=13, Sim=24}
| | Idade_M ≤ 24.500: Não {Não=2, Sim=0}
| Grau_Varicoc = III: Não {Não=11, Sim=2}
A_B_Pre_Qualificado = Normal
| A_B_3M_Qualificado = ?
| | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?: Não {Não=2, Sim=1}
| | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não: Não {Não=3, Sim=1}
| | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Sim {Não=0, Sim=4}
| A_B_3M_Qualificado = Anormal
| | Idade_M > 33: Não {Não=8, Sim=0}
| | Idade_M ≤ 33
| | | Conc_3M_Qualificado = Anormal: Sim {Não=0, Sim=6}
| | | Conc_3M_Qualificado = Normal: Não {Não=4, Sim=2}
| A_B_3M_Qualificado = Normal: Sim {Não=5, Sim=22}</td></tr>
</table>

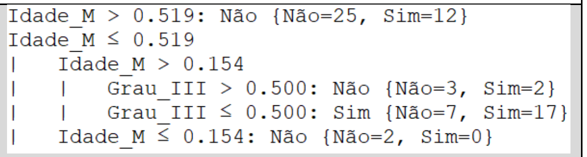
6.1	A_B_3M_Qualificado A_B_Pre_Qualificado Conc_3M_Qualificado Grau_Varicoc Gravidez Idade_M ProfissãoComRiscoDeContacto…	The dataset was filtered by non-missing values and binominal attributes were parsed into numerical, the "Grau_Varicoc" attribute was manually dichotomized and the "Idade_M" attribute normalized to end up with numerical values between 0 to 1 of 85 instances. (see Table 6.2)	RapidMiner´s Decision tree	Simple Validation	80.00%	85.71%	66.67%	75.00%	0.717	 Idade_M > 0.519: Não {Não=25, Sim=12} Idade_M ≤ 0.519 | Idade_M > 0.154 | | Grau_III > 0.500: Não {Não=3, Sim=2} | | Grau_III ≤ 0.500: Sim {Não=7, Sim=17} | Idade_M ≤ 0.154: Não {Não=2, Sim=0} Note: Since the above decision tree has its woman ´s age normalize, we can interpret 0.519=34 years old and 0.154=25 years old. Further more, since the yougest and oldest woman´s we have in this filtered data set goes from 20 to 46, the normalized value 0 is equal to 20 years old and the normalized value 1, is equal to 46 years old.  1-Validation.sampling_type = linear sampling 1-Decision Tree.criterion = accuracy 1-Decision Tree.apply_pruning = true 1-Decision Tree.minimal_size_for_split = 4 1-Decision Tree.minimal_gain = 0.1 1-Decision Tree.minimal_leaf_size = 2 1-Decision Tree.maximal_depth = 20
6.2				Cross Validation	64.71% +/- 11.00%	64.58% +/- 27.24%	47.50% +/- 19.92%	53.20% +/- 20.26%	0.589	 Idade_M > 0.519: Não {Não=25, Sim=12} Idade_M ≤ 0.519 | Idade_M > 0.154 | | Grau_III > 0.500: Não {Não=3, Sim=2} | | Grau_III ≤ 0.500: Sim {Não=7, Sim=17} | Idade_M ≤ 0.154: Não {Não=2, Sim=0}  2 - Cross Validation.number_of_folds = 4 2 - Cross Validation.sampling_type = linear sampling 2-Decision Tree.criterion = accuracy 2-Decision Tree.apply_pruning = true 2-Decision Tree.minimal_size_for_split = 4 2-Decision Tree.minimal_gain = 0.1 2-Decision Tree.minimal_leaf_size = 2 2-Decision Tree.maximal_depth = 20

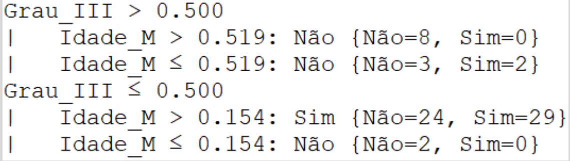
<table>
<tr><td>6.3</td><td rowspan="2"></td><td rowspan="2"></td><td rowspan="2">W-J48</td><td>Simple Validation</td><td>55.00%</td><td>50.00%</td><td>44.44%</td><td>47.06%</td><td>0.439</td><td>
W-J48
J48 unpruned tree

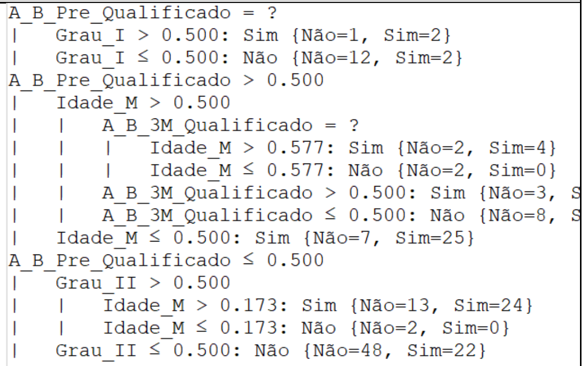
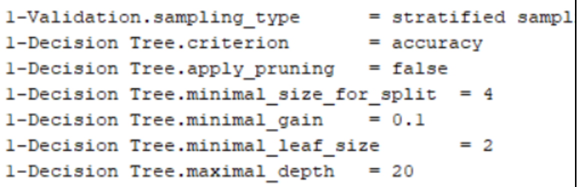
Grau_III <= 0
| Conc_3M_Qualificado <= 0
| | Grau_I <= 0
| | | A_B_Pre_Qualificado <= 0
| | | | A_B_3M_Qualificado <= 0
| | | | | Idade_M <= 0.5: Não (3.0)
| | | | | Idade_M > 0.5: Sim (3.0/1.0)
| | | | A_B_3M_Qualificado > 0: Sim (3.0/1.0)
| | | A_B_Pre_Qualificado > 0: Não (3.0)
| | Grau_I > 0
| | | A_B_3M_Qualificado <= 0
| | | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos <= 0: Sim (4.0)
| | | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos > 0
| | | | | Idade_M <= 0.423077: Não (3.0/1.0)
| | | | | Idade_M > 0.423077: Sim (2.0)
| | | A_B_3M_Qualificado > 0: Não (5.0/1.0)
| Conc_3M_Qualificado > 0
| | A_B_3M_Qualificado <= 0
| | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos <= 0: Sim (13.0/6.0)
| | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos > 0: Não (4.0/1.0)
| | A_B_3M_Qualificado > 0: Sim (12.0/3.0)
Grau_III > 0: Não (13.0/2.0)

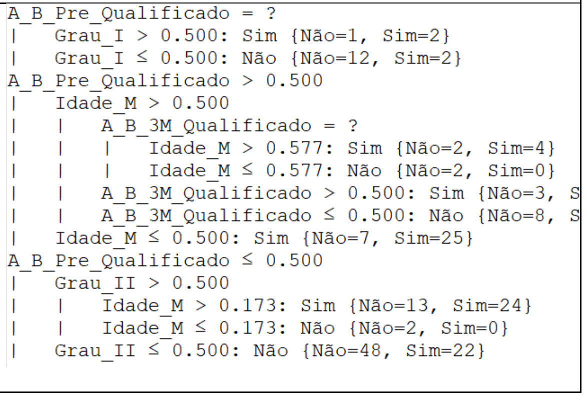
3-W-J48.U = true
3-Validation.sampling_type = linear samplir</td></tr>
<tr><td>6.4</td><td>Cross Validation</td><td>60.29% +/- 11.30%</td><td>56.25% +/- 13.34%</td><td>52.23% +/- 17.47%</td><td>53.72% +/- 14.91%</td><td>0.644 +/- 0.135</td><td>
W-J48
J48 unpruned tree

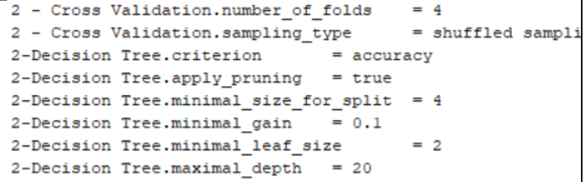
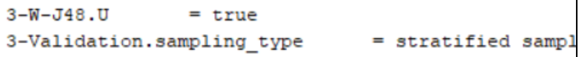
Grau_III <= 0
| Conc_3M_Qualificado <= 0
| | Grau_I <= 0
| | | A_B_Pre_Qualificado <= 0
| | | | A_B_3M_Qualificado <= 0
| | | | | Idade_M <= 0.5: Não (3.0)
| | | | | Idade_M > 0.5: Sim (3.0/1.0)
| | | | A_B_3M_Qualificado > 0: Sim (3.0/1.0)
| | | A_B_Pre_Qualificado > 0: Não (3.0)
| | Grau_I > 0
| | | A_B_3M_Qualificado <= 0
| | | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos <= 0: Sim (4.0)
| | | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos > 0
| | | | | Idade_M <= 0.423077: Não (3.0/1.0)
| | | | | Idade_M > 0.423077: Sim (2.0)
| | | A_B_3M_Qualificado > 0: Não (5.0/1.0)
| Conc_3M_Qualificado > 0
| | A_B_3M_Qualificado <= 0
| | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos <= 0: Sim (13.0/6.0)
| | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos > 0: Não (4.0/1.0)
| | A_B_3M_Qualificado > 0: Sim (12.0/3.0)
Grau_III > 0: Não (13.0/2.0)

4 - Cross Validation.number_of_folds = 4
4 - Cross Validation.sampling_type = stratified sam
4-W-J48.U = true</td></tr>
<tr><td colspan="3">Validation results of the right model found in step 6 with the best parameter´s settings found for the operator called “1-Decision tree” which is related to the highest f-measure found in step 6.1 that is above highlighted in orange color.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>70.59%</td><td>66.67%</td><td>75.00%</td><td>70.59%</td><td>0.750</td><td>
Idade_M > 0.519: Não {Não=25, Sim=12}
Idade_M ≤ 0.519
| Idade_M > 0.154
| | Grau_III > 0.500: Não {Não=3, Sim=2}
| | Grau_III ≤ 0.500: Sim {Não=7, Sim=17}
| Idade_M ≤ 0.154: Não {Não=2, Sim=0}</td></tr>
</table>

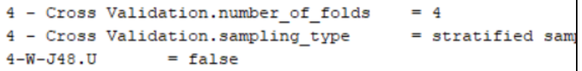
Validation of the model in step 6 with the lowest parameter´s settings (i.e. minimal size for split=4; minimal gain=0.018; minimal leaf size=2; maximal depth=20 and without prunning) and the splitting criterion gain ratio.			RapidMiner´s Decision tree	-	47.06%	46.67%	87.50%	60.87%	0.500	 Grau_III > 0.500
										| Idade_M > 0.519: Não {Não=8, Sim=0}
										| Idade_M ≤ 0.519: Não {Não=3, Sim=2}
										Grau_III ≤ 0.500
										| Idade_M > 0.154: Sim {Não=24, Sim=29}
										| Idade_M ≤ 0.154: Não {Não=2, Sim=0}

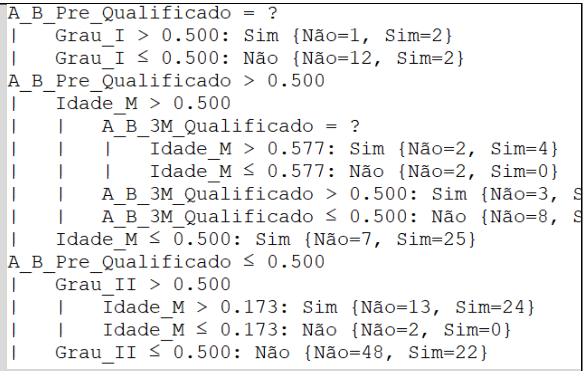
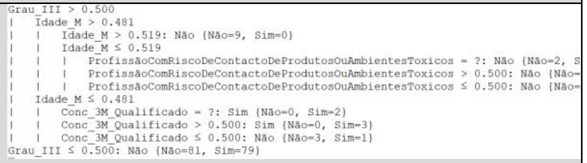
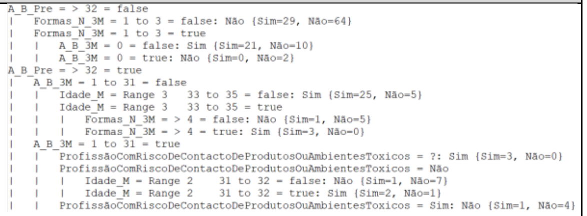
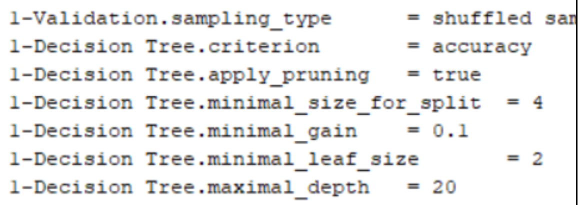
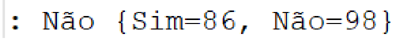
7.1	A_B_3M_Qualificado A_B_Pre_Qualificado Conc_3M_Qualificado Grau_Varicoc Gravidez Idade_M ProfissãoComRiscoDeContacto…	Binominal attributes were parsed into numerical attributes, the “Grau_Varicoc” attribute was manually dichotomized and the “Idade_M” attribute normalized to end up with a data set that only has numerical values between 0 to 1 of 230 instances. (see Table 6.2)	RapidMiner´s Decision tree	Simple Validation	70.91%	66.67%	76.92%	71.43%	0.716	 A_B_Pre_Qualificado = ? | Grau_I > 0.500: Sim {Não=1, Sim=2} | Grau_I ≤ 0.500: Não {Não=12, Sim=2} A_B_Pre_Qualificado > 0.500 | Idade_M > 0.500 | | A_B_3M_Qualificado = ? | | | Idade_M > 0.577: Sim {Não=2, Sim=4} | | | Idade_M ≤ 0.577: Não {Não=2, Sim=0} | | A_B_3M_Qualificado > 0.500: Sim {Não=3, S | | A_B_3M_Qualificado ≤ 0.500: Não {Não=8, S | Idade_M ≤ 0.500: Sim {Não=7, Sim=25} A_B_Pre_Qualificado ≤ 0.500 | Grau_II > 0.500 | | Idade_M > 0.173: Sim {Não=13, Sim=24} | | Idade_M ≤ 0.173: Não {Não=2, Sim=0} | Grau_II ≤ 0.500: Não {Não=48, Sim=22}  1-Validation.sampling_type = stratified sampl 1-Decision Tree.criterion = accuracy 1-Decision Tree.apply_pruning = false 1-Decision Tree.minimal_size_for_split = 4 1-Decision Tree.minimal_gain = 0.1 1-Decision Tree.minimal_leaf_size = 2 1-Decision Tree.maximal_depth = 20

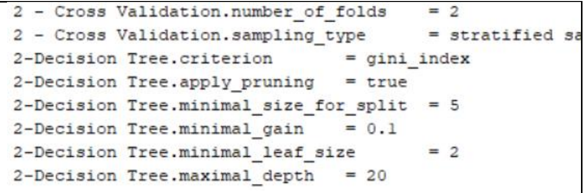
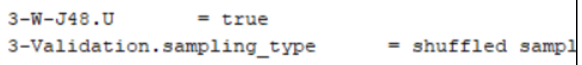
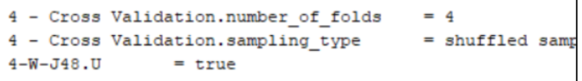
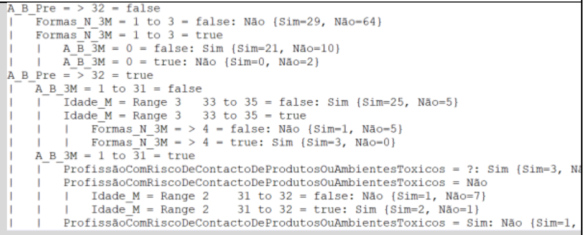
7.2				Cross Validation	60.33% +/- 9.65%	58.10% +/- 4.24%	54.81% +/- 25.62%	53.14% +/- 17.49%	0.629 +/- 0.099	 A_B_Pre_Qualificado = ?
										| Grau_I > 0.500: Sim {Não=1, Sim=2}
										| Grau_I ≤ 0.500: Não {Não=12, Sim=2}
										A_B_Pre_Qualificado > 0.500
										| Idade_M > 0.500
										| | A_B_3M_Qualificado = ?
										| | | Idade_M > 0.577: Sim {Não=2, Sim=4}
										| | | Idade_M ≤ 0.577: Não {Não=2, Sim=0}
										| | A_B_3M_Qualificado > 0.500: Sim {Não=3, S
										| | A_B_3M_Qualificado ≤ 0.500: Não {Não=8, S
										| Idade_M ≤ 0.500: Sim {Não=7, Sim=25}
										A_B_Pre_Qualificado ≤ 0.500
										| Grau_II > 0.500
										| | Idade_M > 0.173: Sim {Não=13, Sim=24}
										| | Idade_M ≤ 0.173: Não {Não=2, Sim=0}
										| Grau_II ≤ 0.500: Não {Não=48, Sim=22}

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>
2 - Cross Validation.number_of_folds   = 4
2 - Cross Validation.sampling_type   = shuffled sampli
2-Decision Tree.criterion   = accuracy
2-Decision Tree.apply_pruning   = true
2-Decision Tree.minimal_size_for_split = 4
2-Decision Tree.minimal_gain   = 0.1
2-Decision Tree.minimal_leaf_size   = 2
2-Decision Tree.maximal_depth   = 20</td></tr>
<tr><td>7.3</td><td></td><td></td><td>W-J48</td><td>Simple Validation</td><td>67.27%</td><td>65.38%</td><td>65.38%</td><td>65.38%</td><td>0.682</td><td>(too lon unprunned tree)

3-W-J48.U   = true
3-Validation.sampling_type   = stratified sampl</td></tr>
<tr><td>7.4</td><td></td><td></td><td></td><td>Cross Validation</td><td>64.13% +/- 3.61%</td><td>60.70% +/- 2.27%</td><td>64.94% +/- 15.03%</td><td>62.09% +/- 7.02%</td><td>0.666 +/- 0.048</td><td>W-J48
J48 pruned tree

Conc_3M_Qualificado <= 0
| A_B_Pre_Qualificado <= 0
| | Grau_II <= 0: Não (52.35/10.94)
| | Grau_II > 0: Sim (20.64/9.28)
| A_B_Pre_Qualificado > 0
| | Idade_M <= 0.5: Sim (15.83/3.32)
| | Idade_M > 0.5: Não (11.23/3.72)
Conc_3M_Qualificado > 0
| A_B_3M_Qualificado <= 0
| | Idade_M <= 0.269231: Não (5.4/0.2)
| | Idade_M > 0.269231
| | | A_B_Pre_Qualificado <= 0: Sim (21.34
| | | A_B_Pre_Qualificado > 0: Não (10.06/
| A_B_3M_Qualificado > 0: Sim (47.15/16.08)


4 - Cross Validation.number_of_folds   = 4
4 - Cross Validation.sampling_type   = stratified sam
4-W-J48.U   = false</td></tr>
</table>

<table>
<tr><td colspan="3">Validation results of the right model found in step 7 with the best parameter´s settings found for the operator called “1-Decision tree” which is related to the highest f-measure found in step 7.1 that is above highlighted in orange color.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>45.65%</td><td>42.31%</td><td>52.38%</td><td>46.81%</td><td>0.465</td><td>
A_B_Pre_Qualificado = ?
| Grau_I > 0.500: Sim {Não=1, Sim=2}
| Grau_I ≤ 0.500: Não {Não=12, Sim=2}
A_B_Pre_Qualificado > 0.500
| Idade_M > 0.500
| | A_B_3M_Qualificado = ?
| | | Idade_M > 0.577: Sim {Não=2, Sim=4}
| | | Idade_M ≤ 0.577: Não {Não=2, Sim=0}
| | A_B_3M_Qualificado > 0.500: Sim {Não=3, S
| | A_B_3M_Qualificado ≤ 0.500: Não {Não=8, S
| Idade_M ≤ 0.500: Sim {Não=7, Sim=25}
A_B_Pre_Qualificado ≤ 0.500
| Grau_II > 0.500
| | Idade_M > 0.173: Sim {Não=13, Sim=24}
| | Idade_M ≤ 0.173: Não {Não=2, Sim=0}
| Grau_II ≤ 0.500: Não {Não=48, Sim=22}</td></tr>
<tr><td colspan="3">Validation of the model in step 7 with the lowest parameter´s settings (i.e. minimal size for split=4; minimal gain=0.018; minimal leaf size=2; maximal depth=20 and without prunning) and the splitting criterion gain ratio.</td><td>RapidMiner´s Decision tree</td><td>-</td><td>54.35%</td><td>unknown</td><td>0.00%</td><td>unknown</td><td>0.581</td><td>
Grau_III > 0.500
| Idade_M > 0.481
| | Idade_M > 0.519: Não {Não=9, Sim=0}
| | Idade_M ≤ 0.519
| | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?: Não {Não=2, S
| | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos > 0.500: Não {Não=
| | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos ≤ 0.500: Não {Não=
| Idade_M ≤ 0.481
| | Conc_3M_Qualificado = ?: Sim {Não=0, Sim=2}
| | Conc_3M_Qualificado > 0.500: Sim {Não=0, Sim=3}
| | Conc_3M_Qualificado ≤ 0.500: Não {Não=3, Sim=1}
Grau_III ≤ 0.500: Não {Não=81, Sim=79}</td></tr>
<tr><td>8.1</td><td rowspan="2">A_B_3M
A_B_Pre
Conc_3M
Conc_6M
Formas_N_3M
Grau_Varicoc
Gravidez
HabitosAlcoolicos_Processado…
HabitosTabagicos_Processado…
Idade_M
ProfissãoComRiscoDeContacto…</td><td rowspan="2">Transformed the numérical attributes into nominal with a user defined discretization that refrects the WHO thresholds for the sperm parameters attributes, the woman´s age quartiles for the Idade_M attribute and the severity</td><td rowspan="2">RapidMiner´s Decision tree</td><td>Simple Validation</td><td>67.27%</td><td>66.67%</td><td>75.86%</td><td>70.97%</td><td>0.710</td><td>
A_B_Pre = > 32 = false
| Formas_N_3M = 1 to 3 = false: Não {Sim=29, Não=64}
| Formas_N_3M = 1 to 3 = true
| | A_B_3M = 0 = false: Sim {Sim=21, Não=10}
| | A_B_3M = 0 = true: Não {Sim=0, Não=2}
A_B_Pre = > 32 = true
| A_B_3M = 1 to 31 = false
| | Idade_M = Range 3   33 to 35 = false: Sim {Sim=25, Não=5}
| | Idade_M = Range 3   33 to 35 = true
| | | Formas_N_3M = > 4 = false: Não {Sim=1, Não=5}
| | | Formas_N_3M = > 4 = true: Sim {Sim=3, Não=0}
| A_B_3M = 1 to 31 = true
| | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?: Sim {Sim=3, Não=0}
| | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não
| | | Idade_M = Range 2   31 to 32 = false: Não {Sim=1, Não=7}
| | | Idade_M = Range 2   31 to 32 = true: Sim {Sim=2, Não=1}
| | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Não {Sim=1, Não=4}

1-Validation.sampling_type = shuffled sam
1-Decision Tree.criterion = accuracy
1-Decision Tree.apply_pruning = true
1-Decision Tree.minimal_size_for_split = 4
1-Decision Tree.minimal_gain = 0.1
1-Decision Tree.minimal_leaf_size = 2
1-Decision Tree.maximal_depth = 20</td></tr>
<tr><td>8.2</td><td>Cross Validation</td><td>61.41% +/- 1.63%</td><td>60.01% +/- 1.90%</td><td>83.67% +/- 4.08%</td><td>69.78% +/- 0.14%</td><td>0.623 +/- 0.012</td><td>
: Não {Sim=86, Não=98}</td></tr>
</table>

		grade was automatically dichotomized; Hence, all the data set was transformed into nominal attribute values.								 2 - Cross Validation.number_of_folds = 2 2 - Cross Validation.sampling_type = stratified sa 2-Decision Tree.criterion = gini_index 2-Decision Tree.apply_pruning = true 2-Decision Tree.minimal_size_for_split = 5 2-Decision Tree.minimal_gain = 0.1 2-Decision Tree.minimal_leaf_size = 2 2-Decision Tree.maximal_depth = 20
8.3			W-J48	Simple Validation	63.64%	66.67%	70.97%	68.75%	0.603	(too long unprunned tree)  3-W-J48.U = true 3-Validation.sampling_type = shuffled sampl
8.4				Cross Validation	63.04% +/- 5.95%	65.74% +/- 5.43%	65.92% +/- 7.51%	65.43% +/- 4.31%	0.627 +/- 0.061	(too long unprunned tree)  4 - Cross Validation.number_of_folds = 4 4 - Cross Validation.sampling_type = shuffled samp 4-W-J48.U = true
Validation results of the right model found in step 8 with the best parameter´s settings found for the operator called "1-Decision tree" which is related to the highest f-measure found in step 8.1 that is above highlighted in orange color.			RapidMiner´s Decision tree	-	50.00%	53.85%	56.00%	54.90%	0.502	 A_B_Pre = > 32 = false | Formas_N_3M = 1 to 3 = false: Não {Sim=29, Não=64} | Formas_N_3M = 1 to 3 = true | | A_B_3M = 0 = false: Sim {Sim=21, Não=10} | | A_B_3M = 0 = true: Não {Sim=0, Não=2} A_B_Pre = > 32 = true | A_B_3M = 1 to 31 = false | | Idade_M = Range 3 33 to 35 = false: Sim {Sim=25, Não=5} | | Idade_M = Range 3 33 to 35 = true | | | Formas_N_3M = > 4 = false: Não {Sim=1, Não=5} | | | Formas_N_3M = > 4 = true: Sim {Sim=3, Não=0} | A_B_3M = 1 to 31 = true | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = ?: Sim {Sim=3, Nã | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Não | | | Idade_M = Range 2 31 to 32 = false: Não {Sim=1, Não=7} | | | Idade_M = Range 2 31 to 32 = true: Sim {Sim=2, Não=1} | | ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos = Sim: Não {Sim=1,
Validation of the model in step 8 with the lowest parameter´s settings (i.e. minimal size for split=4; minimal gain=0.018; minimal leaf size=2; maximal depth=20 and without prunning) and the splitting criterion gain ratio.			RapidMiner´s Decision tree	-	45.65%	50.00%	52.00%	50.98%	0.373	(too long tree)

C.2 Clustering

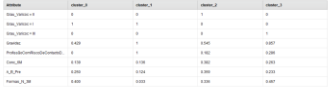
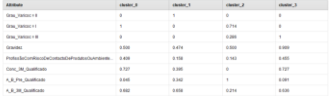
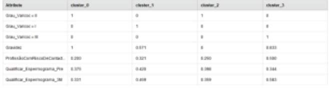
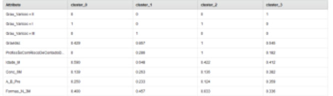
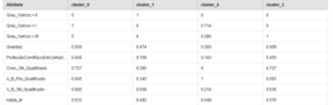
In the below Table C.2 1, we present the first results that were computed with the clustering K-means data mining model depicted in Figure 6.14. These tests were executed in the RapidMiner platform where we have tested several K-mean´s input and parameter values that were in Table 4.11 disclosed.

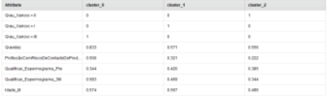
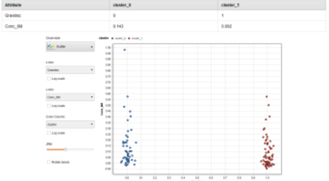
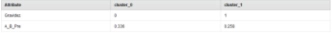
This table discloses the settings of each performed test identified with the id specified under the column named "Test ID" and presents, in the last 3 columns, its corresponding results as follows: under the column named "Davies Bouldin*"*, we present the *Davies Bouldin* index computed by the RapidMiner platform; under the column named "*n*", we showcase the distribution of the patients by the generated clusters for the best models (we present the clustering distribution by indicating that, for example, cluster 1, indicated by C_1, has x patients etc.) and under the column named "Centroid Table", we disclose the generated normalized centroid tables for the best models and when possible, its related scatter plot.

In regards of the *Davies Bouldin* index, the RapidMiner platform multiplies the *Davies Bouldin* index by -1 for maximization purposes and concerning its interpretation, the RapidMiner suggests that we have to consider the absolute value of the generated index and look for the index value that is closer to 0 for the selection of the best clustering models.

In this context, the best computed models suggested by the RapidMiner platform are in Table C.2 1 highlighted in orange and the most interesting ones, in red.

Table C.2 1 K-mean´s results of the model depicted in Figure 6.14

Test ID	Attribute selection	Number of clusters	Numerical measure	Davies Bouldin index	*n*	Centroid Table
1	A_B_Pre, Conc_6M, Formas_N_3M, Grau_Varicoc, ProfissãoComRiscoDeContactoDeProdutosOuAmbientes, Gravidez.	2	ManhattanDistance	-1.438	C$_0$: 7 items	
2		3	ManhattanDistance	-1.054	C$_1$: 3 items	
3		4	ManhattanDistance	-1.097	C$_2$: 11 items	
4		2	EuclideanDistance	-1.438	C$_3$: 7 items	
5		3	EuclideanDistance	-1.054	Total: 28	
6		4	EuclideanDistance	-0.901		
7	A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_Varicoc, ProfissãoComRiscoDeContactoDeProdutosOuAmbientes, Gravidez.	2	ManhattanDistance	-1.786	C$_0$: 22 items	
8		3	ManhattanDistance	-1.473	C$_1$: 38 items	
9		4	ManhattanDistance	-1.420	C$_2$: 14 items	
10		2	EuclideanDistance	-1.786	C$_3$: 11 items	
11		3	EuclideanDistance	-1.473	Total: 85	
12		4	EuclideanDistance	-1.339		
13	Grau_Varicoc, Qualificar_Espermograma_Pre, Qualificar_Espermograma_3M, ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos, Gravidez.	2	ManhattanDistance	-1.399	C$_0$: 20 items	
14		3	ManhattanDistance	-1.099	C$_1$: 28 items	
15		4	ManhattanDistance	-1.097	C$_2$: 16 items	
16		2	EuclideanDistance	-1.399	C$_3$: 12 items	
17		3	EuclideanDistance	-1.099	Total: 76	
18		4	EuclideanDistance	-1.097		
19	Idade_M, A_B_Pre, Conc_6M, Formas_N_3M, Grau_Varicoc, ProfissãoComRiscoDeContactoDeProdutosOuAmbientes, Gravidez.	2	ManhattanDistance	-1.466	C$_0$: 7 items	
20		3	ManhattanDistance	-1.089	C$_1$: 7 items	
21		4	ManhattanDistance	-0.964	C$_2$: 3 items	
22		2	EuclideanDistance	-1.466	C$_3$: 11 items	
23		3	EuclideanDistance	-1.089	Total: 28	
24		4	EuclideanDistance	-0.964		
25	Idade_M, A_B_Pre_Qualificado, A_B_3M_Qualificado, Conc_3M_Qualificado, Grau_Varicoc, ProfissãoComRiscoDeContactoDeProdutosOuAmbientes, Gravidez.	2	ManhattanDistance	-1.802	C$_0$: 22 items	
26		3	ManhattanDistance	-1.488	C$_1$: 38 items	
27		4	ManhattanDistance	-1.435	C$_2$: 14 items	
28		2	EuclideanDistance	-1.802	C$_3$: 11 items	
29		3	EuclideanDistance	-1.488	Total: 85	
30		4	EuclideanDistance	-1.357		
31	Idade_M, Grau_Varicoc, Qualificar_Espermograma_Pre,	2	ManhattanDistance	-1.421	C$_0$: 12 items	
32		3	ManhattanDistance	-1.120	C$_1$: 28 items	

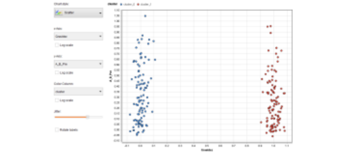
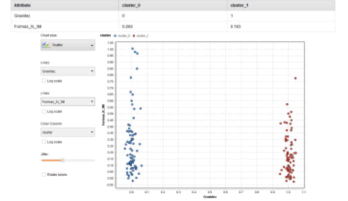
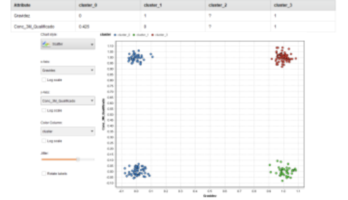
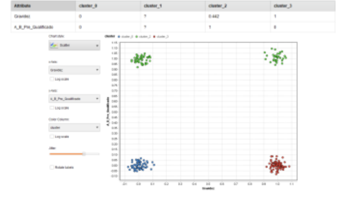
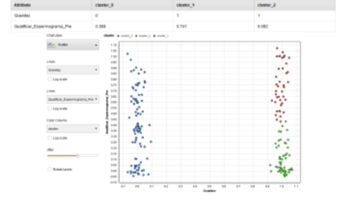
	Qualificar_Espermograma_3M, ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos, Gravidez.				C_2: 36 items Total: 76	
33		4	ManhattanDistance	-1.134		
34		2	EuclideanDistance	-1.421		
35		3	EuclideanDistance	-1.120		
36		4	EuclideanDistance	-1.134		
37	Gravidez vs Conc_6M	2	ManhattanDistance	-0.222	C_0: 50 items C_1: 65 items Total: 115	
38		3	ManhattanDistance	-0.389		
39		4	ManhattanDistance	-0.526		
40		2	EuclideanDistance	-0.222		
41		3	EuclideanDistance	-0.412		
42		4	EuclideanDistance	-0.526		
37	Gravidez vs A_B_Pre	2.	ManhattanDistance	-0.409	C_0: 102 items C_1: 107 items Total: 209	
38		3	ManhattanDistance	-0.432		
39		4	ManhattanDistance	-0.533		
40		2	EuclideanDistance	-0.409		

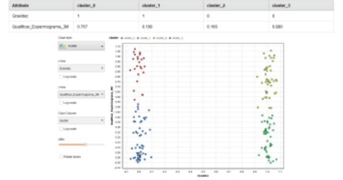
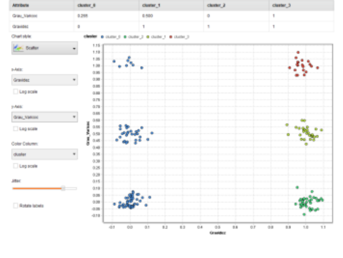
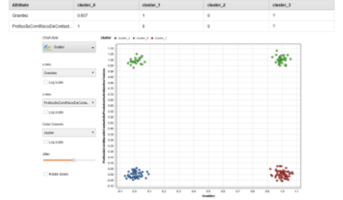
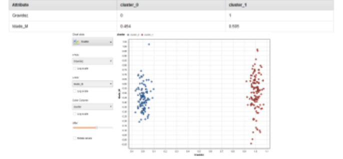
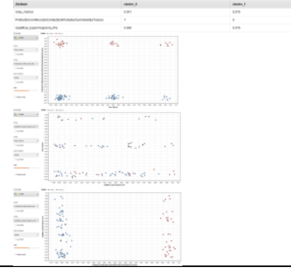
Attribute	cluster_0	cluster_1	cluster_2
Grau_Varicoc = II	0	0	1
Grau_Varicoc = I	0	1	0
Grau_Varicoc = III	1	0	0
Gravidez	0.833	0.571	0.556
ProfissãoComRiscoDeContactoDeProd...	0.500	0.321	0.222
Qualificar_Espermograma_Pre	0.344	0.420	0.385
Qualificar_Espermograma_3M	0.583	0.469	0.344
Idade_M	0.574	0.567	0.495

