INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




ViDi Stat

VITOR HUGO MAGALHAES MOREIRA
Novembro de 2015

POLITÉCNICO
DO PORTO

EDUCATIONAL DATA MINING PLATFORM

VIDI STAT

Vitor Hugo Magalhães Moreira

DISSERTAÇÃO PARA OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA INFORMÁTICA, ÁREA DE ESPECIALIZAÇÃO EM ARQUITECTURA, SISTEMAS E REDES

Este trabalho é o resultado de um trabalho conjunto realizado por Diogo Fernando Barroso Portela Ferreira da Silva e Vítor Hugo Magalhães Moreira

Advisers: Carlos Gomes Ferreira

Co-Adviser: Ângelo Martins

Porto, Outubro 2015

RESUMO

A tese desenvolvida tem como foco fornecer os meios necessários para extrair conhecimento contidos no histórico académico da instituição transformando a informação em algo simples e de fácil leitura para qualquer utilizador.

Com o progresso da sociedade, as escolas recebem milhares de alunos todos os anos que terão de ser orientados e monitorizados pelos dirigentes das instituições académicas de forma a garantir programas eficientes e adequados para o progresso educacional de todos os alunos.

Atribuir a um docente a responsabilidade de actuar segundo o historial académico dos seus alunos não é plausível uma vez que um aluno consegue produzir milhares de registos para análise.

O paradigma de mineração de dados na educação surge com a necessidade de otimizar os recursos disponíveis expondo conclusões que não se encontram visiveis sem uma análise acentuada e cuidada. Este paradigma expõe de forma clara e sucinta os dados estatísticos analisados por computador oferecendo a possibilidade de melhorar as lacunas na qualidade de ensino das instituições.

Esta dissertação detalha o desenvolvimento de uma ferramente de inteligência de negócio capaz de, através de mineração de dados, analisar e apresentar conclusões pertinentes de forma legível ao utilizador.

Palavras-chave: ferramenta inteligência de negócio, mineração de dados, mineração de dados na educação, estatísticas, qualidade de ensino.

ABSTRACT

This thesis focuses on providing the means necessary to assess the knowledge contained in a school's stored data, transforming chaotic infinite rows of value into something simple, easy to read and available for any user.

As the society progresses and schools receives thousands of new students yearly, Educational leaders strive to make their course's programs effective and adequate for its target audience, and while the students data is stored, its unfeasible to burden the teachers with the task to analyze and act on it.

The paradigm of Data Mining arises from the necessity of optimizing the school's available resources by extracting the conclusions hidden behind layers of data, not only by outlining the system's overview, but also by exposing the statistical analysis that supports its claims, providing the necessary proofs that ultimately allow for the educational system to improve.

This dissertation details the development of a Business Intelligence tool capable of mining the educational program's available data and delivering the pertinent conclusions reached to the user.

Keywords: business intelligence, data mining, educational data mining, statistics, educational quality.

THANKS

I would like to thank everyone who was involved directly or indirectly in the successful development of this dissertation, without whom it would have been impossible:

- Professors Carlos Ferreira and Ângelo Martins for all the guidance, helpful insight and knowledge passed down during the past months;
- My fellow colleague who also worked in this project Diogo Silva for his complicity, support and endless nights developing ViDi Stat;
- My closest friends André Macaco, Carlos Fera, João Bimbi, Fernando Pinto and Renato Bilus for their patience, support and feedback given through this adventure.
- My family for the unconditional love and support;

INDEX

FIGURE INDEX

TABLE INDEX

SYMBOLS AND ACRONYMS

ISEP Instituto Superior de Engenharia do Porto

EDM Educational Data Mining

DM Data Mining

BLL Business Logic Layer

DAL Data Access Layer

SSE Sum of Square Errors

OBDC Open Database Connectivity

1. INTRODUCTION

This chapter serves as an introductory note that contextualizes the theme and methods studied in this dissertation. It englobes the project's scope and background, the motivation and goals behind it, as well as the approach adopted and the document's final structure.

This paper's purpose is to show how much potential this project has, discuss the methodologies that will be used to generalize the knowledge acquired so far, present a viable business solution and assess its impact in the business model.

1.1. SCOPE

At a time when every advantage matters and where everything happens really fast, it is important to have all the technological help available.

This Project's purpose is to offer more insight in order to improve the way things are done in **Polytechnic of Porto - School of Engineering** (**ISEP**), by providing a detailed analysis of the student's academic records, starting from as early as 2008.

Despite the level of expertise the teaching process has achieved over the years, no system's perfect and, as such, it's not only possible but quite plausible that there are flaws to analyze and act upon, or at least sectors that could be more effective if done differently. Plus, it is a known fact that presenting information to a team positively influences their performance.

It's not easy to produce the best course possible while taking in consideration all the variables inherent to a system as complex as an educational one. In this particular case, with thousands of students to keep track of, it's no longer plausible to manually handle the data. Therefore, to support decision making at this level, it's important to generalize the knowledge those records hold.

By applying data mining tools it's possible to identify patterns which then can be interpreted in order to make the necessary changes to the course's program and gradually build on it to achieve something new and more appropriate to its reality. It is also possible to reach out and get actual proof of patterns and tendencies that were not yet formally credited, although suspected.

1.2. BACKGROUND

As the technological world progresses, there is a tremendous increase in the amount of data recorded and stored in digital media, so much so that there is no longer a lack of data, but rather of information to act on it, with organizations storing all of their transactions and records and yet starving for knowledge. The solution to this ongoing problem comes in the form of Data Mining.

Data Mining is the process of knowledge discovery in Databases, outlining implicit, valid and potentially useful patterns and representing them in a language which is natural to the user; it makes sense out of all the chaos.

Educational Data Mining merges as a concerned need for the development of techniques, methods and representations that can explore the unique and increasingly larger databases the educational institutions possess, using the achievable solutions to better understand students and the very settings in which they learn.

1.3. MOTIVATION

Since this project was developed by former students of this particular course, there was a great personal enthusiasm and involvement by the team. One of the main sources of motivation is the will to create something that can help education facilities offer the best conditions possible to their students. Modern societies thrives on the success of students to become the future; the main goal is to point out non-obvious flaws so that the whole community can benefit from a better, well-informed, decision making process.

There's an intrinsic need for a company to display data and statistics in meaningful ways, so that it can be easily observed and conclusions can be drawn upon it (SAS, s.d.) (Chamatkar, 2014). Nowadays, unstructured data makes up for 90% percent of the digital universe, and while being valuable, possessing the data is not the same as possessing the knowledge it holds. Data mining is needed to make sense of all of this chaos, transforming raw data into relevant information that'll lead the way towards more likely outcomes.

Nowadays life in educational systems require the educational leaders to navigate through huge stacks of information, concerning a wide range of topics, from assessments data to demographics concerns, forcing the institution administrators to be data literate and make sense of the resources available when it comes to make informed decisions to improve student performances (David Ronka, 2009) (Borah, 2013) (Popelínský, s.d.). ISEP is no different on this matter as it doesn't have a system that provides value over data, making simple insights, such as representing the students by gender, a troublesome matter. To visually present an overview over the current and past state of affairs would be a great improvement by itself.

In this context, with an undeniable need for a quality and effectiveness improvement of education, Educational Data Mining (EDM) rose and consolidated itself as the field that strives to provide the answers and conclusions hidden in the multitude of heterogeneous data sources compiled by the institutions, in this case ISEP. The data mining methods and tools to be applied are able to draw intrinsic knowledge of both teaching and learning process, thus providing the basis needed to make well informed decisions, making for effective education planning.

An early predicament of student failure, or the discovery of a lack of synergy between two subjects, are examples of situations that, if indicted, can get more attention and promote informed changes.

There is also an interest in exploring the potential of educational mining tools, which use is growing worldwide, thus exploring a growing market with respects to concerns shared by the project's team.

1.4. GOALS

The main goal set for this project is to build an auxiliary platform for a better management of ISEP's learning and orientation programs, while making it flexible and customizable enough to be adapted for other educational systems.

This tool intends to show pertinent information in a user friendly way, with an interactive graphic overview of what's happening, making it not only possible, but a lot easier to analyze the data and the correlations to be found in it.

This project aspires to help improve the student's learning process in each subject, giving them a sense of belonging while motivating them to thrive on a system more responsive, supportive and adequate to their needs.

1.5. APPROACH

The first step of this process was to search through the repository of data of the educational environment and pre-process the information as to what is to be salvaged and what isn't, define the standard formats to use and outline the information structure to use in the search of consistent relationships between variables.
With the structure defined and data pre-processing complete its time to use a set of data mining algorithms that help identify the present relationships. The main analysis performed include association rule mining, clustering and social network analysis.
The discoveries found must be validated to avoid overfitting, eliminate noise and anomalies, providing a level of robustness and optimization to the solutions, with known patterns and tendencies used as training examples during this phase. The attainable solutions will then present multiple levels of meaningful hierarchy that can't be determined in advance, but rather by the data properties themselves.
The identified and validated solutions can then be applied to make predictions on future events and behavior on the learning environment studied, thus making this data visualization an important tool to assess an overview of the system and the causes behind the end results. These predictions and explanations are then used to support decision-making processes and policy decisions by the educational leaders.

The user's findings can be downloaded to a local device, with the available data being consulted by the means of a user-friendly interface designed for this very purpose, with visual aids, markers and descriptions guiding the user.

1.6. STRUCTURE

This document is structured as follows: Section 1 introduces the motivations and goals that led to this work; Section 2 is where the state of the art is described; Section 3 describes the business analysis study made for this problem; Section 4 details the experiments made and discusses the results found; Section 5 refers details the requirements, architecture, design and implementation of ViDi Stat; Section 6 discusses the work schedule; Section 7 elaborates on the conclusions; Section 8 contains the references and section 9 contains the appendix.

2. STATE OF THE ART

This project is to deal with a particular set of technologies and themes in order to come to terms with what is proposed.

The main theme focused by this project is Business Intelligence which is a set of techniques and tools that allow users to transform raw data into meaningful and useful information. Business analysis has grown up in the last years and it will keep growing towards the future.

Business Intelligence focus on handling large amounts of data to help identifying, developing and creating new strategic business opportunities. It keeps simple and easy to interpret big chunks of data where users can effectively create and implement business strategies based on the insights given by the outputs that a BI application can provide.

Nowadays Business Intelligence has a great impact in enterprise solutions, it is applied in various business purposes that help monitoring and keeping up with the business pace through, like:

- Analytics, builds quantitative processes to achieve optimal decisions
- Measurement, a hierarchy of performance metrics and benchmarking which supports the progress made for the goals established
- Reporting, to visualize data and other valuable information
- Collaboration platform, software that shares data from both inside and outside the business
- Knowledge management, it helps the companies to “drive” their data for a better management and understanding of the business as a whole.

Business Intelligence plays a big cut by adding the possibility to transform large amounts of raw data about students into valuable and easily manipulated data ready to be incorporated in an application which will portray it to the end user in a friendly and easy to read interface. This development was made possible with the integration of a database to store the valuable data and the presentation means provided by the Google Charts API. The user friendly interface mentioned is a dashboard.

A dashboard is a data visualization tool that displays the status and key performance indicators for a given business logic. They may be customized for a specific role and display metrics targeted to a single point of view, filter or parameter. One of the most valuable assets of a well implemented dashboard is the possibility to be fully customizable and upgradable to work with a complex system in the background while providing valuable stats that sum up the system’s state and history.

For the project being developed, the dashboard will have a direct communication with the database and data mining software which will allow the user to interact with the system and freely perform searches, avoiding pre-defined queries.

2.1. Development Process

2.1.1. Data Mining Development

The current standard model used for a Data Mining process is the Cross Industry Standard Process (CRISP-DM) (Anon., 2015) (Pete Chapman, s.d.), which is a comprehensive methodology that provides a complete blueprint for conducting a project. CRISP-DM cycle's composed by six phases as seen in Figure 1.

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

FIGURE 1 - CRISP-DM CYCLE

2.1.1.1. BUSINESS UNDERSTANDING

The **Business Understanding** phase purpose is to understand the project's goals and requirements, define the needs and expectations, assess the current situation, and identify available technologies and knowledge sources, determine mining goals and produce a plan for the project.

- **Goal –**To extract, outline and identify the hidden patterns and correlations present in the source data. To get value over data.
- **Requirements –**To understand the business logic and to filter the right information to be treated. A data warehouse to store said data.

The data mining tool adopted was Weka, a free and open-source application that provides of the necessary functions for this project. The working plan is present in detail in a later topic.

2.1.1.2. DATA UNDERSTANDING

Data Understanding is when the data is collected and explored, with its gross properties analyzed and described, and their value and quality assessed. This implies extensive exploration and understanding of the raw data. This process is similar with the one described in the 4.3.1 topic.

2.1.1.3. DATA PREPARATION

The third phase, **Data Preparation**, is usually the most extensive one, covering all the necessary activities to construct the final dataset: selection (define selection criteria, explain the inclusion or exclusion of certain data, collect additional info, etc.), cleaning (decide how to deal with missing values, aggregation level, outliers, etc.), construction (derive attributes, decide how the missing attributes can be constructed or imputed, etc.), integration (integrate the sources and store the results on new tables and records), and data formation (rearrange attributes order, reorder the records, etc.). This process is similar to the one described in the 4.3.2 topic.

2.1.1.4. MODELING

Modeling is the phase when the modeling technique is selected, a test design is designed (with a plan intended for training and the dataset divided into training, test and validations sets), and a model is built, described, explained and assessed, interpreting the results in business terms with feedback from domain experts.

2.1.1.5. EVALUATION

The **Evaluation** of the model consists on understanding the data mining results and appreciating their impact on the given mining goals; Reviewing the process, summarizing it and identifying positive and negative actions taken, with respect to the business success criteria; Determine possible next steps, estimate their potential improvements, refine the process plan and/or recommend alternative continuations; And deciding how to process to the next stage, documenting the reasons of said choice.

2.1.1.5.1. DEPLOYMENT

Deployment involves a set of tasks: Determine how the results will be utilized and the knowledge presented in a meaningful way to the customer; Define a monitoring and maintenance plan, referring if the business objectives are likely to change over time and what would happen if the model no longer applied (new project, update current model, etc.); Outline structure and contents of a final report, select the adequate findings to include in it and write it; Review the project, collecting feedback from the end-user, analyzing the process as a whole and documenting it.

2.1.2. SOFTWARE DEVELOPMENT PROCESS

Given the nature of this project, not all of the requirements and technologies to be used were known in earlier stages, conceiving the need to adopt an iterative and dynamic program development methodology, RUP (Rational Unified Process) (Anon., s.d.) (Anon., s.d.) (Rouse, s.d.). This is not a concrete process, but rather an adaptive one, tailored by the elements necessary at a given point, fostering the idea of a progressive and constructive approach to the problem, which is essential for the incremental development of a solution where not all of the requirements are yet known.

RUP has four major phases as shown in Figure **2**:

- Inception
- Elaboration
- Construction
- Transition

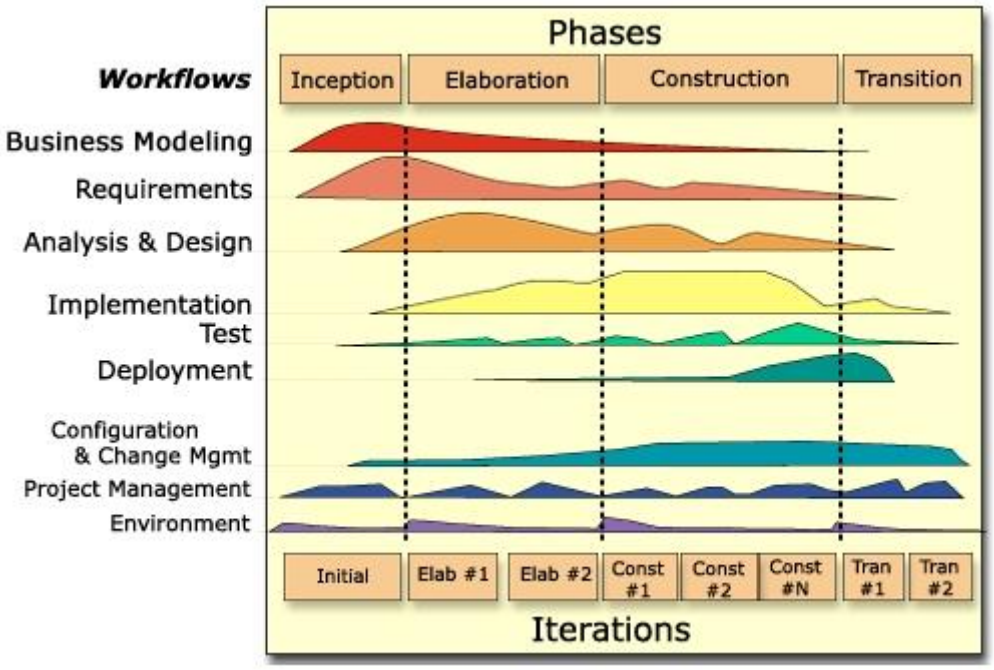

FIGURE 2 - RUP'S PROCESS ARCHITECTURE

2.1.2.1. INCEPTION

The inception phase establishes the system's business case, as well as its scope. To do this, it's necessary to identify the interacting actors and the nature of their interaction (at a high-level). This usually also implies UML diagrams describing the most significant use cases.

2.1.2.2. ELABORATION

The second phase is when a prototype, that addresses the critical use cases defined, is built. Elaboration also identifies the major technical issues of the problem, and further completes the diagrams.

2.1.2.3. CONSTRUCTION

During the construction phase, the remaining components and application features are developed and integrated in the product, which is then thoroughly tested. At the end of this phase, the product should be rightfully documented, integrated in the adequate platforms and ready to be experienced by the hands of the end-users.

2.1.2.4. TRANSITION

The last phase focus on the activities required for the transition of the solution to the user community and the adequate correction of the issues that will arise from their use leading to new, more robust releases.

2.2. RELATED WORK

Data-driven needs aren't new in the market and plenty of projects came up with specific solutions before. By analyzing previous works on the field, it is possible to have a better understanding of the possibilities at hand and how previous works can suggest the best suited development ideas.

Tableau Desktop has an academic program specially for students to develop data mining skills for free, thus specializing more and more professionals in this particular area (Tableau, 2015) (Tableau, 2015).

In 2011 Jaspersoft launched the Jaspersoft Scholars Program, which allows academic institutions to apply for a Jaspersoft BI Suit license free of charge and apply it in their educational curriculum, allowing higher education institutions to access online knowledge and training resources, thus giving the students modern tools and adequate resources to explore educational data mining technologies (Jaspersoft, 2011).

In 2009, Columbus city schools and Nationwide Mutual Insurance worked together to help district educators with the WebFocus tools to analyze student data in an effort to reform the school system, improving student scores, graduation rates, performance and overall

achievements in core subjects (McGee, 2009).This collaboration made the schools more efficient and effective, making for a more proactive rather than reactive system.

There wasn't a lack of data but there was a lack of information to act on it and this collaboration's provided the tools that enabled the educators to create graphical and other reports representing an analysis of student that made for a model of future behavior based on current performance. By assessing student progress and identifying its weaknesses, it was possible to keep track of their responses to the program and intervene when necessary.

An American Midwest grocery store chain which analyzed their sales and discovered that young males who bought diapers also have a predisposition to buy beer (Whitehorn, 2006). After uncovering this, it was easy to extrapolate from the effect to the cause:

- Young males like carousing with friends on the weekends and that involves lots of alcohol, usually beer.
- Most young males only buy diapers after becoming fathers, and acquiring offspring is a known carousing inhibitor.

A newly father goes shopping knowing fully well he is not going to be able to join his friends over the weekend, as he usually did, but there is nothing stopping him from enjoying a beer in the comfort of his house. The retailer then concluded that all he needed to do was to remind the father of that fact and, by moving the beer display closer to the diaper display, he could increase his sales and revenue.

There are free online courses that provide the necessary means to explore educational data mining techniques using R software, thus promoting its use in the area with communities, documentation and example made available for this purpose (Zhao, 2015)

Qlik launched an academic program that offers it's potential to students and teachers that meet their required criteria, thus allowing these individuals to further explore and develop their analytical skills (Qlik, 2015).

A study developed at University of Minho (Silva, n.d.)shows that by using data mining techniques such as Decision trees and Social Networks, it is possible to possible to predict student performance with a good level of accuracy, underlining not only their past grades as

decisive factors, although they proved to me to most important one, but also other relevant features such as number of absences, alcohol consumption or parent's job and education. Although these methods were only applied after the data was collected, the results clearly point that there's more to student performance than meets the eye and enforce the potential such a tool could have, should it be used as part of school management support system.

In 2003, a study in performance prediction (Behrouz Minaei-Bidgoli, 2003) showed it is possible to identify classes of students who use the available resources in similar ways, making it possible for students to improve by sharing their user experience with their fellow colleagues in the same group. This classification also allowed the teachers to devise learning strategies more effectively and efficiently, enforcing the need for clustering and classification techniques to be applied to school systems in a timely manner.

The case study developed at INESC in 2014 (Pedro Strecht, s.d.) set a precedent by using decision trees to build interpretable models able to predict the students in danger of failing. These models can not only let to strategies being devised to prevent the failure, they can also point to some of the reasons that lead to that outcome, offering the necessary tools for a better, more cost-efficient, school system.

2.1.2. Educational Data Mining Software

The following list contains a set of business intelligence software with a lot of different potential and utilities. It is possible to take notice that some of the software detailed above has less user-friendly interface versus more technical tasks and capabilities like Weka, detailed further.
Each of the applications mentioned in this section are have case studies and were applied in EDM, but can also be applied to other business logics.

2.1.2.1. SAS

SAS Enterprise Miner is pattern discovery software to able to perform data analysis, driving a better decision making by the means of descriptive and predictive modelling, thus providing value for the data (SAS, s.d.) (Taylor, 2011) (SAS, s.d.).