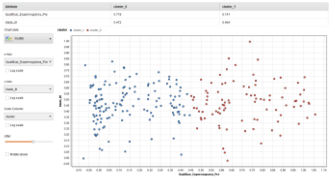
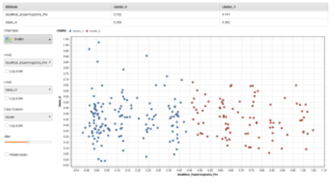
Attribute	cluster_0	cluster_1
Gravidez	0	1
Conc_6M	0.143	0.892

Chart style
Scatter
x-axis
Gravidez
☐ Log scale
y-axis
Conc_6M
☐ Log scale
Color Column
cluster
☐ Log scale
Jitter
☐ Rotate labels

Attribute	cluster_0	cluster_1
Gravidez	0	1
A_B_Pre	0.336	0.258

41		3	EuclideanDistance	-0.432		
42		4	EuclideanDistance	-0.533		
43	Gravidez vs Formas_N_3M	2	ManhattanDistance	-0.307	C_0: 78 items C_1: 74 items Total: 152	
44		3	ManhattanDistance	-0.404		
45		4	ManhattanDistance	-0.489		
46		2	EuclideanDistance	-0.307		
47		3	EuclideanDistance	-0.404		
48		4	EuclideanDistance	-0.511		
49	Gravidez vs Conc_3M_Qualificado	2	ManhattanDistance	-0.938	C_0: 94 items C_1: 40 items C_2: 0 items C_3: 67 items Total: 201	
50		3	ManhattanDistance	-0.441		
51		4	ManhattanDistance	Infinity		
52		2	EuclideanDistance	-0.938		
53		3	EuclideanDistance	-0.441		
54		4	EuclideanDistance	-0.0		
55	Gravidez vs A_B_Pre_Qualificado	2	ManhattanDistance	-0.916	C_0: 59 items C_1: 0 items C_2: 77 items C_3: 73 items Total: 209	
56		3	ManhattanDistance	-0.444		
57		4	ManhattanDistance	Infinity		
58		2	EuclideanDistance	-0.916		
59		3	EuclideanDistance	-0.444		
60		4	EuclideanDistance	Infinity		
61	Gravidez vs A_B_3M_Qualificado	2	ManhattanDistance	-0.973	C_0: 54 items C_1: 92 items C_2: 38 items C_3: 0 items Total: 184	
62		3	ManhattanDistance	-0.443		
63		4	ManhattanDistance	Infinity		
64		2	EuclideanDistance	-0.973		
65		3	EuclideanDistance	-0.443		
66		4	EuclideanDistance	Infinity		
67	Gravidez vs Qualificar_Espermograma_Pre	2	ManhattanDistance	-0.564	C_0: 93 items C_1: 41 items C_2: 57 items Total: 191	
68		3	ManhattanDistance	-0.389		
69		4	ManhattanDistance	-0.472		
70		2	EuclideanDistance	-0.564		
71		3	EuclideanDistance	-0.389		
72		4	EuclideanDistance	-0.472		
73	Gravidez vs Qualificar_Espermograma_3M	2	ManhattanDistance	-0.607	C_0: 43 items	

74		3	ManhattanDistance	-0.431	C_1: 46 items C_2: 55 items C_3: 25 items Total: 169	
75		4	ManhattanDistance	-0.422		
76		2	EuclideanDistance	-0.607		
77		3	EuclideanDistance	-0.431		
78		4	EuclideanDistance	-0.390		
79	Gravidez vs Grau_Varicoc	2	ManhattanDistance	-0.634	C_0: 85 items C_1: 31 items C_2: 38 items C_3: 20 items Total: 174 Without binomial operator	
80		3	ManhattanDistance	-0.443		
81		4	ManhattanDistance	-0.278		
82		2	EuclideanDistance	-0.634		
83		3	EuclideanDistance	-0.443		
84		4	EuclideanDistance	-0.358		
85	Gravidez vs ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos.	2	ManhattanDistance	-0.885	C_0: 56 items C_1: 65 items C_2: 47 items C_3: 0 items Total: 168	
86		3	ManhattanDistance	-0.410		
87		4	ManhattanDistance	Infinity		
88		2	EuclideanDistance	-0.885		
89		3	EuclideanDistance	-0.432		
90		4	EuclideanDistance	Infinity		
91	Gravidez vs Idade_M	2	ManhattanDistance	-0.244	C_0: 107 items C_1: 122 items Total: 229	
92		3	ManhattanDistance	-0.484		
93		4	ManhattanDistance	-0.579		
94		2	EuclideanDistance	-0.244		
95		3	EuclideanDistance	-0.484		
96		4	EuclideanDistance	-0.581		
97	Grau_Varicoc vs Qualificar_Espermograma_Pre vs ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos.	2	ManhattanDistance	-1.752	C_0: 36 items C_1: 70 items Total: 106 Without binomial operator for severity grade	
98		3	ManhattanDistance	-0.958		
99		4	ManhattanDistance	-1.054		
100		2	EuclideanDistance	-0.919		
101		3	EuclideanDistance	-1.049		
102		4	EuclideanDistance	-1.038		

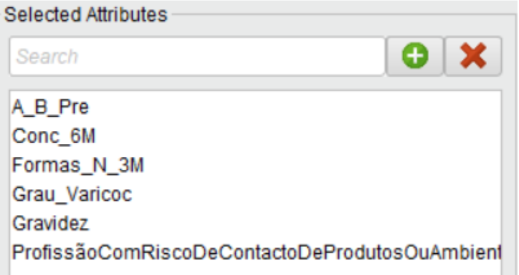
103	Idade_M vs Qualificar_Espermograma_Pre	2	ManhattanDistance	-0.721	C0: 92 items C1: 142 items Total: 234	
104		3	ManhattanDistance	-0.876		
105		4	ManhattanDistance	-1.004		
106		2	EuclideanDistance	-0.721		
107		3	EuclideanDistance	-0.876		
108		4	EuclideanDistance	-0.972		
109	Idade_H vs Qualificar_Espermograma_Pre	2	ManhattanDistance	-0.711	C0: 94 items C1: 144 items Total: 238	
110		3	ManhattanDistance	-0.915		
111		4	ManhattanDistance	-1.017		
112		2	EuclideanDistance	-0.711		
113		3	EuclideanDistance	-0.915		
114		4	EuclideanDistance	-0.941		

C.3 Association

In this section, we present the obtained results with the *FP_Growth* algorithm. These results came by implementing the six modeling steps described in Table 6.6 for the groups of attributes specified in section 5.4.4. Hence, we have tested the built models presented in section 0 by altering the selected attributes set in the operator named “Select to model” and by adjusting the *support* and *confidence* measure as planed in the defined modeling steps. In order to better convey the obtained results, we below present, through several sections, the results obtained in each carried out modeling step. The most interesting rules are in each result´s table identified with a check mark and an exclamation mark - the check mark identifies the rules that are objectively interesting and the exclamation mark, the ones that are subjectively interesting.

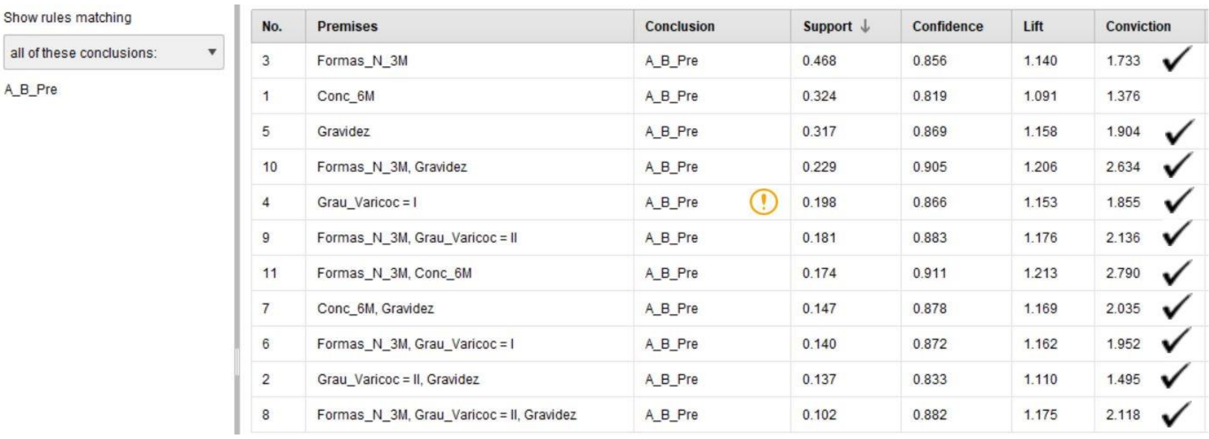
C.3.1 Results of Step 1

Firstly, we have tested the model presented in Figure 6.27 FP-Growth model, with the attributes specified in Figure C.3 1 (first modeling step of the association rule application). The first obtained results with a support=0.1 and confidence=0.8 are presented in Figure C.3 2.



A_B_Pre
Conc_6M
Formas_N_3M
Grau_Varicoc
Gravidez
ProfissãoComRiscoDeContactoDeProdutosOuAmbient

Figure C.3 1 Selected attributes for the steps 1,2 and 3 of the FP_Growth algorithm

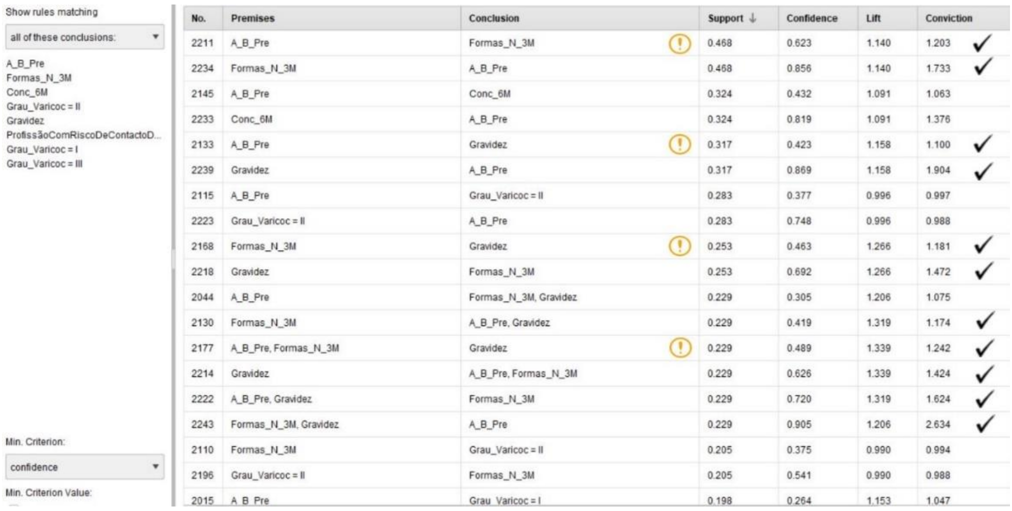


No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction
3	Formas_N_3M	A_B_Pre	0.468	0.856	1.140	1.733 ✓
1	Conc_6M	A_B_Pre	0.324	0.819	1.091	1.376
5	Gravidez	A_B_Pre	0.317	0.869	1.158	1.904 ✓
10	Formas_N_3M, Gravidez	A_B_Pre	0.229	0.905	1.206	2.634 ✓
4	Grau_Varicoc = I	A_B_Pre !	0.198	0.866	1.153	1.855 ✓
9	Formas_N_3M, Grau_Varicoc = II	A_B_Pre	0.181	0.883	1.176	2.136 ✓
11	Formas_N_3M, Conc_6M	A_B_Pre	0.174	0.911	1.213	2.790 ✓
7	Conc_6M, Gravidez	A_B_Pre	0.147	0.878	1.169	2.035 ✓
6	Formas_N_3M, Grau_Varicoc = I	A_B_Pre	0.140	0.872	1.162	1.952 ✓
2	Grau_Varicoc = II, Gravidez	A_B_Pre	0.137	0.833	1.110	1.495 ✓
8	Formas_N_3M, Grau_Varicoc = II, Gravidez	A_B_Pre	0.102	0.882	1.175	2.118 ✓

Figure C.3 2 test 1.1: group 1 - unfiltered - support=0.1 and confidence=0.8

If we interpret the result of the first association rule presented in the above Figure C.3 2 that can be translated into *Formas_N_3M -> A_B_Pre*, we can say that the existence of values greater than zero in the attribute *Formas_N_3M* implies the existence of values greater than zero in the attribute *A_B_Pre* 47% of the times (i.e. the *support* of X union Y written (X U Y), where X=*Formas_N_3M* and Y=*A_B_Pre*, is equal to an *absolute support* of 137 and a *relative support* of 47% since the identified 137 instances divided by our total of 293 instances is equal to 0.468 which means that 47% of the patients of our preprocessed data set have values higher than zero on their sperm morphology at 3 months, as well as on their sperm motility before the treatment). The computed *confidence* value for this association rule enables us to say that the probability of observing a patient with a sperm motility before the treatment (i.e. having a *A_B_Pre* > 0) given a sperm morphology 3 months later (i.e. having a *Formas_N_3M* > 0) is of 86% (i.e. *confidence*(*Formas_N_3M*->*A_B_Pre*) = *confidence*(X->Y) = P(Y|X) = support(XUY) / support(X) = 137/160 = 0.856). Moreover, we see that the "direction" of the rule is different than the other way around since the *conviction* measure is different than the value 1.0, which reinforces the interestingness of the rule. If we assess each sperm parameter independently, we see that 75% (i.e. 220/293=0.750) of the assessed patients had an *A_B_Pre* above 0 and that 55% (i.e. 160/293=0.546) of the patients had a *Formas_N_3M* above 0, which means that the prior probability of having a patient with a sperm motility before the treatment is more likely than having a sperm morphology 3 months later (i.e. *A_B_Pre* = 75% vs *Formas_N_3M* = 55%). However, since the conditional probability of *A_B_Pre* given *Formas_N_3M* is higher than the support of *A_B_Pre*, we have a *lift* value greater than 1.0 which tells us that the *Formas_N_3M* attribute is related with the *A_B_Pre* (i.e. Lift(*Formas_N_3M*->*A_B_Pre*) = Lift(X->Y) = confidence (X->Y)/support(Y) = 0.856/0.75 = 1.14).

Even if all these measures indicate that we have here an interesting rule, in terms of clinical interest, this rule only identifies a pattern on the preprocessed data. In fact, if the rule was in the opposite direction (i.e. *A_B_Pre* ->*Formas_N_3M*) it would be much more interesting because it would provide a predictive information where we could conclude that before the embolization treatment, a patient with a sperm motility above zero implies having 3 months later a sperm morphology also greater than 0. To try to find this rule, we have lowered the *support* and the *confidence* value to 0.00 and ran again the process to identify the metrics related to this rule. Figure C.3 3 presents the obtained results where we can see that the rule *A_B_Pre* ->*Formas_N_3M* appears in the first row. If we analyze its results, we see that it has a lower *confidence* and *conviction* metric than the former assessed *Formas_N_3M->A_B_Pre rule*. However, these values are still acceptable since its metrics are: *support*=0.468, *confidence*=0.623, *lift*=1.140 and *conviction*=1.203.



Show rules matching

all of these conclusions: ▾

A_B_Pre
Formas_N_3M
Conc_6M
Grau_Varicoc = II
Gravidez
ProfissãoComRiscoDeContactoD...
Grau_Varicoc = I
Grau_Varicoc = III

Min. Criterion:

confidence ▾

Min. Criterion Value:

No.	Premises	Conclusion		Support ↓	Confidence	Lift	Conviction	
2211	A_B_Pre	Formas_N_3M	!	0.468	0.623	1.140	1.203	✓
2234	Formas_N_3M	A_B_Pre		0.468	0.856	1.140	1.733	✓
2145	A_B_Pre	Conc_6M		0.324	0.432	1.091	1.063	
2233	Conc_6M	A_B_Pre		0.324	0.819	1.091	1.376	
2133	A_B_Pre	Gravidez	!	0.317	0.423	1.158	1.100	✓
2239	Gravidez	A_B_Pre		0.317	0.869	1.158	1.904	✓
2115	A_B_Pre	Grau_Varicoc = II		0.283	0.377	0.996	0.997	
2223	Grau_Varicoc = II	A_B_Pre		0.283	0.748	0.996	0.988	
2168	Formas_N_3M	Gravidez	!	0.253	0.463	1.266	1.181	✓
2218	Gravidez	Formas_N_3M		0.253	0.692	1.266	1.472	✓
2044	A_B_Pre	Formas_N_3M, Gravidez		0.229	0.305	1.206	1.075	
2130	Formas_N_3M	A_B_Pre, Gravidez		0.229	0.419	1.319	1.174	✓
2177	A_B_Pre, Formas_N_3M	Gravidez	!	0.229	0.489	1.339	1.242	✓
2214	Gravidez	A_B_Pre, Formas_N_3M		0.229	0.626	1.339	1.424	✓
2222	A_B_Pre, Gravidez	Formas_N_3M		0.229	0.720	1.319	1.624	✓
2243	Formas_N_3M, Gravidez	A_B_Pre		0.229	0.905	1.206	2.634	✓
2110	Formas_N_3M	Grau_Varicoc = II		0.205	0.375	0.990	0.994	
2196	Grau_Varicoc = II	Formas_N_3M		0.205	0.541	0.990	0.988	
2015	A_B_Pre	Grau_Varicoc = I		0.198	0.264	1.153	1.047	

Figure C.3 3 test 1.2: group 1 - unfiltered - support=0.0 and confidence=0.0

If we interpret the remaining results presented in Figure C.3 2, we can briefly say that:

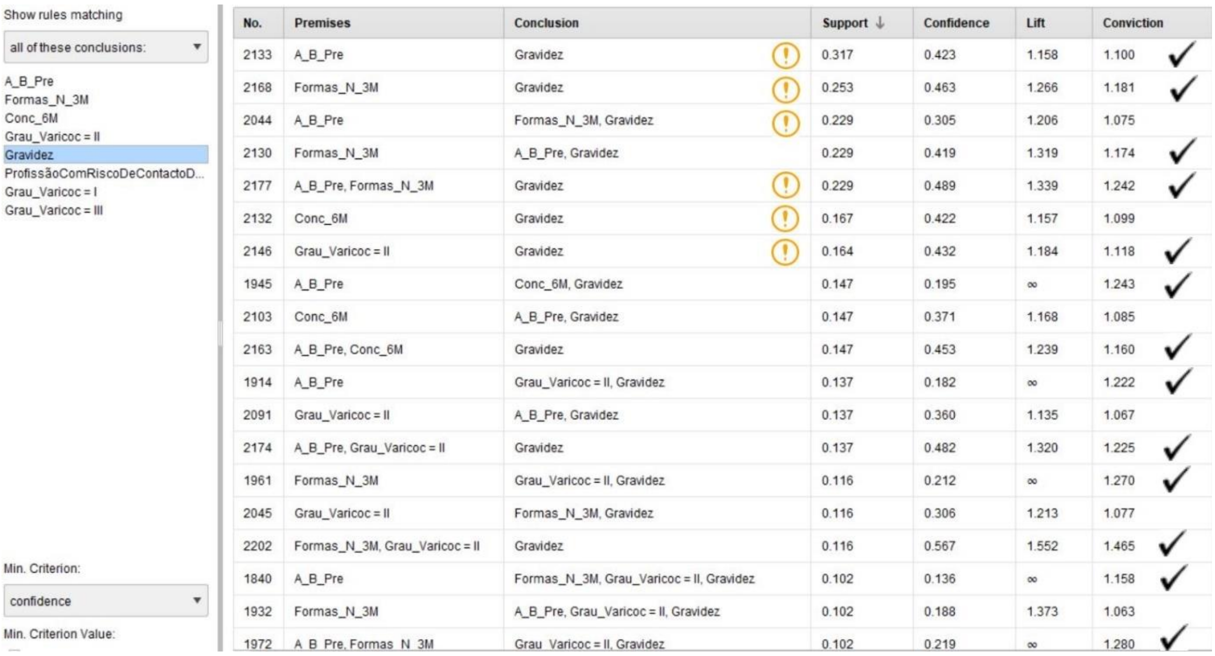
- All generated association rules with a *support* equal or above 0.1 and a *confidence* equal or above 0.8 for the group of attributes: A_B_Pre, Conc_6M, Formas_N_3M, Grau_Varicoc, Gravidez, ProfissãoComRiscoDeContactoDeProdutosOuAmbientes; have the same *consequent* attribute (i.e. the A_B_Pre attribute appears as a conclusion on all generated rules). Therefore, we can say that the most interesting association rules are for this group of attributes the ones with the A_B_Pre attribute as a consequent.
- The second association rule with the highest *support* is the Conc_6M -> A_B_Pre, *support*=0.324, *confidence*=0.819, *lift*=1.091 and *conviction*=1.376. However, this association rule is not interesting: even if the *support* and the *confidence* is acceptable, its *lift* value is too close to 1.0; and hence, it is not an interesting rule because the conditional probability of the event A_B_Pre given Conc_6M is quite equal to the probability of the event A_B_Pre. Furthermore, subjectively, this rule is not interesting.
- The association rule with the highest *confidence*, *lift* and *conviction* is the rule: Formas_N_3M, Conc_6M -> A_B_Pre, *support*=0.174, *confidence*=0.911, *lift*=1.213 and *conviction*=2.790. Its measures tells us that 17% of the patients have the sperm morphology at 3 months and the concentration at 6 months above 0 and that the conditional probability of having an A_B_Pre before treatment above 0 given Formas_N_3M and Conc_6M also above 0 occurs in 91% of the times. Furthermore, this conditional probability does not occur randomly since its lift is above 1.0 and the direction of the rule is objectively interesting due to its high *conviction* measure. Unfortunately, subjectively, it is not an interesting rule due to its direction.
- The only objectively and subjectively interesting rule in the first run is: Grau_Varicoc=I->A_B_Pre, *support*=0.198, *confidence*=0.866, *lift*=1.153 and *conviction*=1.855. This rule tells us that 58 patients of the data set (293 instances * 0.198) have a severity grade equal to I and sperm motility above 0. The conditional probability of having a sperm motility above 0 given a severity grade equal to I is of 87% (i.e. confidence=0.866). Furthermore, we can say that these two attributes are not independent and that the direction of the rule, has a logical implication due to its considerable high conviction value.

If we analyze the results of Figure C.3 3 - which presents the generated results with the *support* and *confidence* measure set 0.0 - we see that we have found other objectively and subjectively interesting rules (e.g. the rules identified with the No. 2133, 2168 and 2177 have the attribute “Gravidez” set as the consequent in all these rules which indicates that they are predictive rules. The smaller number of patients they encompass is 67 patients, i.e, 293 instances * 0.229).

C.3.2 Results of Step 2

In this test, we have tested the model presented in Figure 6.27 FP-Growth model with the same attributes specified in Figure C.3 1, and the *support* and *confidence* values set to 0.0 to look up for subjectively interesting association rules (second step of the association rule application). Since most subjectively interesting rules were previously seen with the “Gravidez” attribute set as the consequent attribute, we have in this test continued to explore those type of rules. Hence, the below Figure C.3 4 presents the generated rules ordered by its *support* in descendant order and filtered by the “Gravidez” attribute as a *consequent*.

Figure C.3 4 test 2.1: group 1 - unfiltered - support=0.0 and confidence=0.0



Show rules matching

all of these conclusions: ▾

A_B_Pre
Formas_N_3M
Conc_6M
Grau_Varicoc = II
Gravidez
ProfissãoComRiscoDeContactoD...
Grau_Varicoc = I
Grau_Varicoc = III

Min. Criterion:

confidence ▾

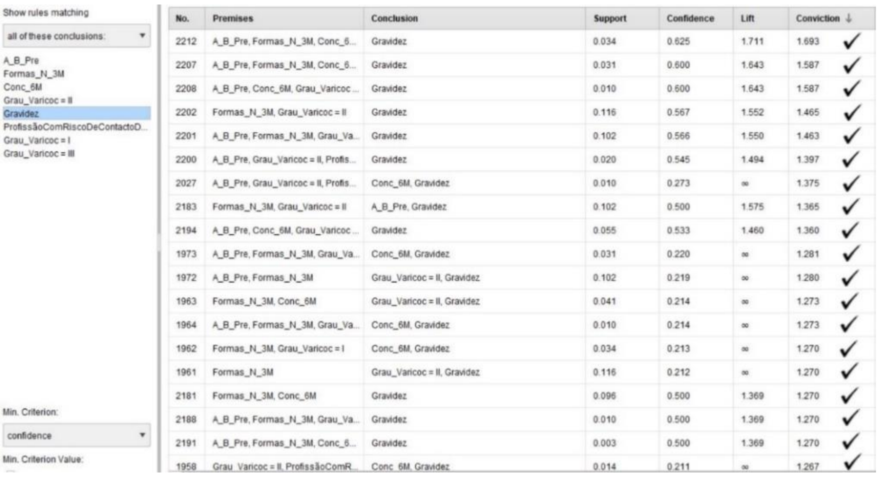
Min. Criterion Value:

No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction
2133	A_B_Pre	Gravidez ⚠	0.317	0.423	1.158	1.100 ✓
2168	Formas_N_3M	Gravidez ⚠	0.253	0.463	1.266	1.181 ✓
2044	A_B_Pre	Formas_N_3M, Gravidez ⚠	0.229	0.305	1.206	1.075
2130	Formas_N_3M	A_B_Pre, Gravidez	0.229	0.419	1.319	1.174 ✓
2177	A_B_Pre, Formas_N_3M	Gravidez ⚠	0.229	0.489	1.339	1.242 ✓
2132	Conc_6M	Gravidez ⚠	0.167	0.422	1.157	1.099
2146	Grau_Varicoc = II	Gravidez ⚠	0.164	0.432	1.184	1.118 ✓
1945	A_B_Pre	Conc_6M, Gravidez	0.147	0.195	∞	1.243 ✓
2103	Conc_6M	A_B_Pre, Gravidez	0.147	0.371	1.168	1.085
2163	A_B_Pre, Conc_6M	Gravidez	0.147	0.453	1.239	1.160 ✓
1914	A_B_Pre	Grau_Varicoc = II, Gravidez	0.137	0.182	∞	1.222 ✓
2091	Grau_Varicoc = II	A_B_Pre, Gravidez	0.137	0.360	1.135	1.067
2174	A_B_Pre, Grau_Varicoc = II	Gravidez	0.137	0.482	1.320	1.225 ✓
1961	Formas_N_3M	Grau_Varicoc = II, Gravidez	0.116	0.212	∞	1.270 ✓
2045	Grau_Varicoc = II	Formas_N_3M, Gravidez	0.116	0.306	1.213	1.077
2202	Formas_N_3M, Grau_Varicoc = II	Gravidez	0.116	0.567	1.552	1.465 ✓
1840	A_B_Pre	Formas_N_3M, Grau_Varicoc = II, Gravidez	0.102	0.136	∞	1.158 ✓
1932	Formas_N_3M	A_B_Pre, Grau_Varicoc = II, Gravidez	0.102	0.188	1.373	1.063
1972	A_B_Pre, Formas_N_3M	Grau_Varicoc = II, Gravidez	0.102	0.219	∞	1.280 ✓

By analyzing the results presented in Figure C.3 4, we can say that the most objectively and subjectively interesting association rule in this test is the first one:

A_B_Pre -> Gravidez, *support*=0.317, *confidence*=0.423, *lift*=1.158 and *conviction*=1.100. This rule tells us that 92 patients from the data set have a sperm motility above 0 before the treatment as well as a pregnancy after the treatment. The conditional probability of getting pregnant having before the treatment a sperm motility above 0 is of 42% (i.e. confidence=0.423).

If we order the generated rules by the *conviction* measure and continue on seeking rules with the “Gravidez” attribute as a consequent, we obtain the results presented in the below Figure C.3 5.



Show rules matching

all of these conclusions: ▾

A_B_Pre
Formas_N_3M
Conc_6M
Grau_Varicoc = II
Gravidez
ProfissãoComRiscoDeContactoD...
Grau_Varicoc = I
Grau_Varicoc = III

Min. Criterion:

confidence ▾

Min. Criterion Value:

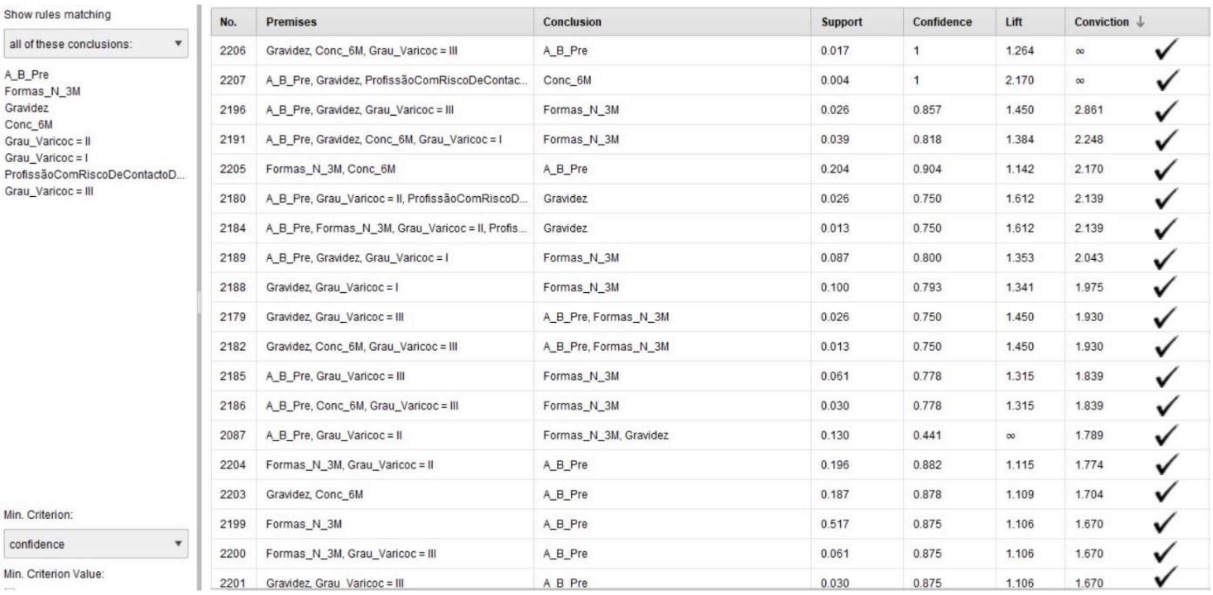
No.	Premises	Conclusion	Support	Confidence	Lift	Conviction ↓
2212	A_B_Pre, Formas_N_3M, Conc_6...	Gravidez	0.034	0.625	1.711	1.693 ✓
2207	A_B_Pre, Formas_N_3M, Conc_6...	Gravidez	0.031	0.600	1.643	1.587 ✓
2208	A_B_Pre, Conc_6M, Grau_Varicoc ...	Gravidez	0.010	0.600	1.643	1.587 ✓
2202	Formas_N_3M, Grau_Varicoc = II	Gravidez	0.116	0.567	1.552	1.465 ✓
2201	A_B_Pre, Formas_N_3M, Grau_Va...	Gravidez	0.102	0.566	1.550	1.463 ✓
2200	A_B_Pre, Grau_Varicoc = II, Profis...	Gravidez	0.020	0.545	1.494	1.397 ✓
2027	A_B_Pre, Grau_Varicoc = II, Profis...	Conc_6M, Gravidez	0.010	0.273	∞	1.375 ✓
2183	Formas_N_3M, Grau_Varicoc = II	A_B_Pre, Gravidez	0.102	0.500	1.575	1.365 ✓
2194	A_B_Pre, Conc_6M, Grau_Varicoc ...	Gravidez	0.055	0.533	1.460	1.360 ✓
1973	A_B_Pre, Formas_N_3M, Grau_Va...	Conc_6M, Gravidez	0.031	0.220	∞	1.281 ✓
1972	A_B_Pre, Formas_N_3M	Grau_Varicoc = II, Gravidez	0.102	0.219	∞	1.280 ✓
1963	Formas_N_3M, Conc_6M	Grau_Varicoc = II, Gravidez	0.041	0.214	∞	1.273 ✓
1964	A_B_Pre, Formas_N_3M, Grau_Va...	Conc_6M, Gravidez	0.010	0.214	∞	1.273 ✓
1962	Formas_N_3M, Grau_Varicoc = I	Conc_6M, Gravidez	0.034	0.213	∞	1.270 ✓
1961	Formas_N_3M	Grau_Varicoc = II, Gravidez	0.116	0.212	∞	1.270 ✓
2181	Formas_N_3M, Conc_6M	Gravidez	0.096	0.500	1.369	1.270 ✓
2188	A_B_Pre, Formas_N_3M, Grau_Va...	Gravidez	0.010	0.500	1.369	1.270 ✓
2191	A_B_Pre, Formas_N_3M, Conc_6...	Gravidez	0.003	0.500	1.369	1.270 ✓
1958	Grau_Varicoc = II, ProfissãoComR...	Conc_6M, Gravidez	0.014	0.211	∞	1.267 ✓

Figure C.3 5 test 2.2: group 1 - unfiltered - support=0.0 and confidence=0.0

If we analyze the generated results presented in the above Figure C.3 5, we see that this run did not gave any objectively and subjectively interesting rules since their support are all low (i.e. lower than 0.116 which is quite lower than 33 patients).

C.3.3 Results of Step 3

In this test, we have tested the model presented in Figure 6.28 FP-Growth model with the same attributes specified in Figure C.3 1. We have at first, maintained the *support* and *confidence* values to 0.0 to look up for objectively and subjectively interesting association rules upon the rows with no missing values in the “Gravidez” attribute (third step of the association rule application). The generated results are in Figure C.3 6 and Figure C.3 7 presented.