SAS's software key benefits are:

- **Modelling –** The modeling tools available shorten the model development time, as well provide an interactive and self-documenting environment that is capable of mapping the entire data mining process.
- **GUI –** SAS's software GUI was designed to be easy to use, guiding business users with little to no statistical skills through the workflow of the data mining tasks they intend to use, thus allowing them to generate their own models which are then displayed in easy-to-read charts, granting the insights needed for a better decision making process.
- **Accuracy –** By using innovative algorithms and industry specific methods, it's possible to use validation metrics to assess the results, as well as easily compare statistics and predictions built with different approaches.

SAS has an industry solution specific to higher education analysis so that qualified educational institutions can meet their needs for data management, analytics, data visualization and reporting (Anon., s.d.). The key features for this solution are:

- **Efficiency –** SAS is a fully integrated system, boosting productivity and efficiency by assuring the current data is reliable and ready for analysis.
- **Simple –** The analytics properties offered do not require the user to know any code, it's all contained into easy-to-use features, making complex data mining and reading simple.
- **Design –** The reports created are interactive, attractive, meaningful and easy to share via web, Microsoft applications or mobile devices. The capabilities of the reporting tool allow the users to fully explore the data on their own terms, making for a very appealing and complete software design.
- **Familiarity –** The obtained results can be exported to environments more familiar with the end users, such as Microsoft Office Applications. This guided mining approach makes it so that even novice users are able to quickly conduct complex queries and share/embed the outputs on well-known formats.

Still on the education level, this company is also known to be partners with Teradata University Network in order to bring their SAS Visual Analytics and SAS Visual Statistics to colleges, for learning purposes (SAS, 2015).

2.1.2.2. WEKA

Weka is a data mining open source application written with the objected oriented language Java, by the University of Waikato in New Zealand, standing for Waikato Environment for Knowledge Analysis (WEKA) (Mark Hall, 2009) (Abernethy, 2010). This tool provides a collection of machine learning algorithms that can be used to draw knowledge from a set of stats, using modules of data preprocessing, classification, clustering, and association rule extraction (Anon., s.d.).

This application contains a GUI for the user to interact with the source files, importing, treating and mining at will, with the Graphical component producing visual results by the means of formatted text, lines and tables, as well as a CLI component, recommended for in-depth usage, with more advanced functionalities, not yet available via the GUI. Weka has a general API so it can be embed with other applications, making it possible to have it performing server-side data-mining tasks (Wan Aezwani Wan Abu Bakar, s.d.).

2.1.2.3. GOOGLECHARTS

Google Charts offers powerful dashboards features, from line charts to complex hierarchical tree maps, which make for a smooth presentation of data on a website (Anon., s.d.). The charts provided by this tool are customizable and interactive, allowing for the creation of complex dashboards as the events exposed are handled, representing a good and reliable source to integrate with educational data mining software (Géryk, 2015).

The platform shown in Figure 3 can be integrated into a project by the importation of its libraries and the presence of JavaScript embedded on the webpage. The appearance from the chart types available can be customized to better fit the needs of a given context, so that different contexts can have a unique look and feel to the user.

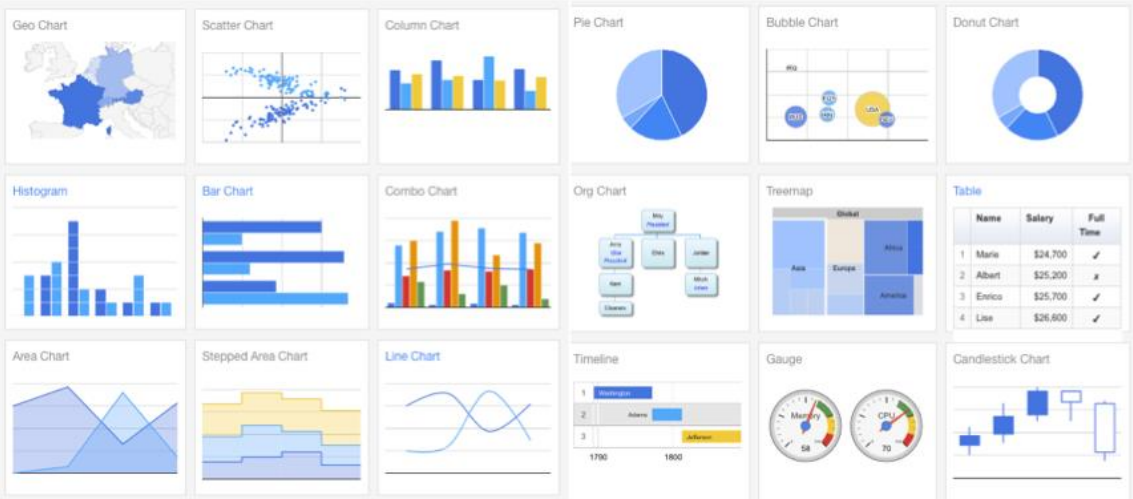

FIGURE 3 GOOGLE CHARTS EXAMPLE

The charts are rendered using HTML5/SVG technology which provide cross-browser compatibility and cross portability compatibility to iPhones, iPads and Android, so that a user will never have to have additional work to successfully interact with the application: It is available in mostly any platform, the only real requirement is a web browser.

2.1.2.4. JASPERSOFT

Jaspersoft is a commercial open source software vendor focused on business intelligence, including data visualization, reporting, and analytics (Jaspersoft, 2011) (Anon., s.d.) (Anon., s.d.). Jaspersoft provides reports, dashboards and analytics easy to use, through a flexible, web-based architecture which can be embedded into other applications.

The platform presented in Figure 4 is easy-to-read and highly interactive, being able to report information from multiple data sources at once. The reports shown are also interactive and capable of being dynamically filtered, sorted, formatted, etc, and stored for further reuse. It is also possible to share the information efficiently by the means of an embedded view that allows for an interactive report to be distributed by mail or inside web applications.

FIGURE 4 JASPERSOFT INTERFACE

As for as input and output is concerned, Jaspersoft is very versatile can create reports from any data source including big data, relational and non-relational datasets, as well as build and publish the reports in the current most widely used formats.

2.1.2.5. QLIK

Qlik provides a platform-based approach that offers valuable insights to the decision making process, having already been applied to the educational data mining world (Senior Technology Advisor, 2009) (Qlik, 2015) (Anon., 2015) (Qlik, 2015) (Cronstrom, 2014). As far as Business Intelligence goes, this applications splits into two different products: QlikView (Traxion Consulting, 2011), which is a tool for situations where prepared business applications are needed as seen in Figure 5, with developers working out the data model, layout, charts and formulas, and delivering the application for the end-users to consume, and Qlik Sense, a tool for situations where the users are more engaged and has the freedom to create a layout on his own, with new visualizations and charts to fit their needs as seen in Figure 6.

FIGURE 5 - QLIK VIEW INTERFACE

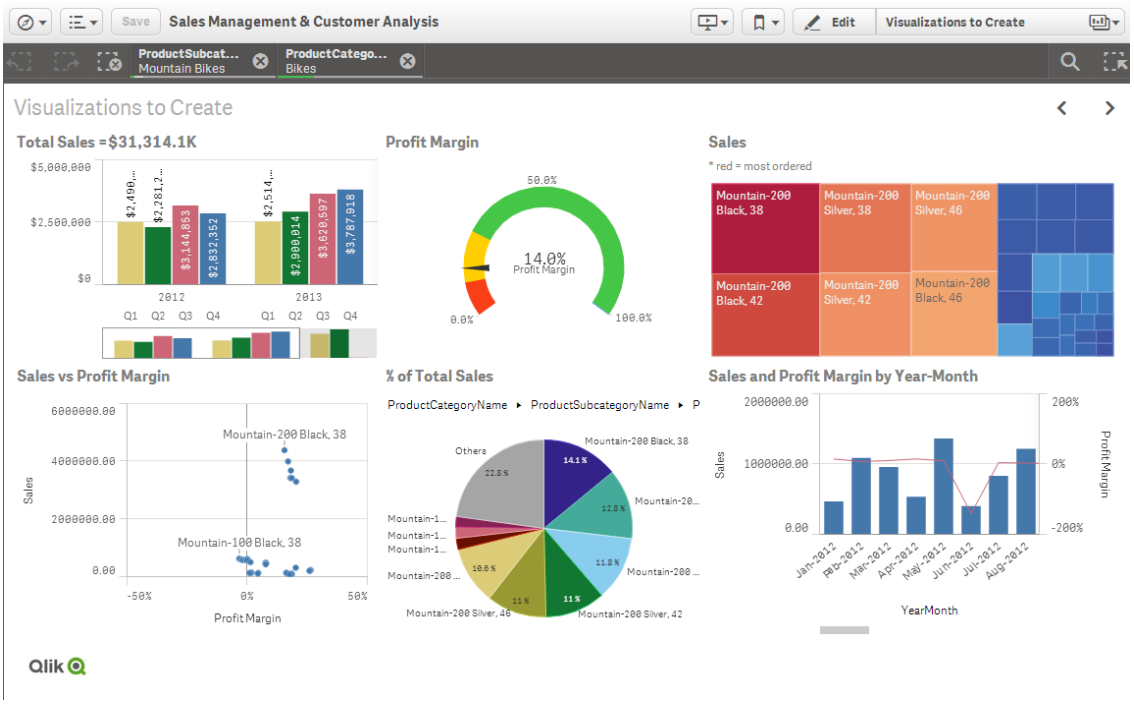

FIGURE 6 - QLIK SENSE INTERFACE

Both products offer total freedom to explore, select, drill down and navigate through the data, with Qlik View offering a guided analytics posture, while Qlik View goes in line with a self-service data discovery posture. This application is used to improve tracking and reporting of student registrations, examination performance and workforce effectiveness. Its indicators are used to adjust operations to market changes and resource requirements, deliver an improved marketing return on a given investment and reduce operational costs through improved reports and efficiency. When applied to a school system, Qlik is capable of:

- Increasing insight of student's demands, records and behaviours.
- Ease the administrative burden associated with data analysis and reporting
- Optimize costs by providing improved visibility on investments
- Improve management strategies by providing analysis on planning and effectiveness
- Increase operational efficiency

2.1.2.6. WEBFOCUS

Information Builders developed WebFocus (White, s.d.) (Variar, 2005), a tool to manage, package and deliver information to internal and external users. The information provided by this application is carefully chosen to fit the best approach for the needs of different kinds of end users as shown in Figure 7.

FIGURE 7 WEBFOCUS INTERFACE

The key points of this software to take into consideration are their high level view of the critical indicators through dashboards and scorecards, the possibility to interact with the data on any device, whether connected to the internet or not, its ease of use, analyse and manipulate data with no previous training required, its integration with desktop formats familiar to the user and its dynamic report distribution that fires real-time alerts, as well as automates the scheduling and delivery of vital information.

2.1.2.7. INFOESCOLAS STATISTICS

Infoescolas is website (Infoescolas, 2015) that grants access to a tool built by the Portuguese government which aids in the search of data regarding grade statistics of schools all over the country, as well as other valuable information about superior education, according to a set of preconditions established by the user using the platform.

The website is able to filter data from various sources delivering a friendly user end interface result that can be read and interpreted easily. It is possible to search for a specific school by using District and County as parameters, selected beforehand by the means of a couple of dropdown lists, which will then returns all the results that fit the criteria.

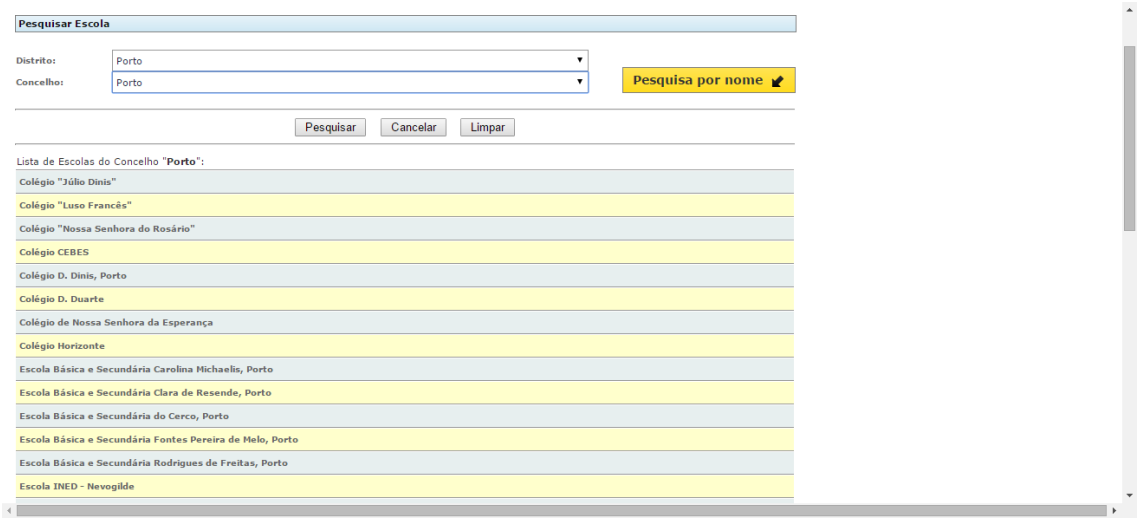

FIGURE 8 INFOESCOLAS SCHOOL FILTER

By looking at Figure 8 it is possible to see a list of schools that were presented for Porto District and Porto County. All the schools that are available and fully operational in this search, can now be drilled down to analyze a set of data as the next Figure demonstrates.

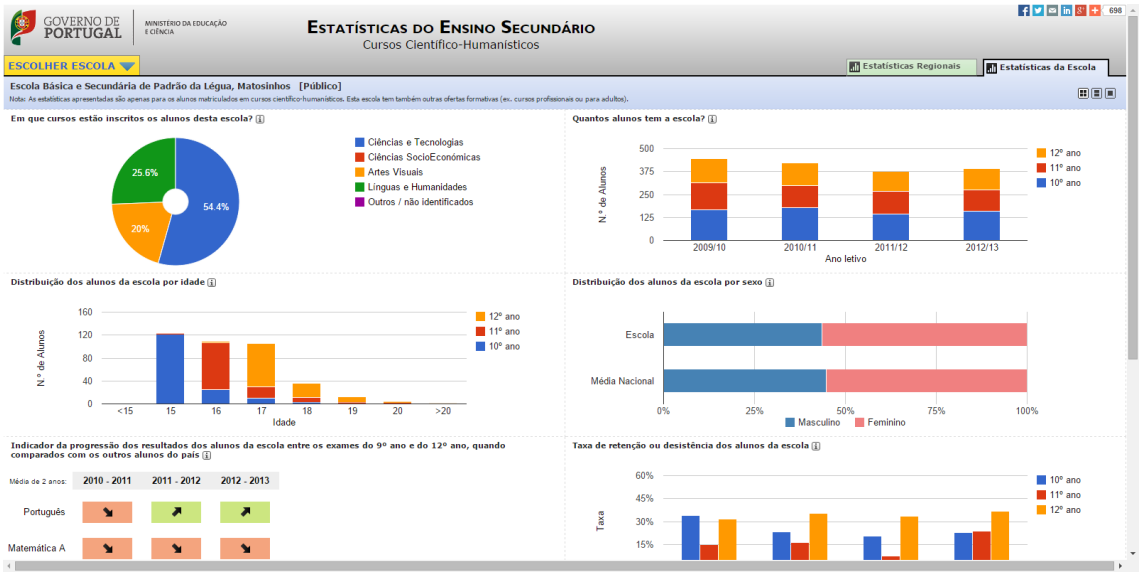

FIGURE 9 INFOESCOLAS DASHBOARD

The information in Figure 9 is useful to an extent but fails to be a reasonable option for a school management standard because the value over data was not fully explored, mostly organizing the students by their most generic features, such as age, gender and course chosen. This tool is very useful for a quick overview, however it has one huge limitation: The statistics and boards displayed are static, it's not possible to successfully interact with the data or mine it any further to gain additional knowledge, flesh out a specific trait or detect underlying patterns, the visual display provided is all there is to it.

2.1.2.8. TABLEAU DESKTOP

Tableau is a market leader in business intelligence and analytical software (Anon., s.d.), specialized in interactive data visualization, allowing their users to easily connect with data and create interactive, sharable dashboards (Tableau, s.d.) (Business Week, 2015).

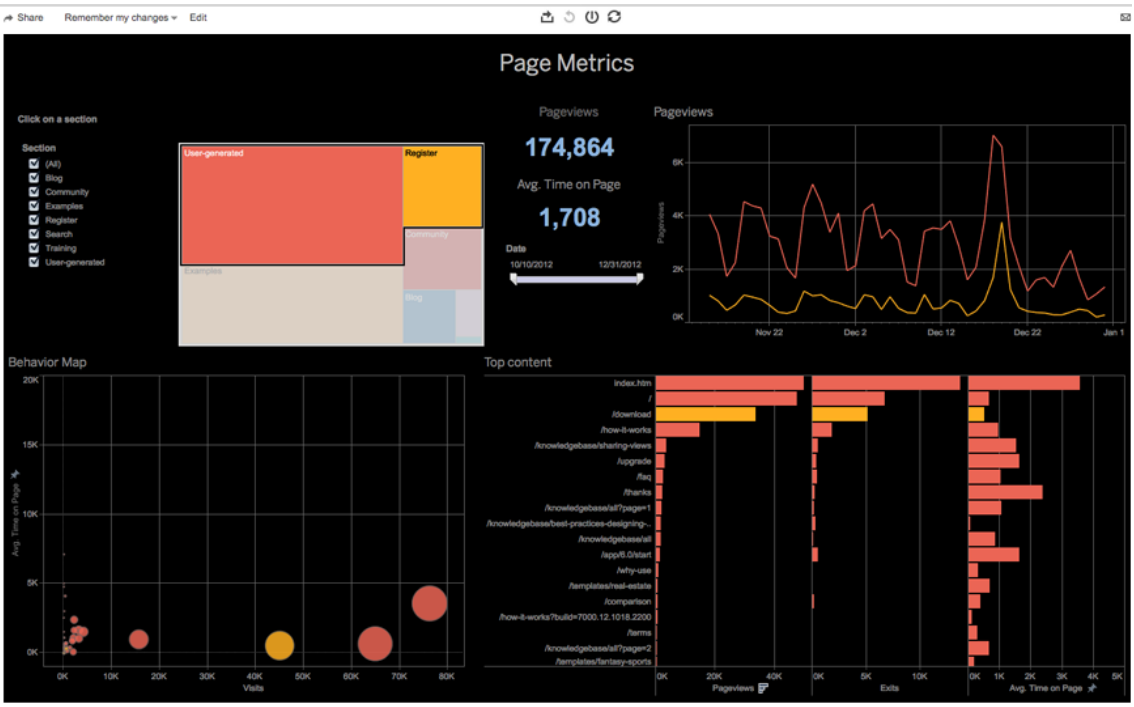

FIGURE 10 - TABLEAU DASHBOARD

The company has developed a range of products to fit their customer's needs:

- **Tableau Desktop –** Easy to use, self-service analytics environment that provides access and insights over data.
- **Tableau Server –** Browser and mobile based platform for data management and scalability to easily share the Tableau Desktop dashboards throughout the organization, promoting enhanced decision-making.
- **Tableau Online –** Cloud-based hosted version of Tableau Server, providing worldwide access.
- **Tableau Mobile –** Mobile version of the platform, optimized for touch technology.
- **Tableau Public –** Cloud-based service that allows general public to interact with data, create interactive visualizations of it and publish it directly to their websites.
- **Tableau Reader –** Desktop application to filter, interact and view author's data details of the visualizations created by Tableau Desktop.

This line of products offers the tools to create rich analyses and share their insights with anyone and anywhere, quickly and efficiently. The provided dashboards that can combine multiple views, enhancing their value, allied with its users not being required to have advanced computation skills to interact with the platform, made its usage intuitive and prone to be adopt.

2.3. Business Intelligence Software

The software being developed may fit in a business intelligence software with great possibilities to become more versatile and complete. The tools presented can offer a deeper look into what can be achieved through data mining software allied with good presentation skills.

2.3.1. R Data Mining

R is a language and environment for statistical computing, used as an integrated suite of software facilities for data manipulation, calculation and graphical display (Anon., s.d.). This software is highly extensible and provides a wide variety of statistical modelling useful for data mining, such as clustering and classification.
The key features of R include a data handling and storage facility, a suite of operators to perform calculations on arrays, an integrated collection of tools for data analysis, a simple enough and well developed programming language (R) capable of withstanding conditionals, cycles and input/output facilities, as well as a graphical component for data analysis and it's display or results exportation. This software's popularity has increased substantially in recent years with the increasing data mining needs by companies (Smith, 2012) (Karl Rexer, 2011).

2.3.2. Yellowfin

Yellowfin (Yellowfin, s.d.) is global business intelligence and analytics software, leading provider of reporting and analytics solutions to the education system, which offers a malleable and agile web-based solution (Yellowfin, 2014) (Council, 2014).

This application show in Figure 11 owes its success to a fine balance between its ease of use for business users and the governance needs of the company's IT department. Other key points of this software are its data discovery module, providing some tools to interact with or combine different datasets, a collaborative BI, allowing its users to share data among themselves for a better, faster, more informed decision making process, and a highly mobile and adaptable system, available in any platform

FIGURE 11 YELLOWFIN INTERFACE

2.3.3. CLARITY

Clarity is a business intelligence and decision support platform written for schools, allowing researchers and statisticians to analyse datasets, giving educational leaders the information required to make informed decisions to drive learning outcomes to good port (Anon., s.d.). This system is currently used by thousands of schools, impacting millions of students worldwide; about one in seven U.S. schools uses this Clarity (Matheson, 2015), which is on its way to becoming an industry standard.