Show rules matching

all of these conclusions: ▾

A_B_Pre
Formas_N_3M
Gravidez
Conc_6M
Grau_Varicoc = II
Grau_Varicoc = I
ProfissãoComRiscoDeContactoD...
Grau_Varicoc = III

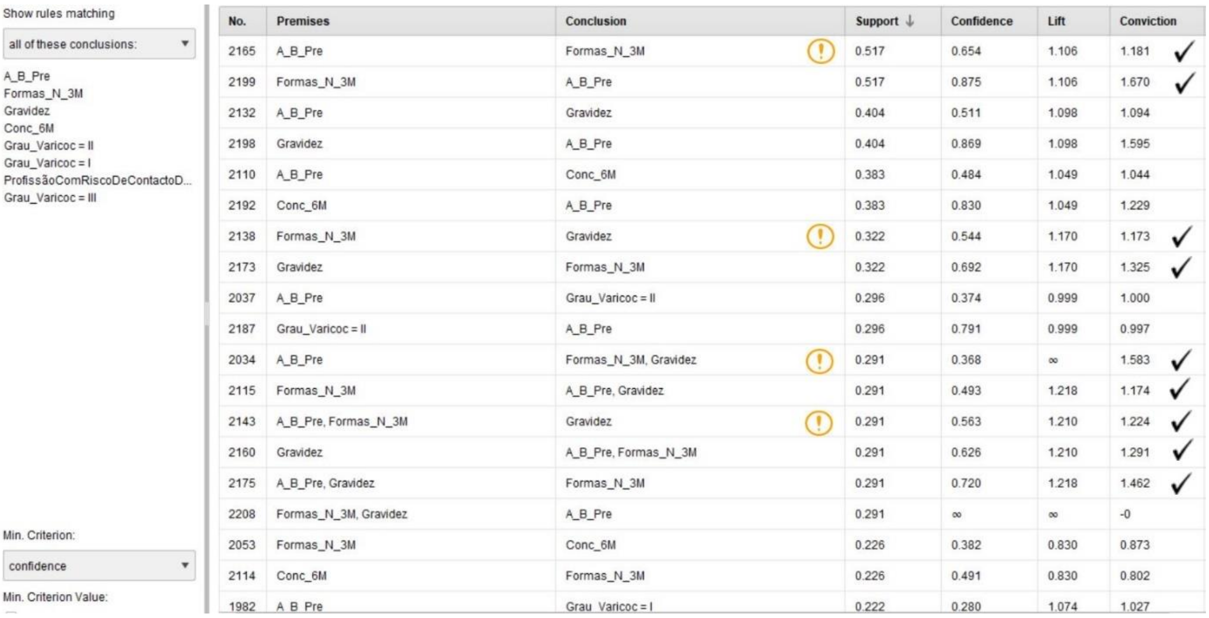
Min. Criterion:

confidence ▾

Min. Criterion Value:

No.	Premises	Conclusion	Support	Confidence	Lift	Conviction ↓	
2206	Gravidez, Conc_6M, Grau_Varicoc = III	A_B_Pre	0.017	1	1.264	∞	✓
2207	A_B_Pre, Gravidez, ProfissãoComRiscoDeContac...	Conc_6M	0.004	1	2.170	∞	✓
2196	A_B_Pre, Gravidez, Grau_Varicoc = III	Formas_N_3M	0.026	0.857	1.450	2.861	✓
2191	A_B_Pre, Gravidez, Conc_6M, Grau_Varicoc = I	Formas_N_3M	0.039	0.818	1.384	2.248	✓
2205	Formas_N_3M, Conc_6M	A_B_Pre	0.204	0.904	1.142	2.170	✓
2180	A_B_Pre, Grau_Varicoc = II, ProfissãoComRiscoD...	Gravidez	0.026	0.750	1.612	2.139	✓
2184	A_B_Pre, Formas_N_3M, Grau_Varicoc = II, Profis...	Gravidez	0.013	0.750	1.612	2.139	✓
2189	A_B_Pre, Gravidez, Grau_Varicoc = I	Formas_N_3M	0.087	0.800	1.353	2.043	✓
2188	Gravidez, Grau_Varicoc = I	Formas_N_3M	0.100	0.793	1.341	1.975	✓
2179	Gravidez, Grau_Varicoc = III	A_B_Pre, Formas_N_3M	0.026	0.750	1.450	1.930	✓
2182	Gravidez, Conc_6M, Grau_Varicoc = III	A_B_Pre, Formas_N_3M	0.013	0.750	1.450	1.930	✓
2185	A_B_Pre, Grau_Varicoc = III	Formas_N_3M	0.061	0.778	1.315	1.839	✓
2186	A_B_Pre, Conc_6M, Grau_Varicoc = III	Formas_N_3M	0.030	0.778	1.315	1.839	✓
2087	A_B_Pre, Grau_Varicoc = II	Formas_N_3M, Gravidez	0.130	0.441	∞	1.789	✓
2204	Formas_N_3M, Grau_Varicoc = II	A_B_Pre	0.196	0.882	1.115	1.774	✓
2203	Gravidez, Conc_6M	A_B_Pre	0.187	0.878	1.109	1.704	✓
2199	Formas_N_3M	A_B_Pre	0.517	0.875	1.106	1.670	✓
2200	Formas_N_3M, Grau_Varicoc = III	A_B_Pre	0.061	0.875	1.106	1.670	✓
2201	Gravidez, Grau_Varicoc = III	A_B_Pre	0.030	0.875	1.106	1.670	✓

Figure C.3 6 test 3.1: group 1 - filtered - support=0.0 and confidence=0.0



Show rules matching

all of these conclusions: ▾

A_B_Pre
Formas_N_3M
Gravidez
Conc_6M
Grau_Varicoc = II
Grau_Varicoc = I
ProfissãoComRiscoDeContactoD...
Grau_Varicoc = III

Min. Criterion:

confidence ▾

Min. Criterion Value:

No.	Premises	Conclusion		Support ↓	Confidence	Lift	Conviction	
2165	A_B_Pre	Formas_N_3M	⚠	0.517	0.654	1.106	1.181	✓
2199	Formas_N_3M	A_B_Pre		0.517	0.875	1.106	1.670	✓
2132	A_B_Pre	Gravidez		0.404	0.511	1.098	1.094	
2198	Gravidez	A_B_Pre		0.404	0.869	1.098	1.595	
2110	A_B_Pre	Conc_6M		0.383	0.484	1.049	1.044	
2192	Conc_6M	A_B_Pre		0.383	0.830	1.049	1.229	
2138	Formas_N_3M	Gravidez	⚠	0.322	0.544	1.170	1.173	✓
2173	Gravidez	Formas_N_3M		0.322	0.692	1.170	1.325	✓
2037	A_B_Pre	Grau_Varicoc = II		0.296	0.374	0.999	1.000	
2187	Grau_Varicoc = II	A_B_Pre		0.296	0.791	0.999	0.997	
2034	A_B_Pre	Formas_N_3M, Gravidez	⚠	0.291	0.368	∞	1.583	✓
2115	Formas_N_3M	A_B_Pre, Gravidez		0.291	0.493	1.218	1.174	✓
2143	A_B_Pre, Formas_N_3M	Gravidez	⚠	0.291	0.563	1.210	1.224	✓
2160	Gravidez	A_B_Pre, Formas_N_3M		0.291	0.626	1.210	1.291	✓
2175	A_B_Pre, Gravidez	Formas_N_3M		0.291	0.720	1.218	1.462	✓
2208	Formas_N_3M, Gravidez	A_B_Pre		0.291	∞	∞	-0	
2053	Formas_N_3M	Conc_6M		0.226	0.382	0.830	0.873	
2114	Conc_6M	Formas_N_3M		0.226	0.491	0.830	0.802	
1982	A_B_Pre	Grau_Varicoc = I		0.222	0.280	1.074	1.027	

Figure C.3 7 test 3.2: group 1 - filtered - support=0.0 and confidence=0.0

If we analyze these results, we can conclude that the rules that are more subjectively and objectively interesting are:

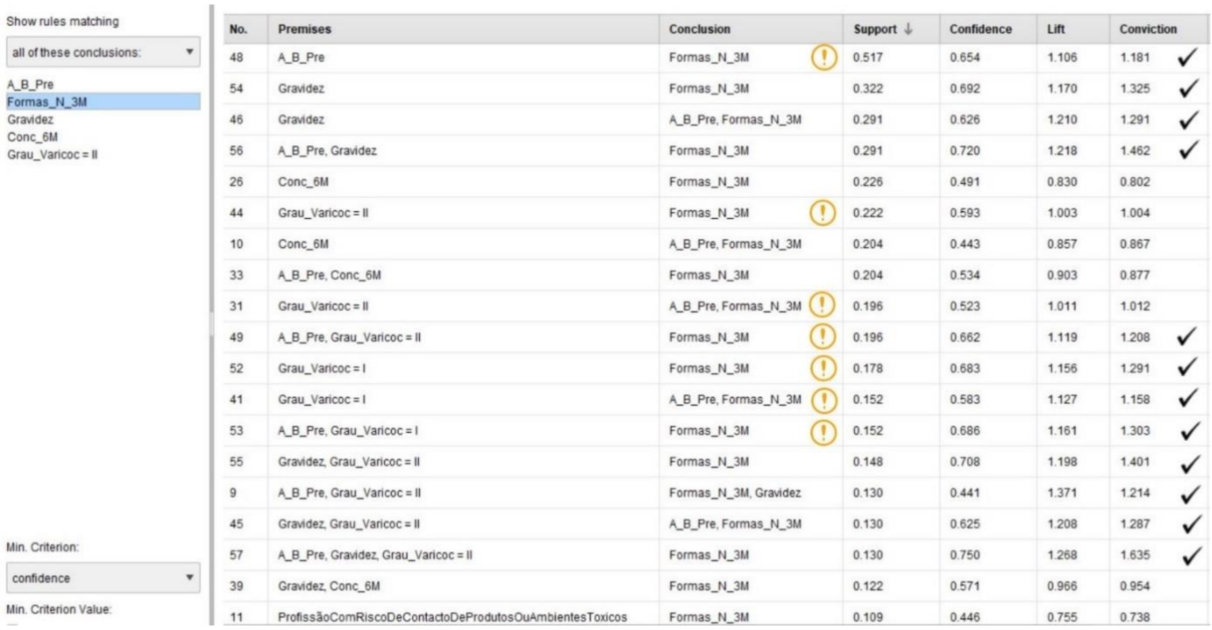
- A_B_Pre -> Formas_N_3M, *support*=0.517, *confidence*=0.654, *lift*=1.106 and *Conviction*=1.181.
- Formas_N_3M -> Gravidez, *support*=0.322, *confidence*=0.544, *lift*=1.170 and *Conviction*=1.173.
- A_B_Pre -> Formas_N_3M, Gravidez, *support*= 0.291, *confidence*=0.368, *lift*=+infinity and *Conviction*=1.583.
- A_B_Pre, Formas_N_3M -> Gravidez, *support*= 0.291, *confidence*=0.563, *lift*=1.210 and *Conviction*=1.224.

Based on the obtained results, we have seen that the most subjectively and objectively interesting rules would appear if we would set our *support* and *confidence* measure as in (Yildirim, 2015) since only one of the subjectively interesting rules presented in Figure C.3 6**Erro! A origem da referência não foi encontrada.** would not be generated with that setting (lowering the confidence to 0.29 would not be objectively interesting). Hence, we have reran the model depicted in Figure 6.28 for the attributes specified in Figure C.3 1 Selected attributes for the steps 1,2 and 3 of the FP_Growth algorithm

and for the following settings: *support*=0.1 and *confidence*=0.4 (forth step of the association rule application).

In order to better identify the most subjectively interesting rules, we have filtered the rules by its *consequent* attributes and have identified at each run the most interesting rules based on the conditions disclosed in the fourth step of the application of the FP_Growth algorithm.

In the below Figure C.3 8, we present the generated results by the FP_Growth algorithm for the consequent attribute “Formas_N_3M”. As we can see, the better rules are the ones identified with the No. 48, 44, 31, 49, 52, 41 and 53 since they all have *support* values above 0.15 and an *antecedent* that has occurred before, or at the same time of, the *consequent*.



Show rules matching

all of these conclusions: ▾

A_B_Pre
Formas_N_3M
Gravidez
Conc_6M
Grau_Varicoc = II

Min. Criterion:

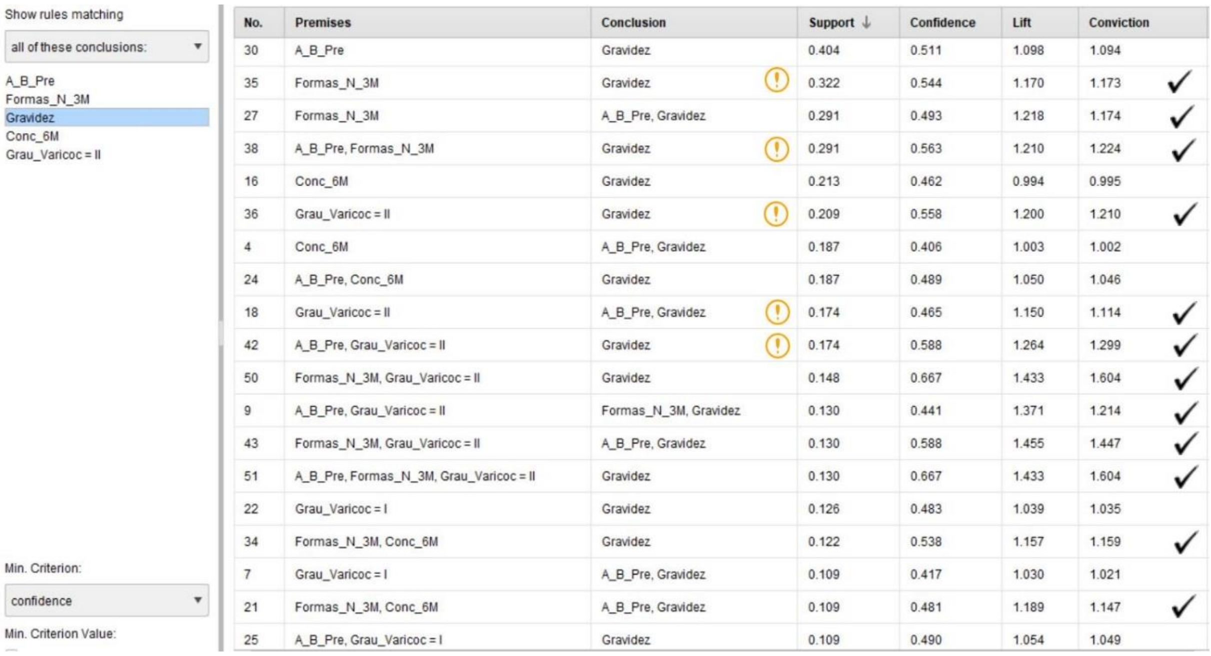
confidence ▾

Min. Criterion Value:

No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction
48	A_B_Pre	Formas_N_3M (!)	0.517	0.654	1.106	1.181 ✓
54	Gravidez	Formas_N_3M	0.322	0.692	1.170	1.325 ✓
46	Gravidez	A_B_Pre, Formas_N_3M	0.291	0.626	1.210	1.291 ✓
56	A_B_Pre, Gravidez	Formas_N_3M	0.291	0.720	1.218	1.462 ✓
26	Conc_6M	Formas_N_3M	0.226	0.491	0.830	0.802
44	Grau_Varicoc = II	Formas_N_3M (!)	0.222	0.593	1.003	1.004
10	Conc_6M	A_B_Pre, Formas_N_3M	0.204	0.443	0.857	0.867
33	A_B_Pre, Conc_6M	Formas_N_3M	0.204	0.534	0.903	0.877
31	Grau_Varicoc = II	A_B_Pre, Formas_N_3M (!)	0.196	0.523	1.011	1.012
49	A_B_Pre, Grau_Varicoc = II	Formas_N_3M (!)	0.196	0.662	1.119	1.208 ✓
52	Grau_Varicoc = I	Formas_N_3M (!)	0.178	0.683	1.156	1.291 ✓
41	Grau_Varicoc = I	A_B_Pre, Formas_N_3M (!)	0.152	0.583	1.127	1.158 ✓
53	A_B_Pre, Grau_Varicoc = I	Formas_N_3M (!)	0.152	0.686	1.161	1.303 ✓
55	Gravidez, Grau_Varicoc = II	Formas_N_3M	0.148	0.708	1.198	1.401 ✓
9	A_B_Pre, Grau_Varicoc = II	Formas_N_3M, Gravidez	0.130	0.441	1.371	1.214 ✓
45	Gravidez, Grau_Varicoc = II	A_B_Pre, Formas_N_3M	0.130	0.625	1.208	1.287 ✓
57	A_B_Pre, Gravidez, Grau_Varicoc = II	Formas_N_3M	0.130	0.750	1.268	1.635 ✓
39	Gravidez, Conc_6M	Formas_N_3M	0.122	0.571	0.966	0.954
11	ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	Formas_N_3M	0.109	0.446	0.755	0.738

Figure C.3 8 test 3.3: group 1 - “Formas_N_3M” as consequent - support=0.1 and confidence=0.4

In the below Figure C.3 9, we present the generated results by the FP_Growth algorithm for the consequent attribute “Gravidez”. We see that the better rules are the rules identified with the No. 35, 38,36,18 and 42.



Show rules matching

all of these conclusions: ▾

A_B_Pre
Formas_N_3M
Gravidez
Conc_6M
Grau_Varicoc = II

Min. Criterion:

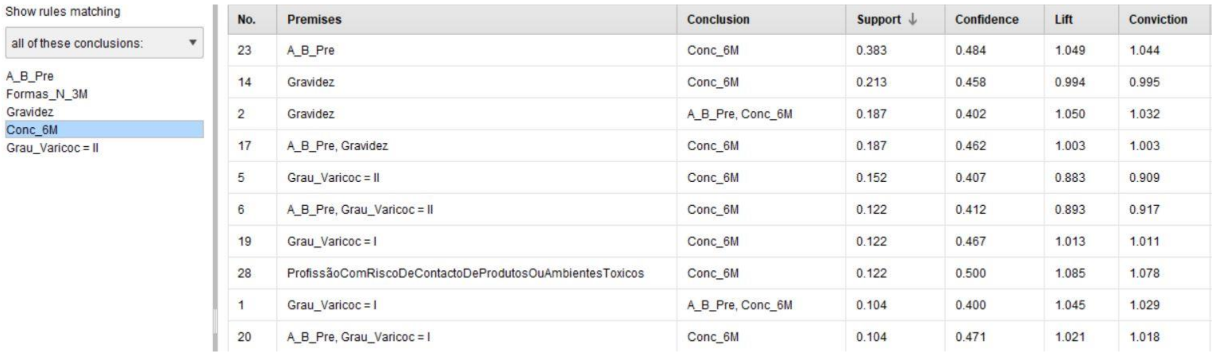
confidence ▾

Min. Criterion Value:

No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction
30	A_B_Pre	Gravidez	0.404	0.511	1.098	1.094
35	Formas_N_3M	Gravidez (!)	0.322	0.544	1.170	1.173 ✓
27	Formas_N_3M	A_B_Pre, Gravidez	0.291	0.493	1.218	1.174 ✓
38	A_B_Pre, Formas_N_3M	Gravidez (!)	0.291	0.563	1.210	1.224 ✓
16	Conc_6M	Gravidez	0.213	0.462	0.994	0.995
36	Grau_Varicoc = II	Gravidez (!)	0.209	0.558	1.200	1.210 ✓
4	Conc_6M	A_B_Pre, Gravidez	0.187	0.406	1.003	1.002
24	A_B_Pre, Conc_6M	Gravidez	0.187	0.489	1.050	1.046
18	Grau_Varicoc = II	A_B_Pre, Gravidez (!)	0.174	0.465	1.150	1.114 ✓
42	A_B_Pre, Grau_Varicoc = II	Gravidez (!)	0.174	0.588	1.264	1.299 ✓
50	Formas_N_3M, Grau_Varicoc = II	Gravidez	0.148	0.667	1.433	1.604 ✓
9	A_B_Pre, Grau_Varicoc = II	Formas_N_3M, Gravidez	0.130	0.441	1.371	1.214 ✓
43	Formas_N_3M, Grau_Varicoc = II	A_B_Pre, Gravidez	0.130	0.588	1.455	1.447 ✓
51	A_B_Pre, Formas_N_3M, Grau_Varicoc = II	Gravidez	0.130	0.667	1.433	1.604 ✓
22	Grau_Varicoc = I	Gravidez	0.126	0.483	1.039	1.035
34	Formas_N_3M, Conc_6M	Gravidez	0.122	0.538	1.157	1.159 ✓
7	Grau_Varicoc = I	A_B_Pre, Gravidez	0.109	0.417	1.030	1.021
21	Formas_N_3M, Conc_6M	A_B_Pre, Gravidez	0.109	0.481	1.189	1.147 ✓
25	A_B_Pre, Grau_Varicoc = I	Gravidez	0.109	0.490	1.054	1.049

Figure C.3 9 test 3.4: group 1 - “Gravidez “ as consequent - support=0.1 and confidence=0.4

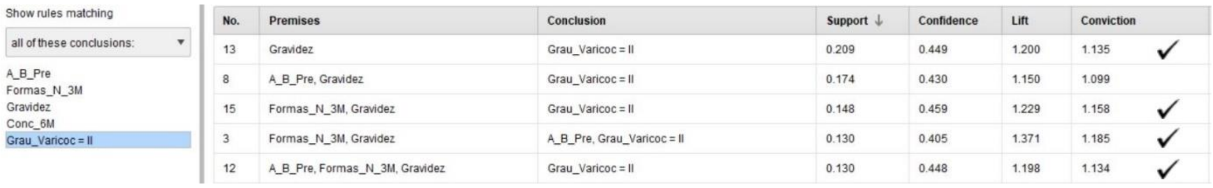
In the below Figure C.3 10, we present the generated results by the FP_Growth algorithm for the consequent attribute “Conc_6M”. As we can see, in this test we were not able to identify objectively interesting rules since all generate rules had a *lift* and a *conviction* measure value below 1.1; and therefore, we have concluded that in this run, we did not have a rule that stood out from the rest.



No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction
23	A_B_Pre	Conc_6M	0.383	0.484	1.049	1.044
14	Gravidez	Conc_6M	0.213	0.458	0.994	0.995
2	Gravidez	A_B_Pre, Conc_6M	0.187	0.402	1.050	1.032
17	A_B_Pre, Gravidez	Conc_6M	0.187	0.462	1.003	1.003
5	Grau_Varicoc = II	Conc_6M	0.152	0.407	0.883	0.909
6	A_B_Pre, Grau_Varicoc = II	Conc_6M	0.122	0.412	0.893	0.917
19	Grau_Varicoc = I	Conc_6M	0.122	0.467	1.013	1.011
28	ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	Conc_6M	0.122	0.500	1.085	1.078
1	Grau_Varicoc = I	A_B_Pre, Conc_6M	0.104	0.400	1.045	1.029
20	A_B_Pre, Grau_Varicoc = I	Conc_6M	0.104	0.471	1.021	1.018

Figure C.3 10 test 3.5: group 1 - “Conc_6M” as consequent - support=0.1 and confidence=0.4

In the below Figure C.3 11, we present the generated results by the FP_Growth algorithm for the consequent attribute “Grau_Varicoc=II”. As we can see, this run did not give subjectively interesting rules since all antecedents occurred after the consequent attribute “Grau_Varicoc=II”. In fact, the severity grade of the varicocele condition (i.e. “Grau_Varicoc”) is known before the *embolization* treatment.



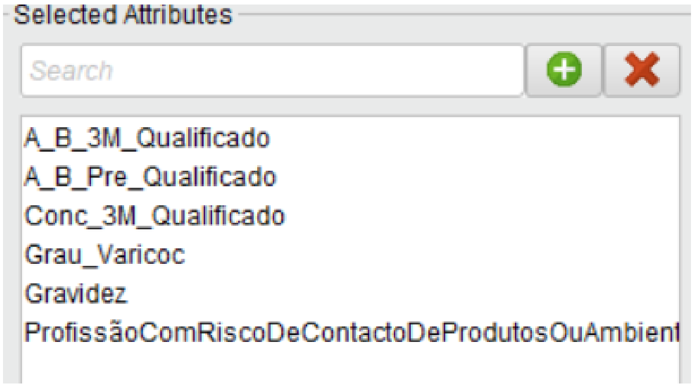
No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction	
13	Gravidez	Grau_Varicoc = II	0.209	0.449	1.200	1.135	✓
8	A_B_Pre, Gravidez	Grau_Varicoc = II	0.174	0.430	1.150	1.099	
15	Formas_N_3M, Gravidez	Grau_Varicoc = II	0.148	0.459	1.229	1.158	✓
3	Formas_N_3M, Gravidez	A_B_Pre, Grau_Varicoc = II	0.130	0.405	1.371	1.185	✓
12	A_B_Pre, Formas_N_3M, Gravidez	Grau_Varicoc = II	0.130	0.448	1.198	1.134	✓

Figure C.3 11 test 3.6: group 1 - “Grau_Varicoc=II” as consequent - support=0.1 and confidence=0.4

C.3.4 Results of Step 4

In this step, we have tested the model presented in Figure 6.28 with the attributes specified in the below Figure C.3 12 to look up for objectively and subjectively interesting association rules upon the rows with no missing values under the “Gravidez” attribute with the following thresholds: *support* >0.1 and a *confidence* > 0.4 (forth step of the association rule application).

In this test, we have obtained the results presented in Figure C.3 13, Figure C.3 14, Figure C.3 15 and Figure C.3 16. Each of these tables depicts the association rules that were respectively computed for the following *consequent* attributes: “Gravidez”, “Conc_3M_Qualificado”, “A_B_3M_Qualificado” and “Grau_Varicoc=II”. As we can see, the *consequent* “ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos” does not here appear because association rules with this consequent attribute were not computed in this test with the specified settings.

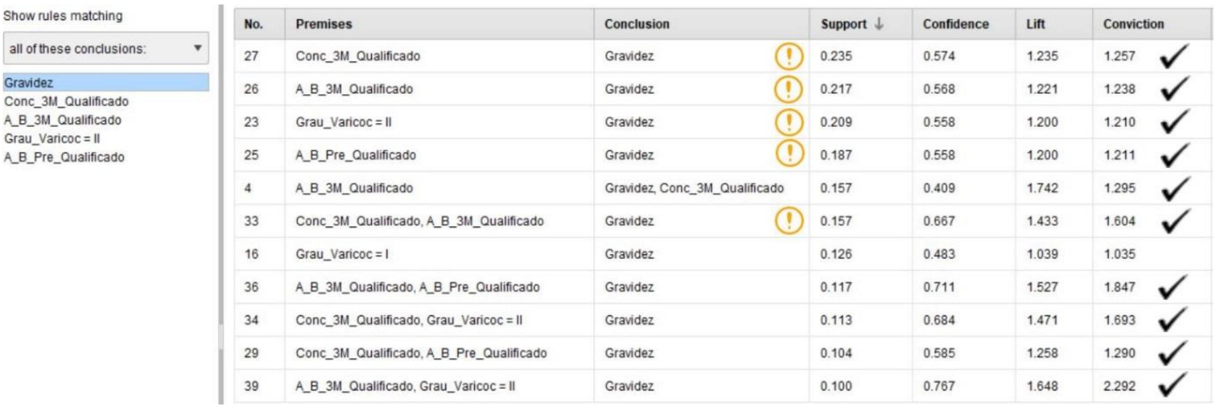


Selected Attributes

A_B_3M_Qualificado
A_B_Pre_Qualificado
Conc_3M_Qualificado
Grau_Varicoc
Gravidez
ProfissãoComRiscoDeContactoDeProdutosOuAmbient

Figure C.3 12 Selected attributes for the step 4 of the FP_Growth algorithm

For the consequent attribute “Gravidez”, the most interesting rules were the ones with the No. 27, 26, 23, 25 e 33 (see Figure C.3 13) since they at least encompass 36 patients of the data set. The rule that presented the highest *support* encompassed 54 patients (i.e. a support value of 0.235) which is not a very high *support* value.



Show rules matching

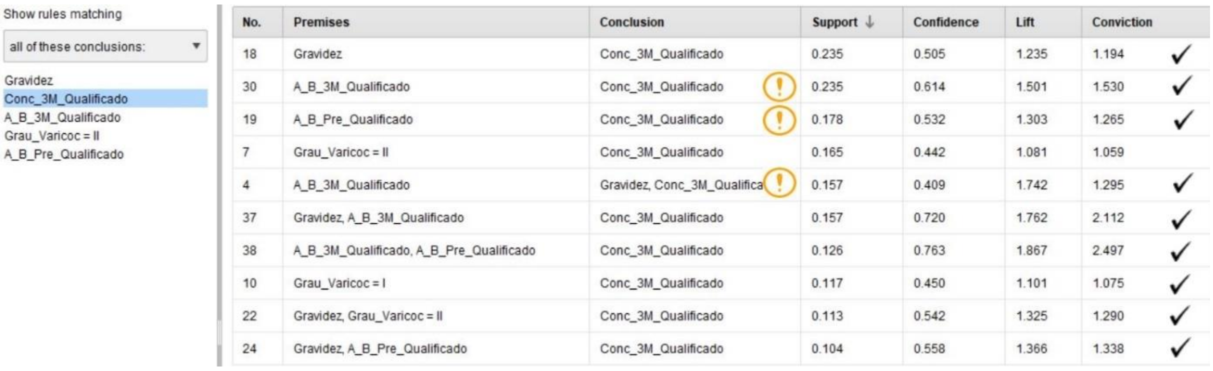
all of these conclusions:

Gravidez
Conc_3M_Qualificado
A_B_3M_Qualificado
Grau_Varicoc = II
A_B_Pre_Qualificado

No.	Premises	Conclusion		Support ↓	Confidence	Lift	Conviction	
27	Conc_3M_Qualificado	Gravidez	!	0.235	0.574	1.235	1.257	✓
26	A_B_3M_Qualificado	Gravidez	!	0.217	0.568	1.221	1.238	✓
23	Grau_Varicoc = II	Gravidez	!	0.209	0.558	1.200	1.210	✓
25	A_B_Pre_Qualificado	Gravidez	!	0.187	0.558	1.200	1.211	✓
4	A_B_3M_Qualificado	Gravidez, Conc_3M_Qualificado		0.157	0.409	1.742	1.295	✓
33	Conc_3M_Qualificado, A_B_3M_Qualificado	Gravidez	!	0.157	0.667	1.433	1.604	✓
16	Grau_Varicoc = I	Gravidez		0.126	0.483	1.039	1.035	
36	A_B_3M_Qualificado, A_B_Pre_Qualificado	Gravidez		0.117	0.711	1.527	1.847	✓
34	Conc_3M_Qualificado, Grau_Varicoc = II	Gravidez		0.113	0.684	1.471	1.693	✓
29	Conc_3M_Qualificado, A_B_Pre_Qualificado	Gravidez		0.104	0.585	1.258	1.290	✓
39	A_B_3M_Qualificado, Grau_Varicoc = II	Gravidez		0.100	0.767	1.648	2.292	✓

Figure C.3 13 test 4.1: group 2 - “Gravidez” as consequent - support=0.1 and confidence=0.4

For the consequent attribute “Conc_3M_Qualificado”, we have seen that the rules with the No. 30, 19 and 4 were in this run the most interesting ones. These rules are in the below Figure C.3 14 depicted with the check mark and the exclamation mark as in the other runs.



Show rules matching

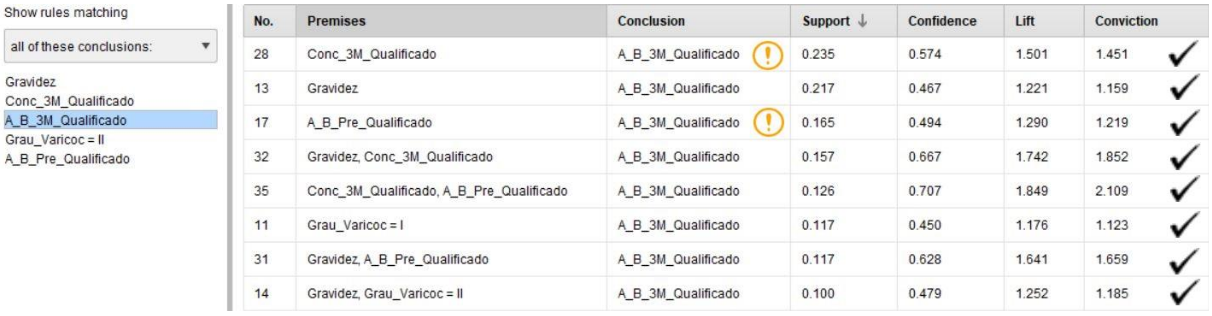
all of these conclusions:

Gravidez
Conc_3M_Qualificado
A_B_3M_Qualificado
Grau_Varicoc = II
A_B_Pre_Qualificado

No.	Premises	Conclusion		Support ↓	Confidence	Lift	Conviction	
18	Gravidez	Conc_3M_Qualificado		0.235	0.505	1.235	1.194	✓
30	A_B_3M_Qualificado	Conc_3M_Qualificado	!	0.235	0.614	1.501	1.530	✓
19	A_B_Pre_Qualificado	Conc_3M_Qualificado	!	0.178	0.532	1.303	1.265	✓
7	Grau_Varicoc = II	Conc_3M_Qualificado		0.165	0.442	1.081	1.059	
4	A_B_3M_Qualificado	Gravidez, Conc_3M_Qualifica	!	0.157	0.409	1.742	1.295	✓
37	Gravidez, A_B_3M_Qualificado	Conc_3M_Qualificado		0.157	0.720	1.762	2.112	✓
38	A_B_3M_Qualificado, A_B_Pre_Qualificado	Conc_3M_Qualificado		0.126	0.763	1.867	2.497	✓
10	Grau_Varicoc = I	Conc_3M_Qualificado		0.117	0.450	1.101	1.075	✓
22	Gravidez, Grau_Varicoc = II	Conc_3M_Qualificado		0.113	0.542	1.325	1.290	✓
24	Gravidez, A_B_Pre_Qualificado	Conc_3M_Qualificado		0.104	0.558	1.366	1.338	✓

Figure C.3 14 test 4.2: group 2 - “Conc_3M_Qualificado” as consequent - support=0.1 and confidence=0.4

For the consequent attribute “A_B_3M_Qualificado”, the algorithm has generated the rules presented in Figure C.3 15 where we can see that the rules with the No. 28 and 17 were the most interesting ones.



Show rules matching

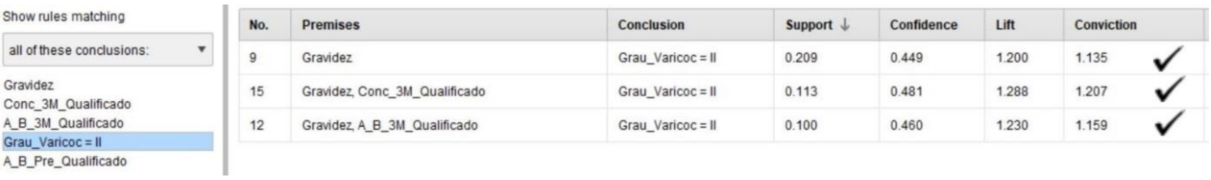
all of these conclusions: ▾

Gravidez
Conc_3M_Qualificado
A_B_3M_Qualificado
Grau_Varicoc = II
A_B_Pre_Qualificado

No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction	
28	Conc_3M_Qualificado	A_B_3M_Qualificado !	0.235	0.574	1.501	1.451	✓
13	Gravidez	A_B_3M_Qualificado	0.217	0.467	1.221	1.159	✓
17	A_B_Pre_Qualificado	A_B_3M_Qualificado !	0.165	0.494	1.290	1.219	✓
32	Gravidez, Conc_3M_Qualificado	A_B_3M_Qualificado	0.157	0.667	1.742	1.852	✓
35	Conc_3M_Qualificado, A_B_Pre_Qualificado	A_B_3M_Qualificado	0.126	0.707	1.849	2.109	✓
11	Grau_Varicoc = I	A_B_3M_Qualificado	0.117	0.450	1.176	1.123	✓
31	Gravidez, A_B_Pre_Qualificado	A_B_3M_Qualificado	0.117	0.628	1.641	1.659	✓
14	Gravidez, Grau_Varicoc = II	A_B_3M_Qualificado	0.100	0.479	1.252	1.185	✓

Figure C.3 15 test 4.3: group 2 - “A_B_3M_Qualificado” as consequent - support=0.1 and confidence=0.4

For the consequent attribute “Grau_Varicoc=II”, the algorithm has generated the rules presented in Figure C.3 16. As we can see, in this run, in spite of all rules being objectively interesting, none of them is subjectively interesting.