This software combines the data collected from students, teachers, school and staff to create action plans for implementing technologies and strategies, allowing the administrators to make better informed decisions regarding where to direct their funding, thus making the system more financially efficient.

The Clarity platform relies on a combination of human expertise and computation techniques. The data worked on is poured into the platform by data scientists and researchers, in the form of multiple reports, case studies and papers, surveys and questionnaires, completed in the school being analysed, and third-party-data regarding the school's proficiency in certain areas.

Clarity is then able to identify what works and what doesn't, pointing out the solutions which led to better student performance.

The benefit of Clarity is making value of data, pointing the schools in the right direction, while saving their administrators the burden and time necessary to conduct their own researchers and/or the costs associated with hiring external consultants.

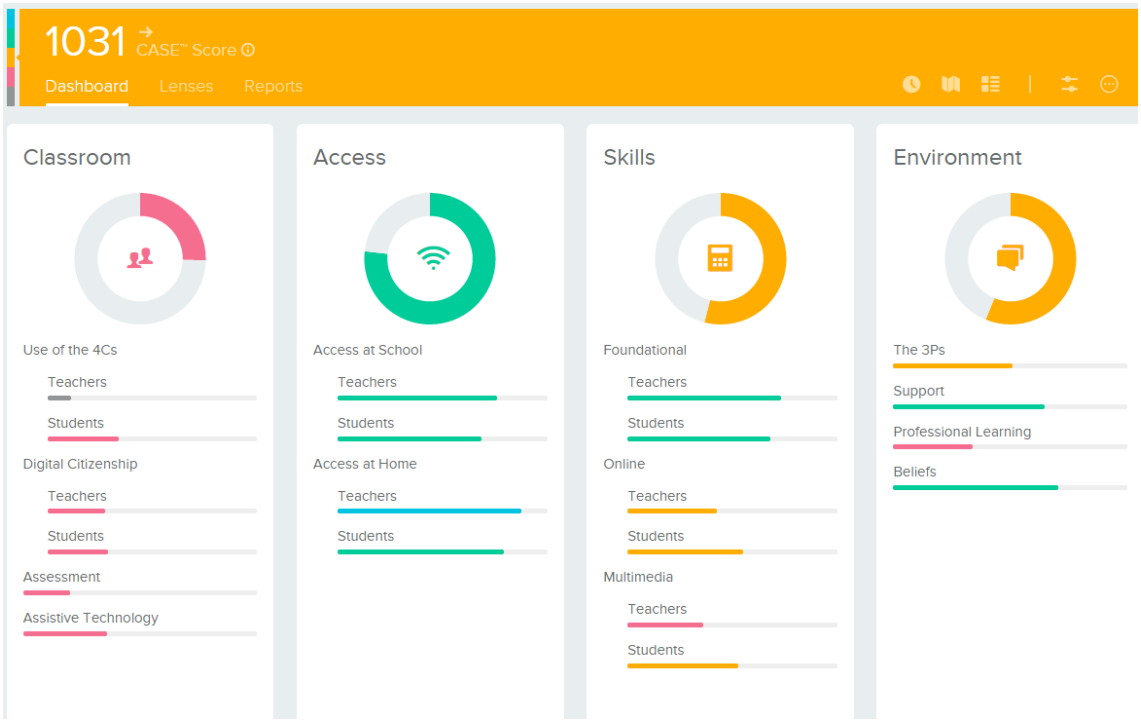

FIGURE 12 CLARITY DASHBOARDS

The value over data mined by this platform is presented with dashboards as shown in Figure 12, displaying the key factors the educational leaders should take into account, such as at-risk students, statistics and pointers about the like hood of a student dropping out, the best practices to adopt while dealing with a set of students, etc, thus seamlessly granting research findings and conclusions for a particular school's environment.

Clarity also features a set of tools as seen in Figure 13 so that the dashboards can be dynamically organized and filtered to standards that best suit the user's needs.

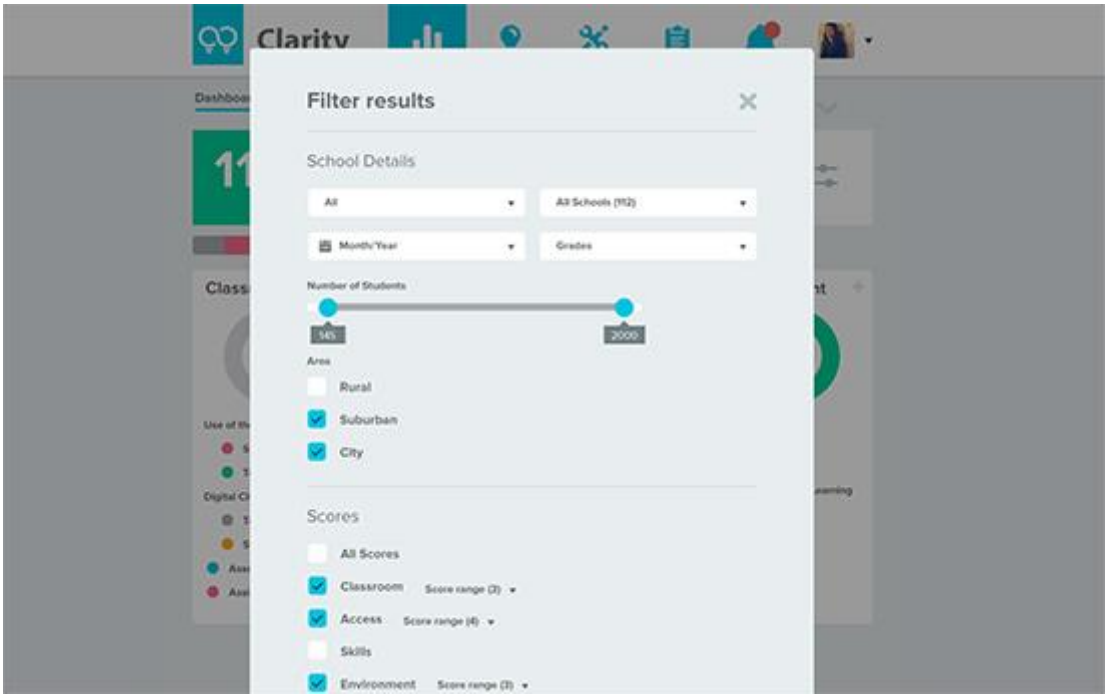

FIGURE 13 CLARITY FILTERS

2.3.4. PENTAHO

Pentaho is a BI platform written in java that addresses the barriers that keep a company from getting value from their data by simplifying, preparing and blending datasets while offering the users a spectrum of tools to analyze, visualize, predict future behavior, explore, and build reports on it (Pentaho, s.d.) (Moody, 2010) (Morgner, 2010). With this system, any member of the team, from developers to regular users, can quickly access a user-friendly system overview on a set of dashboards with pertinent information, transforming data into value.

This applications presents the data by using interactive intuitive dashboards combining a wide range of displays, colors and symbols as displayed in Figure 14.

The Pentaho platform uses a combination of rules, services, assured messaging, workflow, clustering and auditing to provide improved performance by performing tasks such as reducing spam, freeing up memory occupied by excluded useless data, reducing overhead of redundant reports or successfully handling each exception (Anon., s.d.) (Anon., 2006).

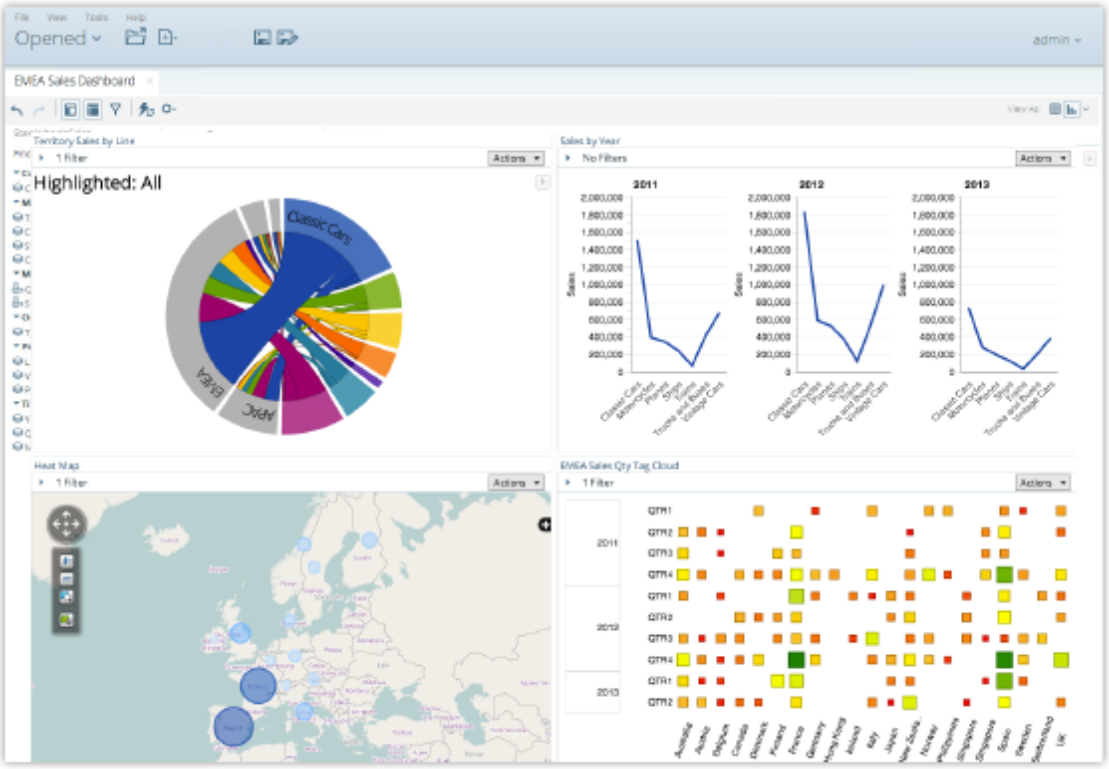

FIGURE 14 PENTAHO INTERFACE

Pentaho has established itself as a leading data integration and business analytics Company with an open-source based platform that provides native integration to all big data sources, regardless of analytics requirements or deployment environment. It was also the first major BI vendor to introduce big data capabilities and is currently leading the charge in big data analytics and integration.

2.3.5. WIZDEE

Wizdee is a business intelligence software (Wizdee, s.d.) very similar to the one this project aims to develop, providing an interface that allows the customers to make queries through a language that is more natural to them and displaying the results on an interactive dashboard that fits their needs. The Neuro-linguistic programming (NLP) component indexed to this tool provides support to query the database with natural language, whether it is vocal or textual.

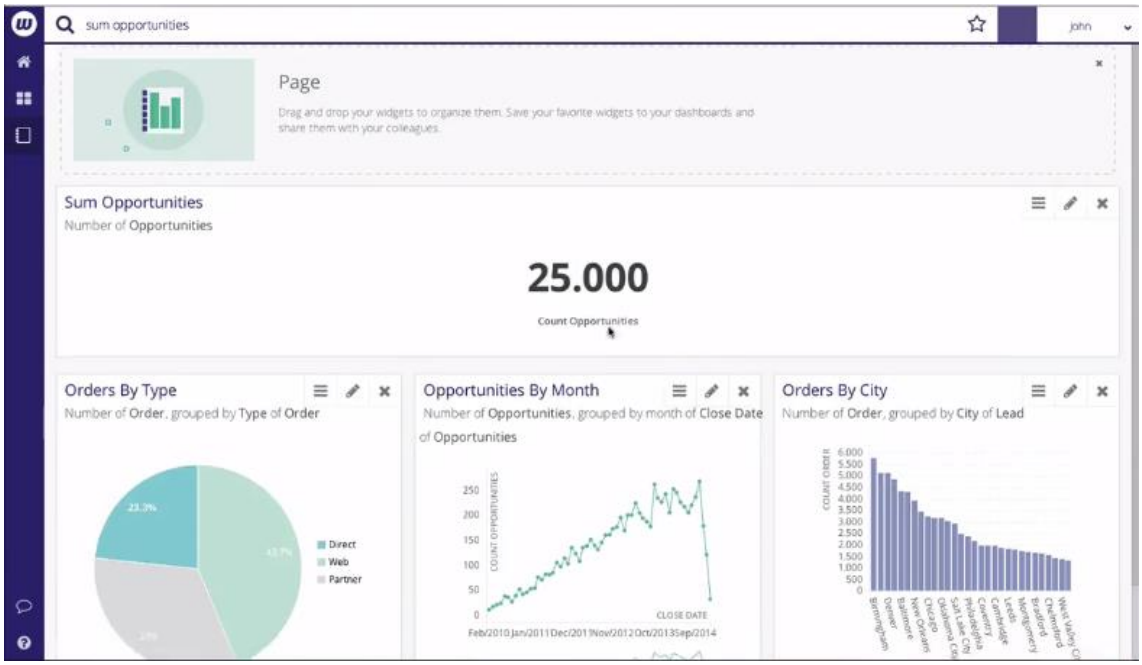

FIGURE 15 WIZDEE QUERY EXAMPLE

Once a query is handled, the UI creates a page with the results ordered and organized by different parameters and representations, providing a complete insight over the selected data. This interface is customizable and intuitive, making it possible to portray complex results in a user-friendly way, conceived to be used by anyone, anywhere. The multiple dashboards seen above can be moved around, edited and deleted freely, making data exploration as simple as moving files on the desktop.

Other than its NPL process and interface design, one of the main strengths Wizdee is its ability to customize the information contained in the dashboards, as well as their own preferred view of the dashboard, successfully creating the display that best meets the user needs and holds all of the pertinent data, making for a dynamic experience of exploring and visualizing the process as a whole.

2.3.6. FLEXIDASH

Flexidash is a platform designed to seamlessly portray what is going on a given environment, in a perceptible way (Flexidash, s.d.).

FIGURE 16 FLEXIDASH DASHBOARDS

As seen in Figure 16 This particular tool can provide multiple customizable dashboards for a given dataset and was designed so that it could be easily and quickly deployed in different companies, making it possible to centralize all of the company's valuable and relevant data, for a given context, in a single screen, which can be accessed anywhere as long as there's an internet connection.

R is currently available as a free software under the terms of the Free Software Foundation's GNU General Public License and runs on a wide variety of UNIX platforms, MacOS and Windows, being a viable alternative to the Weka system adopted in this project.

3. Business Analysis

Nowadays society is technology driven and companies are forced to become increasingly effective in order to keep up with a continuously demanding market. The raw amounts of data available contain the knowledge needed to reach the required level of efficiency, but as records multiply, it is not humanly feasible to keep track of the progress.

The tremendous increase in the amount of recorded and stored data in this digital era, and the investments made in building and managing said records, lead to an intrinsic need to make sense of all of that chaos as a return investment beyond basic reports and statistics. This demand made companies turn to data mining to examine their records for insights into business transactions and operations.

3.1. Understanding Data Mining

Data mining is the knowledge discovery process responsible for drawing valuable information and conclusions, otherwise hardly detected (if at all), from large datasets (SAS, s.d.) (Frand, s.d.) (Albion Research, s.d.). The main goal of this technology is to automatically extract previously unknown and interesting patterns in the given data, transforming it into a more organized and understandable structure for further use. It is imperial for this project to use a data mining tool to summarize the knowledge contained in its database.

Generally speaking, data mining is the process of analyzing data from different perspectives and summarizing the findings into useful and objective information. Companies have analyzed market search reports for years and, with the continuous improvement in processing power, disk storage and software quality, the accuracy of the analysis is improving, while the associated cost is dropping.

This kind of data analysis is a complex process to implement in a company, but the results are highly rewarding, so much so that this particular field of expertise is one of the most requested at the moment (Dell, 2015).This process can be applied to anywhere in a business or organization if there's an interest to identify and exploit predictable outcomes.

Data mining software is but a number of analytical tools and algorithms designed to analyze a given dataset though various angles, categorizing and summarizing the findings with a level of automatism. Technically speaking, this process finds correlations and patterns in large relational databases.

In order to link the raw data with the mining tools, a number of procedures is taken, following the CRISP-DM model.

3.2. Educational Data Mining

The educational system is becoming increasingly demanding, with society's youngest minds requiring more resources with which to educate themselves, and school administrating boards being expected to meet these demands seamlessly. With this progressively tougher operating conditions, producing the required reports and analysis takes up a significant amount of time and resources (Educational Data Mining, 2015).

In order to reduce this burden and even improve future operations and academic performance, reducing management complexity and outlining the insights that matter, business intelligence systems started looking increasingly attractive to educational institutions.

By implementing a business intelligence and analytics initiative, school administrators construct a pipeline to instantly updated reports, thus meeting their reporting requirements faster and easier, diminishing the administrative burden, while fostering for a wider culture of accountability and data driven decision making.

Educational Data Mining (EDM) (Editechreview, 2013) (Yacef, 2009) (Yellowfin, 2014) (Silva, n.d.) emerges as a discipline concerned with developing the means to understand and give an appropriate answer the unique settings the students experience. EDM focus on developing the methods, tools, algorithms that best flesh out the underlying patterns of a given environment, with its main goals being:

- **Predicting –** By creating student models that measures the student's characteristics, predisposition, attitude, user experience and satisfaction, making future learning behavior predictable.

- **Improving domains –** By continuously applying EMD's techniques, dynamically discover new models or possible improvements to existing one, optimizing the whole process.
- **Support analysis –** Study the effects of the different kinds of pedagogical support that can be achieved through learning systems.
- **Improving knowledge –** By building models of students, domain, technology or software used, and the field of the EDM research, it's possible to make advancements in scientific knowledge about learners and learning.

3.2.1. WEKA

Weka is a data mining open source application written with the objected oriented language java, by the University of Waikato in New Zealand, standing for Waikato Environment for Knowledge Analysis (WEKA) (Mark Hall, 2009). This tool provides a collection of machine learning algorithms that can be used to draw knowledge from a set of stats which can be implemented in educational data. Using modules of data preprocessing, classification, clustering, and association rule extraction, EDM can be improved with tools like Weka, in order to provide better quality of service for educational institutions.

An open source application with very good capabilities of studying and aiding academic institutions in finding and improving their own art of teaching students. An improvement to the overall quality of service of academic skills of a Country contribute for a huge development of their culture.

EDM can be enhanced with Weka and it is possible to see real results in this document with real life data retrieved from the academic institution ISEP.

3.2.2. ASSOCIATION RULES

Association rule mining (Pang-Ning Tan, 2006) is finding frequent patterns, associations, correlations or causal structures in a given set of transactions. This type of mining makes it possible to understand, and even predict more accurately, the most likely next step (Rakesh Agrawal, 1993) (Rouse, 2011).For instance: by understanding the buying habits of a group of

customers, it is possible to predict their future purchases; a customer who bought a cellphone is probably going to buy a case for it at some point.

An association rule has two parts (Wikibooks, 2015): the antecedent and the consequent, and can be read as: “If antecedent, then consequent”, the antecedent being an item found in the data and the consequent the item that is found in combination with the first. Both the antecedent and the consequent can be composed by one or more items, with these associations being less and less vulgar as the number of item grows (fewer transactions) and usually discarded because they don’t meet the minimum thresholds, proving to be anomalies and, therefore, unreliable to point towards the big picture.

In order to flesh out the relations in the dataset, the association rule algorithm uses different measures of interestingness. For an X -> Y rule:

- **Support –** Indicates the proportion of transactions which contain the itemset X; how frequent it is.
- **Confidence –** For an itemset X, indicates the proportion of transactions Y was also present; the number of times that rule was proven to be right
- **Lift –** Indicates the ratio of the observed support to what would be expected if X and Y were independent.
- **Conviction –** Indicates the frequency in which X occurs without Y; how often the rule makes an incorrect prediction.

After settling the parameters that best fit the associations being looked for, the results are straightforward and easy to read. Rule example: “Customers who purchase Barbie dolls have a 60 % likehood of also purchasing candy”. This indicates that out of all customers who bought Barbie Dolls, 60 % also bought candy.
There is no right/wrong values for the chosen parameters and it’s rather up for the user to decide on what those values should be, given the nature and context of the results being looked for, and adopt a trial an error approach to the problem.

3.2.2.1. APRIORI

The association rule algorithm used was Apriori, which receives the support and confidence minimum thresholds defined by the user and then has its work decomposed into two sub problems:

- Find the itemsets which occurrences exceed the predefined threshold by performing a breadth-first search, (starting at the tree root and exploring its neighbor nodes before moving to a different level's neighbors), and then pruning the candidates that are infrequent (don't fit the threshold).

  In order to improve performance, this algorithm reduces the number of generated candidates, by pruning all supersets of an infrequent item. The other way around is also applicable, by assuming all that all subsets of a frequent itemset, are also frequent. For instance: If X is an infrequent item, no itemset containing X can ever be frequent, thus deeming all candidates containing X to be impracticable. If XY is a frequent itemset, then both X and Y are frequent itemsets on their own.

- Generate association rules with the predefined threshold for the candidates generated in the first problem. This is accomplished by removing the antecedent's last item and assessing whether or not the association rule fits the given constraints. This process repeats itself until the antecedent becomes empty. For instance, the itemset {A, B, C} would iterate as {A, B} -> C and then as {A} -> B, with the relevant rules being outputted.

The Apriori algorithm pseudo code can be seen in Table 1.

TABLE 1 - APRIORI ALGORITHM PSEUDO CODE

```
Ck: Candidate itemset of size k
Lk: frequent itemset of size k

L1 = {frequent items};

for ( k = 1; L k != Ø; k++) do

        begin

        Ck+1 = candidates generated from Lk;

                for each transaction t in database do

                increment the count of all candidates in Ck+1 that are contained in t

                Lk+1 = candidates in Ck+1 with min_support
```

```
        end
return ∪ k Lk;
```

Example: Let's consider the training set displayed in Table 2.

TABLE 2 - DATA MINING TRAINING SET

Transcation ID	Transaction List of Items
T100	I1, I2, I5
T101	I2, I4
T102	I2, I3
T103	I1, I2, I4
T104	I1, I3
T105	I2, I3
T106	I1, I3
T107	I1, I2 ,I3, I5
T108	I1, I2, I3

The training set contains a total of 9 transactions. Let's assume the support and confidence parameters set for this experiment are 0.2 and 0.7, respectively.

The first step is to find out the frequent itemsets that fit the support threshold, as seen in Figure 17.

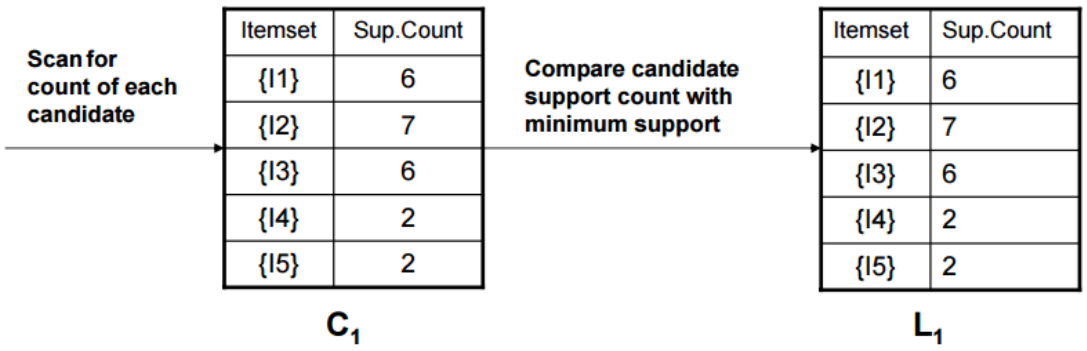

FIGURE 17 - FINDING FREQUENT ITEMSETS OF SIZE 1

In the first algorithm iteration, each of the items is a candidate itemset. L_1 is the set of candidates of size 1 that satisfy the minimum support threshold.

The second step is to generate the frequent itemsets of size 2 that fit the minimum support threshold as seen in Figure 18.

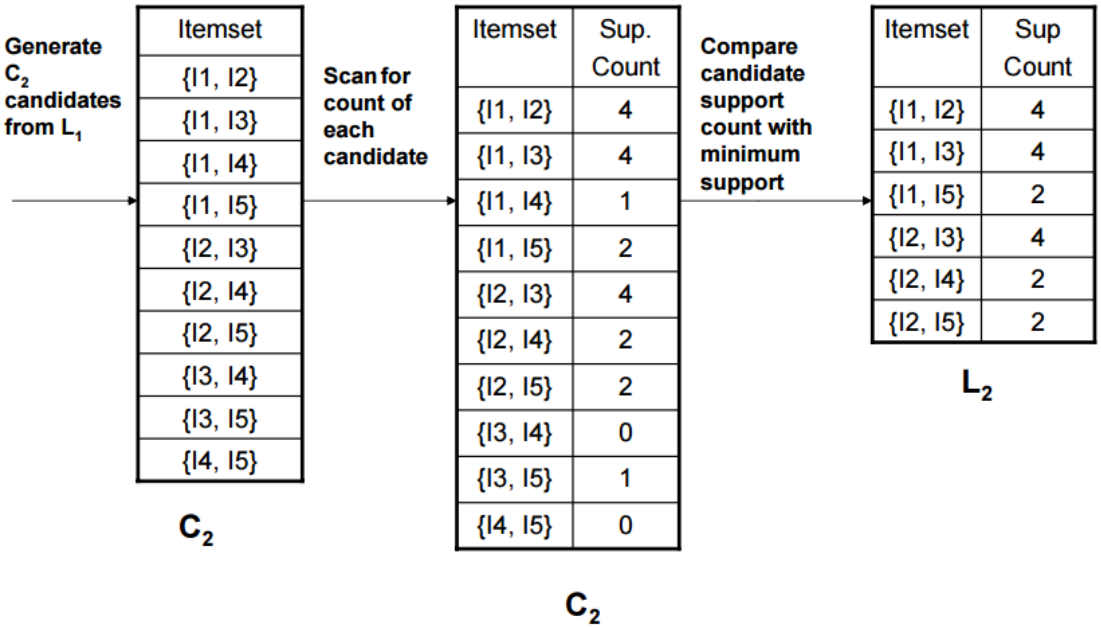

FIGURE 18 - FINDING FREQUENT ITEMSETS OF SIZE 2

In the second step, the algorithm joins the itemsets found in L_1 in order to generate a set of itemsets of size 2, then the transactions are scanned again, this time to count for the candidates present in C_2. Finally, the set of frequent itemsets of size 2 is compared with the defined minimum support threshold, creating L_2.

The third step of this process is to generate the set of candidate itemsets of size 3 as seen in Figure 19

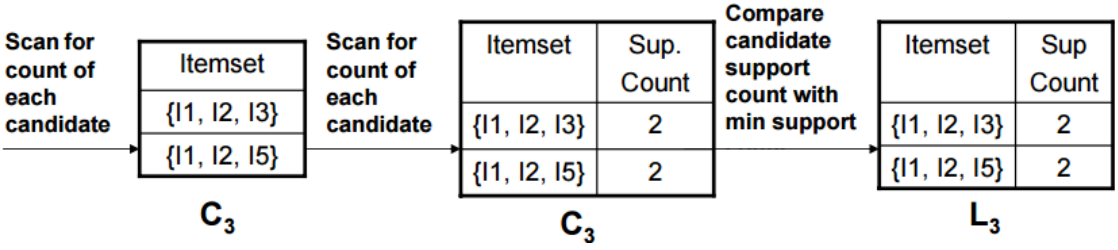

FIGURE 19 – FINDING FREQUENT ITEMSETS OF SIZE 3

The generation of candidate itemsets of size is done by joining the values of L_2:

- $C_3 = L_2$ Join L_2 = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}

With initial C_3 generated, it's time to apply the Apriori property to prune the result, reducing its size and therefor optimizing the computer calculations. The Apriori property states that all subsets of a frequent itemset must also be frequent. For example, in order for {I1, I2, I3} to be frequent, the size 2 itemsets {I1, I2}, {I1, I3} and {I2, I3} must be frequent. Only {I1, I2, I3} and {I1, I2, I5} didn't violate this property, thus being the only frequent itemsets present in L_3.
In the following step the algorithm joins the values of L_3 to generate a size 4 itemset, however the itemset found ({I1, I2, I3, I5}) is too pruned via the Apriori property, since the subset {I2, I3, I5} is not frequent. With this step terminated, all the frequent itemsets were found and it's time to generate association rules that contain them.
The next step is to generate all non-empty subsets **S** of the frequent itemsets **I** found. For each of the subsets, the rule **S -> (I-S)** is generated if the support count of **I** divided by the support count of **S** is greater or equal to the minimum confidence threshold: **support (I) / support (s) >= confidence.**

In this example, the frequent itemsets found are: {{I1}, {I2}, {I3}, {I4}, {I5}, {I1,I2}, {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}, {I1,I2,I3} and {I1,I2,I5}}. For the example itemset {I1, I2, I5}, the non-empty subsets are {I1, I2}, {I1, I5}, {I2, I5}, {I1}, {I2} and {I5}, outputting the following rules:

- Candidate Rule 1: **I1** and **I2 -> I5**
  Rule confidence: **Support {I1, I2, I5} / Support {I1, I2}** = 2/4 = 50%
  Rule is discarded
- Candidate Rule 2: **I1** and **I5 -> I2**
  Rule confidence: **Support {I1, I2, I5} / Support {I1, I5}** = 2/2 = 100%
  Rule fits the selected parameters.
- Candidate Rule 3: **I2** and **I5 -> I1**
  Rule confidence: **Support {I1, I2, I5} / Support {I2, I5}** = 2/2 = 100%
  Rule fits the selected parameters.
- Candidate Rule 4: **I1 -> I2 and I5**
  Rule confidence: **Support {I1, I2, I5} / Support {I1}** = 2/6 = 33%
  Rule is discarded
- Candidate Rule 5: **I2 -> I1 and I5**
  Rule confidence: **Support {I1, I2, I5} / Support {I2}** = 2/7 = 29%

Rule is discarded

- Candidate Rule 6: **I5 -> I1 and I2**
  Rule confidence: **Support {I1, I2, I5} / Support {I5}** = 2/2 = 100%
  Rule fits the selected parameters.

The second, third and sixth rules found for the frequent itemset {I1, I2, I5} fit the current support and confidence thresholds and are therefore present in the final results.

3.2.2.2. CLUSTERING

Clustering (Pang-Ning Tan, 2006)is a method by which large sets of data are grouped into smaller sets of data objects similar to one another.

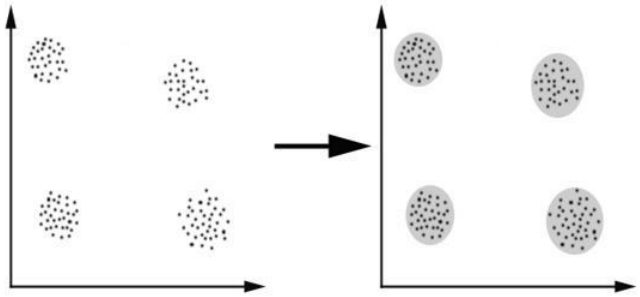

FIGURE 20 - CLUSTERING - 4 CLUSTERS

By clustering a dataset, it is possible to identify patterns that point to natural groupings that have common attributes or tendencies. This is common practice when it comes to identifying and classifying different classes, according to a given parameter. Example: Targeted marketing programs for distinct customer groups. In biology can be used to classify plants given their features.

There are two main groups of clustering algorithms:

1. **Hierarchical**
   a. Agglomerative
   b. Divisive

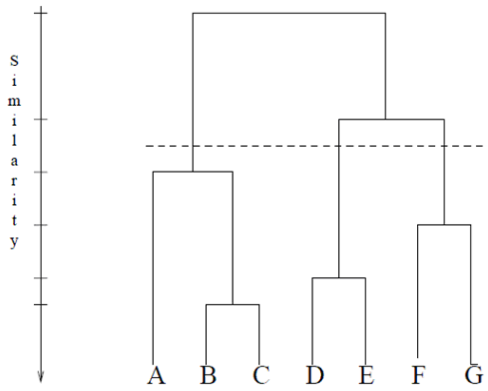

FIGURE 21 - HIERARCHICAL CLUSTERING

2. **Partitive**
    a. K Means
    b. Self-Organizing Map

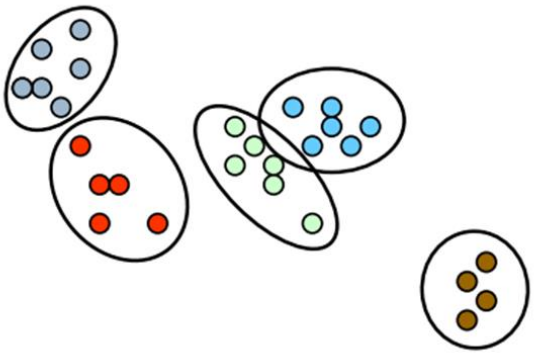

FIGURE 22 - PARTITIVE CLUSTERING

The core goal of clustering is the similarity between two objects so that clusters can be formed from objects with high similarity within clusters and low similarity between clusters. (Sayad, n.d.) A distance measure is used in order to ensure similarity and dissimilarity between objects by returning a value for pair of objects. The common distances used for clustering are:

- Euclidean
- Manhattan
- Minkowski

Since clustering is ambiguous, it is possible to cluster the same set multiple times with different parameters and considerations, and draw multiple valid conclusions of it (Estivill-Castro, 2002). This is a common practice when clustering. There are no right/wrong values for the parameters

with a trial and error approach being standard and the user responsible for choosing the requisites that best fit the context and nature of the results being searched for.
The algorithm adopted for the project solution and clustering the datasets was Simple K-Means and Euclidean distance was used for most cases.

3.2.2.2.1. SIMPLE K-MEANS

Simple K-Means is one of the most popular and simple clustering techniques. The user start by inserting K initial centroids as K is the number of clusters desired. Each point of the data set will be assigned to the closest centroid forming a collection of points (cluster). The centroid of each cluster is updated based on the points assigned to the cluster. This action consists in assigning points and updating them repeatedly until no points changes clusters or until the centroids remain the same. (Sayad, 2010-2015). The pseudo-code in Table 3 describes this action.

TABLE 3 - K-MEANS ALGORITHM

1: Select **K** points as initial centroids
2: Do
3: Form **K** clusters by assigning each point to its closest centroid
4: Recompute the centroid of each cluster
5: While Centroids do not change

This method produces ***k*** clusters with the best possible distinction and starts by computing the greatest distance between them. (Naik, 2014) Its main objective is to reduce the squared error function:

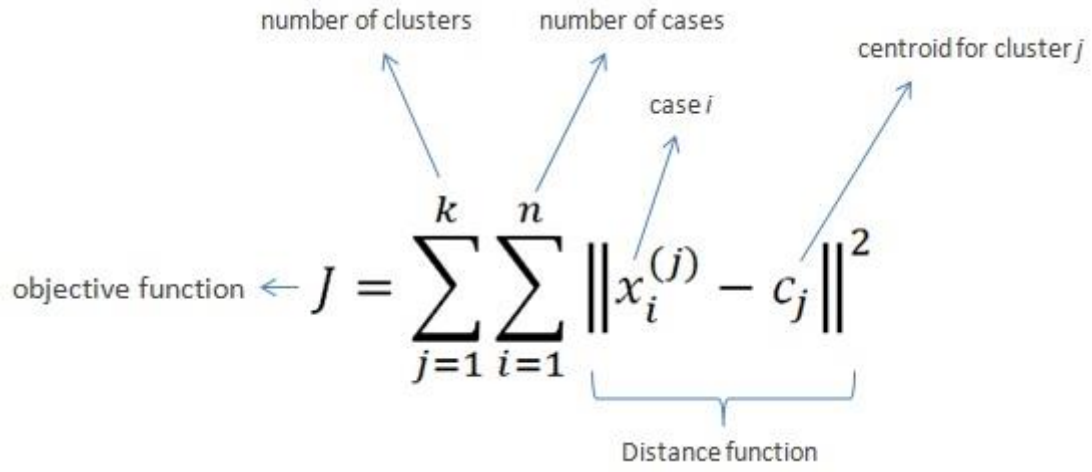

FIGURE 23 - K-MEANS ERROR FUNCTION

By receiving a set of data points **D** and an integer number of clusters **K**, this algorithm sets out to find **K** cluster center points, so as to minimize the mean square distance from each data point to its nearest center (Tapas Kanungo, s.d.) (Kiri Wagstaf, 2001) (McCullock, 2012). Basically, K-Means workflow is as follows:

- Each instance is assigned to its nearest cluster center (centroid).
- Each cluster center is updated to be the means of the data instances that compose it.

K-means can also be referred as Lloyd's algorithm and it works under iterative refining technique understood as:

- **Assignment -** consists in assigning each instance to its nearest mean, which mathematically means it partitions the observed instances according to Voronoi diagram. The formula seen in Figure 24 is applied in assignment set where each x_p is assigned to exactly one $S_i^{(t)}$.

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \leq j \leq k\},$$

FIGURE 24 - ASSIGNEMENT FORMULA

- **Update -** where the means are calculated as the new centroids for the indicated number of clusters. Assignment set is follow by the update set formula seen in Figure 25.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

FIGURE 25 - UPDATE FORMULA

There is no guarantee that the global optimum case is matched as the assignment set is no longer processed when the algorithm hits convergence.

Figure 26 demonstrates the operation of K-Means algorithm where three initial clusters are inserted and their centroids and points are assigned in four assignment-update steps. The centroids are indicated by '+' symbol and points that belong to the same cluster have the same shape. Each step in Figure 26 shows the centroids at the start of the step and the assignment of points to those centroids.

In Step 1 at Figure 26 points are assigned to the first centroids which are all in the large group of the data set. Points are then assigned to a centroid that will be updated in the next steps. Points from the same cluster have the same shape as shown in Figure 26 (Step1). After assigning points to the centroid, the centroid is then updated.

At Step 2 in Figure 26 points are assigned to the updated centroids and the centroids are updated again. It is possible to see that clusters start to be partitioned like in a Voronoi diagram that will be detailed in 3.2.2.2.1.6.

This action is repeated at step 3 in Figure 26 where is possible to see that two of the centroids have already moved several times to two small groups of points. (Sayad, 2010-2015) (MateuCC, n.d.)

Step 4 in Figure 26 shows the end of the algorithm operation where the centroids have already been identified to their grouping of points.

The algorithm ends when convergence is hit as illustrated in step 4 in Figure 26, i.e., a state where all the points are not shifting from one cluster to another, and hence, the centroids do not change. (MING-CHUAN HUNG, 2005)

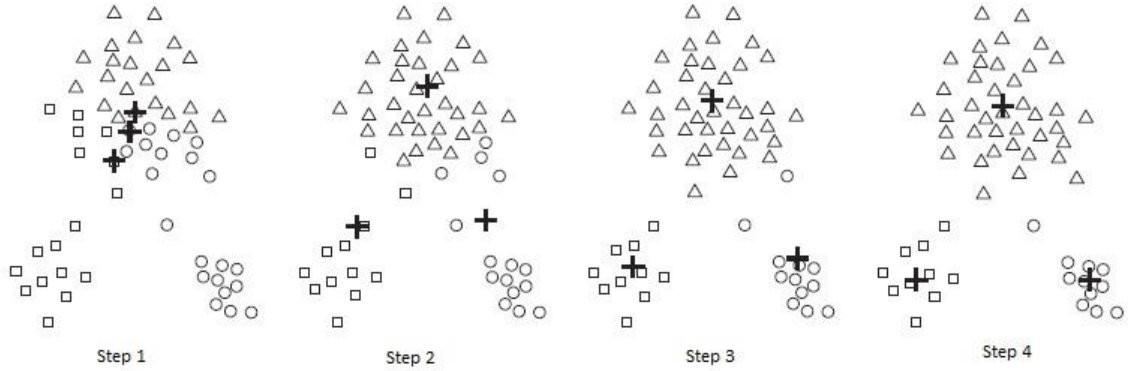

FIGURE 26- K-MEANS ALGORITHM STEPS

There is no guarantee that convergence will reach the global optimum value, k-means usually requires a few test runs to assure that the best output is obtained. Training and understanding is the key when applying clustering algorithms like simple K-means.

One of the issues of this method has, is how to choose a value for K, if there's already a known optimal value, then it should be used, however this is not the case more often than not. This issue, aligned with the fact that clustering itself is an ambiguous concept makes for a trial and error approach, whether by starting with a large value for K and progressively remove centroids (reducing K), or starting with one cluster and keep splitting it until the results are acceptable.

There is no one true approach that works better than the others, it's up to the developer/user best judgment to try and find the optimal balance between the number of clusters and their average variance, maximizing the first and minimizing the latter. K-Means is relatively an efficient method unfortunately there is no global theoretical method to find the optimal number of clusters. A practical approach is to run the algorithm several time switch different configurations and compare the outcomes. This technique was applied and is detailed in the experiments section.

3.2.2.2.1.1. DISTANCE FUNCTIONS

Simple K-Means has distance functions that can be applied to calculate the clusters. The distances applied in this thesis are:

- Euclidean Distance is a straight line distance between 2 points in Euclidean space. (Minitab® 17 Support, 2015) (Garg, July 2012)
- Manhattan Distance is the sum of absolute distances, so that outliers receive less weight than they would if the Euclidean method were used. (Garg, July 2012)
- Minkowski Distance commonly used for higher dimensional data (MateuCC, n.d.).

Euclidean Distance is the ordinary distance between two points that could be measured with a ruler and it uses Pythagorean formula. (Garg, July 2012) (Moore, 2005)
The distance between 2 points in Euclidean space can determined this way (Dell, n.d.):

- Point A – (2,2)
- Point B – (3,4)

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

FIGURE 27 - EUCLIDEAN DISTANCE FORMULA

By applying formula in Figure 27 the distance between point A and B can be determined:

$dist(x,y) = \sqrt{(2-3)^2 + (2-4)^2} = (1,4)$

After applying the formula the Euclidean distance between point A and B is (1, 4). This process is repeated in steps 3 and 4 of the clustering algorithm calculating the distance to the centroid and also the sum of squared errors.
Manhattan distance is the distance between two points in a grid base on a strictly horizontal or vertical path. It is the sum of the horizontal and vertical components. (Garg, July 2012) (Moore, 2005)

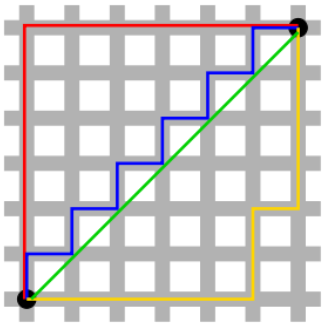

FIGURE 28 - MANHATTAN DISTANCE

Figure 28 red line is the Manhattan distance calculated for a case, where green is the diagonal between the two points and blue and yellow are other possible Manhattan distances calculated.

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$

FIGURE 29 - MANHATTAN DISTANCE FORMULA

By applying formula in Figure 29 for vector A (1, 2) and vector B (2, 3) it is possible to determine the Manhattan distance between both points:

$$d_1 = |\,1 - 2\,| + |\,2 - 3\,| = 2$$

Manhattan distance between point A and B is 2.

Figure 30 provides a good distinction between both distances.