Show rules matching

all of these conclusions: ▾

Gravidez
Conc_3M_Qualificado
A_B_3M_Qualificado
Grau_Varicoc = II
A_B_Pre_Qualificado

No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction	
9	Gravidez	Grau_Varicoc = II	0.209	0.449	1.200	1.135	✓
15	Gravidez, Conc_3M_Qualificado	Grau_Varicoc = II	0.113	0.481	1.288	1.207	✓
12	Gravidez, A_B_3M_Qualificado	Grau_Varicoc = II	0.100	0.460	1.230	1.159	✓

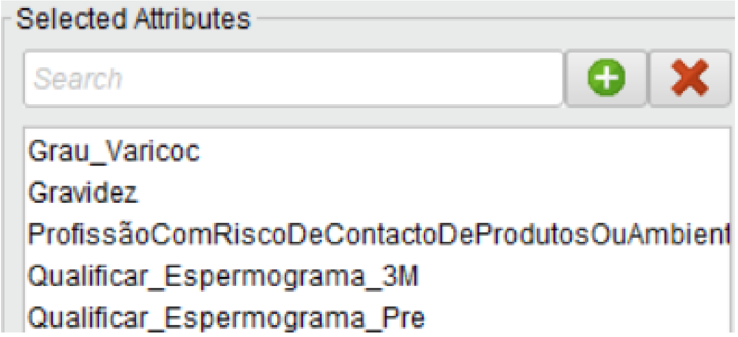
Figure C.3 16 test 4.4: group 2 - “Grau_Varicoc=II” as consequent - support=0.1 and confidence=0.4

Since the attribute “A_B_Pre_Qualificado” does not enable subjectively interesting rules as a consequent attribute, we have not presented its related results.

C.3.5 Results of Step 5

In this test, we have tested the model presented in Figure 6.28 with the attributes specified in Figure C.3 17 to look up for objectively and subjectively interesting association rules upon the rows with no missing values in the “Gravidez” attribute with the following thresholds: *support* >0.1 and a *confidence* > 0.4 (fifth step of the association rule application).

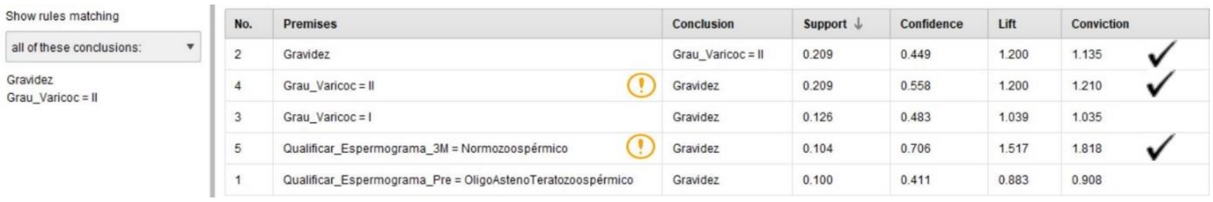
In this test, the obtained results for all tested attributes as consequents are in Figure C.3 18 presented.



Selected Attributes

Grau_Varicoc
Gravidez
ProfissãoComRiscoDeContactoDeProdutosOuAmbient
Qualificar_Espermograma_3M
Qualificar_Espermograma_Pre

Figure C.3 17 Selected attributes for the step 5 of the FP_Growth algorithm



Show rules matching

all of these conclusions:

Gravidez
Grau_Varicoc = II

No.	Premises	Conclusion	Support ↓	Confidence	Lift	Conviction	
2	Gravidez	Grau_Varicoc = II	0.209	0.449	1.200	1.135	✓
4	Grau_Varicoc = II	Gravidez	0.209	0.558	1.200	1.210	✓
3	Grau_Varicoc = I	Gravidez	0.126	0.483	1.039	1.035	
5	Qualificar_Espermograma_3M = Normozoospérmico	Gravidez	0.104	0.706	1.517	1.818	✓
1	Qualificar_Espermograma_Pre = OligoAstenoTeratozoospérmico	Gravidez	0.100	0.411	0.883	0.908	

Figure C.3 18 test 5.1: group 3 - support=0.1 and confidence=0.4

This test has generated less rules than the other ones. However, it has computed two objectively and subjectively interesting rules since we have considered one of them as a subjectively interesting rule in spite of its very low *support* (i.e. Qualificar_Espermograma_3M=Normozoospérmico -> Gravidez, *support*=0.104, *confidence*=0.706, *lift*=1.517, *conviction*=1.818) due to its other good measure values and clinical interest.

C.3.6 Results of Step 6

In Table C.3 1, we present all generated association rules by the model depicted in Figure 6.29 that applies the FP-Growth algorithm with the following discretized or dichotomized attributes upon the data set filtered by non-missing values in the "Gravidez" attribute (n=230):

> Idade_M, Grau_Varicoc, Conc_3M , Conc_6M, A_B_pré, A_B_3M, Formas_N_3M,
> ProfissãoComRiscoDeContactoDeProdutosOuAmbientes,
> HabitosTabagicos_Processado_Simplificado,
> HabitosAlcoolicos_Processado_Simplificado,
> Gravidez, PMA, Gravidez_espontanea.

Table C.3 2 discloses all generated results by the same model depicted in Figure 6.29, but this time, the data set was filtered by "Gravidez" = "Sim" and the "Gravidez" attribute was then excluded from the test with the "Feature Selection" operator (n=107). This model was also applied upon the semen categorizations (i.e. the 6th group of attributes) but the identified interesting rules were already encompassed in Table B.3:17. Therefore, these identified interesting rules did not encompass semen categorizations.

The association rules are ordered in descending order by the *support* value and the rules with a p<=0.10 have its Chi-square values highlighted in bold.

Table C.3 1 test 6.1: group 4 & 5 - support=0.1 and confidence=0.4 - “Gravidez” = “Sim” OR “Não” (n=230)

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
157	Gravidez	PMA	0,287	0,617	2,15	1,861	**106,4718**
193	PMA	Gravidez	0,287	1	2,15	∞	**106,4677**
97	Gravidez	Conc_3M = > 15	0,235	0,505	1,235	1,194	**7,647794**
142	Conc_3M = > 15	Gravidez	0,235	0,574	1,235	1,257	**7,646138**
143	Conc_3M = > 15	A_B_3M = > 32	0,235	0,574	1,501	1,451	**24,78013**
155	A_B_3M = > 32	Conc_3M = > 15	0,235	0,614	1,501	1,53	**24,7786**
132	Conc_3M = > 15	Formas_N_3M = > 4	0,23	0,564	1,729	1,545	**40,74905**
176	Formas_N_3M = > 4	Conc_3M = > 15	0,23	0,707	1,729	2,016	**40,77189**
69	Gravidez	A_B_Pre = 1 to 31	0,217	0,467	1,024	1,02	0,096412
70	Gravidez	A_B_3M = > 32	0,217	0,467	1,221	1,159	**6,039175**
73	A_B_Pre = 1 to 31	Gravidez	0,217	0,476	1,024	1,021	0,096413
133	A_B_3M = > 32	Gravidez	0,217	0,568	1,221	1,238	**6,040879**
61	Gravidez	Gravidez_espontanea	0,213	0,458	2,15	1,452	**71,58172**
194	Gravidez_espontanea	Gravidez	0,213	1	2,15	∞	**71,5864**
50	Gravidez	Grau_Varicoc = II	0,209	0,449	1,2	1,135	**4,789931**
60	A_B_Pre = 1 to 31	A_B_3M = 1 to 31	0,209	0,457	1,348	1,217	**12,0398**
122	Grau_Varicoc = II	Gravidez	0,209	0,558	1,2	1,21	**4,788592**
156	A_B_3M = 1 to 31	A_B_Pre = 1 to 31	0,209	0,615	1,348	1,413	**12,03033**
49	A_B_Pre = 1 to 31	Conc_3M = > 15	0,204	0,448	1,095	1,07	1,201681
88	Conc_3M = > 15	A_B_Pre = 1 to 31	0,204	0,5	1,095	1,087	1,202172
111	A_B_3M = > 32	Formas_N_3M = > 4	0,204	0,534	1,638	1,446	**27,99353**
162	Formas_N_3M = > 4	A_B_3M = > 32	0,204	0,627	1,638	1,654	**28,00112**
36	A_B_Pre = 1 to 31	Conc_3M = 0,01 to 14,9	0,196	0,429	1,095	1,065	1,124756
89	Conc_3M = 0,01 to 14,9	A_B_Pre = 1 to 31	0,196	0,5	1,095	1,087	1,124632
3	Gravidez	A_B_Pre = > 32	0,187	0,402	1,2	1,112	**4,031012**
124	A_B_Pre = > 32	Gravidez	0,187	0,558	1,2	1,211	**4,030461**
125	Formas_N_3M = > 4	Gravidez	0,183	0,56	1,204	1,215	**4,040174**
41	Conc_3M = > 15	A_B_Pre = > 32	0,178	0,436	1,303	1,18	**7,326234**
59	Conc_3M = 0,01 to 14,9	A_B_3M = 1 to 31	0,178	0,456	1,343	1,214	**8,907022**

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
108	A_B_3M = 1 to 31	Conc_3M = 0,01 to 14,9	0,178	0,526	1,343	1,283	**8,910878**
110	A_B_Pre = > 32	Conc_3M = > 15	0,178	0,532	1,303	1,265	**7,32634**
19	Conc_3M = > 15	Formas_N_3M = 1 to 3	0,17	0,415	1,564	1,256	**18,33568**
87	HabitosTabagicos_Processado_Simplificado	Gravidez	0,17	0,494	1,061	1,056	0,391233
116	Idade_M = Range 1 <31	Gravidez	0,17	0,549	1,181	1,187	**2,935959**
165	Formas_N_3M = 1 to 3	Conc_3M = > 15	0,17	0,639	1,564	1,64	**18,31982**
5	Conc_3M = > 15	Grau_Varicoc = II	0,165	0,404	1,081	1,051	0,621694
27	Conc_3M = 0,01 to 14,9	Gravidez	0,165	0,422	0,908	0,926	1,085252
38	A_B_3M = > 32	A_B_Pre = > 32	0,165	0,432	1,29	1,171	**6,018565**
43	Grau_Varicoc = II	Conc_3M = > 15	0,165	0,442	1,081	1,059	0,621761
86	A_B_Pre = > 32	A_B_3M = > 32	0,165	0,494	1,29	1,219	**6,020406**
26	A_B_3M = > 32	A_B_Pre = 1 to 31	0,161	0,42	0,921	0,938	0,748031
37	Grau_Varicoc = II	A_B_Pre = 1 to 31	0,161	0,43	0,942	0,954	0,388916
72	A_B_3M = 1 to 31	Conc_3M = > 15	0,161	0,474	1,161	1,125	2,11584
1	Conc_3M = 0,01 to 14,9	Conc_6M = 0,01 to 14,9	0,157	0,4	1,533	1,232	**14,9041**
10	A_B_3M = > 32	Gravidez, Conc_3M = > 15	0,157	0,409	1,742	1,295	**24,20627**
147	Conc_6M = 0,01 to 14,9	Conc_3M = 0,01 to 14,9	0,157	0,6	1,533	1,522	**14,89182**
168	Gravidez, Conc_3M = > 15	A_B_3M = > 32	0,157	0,667	1,742	1,852	**24,18703**
169	Conc_3M = > 15, A_B_3M = > 32	Gravidez	0,157	0,667	1,433	1,604	**11,55926**
182	Gravidez, A_B_3M = > 32	Conc_3M = > 15	0,157	0,72	1,762	2,112	**25,7332**
44	HabitosTabagicos_Processado_Simplificado	Conc_3M = 0,01 to 14,9	0,152	0,443	1,132	1,093	1,345894
45	HabitosTabagicos_Processado_Simplificado	A_B_3M = > 32	0,152	0,443	1,158	1,108	1,858191
51	A_B_3M = 1 to 31	Gravidez	0,152	0,449	0,965	0,97	0,125472
67	Formas_N_3M = > 4	A_B_Pre = 1 to 31	0,152	0,467	1,022	1,019	0,045199
55	Formas_N_3M = > 4	Conc_3M = > 15, A_B_3M = > 32	0,148	0,453	1,931	1,4	**29,64919**
121	Formas_N_3M = 1 to 3	A_B_Pre = 1 to 31	0,148	0,557	1,221	1,228	**3,409871**
164	Conc_3M = > 15, A_B_3M = > 32	Formas_N_3M = > 4	0,148	0,63	1,931	1,82	**29,64184**
166	Conc_3M = > 15, Formas_N_3M = > 4	A_B_3M = > 32	0,148	0,642	1,677	1,722	**19,59001**
183	A_B_3M = > 32, Formas_N_3M = > 4	Conc_3M = > 15	0,148	0,723	1,77	2,138	**24,23789**
23	HabitosTabagicos_Processado_Simplificado	A_B_Pre = 1 to 31	0,143	0,418	0,915	0,933	0,726756

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
24	HabitosTabagicos_Processado_Simplificado	Conc_3M = > 15	0,143	0,418	1,022	1,016	0,04006
25	HabitosTabagicos_Processado_Simplificado	Idade_M = Range 1 <31	0,143	0,418	1,353	1,187	**6,662623**
42	Formas_N_3M = > 4	Gravidez, Conc_3M = > 15	0,143	0,44	1,874	1,366	**25,95576**
66	Idade_M = Range 1 <31	HabitosTabagicos_Processado_Simplificado	0,143	0,465	1,353	1,227	**6,664958**
90	PMA	A_B_Pre = 1 to 31	0,143	0,5	1,095	1,087	0,698708
91	PMA	Conc_3M = > 15	0,143	0,5	1,223	1,183	**3,168379**
93	PMA	Gravidez, A_B_Pre = 1 to 31	0,143	0,5	2,3	1,565	**43,24938**
94	Gravidez, PMA	A_B_Pre = 1 to 31	0,143	0,5	1,095	1,087	0,698708
95	PMA	Gravidez, Conc_3M = > 15	0,143	0,5	2,13	1,53	**36,08568**
96	Gravidez, PMA	Conc_3M = > 15	0,143	0,5	1,223	1,183	**3,168379**
153	Gravidez, Conc_3M = > 15	Formas_N_3M = > 4	0,143	0,611	1,874	1,733	**25,97041**
154	Gravidez, Conc_3M = > 15	PMA	0,143	0,611	2,13	1,834	**36,09594**
161	Conc_3M = > 15, Formas_N_3M = > 4	Gravidez	0,143	0,623	1,338	1,417	**6,820843**
167	Gravidez, A_B_Pre = 1 to 31	PMA	0,143	0,66	2,3	2,097	**43,26733**
190	Gravidez, Formas_N_3M = > 4	Conc_3M = > 15	0,143	0,786	1,922	2,759	**30,08562**
195	A_B_Pre = 1 to 31, PMA	Gravidez	0,143	1	2,15	∞	**44,13477**
196	Conc_3M = > 15, PMA	Gravidez	0,143	1	2,15	∞	**44,13477**
6	HabitosTabagicos_Processado_Simplificado	Grau_Varicoc = II	0,139	0,405	1,083	1,052	0,494587
7	HabitosTabagicos_Processado_Simplificado	A_B_Pre = > 32	0,139	0,405	1,21	1,118	2,666605
20	A_B_Pre = > 32	HabitosTabagicos_Processado_Simplificado	0,139	0,416	1,21	1,123	2,666701
107	Formas_N_3M = 1 to 3	Gravidez	0,139	0,525	1,128	1,125	1,181455
4	A_B_Pre = > 32	Grau_Varicoc = II	0,135	0,403	1,077	1,048	0,410727
17	Formas_N_3M = > 4	Grau_Varicoc = II	0,135	0,413	1,105	1,067	0,734919
98	Formas_N_3M = 1 to 3	A_B_3M = 1 to 31	0,135	0,508	1,499	1,344	**10,62537**
100	Grau_Varicoc = I	A_B_Pre = 1 to 31	0,135	0,517	1,132	1,124	1,190588
28	Idade_M = Range 1 <31	A_B_Pre = 1 to 31	0,13	0,423	0,926	0,941	0,469937
29	Idade_M = Range 1 <31	Conc_3M = 0,01 to 14,9	0,13	0,423	1,08	1,054	0,420493
30	Idade_M = Range 1 <31	A_B_3M = > 32	0,13	0,423	1,104	1,069	0,685588
56	PMA	HabitosTabagicos_Processado_Simplificado	0,13	0,455	1,323	1,204	**5,03135**
57	PMA	Gravidez, HabitosTabagicos_Processado_Simplificado	0,13	0,455	2,681	1,522	**53,13852**

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
58	Gravidez, PMA	HabitosTabagicos_Processado_Simplificado	0,13	0,455	1,323	1,204	**5,03135**
188	Gravidez, HabitosTabagicos_Processado_Simplificado	PMA	0,13	0,769	2,681	3,09	**53,1795**
203	HabitosTabagicos_Processado_Simplificado, PMA	Gravidez	0,13	1	2,15	∞	**39,52299**
8	Idade_M = Range 1 <31	Conc_3M = > 15	0,126	0,408	0,999	1	7,09E-05
80	Grau_Varicoc = I	Gravidez	0,126	0,483	1,039	1,035	0,107258
112	Conc_3M = > 15, A_B_3M = > 32	A_B_Pre = > 32	0,126	0,537	1,604	1,437	**12,94614**
177	Conc_3M = > 15, A_B_Pre = > 32	A_B_3M = > 32	0,126	0,707	1,849	2,109	**22,25825**
186	A_B_3M = > 32, A_B_Pre = > 32	Conc_3M = > 15	0,126	0,763	1,867	2,497	**23,63484**
31	PMA	A_B_3M = > 32	0,122	0,424	1,109	1,072	0,683296
32	PMA	Gravidez, A_B_3M = > 32	0,122	0,424	1,952	1,359	**23,36669**
33	Gravidez, PMA	A_B_3M = > 32	0,122	0,424	1,109	1,072	0,683296
62	Formas_N_3M = 1 to 3	A_B_3M = > 32	0,122	0,459	1,2	1,141	2,063062
68	Conc_6M = 0,01 to 14,9	A_B_Pre = 1 to 31	0,122	0,467	1,022	1,019	0,033124
79	Idade_M = Range 4 <36	A_B_Pre = 1 to 31	0,122	0,483	1,057	1,051	0,212503
101	Gravidez, Conc_3M = > 15	A_B_Pre = 1 to 31	0,122	0,519	1,136	1,129	1,099657
102	Gravidez, Conc_3M = > 15	Gravidez_espontanea	0,122	0,519	2,434	1,634	**39,39073**
126	Gravidez, A_B_Pre = 1 to 31	Conc_3M = > 15	0,122	0,56	1,37	1,344	**6,063453**
127	Gravidez, A_B_3M = > 32	Formas_N_3M = > 4	0,122	0,56	1,717	1,532	**15,94067**
128	Gravidez, A_B_3M = > 32	PMA	0,122	0,56	1,952	1,621	**23,35803**
129	Gravidez, A_B_3M = > 32	Gravidez_espontanea	0,122	0,56	2,629	1,789	**46,01329**
134	Gravidez_espontanea	Conc_3M = > 15	0,122	0,571	1,398	1,38	**6,834993**
135	Gravidez_espontanea	A_B_3M = > 32	0,122	0,571	1,494	1,441	**9,434733**
136	Gravidez_espontanea	Gravidez, Conc_3M = > 15	0,122	0,571	2,434	1,786	**39,3879**
137	Gravidez, Gravidez_espontanea	Conc_3M = > 15	0,122	0,571	1,398	1,38	**6,834993**
138	Gravidez_espontanea	Gravidez, A_B_3M = > 32	0,122	0,571	2,629	1,826	**46,01233**
139	Gravidez, Gravidez_espontanea	A_B_3M = > 32	0,122	0,571	1,494	1,441	**9,434733**
145	A_B_Pre = 1 to 31, Conc_3M = > 15	Gravidez	0,122	0,596	1,281	1,323	**4,067033**
146	A_B_3M = > 32, Formas_N_3M = > 4	Gravidez	0,122	0,596	1,281	1,323	**4,067033**
150	Idade_M = Range 2 31 to 32	Gravidez	0,122	0,609	1,308	1,367	**4,762119**
151	Conc_6M = > 15	Conc_3M = > 15	0,122	0,609	1,489	1,511	**9,534788**

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
170	Gravidez, Formas_N_3M = > 4	A_B_3M = > 32	0,122	0,667	1,742	1,852	**17,588**
197	Conc_3M = > 15, Gravidez_espontanea	Gravidez	0,122	1	2,15	∞	**36,75285**
199	A_B_3M = > 32, PMA	Gravidez	0,122	1	2,15	∞	**36,75285**
200	A_B_3M = > 32, Gravidez_espontanea	Gravidez	0,122	1	2,15	∞	**36,75285**
9	PMA	Idade_M = Range 1 <31	0,117	0,409	1,325	1,17	**4,346357**
11	PMA	Gravidez, Idade_M = Range 1 <31	0,117	0,409	2,413	1,405	**37,55266**
12	Gravidez, PMA	Idade_M = Range 1 <31	0,117	0,409	1,325	1,17	**4,346357**
52	Grau_Varicoc = I	Conc_3M = > 15	0,117	0,45	1,101	1,075	0,569828
53	Grau_Varicoc = I	A_B_3M = > 32	0,117	0,45	1,176	1,123	1,551568
113	Gravidez, A_B_3M = > 32	A_B_Pre = > 32	0,117	0,54	1,613	1,446	**12,03063**
117	Gravidez_espontanea	Grau_Varicoc = II	0,117	0,551	1,474	1,394	**8,316308**
118	Gravidez_espontanea	Gravidez, Grau_Varicoc = II	0,117	0,551	2,64	1,762	**43,98705**
119	Gravidez, Gravidez_espontanea	Grau_Varicoc = II	0,117	0,551	1,474	1,394	**8,316308**
131	Gravidez, Grau_Varicoc = II	Gravidez_espontanea	0,117	0,562	2,64	1,799	**43,98779**
163	Gravidez, A_B_Pre = > 32	A_B_3M = > 32	0,117	0,628	1,641	1,659	**13,41402**
174	Gravidez, Idade_M = Range 1 <31	PMA	0,117	0,692	2,413	2,317	**37,57123**
178	A_B_3M = > 32, A_B_Pre = > 32	Gravidez	0,117	0,711	1,527	1,847	**10,96296**
202	Grau_Varicoc = II, Gravidez_espontanea	Gravidez	0,117	1	2,15	∞	**35,047**
207	Idade_M = Range 1 <31, PMA	Gravidez	0,117	1	2,15	∞	**35,047**
35	Formas_N_3M = 1 to 3	HabitosTabagicos_Processado_Simplificado	0,113	0,426	1,241	1,144	2,520856
78	Gravidez, Conc_3M = > 15	Grau_Varicoc = II	0,113	0,481	1,288	1,207	**3,491524**
114	Gravidez, Grau_Varicoc = II	Conc_3M = > 15	0,113	0,542	1,325	1,29	**4,429485**
115	Gravidez, Grau_Varicoc = II	PMA	0,113	0,542	1,888	1,556	**19,23664**
148	Gravidez, A_B_Pre = > 32	PMA	0,113	0,605	2,107	1,804	**26,07485**
172	Conc_3M = > 15, Grau_Varicoc = II	Gravidez	0,113	0,684	1,471	1,693	**8,775934**
173	Gravidez, Conc_3M = 0,01 to 14,9	PMA	0,113	0,684	2,384	2,258	**35,07919**
198	Conc_3M = 0,01 to 14,9, PMA	Gravidez	0,113	1	2,15	∞	**33,69617**
201	Grau_Varicoc = II, PMA	Gravidez	0,113	1	2,15	∞	**33,69617**
205	A_B_Pre = > 32, PMA	Gravidez	0,113	1	2,15	∞	**33,69617**
13	Formas_N_3M = 1 to 3	PMA	0,109	0,41	1,428	1,208	**6,144852**

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
14	Formas_N_3M = 1 to 3	Gravidez, PMA	0,109	0,41	1,428	1,208	**6,144852**
21	Conc_6M = 0,01 to 14,9	Gravidez	0,109	0,417	0,896	0,917	0,766427
22	Grau_Varicoc = I	Conc_3M = 0,01 to 14,9	0,109	0,417	1,065	1,043	0,221305
65	Idade_M = Range 3 33 to 35	A_B_Pre = 1 to 31	0,109	0,463	1,014	1,012	0,011664
71	Conc_3M = > 15, Formas_N_3M = > 4	A_B_Pre = 1 to 31	0,109	0,472	1,033	1,029	0,063278
92	Gravidez, A_B_Pre = 1 to 31	A_B_3M = 1 to 31	0,109	0,5	1,474	1,322	**7,395121**
103	A_B_Pre = 1 to 31, A_B_3M = 1 to 31	Gravidez	0,109	0,521	1,12	1,116	0,762133
104	A_B_Pre = 1 to 31, A_B_3M = 1 to 31	Conc_3M = 0,01 to 14,9	0,109	0,521	1,331	1,27	**4,288109**
109	A_B_Pre = 1 to 31, Conc_3M = > 15	Formas_N_3M = > 4	0,109	0,532	1,631	1,44	**11,42317**
120	A_B_Pre = 1 to 31, Conc_3M = 0,01 to 14,9	A_B_3M = 1 to 31	0,109	0,556	1,638	1,487	**11,73102**
152	Conc_3M = 0,01 to 14,9, A_B_3M = 1 to 31	A_B_Pre = 1 to 31	0,109	0,61	1,336	1,393	**4,746663**
179	Gravidez, A_B_3M = 1 to 31	A_B_Pre = 1 to 31	0,109	0,714	1,565	1,902	**11,09851**
180	Gravidez, A_B_3M = 1 to 31	PMA	0,109	0,714	2,489	2,496	**36,95628**
181	A_B_Pre = 1 to 31, Formas_N_3M = > 4	Conc_3M = > 15	0,109	0,714	1,748	2,07	**16,00958**
189	Gravidez, Formas_N_3M = 1 to 3	PMA	0,109	0,781	2,723	3,26	**44,54074**
204	A_B_3M = 1 to 31, PMA	Gravidez	0,109	1	2,15	∞	**32,35746**
208	PMA, Formas_N_3M = 1 to 3	Gravidez	0,109	1	2,15	∞	**32,35746**
2	Grau_Varicoc = I	Formas_N_3M = > 4	0,104	0,4	1,227	1,123	2,014075
18	Idade_M = Range 4 <36	Conc_3M = > 15	0,104	0,414	1,012	1,009	0,007692
46	Gravidez, Conc_3M = > 15	A_B_Pre = > 32	0,104	0,444	1,328	1,197	**3,80155**
47	Gravidez, Conc_3M = > 15	A_B_3M = > 32, Formas_N_3M = > 4	0,104	0,444	2,175	1,432	**24,91402**
48	Conc_3M = > 15, A_B_3M = > 32	Gravidez, Formas_N_3M = > 4	0,104	0,444	2,434	1,471	**32,27829**
54	Conc_3M = > 15, Formas_N_3M = > 4	Gravidez, A_B_3M = > 32	0,104	0,453	2,083	1,43	**22,34103**
77	Gravidez, A_B_3M = > 32	Conc_3M = > 15, Formas_N_3M = > 4	0,104	0,48	2,083	1,48	**22,34282**
83	Gravidez_espontanea	Formas_N_3M = > 4	0,104	0,49	1,502	1,321	**7,561307**
84	Gravidez_espontanea	Gravidez, Formas_N_3M = > 4	0,104	0,49	2,682	1,602	**39,19055**
85	Gravidez, Gravidez_espontanea	Formas_N_3M = > 4	0,104	0,49	1,502	1,321	**7,561307**
99	A_B_3M = > 32, Formas_N_3M = > 4	Gravidez, Conc_3M = > 15	0,104	0,511	2,175	1,564	**24,91782**
105	Conc_6M = > 15	Gravidez	0,104	0,522	1,121	1,118	0,730129
106	Conc_6M = > 15	A_B_Pre = > 32	0,104	0,522	1,558	1,391	**8,977678**

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
123	Gravidez, A_B_Pre = > 32	Conc_3M = > 15	0,104	0,558	1,366	1,338	**4,874054**
140	Gravidez, Formas_N_3M = > 4	Gravidez_espontanea	0,104	0,571	2,682	1,836	**39,19622**
141	Gravidez, Formas_N_3M = > 4	Conc_3M = > 15, A_B_3M = > 32	0,104	0,571	2,434	1,786	**32,2824**
144	Conc_3M = > 15, A_B_Pre = > 32	Gravidez	0,104	0,585	1,258	1,29	**2,877374**
171	Gravidez, Conc_3M = > 15, A_B_3M = > 32	Formas_N_3M = > 4	0,104	0,667	2,044	2,022	**22,43085**
175	Conc_3M = > 15, A_B_3M = > 32, Formas_N_3M = > 4	Gravidez	0,104	0,706	1,517	1,818	**9,245483**
184	Gravidez, Conc_3M = > 15, Formas_N_3M = > 4	A_B_3M = > 32	0,104	0,727	1,901	2,264	**19,30141**
192	Gravidez, A_B_3M = > 32, Formas_N_3M = > 4	Conc_3M = > 15	0,104	0,857	2,097	4,139	**26,42035**
206	Formas_N_3M = > 4, Gravidez_espontanea	Gravidez	0,104	1	2,15	∞	**30,70089**
15	ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	Conc_3M = 0,01 to 14,9	0,1	0,411	1,05	1,033	0,118918
16	ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	Idade_M = Range 1 <31	0,1	0,411	1,33	1,173	**3,601817**
34	Conc_3M = > 15, A_B_3M = > 32	A_B_Pre = 1 to 31	0,1	0,426	0,933	0,947	0,26611
39	Conc_3M = > 15, Formas_N_3M = > 4	Grau_Varicoc = II	0,1	0,434	1,161	1,106	1,065586
40	Conc_3M = > 15, Formas_N_3M = > 4	A_B_Pre = > 32	0,1	0,434	1,296	1,175	**3,037717**
63	Gravidez, A_B_Pre = 1 to 31	Grau_Varicoc = II	0,1	0,46	1,23	1,159	2,019055
64	Gravidez, A_B_3M = > 32	Grau_Varicoc = II	0,1	0,46	1,23	1,159	2,019055
74	Gravidez, Grau_Varicoc = II	A_B_Pre = 1 to 31	0,1	0,479	1,05	1,043	0,127271
75	Gravidez, Grau_Varicoc = II	A_B_3M = > 32	0,1	0,479	1,252	1,185	2,388063
76	A_B_Pre = 1 to 31, A_B_3M = 1 to 31	Conc_3M = > 15	0,1	0,479	1,172	1,135	1,240931
81	A_B_Pre = 1 to 31, Conc_3M = > 15	A_B_3M = > 32	0,1	0,489	1,279	1,209	**2,848842**
82	A_B_Pre = 1 to 31, Conc_3M = > 15	A_B_3M = 1 to 31	0,1	0,489	1,443	1,294	**5,947661**
130	Conc_3M = > 15, A_B_Pre = > 32	Formas_N_3M = > 4	0,1	0,561	1,72	1,535	**12,51905**
149	Conc_3M = > 15, Grau_Varicoc = II	Formas_N_3M = > 4	0,1	0,605	1,856	1,707	**16,1392**
158	A_B_Pre = 1 to 31, Grau_Varicoc = II	Gravidez	0,1	0,622	1,336	1,413	**4,333393**
159	A_B_Pre = 1 to 31, A_B_3M = > 32	Conc_3M = > 15	0,1	0,622	1,521	1,563	**8,274914**
160	Conc_3M = > 15, A_B_3M = 1 to 31	A_B_Pre = 1 to 31	0,1	0,622	1,362	1,436	**4,853255**
185	Grau_Varicoc = II, Formas_N_3M = > 4	Conc_3M = > 15	0,1	0,742	1,815	2,291	**16,45554**
187	A_B_3M = > 32, Grau_Varicoc = II	Gravidez	0,1	0,767	1,648	2,292	**12,60583**
191	A_B_Pre = > 32, Formas_N_3M = > 4	Conc_3M = > 15	0,1	0,852	2,084	3,991	**24,85407**

Table C.3 2 test 6.1.1: group 4 & 5 - support=0.1 and confidence=0.4 – “Gravidez” = “Sim” (n=107)

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
362	A_B_3M = > 32	Conc_3M = > 15	0,336	0,72	1,427	1,769	**17,38443832**
323	Conc_3M = > 15	A_B_3M = > 32	0,336	0,667	1,427	1,598	**17,38053149**
398	Formas_N_3M = > 4	Conc_3M = > 15	0,308	0,786	1,557	2,312	**21,80645934**
142	PMA	Conc_3M = > 15	0,308	0,5	0,991	0,991	0,014158159
143	PMA	A_B_Pre = 1 to 31	0,308	0,5	1,07	1,065	0,73777595
285	Conc_3M = > 15	Formas_N_3M = > 4	0,308	0,611	1,557	1,562	**21,79477199**
284	Conc_3M = > 15	PMA	0,308	0,611	0,991	0,985	0,014165584
322	A_B_Pre = 1 to 31	PMA	0,308	0,66	1,07	1,127	0,738495732
390	HabitosTabagicos_Processado_Simplificado	PMA	0,28	0,769	1,247	1,66	**6,01347209**
93	PMA	HabitosTabagicos_Processado_Simplificado	0,28	0,455	1,247	1,165	**6,000450838**
41	PMA	A_B_3M = > 32	0,262	0,424	0,908	0,925	1,283116838
180	Conc_3M = > 15	A_B_Pre = 1 to 31	0,262	0,519	1,11	1,106	1,159090091
181	Conc_3M = > 15	Gravidez_espontanea	0,262	0,519	1,132	1,126	1,609187652
324	Formas_N_3M = > 4	A_B_3M = > 32	0,262	0,667	1,427	1,598	**11,07638486**
228	A_B_3M = > 32	PMA	0,262	0,56	0,908	0,871	1,28130713
229	A_B_Pre = 1 to 31	Conc_3M = > 15	0,262	0,56	1,11	1,126	1,15898953
238	Gravidez_espontanea	Conc_3M = > 15	0,262	0,571	1,132	1,156	1,608969083
239	Gravidez_espontanea	A_B_3M = > 32	0,262	0,571	1,223	1,243	**3,951161925**
230	A_B_3M = > 32	Gravidez_espontanea	0,262	0,56	1,223	1,232	**3,951418507**
231	A_B_3M = > 32	Formas_N_3M = > 4	0,262	0,56	1,427	1,381	**11,07881944**
18	PMA	Idade_M = Range 1 <31	0,252	0,409	1,122	1,075	1,46635156
302	A_B_Pre = > 32	A_B_3M = > 32	0,252	0,628	1,344	1,432	**7,443204994**
236	Grau_Varicoc = II	Gravidez_espontanea	0,252	0,562	1,228	1,239	**3,815525892**
224	Gravidez_espontanea	Grau_Varicoc = II	0,252	0,551	1,228	1,228	**3,815449891**
208	A_B_3M = > 32	A_B_Pre = > 32	0,252	0,54	1,344	1,3	**7,441259104**
349	Idade_M = Range 1 <31	PMA	0,252	0,692	1,122	1,245	1,467874555
130	Conc_3M = > 15	Grau_Varicoc = II	0,243	0,481	1,073	1,063	0,473022919
209	Grau_Varicoc = II	PMA	0,243	0,542	0,878	0,836	2,087846344
210	Grau_Varicoc = II	Conc_3M = > 15	0,243	0,542	1,073	1,081	0,47300894
279	A_B_Pre = > 32	PMA	0,243	0,605	0,98	0,969	0,046351691