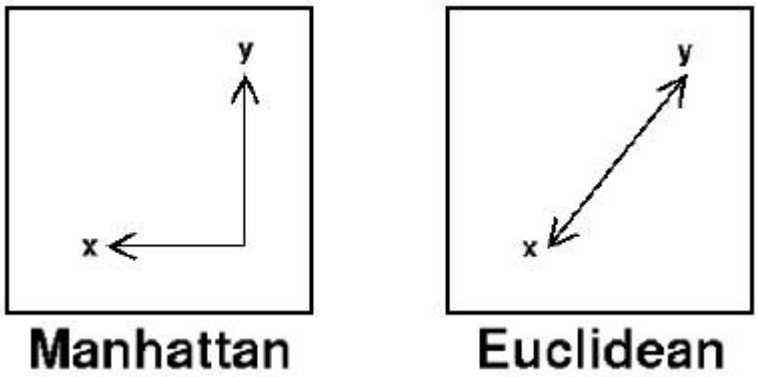

FIGURE 30 - EUCLIDEAN DISTANCE VS MANHATTAN DISTANCE

3.2.2.2.1.2. NUMBER OF CLUSTERS

The number of clusters is inserted by the user when configuring the clustering algorithm. It does not require to be calculated however, preprocessing this information may bring better results. The user must be aware that the data set he is about to cluster must be minimally understood by him.
Cluster data without knowing what variables represent and the different values they may obtain is not a good to obtain conclusions.

3.2.2.2.1.3. CENTROID

As referred in the previous sections, centroid is a collection of points assigned to himself. It is updated regularly while the clustering algorithm hits convergence, when no points are moving from one cluster to another (Minitab® 17 Support, 2015).

To assign points to the closest centroid, a proximity measure is used to quantify the notion of being close for the data set working on. Like referred above, a distance function is used for this action like Euclidean distance, calculating the similarity of a point to each cluster.
The formula for centroids is represented in Figure 31 where:

- m_i is the number of objects in the i^{th} cluster.
- m is the number of objects in the data set.
- C_i is the i^{th} cluster.
- c_i is the centroid of cluster C_i.
- X an object.

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

FIGURE 31 - CENTROID FORMULA

To illustrate centroid formula, four three dimension points will be used to obtain the centroid of a cluster:

- (2,6)
- (3,2)

- (1,2)
- (6,2)

Applying the formula in Figure 31 the result is:
((2+3+1+7) / 4), (5+2+2+2)) = (3, 3)
(3, 3) is the centroid for the cluster just calculated with the four three dimensions points used in the equation.

3.2.2.2.1.4. SUM SQUARED ERRORS (SSE)

The sum of squared errors, also known as scatter, is called as the measure of the quality of a clustering. (Rajaraman, n.d.)
The Euclidean distance for each data point is calculated and then the total sum of squared errors is computed. The lower the sum of squared errors the better the output which is the main goal of running K-Means several times for the same data set. Finding the lowest SSE value is the key to obtain the best solution since this means that the centroids provide a better representation of the instances in the clusters.

The formula in Figure 32 represents how the sum of squared errors is obtained where:

- C_i is the i^{th} cluster.
- c_i is the centroid of cluster C_i.
- K the number of clusters.
- X an object.

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

FIGURE 32 - SSE FORMULA

In Figure 32 ***dist*** represents the standard Euclidean distance between two objects in Euclidean space. The centroid that minimizes the SSE of the cluster is the mean which can be obtained by looking at Figure 31 where:

Repeating steps 3 and 4 of the K-means algorithm will result in an attempt to lower the SSE where step 3 is forming the clusters by assigning points to their nearest cluster and step 4 recomputes the centroids to further minimize the SSE. (Rajaraman, n.d.)
An example will be provided in the next sections of this document.

3.2.2.2.1.5. SEEDS

The seed value is used in clusters as the initial point to start applying the clustering algorithm. This value must be inserted by the user in cluster configuration and it is the first attribute that should be look into when experimenting the clustering. Different seeds provide different outputs for the clustering algorithm and the best result should be chosen, attending to the lower value of sum squared errors. (Arthur, n.d.) (Sang Su Lee, n.d.)

3.2.2.2.1.6. VORONOI DIAGRAM

Consists in partitioning into regions based on distance to points in a specific subset of data. Points can be seen as seeds and for each seed there is a corresponding region filled with points closer to that seed than any other. Regions can be seen as Voronoi cells.

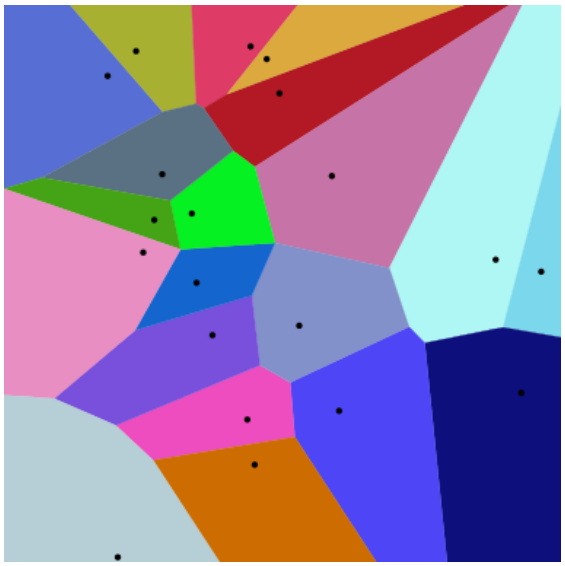

FIGURE 33 - VORONOI DIAGRAM - EUCLIDEAN DISTANCE

Basically a Voronoi diagram is a pair of points that are closed together with a point acting as the "middle" forming Voronoi cells that are delimited by a drawing line. Figure 33 and Figure 34 illustrate this behavior.

FIGURE 34 - VORONOI DIAGRAM - MANHATTAN DISTANCE

3.2.2.3. SOCIAL NETWORKS

A social network is a theoretical structure used in social sciences to study the relationships between individuals, groups or organizations. A social network is composed by a set of nodes and a set of ties that connect them as seen in Figure 35.

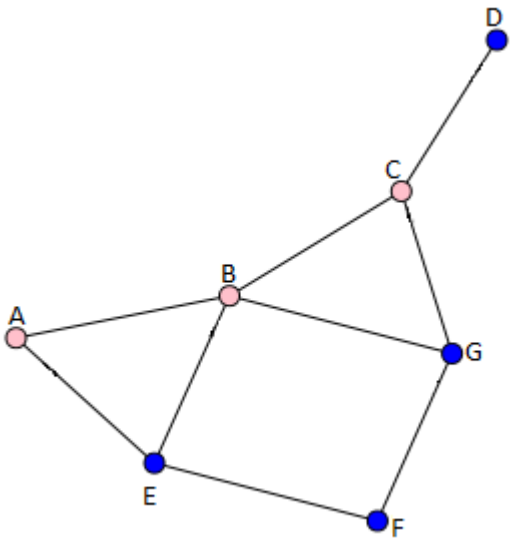

FIGURE 35 - SOCIAL NETWORK

By studying these structures it is possible to identify local and global patterns, tendencies, locate influential or important nodes and examine the networks relationships and its dynamics (Stephen P. Borgatti, 2009) (John Scott, 2011).

The social network approach is always relational, studying the properties of the relations between nodes rather than the nodes own properties. These ties through which the interactions are represented point out the convergence of the various contacts of a given node and are often given a weight value to represent the connection's strength, while the nodes are given a value defining how relevant they are, as seen in Figure 36

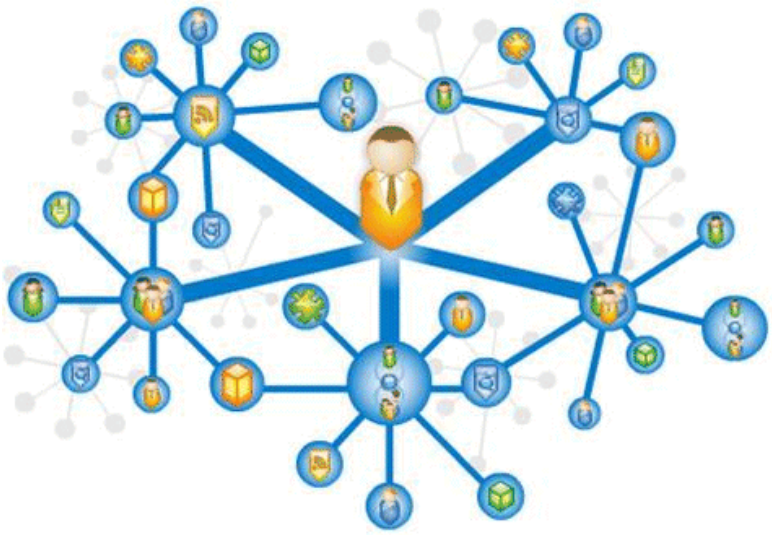

FIGURE 36 - WEIGHTED SOCIAL NETWORK

3.2.2.3.1. VIS.JS

This application uses a library called Vis.js is used in order to represent and interact with Social networks, offering the user the best overview in order to represent the analysis and statistics obtained from it.

Vis.js is a browser based open source visualization library, developed and licensed by Apache 2.0 and MIT, and it is available at *Github*. The library is designed to be easy to use, handle large amounts of dynamic data, and enable manipulation of the data.

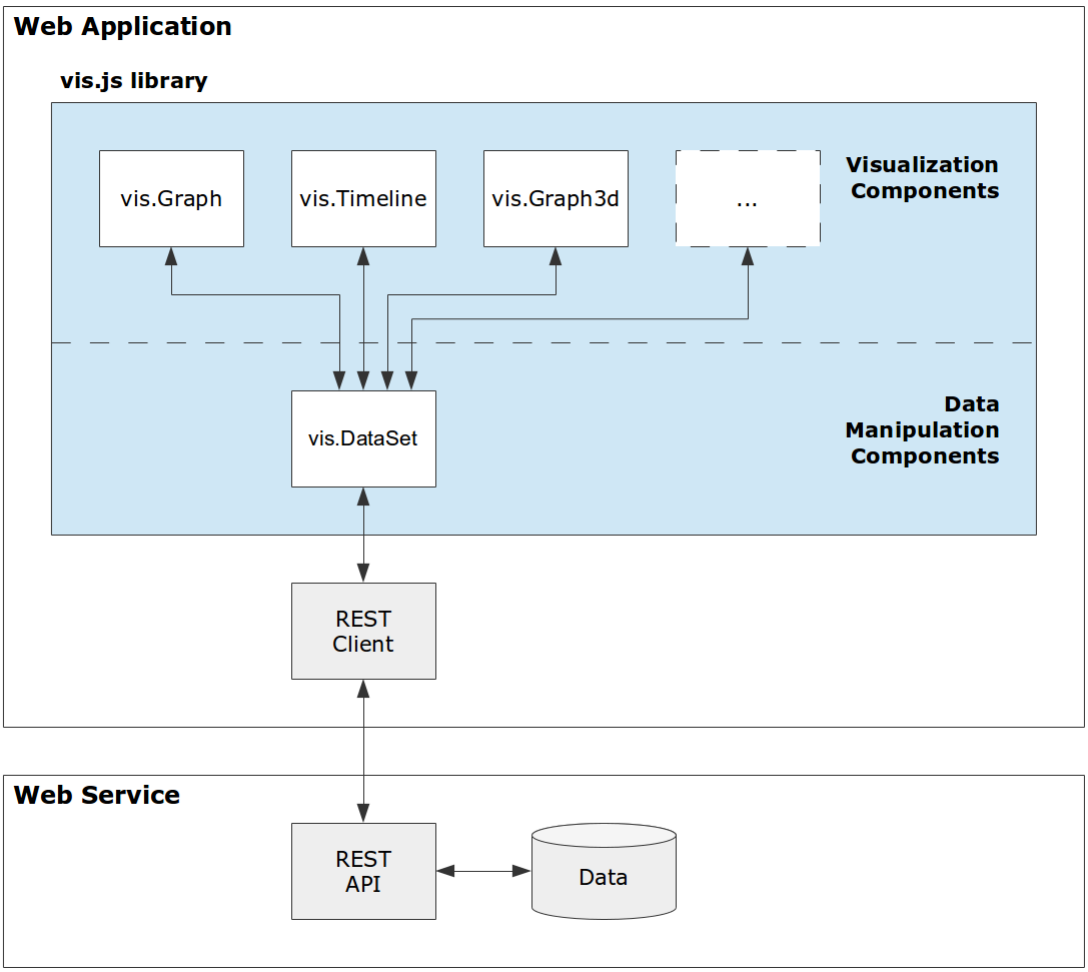

FIGURE 37- VIS.JS LIBRARY

3.3. PATTERNS AND STATISTICS

A pattern is discernible regularity and, as such, the elements of a pattern repeat in a predictable manner; to detect patterns, it's necessary to have the environment's statistics, with which one can interpret, present and organize information. This is the very basis of the service Weka provides.

By using a series of algorithms of data pre-processing, classification, regression, clustering and association rules, it is possible to take a system's raw data and stats and analyze its current state, as well as to predict its future behavior. In an educational institution such as ISEP, it's feasible to discover important underlying patterns such as:

- ✓ Students who fail at subject X usually fail at subject Y.
- ✓ Although X and Y are similar subjects, their approval rates are a lot different.
- ✓ Students usually have a grade much lower than their GPA to subject X
- ✓ On subject X, students achieve far better results during the school year than they do on the final exam.
- ✓ Given the same subject, students of teacher X have far better results than students of teacher Y.

This kind of drawn conclusions can give important pointers as for how to manage and improve on a dynamic system. Taking these patterns as examples:

- ✓ There's an underlying correlation between X and Y and its plausible to make a precedence rule, meaning that only student's approved in X can enroll in Y, saving time and money.
- ✓ This can act as a warning that these subject's standards aren't as similar as previously expected.
- ✓ A warning that something is wrong with X's current program.
- ✓ Can mean that the school year doesn't quite fit the final's exam standard, or that the final exam isn't adequate.
- ✓ Might point that X's teaching techniques aren't the most adequate, or that teacher Y has discovered a better way to approach the subject.

In any of the previously stated cases, once a pattern is discovered, there's room for further analysis, improvements and optimizations, obtaining better results while saving company resources. These statistics and patterns also make it easier to grant a greater level of transparency.

With these considerations in mind, the patterns and statistics available to the users are decide with the upmost care for their needs, while protecting sensitive and/or confidential data from being revealed.

3.4. Pearson Correlation

The Pearson Correlation is the most common statistical method used in stats to assess and measure a possible linear association (correlation) between two continuous variables, being

able to calculate the strength of the linear relationship between sets of data (Andale, 2015) (Strangroom, 2015) (Lane, s.d.). In simpler terms, the correlation attempts to find a line by which it can represent the data.

For a correlation between variables x and y, the formula for calculating the sample Pearson's correlation coefficient is given by the formula shown in Figure 38.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

FIGURE 38 - PEARSON CORRELATION FORMULA

The coefficient (r) will vary between -1 (perfect negative linear relationship) and 1 (perfect positive linear relationship), with values closer to 0 (no linear relationship) representing a higher variation of data points around the correlation line. As far as the level of correlation goes, it's possible to identify three levels:

- **High Correlation**: [0.5, 1.0] or [-0.5, -1.0]
- **Medium Correlation:** [0.3, 0.5] or [-0.3, -0.5]
- **Low Correlation:** [0.1, 0.3] or [-0.1,-0.3]

If the correlation is positive, it indicates that as one variable increases or decreases, the other tends to increase or decrease with it, whereas in a negative correlation as one variable increases, the other decreases (and vice-versa). These scenarios are presented in Figure 39

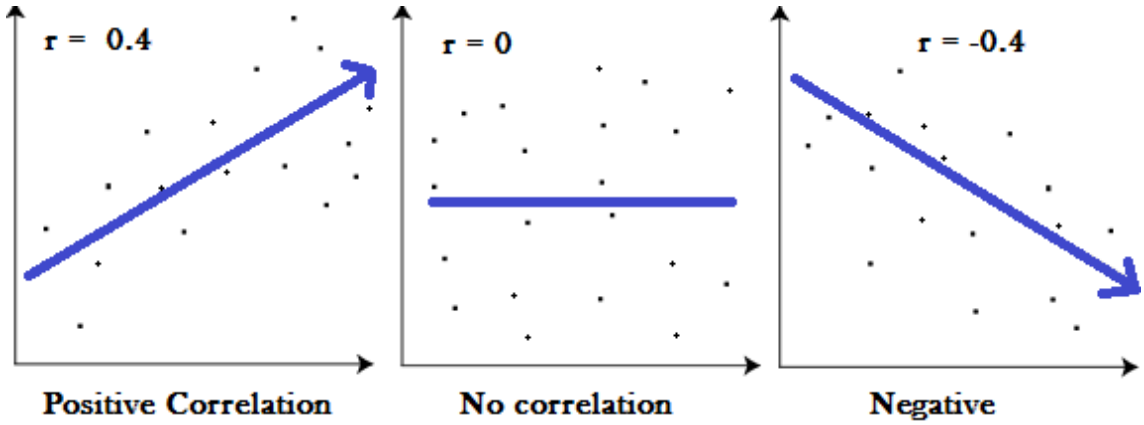

FIGURE 39 - PEARSON CORRELATION TYPES

By applying this technique in an educational data mining scenario it's possible to identify relationships between clustered instances, performance at given subjects, etc.

3.5. COSINE SIMILARITY

The cosine similarity is an arbitrary mathematical calculation of the similarity between two vectors 'a' and 'b' (Anon., s.d.) (O'Connor, 2012) as seen in Figure 40

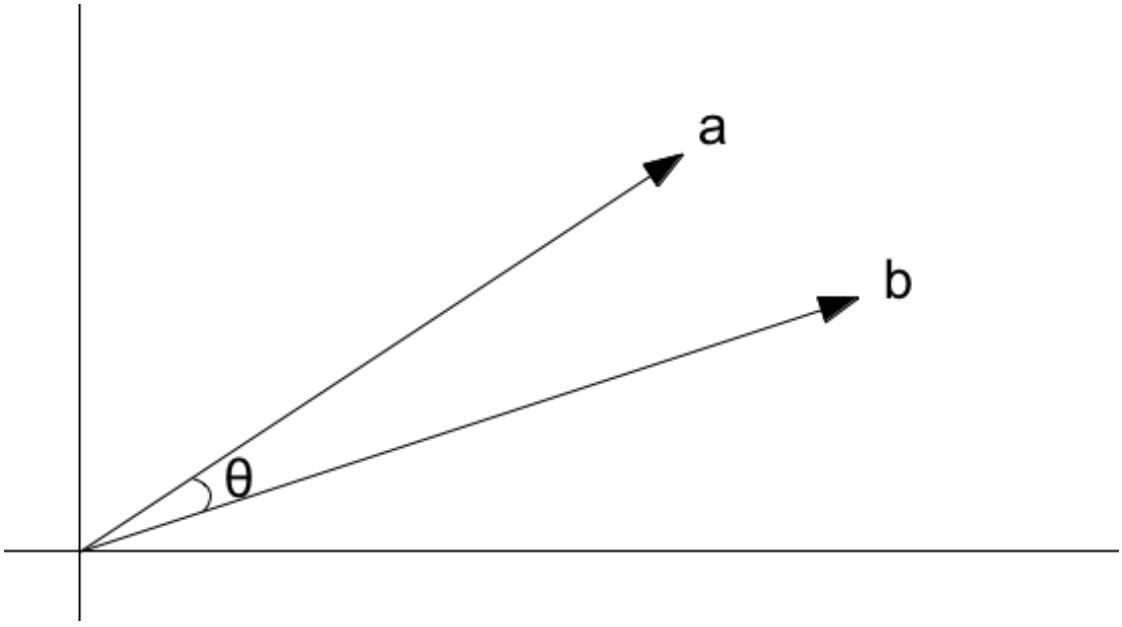

FIGURE 40 – COSINE SIMILARITY ANGLE

This similarity is calculated by studying the cosine value of the angle formed by these vectors as seen in Figure 41

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

FIGURE 41- COSINE SIMILARITY FORMULA

The cosine value being analyzed is comprehended between 0 (90°) and 1 (0°) as shown in Figure 42.

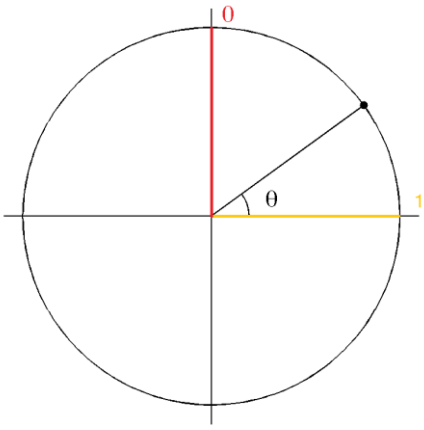

FIGURE 42 - COSINE SIMILARITY INTERVAL

This method compares the vectors with regards to their orientation, rather than their magnitude:

- Two vectors with the same orientation have a cosine similarity of 1
- Two vectors at 90° have a cosine similarity of 0

This is useful in educational data mining to assert the cohesion between nodes or within clusters.

3.6. DATABASE

The raw data to be processed by the developed application is kept in a database. This database is a key component in the application structure, providing the stored information that, when combined with the statistical and data mining technologies implemented, make it possible for the end user to access a better overview over the whole system, with data being transformed into knowledge.

Information is processed data on which to base decisions, and one of the application main goals is to improve the quality of the overall view over academic performance.
The business knowledge can be greatly improved by this application, in which the database assumes a main role by enabling a secure connection between the data and its transformation utilities provided to the user, without compromising the necessary level of confidentiality associated with the sensitive data, such as personal details.

3.6.1. AVERAGE

Grades have a strong impact in the quality of the service offered to the user has they allow to build and present conclusions to the end user.
Average grades were obtained through a set of actions in excel in order to provide the final result. A more complex excel formula was built for this application in order to retrieve crucial data that detains important value to the user.
After grouping and ordering all students correctly, removing duplicate fields, an average function was applied to the student's frequency and exam grades in order to retrieve their average scores per subject.

3.6.2. GENDER

Gender also adds value to the application, offering more options to consult data. Besides enabling a gender distribution overview, filtering students by this attribute may also point towards gender dissimilarities or patterns, thus providing important information for the improvement of the academic system.

3.6.3. APPROVAL RATE PERCENTAGE

The grades percentages assess the overall performance of the students at a given subject, thus effectively measuring how successful the adopted program is and whether or not the end results were expected.

4. DATA ANALYSIS AND EXPERIMENTS

This chapter explores the data analysis performed prior to the software development, as well as the most pertinent experiments performed during the analysis.

4.1. DATA PREPROCESSING

Data preprocessing is an essential phase for it prepares the data so that the information contained in it is best exposed to the mining tool. The real world data is not always clean, clear, complete or adequate for a mining process, and its preparation can format it to work with a given software or method, while dealing with possible inconsistencies found in it, such as errors or outliers.

The first step of data preprocessing was to remove any details concerning confidential and sensitive data, thus refining the information while taking ISEP's concerns into consideration. Measures were taken in order to anonymize all of the student's personal info.

With the sensitive data taken care of, there was a need to assure consistency throughout the multiple data sources available, because although the data was continually gathered from as early as 2008, the information was not always kept in the same molds. In order to join all of the data sources, the following modifications were made:

- **Redundancy –** Redundant data was deleted to assure consistency and prevent duplicate results, promoting efficiency.
- **Selection –** The data to be treated was selected, discarding non pertinent columns that didn't add any real value to the final set.
- **Transformation –** Data was actively transformed to fit the Weka source requirements, transforming colons into dots, semicolons into colons and adding attributes that'd symbolize specific empty fields, normalizing and segmenting numerical attributes and calculating pertinent missing columns. Example: The percentage of approvals wasn't known, but the total number of students and the number of approved students was disclosed, making the calculus possible.
- **Adaptation –** This was a continuous process as the data was adapted using the Weka tools available, so that it'd fit the requirements of the algorithms to be used at the time. The adaptation of a data source consists mostly on trimming, converting terms or grouping up the data and was made before and during each trial, with the tester

deciding what to do based on what is being looked for. Example: When many different types of failures are present in the result, such as SMNF, SMF,SMS,NC and NF, these records are converted into the simpler 'Failure' term, thus joining in all instances for what they have in common and obtaining more results regarding student negative performances.

A dummy file for Weka is presented in Table 4.

TABLE 4 - DUMMY WEKA FILE

```
@relation dummy-set

@attribute Dummy1 numeric
@attribute Dummy2 numeric
@attribute Dummy3 numeric

@data

1,0,0
1,1,1
1,0,0
1,1,1
1,0,1
```

4.2. ASSOCIATION RULE

The only association rule algorithm considered for this project is that of Apriori, but the platform is prepared to assume more algorithms in the future. This method takes order of occurrence into account and identifies a rule for each different transaction, using support and confidence as parameters to assess the interestingness of its rules, with the first being the proportion of transactions that contain the itemset, and the latter being the proportion of times the transaction's items appear together and in the same order.

4.2.1. ASSOCIATION IN WEKA

The association rules process in Weka is described in this section.

When the association rule tab is selected, it is possible to selected which association algorithm should be applied. As seen in Figure *43* the algorithm chosen was Apriori.

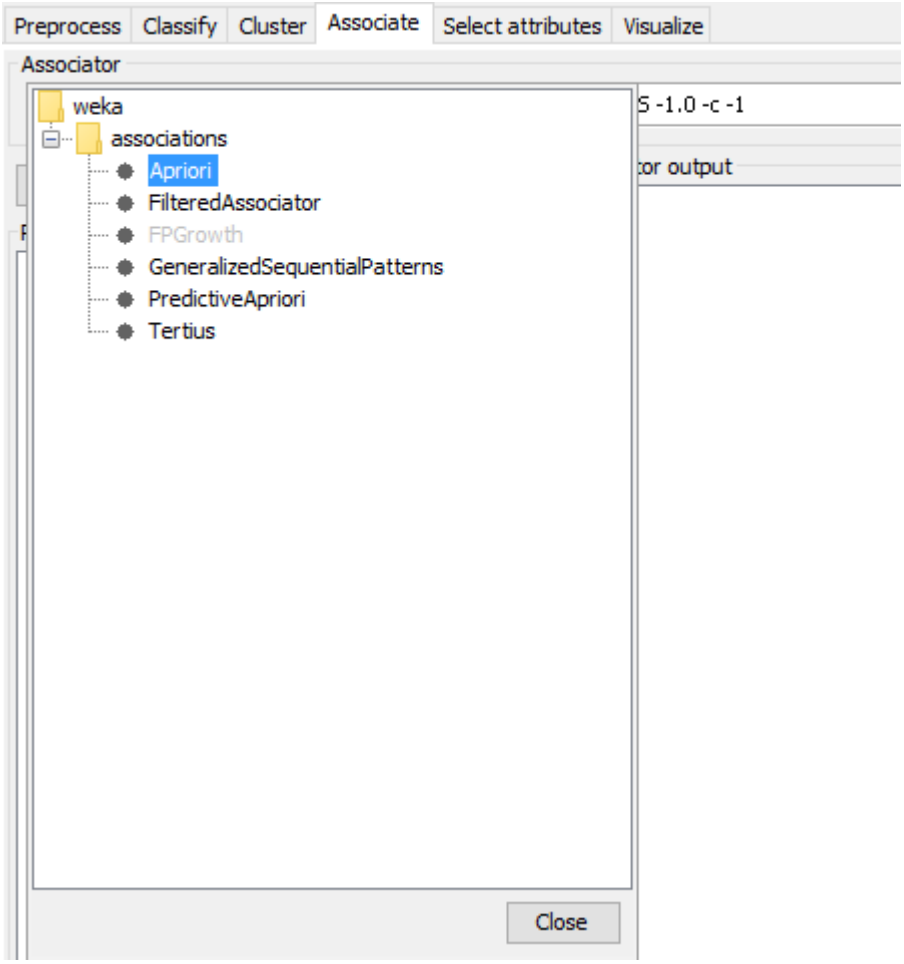

FIGURE 43 - WEKA ASSOCIATION ALGORITHMS

After choosing the algorithm to apply, it is possible to choose the parameters and values to consider as seen in Figure 44.

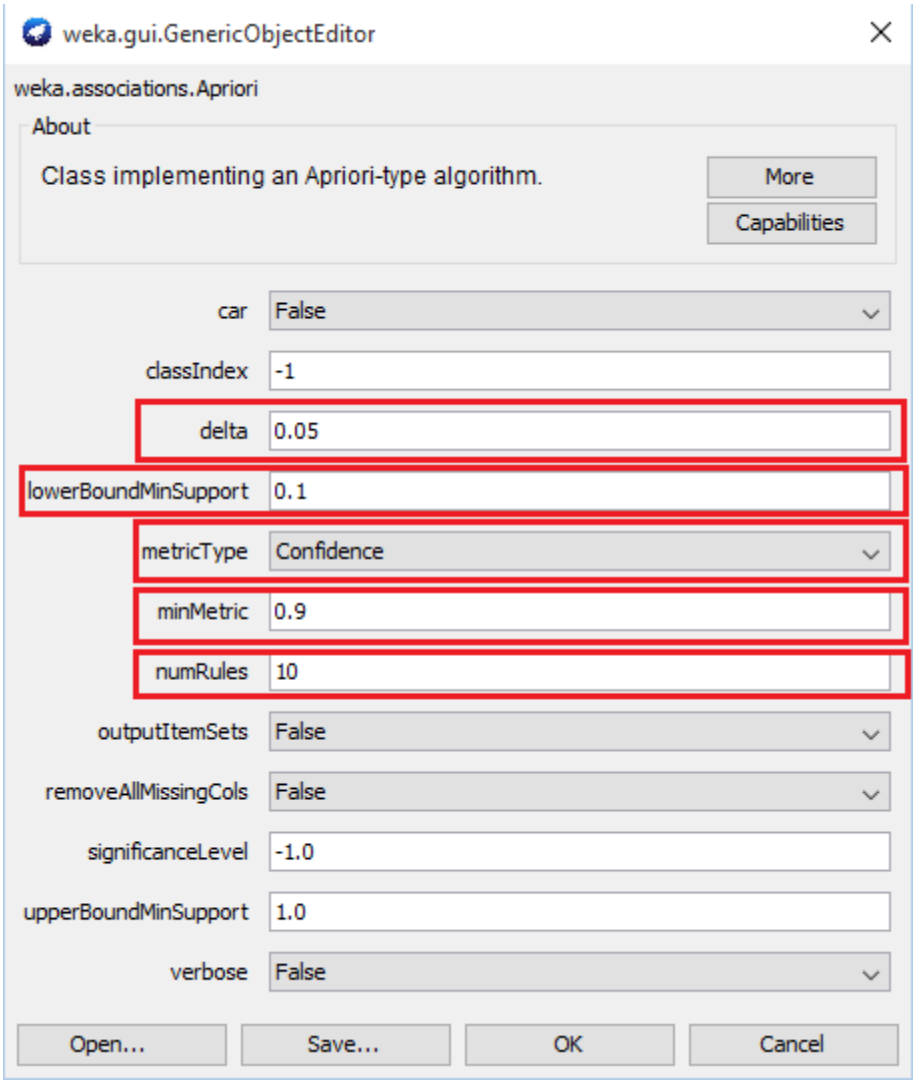

FIGURE 44 - APRIORI PARAMETERS WEKA INTERFACE

The main features to choose from the configuration window are:

- **Delta –** Support is iteratively decreased by this factor, until min support is reached or the required number of rules has been generated.
- **Lower Bound Min. Support –** Defines the minimum support threshold of the rules
- **Metric Type –** Defines the metric to apply. For this project, the only metric applied was the Confidence.
- **Min. Metric –** Indicates the minimum threshold for the chosen metric.
- **Num. Rules –** Indicates the maximum number of rules to generate

With the metrics and their values set, the algorithm execution generates an output such as the one seen in *Figure 45* .

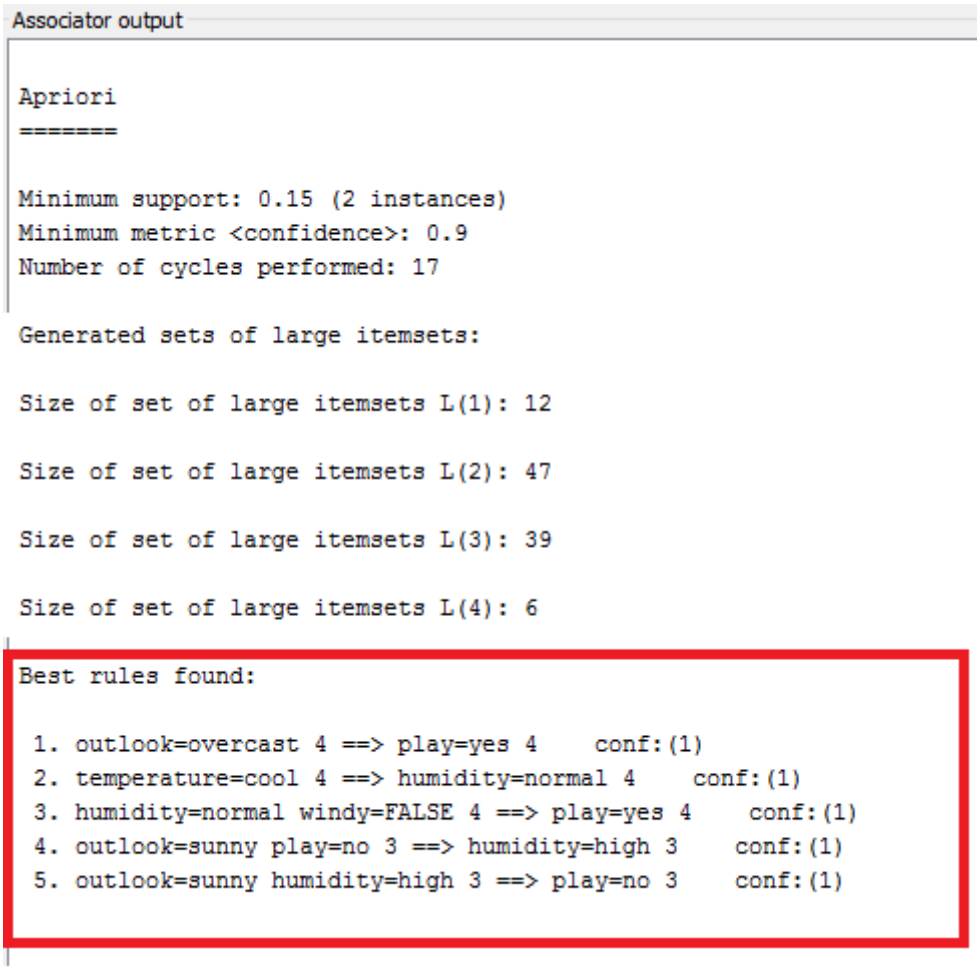

FIGURE 45 – WEKA APRIORI OUTPUT

The output starts by stating the dataset chosen and its attributes, then proceeds to refer the algorithm used, the metrics applied and the number and size of the itemsets found. This information is already known and therefore not important for the study; the valuable information that follows is the generated rules.

The generated rules are numbered from first found to last found and are structured as seen in Figure 46

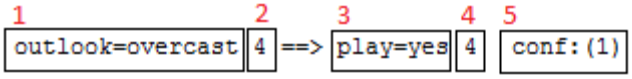

FIGURE 46- WEKA RULE STRUCTURE

1) **Antecedent –** The rules antecedent itemset
2) **Antecedent frequency –** Number of times the antecedent itemset has occurred.
3) **Consequent –** Consequent itemset

4) **Consequent frequency -** Number of times the consequent itemset has occurred after the antecedent itemset
5) **Confidence –** Rule's confidence (Consequent frequency / Antecedent Frequency)

4.2.2. Experiments Overview

Hundreds of experiments were conducted, generating thousands of different results to analyze. All of these experiments were completed with the Apriori Algorithm, and only the values of confidence and support were changed.
The association rules experiments had two different scenarios:

- The first scenario combines the data of student grades throughout multiple subjects in order to find any correlations between the performances observed in said subjects.
- The second one separates each single subject and analyzes data regarding the students that enrolled in that subject, fleshing out patterns familiar to that particular subject.

4.2.3. Multiple subjects Scenario

This scenario meant to study the synergy between related subjects, comparing student performance, fleshing out the most common grades and objectifying relationships or the lack of them where expected.

4.2.3.1. Data

For these experiments, the data analyzed was that of students who had already completed the course, which implies that all grades are positive. This ensured that the number of student records was the same for all tests and trials and that the population being observed was always the same, thus providing an overall view over student's normal behavior throughout the course.

Only the student's grades at these subjects were considered in these trials. These subjects can be consulted in Table 5.

TABLE 5 - COMPUTER ENGINEERING SUBJECTS

First Year

First Semester	Second Semester
ALGAN	PPROG
AMATA	ESOFT
PRCMP	MATCP
APROG	MATDISC
LAPR1	LAPR2

Second Year

First Semester	Second Semester
ARQCP	EAPLI
ESINF	LPROG
BDDAD	RCOMP
FSIAP	SCOMP
LAPR3	LAPR4

Third Year

First Semester	Second Semester
ALGAV	IARTI
ARQSI	CORGA
COMPA	PESTI
GESTA	
ASIST	
LAPR5	

4.2.3.2. CONFIGURATION

These experiments were conducted with:

- Apriori Algorithm
- Support values ranging from 0.1 to 0.4
- Confidence values ranging from 0.35 to 0.6.
- Lift, conviction and leverage weren't used.

4.2.3.3. RESULTS

This section summarizes the most relevant correlations found between the performances of multiple subjects.

4.2.3.3.1. Experiment 1 – Math Related Subjects

Experiment 1 isolated all math related subject to study the relationship between its grades. The analyzed dataset was as follows:

- **ALGAN –** Student's grades at subject
- **AMATA -** Student's grades at subject

- **FSIAP -** Student's grades at subject
- **MATCP -** Student's grades at subject
- **MATDISC -** Student's grades at subject

This trial meant studies the synergy between subjects that mostly deal with mathematical skills.

As seen in Table 98, Table 99 and Table 100, there's a correlation between the lowest scores of ALGAN, AMATA and FSIAP, however student's overall performance at these subjects is poor, indicating difficulties with the subject's programs, rather than a relationship between them.
Grades between 10 and 11 represent:

- 69% of FSIAP's grades;
- 59% of ALGAN's grades;
- 53% of AMATA's grades.
- 38% of MDISC'S grades
- 37% of MATCP's grades

MATCP and MDISC only show up with lower support parameters, pointing out the fact that, although most students get minimal scores at these subjects too, their overall performance is much better than that of the previous ALGAN, AMATA and FSIAP.

4.2.3.3.2. Experiment 2 – LAPR Subjects

This experiment isolated all LAPR subjects to study the relationship between its grades. The analyzed dataset was as follows:

- **LAPR1 –** Student's grades at subject
- **LAPR2 -** Student's grades at subject
- **LAPR3 -** Student's grades at subject
- **LAPR4 -** Student's grades at subject
- **LAPR5 -** Student's grades at subject

This studies the synergy between LAPR subjects, which main focus is teamwork, project planning and development.

As seen in Table 101 and Table 102, rules for this dataset were only found with low support values, however LAPR2 and LAPR3 grades look closely related.

Given the fact that LAPR subjects are only found with low support values, the connection between these subjects is not as obvious as expected. The results also show that students perform better with the two final LAPR projects than they do with the first three.

4.2.3.3.3. Experiment 3 – Structure and Planning Subjects

This experiment isolated all structure and planning subjects. The analyzed dataset is as follows:

- **BDDAD –** Student's grades at subject
- **EAPLI -** Student's grades at subject
- **ESOFT -** Student's grades at subject
- **LAPR3 -** Student's grades at subject
- **LAPR4 -** Student's grades at subject
- **LAPR5 –** Student's grades at subject
- **PESTI –** Student's grades at subject

This studied the synergy between structure and planning subjects, which focus on project planning and structure, defining and applying patterns, methods and behaviors that represent engineering good practices to adopt.

As seen in Table 103, Table 104 and Table 105 students perform a lot better at PESTI than they do at the subjects that teach the skills required to develop PESTI; it is also clear that students don't perform as well at ESOFT as they do in close related subjects. The gap between these scores is alarming.

Although a little better, the scores of LAPR3 are usually related to the ones of BDDAD, whereas LAPR4 and LAPR5 aren't as present in the results as expected.

Most students only achieve the bare minimum in ESOFT and BDDAD with grades between 10 and 11 representing 69 % of ESOFT's grades and 57% of BDDAD's grades. This fact compromises the reliability of the connections found with these subjects.

4.2.3.3.4. Experiment 4 – Entrepreneurship and Management Subjects

This experiment isolated both entrepreneurship and management subjects to study the relationship between its grades. The analyzed dataset was as follows:

- **GESTA–** Student's grades at subject
- **CORGA-** Student's grades at subject

This studied the synergy between management and entrepreneurship subjects, which focus on how to manage a company, how to sell oneself or a product and how to adapt to the market, as well as teamwork and leading practices to adopt.

TABLE 6 A.R EXPERIMENT – ENTREPRENEURSHIP AND MANAGEMENT SUBJECTS SYNERGY

Support	**Confidence**
0.1	0.4
Result	**Description**
CORGA='(12-13]' 152 ==> GESTA='(10-11]' 67 conf:(0.44)	Scores between 12 and 13 at CORGA usually related to scores between 10-11 at GESTA
GESTA='(12-13]' 189 ==> CORGA='(14-15]' 83 conf:(0.44)	Scores between 12 and 13 at GESTA usually related to scores between 14-15 at CORGA
GESTA='(10-11]' 182 ==> CORGA='(14-15]' 77 conf:(0.42)	Scores between 10-11 at GESTA usually related to scores between 14-15 at CORGA
CORGA='(14-15]' 203 ==> GESTA='(12-13]' 83 conf:(0.41)	Scores between 14-15 at CORGA usually related to scores between 12 and 13 at GESTA
CORGA='(12-13]' 152 ==> GESTA='(12-13]' 61 conf:(0.4)	Scores between 12 and 13 at CORGA usually related to scores between 12 and 13 at GESTA

As seen in Table 6, the association rules found for these subjects were only found with very low support parameters, compromising their reliability. These rules also point out that, despite the similarities, CORGA's scores are usually higher than those of GESTA.

4.2.3.3.5. Experiment 5 – Artificial intelligence Subjects

This experiment isolated both artificial intelligence subjects to study the relationship between its grades. The analyzed dataset was as follows:

- **ALGAV–** Student's grades at subject
- **IARTI-** Student's grades at subject

This allowed to study the synergy between artificial intelligence subjects, which focus on recursive problem solving, reasoning and knowledge representation.

TABLE 7 A.R EXPERIMENT – ARTIFICIALL INTELLIGENCE SUBJECTS SYNERGY

Support	**Confidence**
0.2	0.4
Result	**Description**
ALGAV='(10-11]' 172 ==> IARTI='(10-11]' 102 conf:(0.59)	Scores between 10-11 at ALGAV usually related to scores between 10-11 at IARTI
IARTI='(10-11]' 204 ==> ALGAV='(10-11]' 102 conf:(0.5)	Scores between 10-11 at IARTI usually related to scores between 10-11 at ALGAV

TABLE 8 A.R EXPERIMENT – ARTIFICIALL INTELLIGENCE SUBJECTS SYNERGY

Support	**Confidence**
0.1	0.35
Result	**Description**
ALGAV='(10-11]' 172 ==> IARTI='(10-11]' 102 conf:(0.59)	Scores between 10-11 at ALGAV usually related to scores between 10-11 at IARTI
IARTI='(10-11]' 204 ==> ALGAV='(10-11]' 102 conf:(0.5)	Scores between 10-11 at IARTI usually related to scores between 10-11 at ALGAV
ALGAV='(12-13]' 163 ==> IARTI='(10-11]' 70 conf:(0.43)	Scores between 12 and 13 at ALGAV usually related to scores between 10-11 at IARTI

IARTI='(12-13]' 151 ==> ALGAV='(12-13]' 60 conf:(0.4)	Scores between 12 and 13 at IARTI usually related to scores between 12 and 13 at ALGAV
ALGAV='(12-13]' 163 ==> IARTI='(12-13]' 60 conf:(0.37)	Scores between 12 and 13 at ALGAV usually related to scores between 12 and 13 at IARTI

As seen in Table 7 and Table 8, meaningful association rules were only found with low support parameters and mostly relating the lowest scores of each subject, with grades between 10 and 11 representing 45 % of IARTI's grades and 38 % of ALGAV's grades, however the results point out a close correlation between the grades of the two subjects, as expected.