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
340	Conc_3M = 0.01 to 14.9	PMA	0,243	0,684	1,109	1,213	1,127383383
396	Formas_N_3M = 1 to 3	PMA	0,234	0,781	1,267	1,752	**5,243841981**
359	A_B_3M = 1 to 31	A_B_Pre = 1 to 31	0,234	0,714	1,529	1,864	**12,78822724**
358	A_B_3M = 1 to 31	PMA	0,234	0,714	1,158	1,341	2,094053694
145	A_B_Pre = 1 to 31	A_B_3M = 1 to 31	0,234	0,5	1,529	1,346	**12,7992338**
423	A_B_3M = > 32, Formas_N_3M = > 4	Conc_3M = > 15	0,224	0,857	1,698	3,467	**18,79852427**
77	Conc_3M = > 15	A_B_Pre = > 32	0,224	0,444	1,106	1,077	0,821005046
365	Conc_3M = > 15, Formas_N_3M = > 4	A_B_3M = > 32	0,224	0,727	1,556	1,953	**12,91794034**
327	Conc_3M = > 15, A_B_3M = > 32	Formas_N_3M = > 4	0,224	0,667	1,698	1,822	**17,05320778**
227	A_B_Pre = > 32	Conc_3M = > 15	0,224	0,558	1,106	1,121	0,82101426
240	Formas_N_3M = > 4	Gravidez_espontanea	0,224	0,571	1,248	1,265	**3,583052094**
243	Formas_N_3M = > 4	Conc_3M = > 15, A_B_3M = > 32	0,224	0,571	1,698	1,548	**17,05003829**
141	Gravidez_espontanea	Formas_N_3M = > 4	0,224	0,49	1,248	1,191	**3,582452126**
127	A_B_3M = > 32	Conc_3M = > 15, Formas_N_3M = > 4	0,224	0,48	1,556	1,33	**12,91131212**
80	Conc_3M = > 15	A_B_3M = > 32, Formas_N_3M = > 4	0,224	0,444	1,698	1,329	**18,79340159**
101	A_B_Pre = 1 to 31	Grau_Varicoc = II	0,215	0,46	1,025	1,021	0,047779935
102	A_B_3M = > 32	Grau_Varicoc = II	0,215	0,46	1,025	1,021	0,047779935
122	Grau_Varicoc = II	A_B_Pre = 1 to 31	0,215	0,479	1,025	1,023	0,047779456
123	Grau_Varicoc = II	A_B_3M = > 32	0,215	0,479	1,025	1,023	0,047779456
95	Grau_Varicoc = II	Formas_N_3M = > 4	0,206	0,458	1,168	1,121	1,592487932
342	Formas_N_3M = 1 to 3	A_B_Pre = 1 to 31	0,206	0,688	1,471	1,705	**8,913996316**
187	Formas_N_3M = > 4	PMA	0,206	0,524	0,849	0,805	2,54815416
188	Formas_N_3M = > 4	Grau_Varicoc = II	0,206	0,524	1,168	1,158	1,591797657
69	A_B_Pre = 1 to 31	Formas_N_3M = 1 to 3	0,206	0,44	1,471	1,252	**8,918071832**
430	Conc_3M = > 15, A_B_Pre = > 32	A_B_3M = > 32	0,196	0,875	1,873	4,262	**20,6384023**
28	A_B_Pre = 1 to 31	Gravidez_espontanea	0,196	0,42	0,917	0,935	0,545055739
29	A_B_3M = > 32	Idade_M = Range 1 <31	0,196	0,42	1,152	1,096	1,241129836
47	Gravidez_espontanea	A_B_Pre = 1 to 31	0,196	0,429	0,917	0,932	0,545101758
48	Gravidez_espontanea	A_B_Pre = > 32	0,196	0,429	1,066	1,047	0,26405212
393	A_B_3M = > 32, A_B_Pre = > 32	Conc_3M = > 15	0,196	0,778	1,541	2,229	**10,75391139**
377	PMA, A_B_3M = > 32	Conc_3M = > 15	0,196	0,75	1,486	1,981	**9,111420117**
138	A_B_Pre = > 32	Gravidez_espontanea	0,196	0,488	1,066	1,059	0,264141629

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
201	Idade_M = Range 1 <31	A_B_3M = > 32	0,196	0,538	1,152	1,154	1,241408811
305	PMA, Conc_3M = > 15	A_B_3M = > 32	0,196	0,636	1,362	1,465	**5,471731882**
266	Conc_3M = > 15, A_B_3M = > 32	PMA	0,196	0,583	0,946	0,92	0,253792237
268	Conc_3M = > 15, A_B_3M = > 32	A_B_Pre = > 32	0,196	0,583	1,452	1,436	**7,427703034**
150	Formas_N_3M = > 4	A_B_Pre = > 32	0,196	0,5	1,244	1,196	**2,760215478**
139	A_B_Pre = > 32	Formas_N_3M = > 4	0,196	0,488	1,244	1,187	**2,760162663**
140	A_B_Pre = > 32	Conc_3M = > 15, A_B_3M = > 32	0,196	0,488	1,452	1,297	**7,428081418**
30	A_B_3M = > 32	PMA, Conc_3M = > 15	0,196	0,42	1,362	1,192	**5,470252325**
31	A_B_3M = > 32	Conc_3M = > 15, A_B_Pre = > 32	0,196	0,42	1,873	1,337	**20,62547667**
1	A_B_3M = > 32	HabitosTabagicos_Processado_Simplificado	0,187	0,4	1,097	1,059	0,507243467
17	Gravidez_espontanea	Conc_3M = > 15, A_B_3M = > 32	0,187	0,408	1,213	1,121	2,081884348
360	Conc_3M = > 15, Gravidez_espontanea	A_B_3M = > 32	0,187	0,714	1,529	1,864	**9,308221646**
361	A_B_3M = > 32, Gravidez_espontanea	Conc_3M = > 15	0,187	0,714	1,415	1,734	**6,660259585**
175	HabitosTabagicos_Processado_Simplificado	A_B_3M = > 32	0,187	0,513	1,097	1,093	0,50728938
226	Conc_3M = > 15, A_B_3M = > 32	Gravidez_espontanea	0,187	0,556	1,213	1,22	2,081937304
348	Grau_Varicoc = I	PMA	0,187	0,69	1,118	1,235	0,892949165
6	A_B_3M = > 32	Conc_3M = > 15, Gravidez_espontanea	0,187	0,4	1,529	1,231	**9,31371993**
433	A_B_Pre = > 32, Formas_N_3M = > 4	Conc_3M = > 15	0,178	0,905	1,793	5,201	**16,79002556**
428	PMA, Formas_N_3M = > 4	Conc_3M = > 15	0,178	0,864	1,711	3,632	**14,3169068**
407	Conc_3M = > 15, A_B_Pre = > 32	Formas_N_3M = > 4	0,178	0,792	2,017	2,916	**20,74275356**
256	Conc_3M = > 15, Formas_N_3M = > 4	PMA	0,178	0,576	0,933	0,903	0,346596825
276	Formas_N_3M = 1 to 3	Conc_3M = > 15	0,178	0,594	1,177	1,219	1,461419414
255	PMA, Conc_3M = > 15	Formas_N_3M = > 4	0,178	0,576	1,467	1,432	**6,746821938**
257	Conc_3M = > 15, Formas_N_3M = > 4	A_B_Pre = > 32	0,178	0,576	1,433	1,41	**6,0302895**
91	Formas_N_3M = > 4	PMA, Conc_3M = > 15	0,178	0,452	1,467	1,263	**6,750864185**
92	Formas_N_3M = > 4	Conc_3M = > 15, A_B_Pre = > 32	0,178	0,452	2,017	1,416	**20,76439185**
76	A_B_Pre = > 32	Conc_3M = > 15, Formas_N_3M = > 4	0,178	0,442	1,433	1,239	**6,03287408**
425	A_B_Pre = > 32, Formas_N_3M = > 4	A_B_3M = > 32	0,168	0,857	1,834	3,729	**15,91813549**
414	A_B_Pre = 1 to 31, Formas_N_3M = 1 to 3	PMA	0,168	0,818	1,326	2,107	**4,732651162**
397	A_B_3M = > 32, Grau_Varicoc = II	Gravidez_espontanea	0,168	0,783	1,709	2,493	**12,42399275**
103	Idade_M = Range 1 <31	Conc_3M = > 15	0,168	0,462	0,915	0,92	0,450533775
104	HabitosTabagicos_Processado_Simplificado	Conc_3M = > 15	0,168	0,462	0,915	0,92	0,450533775

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
381	Gravidez_espontanea, Formas_N_3M = > 4	A_B_3M = > 32	0,168	0,75	1,605	2,131	**9,916900886**
363	PMA, Formas_N_3M = 1 to 3	A_B_Pre = 1 to 31	0,168	0,72	1,541	1,903	**8,358682925**
330	Gravidez_espontanea, Grau_Varicoc = II	A_B_3M = > 32	0,168	0,667	1,427	1,598	**5,764484894**
331	A_B_3M = > 32, A_B_Pre = > 32	Formas_N_3M = > 4	0,168	0,667	1,698	1,822	**11,3545657**
215	PMA, Conc_3M = > 15	A_B_Pre = 1 to 31	0,168	0,545	1,167	1,172	1,165173957
216	PMA, A_B_Pre = 1 to 31	Conc_3M = > 15	0,168	0,545	1,081	1,09	0,318092522
310	A_B_3M = > 32, Gravidez_espontanea	Grau_Varicoc = II	0,168	0,643	1,433	1,544	**5,775094885**
311	A_B_3M = > 32, Gravidez_espontanea	Formas_N_3M = > 4	0,168	0,643	1,638	1,701	**9,954711987**
312	A_B_3M = > 32, Formas_N_3M = > 4	Gravidez_espontanea	0,168	0,643	1,404	1,518	**5,219014412**
313	A_B_3M = > 32, Formas_N_3M = > 4	A_B_Pre = > 32	0,168	0,643	1,6	1,675	**9,15379086**
309	Conc_3M = > 15, A_B_Pre = 1 to 31	PMA	0,168	0,643	1,042	1,073	0,107581132
237	Formas_N_3M = 1 to 3	PMA, A_B_Pre = 1 to 31	0,168	0,562	1,824	1,581	**13,795207**
217	PMA, A_B_Pre = 1 to 31	Formas_N_3M = 1 to 3	0,168	0,545	1,824	1,542	**13,79532989**
176	A_B_3M = 1 to 31	Conc_3M = 0.01 to 14.9	0,168	0,514	1,448	1,328	**5,738379038**
118	Conc_3M = 0.01 to 14.9	A_B_3M = 1 to 31	0,168	0,474	1,448	1,279	**5,737821662**
55	Formas_N_3M = > 4	A_B_3M = > 32, Gravidez_espontanea	0,168	0,429	1,638	1,292	**9,947750882**
56	Formas_N_3M = > 4	A_B_3M = > 32, A_B_Pre = > 32	0,168	0,429	1,698	1,308	**11,34381323**
27	A_B_Pre = > 32	A_B_3M = > 32, Formas_N_3M = > 4	0,168	0,419	1,6	1,27	**9,147153435**
64	Idade_M = Range 1 <31	Grau_Varicoc = II	0,159	0,436	0,972	0,977	0,039168676
65	HabitosTabagicos_Processado_Simplificado	Grau_Varicoc = II	0,159	0,436	0,972	0,977	0,039168676
66	HabitosTabagicos_Processado_Simplificado	A_B_Pre = > 32	0,159	0,436	1,085	1,06	0,298112905
67	Idade_M = Range 1 <31	Conc_3M = 0.01 to 14.9	0,159	0,436	1,227	1,143	1,744467381
87	Conc_3M = 0.01 to 14.9	A_B_Pre = 1 to 31	0,159	0,447	0,957	0,964	0,095733221
88	Conc_3M = 0.01 to 14.9	Idade_M = Range 1 <31	0,159	0,447	1,227	1,15	1,744428
136	A_B_3M = 1 to 31	Conc_3M = > 15	0,159	0,486	0,962	0,963	0,076706057
356	Conc_6M = > 15	Conc_3M = > 15	0,159	0,708	1,404	1,698	**5,145117602**
357	Gravidez_espontanea, Formas_N_3M = > 4	Conc_3M = > 15	0,159	0,708	1,404	1,698	**5,145117602**
178	Conc_3M = > 15, Formas_N_3M = > 4	Gravidez_espontanea	0,159	0,515	1,125	1,118	0,630417592
337	PMA, A_B_3M = 1 to 31	A_B_Pre = 1 to 31	0,159	0,68	1,455	1,665	**5,931618107**
194	Formas_N_3M = 1 to 3	A_B_3M = > 32	0,159	0,531	1,137	1,136	0,752143996
280	Idade_M = Range 2 31 to 32	PMA	0,159	0,607	0,984	0,975	0,015652734
281	Conc_3M = > 15, Gravidez_espontanea	Formas_N_3M = > 4	0,159	0,607	1,547	1,546	**7,337335508**

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
336	Conc_6M = 0.01 to 14.9	PMA	0,159	0,68	1,102	1,197	0,547444358
338	A_B_Pre = 1 to 31, A_B_3M = 1 to 31	PMA	0,159	0,68	1,102	1,197	0,547444358
177	PMA, A_B_Pre = 1 to 31	A_B_3M = 1 to 31	0,159	0,515	1,575	1,388	**7,676583129**
355	Conc_6M = > 15	PMA	0,159	0,708	1,148	1,314	1,092226148
137	A_B_3M = 1 to 31	PMA, A_B_Pre = 1 to 31	0,159	0,486	1,575	1,345	**7,676747005**
7	Formas_N_3M = > 4	Conc_3M = > 15, Gravidez_espontanea	0,159	0,405	1,547	1,24	**7,338534949**
431	Conc_3M = > 15, HabitosTabagicos_Processado_Simplificado	PMA	0,15	0,889	1,441	3,449	**6,802529918**
19	Idade_M = Range 1 <31	A_B_Pre = > 32	0,15	0,41	1,021	1,014	0,01826766
20	Idade_M = Range 1 <31	HabitosTabagicos_Processado_Simplificado	0,15	0,41	1,126	1,078	0,561194768
21	HabitosTabagicos_Processado_Simplificado	Idade_M = Range 1 <31	0,15	0,41	1,126	1,078	0,561194768
22	HabitosTabagicos_Processado_Simplificado	Conc_3M = 0.01 to 14.9	0,15	0,41	1,155	1,093	0,81619263
432	A_B_3M = > 32, A_B_Pre = > 32, Formas_N_3M = > 4	Conc_3M = > 15	0,15	0,889	1,761	4,458	**12,82286848**
32	Conc_3M = 0.01 to 14.9	Grau_Varicoc = II	0,15	0,421	0,939	0,952	0,179109141
33	Conc_3M = 0.01 to 14.9	HabitosTabagicos_Processado_Simplificado	0,15	0,421	1,155	1,098	0,816122175
418	Conc_3M = > 15, A_B_Pre = > 32, Formas_N_3M = > 4	A_B_3M = > 32	0,15	0,842	1,802	3,374	**13,08454235**
384	Gravidez_espontanea, A_B_Pre = > 32	A_B_3M = > 32	0,15	0,762	1,63	2,237	**9,137764469**
385	Conc_3M = > 15, A_B_3M = > 32, A_B_Pre = > 32	Formas_N_3M = > 4	0,15	0,762	1,941	2,551	**15,00873286**
386	A_B_Pre = > 32, Formas_N_3M = > 4	Conc_3M = > 15, A_B_3M = > 32	0,15	0,762	2,265	2,787	**21,27652296**
134	Conc_3M = > 15, Formas_N_3M = > 4	Grau_Varicoc = II	0,15	0,485	1,081	1,07	0,255797176
366	Grau_Varicoc = II, Formas_N_3M = > 4	Conc_3M = > 15	0,15	0,727	1,441	1,816	**5,50823679**
195	PMA, HabitosTabagicos_Processado_Simplificado	Conc_3M = > 15	0,15	0,533	1,057	1,061	0,13849111
332	Conc_3M = > 15, A_B_Pre = > 32	A_B_3M = > 32, Formas_N_3M = > 4	0,15	0,667	2,548	2,215	**26,3793108**
333	Conc_3M = > 15, A_B_3M = > 32, Formas_N_3M = > 4	A_B_Pre = > 32	0,15	0,667	1,659	1,794	**9,065034987**
244	Conc_3M = > 15, Gravidez_espontanea	Grau_Varicoc = II	0,15	0,571	1,274	1,287	2,324742471
296	Conc_3M = > 15, Grau_Varicoc = II	Gravidez_espontanea	0,15	0,615	1,344	1,409	**3,445772149**
297	Conc_3M = > 15, Grau_Varicoc = II	Formas_N_3M = > 4	0,15	0,615	1,568	1,579	**7,186227641**
273	Gravidez_espontanea, Grau_Varicoc = II	Conc_3M = > 15	0,15	0,593	1,174	1,216	1,119562564
274	A_B_3M = > 32, A_B_Pre = > 32	Gravidez_espontanea	0,15	0,593	1,294	1,331	2,649126193
275	A_B_3M = > 32, A_B_Pre = > 32	Conc_3M = > 15, Formas_N_3M = > 4	0,15	0,593	1,921	1,698	**13,72294078**
246	A_B_3M = > 32, Gravidez_espontanea	A_B_Pre = > 32	0,15	0,571	1,422	1,396	**4,555377695**
252	A_B_3M = > 32, Formas_N_3M = > 4	Conc_3M = > 15, A_B_Pre = > 32	0,15	0,571	2,548	1,81	**26,38543707**
364	ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	PMA	0,15	0,727	1,179	1,405	1,433511765

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
133	PMA, Conc_3M = > 15	HabitosTabagicos_Processado_Simplificado	0,15	0,485	1,33	1,234	**2,99462978**
135	Conc_3M = > 15, Formas_N_3M = > 4	A_B_3M = > 32, A_B_Pre = > 32	0,15	0,485	1,921	1,451	**13,72577226**
85	Conc_3M = > 15, A_B_3M = > 32	A_B_Pre = > 32, Formas_N_3M = > 4	0,15	0,444	2,265	1,447	**21,30010683**
23	HabitosTabagicos_Processado_Simplificado	PMA, Conc_3M = > 15	0,15	0,41	1,33	1,173	**2,995888169**
26	Conc_3M = > 15, A_B_3M = > 32	Grau_Varicoc = II	0,14	0,417	0,929	0,945	0,222031624
49	A_B_3M = 1 to 31	Grau_Varicoc = II	0,14	0,429	0,955	0,965	0,08560725
406	Conc_3M = > 15, Formas_N_3M = 1 to 3	A_B_Pre = 1 to 31	0,14	0,789	1,689	2,53	**9,605944081**
158	PMA, HabitosTabagicos_Processado_Simplificado	A_B_3M = > 32	0,14	0,5	1,07	1,065	0,178854776
179	Grau_Varicoc = I	Conc_3M = > 15	0,14	0,517	1,025	1,026	0,025274194
339	A_B_Pre = 1 to 31, Formas_N_3M = 1 to 3	Conc_3M = > 15	0,14	0,682	1,351	1,557	**3,471242997**
196	Idade_M = Range 2 31 to 32	Conc_3M = > 15	0,14	0,536	1,062	1,067	0,148176404
199	Conc_3M = > 15, A_B_Pre = 1 to 31	Grau_Varicoc = II	0,14	0,536	1,194	1,188	1,159735586
225	A_B_3M = > 32, A_B_Pre = > 32	PMA	0,14	0,556	0,901	0,862	0,568780329
258	PMA, A_B_Pre = > 32	A_B_3M = > 32	0,14	0,577	1,235	1,259	1,66003032
259	Conc_3M = > 15, Grau_Varicoc = II	A_B_Pre = 1 to 31	0,14	0,577	1,235	1,259	1,66003032
260	Conc_3M = > 15, Grau_Varicoc = II	A_B_3M = > 32	0,14	0,577	1,235	1,259	1,66003032
320	A_B_Pre = 1 to 31, Grau_Varicoc = II	Conc_3M = > 15	0,14	0,652	1,292	1,424	2,54141259
321	A_B_3M = > 32, Grau_Varicoc = II	Conc_3M = > 15	0,14	0,652	1,292	1,424	2,54141259
197	PMA, A_B_3M = > 32	A_B_Pre = > 32	0,14	0,536	1,333	1,288	**2,821055061**
198	PMA, A_B_3M = > 32	HabitosTabagicos_Processado_Simplificado	0,14	0,536	1,47	1,369	**4,795459823**
200	Conc_3M = > 15, A_B_Pre = 1 to 31	Formas_N_3M = 1 to 3	0,14	0,536	1,791	1,51	**10,10859177**
378	A_B_3M = > 32, HabitosTabagicos_Processado_Simplificado	PMA	0,14	0,75	1,216	1,533	1,844016605
116	Formas_N_3M = 1 to 3	A_B_3M = 1 to 31	0,14	0,469	1,433	1,267	**4,153244675**
117	Formas_N_3M = 1 to 3	Conc_3M = > 15, A_B_Pre = 1 to 31	0,14	0,469	1,791	1,39	**10,10672084**
50	A_B_3M = 1 to 31	Formas_N_3M = 1 to 3	0,14	0,429	1,433	1,227	**4,152535393**
42	PMA, Conc_3M = > 15	Grau_Varicoc = II	0,131	0,424	0,946	0,958	0,113310544
43	PMA, Conc_3M = > 15	A_B_Pre = > 32	0,131	0,424	1,056	1,039	0,100649654
45	PMA, A_B_Pre = 1 to 31	Grau_Varicoc = II	0,131	0,424	0,946	0,958	0,113310544
415	A_B_Pre = > 32, HabitosTabagicos_Processado_Simplificado	PMA	0,131	0,824	1,335	2,171	**3,660309124**
416	Conc_3M = > 15, A_B_3M = 1 to 31	A_B_Pre = 1 to 31	0,131	0,824	1,762	3,019	**10,31706472**
417	Conc_3M = > 15, Gravidez_espontanea, Formas_N_3M = > 4	A_B_3M = > 32	0,131	0,824	1,762	3,019	**10,31706472**
392	A_B_Pre = 1 to 31, A_B_3M = > 32	Conc_3M = > 15	0,131	0,778	1,541	2,229	**6,465474184**

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
115	PMA, HabitosTabagicos_Processado_Simplificado	A_B_Pre = > 32	0,131	0,467	1,161	1,121	0,727654062
394	Conc_3M = > 15, Idade_M = Range 1 <31	A_B_3M = > 32	0,131	0,778	1,664	2,397	**8,387505665**
395	A_B_3M = > 32, Gravidez_espontanea, Formas_N_3M = > 4	Conc_3M = > 15	0,131	0,778	1,541	2,229	**6,465474184**
132	Grau_Varicoc = I	A_B_Pre = 1 to 31	0,131	0,483	1,033	1,03	0,038082437
146	Idade_M = Range 2 31 to 32	A_B_3M = > 32	0,131	0,5	1,07	1,065	0,163275092
147	Idade_M = Range 2 31 to 32	Gravidez_espontanea	0,131	0,5	1,092	1,084	0,271551783
151	Idade_M = Range 2 31 to 32	A_B_Pre = > 32	0,131	0,5	1,244	1,196	1,519866451
156	PMA, A_B_3M = > 32	Formas_N_3M = > 4	0,131	0,5	1,274	1,215	1,84229374
157	A_B_3M = > 32, Formas_N_3M = > 4	PMA	0,131	0,5	0,811	0,766	2,181533723
160	Conc_3M = > 15, A_B_Pre = 1 to 31	A_B_3M = > 32	0,131	0,5	1,07	1,065	0,163275092
161	Conc_3M = > 15, A_B_Pre = 1 to 31	Gravidez_espontanea	0,131	0,5	1,092	1,084	0,271551783
162	Conc_3M = > 15, Gravidez_espontanea	A_B_Pre = 1 to 31	0,131	0,5	1,07	1,065	0,163275092
164	A_B_3M = > 32, Formas_N_3M = > 4	Grau_Varicoc = II	0,131	0,5	1,115	1,103	0,408431544
350	Conc_3M = > 15, A_B_3M = > 32, Gravidez_espontanea	Formas_N_3M = > 4	0,131	0,7	1,783	2,025	**9,761935065**
182	Gravidez_espontanea, Grau_Varicoc = II	Formas_N_3M = > 4	0,131	0,519	1,321	1,262	2,408942324
202	PMA, Grau_Varicoc = II	Conc_3M = > 15	0,131	0,538	1,067	1,073	0,157230528
203	Conc_3M = > 15, Grau_Varicoc = II	PMA	0,131	0,538	0,873	0,83	0,89208364
204	PMA, A_B_Pre = > 32	Conc_3M = > 15	0,131	0,538	1,067	1,073	0,157230528
205	PMA, Grau_Varicoc = II	A_B_Pre = 1 to 31	0,131	0,538	1,152	1,154	0,697207009
326	A_B_Pre = 1 to 31, Gravidez_espontanea	Conc_3M = > 15	0,131	0,667	1,321	1,486	**2,748200625**
328	A_B_3M = > 32, Idade_M = Range 1 <31	Conc_3M = > 15	0,131	0,667	1,321	1,486	**2,748200625**
233	PMA, Formas_N_3M = 1 to 3	Conc_3M = > 15	0,131	0,56	1,11	1,126	0,402539487
235	A_B_Pre = 1 to 31, A_B_3M = 1 to 31	Conc_3M = > 15	0,131	0,56	1,11	1,126	0,402539487
304	Idade_M = Range 3 33 to 35	A_B_Pre = 1 to 31	0,131	0,636	1,362	1,465	**3,186407615**
306	PMA, Formas_N_3M = > 4	A_B_3M = > 32	0,131	0,636	1,362	1,465	**3,186407615**
307	Grau_Varicoc = II, Formas_N_3M = > 4	A_B_3M = > 32	0,131	0,636	1,362	1,465	**3,186407615**
308	Grau_Varicoc = II, Formas_N_3M = > 4	Gravidez_espontanea	0,131	0,636	1,39	1,491	**3,561054981**
267	Conc_3M = > 15, A_B_Pre = > 32	PMA	0,131	0,583	0,946	0,92	0,145233261
269	Gravidez_espontanea, Formas_N_3M = > 4	Grau_Varicoc = II	0,131	0,583	1,3	1,323	2,269386918
270	Conc_3M = > 15, A_B_3M = > 32, Formas_N_3M = > 4	Gravidez_espontanea	0,131	0,583	1,274	1,301	1,964302058
283	A_B_3M = > 32, Grau_Varicoc = II	Formas_N_3M = > 4	0,131	0,609	1,551	1,552	**5,755681642**
282	A_B_Pre = 1 to 31, Grau_Varicoc = II	PMA	0,131	0,609	0,987	0,979	0,007984346

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
271	Gravidez_espontanea, Formas_N_3M = > 4	Conc_3M = > 15, A_B_3M = > 32	0,131	0,583	1,734	1,593	**8,462564455**
303	Idade_M = Range 3 33 to 35	PMA	0,131	0,636	1,032	1,054	0,045648402
232	Conc_6M = 0.01 to 14.9	Conc_3M = 0.01 to 14.9	0,131	0,56	1,577	1,466	**5,989857376**
234	PMA, A_B_3M = 1 to 31	Conc_3M = 0.01 to 14.9	0,131	0,56	1,577	1,466	**5,989857376**
206	PMA, A_B_Pre = > 32	HabitosTabagicos_Processado_Simplificado	0,131	0,538	1,477	1,377	**4,489666743**
207	PMA, Conc_3M = 0.01 to 14.9	A_B_3M = 1 to 31	0,131	0,538	1,646	1,458	**6,978598542**
163	Conc_3M = > 15, A_B_Pre = 1 to 31	A_B_3M = 1 to 31	0,131	0,5	1,529	1,346	**5,165289658**
169	Conc_3M = > 15, Gravidez_espontanea	A_B_3M = > 32, Formas_N_3M = > 4	0,131	0,5	1,911	1,477	**11,17142185**
170	A_B_3M = > 32, Gravidez_espontanea	Conc_3M = > 15, Formas_N_3M = > 4	0,131	0,5	1,621	1,383	**6,53395741**
171	A_B_3M = > 32, Formas_N_3M = > 4	Conc_3M = > 15, Gravidez_espontanea	0,131	0,5	1,911	1,477	**11,17142185**
376	Conc_3M = > 15, Formas_N_3M = 1 to 3	PMA	0,131	0,737	1,195	1,456	1,415316865
391	Conc_3M = 0.01 to 14.9, A_B_3M = 1 to 31	PMA	0,131	0,778	1,261	1,724	2,377192074
68	Formas_N_3M = 1 to 3	PMA, Conc_3M = > 15	0,131	0,438	1,419	1,229	**3,578902334**
44	PMA, Conc_3M = > 15	Formas_N_3M = 1 to 3	0,131	0,424	1,419	1,217	**3,578971786**
46	Conc_3M = > 15, Formas_N_3M = > 4	A_B_3M = > 32, Gravidez_espontanea	0,131	0,424	1,621	1,282	**6,534951361**
4	A_B_3M = 1 to 31	PMA, Conc_3M = 0.01 to 14.9	0,131	0,4	1,646	1,262	**6,980873662**
5	A_B_3M = 1 to 31	Conc_3M = > 15, A_B_Pre = 1 to 31	0,131	0,4	1,529	1,231	**5,166306725**
437	PMA, A_B_3M = > 32, Formas_N_3M = > 4	Conc_3M = > 15	0,121	0,929	1,84	6,935	**11,52958556**
8	Formas_N_3M = 1 to 3	Conc_3M = 0.01 to 14.9	0,121	0,406	1,144	1,086	0,518225697
429	A_B_Pre = 1 to 31, HabitosTabagicos_Processado_Simplificado	PMA	0,121	0,867	1,405	2,874	**4,587511741**
63	PMA, HabitosTabagicos_Processado_Simplificado	A_B_Pre = 1 to 31	0,121	0,433	0,927	0,94	0,193830101
89	Grau_Varicoc = I	Formas_N_3M = > 4	0,121	0,448	1,142	1,101	0,515367025
90	Grau_Varicoc = I	A_B_3M = 1 to 31	0,121	0,448	1,37	1,22	2,633734159
109	Idade_M = Range 2 31 to 32	A_B_Pre = 1 to 31	0,121	0,464	0,994	0,994	0,001189651
110	Idade_M = Range 2 31 to 32	HabitosTabagicos_Processado_Simplificado	0,121	0,464	1,274	1,186	1,623338599
112	Conc_3M = > 15, Gravidez_espontanea	A_B_Pre = > 32	0,121	0,464	1,155	1,117	0,608945405
131	PMA, Idade_M = Range 1 <31	Conc_3M = 0.01 to 14.9	0,121	0,481	1,356	1,244	2,50555342
155	PMA, Conc_3M = 0.01 to 14.9	A_B_Pre = 1 to 31	0,121	0,5	1,07	1,065	0,146832153
183	PMA, Formas_N_3M = 1 to 3	A_B_3M = > 32	0,121	0,52	1,113	1,11	0,363330444
341	PMA, Conc_3M = > 15, Formas_N_3M = > 4	A_B_3M = > 32	0,121	0,684	1,464	1,687	**4,341684022**
211	Conc_6M = > 15	A_B_Pre = > 32	0,121	0,542	1,348	1,305	2,504435117
212	Conc_3M = > 15, A_B_Pre = > 32	Gravidez_espontanea	0,121	0,542	1,183	1,183	0,870824198

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
213	Gravidez_espontanea, Formas_N_3M = > 4	A_B_Pre = > 32	0,121	0,542	1,348	1,305	2,504435117
214	Conc_3M = > 15, A_B_3M = > 32, Formas_N_3M = > 4	PMA	0,121	0,542	0,878	0,836	0,738357554
299	Gravidez_espontanea, A_B_Pre = > 32	Formas_N_3M = > 4	0,121	0,619	1,577	1,595	**5,592636226**
300	A_B_Pre = > 32, Formas_N_3M = > 4	Gravidez_espontanea	0,121	0,619	1,352	1,423	**2,720268713**
301	PMA, Conc_3M = > 15, A_B_3M = > 32	Formas_N_3M = > 4	0,121	0,619	1,577	1,595	**5,592636226**
272	PMA, Formas_N_3M = > 4	Conc_3M = > 15, A_B_3M = > 32	0,121	0,591	1,756	1,622	**7,986866708**
298	Gravidez_espontanea, A_B_Pre = > 32	Conc_3M = > 15	0,121	0,619	1,227	1,3	1,363887626
159	PMA, Conc_3M = 0.01 to 14.9	Idade_M = Range 1 <31	0,121	0,5	1,372	1,271	**2,710624362**
387	A_B_Pre = 1 to 31, Conc_3M = 0.01 to 14.9	PMA	0,121	0,765	1,24	1,629	1,86497747
388	A_B_3M = > 32, Formas_N_3M = 1 to 3	PMA	0,121	0,765	1,24	1,629	1,86497747
389	Idade_M = Range 1 <31, Conc_3M = 0.01 to 14.9	PMA	0,121	0,765	1,24	1,629	1,86497747
111	PMA, A_B_3M = > 32	Formas_N_3M = 1 to 3	0,121	0,464	1,552	1,308	**4,905036784**
113	PMA, A_B_3M = > 32	Conc_3M = > 15, Formas_N_3M = > 4	0,121	0,464	1,505	1,291	**4,290669224**
114	A_B_3M = > 32, Formas_N_3M = > 4	PMA, Conc_3M = > 15	0,121	0,464	1,505	1,291	**4,290669224**
9	Formas_N_3M = 1 to 3	PMA, A_B_3M = > 32	0,121	0,406	1,552	1,243	**4,90392545**
2	PMA, HabitosTabagicos_Processado_Simplificado	Grau_Varicoc = II	0,112	0,4	0,892	0,919	0,394595122
3	PMA, HabitosTabagicos_Processado_Simplificado	Formas_N_3M = > 4	0,112	0,4	1,019	1,012	0,009707018
435	Conc_3M = > 15, Gravidez_espontanea, A_B_Pre = > 32	A_B_3M = > 32	0,112	0,923	1,975	6,925	**12,32469381**
436	Gravidez_espontanea, A_B_Pre = > 32, Formas_N_3M = > 4	A_B_3M = > 32	0,112	0,923	1,975	6,925	**12,32469381**
24	Grau_Varicoc = I	A_B_3M = > 32	0,112	0,414	0,886	0,909	0,452337822
25	Grau_Varicoc = I	Conc_3M = 0.01 to 14.9	0,112	0,414	1,165	1,1	0,595556748
424	A_B_3M = > 32, Idade_M = Range 2 31 to 32	Conc_3M = > 15	0,112	0,857	1,698	3,467	**7,986218699**
426	Formas_N_3M = > 4, Idade_M = Range 1 <31	A_B_3M = > 32	0,112	0,857	1,834	3,729	**9,814403001**
427	PMA, Conc_3M = > 15, A_B_Pre = > 32	A_B_3M = > 32	0,112	0,857	1,834	3,729	**9,814403001**
51	PMA, A_B_3M = > 32	A_B_Pre = 1 to 31	0,112	0,429	0,917	0,932	0,228947652
52	PMA, A_B_3M = > 32	Idade_M = Range 1 <31	0,112	0,429	1,176	1,112	0,672519009
53	Conc_3M = > 15, A_B_Pre = 1 to 31	Formas_N_3M = > 4	0,112	0,429	1,092	1,063	0,20704367
54	Idade_M = Range 2 31 to 32	Conc_3M = > 15, A_B_3M = > 32	0,112	0,429	1,274	1,161	1,440934524
57	A_B_3M = > 32, Formas_N_3M = > 4	Idade_M = Range 1 <31	0,112	0,429	1,176	1,112	0,672519009
410	Conc_3M = > 15, Idade_M = Range 2 31 to 32	A_B_3M = > 32	0,112	0,8	1,712	2,664	**7,745839576**
411	Grau_Varicoc = II, A_B_3M = 1 to 31	A_B_Pre = 1 to 31	0,112	0,8	1,712	2,664	**7,745839576**
78	PMA, Idade_M = Range 1 <31	A_B_3M = > 32	0,112	0,444	0,951	0,959	0,075898097