4.2.3.3.6. Experiment 6 – Computation Subjects

This experiment isolated computation subjects to study the relationship between its grades. The analyzed dataset was as follows:

- **ARQCP–** Student's grades at subject
- **ASIST-** Student's grades at subject
- **COMPA -** Student's grades at subject
- **PRCMP -** Student's grades at subject
- **SCOMP -** Student's grades at subject

This studied the synergy between computation subjects, which focus on computer components work and communicate with one another, as well as how to program efficiently.

TABLE 9 A.R EXPERIMENT – COMPUTATION SUBJECTS SYNERGY

Support	**Confidence**
0.2	0.4
Result	**Description**
ARQCP='(10-11]' 151 ==> ASIST='(10-11]' 106 conf:(0.7)	Scores between 10-11 at ARQCP usually related to scores between 10-11 at ASIST
SCOMP='(10-11]' 176 ==> ASIST='(10-11]' 123 conf:(0.7)	Scores between 10-11 at SCOMP usually related to scores between 10-11 at ASIST

RCOMP='(11-12]' 172 ==> ASIST='(10-11]' 110 conf:(0.64)	Scores between 11 and 12 at RCOMP usually related to scores between 10-11 at ASIST
COMPA='(10-11]' 169 ==> ASIST='(10-11]' 108 conf:(0.64)	Scores between 10-11 at COMPA usually related to scores between 10-11 at ASIST
PRCMP='(11-12]' 177 ==> ASIST='(10-11]' 98 conf:(0.55)	Scores between 11 and 12 at PRCMP usually related to scores between 10-11 at ASIST
ASIST='(10-11]' 257 ==> SCOMP='(10-11]' 123 conf:(0.48)	Scores between 10-11 at ASIST usually related to scores between 10-11 at SCOMP
ASIST='(10-11]' 257 ==> RCOMP='(11-12]' 110 conf:(0.43)	Scores between 10-11 at ASIST usually related to scores between 11 and 12 at RCOMP
ASIST='(10-11]' 257 ==> COMPA='(10-11]' 108 conf:(0.42)	Scores between 10-11 at ASIST usually related to scores between 10-11 at COMPA
ASIST='(10-11]' 257 ==> ARQCP='(10-11]' 106 conf:(0.41)	Scores between 10-11 at ASIST usually related to scores between 10-11 at ARQCP

As seen in Table 9, the rules were only detectable with low support parameters and only relating the lowest scores of the analyzed subjects, with grades between 10 and 11 representing 55% of ASIST's grades, 39% of SCOMP's grades, 37% of COMPA's grades and 33% of ARQCP's grades. Despite not being entirely reliable for only relating the lowest scores, these results indicate a level of correlation between the subject's scores.

4.2.3.3.7. Experiment 6 – Programming Subjects

This experiment isolated programming subjects to study the relationship between its grades. The analyzed dataset was as follows:

- **APROG–** Student's grades at subject
- **ARQCP-** Student's grades at subject
- **ARQSI -** Student's grades at subject
- **COMPA -** Student's grades at subject
- **ESINF -** Student's grades at subject
- **LPROG -** Student's grades at subject
- **PPROG -** Student's grades at subject
- **SCOMP -** Student's grades at subject

- **SGRAI -** Student's grades at subject

This studies the synergy between programming subjects, which focus on software development logic and techniques, approaching many different languages and their applications.

As seen in tables Table 107 and Table 107 this experiment only showed results with low support parameters, and the relationship between subjects is only observable between their lowest and most common grades, with grades between 10 and 11 representing 41% of SGRAI's grades, 39% of LPROG's grades, 38% of SCOMP's grades, 37% of COMPA's grades and 33% of ARQCP's grades.

4.2.4. INDIVIDUAL SUBJECTS SCENARIO

This scenario meant to study student performance at each subject, taking into account the students situations pointing out usual trends and repetitive behavior.
Most of the subjects analyzed have the majority of their records associated with a given school year, which means their association rules are more representative of how it worked during that period of time rather than they are of the subject performance in general. However, it was possible to draw useful conclusions from this study. It is also clear that the more data available, the more accurate the extracted knowledge will be.

4.2.4.1. DATA

For these experiments, the data analyzed was that of student's performance throughout the subjects, with both approved and failed student's behavior being considered. This process implied that the number of student records wasn't always the same for all tests and trials and neither was the population being observed.
These trials were run for each of the course's subjects described in Table 5
The variables analyzed for each subject were:

- **Class –** Class the student was enrolled in.
- **Exam-** Grade obtained at subject's exam
- **Enrollment type –** Type of enrollment at subject – Whether Integral, Partial, Erasmus, Mobile, Free, Extraordinaire or Extra-curricular

- **Final Grade –** Final grade obtained at subject
- **Frequency_Grade–** Grade obtained during
- **Grade Difference –** Difference between the grades obtained during school year and final exam.
- **Enrollment Year –** Year student enrolled in the subject
- **Season –** Season in which the student's result was obtained – Whether NM or RE
- **Schedule –** Schedule of student at subject – Whether Diurnal (D) or Nocturnal (N)
- **Result –** Student's final result at subject

4.2.4.2. CONFIGURATION

These experiments were conducted with:

- Apriori Algorithm
- Support value of 0.2
- Confidence value of 0.4
- Lift, conviction and leverage weren't used.

4.2.4.3. RESULTS

This section summarizes the most relevant patterns found in each of the subjects.

4.2.4.3.1. Experiment 7 – Algan

TABLE 10 A.R EXPERIMENT – ALGAN PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Res=Aprov 252 ==> Exam Grade=NR 219 conf:(0.87)	87% of approvals didn't include exam evaluation
Exam Grade=NR 257 ==> Res=Aprov 219 conf:(0.85)	85% of student's who weren't evaluated by exam, were approved

Enrollment Year=13-14 230 ==> Res=Aprov 187 conf:(0.81)	81% of students who enrolled between 2013-2014 were approved
Season=Normal 278 ==> Res=Aprov 220 conf:(0.79	79% of students in normal season were approved
Frequency Grade=NR 112 ==> Res=Failed 79 conf:(0.71)	71% of students without frequency grade failed
Res=Failed 117 ==> Frequency Grade=NR 79 conf:(0.68)	68% of failed students had no frequency grade

Regarding the result seen in table 17, it is important to note that:

- The exam isn't required to complete the subject
- The majority of approved students is approved with a frequency grade
- Most students who can't complete the subject with a frequency grade, can't complete it with the exam either
- 68 % of total students were approved
- 62 % of available records were of students who enrolled between 2013 and 2014

4.2.4.3.2. Experiment 8 - ALGAV

TABLE 11 A.R EXPERIMENT – ALGAV PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Result=Aprov 140 ==> Season=Normal 121 conf:(0.86)	86% of Approvals were obtained in normal season
Result=Aprov 140 ==> Schedule=D 113 conf:(0.81)	81 % of approvals were obtained in a Diurnal Schedule
Enrollment Year=11-12 80 ==> Res=Aprov. 60 conf:(0.75)	75% of students who enrolled between 2011 and 2012 were approved
Season=Appeal 86 ==> Res=Failed 67 conf:(0.78)	78% of students who got their results in appeal season, failed

Season=Appeal 86 ==> Res=SMS 57 conf:(0.66)	66% of students who got their results in appeal season, failed by SMS
Enrollment Year=9-10 105 ==> Season=Normal 62 conf:(0.59)	Only 59% of students who enrolled between 2009 and 2010, got their final grades in normal season
Enrollment Year=9-10 105 ==> Res=Failed 61 conf:(0.58)	58% of students who enrolled between 2009 and 2010, failed
Res=Failed 127 ==> Season=Appeal 67 conf:(0.53)	53% of failed students, got their final grades in appeal season
Res=Failed 127 ==> Exam Grade=FT 56 conf:(0.44)	44% of failed students, didn't show up at the exam
Res=Aprov 140 ==> Enrollment Year=11-12 60 conf:(0.43)	43% of approved students, enrolled between 2011 and 2012

Regarding the results seen on table 18, it is important to note that

- 86% of failures were by SMS
- 68 % of final results were obtained in normal season
- 52% of students were approved
- 39% of students enrolled between 2009 and 2010
- 30% of students enrolled between 2011 and 2012

4.2.4.3.3. Experiment 9 – AMATA

TABLE 12 A.R EXPERIMENT – AMATA PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Res=Aprov. 207 ==> Season=Normal 171 conf:(0.83)	83% of approvals were completed in normal season
Res=Aprov. 207 ==> Enrollment Year=13-14 170 conf:(0.82)	82% of approved students enrolled between 2013/2014

Enrollment Year=11-12 137 ==> Res=Failed 107 conf:(0.78)	78% of students who enrolled between 2011 and 2012, failed
Enrollment Year=13-14 233 ==> Res=Aprov. 170 conf:(0.73)	73% of students who enrolled between 2013/2014 were approved
Res=Failed 264 ==> Season=Normal 177 conf:(0.67)	67% of failures are achieved in normal season
Res=Failed 264 ==> Exam Grade=FT 110 conf:(0.42)	42% of failed students didn't show up for the exam

Regarding the results seen in table 19, it is important to note that:

- Exam is not required to conclude the subject
- 74% of final grades were handed in normal season
- 50% of students enrolled between 2013/2014
- 44% of students were approved
- 32% of failures were by NC
- 30% of failures were by NF
- 30% of students enrolled between 2011 and 2012

4.2.4.3.4. Experiment 10 – APROG

TABLE 13 A.R EXPERIMENT – APROG PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Final Grade Difference=-1-2 91 ==> Res=Aprov. 89 conf:(0.98)	98% of students who dropped between 1 and 2 in the exam, were approved
Res=Aprov. 214 ==> Enrollment Year=13-14 150 conf:(0.7)	70% of approvals were by students who enrolled between 2013/2014
Enrollment Year=13-14 229 ==> Res=Aprov. 150 conf:(0.66)	66% of students who enrolled between 2013/2014 were approved

Res=Failed 171 ==> Enrollment Year=13-14 95 conf:(0.56)	56% of failures were by students enrolled between 2013/2014
Res=Aprov. 214 ==> Final Grade Difference=-1-2 89 conf:(0.42)	42% of approved students, dropped their grades between 1 and 2 in the exam

Regarding the results seen in table 20, it's important to note that:

- 65% of students enrolled between 2013/2014
- 61% of students were approved
- Out of the students who did the exam, 53 % dropped their grades
- 51% of failures were by SMNF
- 31% of students dropped their grades in the exam

4.2.4.3.5. Experiment 11 – ARQCP

TABLE 14 A.R EXPERIMENT – ARQCP PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Res=Aprov 128 ==> Season=Normal 91 conf:(0.71)	71% of approvals were obtained in normal season
Res=Aprov. 128 ==> Enrollment Year=11-12 88 conf:(0.69)	69% of approvals were by students who enrolled between 2011 and 2012
Season=Appeal 116 ==> Res=Failed 79 conf:(0.68)	68% of students in appeal season, fail
Season=Appeal 116 ==> Final Grade=SMR 76 conf:(0.66)	66 % of students in appeal season, have SMR grade
Enrollment Year=11-12 198 ==> Res=Failed 110 conf:(0.56)	56% of students who enrolled between 2011 and 2012 failed
Enrollment Year=11-12 198 ==> Res=Aprov. 88 conf:(0.44)	44% of students who enrolled between 2011 and 2012, were approved

Regarding the results seen in table 21, it's important to note that:

- 61% of students enrolled between 2011 and 2012
- 45% of failures were by SMR
- 40% of students were approved
- 36% of students got their final results in appeal season

4.2.4.3.6. Experiment 12 – ARQSI

TABLE 15 A.R EXPERIMENT – ARQSI PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Enrollment Year=11-12 74 ==> Res=Aprov. 57 conf:(0.77)	77% of students who enrolled between 2011 and 2012 were approved
Res=Aprov 130 ==> Season=Normal 83 conf:(0.64)	64% of approvals were obtained in normal season
Season=Appeal 74 ==> Res=Aprov 47 conf:(0.64)	64% of students who got their final grades in appeal season, were approved
Res=Aprov. 130 ==> Final Grade=11-12 67 conf:(0.52)	52% of students who were approved got grades between 11 and 12
Res=Aprov. 130 ==> Exam Grade=7-8 52 conf:(0.4)	40% of approved students has a grade between 7-8 in the exam

Regarding the results seen in table 22, it is important to note that:

- 69% of failures were by SMS
- 58% of students were approved
- 35% of students enrolled between 2009 and 2010
- 33% of students enrolled between 2011 and 2012
- 30% of the final grades were between 11 and 12
- 30% of the final grades were SMS

4.2.4.3.7. Experiment 13 – ASIST

TABLE 16 A.R EXPERIMENT – ASIST PERFORMANCE

Support	Confidence
0.2	0.4

Result	Description
Enrollment Year=11-12 82 ==> Res=Aprov. 74 conf:(0.9)	90% of students who enrolled between 2011 and 2012 were approved
Res=Aprov. 159 ==> Season=Normal 133	84% of approvals were obtained in normal season
Frequency Grade=13-14 53 ==> Final Grade Difference=-1-2 43 conf:(0.81)	81% of students with frequency grades between 13-14, dropped their grades in the exam by 1-2 values
Enrollment Year=9-10 75 ==> Res=Aprov. 60 conf:(0.8)	80% of students who enrolled between 2009 and 2010 were approved
Res=Aprov. 159 ==> Final Grade Difference=-1-2 86 conf:(0.54)	54 % of approved students dropped their grades between 1 and 2 in the exam
Enrollment Year=11-12 82 ==> Res=Aprov. Final Grade Difference=-1-2 42 conf:(0.51)	51% of students who enrolled between 2011 and 2012, were approved but had their grades drop between 1 and 2 in the exam
Res=Aprov. 159 ==> Final Grade=11-12 76 conf:(0.48)	48% of approvals were obtained by grades between 11 and 12

Regarding the results seen in table 22, it is important to note that:

- 79% of students were approved
- Of all students who did the exam, 65% dropped their grades
- 43% of students dropped between 1 and 2 in the exam
- 41% of students enrolled between 2011 and 2012
- 37% of students enrolled between 2009 and 2010
- 38% of grades were between 11 and 12

4.2.4.3.8. Experiment 14 – BDDAD

Regarding the results seen in table 23, it is important to note that:

- Out of all approved students with frequency grades, 94% dropped their grades in the exam
- 68% of students enrolled between 2011 and 2012
- 62% of students didn't get a frequency grade
- 54% of students got their results in normal season
- 54% of students were approved

4.2.4.3.9. Experiment 15 – COMPA

Regarding the results seen in table 24, it is important to note that:

- 57% of students got their results in normal season
- 58% of students were approved
- 36% of students enrolled between 2009 and 2010
- 35% of students enrolled between 2011 and 2012
- 54% of approvals were obtained by bare minimum grade: 10

4.2.4.3.10. Experiment 16 – CORGA

TABLE 17 A.R EXPERIMENT – CORGA PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Enrollment Year=11-12 85 ==> Res=Aprov. 84 conf:(0.99)	99% of students who enrolled between 2011 and 2012 were approved
Res=Aprov. 148 ==> Season=Normal 143 conf:(0.97)	97% of approved students get their final results in normal season
Enrollment Year=9-10 44 ==> Res=Aprov. 39 conf:(0.89)	89% of students who enrolled between 2009 and 2010, were approved

Final Grade=13-14 62 ==> Final Grade Difference=-1-2 43 conf:(0.69)	69% of students with final grades between 13-14, dropped their grades in the exam by 1-2 values
Frequency Grade=15-16 76 ==> Enrollment Year=11-12 46 conf:(0.61)	61% of students with frequency grades between 15-16, enrolled between 2011 and 2012
Final Grade=13-14 62 ==> Enrollment Year=11-12 37 conf:(0.6)	60% of students with final grade between 13-14, enrolled between 2011 and 2012
Final Grade=13-14 62 ==> Frequency Grade=15-16 36 conf:(0.58)	58% of students with final grades between 13-14, have frequency grades between 15-16
Res=Aprov. 148 ==> Enrollment Year=11-12 84 conf:(0.57)	57% of approved students enrolled between 2011 and 2012
Enrollment Year=11-12 85 ==> Final Grade Difference=-1-2 38 conf:(0.45)	45% of students who enrolled between 2011 and 2012, dropped their grades between 1 and 2 in the exam
Res=Aprov. 148 ==> Final Grade=13-14 62 conf:(0.42)	42% of approvals were obtained by grades between 13-14

Regarding the results seen in table 25, it is important to note that:

- 97% of students got their final grades in normal season
- 90% of students were approved
- 52% of students enrolled between 2011 and 2012
- 50% of students dropped their grades in the exam
- 46% of students had frequency grades between 15-16
- 38% of students had final grade between 13-14
- 34% of students had frequency grades between 17-18
- 38% of students dropped between 1 and 2 in the exam

4.2.4.3.11. Experiment 17 – EAPLI

TABLE 18 A.R EXPERIMENT – EAPLI PERFORMANCE

Support	Confidence
0.2	0.4
Result	**Description**
Res=Aprov. 209 ==> Season=Normal 163 conf:(0.78)	78% of approved students got their grades in normal season
Enrollment Year=11-12 203 ==> Res=Aprov. 157 conf:(0.77)	77% of students who enrolled between 2011 and 2012, were approved
Result=Aprov 209 ==>Enrollment Year=11-12 157 conf:(0.75)	75% of approvals were by students who enrolled between 2011 and 2012
Res=Failed 98 ==> Season=Normal 66 conf:(0.67)	67% of failed students, got their final results in normal season
Final Grade=11-12 103 ==> Enrollment Year=11-12 71 conf:(0.69)	69% of students with grades between 11 and 12, enrolled between 2011 and 2012
Final Grade=11-12 103 ==> Exam Grade=10 65 conf:(0.63)	63% of students with final grades between 11 and 12, obtained grade 10 in exam
Res=Aprov. 209 ==> Final Grade=11-12 103 conf:(0.49)	49% of approvals were by grades between 11 and 12
Result=Aprov 209 ==> Difference=-1-2 89 conf:(0.43)	43% of approved students dropped between 1 and 2 in the exam

Regarding the results seen in table 26, it is important to note that:

- 92% of approved students who had frequency grades, dropped their grades in the exam (51% of total students)
- 68% of students were approved
- 66% of students enrolled between 2011 and 2012
- 49% of approved students got their final grades between 11 and 12

4.2.4.3.12. Experiment 18 – ESINF

TABLE 19 A.R EXPERIMENT – ESINF PERFORMANCE

Support	Confidence
0.2	0.4
Result	**Description**
Res=Aprov. 131 ==> Enrollment Year=11-12 102 conf:(0.78)	78% of approved students enrolled between 2011 and 2012
Season=Appeal 180 ==> Enrollment Year=11-12 107 conf:(0.59)	59% of students who got their grades in appeal season, enrolled between 2011 and 2012
Season=Appeal 180 ==> Result=Failed 119 conf:(0.66)	66% of students who got their grades in appeal season failed
Frequency_Grade=NR 180 ==>Enrollment Year=11-12 107 conf:(0.59)	59% of students who didn't get frequency grade enrolled between 2011 and 2012
Res=Failed 212 ==> Season=Appeal 119 conf:(0.56)	56% of failed students got their final grades in appeal season
Enrollment Year=11-12 199 ==> Season=Appeal 107 conf:(0.54)	54% of students who enrolled between 2011 and 2012, got their final grades in appeal season
Res=Aprov. 131 ==> Season=Normal 70 conf:(0.53)	53% of approvals are obtained in normal season
Enrollment Year=11-12 199 ==> Res=Aprov. 102 conf:(0.51)	51% of students who enrolled between 2011 and 2012 were approved
Enrollment Year=11-12 199 ==> Season=Normal 92 conf:(0.46)	46% of students who enrolled between 2011 and 2012, got their final grades in normal season
Res=Failed 212 ==> Enrollment Year=11-12 97 conf:(0.46)	46% of failed students enrolled between 2011 and 2012
Exam Grade=NR 160 ==> Res=Aprov. 69 conf:(0.43)	43% of students who didn't perform the exam were approved
Res=Failed 212 ==> Exam Grade=NR 91 conf:(0.43)	43% of failed students, didn't complete the exam

Regarding the results seen in table 27, it is important to note that:

- Exam not required to complete subject.
- 58% of students enrolled between 2011 and 2012
- 52% of students didn't have frequency grades
- 48% of students got their grades in normal season
- 40% of failures were by SMS
- 38% of students were approved

4.2.4.3.13. Experiment 19 – ESOFT

TABLE 20 A.R EXPERIMENT – ESOFT PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Result=Aprov 167 ==>Enrollment Year=13-14 120 conf:(0.72)	72% of approvals were obtained by students who enrolled between 2013/2014
Exam Grade=NR 229 ==> Res=SMNF 126 conf:(0.55)	55% of students who didn't complete the exam, failed by SMNF
Res=Aprov. 167 ==> Final Grade Difference=-1-2 87 conf:(0.52)	52% of approved students dropped their grades between 1 and 2 in the exam
Exam=NR 229 ==>RES=NF 103 conf:(0.45)	45% of students who didn't do the exam, failed by NF
Enrollment Year=13-14 259 ==> Res=Aprov. 120 conf:(0.46)	46% of students who enrolled between 2013/2014 were approved