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
79	PMA, Idade_M = Range 1 <31	A_B_Pre = > 32	0,112	0,444	1,106	1,077	0,272019744
412	A_B_3M = 1 to 31, Formas_N_3M = 1 to 3	A_B_Pre = 1 to 31	0,112	0,8	1,712	2,664	**7,745839576**
81	Gravidez_espontanea, Grau_Varicoc = II	A_B_Pre = 1 to 31	0,112	0,444	0,951	0,959	0,075898097
82	A_B_3M = > 32, A_B_Pre = > 32	Grau_Varicoc = II	0,112	0,444	0,991	0,993	0,002373255
413	PMA, A_B_3M = > 32, A_B_Pre = > 32	Conc_3M = > 15	0,112	0,8	1,585	2,477	**6,074988002**
105	PMA, Grau_Varicoc = II	HabitosTabagicos_Processado_Simplificado	0,112	0,462	1,266	1,18	1,392140141
106	PMA, A_B_Pre = > 32	Formas_N_3M = > 4	0,112	0,462	1,176	1,128	0,686282391
107	PMA, A_B_Pre = > 32	Idade_M = Range 1 <31	0,112	0,462	1,266	1,18	1,392140141
108	PMA, A_B_Pre = > 32	Conc_3M = > 15, A_B_3M = > 32	0,112	0,462	1,372	1,232	2,405582297
124	Conc_6M = 0.01 to 14.9	A_B_Pre = 1 to 31	0,112	0,48	1,027	1,024	0,020832214
380	A_B_Pre = 1 to 31, Formas_N_3M = > 4	Conc_3M = > 15	0,112	0,75	1,486	1,981	**4,521027089**
128	A_B_Pre = 1 to 31, A_B_3M = 1 to 31	Grau_Varicoc = II	0,112	0,48	1,07	1,06	0,129819307
382	A_B_3M = > 32, Gravidez_espontanea, A_B_Pre = > 32	Conc_3M = > 15	0,112	0,75	1,486	1,981	**4,521027089**
383	A_B_3M = > 32, Gravidez_espontanea, A_B_Pre = > 32	Formas_N_3M = > 4	0,112	0,75	1,911	2,43	**10,07040459**
148	Conc_6M = > 15	Grau_Varicoc = II	0,112	0,5	1,115	1,103	0,332093705
351	PMA, Conc_6M = > 15	Conc_3M = > 15	0,112	0,706	1,399	1,684	**3,272145548**
354	Grau_Varicoc = II, A_B_Pre = > 32	A_B_3M = > 32	0,112	0,706	1,511	1,811	**4,620255326**
184	A_B_Pre = 1 to 31, Grau_Varicoc = II	Gravidez_espontanea	0,112	0,522	1,139	1,133	0,477785409
186	A_B_3M = > 32, Grau_Varicoc = II	A_B_Pre = > 32	0,112	0,522	1,298	1,251	1,746060363
218	PMA, Formas_N_3M = > 4	A_B_Pre = > 32	0,112	0,545	1,357	1,316	2,367504479
329	Conc_3M = > 15, HabitosTabagicos_Processado_Simplificado	A_B_3M = > 32	0,112	0,667	1,427	1,598	**3,455228784**
334	A_B_3M = > 32, Gravidez_espontanea, Formas_N_3M = > 4	A_B_Pre = > 32	0,112	0,667	1,659	1,794	**6,305126557**
335	A_B_3M = > 32, A_B_Pre = > 32, Formas_N_3M = > 4	Gravidez_espontanea	0,112	0,667	1,456	1,626	**3,795660263**
241	A_B_3M = > 32, Idade_M = Range 1 <31	PMA	0,112	0,571	0,926	0,894	0,229964241
242	A_B_Pre = > 32, Formas_N_3M = > 4	PMA	0,112	0,571	0,926	0,894	0,229964241
245	A_B_Pre = 1 to 31, Gravidez_espontanea	Grau_Varicoc = II	0,112	0,571	1,274	1,287	1,592102498
249	Conc_3M = > 15, A_B_3M = > 32, A_B_Pre = > 32	PMA	0,112	0,571	0,926	0,894	0,229964241
250	Conc_3M = > 15, A_B_3M = > 32, A_B_Pre = > 32	Gravidez_espontanea	0,112	0,571	1,248	1,265	1,354378079
277	A_B_3M = > 32, HabitosTabagicos_Processado_Simplificado	Conc_3M = > 15	0,112	0,6	1,189	1,238	0,893596671
278	Conc_3M = > 15, A_B_3M = > 32, Gravidez_espontanea	A_B_Pre = > 32	0,112	0,6	1,493	1,495	**4,010288422**
247	A_B_3M = > 32, Idade_M = Range 1 <31	Formas_N_3M = > 4	0,112	0,571	1,456	1,417	**3,502770131**
248	PMA, Conc_3M = > 15, A_B_3M = > 32	A_B_Pre = > 32	0,112	0,571	1,422	1,396	**3,119755542**

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
251	Gravidez_espontanea, A_B_Pre = > 32	Conc_3M = > 15, A_B_3M = > 32	0,112	0,571	1,698	1,548	**6,444840182**
253	Gravidez_espontanea, A_B_Pre = > 32	A_B_3M = > 32, Formas_N_3M = > 4	0,112	0,571	2,184	1,723	**12,95669574**
254	A_B_Pre = > 32, Formas_N_3M = > 4	A_B_3M = > 32, Gravidez_espontanea	0,112	0,571	2,184	1,723	**12,95669574**
325	A_B_Pre = 1 to 31, A_B_3M = > 32	PMA	0,112	0,667	1,081	1,15	0,228246616
219	PMA, Formas_N_3M = > 4	HabitosTabagicos_Processado_Simplificado	0,112	0,545	1,497	1,398	**3,913688313**
220	A_B_Pre = 1 to 31, Formas_N_3M = 1 to 3	A_B_3M = 1 to 31	0,112	0,545	1,668	1,48	**5,993542303**
185	A_B_Pre = 1 to 31, Grau_Varicoc = II	A_B_3M = 1 to 31	0,112	0,522	1,595	1,407	**5,034109479**
154	Conc_6M = > 15	PMA, Conc_3M = > 15	0,112	0,5	1,621	1,383	**5,312729043**
352	Conc_3M = > 15, Conc_6M = > 15	PMA	0,112	0,706	1,144	1,303	0,674327592
353	Grau_Varicoc = II, HabitosTabagicos_Processado_Simplificado	PMA	0,112	0,706	1,144	1,303	0,674327592
166	Conc_3M = > 15, A_B_Pre = > 32	PMA, A_B_3M = > 32	0,112	0,5	1,911	1,477	**9,083428862**
168	Conc_3M = > 15, A_B_Pre = > 32	A_B_3M = > 32, Gravidez_espontanea	0,112	0,5	1,911	1,477	**9,083428862**
174	Gravidez_espontanea, Formas_N_3M = > 4	A_B_3M = > 32, A_B_Pre = > 32	0,112	0,5	1,981	1,495	**10,03513075**
379	A_B_Pre = > 32, Idade_M = Range 1 <31	PMA	0,112	0,75	1,216	1,533	1,410470381
125	PMA, A_B_3M = 1 to 31	Formas_N_3M = 1 to 3	0,112	0,48	1,605	1,348	**5,08573171**
126	PMA, Formas_N_3M = 1 to 3	A_B_3M = 1 to 31	0,112	0,48	1,467	1,294	**3,453916448**
129	A_B_Pre = 1 to 31, A_B_3M = 1 to 31	Formas_N_3M = 1 to 3	0,112	0,48	1,605	1,348	**5,08573171**
83	A_B_3M = > 32, A_B_Pre = > 32	PMA, Conc_3M = > 15	0,112	0,444	1,441	1,245	**3,126286012**
84	A_B_3M = > 32, A_B_Pre = > 32	Conc_3M = > 15, Gravidez_espontanea	0,112	0,444	1,698	1,329	**6,226729441**
86	A_B_3M = > 32, A_B_Pre = > 32	Gravidez_espontanea, Formas_N_3M = > 4	0,112	0,444	1,981	1,396	**10,03484882**
408	Formas_N_3M = > 4, HabitosTabagicos_Processado_Simplificado	PMA	0,112	0,8	1,297	1,916	2,473203538
409	A_B_3M = 1 to 31, Formas_N_3M = 1 to 3	PMA	0,112	0,8	1,297	1,916	2,473203538
58	PMA, A_B_3M = > 32	Conc_3M = > 15, A_B_Pre = > 32	0,112	0,429	1,911	1,357	**9,08214394**
59	Conc_3M = > 15, Gravidez_espontanea	A_B_3M = > 32, A_B_Pre = > 32	0,112	0,429	1,698	1,308	**6,226572561**
60	A_B_3M = > 32, Gravidez_espontanea	Conc_3M = > 15, A_B_Pre = > 32	0,112	0,429	1,911	1,357	**9,08214394**
61	A_B_3M = > 32, Gravidez_espontanea	A_B_Pre = > 32, Formas_N_3M = > 4	0,112	0,429	2,184	1,407	**12,95466284**
62	A_B_3M = > 32, Formas_N_3M = > 4	Gravidez_espontanea, A_B_Pre = > 32	0,112	0,429	2,184	1,407	**12,95466284**
434	PMA, A_B_Pre = > 32, Formas_N_3M = > 4	Conc_3M = > 15	0,103	0,917	1,816	5,944	**9,195737362**
10	PMA, Idade_M = Range 1 <31	Conc_3M = > 15	0,103	0,407	0,807	0,836	1,374030758
11	PMA, Idade_M = Range 1 <31	A_B_Pre = 1 to 31	0,103	0,407	0,872	0,899	0,519886828
12	PMA, Idade_M = Range 1 <31	HabitosTabagicos_Processado_Simplificado	0,103	0,407	1,118	1,072	0,288959011

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
13	A_B_3M = > 32, A_B_Pre = > 32	Idade_M = Range 1 <31	0,103	0,407	1,118	1,072	0,288959011
14	A_B_3M = > 32, A_B_Pre = > 32	HabitosTabagicos_Processado_Simplificado	0,103	0,407	1,118	1,072	0,288959011
15	Gravidez_espontanea, Grau_Varicoc = II	Conc_3M = > 15, A_B_3M = > 32	0,103	0,407	1,211	1,12	0,817053867
34	PMA, Grau_Varicoc = II	A_B_3M = 1 to 31	0,103	0,423	1,293	1,166	1,437564632
35	PMA, Conc_3M = 0.01 to 14.9	HabitosTabagicos_Processado_Simplificado	0,103	0,423	1,161	1,102	0,511689826
36	PMA, Conc_3M = 0.01 to 14.9	Formas_N_3M = 1 to 3	0,103	0,423	1,415	1,215	2,529274193
37	Conc_3M = > 15, Grau_Varicoc = II	A_B_3M = 1 to 31	0,103	0,423	1,293	1,166	1,437564632
38	PMA, A_B_Pre = > 32	Conc_3M = > 15, Formas_N_3M = > 4	0,103	0,423	1,372	1,199	2,124374398
419	Conc_3M = 0.01 to 14.9, Formas_N_3M = 1 to 3	PMA	0,103	0,846	1,372	2,491	**3,301434515**
420	Formas_N_3M = > 4, Grau_Varicoc = I	Conc_3M = > 15	0,103	0,846	1,677	3,22	**6,921159619**
421	Conc_3M = > 15, Gravidez_espontanea, A_B_Pre = > 32	Formas_N_3M = > 4	0,103	0,846	2,156	3,949	**12,80108921**
422	Gravidez_espontanea, A_B_Pre = > 32, Formas_N_3M = > 4	Conc_3M = > 15	0,103	0,846	1,677	3,22	**6,921159619**
70	Conc_6M = 0.01 to 14.9	HabitosTabagicos_Processado_Simplificado	0,103	0,44	1,207	1,135	0,803876078
71	PMA, A_B_3M = 1 to 31	Conc_3M = > 15	0,103	0,44	0,872	0,885	0,545732692
72	PMA, A_B_3M = 1 to 31	Grau_Varicoc = II	0,103	0,44	0,981	0,985	0,009601823
73	PMA, Formas_N_3M = 1 to 3	Conc_3M = 0.01 to 14.9	0,103	0,44	1,239	1,152	1,028709619
74	A_B_Pre = 1 to 31, A_B_3M = 1 to 31	Conc_3M = 0.01 to 14.9	0,103	0,44	1,239	1,152	1,028709619
400	Formas_N_3M = > 4, Idade_M = Range 1 <31	Conc_3M = > 15	0,103	0,786	1,557	2,312	**5,103626186**
401	PMA, Conc_3M = > 15, Formas_N_3M = 1 to 3	A_B_Pre = 1 to 31	0,103	0,786	1,681	2,486	**6,571946711**
402	PMA, Conc_3M = > 15, A_B_Pre = > 32	Formas_N_3M = > 4	0,103	0,786	2,002	2,835	**10,47188135**
94	Conc_6M = > 15	Gravidez_espontanea	0,103	0,458	1,001	1,001	2,61853E-05
403	A_B_3M = > 32, Grau_Varicoc = II, Formas_N_3M = > 4	Conc_3M = > 15	0,103	0,786	1,557	2,312	**5,103626186**
97	Conc_3M = > 15, A_B_3M = > 32, Formas_N_3M = > 4	Grau_Varicoc = II	0,103	0,458	1,022	1,018	0,012201814
404	A_B_3M = > 32, Grau_Varicoc = II, Formas_N_3M = > 4	Gravidez_espontanea	0,103	0,786	1,716	2,53	**6,991428826**
405	Gravidez_espontanea, Grau_Varicoc = II, Formas_N_3M = > 4	A_B_3M = > 32	0,103	0,786	1,681	2,486	**6,571946711**
369	Grau_Varicoc = II, A_B_3M = 1 to 31	Conc_3M = > 15	0,103	0,733	1,453	1,857	**3,654671405**
370	Formas_N_3M = > 4, HabitosTabagicos_Processado_Simplificado	Conc_3M = > 15	0,103	0,733	1,453	1,857	**3,654671405**
371	Conc_3M = > 15, Grau_Varicoc = I	Formas_N_3M = > 4	0,103	0,733	1,868	2,278	**8,51192947**
372	Gravidez_espontanea, Idade_M = Range 1 <31	A_B_3M = > 32	0,103	0,733	1,569	1,998	**4,965953892**
374	Conc_3M = > 15, A_B_3M = > 32, Grau_Varicoc = II	Gravidez_espontanea	0,103	0,733	1,601	2,033	**5,335973283**
144	Idade_M = Range 3 33 to 35	Conc_3M = > 15	0,103	0,5	0,991	0,991	0,002289834

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
375	Conc_3M = > 15, A_B_3M = > 32, Grau_Varicoc = II	Formas_N_3M = > 4	0,103	0,733	1,868	2,278	**8,51192947**
149	Idade_M = Range 3 33 to 35	Grau_Varicoc = II	0,103	0,5	1,115	1,103	0,298484006
152	Idade_M = Range 3 33 to 35	Formas_N_3M = > 4	0,103	0,5	1,274	1,215	1,34635834
153	ProfissãoComRiscoDeContactoDeProdutosOuAmbientesToxicos	HabitosTabagicos_Processado_Simplificado	0,103	0,5	1,372	1,271	2,202773762
345	A_B_Pre = > 32, Idade_M = Range 1 <31	A_B_3M = > 32	0,103	0,688	1,471	1,705	**3,67226173**
346	Conc_3M = > 15, Gravidez_espontanea, Grau_Varicoc = II	A_B_3M = > 32	0,103	0,688	1,471	1,705	**3,67226173**
347	Conc_3M = > 15, Grau_Varicoc = II, Formas_N_3M = > 4	A_B_3M = > 32	0,103	0,688	1,471	1,705	**3,67226173**
189	A_B_3M = > 32, Idade_M = Range 1 <31	Gravidez_espontanea	0,103	0,524	1,144	1,138	0,458778955
190	A_B_3M = > 32, Idade_M = Range 1 <31	A_B_Pre = > 32	0,103	0,524	1,303	1,256	1,616657394
221	PMA, Grau_Varicoc = I	A_B_Pre = 1 to 31	0,103	0,55	1,177	1,184	0,677571547
222	A_B_3M = > 32, HabitosTabagicos_Processado_Simplificado	A_B_Pre = > 32	0,103	0,55	1,369	1,329	2,254472797
223	Conc_3M = > 15, A_B_3M = > 32, Gravidez_espontanea	Grau_Varicoc = II	0,103	0,55	1,226	1,225	1,024581125
316	Conc_3M = > 15, A_B_3M = 1 to 31	Grau_Varicoc = II	0,103	0,647	1,442	1,562	**3,221096926**
317	A_B_Pre = 1 to 31, Conc_3M = 0.01 to 14.9	A_B_3M = 1 to 31	0,103	0,647	1,978	1,907	**9,419460678**
319	Conc_3M = > 15, Gravidez_espontanea, Formas_N_3M = > 4	A_B_Pre = > 32	0,103	0,647	1,61	1,695	**5,064770445**
261	Conc_3M = > 15, Formas_N_3M = 1 to 3	A_B_3M = > 32	0,103	0,579	1,239	1,265	1,160231039
288	Conc_3M = > 15, Idade_M = Range 1 <31	Formas_N_3M = > 4	0,103	0,611	1,557	1,562	**4,347281413**
264	Conc_3M = > 15, A_B_Pre = > 32, Formas_N_3M = > 4	PMA	0,103	0,579	0,939	0,91	0,138563799
265	Conc_3M = > 15, A_B_Pre = > 32, Formas_N_3M = > 4	Gravidez_espontanea	0,103	0,579	1,264	1,287	1,363985946
289	Conc_3M = > 15, HabitosTabagicos_Processado_Simplificado	Formas_N_3M = > 4	0,103	0,611	1,557	1,562	**4,347281413**
291	PMA, Conc_3M = > 15, A_B_Pre = 1 to 31	Formas_N_3M = 1 to 3	0,103	0,611	2,043	1,802	**10,06987546**
294	A_B_3M = > 32, Gravidez_espontanea, Grau_Varicoc = II	Formas_N_3M = > 4	0,103	0,611	1,557	1,562	**4,347281413**
286	Idade_M = Range 4 <36	Gravidez_espontanea	0,103	0,611	1,334	1,394	2,045281719
287	Conc_3M = > 15, Idade_M = Range 1 <31	PMA	0,103	0,611	0,991	0,985	0,00282553
290	Conc_3M = 0.01 to 14.9, A_B_3M = 1 to 31	A_B_Pre = 1 to 31	0,103	0,611	1,308	1,37	1,804127204
292	PMA, A_B_Pre = 1 to 31, Formas_N_3M = 1 to 3	Conc_3M = > 15	0,103	0,611	1,211	1,274	0,983585597
293	A_B_3M = > 32, Gravidez_espontanea, Grau_Varicoc = II	Conc_3M = > 15	0,103	0,611	1,211	1,274	0,983585597
262	Conc_3M = > 15, Formas_N_3M = 1 to 3	PMA, A_B_Pre = 1 to 31	0,103	0,579	1,877	1,643	**7,943595254**
295	A_B_3M = > 32, Gravidez_espontanea, Formas_N_3M = > 4	Grau_Varicoc = II	0,103	0,611	1,362	1,418	2,313000382
263	PMA, Conc_3M = > 15, Formas_N_3M = > 4	A_B_Pre = > 32	0,103	0,579	1,441	1,421	**3,024562244**
314	Conc_3M = > 15, A_B_3M = 1 to 31	PMA	0,103	0,647	1,049	1,086	0,078287512
315	A_B_3M = > 32, Formas_N_3M = 1 to 3	Conc_3M = > 15	0,103	0,647	1,282	1,403	1,641537751

No.	Antecedent	Consequent	Support	Confidence	Lift	Conviction	x^2
318	A_B_Pre = > 32, HabitosTabagicos_Processado_Simplificado	A_B_3M = > 32	0,103	0,647	1,385	1,509	2,63264003
191	A_B_Pre = > 32, Formas_N_3M = > 4	PMA, Conc_3M = > 15	0,103	0,524	1,698	1,452	**5,692629799**
192	Gravidez_espontanea, A_B_Pre = > 32	Conc_3M = > 15, Formas_N_3M = > 4	0,103	0,524	1,698	1,452	**5,692629799**
193	A_B_Pre = > 32, Formas_N_3M = > 4	Conc_3M = > 15, Gravidez_espontanea	0,103	0,524	2,002	1,55	**9,318182297**
343	Idade_M = Range 1 <31, HabitosTabagicos_Processado_Simplificado	PMA	0,103	0,688	1,115	1,226	0,401440681
344	HabitosTabagicos_Processado_Simplificado, Conc_3M = 0.01 to 14.9	PMA	0,103	0,688	1,115	1,226	0,401440681
165	A_B_Pre = 1 to 31, Formas_N_3M = 1 to 3	PMA, Conc_3M = > 15	0,103	0,5	1,621	1,383	**4,775051806**
167	PMA, Formas_N_3M = > 4	Conc_3M = > 15, A_B_Pre = > 32	0,103	0,5	2,229	1,551	**12,1257685**
172	Grau_Varicoc = II, Formas_N_3M = > 4	Conc_3M = > 15, A_B_3M = > 32	0,103	0,5	1,486	1,327	**3,325034253**
173	Grau_Varicoc = II, Formas_N_3M = > 4	A_B_3M = > 32, Gravidez_espontanea	0,103	0,5	1,911	1,477	**8,164136178**
367	A_B_Pre = 1 to 31, Idade_M = Range 1 <31	PMA	0,103	0,733	1,189	1,437	1,004484393
368	Grau_Varicoc = II, A_B_3M = 1 to 31	PMA	0,103	0,733	1,189	1,437	1,004484393
373	Conc_3M = > 15, A_B_Pre = 1 to 31, Formas_N_3M = 1 to 3	PMA	0,103	0,733	1,189	1,437	1,004484393
119	A_B_3M = > 32, Grau_Varicoc = II	Conc_3M = > 15, Gravidez_espontanea	0,103	0,478	1,828	1,415	**7,134187781**
120	A_B_3M = > 32, Grau_Varicoc = II	Conc_3M = > 15, Formas_N_3M = > 4	0,103	0,478	1,551	1,326	**3,974853321**
121	A_B_3M = > 32, Grau_Varicoc = II	Gravidez_espontanea, Formas_N_3M = > 4	0,103	0,478	2,132	1,487	**10,88366034**
96	Conc_3M = > 15, A_B_Pre = > 32	PMA, Formas_N_3M = > 4	0,103	0,458	2,229	1,467	**12,12671949**
98	Conc_3M = > 15, A_B_Pre = > 32	Gravidez_espontanea, Formas_N_3M = > 4	0,103	0,458	2,043	1,432	**9,758823821**
399	A_B_Pre = 1 to 31, Grau_Varicoc = I	PMA	0,103	0,786	1,274	1,788	1,951210608
99	Gravidez_espontanea, Formas_N_3M = > 4	Conc_3M = > 15, A_B_Pre = > 32	0,103	0,458	2,043	1,432	**9,758823821**
100	Gravidez_espontanea, Formas_N_3M = > 4	A_B_3M = > 32, Grau_Varicoc = II	0,103	0,458	2,132	1,449	**10,88417529**
75	PMA, Formas_N_3M = 1 to 3	Conc_3M = > 15, A_B_Pre = 1 to 31	0,103	0,44	1,681	1,318	**5,377324192**
39	Conc_3M = > 15, Grau_Varicoc = II	A_B_3M = > 32, Gravidez_espontanea	0,103	0,423	1,617	1,28	**4,644910954**
40	Conc_3M = > 15, Grau_Varicoc = II	A_B_3M = > 32, Formas_N_3M = > 4	0,103	0,423	1,617	1,28	**4,644910954**
16	Gravidez_espontanea, Grau_Varicoc = II	A_B_3M = > 32, Formas_N_3M = > 4	0,103	0,407	1,557	1,246	**3,980653522**

Appendix D: Published work

Published in the abstract book of the European Journal of Clinical Investigation and presented by the first author in the 53rd Annual Scientific Meeting of the European Society for Clinical Investigation in 22/05/19 (can be see in **https://doi.org/10.1111/eci.13108** at S5-O3).

Data Mining Applied to the Varicocele Condition

Judith Santos-Pereira [1], Ana Paula Sousa [2], João Ramalho-Santos [3], Jorge Bernardino [4]

[1] ISEC, Polytechnic Institute of Coimbra, Coimbra, Portugal; [2] Biology of Reproduction & Stem Cell Group, Center for Neuroscience and Cell Biology, University of Coimbra, Reproductive Medicine Unit, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal; [3] Biology of Reproduction & Stem Cell Group, Center for Neuroscience and Cell Biology, University of Coimbra, Department of Life Sciences, University of Coimbra, Coimbra, Portugal; [4] ISEC, Polytechnic Institute of Coimbra, CISUC, University of Coimbra, Coimbra, Portugal

Background: varicocele is manifested by an abnormal dilation of the veins within the scrotum. Its prevalence is related to 40% of the males treated for infertility where male factors encompass 50% of infertility causes. Its correction can be achieved with the radiological embolization technique that introduces substances into the circulation to devitalize the enlarged veins. The aim of this study was to identify data patterns on patient’s data that have undergone varicocele embolization with Data Mining since, to the best of our knowledge, this advanced data analysis technique has not been yet applied upon this highly prevalent condition.

Materials and methods: Data analysis was carried out upon a preprocessed data set of 293 men from infertile couples described using 64 features that have undergone varicocele embolization between January 2007 and April 2016. Data mining was achieved by following the Crisp-DM methodology with the application of the most commonly applied Data Mining algorithms (i.e. C4.5, K-Means and FP-Growth).

Results: The K-Means algorithm was the most effective with the following features, where statistical significance between the computed centroid values were with the ANOVA test calculated: male patient´s age (p=0.778); normality of the sperm concentration 3 months after the treatment (p<0.001); normality of the sperm progressive motility before (p<0.001) and 3 months after the treatment (p=0.011); varicocele severity grade (p<0.001); presumed occupational exposure (p=0.007) and pregnancy outcome (p=0.030). The resultant data set was of 85 couples partitioned into 4 clusters with the Manhattan distance.

Conclusions: This clinical investigation enlightened the possibility that infertile male patients with a high varicocele severity grade rarely conceive and that the frequency of patients with normal sperm concentrations 3 months after the varicocele embolization is much higher in clusters where fewer male patients work in putative hazardous environments.

Appendix E: Paper to be submitted

Preprint to be submitted to an International Journal.

Data Mining Applied to the varicocele Embolization

Judith Santos-Pereira 1*, Ana Paula Sousa 2,3, João Ramalho-Santos 2,4, Jorge Bernardino 1,5

1 ISEC, Polytechnic Institute of Coimbra, Coimbra, Portugal, santosj@hotmail.ca
2 Biology of Reproduction & Stem Cell Group, CNC, University of Coimbra, Coimbra, Portugal
3 Reproductive Medicine Unit, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal, anapauladesousa.apms@gmail.com
4 Department of Life Sciences, University of Coimbra, Coimbra, Portugal, jramalho@uc.pt
5 Centre of Informatics and Systems, University of Coimbra, Coimbra, Portugal, jorge@isec.pt
*Corresponding author´s institution: ISEC, Polytechnic of Coimbra, Rua Pedro Nunes, Quinta da Nora, 3030-190, Coimbra, Portugal, +351 239 790 200, santosj@hotmail.ca

Declarations of interest: None.
Funding sources: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abstract

Varicocele is manifested by an abnormal dilation of the veins within the scrotum. Its prevalence is related to 40% of the males treated for infertility where male factors encompass 50% of infertility causes. Its correction can be achieved with the radiological embolization technique which has been seen to positively impact pregnancy rates. To leverage ongoing investigations, this work aims to predict the success of the varicocele embolization through the pregnancy outcome and identify data patterns with data mining techniques since, to the best of our knowledge, an advanced data analysis technique has not been yet applied upon the correction of this highly prevalent condition. In this context, we have followed the CRISP-DM methodology and used the RapidMiner platform to apply the most commonly used data mining techniques in the health care domain; i.e., classification, clustering and association rule. These techniques were applied upon a dataset of 293 men from infertile couples that have undergone varicocele embolization. Our main findings suggest that among non azoospermic infertile embolized patients, the success of the varicocele embolization could be predicted with a F-measure of 77.78% through the male patient partner age and influenced by some of the male patient´s features. In a data mining application point of view, this work enabled us to conclude that knowledge discovery can be potentiated by how data mining techniques are applied; and hence, this paper provides a practical guideline for similar studies due to its interesting outcomes validated by clinical experts.

Keywords: Knowledge discovery, Decision tree, Clustering, Association Rule, varicocele embolization

1. Introduction

High dimensionality, noisy and missing values are some of the general medical data challenges that require a good time investment to preprocess it and identify data modeling steps that potentiate knowledge discovery. In this context, this paper aims to provide guidance to other similar works by sharing how we have carried out the knowledge discovery process that has computed our most interesting data mining results that we also disclose.

Data Mining is the process of discovering interesting data patterns (Han, Kamber, & Pei, 2012) where standard statistical exploratory data analysis procedures - traditional statistics - could not discover useful or new insights (Hand, Blunt, Kelly, & Adams, 2000). In our era, traditional statistics is viewed as the primary data analysis technique and data mining as the secondary technique due to its strengths and rapid developments (Hand, 1998). While the groundwork of both techniques is mathematics, data mining extends it with other subjects such as machine learning, database systems and visualization which brings important gains over the traditional statistics techniques (Tekieh & Raahemi, 2015). The main advantages of data mining over the traditional statistical techniques are its capability to analyze different types of data (i.e. numbers, names, severity degrees etc.), as well as its ability to perform non hypothetical deductive analysis. In fact, this study has benefited by this last advantage since we have found new insights on the varicocele embolization domain.

The varicocele condition is characterized by the dilation of the veins of the spermatic cord (Arif *et al*., 2018). The McGraw-Hill Concise Dictionary of Modern Medicine ("varicocele Definition," 2002) state that the varicocele condition is linked to infertility in 40% of males treated for infertility. By having in mind that male infertility factors are responsible for 50% (Kirby, Wiener, Rajanahally, Crowell, & Coward, 2016) of infertility causes, the importance of assessing data patients with a condition with such prevalence is clear, given that infertility affects an estimated 15% of couples globally (Agarwal, Mulgund, Hamada, & Chyatte, 2015).

The varicocele correction can be achieved through surgery or radiologically. Over the last decades, radiological techniques such as embolization has become increasingly popular as a less invasive technique. This technique introduces substances such as coils, sclerosants or glue into the circulation to devitalize the enlarged veins (Lippincott Williams & Wilkins, 2012).

Related works have shown that embolization successfully corrects varicocele, whether it is with coils or glue, in an average of 92% (Makris *et al*., 2018). Furthermore, varicocele correction is considered has an important procedure in the treatment of infertility (Samplaski, Lo, Grober, Zini, & Jarvi, 2017) since it has been seen to increase the pregnancy rate of those undergoing varicocele correction compared with those with untreated varicocele (Kirby *et al*., 2016). In terms of patient´s features that were analyzed in related work, we have seen that sperm parameters, as well as its categorizations were the mostly studied (Çayan & Akbay, 2018) (Samplaski *et al*., 2017) (Makris *et al*., 2018). Moreover, external factors such as previous diseases (Niederberger, 2015), occupation, drinking and smoking habits have also been assessed (Delavar *et al*., 2014), as well as the condition laterality (DeWitt *et al*., 2018) and its severity grade (Aza Mohammed & Frank Chinegwundoh, 2009). To the best of our knowledge, all these patient´s features were in the varicocele domain uniquely assessed with traditional statistics and never analyzed in the perspective of predicting or defining patterns for the varicocele embolization success which raised the need for further investigations with an advanced data analysis technique such as data mining to identify new insights in this such important infertility treatment.

In this context, the aim of this work is to predict the success of the varicocele embolization through the pregnancy outcome and identify data patterns that can contribute to the ongoing varicocele researches.

To tackle these data mining goals, we have applied upon a varicocele embolized population the most commonly used data mining techniques in the healthcare domain: classification, with the decision tree algorithm; clustering, with the K-Means algorithm and association, with the FP-Growth algorithm. In spite of these data mining techniques having a proven applicability in the healthcare domain (Tekieh & Raahemi, 2015; Ahmad, Qamar, Qasim, & Rizvi, 2015; Tomar & Agarwal, 2013), they are rarely applied together in the medical treatment domain as seen in the meta-analysis carried out in Esfandiari, Babavalian, Moghadam, & Tabar (2014) work. Accordingly, the main contributions of this paper are:

- Identify measures that had leverage knowledge discovery;
- Contribute to the ongoing research on varicocele embolization;
- Leverage the findings in the global field of male infertility.

The remainder of this paper is organized as follows. Section 2 specifies the materials used and describes the knowledge discovery process that was followed to achieve our most interesting results. Section 3 presents the best obtained results (i.e., the ones that have achieved the highest performance and clinical interest). Section 4 discusses the followed knowledge discovery process by exploring its utility through the identification of measures that were seen potentiators of knowledge discovery. Section 5 briefly disclose the context of this work by enhancing the difference of the followed methodology from related works; and finally, section 6 presents conclusions and future work.

2. Material and methodology

In order to better convey how this work was carried out, this section describes the materials that were used and the methodologies that were followed to achieve the results disclosed in the following section 3.

2.1 Materials

In this section, we describe the studied population with its analyzed attributes and we specify the electronic tools that were used to achieve the data mining goals set.

2.1.1 Studied Population

Data analysis was carried out upon a preprocessed data set of 293 infertile couples (i.e. couples that were unable to get pregnant after 1 year of regular intercourse) where male partners had undergone varicocele embolization between January 2007 and April 2016 in the Portuguese public hospital called Centro Hospitalar e Universitário de Coimbra (CHUC) with the aim of improving their chances of conceiving.

All male partners had undergone a semen analysis according to the fifth edition of the World Health Organization (WHO) laboratory manual for the examination and processing of human semen (WorldHealthOrganization, 2010) before and at 3, 6, 12 months after the varicocele embolization treatment, with a previous sexual abstinence of 3 to 4 days. Furthermore, couples were followed in fertility appointments where some of the female partners have undergone Assisted Reproduction Techniques (ART) procedures such as intrauterine insemination (IUI), in vitro fertilization (IVF), intracytoplasmic sperm injection (ICSI) or intracytoplasmic morphologically selected sperm injection (IMSI) to conceive. However, some couples were able to achieve pregnancy spontaneously.

In this context, we have collected 39 patient’s attributes which were seen studied in the male infertility domain and have subsequently generated 25 attributes. Table 1 describes all these 64 analyzed attributes where the first 39 listed below are the ones that were collected in the CHUC and the following ones, are the ones that were subsequently generated. Under the “Attribute value” column, we disclose the range of values (for numeric continuous attributes), the values (for binominal or ordinal attributes) or some values (for nominal attributes) recorded in each corresponding attribute specified under the “Attribute name” column. The indication of the attribute type is under the column named “Type” indicated with the following abbreviations: “Nu” (i.e. Numeric); “Bi” (i.e. Binominal); “No” (i.e. Nominal) or “Or” (i.e. Ordinal). The collected attributes that were seen with a good data quality by the RapidMiner platform (i.e. without having the same attribute value in all instances or/add a lot of missing values) after the data preparation step, have a check symbol (i.e. ✓) under this column. Note that the attributes which could associate a particular person were either deleted or deidentified to ensure confidentiality of the intervenient.

Table 1
List of assessed attributes

ID	Attribute name	Description	Type	Attribute value
1	Man age	Age of the male patient at embolization time	Nu ✓	23-54
2	Woman age	Age of the patient´s partner at embolization	Nu ✓	20-46
3	Infertility time	Months the couple have been trying to conceive	Nu	4-192
4	Type of infertility	Patient´s partner first or second pregnancy	Bi	Primary, Secondary
5	Woman infertility factor	Patient´s partner diagnosed infertility cause	No	Anovulation
6	Man infertility factor	Male patient diagnosed infertility cause	No ✓	Azoospermia, OAT
7	Smoking habit	Male patient smoking habits	No ✓	4 cigarettes per day
8	Drinking habit	Male patient drinking habits	No ✓	Socially, Rarely
9	Surgeries	Male patient surgeries before treatment	No ✓	Hernioplasty
10	Diseases	Male patient diseases before treatment	No ✓	Left Epididymis cyst
11	Occupation	Male patient occupation before treatment	No ✓	Factory worker
12	Severity grade	varicocele severity grade before treatment	Or ✓	I, II, III
13	Laterality	Scrotum site of the varicocele condition	No ✓	Left, Right, Both
14	Testis volume	Categorization of the patient´s testis volume	No	Above 20cc, Normal
15	Embolization date	Date of the embolization treatment	Nu	01/17/2007-04/28/2016
16	Embolized laterality	Treated scrotum laterality	No	Left, Right, Both
17	Material of Embolization	Material used during the treatment	No	Coils, Glue
18	Complications	Complications after the embolization treatment	No	None, Pain
19	Repeat embolization	Whether the patient would repeat the treatment	No	Unknown, Yes, No
20	Reason to not repeat	Reason told for not repeating the treatment	No	Unknown, Pain
21	Concentration before treatment	Concentration of spermatozoa before	Nu	0-220
22	Concentration at 3 months	Concentration of spermatozoa at 3 months	Nu ✓	0-170
23	Concentration at 6 months	Concentration of spermatozoa at 6 months	Nu ✓	0-160
24	Concentration at 12 months	Concentration of spermatozoa at 12 months	Nu ✓	0-80
25	Progressive motility before treatment	Percentage of fast/slow spermatozoa before	Nu ✓	0-89
26	Progressive motility at 3 months	Percentage of fast/slow spermatozoa at 3 months	Nu ✓	0-94
27	Progressive motility at 6 months	Percentage of fast/slow spermatozoa at 6 months	Nu ✓	0-83
28	Progressive motility at 12 months	Percentage of fast/slow spermatozoa at 12 months	Nu ✓	0-83
29	Morphology before treatment	Percentage of normal spermatozoa before	Nu ✓	0-38
30	Morphology at 3 months	Percentage of normal spermatozoa at 3 months	Nu ✓	0-21
31	Morphology at 6 months	Percentage of normal spermatozoa at 6 months	Nu	0-21
32	Morphology at 12 months	Percentage of normal spermatozoa at 12 months	Nu	1-10

ID	Attribute name	Description	Type	Attribute value
33	Pregnancy outcome	Couple got or not pregnant after embolization	Bi ✓	No, Yes
34	Number of pregnancies	Number of pregnancies had after embolization	Nu	0-3
35	Birth	Couple got or not a birth after embolization	Bi	No, Yes
36	Number of alive babies	Number of alive babies born after embolization	Nu	0-3
37	Time took to conceive	Number of months after embolization	Nu	0-79
38	ART	Patient´s partner got pregnant with ART	Bi	No, Yes
39	Spontaneous pregnancy	Patient´s partner got pregnant spontaneously	Bi	No, Yes
40	Preprocessed smoking habit	Male patient smokes or not	Bi	No, Yes
41	Preprocessed drinking habit	Male patient drinks or not	Bi	No, Yes
42	Preprocessed surgeries	Male patient got surgeries before treatment	Bi	No, Yes
43	Preprocessed diseases	Male patient got diseases before treatment	No	Epididymis
44	Hazardous occupation	Male patient works or not in a toxic environment	Bi	No, Yes
45	Altered before	Number of altered sperm parameters before	Nu	0, 1, 2, 3
46	Altered at 3 months	Number of altered sperm parameters at 3 months	Nu	0, 1, 2, 3
47	Altered at 6 months	Number of altered sperm parameters at 6 months	Nu	0, 1, 2, 3
48	Altered at 12 months	Number of altered sperm parameters at 12 months	Nu	0, 1, 2, 3
49	Semen classification before treatment	Semen classification before treatment	No	OAT
50	Semen classification at 3 months	Semen classification 3 months after treatment	No	Normozoospermia
51	Semen classification at 6 months	Semen classification 6 months after treatment	No	Azoospermia
52	Semen classification at 12 months	Semen classification 12 months after treatment	No	Azoospermia
53	Concentration category before treatment	Normality of the concentration value before	Bi	Abnormal, Normal
54	Concentration category at 3 months	Normality of the concentration value at 3 months	Bi	Abnormal, Normal
55	Concentration category at 6 months	Normality of the concentration value at 6 months	Bi	Abnormal, Normal
56	Concentration category at 12 months	Normality of the concentration value at 12 months	Bi	Abnormal, Normal
57	Progressive motility category before	Normality of the motility value before	Bi	Abnormal, Normal
58	Progressive motility category at 3 months	Normality of the motility value at 3 months	Bi	Abnormal, Normal
59	Progressive motility category at 6 months	Normality of the motility value at 6 months	Bi	Abnormal, Normal
60	Progressive motility category at 12 months	Normality of the motility value at 12 months	Bi	Abnormal, Normal
61	Morphology category before treatment	Normality of the morphology value before	Bi	Abnormal, Normal
62	Morphology category at 3 months	Normality of the morphology value at 3 months	Bi	Abnormal, Normal
63	Morphology category at 6 months	Normality of the morphology value at 6 months	Bi	Abnormal, Normal
64	Morphology category at 12 months	Normality of the morphology value at 12 months	Bi	Abnormal, Normal

2.1.2 Tools

In terms of software, this study used the software tools that we below specify for each following purpose of use:

- Data collection and preparation: Microsoft Excel 2016, Home and Student Edition.
- Data integration: Microsoft SQL Server Management Studio 2012.
- Data Analysis (Statistical & Mining): RapidMiner Studio Educational platform, version 8.1.001.

RapidMiner was the data mining tool that was selected since related works (Al-odan & Saud, 2015) (Almeida, Gruenwald, & Bernardino, 2016) (Sharma, Singh, & Khatri, 2016) and consulting companies such as Gartner highly ranks this tool.

This data mining tool is a data science software platform developed by the company of the same name. It was formerly known as YALE (Yet Another Learning Environment) and was developed at the Artificial Intelligence Unit of the Technical University of Dortmund, Germany, that has its initial release in 2006. This software platform as a Free Edition that can be used on data sets up to 10 000 rows with a limit of 1 logical processor that is distributed under the AGPL license - AGPL is a license that can be attributed to open source software that can be run over a network. However, for academic use, the RapidMiner enables the use of a one year unlimited version which is the one that we have used.

RapidMiner is written in the Java programming language and provides a GUI to design and execute analytical processes. These processes are built with the mean of drag and drop components that can apply data transformation tasks, descriptive statistical tests and Data Mining algorithms already implemented to the data. These components are in RapidMiner called “operators” and the connection of several operators is called visual composition framework (VCF). Each operator has several parameters that can be configured and always has input and output ports to respectively receive the data from the previous operator and send it to the next operator. Through this experience, we have identified that the main advantages of RapidMiner are its capabilities to optimize the data modeling process through its intuitive operators, as well as the good support delivered by its vast and active online community.

2.2 Methodology

This study was carried out by following the methodology called cross-industry standard process for data mining projects (CRISP-DM). This methodology encompasses a set of six phases that can be executed recursively and be briefly described as covering the following tasks (Chapman *et al*., 2000):

- Business understanding – determine the business objectives and data mining goals;
- Data understanding – collect, describe and explore the data and verify its quality;
- Data preparation – select, clean, construct, integrate and format the previously understood data;
- Modeling – select modeling techniques, define the training/testing design, build the data mining models and assess them with performance measures;
- Evaluation – assess models with respect to business objectives and domain expertise;
- Deployment – results deployment (e.g. simple report).

Therefore, this study has begun with the understanding of the male infertility domain, as well as the identification of the needs/objectives of the ongoing clinical investigations that were converted into the data mining goals set (i.e. aims of this work); followed by the understanding of the collected data; the preparation of said data; the assessment of the execution of a set of defined data modeling steps implemented with models built in the RapidMiner platform; the evaluation of the models´ compliance with the business objectives and clinical expectations, and at last, the disclosure of findings.

In order to showcase a practical guideline that potentiates knowledge discovery, in this subsection we explain how the data understanding, the data preparation, and the modelling phase were carried out during this data mining application since they are the steps that young data scientists usually seek for guidance. In order to better convey the executed modelling phase, we also disclose how each selected data mining algorithm was applied.

2.2.1 Data Understanding

The data of the first 39 attributes listed in Table 1 was collected from the information technology systems, patient medical dossiers and semen analysis reports of the CHUC which were all integrated with the Microsoft SQL Server and then exported to an Excel file for further assessments.

Afterwards, these attributes were described and explored statistically as suggested in Han *et al*. (2012) to successfully prepare the data for the mining process. In fact, these authors state that one should analysis the central tendency of the data (i.e. compute the Mean, Median and Mode of each attribute), as well as its dispersion (i.e. compute the minimum value (Min), the value of the first quartile (Q1), the value of the third quartile (Q3), the maximum value (Max) and the standard deviation (SD) of each attribute). Hence, this study has computed all these statistical measures and complemented its assessment by generating graphs such as time series, box plots and histograms with the RapidMiner platform and the Excel software to better understand the collected attributes.

At last, we have verified the quality of the understood data as suggested in Thatipamula (2013) and Maydanchik (2007) to assess if the collected data was trustworthy and ready to respond to the Data Mining goals set. The collected data was checked regarding the compliance of the requirements of the following key data dimensions that can be briefly described as the following (Thatipamula, 2013):

- Completeness - having all attributes needed and in a usable state to tackle the Data Mining goals;
- Consistency - showing data coherence between attributes and without duplicated instances;
- Conformity - data complies with a specific format that is the same across all instances;
- Accuracy - having the correct data;
- Integrity - having the correct data linkage (e.g. the recorded male patient partner is correct).

The Completeness, Consistency and Conformity were assessed by verifying whether the provided data complied with the key dimension´s requirements where the previously carried out statistical analysis had given us some insights. In the other hand, the Accuracy and Integrity were assessed by validating whether the collected data was coherent across the several information technology systems from which the data was retrieve from since several people were involved in its collection.

2.2.2 Data Preparation

After the Data Understanding phase, we have seen that several data preparation tasks were needed to tackle the Data Mining goals set; and hence, all CRISP-DM tasks proposed for this phase were carried out to leverage its data quality which has begun with the selection and the cleaning of the data during the assessment of its key data dimensions: during the selection task, we have mainly deleted duplicated and empty instances and during the cleaning task, we have corrected the data based on all the requirements of the key data dimensions. At last, we were not able to fill all missing values due to the inexistence of some of them in the medical information systems but nearly half of the attributes values were on average filled (55.86%). This study has not used a technique to fill the remaining missing values since we aimed to analyze them exactly as they were - data estimation techniques could bias the findings due to our small sample size.

The data construction and integration tasks encompassed the production of derived attributes based on: the simplification (e.g. attribute with the id 40 to 44 of the previous Table 1); the aggregation (e.g. attribute with the id 45 to 52 of the previous Table 1) and the categorization (e.g. attribute with the id 53 to 64 of the previous Table 1) of existing attributes. All these attributes cover the 25 subsequently generated attributes.

These attributes mainly handle sperm parameter data to enable the mining of this such relevant information in the male infertility domain on different perspectives; and hence, potentiate knowledge discovery. Therefore, we have transformed the collected numerical sperm parameter attributes into categorized ones based on the lower reference limits defined by the WHO (WorldHealthOrganization, 2010). These lower reference limits, also called thresholds, state that a male patient is considered with normospermia (i.e. with normal sperm parameter values) when having the following semen characteristics:

- Sperm concentration equal or above 15 million/ml;
- Sperm progressive motility with at least 32%;
- Sperm morphology with at least 4%.

These reference limits enabled us to categorize the sperm parameters values on whether or not its values are normal (i.e. attribute with the id 53 to 64 of the previous Table 1), as well as classify the semen (i.e. attribute with the id 49 to 52 of the previous Table 1) and indicate the number of altered sperm parameters seen in the semen (i.e. attribute with the id 45 to 48 of the previous Table 1) as defined in Table 2.

Table 2

Description of semen classifications and number of altered sperm parameters

Semen classification	Semen characteristics	Number of altered parameters
Normozoospermia	Sperm concentration =>15 million/mL Sperm progressive motility => 32% Sperm morphology => 4%	0
Oligozoospermia	Sperm concentration < 15 million/mL Sperm progressive motility => 32% Sperm morphology => 4%	1
OligoAstenozoospermia	Sperm concentration < 15 million/mL Sperm progressive motility < 32% Sperm morphology => 4%	2
OligoTeratozoospermia	Sperm concentration < 15 million/mL Sperm progressive motility => 32% Sperm morphology < 4%	2
Asthenozoospermia	Sperm concentration => 15 million/mL Sperm progressive motility < 32% Sperm morphology => 4%	1
AsthenoTeratozoospermia	Sperm concentration => 15 million/mL Sperm progressive motility < 32% Sperm morphology < 4%	2

Semen classification	Semen characteristics	Number of altered parameters
Teratozoospermia	Sperm concentration => 15 million/mL Sperm progressive motility => 32% Sperm morphology < 4%	1
OligoAstenoTeratozoospermia	Sperm concentration 15 < million/mL Sperm progressive motility < 32% Sperm morphology < 4%	3
Azoospermia	Sperm concentration = 0 million/mL Sperm motility does not exist Sperm morphology does not exist	1

Next, the data was reexplored statistically - as carried out during the data understanding phase since the data is in this stage trustworthy to guide our mining process - and then, have extended our statistical analysis by applying the Pearson correlation and the ANOVA statistical test upon the collected continuous numerical attributes and the Chi-square test upon the collected nominal attributes to identify the attributes that are more related with the pregnancy outcome. Furthermore, for each of these attributes, we have analyzed the Stability criterion (i.e. the proportion of the most frequent value), as well as have identified their proportion of Missing values. Hence, the attributes were in this study selected based on the following selection criteria where their thresholds, below specified under parentheses, were the ones suggested by the RapidMiner platform:

- Low value on the Significance probability computed with ANOVA and Chi-square (i.e. <0.05);
- High value on the Correlation with the Pregnancy outcome (i.e. > 0.01);
- Low rate on the Stability (i.e. <90%) and Missing criteria (i.e. <70%).

Since regular attributes should not be correlated, we have separated the selected attributes into several groups to overcome this situation which also lead us to the execution of several modelling steps.

At last, we have formatted the selected data to comply with the specifications of each data mining technique (e.g. transform all nominal attributes into numerical and afterwards normalize all attributes between 0 and 1 to apply the K-Means algorithm).

2.2.3 Decision tree

Decision tree is an umbrella name for a set of algorithms that computes decision trees. Decision trees are very popular for the application of the classification data mining technique and in the health care domain, they are the most commonly used due to their higher result ´s interpretability (Esfandiari *et al*., 2014). They are mainly applied to identify the most interesting attributes that one should use to mine and/or to predict the conditional probability of an outcome, recorded under a special attribute called label, based on its historical records. The most commonly used algorithm is the C4.5 algorithm.

Classifiers as the decision tree algorithm C4.5, began with the attribute that promotes the highest gain of information by placing it at its root and its ramification is guided with the entropy measure. In fact, the C4.5 algorithm aims to decrease the entropy through the downward splitting of the nodes; and hence, choose as attribute nodes, the one that produces the purest daughter nodes (i.e. entropy equal to zero) to compute the smallest tree as soon as possible (Witten, Frank, & Hall, 2011). Therefore, the C4.5 algorithm works as follows: after splitting a node and testing whether the entropy of the next node is lesser than the entropy before splitting and if this value is the least as compared to all possible test-cases for splitting, then the node is split into its purest constituents (i.e. attribute values). This assessment is recursively performed with the remaining attributes until all leaf nodes are pure (i.e. leaf nodes with instances belonging to one class, such as: “Pregnancy outcome”= “Yes” or “Pregnancy outcome”= “No”), or until it is not possible to further on split because the entropy is equal to 1. In other words, as Witten *et al*. (2011) states, this algorithm works top-down, seeking at each stage for an attribute that can better split the classes (i.e. yields the highest gain of information at each stage). The gain of information is in the C4.5 algorithm computed with the Gain ratio measurement which is an extension of the information gain measure used by the ID3 algorithm (Han *et al*., 2012).

Since decision trees are supervised learners, most decision trees algorithms need a binomial attribute as a label attribute and some implementations, require non-missing values under this special attribute. Hence, the application of this algorithm entails to select a label attribute, transform it as binomial (if it is not yet binomial), filter the rows of the data set by non-missing value under the label attribute, apply some optimizations (e.g., other algorithms, attribute discretization, attribute normalization etc.), train and test the decision tree algorithm with different settings (i.e. different parameter values) and then, validate the model that has previously computed the highest elected performance measures.

Regarding the test design that we have followed, we have implemented a test design that splits the data set into 3 parts; and therefore, called the 3-parts test design, which partitions the data set into 80% for training/testing and 20% for validation where in the 80% part, 70% is taken to train and the remaining 30%, to test the data set.

In Guh, Wu and Weng (2011) work, they have divided their data set of 5275 instances into only two parts. In fact, they have used 80% for training and 20% for testing; and hence, they have not validated their model. However, the CRISP-DM methodology (Chapman *et al*., 2000) suggests a 3-part test design to avoid test overfitting which especially occurs with small data sets.

The first 80%/20% data set partitioning was executed with the RapidMiner´s Split data operator. This partitioning was performed with the Stratified sampling type to ensure that we have in each subset the same number of instances classified as “Yes” and “No” to promote balanced subsets. Afterwards, the 80% training/testing subset was further on partitioned into 70% training and 30% testing with the following operators:

A.Split Validation - operator that splits the data set with a single iteration.

B.Cross Validation - operator that performs several split validations.

The Split Validation operator has partitioned this last 80% subset with several sampling types (e.g. Linear, Shuffle and Stratified) to test the one that delivered the best performance.

On the other hand, Cross Validation, which was already used in related works to predict seminal quality (Gil, Girela, De Juan, Gomez-Torres, & Johnsson, 2012), was in this study also applied with its corresponding RapidMiner´s operator which performs several Split Validations. In fact, this operator splits the data set into k sub-datasets and keeps one sub-dataset for testing and the remaining ones for training. Next, it recursively selects another sub-dataset for testing and considers the remaining ones for training. This test is done k times (i.e. until all sub-datasets were at least 1 time a testing dataset) and several k values can be tested. In this study, we have tested the model with k=2 to k=4 since the default number of folds set by the RapidMiner platform is 4 and our small number of training/testing subset lead us to also test the model with larger subsets (i.e. lower number of k folds)

Since we have applied two decision tree algorithms (i.e. the decision tree from the RapidMiner platform and the java application of the C4.5 algorithm called W-J48), all built decision tree models have executed the following ordered testing steps:

1. Test the RapidMiner´s Decision tree algorithm within a Split Validation.
2. Test the RapidMiner´s Decision tree algorithm within a Cross Validation.
3. Test the W-J48 algorithm within a Split Validation operator.
4. Test the W-J48 algorithm within a Cross Validation operator.

The training of the Decision tree algorithm entailed its application on several groups of attributes with the variation of the parameters shown in Table 3 at each previously listed testing steps. These varied parameters were selected based on the guidelines explained in Bala Deshpande (2012). The variation of the model parameters involved the execution of 8664 tests per modeling step of the decision tree algorithm: in each modeling step, we have carried out 2160 tests for the decision tree algorithm ran within a simple validation; 6480 tests for the decision tree algorithm ran within a cross validation; 6 tests for the W-J48 algorithm ran within a simple validation and 18 tests for the W-J48 algorithm ran within a cross validation. These tests were executed within the RapidMiner´s "Optimize Parameters" operator which returns the best model that we have at last validated.

Table 3

Parameters varied through decision tree training

Related Operator	Parameter Name	Tested Values
Decision tree	Criterion	*Information_gain*; *Gain_ratio*; *Gini_index*; *Accuracy*.
Decision tree	Minimal size for split	4; 5; 6.
Decision tree	Minimal gain	0.100; 0.140; 0.180; 0.220; 0.260; 0.300.
Decision tree	Minimal leaf size	2;3;4;5; 6.
Decision tree & W-J48	Apply pruning	Yes; No.
Split & Cross Validation	Sampling Type	Linear sampling; Shuffled sampling; Stratified sampling.
Cross Validation	Number of folds	2;3;4.

Evaluation measures called performance metrics or error rates, were used to elect the right and best model to tackle the goals of this study. Since we aim to predict the success of the varicocele embolization, we had more interest on the model´s success to classify positive classes; and hence, we have focused on assessing performance measures that not only indicates how accurate and worth the model is through the respective Accuracy and AUC measure, but also, how capable the model is to classify positive classes. In this context, we have elected the performance measures that we below disclose:

- Accuracy - The proportion of instances classified correctly among the total number of instances;
- Precision - The proportion of instances classified correctly as positive among the number of instances classified/predicted as positively;
- Recall - The proportion of instances classified correctly as positive among the number of instances with the label attribute set to positive (i.e. “Pregnancy outcome” = “Yes”);
- F-Measure - The harmonic mean of Precision and Recall;
- AUC - The probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

However, the metrics that were in this study determinant in the choice of the right Data Mining model for the prediction of the embolization success were, in the following order: The F-Measure, The AUC, the Recall and the Accuracy metric because the F-Measure is a complete performance measure to assess how well the positive classes are being classified in contrast to the Accuracy measure that does not tell us the percentage of positive classes that were correctly classified. However, we have seen that the Accuracy measure is in related works mostly considered as the determinant measure; and hence, our model selection criteria also entails its assessment.

Concerning the work of Guh *et al*. (2011), these authors have used the following measures to assess its Decision tree model: Accuracy, Recall and Specificity. Hence, the AUC measure was not used, which we see as a drawback since the AUC measure allows to assess if a model is worth applying (i.e. if its prediction is not random).

Since all studied related works that have applied data mining techniques to sperm parameters have obtain Accuracies above 73% during its training/testing, this study considers that an Accuracy above this value during its training/testing is also acceptable, as well as a fair AUC value above 0.70. However, we aim to also achieve these performance values during the model´s validation.

2.2.4 K-Means

K-Means is a commonly used Data Mining algorithm for the application of the clustering data mining technique (Han *et al*., 2012). Its aim is to partition a data set into k groups of similar instances (called clusters) to find a categorization for the studied object (e.g. embolized patients). Its partitioning is performed with an agglomerative technique and not a hierarchical one; and hence, its results are disclosed in a so called centroid table which records the mean point value of each cluster per attribute (i.e. centroid) that can be visually interpreted with a centroid series plot. The K-Means algorithm only mines filled numeric attributes that should be normalized and works as follows (Han *et al*., 2012):

1. Arbitrarily choose k attribute values as initial cluster centers;

2. Assign each instance to the cluster to which the instance is the most similar with based on its shorter defined spatial distance to the centroid;
3. Update the cluster means with the newly assign instance;
4. Repeat step 2 and 3 until there is no change on the centroid value.

The K-Means algorithm works well to find spherical-shaped clusters in small to medium-size data sets without outliers (Han *et al*., 2012). Since our data set is small and preprocessed, we have applied it to identify data patterns that could categorize/describe the most successful varicocele embolized patients.

The training of the K-Means algorithm entailed its application on the group of attributes that delivered the largest number of filled instances with the variation of the number of clusters, that went from 2 to 4, and the type of spatial distance (i.e. Euclidean and Manhattan distance). We have at last validated the found insights into a separate data subset to formulate conclusions.

Clustering results were assessed externally and internally as follows: externally, by analyzing the generated centroid table/plot, as previously said, and internally, by assessed the distance similarity index called Davies Bouldin which is indicated for crisp/hard clusters (i.e. clusters where each instance only falls within one cluster). In fact, we have used the Davies Bouldin index to identify the optimal number of clusters one should use. This index specifies the density and the separation between clusters; i.e., the closer the absolute index is to 0, the better, since it expresses that the generated clusters have a low intra-cluster distance and a high inter-cluster distance. However, this index does not inform if the generated clusters have interesting insights so we have then carefully checked if the generated centroid table had a clinically interesting patient classification that could fulfill the data mining goal set.

2.2.5 FP-Growth

During the data understanding phase, we have found that we were more successful at identifying the most correlated attributes with the label attribute with the Chi-square statistical test than the Pearson correlation. Hence, our idea to apply a data mining algorithm that would identify frequent item sets has flourished and made us apply an algorithm from the association rule data mining technique to not only find data patterns, but also rules that could predict the success of the embolization treatment to complement the findings of the decision tree algorithm. In fact, as experienced, decision trees are "greedy" algorithms that hardly provide interesting results in small data sets. For this reason, we have sought which association rule algorithm can be applied to our data set, and have seen that the RapidMiner platform had the association rule algorithm called FP-Growth.

The FP-Growth algorithm aims to find frequent patterns and interesting relationships among the data set attributes. This algorithm is an optimization of the APRIORI algorithm since it has the ability to only perform two scans of the data set to identify the most frequent item sets (i.e. the first scan is to detect the frequency of each attribute and the other one, is to build the FP-Tree).

The FP-Growth algorithm starts by scanning the data set to find the frequent single items. Then, it sorts them to construct a tree that presents the association between the frequent attributes with the indication of their support in each node and at last, mines the generated conditional pattern-base recursively to identify frequent patterns and then formulate association rules (Han *et al*., 2012).

The FP-Growth algorithm requires that all input attributes have to be binomial, and to better interpret the result it is a good practice to map the attribute values. In contrast to the K-Means algorithm that cannot accept data sets with empty values, this algorithm can. However, the algorithm does not consider the missing attributes values since this algorithm seeks to count frequencies (i.e. count attribute values set to TRUE). Nevertheless, this point is useful for our type of data set since it has a low number of instances with all attributes filled. Furthermore, this technique is widely used in the bioinformatics field which reinforced its selection to tackle the data mining goals set.

As carried out in the healthcare work of Yildirim (2015), association rules were evaluated objectively (i.e. through the computed rules´ measures) and subjectively (i.e. through the evaluation of the clinical sense and interest of the rule for the studied domain).

Objectively, the rules were assessed through their computed Support, Confidence, Lift and Conviction measure since a high Support indicates that the rule occurs frequently; if it has a high Confidence, it reveals that its conditional probability is high; if the Lift measure presents a different measure than 1, it means that the attributes covered in the rule are related with each other - which means that the generated rule can be considered as interesting - and if the Conviction measure is different than 1, it means that the rule direction has an implication; and hence, it also contributes for its interestingness.

Subjectively, we have selected the rules that enabled to fulfill the prediction of the varicocele embolization success.

Consequently, the selection of the most objectively and subjectively interesting association rules entailed the assessment of each computed association rule by the following conditions which can be seen as our pruning conditions based on Yildirim (2015) work:

- Objectively interesting:
    - Support > = 0.1
    - Confidence > = 0.4
    - Lift and Conviction > = 1.1
- Subjectively interesting:
    - The Antecedent occurred before, or at the same time of the Consequent.
    - Support > = 0.15 which means that the rule encompasses at least 35 patients.

Rule pruning is recommended in the mining of health care data sets since this type of data tend to generate a large number of rules with low Support. In fact, in Shukla, Patel and Sen (2014) – a study that performs a review on the application of data mining techniques in the health care domain – the authors state that in the health care domain we tend to have a significant fraction of association rules that are irrelevant and that the most relevant rules often appear with high quality metrics but with a low Support. We believe that this is the reason why in Yildirim

(2015), the Support was set to 1% and the Confidence to 40%. Furthermore, after our first application of the FP-Growth algorithm, we have seen that a Support=0.1 and a Confidence=0.8, as we have initially thought, would not generate subjectively interesting rules so we have lowered this threshold to 0.0 and after that, have seen that a support equal to 0.1 and a confidence equal to 0.4 was in our case also enough to identify the most objectively and subjectively interesting rules as in Yildirim (2015) work.

2.2.6 Modeling steps

The designated findings disclosed in the next section 3 were achieved with the execution of the following 7 modeling steps in this specific order which encompassed our modeling strategy:

1. Apply the Decision tree algorithms upon all collected attributes that have a good data quality, and afterwards, only upon the first group of selected attributes;
2. Apply the K-Means algorithm upon the group of selected attributes that produces the largest data sample, seek for interesting data patterns through the assessment of its generated centroid tables and plots, and at last, validate the identified patterns in another subset of data;
3. Apply the FP-Growth algorithm upon all groups of selected attributes, select the objectively and subjectively interesting rules based on the pruning criteria disclosed in the previous section 2.2.5 and then, select the ones with the highest Support and Confidence for each data mining goal set;
4. Continue the Decision tree application by applying its models upon the remaining groups of selected attributes and afterwards, upon the data set prepared for the training of the K-Means algorithm;
5. Select the decision tree model with the highest F-measure and clinical coherence;
6. Analyze all interesting findings as a whole by checking their coherence;
7. Formulate conclusions for the most relevant attributes and validate them with clinical experts to approve the built models.

3. Results

By analyzing the filled preprocessed dataset with descriptive statistics, we were able to see that 81.65% of our embolized patients had the varicocele condition on their left testicle. Furthermore, out of the 293 assessed patients, we were able to retrieve the varicocele severity grade of 211 patients: 111 patients (52.61%) had a moderate severity grade (i.e. Severity grade = II), 67 patients (31.75%) had a mild severity grade (i.e. Severity grade = I) and 33 patients (15.64%) had a severe severity grade (i.e. Severity grade = III). Hence, in most cases, the varicocele condition was diagnosed with a severity grade of I or II. Male patients´ ages went from 23 to 54 years old with a mean of 34.43 and a standard deviation of ± 5.215 years old. Patient´s partners have a lower age mean (32.22), as well as a shorter standard deviation (± 4,399) which indicates that, on average, the woman patient is younger than the male patient. On average, couples arrive to the medical infertility appointment with a 39-month (± 28.87) infertility time span and 80% of them, are related to the first pregnancy of the male patient´s partner. From the 293 assessed patients, we were able to know the pregnancy outcome of 230 of them. Out of these patients, 46.52% were successful (i.e. 107/230) and 53.47% (i.e. 123/230) were not. From these 107 successful couples, 61.7% (i.e. 66/107) got pregnant with an ART procedure and 45.8% (i.e. 49/107) got pregnant spontaneously. Before the embolization treatment, the most common semen classification was Oligoastenoteratozoospermia encompassing 26,89% of the 238 male patients that were possible to categorize. This shows that varicocele clearly reduces patients' ability to achieve pregnancy, as they produce very few sperm (oligo), that are not very motile (astheno) and have abnormal morphology (terato). However, 3 months after the varicocele embolization, we see that the most common semen classification is Normozoospermia with 19.90% (41/206) by increasing in 14% which enabled us to fulfill the data mining goals set which aims to predict and describe the success of the varicocele embolization; and hence, this section begins by disclosing the attributes that were selected to tackle these data mining goals and in the following sections, showcase the best obtained results grouped by each applied data mining algorithm.

3.1Selected attributes

Based on the selection criteria exposed at the end of section 2.2.2, we have selected the attributes statistically described in the below Table 4. For predictive purposes, we have used the "Pregnancy outcome" attribute as the label attribute due to its quite balance characteristic (i.e. 46.52% "Yes" vs 53.47% "No").

Table 4
Selected attributes for data mining applications

Selection ID	Attribute Name	Significance (*p*)	Correlation (*r*)	Stability	Missing
1	Woman age	ANOVA 0.018	0.156	11.27%	3.07%
2	Severity grade	Chi-square 0.049		52.61%	27.99%
3	Concentration at 6 months	ANOVA 0.015	-0.161	11.45%	55.29%
4	Progressive motility before treatment	ANOVA 0.018	-0.155	12.35%	14.33%
5	Morphology at 3 months	ANOVA 0.004	-0.186	16.11%	38.57%
6	Hazardous Occupation	Chi-square 0.023		63.86%	31.06%
7	Semen classification before treatment	Chi-square 0.017		26.89%	18.77%
8	Semen classification at 3 months	Chi-square 0.018		19.90%	29.69%
9	Concentration category at 3 months	Chi-square 0.017		54.69%	16.38%
10	Progressive motility category before treatment	Chi-square 0.027		62.55%	14.33%
11	Progressive motility category at 3 months	Chi-square 0.022		53.46%	25.94%

If we analyze the selected attributes, we see that the production of derived attributes during the data preparation CRISP-DM phase was worthwhile since almost half of the identified attributes were from the 25 subsequently generated attributes. Furthermore, we have seen that the ANOVA statistical test was helpful to select numerical continuous relevant attributes since their corresponding Pearson correlation values were seen despicable with the "Pregnancy outcome" label attribute. Moreover, the Missing criteria unveiled that our modeling data mining process would not be facilitated which guided us to the application of different techniques such as the association data mining technique that can manage missing values. Additionally, due to the attribute correlation between some of the selected attributes, we have defined the following groups of attributes to separately test our models:

A.Severity grade, Concentration at 6 months, Progressive motility before treatment, Morphology at 3 months, Hazardous Occupation, Pregnancy outcome.

B.Severity grade, Concentration category at 3 months, Progressive Motility category before treatment, Progressive Motility category at 3 months, Hazardous Occupation, Pregnancy outcome.

C.Severity grade, Semen classification before treatment, Semen classification at 3 months, Hazardous Occupation, Pregnancy outcome.

Hence, all these data specificities guided us on the construction of the modeling steps disclosed in section 2.2.6.

If we analyze the defined groups of attributes, we see that these groups of attributes focus on seeking interesting findings only upon relevant male patient attributes but to potentiate knowledge discovery, we have extended our assessment to the remaining identified attribute. This step was seen helpful to tackle the predictive modeling task. In fact, the most efficient and interesting computed decision tree model encompassed the “woman age” attribute. Regarding the group of attributes that has generated the most interesting results, we have found that it was the group B; and hence, next sections mainly disclose its related results.

3.2 Classification

The performance measures of the best result obtained at each decision tree testing step described in section 2.2.3 and applied upon different groups of attributes specified under the column named “Attribute” of Table 5 are in this last table disclosed. During the first testing steps showcased in Table 5 (i.e. 1.1 to 1.4), we aimed to identify relevant attributes to complement the attributes listed in Table 4 by applying the 4 decision tree testing steps (i.e. apply the RapidMiner´s decision tree within a simple validation, then, a cross validation etc.) upon the collected attributes that were seen with a good data quality by the RapidMiner platform (i.e. the ones checked in Table 1); the following testing steps showcased in Table 5 (i.e. 2.1 to 4.4) sought to apply these same 4 decision tree testing steps upon the defined groups of attributes disclosed previously in section 3.1; and at last (i.e. steps 5.1 to 5.4), aimed to apply these decision tree testing steps upon the data subset that had delivered the most interesting insights during the K-Means training; i.e. the attribute group B preprocessed as follows: filtered by non-missing values; binominal attributes parsed into numerical and the “Severity grade” attribute manually dichotomized to end up with a normalized and numerical subset of 0 and 1 that encompassed 85 instances.

The step ID 1.1 to 1.4 has not identified new interesting attributes, which left us with the ones listed in Table 4. If we analyze the following tests and focus on the computed F-measures, we see that the Step ID 5.1 disclosed in Table 5 has computed the highest value (i.e. 75%) which made us validate it upon the remaining 20% of the dataset. This validation has computed the performance measures disclosed in the next Table 6 under the corresponding test ID 5.1 where we can see an F_Measure=70.59% and AUC=0.750.

Table 5
Performance measures of the best decision tree testing steps

ID	Attribute	Accuracy	Precision	Recall	F-Measure	AUC
1.1	Good quality	67.27%	60.71%	70.83%	65.38%	0.680
1.2		58.14%	57.38%	40.23%	47.62%	0.575
1.3		60.00%	unknown	0.00%	unknown	0.500
1.4		54.35%	58.33%	7.95%	14.29%	0.498
2.1	A	67.27%	68.18%	57.69%	62.50%	0.701
2.2		60.33%	65.84%	38.57%	45.88%	0.638
2.3		61.82%	58.82%	41.67%	48.78%	0.603
2.4		58.70%	57.88%	48.16%	49.60%	0.570
3.1	B	74.55%	73.08%	73.08%	73.08%	0.747
3.2		62.50%	60.28%	58.12%	59.16%	0.613
3.3		63.64%	57.14%	36.36%	44.44%	0.606
3.4		63.59%	61.63%	58.28%	59.85%	0.621
4.1	C	67.27%	57.89%	52.38%	55.00%	0.616
4.2		61.92%	59.72%	54.61%	56.88%	0.569
4.3		70.91%	70.00%	58.33%	63.64%	0.739
4.4		66.30%	66.11%	58.12%	60.95%	0.684
5.1	B preprocessed	80.00%	85.71%	66.67%	75.00%	0.717
5.2		64.71%	64.58%	47.50%	53.20%	0.589
5.3		55.00%	50.00%	44.44%	47.06%	0.439
5.4		60.29%	56.25%	52.23%	53.72%	0.644

Table 6
Performance measures of the validation of the best decision tree testing steps

ID	Accuracy	Precision	Recall	F-Measure	AUC
1.1	54.35%	unknown	0.00%	unknown	0.500
2.1	56.52%	52.38%	52.38%	52.38%	0.509
3.1	47.83%	44.00%	52.38%	47.83%	0.554
4.3	56.52%	53.33%	38.10%	44.44%	0.604
5.1	70.59%	66.67%	75.00%	70.59%	0.750

In spite of the acceptable generated performance measures in the 5.1 test, it was seen that the decision tree had a clinically incoherent leaf that made us prune it to diminish the seen data overfit; and hence, its performance validation measures were seen increased; i.e., Accuracy=76.47% , Precision=70.00% , Recall= 87.50%, F-Measure to 77.78%, and AUC = 0.771 which surpassed the Accuracy threshold initially established of 73%. The validated decision tree can be seen in Fig. 1 with its confusion matrix depicted in Table 7 which can be described as:

- Woman age > 33: Pregnancy outcome =No { No=25, Yes=12}

- Woman age <= 33: Pregnancy outcome =Yes { No=12, Yes=19}

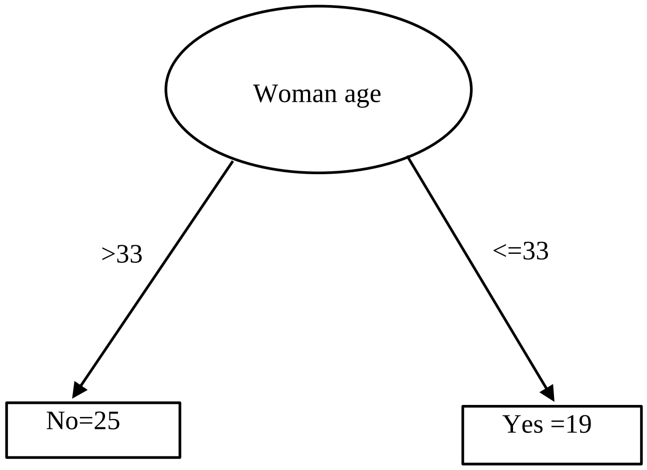


Fig. 1. Best computed decision tree

Table 7
Definition of the confusion matrix during the decision tree´s validation

Actual	Predicted	
	Yes	No
Yes	7	1
No	3	6

Its identified optimized parameter values were:

Splitting criterion: accuracy;
Minimal size for split: 4;
Minimal gain: 0.1;
Minimal leaf size: 2;
Pruning: True;

The generated decision tree, in spite of not elaborated, enables us to conclude that most women below and equal to 33 years old are able to get pregnant (i.e. 61.29% (19/(12+19))) in contrast to women above 33 (i.e. 32.43% (12/(25+12))) which indicates that the male patient´s partner age prevails upon the varicocele condition since the male patient´s partner age was selected for the tree root to the detriment of all other assessed male patients features. This situation was also seen in the unpruned decision tree of test 5.1 and this clinical finding goes along with the male infertility-related work (Williams & Alderman, 2001); and hence, its result served as a validator of what we already expected.

3.3Clustering

Since the group of attributes B generates the largest sample of filled data (i.e. 85 instances), we have applied the K-Means algorithm upon this set of attributes. The most interesting identified patient´s categorization was retrieved from the series plot depicted in Fig. 2, which was generated from the centroid table showcased in Table 8. Please note that the centroids depicted in Table 8 can be consider as relative frequencies for the value 1 since all values of the data set were for the K-Means application normalized between the value 0 to 1.

The choice of the number of clusters was based on the lowest Davies Bouldin index computed by the RapidMiner during the optimization of the K-Means model parameters which as computed a value of -1.330 for 4 clusters generated with the Manhattan distance. This partitioning lead us to the following cluster distribution: Cluster 1: 21 instances; Cluster2: 38 instances; Cluster 3: 14 instances; Cluster 4: 12 instances.

In this case, this high Davies Bouldin index can be explained by the high number of assessed attributes which difficult the formation of completely separated clusters as Fig. 2 illustrates. Hence, this K-Mean result was seen as a starter to further on individually assess the subset of attributes that constitutes the identified interesting insights. In this case, we have identified these 2 interesting insights:

- The varicocele severity grade vs the pregnancy outcome;
- The pregnancy outcome vs the hazardous occupation, the concentration category at 3 months and the progressive motility category before and at 3 months after treatment.

The attributes related with the first insight encompass 174 filled instances which enabled us to validate the found insight upon the remaining 89 instances (i.e. 174 total instances - 85 training instances). During this validation, we have seen that the relative frequency of successful patients is higher on patients with a moderate to mild varicocele severity grades (i.e. 58,3% and 46,4% respectively) in contrast to the ones with a severe severity grade. However, the pregnancy rate related with this last severity grade was seen during the model ´s validation not as low as during the model´s training (i.e. validation = 38.5% vs training = 8.3%).

Regarding the second insight, we had 126 filled instances under its attributes which enabled us to validate it upon a subset of 41 instances (i.e. 126 total instances - 85 training instances). This validation enabled us to say that the relative frequency of patients with normal sperm progressive motility before the varicocele embolization is higher in clusters where fewer male patients work in putative hazardous environments and 68.20% of them, were able to get pregnant after the treatment in contrast to the other ones where only 10,5% were able to be successful.

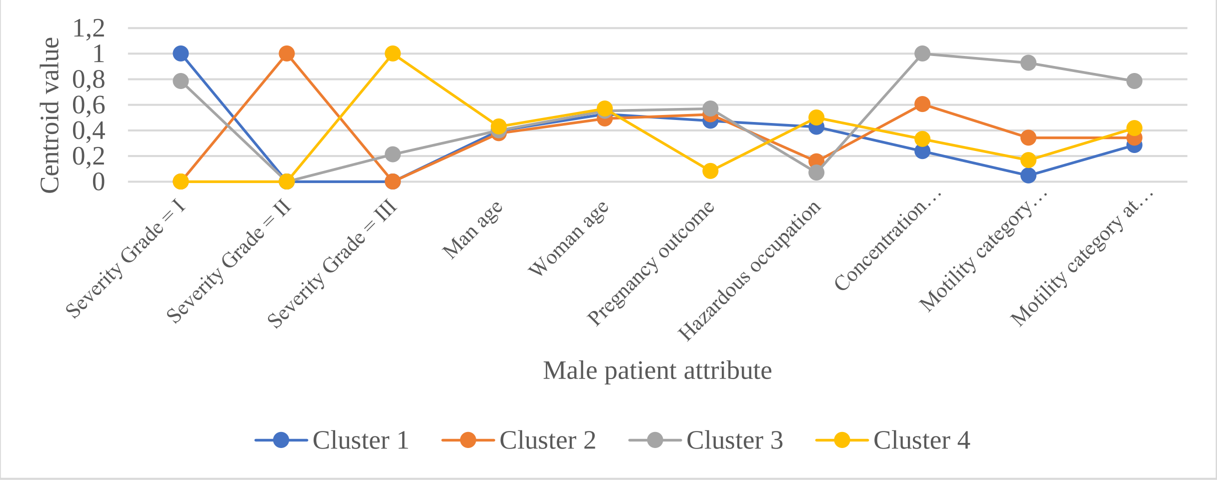


Fig. 2. K-Means computed series plot

Table 8
Centroid table

	Cluster 1 n=21	Cluster 2 n=38	Cluster 3 n=14	Cluster 4 n=12
Severity grade I	1	0	0.786	0
Severity grade II	0	1	0	0
Severity grade III	0	0	0.214	1
Pregnancy outcome	0.476	0.526	0.571	0.083
Hazardous occupation	0.429	0.158	0.071	0.500
Concentration category at 3 months	0.238	0.605	1	0.333
Progressive Motility category before treatment	0.048	0.342	0.929	0.167
Progressive Motility category at 3 months	0.286	0.342	0.786	0.417

3.4Association

After applying the FP-Growth algorithm upon all groups of attributes, we have identified the following association rules as the ones with the highest Support and Confidence value for the "Pregnancy outcome" attribute:

- **Morphology at 3 months > 0% -> Pregnancy outcome=Yes** (n=230)
  support=0.322, confidence=0.544, lift=1.170, Conviction=1.173
- **Concentration category at 3 months = Normal, Progressive motility category at 3 months = Normal -> Pregnancy outcome=Yes** (n=230)
  support=0.157, confidence=0.667, lift=1.433, Conviction=1.604

These association rules indicate that:

- The conditional probability of a woman getting pregnant given a partner with a sperm morphology 3 months after the treatment greater than 0% is of 54.4% (i.e. 74/136) and both situations occur 32.2% of the times.
- The conditional probability of a woman getting pregnant given a partner with a normal sperm concentration and progressive motility 3 months after the treatment is of 67% (i.e. 36/54) and these events occur 15.7% of the time.

Moreover, to seek for more data patterns, we have identified the association rules with the highest Support and Confidence value for all attributes. The following rules were identified:

- **Progressive Motility before treatment>0% -> Morphology at 3 months > 0%** (n=230)
  support=0.517, confidence=0.654, lift=1.106, Conviction=1.181
- **Severity grade=I -> Progressive Motility before treatment>0%** (n=293)
  support=0.198, confidence=0.866, lift=1.153, conviction=1.855

These association rules can be interpreted as:

- The conditional probability of observing 3 months after the embolization treatment a sperm morphology greater than 0% given a sperm progressive motility before the embolization treatment also greater than 0% is of 65.4% (i.e. 119/182) and both situations occur 51.7% of the times;

- The conditional probability of observing a sperm progressive motility before the embolization treatment greater than 0% given a low severity grade of the varicocele condition is of 86.6% (i.e. 58/67) and both situations occur 19.8% of the times.

4. Discussion

From the last KDnuggets poll that inquires the main methodology used in Data Mining projects (Piatetsky, 2014), the CRISP-DM methodology was seen as the mostly used (43%), followed by the data scientist´s own methodology (27.5%). Based on these results, we have selected the CRISP-DM methodology and have seen that it has guided us to a greater knowledge discovery.

Regarding the data understanding and the modeling phases, they were carried out with the RapidMiner platform version 8.1.001. In fact, from the last KDnuggets poll (Piatetsky, 2018a), that has inquired 2025 participants on which Data Science tools they were using, the RapidMiner platform was seen the mostly used rising from 33% in 2017 to 52,7% in 2018. Furthermore, the application fields of the data mining techniques have also been inquired in this same site (Piatetsky, 2018b) and the healthcare domain was in the last poll back in 2017 seen as gaining popularity by moving to the fourth position with a 13% share from the 2016 fifth position with a 12% share which supports the usefulness of this practical guideline for a rising number of data scientists.

The first step that we have done after understanding the research aims, was the assessment of the possibility of applying Data Mining techniques upon the provided data set. To do so, the volume and quality of the data set were assessed.

Regarding data volume, the initial data set had 320 instances and 32 attributes which was at first sight seen as small on the matter of its number of instances. However, after analyzing the literature review in Makris *et al.* (2018) which studies 30 clinical investigations on the varicocele embolization domain, we have seen that the provided data set had an interesting volume of data since related works were in average of 117 patients (± 102 patients). Hence, even the 230 preprocessed instances with non-missing values under the pregnancy outcome attribute - that were the instances mostly analyzed during this study - remained a good volume of data. Regarding the application of data mining techniques, we have seen that it was possible since reviewed related works described in the next section 5 managed similar volume of data and we were able to identify interesting findings.

Data quality was assessed with key data quality dimensions to check if the provided attributes were directly usable to tackle the data mining goals set (i.e. completeness), as well as coherent (i.e. consistency), rightly formatted (i.e. conformity), correct (i.e. accurate with the available information systems) and correctly linked (i.e. integrity) as suggested in (Arkady Maydanchik, 2007). Most attributes were validated/filled/corrected with the available information systems of the CHUC which enabled us to increase it completeness by going in average from a 55.86% to a 70.40% filled dataset. Moreover, during this process we have also looked up for other patient´s information that we could collect, based on the ones studied in related works, which also helped us to potentiate knowledge discovery. In fact, the male patient´s occupation was one of these attributes that was at last seen encompassed by one of our most interesting findings (i.e. second insight identified during the application of the K-Means algorithm).

After preprocessing the provided data set, our main concern was to statistically analyze the preprocessed dataset. Statistical results helped us to overcome some encountered difficulties. In fact, it enabled us to elect a balanced label attribute, as well as select the attributes that were more related with it since the Pearson correlations were all seen as low with the selected label attribute.

Regarding the label attribute, we have selected the pregnancy outcome attribute as in Guh *et al*. (2011) since it delivered the most balanced data set. Concerning the identification of the most statistically significant attributes, we have seen with the ANOVA and the Chi-square test that the following attributes were the most related with the pregnancy outcome: Woman age (ANOVA p=0.018); Severity grade (Chi-square p=0.049); Concentration at 6 months (ANOVA p=0.015); Progressive motility before treatment (ANOVA p=0.018); Morphology at 3 months (ANOVA p=0.004); Concentration category at 3 months (Chi-square p=0.017); Progressive Motility category before treatment (Chi-square p=0.027); Progressive Motility category at 3 months (Chi-square p=0.022); Semen classification before treatment (Chi-square p=0.017); Semen classification at 3 months (Chi-square p=0.018) and Hazardous Occupation (Chi-square p=0.023). As these attributes reveal, several data transformations were carried out upon the provided and preprocessed data set which showed to potentiate knowledge discovery. These data transformations were: dichotomization of the severity grade, normalization of the numeric attributes and transformation of the numerical attributes into different nominal attributes.

To maximize knowledge discovery, we have selected the most commonly applied data mining techniques in the healthcare industry (i.e. classification, clustering and association) with their well tested algorithm based on (Tekieh & Raahemi, 2015), (Ahmad *et al*., 2015) and (Tomar & Agarwal, 2013). Thereby, these data mining techniques were applied with the following algorithms: classification, with the RapidMiner´s Decision tree algorithm and the W-J48 java implementation of the C4.5 algorithm; clustering, with the K-Means algorithm and association rule, with the FP-Growth algorithm.

All these algorithms were mainly trained upon the identified attributes that were seen related with the pregnancy outcome by varying its main parameters. This task was achieved with the “optimized parameter” operator that helped us to automatically loop the several model parameters in order to select the better ones based on its performance measures (i.e. mainly the F-measure along with the Accuracy and the AUC measure). This “optimized parameter” operator was very useful since during our first modeling phase (i.e. modeling step 1 disclosed in section 2.2.6) we struggled to find a decision tree with even 1 level. Hence, when we have sought a solution that could optimize the training process by exhaustively train/test the algorithms, we have found this operator which enabled us to also maximize knowledge discovery.

Since knowledge discovery was difficult with the decision tree algorithm, we have applied the clustering and the association rule technique in an early modeling stage to bring another understanding of the data that could help us on our search for the predictive model. This is the reason behind the order of the exposed modeling step disclosed in 2.2.6, where we see that we have interrupt the decision tree application at its first stage to try other data mining techniques in the modeling step 2 and 3 to further on continue with its training in the modeling step 4.

This modeling strategy was seen successful since the most interesting knowledge discovery was achieved during the K-Means application which gave us the idea to train the decision tree algorithm on this same preprocessed dataset during the modeling step 4 which in turn, also gave us the predictive model we have sought (decision tree of Fig. 1) since it surpassed the threshold defined of 73% based on related works and its findings, in spite of modest, were coherent with clinical expertise. This outcome makes us say that even if the aim of a data mining project is only to predict an outcome, it is always useful to also use descriptive data mining techniques (i.e. clustering and association) to better understand the mined data; and hence, better guide us through a greater modeling strategy.

Furthermore, through this data mining experiment we have seen that due to the small and missing data that we had, it is understandable that it is more achievable to extract interesting knowledge with a K-Means algorithm, that is less influenced by missing data since it seeks to group the data through similarities between data points, than with a decision tree algorithm, that tries to train/test upon missing values; and hence, struggles to select the attribute that promotes the highest gain of information for its decision tree. Hence, due to this experience, we can say that it is important to first identify the most commonly applied data mining techniques in a research domain and then, apply them as a whole, since the different techniques can complement each other and potentiate interesting knowledge discovery.

Regarding the association rule technique, we have found that it is a good technique to identify attributes or relations that are interesting (i.e. mainly with the highest support and/or confidence). This technique clearly depicts one of the advantages of Data Mining, which is to be an inductive technique and not an hypothetic-deductive technique as statistical analysis is. Therefore, it is, in our point of view, an interesting technique to begin with during the data understanding modeling phase, even before inferential statistics, to identify relations or attributes that we might want to assess statistically later on - when we are not able to formulate a hypothesis – or detect interesting data patterns. Furthermore, we have also seen that the generated association rules were useful to complement predictive decision tree findings since we have identified clinically more interesting conclusions with the FP-growth than with the decision tree algorithms.

Regarding the testing of the decision tree algorithms, we have followed a 3-part test design (i.e. 80% for training/testing and 20% for validation). However, most related works as Guh *et al.* (2011) only implement a 2-part test design; i.e. without validation. Based on the performance measures computed at the model´s testing and validation recorded in the previous Table 5 and Table 6, respectively, we can discuss the benefit of the followed test design: Through the several generated decision tree models, we have seen that the only model that has computed an acceptable AUC during the validation was the one computed during the decision tree testing steps 5.1 that had an AUC = 0.750. Nevertheless, we believe that its non-missing values characteristic has also contributed to its acceptable result. In contrast, all other models, which had some missing values, have failed with an AUC going from 0.500 to 0.604 during validation. Moreover, if we would not have tested the models with a 3-part test design, we would not have been able to say that the elected model was stable, but more importantly, that we were not misled by the performance measures obtained during its training/testing. In fact, we had models that had computed acceptable AUC and F-measures during training, but failed during validation (e.g. the decision tree testing step 3.1 had a training/testing AUC=0.747 and a corresponding validation AUC=0.554). Additionally, to further on compare these two test designs, we have run the best-found model (i.e. model corresponding to the 5.1 test) within a 2-part test design of 70% training and 30% testing and seen that its performance measures were slightly increased in comparison to the ones computed at the model´s validation. In fact, its computed performance values were at the 2-part test design the following: Accuracy=80.77%, Precision= 70.00%, Recall=77,78%, F-measure=73.68% and AUC=0.801. Consequently, based on all these aspects, we consider that the 3-part test design is in fact better to follow as suggested by the CRISP-DM methodology (Chapman *et al*., 2000), as well as other data scientists (Deshpande, 2012; Mierswa, 2012) to overcome the seen test overfitting.

5. Related Work

The varicocele condition has been widely covered and assessed with statistical techniques which not only helped us through the election of the information we could collect to potentiate domain related and clinically coherent findings, but also supported the assessment of the interestingness of the computed results. Unfortunately, we were not able to find a study with the application of data mining techniques in the field of the varicocele embolization, nor the varicocele itself, which lead us to seek for works which applies data mining techniques to sperm parameters or infertility data in general. In this section, we disclose the context in which this work was carried out.

By analyzing the identified data mining applications carried out in the infertility domain; i.e. (Sahoo & Kumar, 2014; Bidgoli, Komleh, & Mousavirad, 2015; Gil *et al*., 2012; Guh *et al*., 2011; Chen, Hsu, Cheng, & Li, 2009), we have seen that most of them have used a feature selection technique to potentiate the performance of their models. In fact, some of them have applied several feature selection techniques (Sahoo & Kumar, 2014) such as Support Vector machine (SVM), neural network (NN), evolutionary logic regression (LR), Support Vector machine with particle swarm optimization (SVM+PSO), principal component analysis (PCA), Chi-square test, Student´s T test and correlation or a genetic algorithm (GA) (Guh *et al*., 2011), and others, only have used one technique such as a decision tree algorithm (DT) (Gil *et al*., 2012) or clinical expertise (Chen *et al*., 2009) to select their patient features to mine. Similarly to Sahoo & Kumar (2014) and Guh *et al*. (2011), we have used several feature selection techniques but have focused our choice on the ones mostly used in the healthcare domain; i.e., the statistical ones such as the Pearson correlation, the Chi-square test and the ANOVA test, as well as the decision tree algorithm. In contrast to Guh *et al*. (2011) work, we have not initialy used clinical expertise to pre-select the attributes before the application of the feature selection techniques since it could discard an eventual unexpected interesting patient feature. However, we have pre-selected our collected attributes based on the fulfillment of the requirements of the key data dimensions to ensure that we have a trustworthy data set and result.

Regarding the modeling tasks, we have seen that only two out of the five identified studies have balanced its data set, four out of the five studies have applied the MLP algorithm and three of them, have also applied the SVM algorithm. Although several algorithms were used, the ones that gave the best accuracy were: Support Vector machine (SVM); Particle Swarm Optimization (PSO), Multilayer perceptron (MLP) and Decision Tree (DT). However, they have only used the data mining classification technique which is the biggest difference from our modeling process that has applied several data mining technique; i.e., the classification, the clustering and the association rule technique which has formulated our modeling strategy.

The study that has applied data mining algorithms upon sperm parameter values to predict a treatment outcome - similarly to what we have done - was the work carried out by Guh *et al*. (2011). This work aims to predict IVF success based on couple´s features by following a knowledge discovery process that significantly draws from the CRISP-DM methodology and applying the C4.5 decision tree algorithm, which is similar to ours. However, what mainly differentiates our work from this one is that we have exhaustively trained our decision tree models by varying its main parameters, tested it within a 3-part test design and evaluated it also with the AUC performance measure to potentiate the election of a more stable and useful model.

6. Conclusion

This study has analyzed a data set of 293 varicocele embolized infertile male patients in the CHUC described with 64 patients features (e.g., male patient age, male patient partner age, varicocele severity grade, male patient occupation and sperm parameter values collected before and after the treatment) by using data mining techniques. More precisely, it has predicted its success through the pregnancy outcome with the RapidMiner´s decision tree algorithm and the W-J48 Java implementation of the C4.5 algorithm; and then, identified interesting data patterns with the K-Means and FP-Growth algorithms which guided us through the election of the best models, as well as the discovery of interesting results that suggest that the success of the varicocele embolization could be positively influenced by: a younger male patient partner, a moderate to low varicocele severity grade, a male patient occupation that is not in contact with putative toxic environments or products and a normal sperm progressive motility before the treatment. Furthermore, we have seen that the other assessed male patient´s lifestyle habits such as the drinking and smoking habit or a previous disease or surgery, does not influence the success of the treatment. These findings were seen relevant by clinical experts and contributed to on-going research. However, these results require a greater clinical assessment and discussion with the measurement of their statistical significance which we aim to further on carry out but was not the aim of this present paper which focused on identifying measures that can leverage knowledge discovery on healthcare data sets. In this context, we have shared how the knowledge discovery process was in this work carried out and at last, have identified through its discussion that the following measures had leverage our knowledge discovery process:

- Follow the CRISP-DM methodology;
- Fill/validate the provided data;
- Collect more data;

- Use feature selection techniques that are mostly use in the studied domain;
- Apply the most commonly applied data mining techniques in the studied domain even if the aim is to only predict an outcome;
- Optimize the training of the models when possible;
- Follow a 3-parts test design;
- Not only focus on the performance of the models but also on its interestingness.

All these measures are in our point of view important contributions for further data mining projects in the healthcare field, since healthcare data sets are commonly known to be difficult to mine due to their characteristics.

As future work, we would like to further on explore these identified measures by applying them in other healthcare data sets to at last being able to formulate a practical guideline for the modeling of healthcare data sets, as well as apply other data mining algorithms such as the DBSCAN clustering technique, that is not influenced by outliers, the SVM, PSO and MLP algorithm, that have shown in related works to provide good performance measures, as well as deep learning algorithms to see if we could achieve more interesting findings.

References

Agarwal, A., Mulgund, A., Hamada, A., & Chyatte, M. R. (2015). A unique view on male infertility around the globe. Reproductive Biology and Endocrinology : RB&E, 13, 37. https://doi.org/10.1186/s12958-015-0032-1

Ahmad, P., Qamar, S., Qasim, S., & Rizvi, A. (2015). Techniques of Data Mining In Healthcare: A Review. International Journal of Computer Applications, 120(15), 975–8887. Retrieved from https://pdfs.semanticscholar.org/8228/9448146b86c6160bf5225bd5e3cea35a8c57.pdf

Al-odan, H. A., & Saud, A. A. A. K. (2015). Open Source Data Mining Tools. In 1st International Conference on Electrical and Information Technologies ICEIT'2015 Open (pp. 369–374). https://doi.org/10.1109/EITech.2015.7162956

Almeida, P., Gruenwald, L., & Bernardino, J. (2016). Evaluating Open Source Data Mining Tools for Business. Proceedings of the 5th International Conference on Data Management Technologies and Applications, (Data), 87–94. https://doi.org/10.5220/0005939900870094

Arif, C., Kotoulas, K., Georgellis, C., Frigkas, K., Bantis, A., & Patris, E. (2018). Two Case Reports of varicocele Rupture during Sexual Intercourse and Review of the Literature. Case Reports in Urology, 2018, 1–6. https://doi.org/10.1155/2018/4068174

Arkady Maydanchik. (2007). Data Quality Assessment. Technics Publications, LLC.

Aza Mohammed, & Frank Chinegwundoh. (2009). Testicular varicocele: An Overview. Urologia Internationalis. https://doi.org/10.1159/000218523

Bidgoli, A. A., Komleh, H. E., & Mousavirad, S. (2015). Seminal Quality Prediction using Optimized Artificial Neural Network with Genetic Algorithm. In Conference: 9th International Conference on Electrical and Electronics Engineering(ELECO 2015)At: Bursa, Turkey. https://doi.org/10.1109/ELECO.2015.7394596

Çayan, S., & Akbay, E. (2018). Fate of Recurrent or Persistent varicocele in the Era of Assisted Reproduction Technology: Microsurgical Subinguinal Redo varicocelectomy Versus Observation. Urology, 0(0). https://doi.org/10.1016/j.urology.2018.03.046

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0. CRISP-DM Consortium, 76.

Chen, C.-C., Hsu, C.-C., Cheng, Y.-C., & Li, S.-T. (2009). Knowledge Discovery on In Vitro Fertilization Clinical Data Using Particle Swarm Optimization. Bioinformatics and BioEngineering, 2009. BIBE &#039;09. Ninth IEEE International Conference On, 278–283. https://doi.org/10.1109/BIBE.2009.36

Delavar, M., Haydari, F., Mahdinejad, N., Abedi, S., Shafi, H., & Esmaeilzadeh, S. (2014). Prevalence of varicocele among primary and secondary infertile men: Association with occupation, smoking and drinking alcohol. North American Journal of Medical Sciences, 6(10), 532. https://doi.org/10.4103/1947-2714.143285

Deshpande, B. (2012). How to choose optimal decision tree model parameters in Rapidminer. Retrieved November 22, 2018, from http://www.simafore.com/blog/bid/107076/How-to-choose-optimal-decision-tree-model-parameters-in-Rapidminer

DeWitt, M. E., Greene, D. J., Gill, B., Nyame, Y., Haywood, S., & Sabanegh, E. (2018). Isolated Right varicocele and Incidence of Associated Cancer. Urology, 0(0). https://doi.org/10.1016/j.urology.2018.03.047

Esfandiari, N., Babavalian, M. R., Moghadam, A. M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications, 41(9), 4434–4463. https://doi.org/10.1016/j.eswa.2014.01.011

Gil, D., Girela, J. L., De Juan, J., Gomez-Torres, M. J., & Johnsson, M. (2012). Predicting seminal quality with artificial intelligence methods. Expert Systems with Applications, 39(16), 12564–12573. https://doi.org/10.1016/j.eswa.2012.05.028

Guh, R. S., Wu, T. C. J., & Weng, S. P. (2011). Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes. Expert Systems with Applications, 38(4), 4437–4449. https://doi.org/10.1016/j.eswa.2010.09.112

Han, J., Kamber, M., & Pei, J. (Computer scientist). (2012). Data mining : concepts and techniques. Elsevier/Morgan Kaufmann.

Hand, D. J. (1998). Data Mining: Statistics and More? The American Statistician, 52(2), 112–118. https://doi.org/10.1080/00031305.1998.10480549

Hand, D. J., Blunt, G., Kelly, M. G., & Adams, N. M. (2000). Data Mining for Fun and Profit. Statistical Science, 15(2), 111–131. https://doi.org/10.1214/ss/1009212753

Kirby, E. W., Wiener, L. E., Rajanahally, S., Crowell, K., & Coward, R. M. (2016). Undergoing varicocele repair before assisted reproduction improves pregnancy rate and live birth rate in azoospermic and oligospermic men with a varicocele: a systematic review and meta-analysis. Fertility and Sterility, (August), 3–8. https://doi.org/10.1016/j.fertnstert.2016.07.1093

Lippincott Williams & Wilkins (Ed.). (2012). Medical Dictionary for the Health Professions and Nursing. Julie K. Stegman. Retrieved from https://medical-dictionary.thefreedictionary.com/_/cite.aspx?url=https%3A%2F%2Fmedical-dictionary.thefreedictionary.com%2Fembolization&word=embolization&sources=MillerKeane,wkMed,dorland,MGH_Med,wkHP,gem,evPod,vet,iMedix

Makris, G. C., Efthymiou, E., Little, M., Boardman, P., Anthony, S., Uberoi, R., & Tapping, C. (2018). Safety and effectiveness of the different types of embolic materials for the treatment of testicular varicoceles: a systematic review. The British Journal of Radiology, 20170445. https://doi.org/10.1259/bjr.20170445

Mierswa, I. (2012). Optimize selection, how to get the resulting best model? Retrieved November 23, 2018, from https://community.rapidminer.com/discussion/16851/solved-optimize-selection-how-to-get-the-resulting-best-model

Niederberger, C. (2015). Re: Infertility etiologies are genetically and clinically linked with other diseases in single meta-diseases. Journal of Urology, 194(6), 1712. https://doi.org/10.1016/j.juro.2015.09.045

Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved July 31, 2018, from https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Piatetsky, G. (2018a). Data Science tools poll. Retrieved February 25, 2019, from https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

Piatetsky, G. (2018b). Where Analytics, Data Science, Machine Learning Were Applied: Trends and Analysis. Retrieved June 3, 2019, from https://www.kdnuggets.com/2018/04/poll-analytics-data-science-ml-applied-2017.html

Sahoo, A. J., & Kumar, Y. (2014). Seminal quality prediction using data mining methods. Technology and Health Care, 22(4), 531–545. https://doi.org/10.3233/THC-140816

Samplaski, M. K., Lo, K. C., Grober, E. D., Zini, A., & Jarvi, K. A. (2017). varicocelectomy to upgrade semen quality to allow couples to use less invasive forms of assisted reproductive technology. Fertility and Sterility, 108(4), 609–612. https://doi.org/10.1016/j.fertnstert.2017.07.017

Sharma, R., Singh, S., & Khatri, S. (2016). Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey. In Second International Conference on Computational Intelligence & Communication Technology. https://doi.org/10.1109/CICT.2016.142

Shukla, D. P., Patel, S., & Sen, A. (2014). A Literature Review in Health Informatics Using Data Mining Techniques Keywords Data mining, frequent patterns, data mining techniques, medical data mining. Retrieved from http://cmapspublic2.ihmc.us/rid=1P0PJQH3F-1WTS7CH-273X/Shukla *et al*. - 2014 - A literature review in health informatics using da.pdf

Tekieh, M. H., & Raahemi, B. (2015). Importance of Data Mining in Healthcare. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM ’15, 1057–1062. https://doi.org/10.1145/2808797.2809367

Thatipamula, S. (2013). Data Done Right: 6 Dimensions of Data Quality (Part 1) - Smartbridge. Retrieved July 22, 2018, from https://smartbridge.com/data-done-right-6-dimensions-of-data-quality-part-1/

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241–266. https://doi.org/10.14257/ijbsbt.2013.5.5.25

varicocele Definition. (2002). Retrieved from https://medical-dictionary.thefreedictionary.com/varicocele

Williams, R. S., & Alderman, J. (2001). Predictors of success with the use of donor sperm. American Journal of Obstetrics and Gynecology, 185(2), 332–337. https://doi.org/10.1067/mob.2001.116733

Witten, I. H. (Ian H. ., Frank, E., & Hall, M. A. (Mark A. (2011). Data mining : practical machine learning tools and techniques. Morgan Kaufmann.

WorldHealthOrganization. (2010). WHO laboratory manual for the examination and processing of human semen. World Health Organization (Fifth edit). World Health Organization. https://doi.org/10.1038/aja.2008.57

Yildirim, P. (2015). Association Patterns in Open Data to Explore Ciprofloxacin Adverse Events. Applied Clinical Informatics, 6(4), 728–747. https://doi.org/10.4338/ACI-2015-06-RA-0076