Regarding the results seen in table 28, it is important to note that:

- 91% of students got their results in normal season
- 62% of students enrolled between 2013/2014
- 55% of students didn't complete the exam
- 50% of failures were by SMNF
- 40% of students were approved

- 25% of students enrolled between 2011 and 2012

4.2.4.3.14. Experiment 20 – FSIAP

Regarding the results seen in table 29, it is important to note that:

- 92% of student's with frequency grade, who did exam, dropped their grades
- 87% of students with frequency grades who completed the exam, dropped their grades
- 72% of grades were obtained in normal season
- 67% of students didn't complete the exam
- 54% of students enrolled between 2011 and 2012
- 38% of student's didn't have frequency grades
- 30% of students didn't have a class
- 24% of students enrolled between 2009 and 2010
- 24% of students did not have a class

4.2.4.3.15. Experiment 21 – GESTA

TABLE 21 A.R EXPERIMENT – GESTA PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Enrollment Year=11-12 81 ==> Res=Aprov. 78 conf:(0.96)	96% of students who enrolled between 2011 and 2012 were approved
Exam Grade=9-10 34 ==> Res=Aprov. 32 conf:(0.94)	94% of students with grades between 9 and 10 in the exam were approved
Season=Normal 113 ==> Res=Aprov. 104 conf:(0.92)	92% of students who got their results in normal season were approved
Season=Appeal 39 ==> Res=Aprov. 34 conf:(0.87)	87% of students who got their results in appeal season were approved
Enrollment Year=9-10 44 ==> Res=Aprov. 38 conf:(0.86)	86% of students who enrolled between 2009 and 2010 were approved

Frequency Grade=NR 45 ==> Res=Aprov. 36 conf:(0.8)	80% of students without frequency grade were approved
Result=Aprov 139 ==>Enrollment Year=11-12 78 conf:(0.56)	56% of approved students enrolled between 2011 and 2012

Regarding the results seen in table 30, it is important to note that:

- 91% of students were approved
- 82% of grades were obtained normal season
- 53% of students enrolled between 2011 and 2012
- 30% of students didn't have frequency grades

4.2.4.3.16. Experiment 21 – IARTI

Regarding the results seen in table 31, it is important to note that:

- 67% of students didn't have a frequency grade
- 60% of students were approved
- 51% of students got their final grades in normal season
- 36% of students enrolled between 2009 and 2010
- 33% of students enrolled between 2011 and 2012
- 33% of students didn't complete the exam
- 30% of students had grades between 9 and 10
- 25% of students had grades between 11 and 12

4.2.4.3.17. Experiment 22 – LAPR1

TABLE 22 A.R EXPERIMENT – LAPR1 PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**

Difference=1-2 77 ==>Enrollment Year=13-14 69 conf:(0.9)	90% of students who raised their grades between 1 and 2 in the exam, enrolled between 2013/2014
Exam=15-16 109 ==>Enrollment Year=13-14 93 conf:(0.85)	85% if students who got between 15-16 in the exam, enrolled between 2013/2014
Enrollment Year=13-14 249 ==> Result=Aprov 214 conf:(0.85)	86% of students who enrolled between 2013/2014 were approved
Res=Aprov. 265 ==> Enrollment Year=13-14 214 conf:(0.81)	81% of approvals were by students who enrolled between 2013/2014
Result=Aprov 265 ==> Difference=0.0 154 conf:(0.58)	58% of approved students didn't change their final grade in the exam
Enrollment Year=13-14 249 ==> Difference=0.0 125 conf:(0.50)	50% of students who enrolled between 2013/2014 didn't change their grade in the exam
Result=Aprov 265 ==> Exam=15-16 107 conf:(0.4)	40% of approved students, got grades between 15-16 in the exam

Regarding the results seen in table 32, it is important to note that:

- 98% of students got their final grades in normal season
- 80% of students were approved
- 75% of students enrolled between 2013/2014
- 47% of students didn't change their grades in the exam
- 33% of students got grades between 15-16 in the exam

4.2.4.3.18. Experiment 23 – LAPR2

TABLE 23 A.R EXPERIMENT – LAPR2 PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**

Exam Grade=17-18 136 ==> Res=Aprov. 126 conf:(0.93)	93% of students who got grades between 17-18 in the exam were approved
Exam Grade=17-18 136 ==> Final Grade Difference=1-2 105 conf:(0.77)	77% of students who got exam grades between 17-18, improved their final grade between 1 and 2
Result=Aprov 229 ==>Enrollment Year=13-14 170 conf:(0.74)	74% of approved students enrolled between 2013/2014
Final Grade Difference=1-2 170 ==> Enrollment Year=13-14 123 conf:(0.72)	70% of students who improved their grade between 1 and 2 in the exam, enrolled between 2013/2014
Result=Aprov 229 ==> Difference=1-2 159 conf:(0.69)	69% of approved students raise their grades between 1 and 2 in the exam
Enrollment Year=13-14 258 ==> Res=Aprov. 170 conf:(0.66)	66% of students who enrolled between 2013/2014 were approved
Enrollment Year=13-14 Res=Aprov. 170 ==> Exam Grade=17-18 95 conf:(0.56)	56% of students approved between 2013/2014, had a grade between 17-18 in the exam
Res=Aprov. 229 ==> Exam Grade=17-18 126 conf:(0.55)	55% of approved students had grades between 17-18 in the exam
Enrollment Year=13-14 258 ==> Final Grade Difference=1-2 123 conf:(0.48)	48% of students who enrolled between 2013/2014, raised their grades between 1 and 2 in the exam
Result=Failed 196 ==>Enrollment Year=13-14 90 conf:(0.46)	46% of failed students enrolled between 2013/2014

Regarding the results seen in table 33, it is important to note that:

- 62% of students enrolled between 2013/2014
- 56% of approved students grades were between 10-12
- 55% of students were approved
- 55% of exam grades were between 17-18

4.2.4.3.19. Experiment 24 – LAPR3

TABLE 24 A.R EXPERIMENT – LAPR3 PERFORMANCE

Support	Confidence
0.2	0.4
Result	**Description**
Enrollment Year=11-12 188 ==> Res=Aprov. 160 conf:(0.85)	85% of students who enrolled between 2011 and 2012 were approved
Final_Grade=11-12 72 ==>Enrollment Year=11-12 57 conf:(0.79)	79% of grades between 11 and 12 were obtained by students who enrolled between 2011 and 2012
Result=Aprov 213 ==>Enrollment Year=11-12 160 conf:(0.75)	75% of approved students enrolled between 2011 and 2012
Final_Grade=13-14 73 ==>Enrollment Year=11-12 53 conf:(0.73)	73% of grades between 13-14 were obtained by students who enrolled between 2013/2014

Regarding the results seen in table 34, it is important to note that:

- 84% of students were approved
- 74% of students enrolled between 2011 and 2012

4.2.4.3.20. Experiment 25 – LAPR4

TABLE 25 A.R EXPERIMENT – LAPR4 PERFORMANCE

Support	Confidence
0.2	0.4
Result	**Description**
Enrollment Year=11-12 205 ==> Result=Aprov 184 conf:(0.9)	90% of students who enrolled between 2011 and 2012 were approved
Result=Aprov 252 ==>Enrollment Year=11-12 184 conf:(0.73)	73% of approved students enrolled between 2011 and 2012

Final_Grade=13-14 113 ==>Enrollment Year=11-12 79 conf:(0.7)	70% of grades between 13-14 were by students who enrolled between 2011 and 2012
Result=Aprov 252 ==>Final_Grade=13-14 113 conf:(0.45)	45% of approved students had grades between 13-14

Regarding the results seen in table 35 it is important to note that:

- This subject does not require exam to be completed.
- 85% of students were approved
- 69% of students enrolled between 2011 and 2012
- 45% of approved students had grades between 13-14

4.2.4.3.21. Experiment 26 – LAPR5

TABLE 26 A.R EXPERIMENT – LAPR5 PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Final_Grade=15-16 40 ==> Difference=0.0 37 conf:(0.93)	93% of students with grade between 15-16, didn't change it in the exam
Enrollment Year=11-12 70 ==> Result=Aprov 64 conf:(0.91)	91% of students who enrolled between 2011 and 2012 were approved
Frequency_Grade=15-16 46 ==>Final_Grade=15-16 40 conf:(0.87)	87% of students with a frequency grade between 15-16, got the same final grade
Enrollment Year=9-10 69 ==> Res=Aprov. 52 conf:(0.75)	75% of students who enrolled between 2009 and 2010 were approved
Enrollment Year=11-12 70 ==> Difference=0.0 45 conf:(0.64)	64% of students who enrolled between 2011 and 2012 didn't change their grades in the exam
Result=Aprov 141 ==> Difference=0.0 88 conf:(0.62)	62% of approved students didn't change their grade in the exam

Result=Aprov 141 ==>Enrollment Year=11-12 64 conf:(0.45)	45% of approved students enrolled between 2011 and 2012

Regarding the results seen in table 36, it is important to note that:

- 78% of students were approved
- 39% of students enrolled between 2009 and 2010
- 39% of students enrolled between 2011 and 2012

4.2.4.3.22. Experiment 27 – LPROG

Regarding the results seen in table 37, it is important to note that:

- Out of all students who changed their grades in the exam, 95 % dropped their grades.
- 65% of students got their results in normal season
- 66% of students enrolled between 2011 and 2012
- 54% of students didn't have frequency grades
- 53% of students were approved
- 47% of approved students had grades between 11 and 12

4.2.4.3.23. Experiment 28 – MATCP

TABLE 27 A.R EXPERIMENT – MATCP PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Result=Aprov 264 ==> Season=Normal 243 conf:(0.92)	92% of approved students got their results in normal season
Res=Aprov. 264 ==> Exam Grade=NR 242 conf:(0.92)	92% of approved students didn't complete the exam
Res=Failed 197 ==> Season=Normal 147 conf:(0.75)	75% of failed students get their final results in normal season

Enrollment Year=13-14 252 ==> Res=Aprov. 180 conf:(0.71)	71% of student's who enrolled between 2013/2014 were approved
Result=Aprov 264 ==>Enrollment Year=13-14 180 conf:(0.68)	68% of approved students enrolled between 2013/2014
Res=Failed 197 ==> Exam Grade=NR 132 conf:(0.67)	67% of failed students didn't complete the exam
Exam Grade=NR 374 ==> Res=Aprov. 242 conf:(0.65)	65% of students who don't complete the exam are approved
Season=Normal 390 ==> Res=Aprov. 243 conf:(0.62)	62% of students who get their final results in normal season were approved

Regarding the results seen in table 38, it is important to note that:

- 84% of students got their results in normal season
- 81% of students didn't complete the exam
- 57% of students were approved
- 55% of students enrolled between 2013/2014

4.2.4.3.24. Experiment 29 – MATDSC

TABLE 28 A.R EXPERIMENT – MATDSC PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Result=Aprov 244 ==> Season=Normal 207 conf:(0.85)	85% of approved students got their results in normal season
Res=Failed 198 ==> Season=Normal 145 conf:(0.73)	73% of failed students got their results in normal season
Result=Aprov 244 ==>Enrollment Year=13-14 165 conf:(0.68)	68% of approved students enrolled between 2013/2014
Enrollment Year=13-14 250 ==> Res=Aprov. 165 conf:(0.66)	66% of students who enrolled between 2013/2014 were approved

Season=Normal 352 ==> Res=Aprov. 207 conf:(0.59)	41% of students who got their final grades in normal season were approved
Res=Aprov. 244 ==> Final Grade Difference=1-2 104 conf:(0.43)	43% of approved students improved their grades by 1-2 values in the exam

Regarding the results seen in table 39 it is important to note that:

- 79% of students got their final results in normal season
- 56% of students enrolled between 2013/2014
- 55% of students were approved

4.2.4.3.25. Experiment 30 – PESTI

TABLE 29 A.R EXPERIMENT – PESTI PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Enrollment Year=9-10 87 ==> Res=NC 69 conf:(0.79)	86% of students who enrolled between 2009 and 2010 failed by NC

Regarding the results seen in table 40, it is important to note that:

- 36% of students enrolled between 2009 and 2010
- 20% of students were approved

4.2.4.3.26. Experiment 31 – PPROG

TABLE 30 A.R EXPERIMENT – PPROG PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**

Res=Aprov. 287 ==> Season=Normal 284 conf:(0.99)	99% of approvals are obtained in normal season
Enrollment Year=13-14 255 ==> Res=Aprov. 185 conf:(0.73)	73% of students who enrolled between 2013/2014 were approved
Season=Normal 416 ==> Res=Aprov. 284 conf:(0.68)	68% of final results obtained in normal season, are approvals
Result=Aprov 287 ==>Enrollment Year=13-14 185 conf:(0.64)	64% of approved students enrolled between 2013/2014

Regarding the results seen in table 41, it is important to note that:

- 94% of students got their final grades in normal season
- 65% of students were approved
- 58% of students enrolled between 2013/2014

4.2.4.3.27. Experiment 32 – PRCMP

TABLE 31 A.R EXPERIMENT – PRCMP PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Result=Aprov 226 ==>Enrollment Year=13-14 182 conf:(0.81)	81% of approved students enrolled between 2013/2014
Enrollment Year=13-14 238 ==> Res=Aprov. 182 conf:(0.76)	76% of students who enrolled between 2013/2014 were approved
Res=Failed 115 ==> Season=Normal 80 conf:(0.7)	70% of failed students are failed in normal season
Res=Failed 115 ==> Final Grade=SMNF 73 conf:(0.63)	63% of failures are by SMNF
Res=Failed 115 ==> Exam Grade=NR 69 conf:(0.6)	60% of failed students didn’t complete the exam

Res=Aprov. 226 ==> Final Grade Difference=-1-2 99 conf:(0.44)	44% of approved students dropped their grades between 1 and 2 in the exam

Regarding the results seen in table 42, it is important to note that:

- 84% of students got their final grades in normal season
- 70% of students enrolled between 2013/2014
- 67% of students were approved

4.2.4.3.28. Experiment 33 – RCOMP

TABLE 32 A.R EXPERIMENT – RCOMP PERFORMANCE

Support	**Confidence**	
0.2	0.4	
Result	**Description**	**Notes**
Exam Grade=9-10 61 ==> Season=Normal Res=Aprov. 54 conf:(0.89)	89% of students with exam grades between 9 and 10 were approved in normal season	
Exam Grade=7-8 91 ==> Res=Aprov. 77 conf:(0.85)	85% of students with grades between 7-8 in the exam, were approved	
Res=Aprov 175 ==>Enrollment Year=11-12 141 conf:(0.81)	81% of approved students enrolled between 2011 and 2012	
Exam Grade=7-8 91 ==>Enrollment Year=11-12 70 conf:(0.77)	77% of students with grades between 7-8 in the exam, enrolled between 2011 and 2012	
Enrollment Year=11-12 201 ==> Res=Aprov. 141 conf:(0.7)	70% of students who enrolled between 2011 and 2012 were approved	
Exam Grade=7-8 91 ==> Season=Normal Res=Aprov. 63 conf:(0.69)	69% of students with exam grades between 7-8 were approved in normal season	
Res=Aprov. 175 ==> Final Grade Difference=-1-2 111 conf:(0.63)	63% of approved students dropped their grades between 1 and 2 in the exam	
Season=Normal 202 ==> Res=Aprov. Final Grade Difference=-1-2 111 conf:(0.55)	55% of students who get their results in normal season are approved but drop their grades between 1 and 2 in the exam	

Enrollment Year=11-12 201 ==> Res=Aprov. Final Grade Difference=-1-2 91 conf:(0.45)	45% of students who enrolled between 2011 and 2012 were approved but dropped their grades between 1 and 2 in the exam
Res=Aprov. 175 ==> Exam Grade=7-8 77 conf:(0.44)	44% of approved students have exam grades between 7-8

Regarding the results seen in table 43, it is important to note that:

- 86% of failures were by SMR
- Out of all students who had their grades changed by the exam, 83% dropped their grades
- 75% of students got their results in normal season
- 75% of students enrolled between 2011 and 2012
- 65% of students were approved
- 34% of exams had grades between 7-8

4.2.4.3.29. Experiment 34 – SCOMP

TABLE 33 A.R EXPERIMENT – SCOMP PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Season=Normal 161 ==> Res=Aprov. 129 conf:(0.8)	80% of approved students got their results in normal season
Enrollment Year=11-12 147 ==> Res=Aprov. 110 conf:(0.75)	75% of students who enrolled between 2011 and 2012 were approved
Enrollment Year=11-12 133 ==> Season=Normal Res=Aprov 90 conf:(0.68)	68% of students who enrolled between 2011 and 2012 were approved in normal season
Res=Aprov. 167 ==> Final Grade=11-12 69 conf:(0.41)	41% of approved students had grades between 11 and 12

Regarding the results seen in table 44, it is important to note that:

- 71% of students were approved
- 68% of students got their results in normal season
- 62% of students enrolled between 2011 and 2012
- 53% of students with frequency grades that went to exam, dropped their grades
- 51% of students who failed normal season and went to appeal season, were approved

4.2.4.3.30. Experiment 35 – SCOMP

TABLE 34 A.R EXPERIMENT – SGRAI PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Enrollment Year=11-12 80 ==> Res=Aprov. 70 conf:(0.88)	88% of students who enrolled between 2011 and 2012 were approved
Enrollment Year=11-12 80 ==> Season=Normal 69 conf:(0.86)	86% of students who enrolled between 2011 and 2012 got their final results in normal season
Result=Aprov 170 ==> Season=Normal 137 conf:(0.81)	81% of approved students got their results in normal season
Enrollment Year=9-10 96 ==> Season=Normal 63 conf:(0.66)	66% of students who enrolled between 2009 and 2010 got their final results in normal season
Enrollment Year=9-10 96 ==> Res=Aprov. 62 conf:(0.65)	65% of students who enrolled between 2009 and 2010 were approved
Season=Normal 191 ==> Res=Aprov. 137 conf:(0.72)	72% of students who got their results in normal season, were approved

Regarding the results seen in table 45, it is important to note that:

- 83% of students with frequency grades that went to exam, dropped their grades in the exam
- 75% of the students got their results in normal season

- 67% of students were approved
- 38% of students enrolled between 2009 and 2010
- 32% of students enrolled between 2011 and 2012

4.3. CLUSTERS

4.3.1. CLUSTERING IN WEKA

Weka clustering is described in this section.

In Figure 47 the clustering panel is demonstrated with simple K-Means configuration open in order to see which fields the user can set.

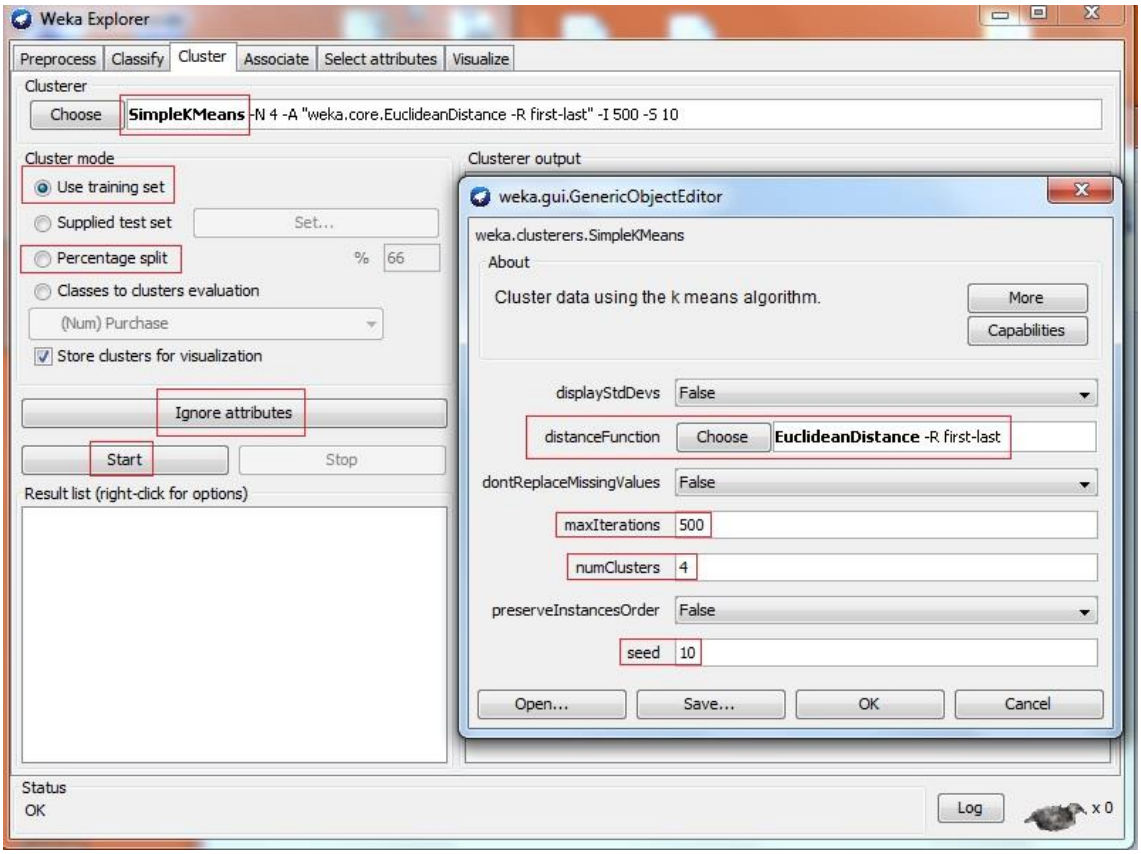

FIGURE 47 - WEKA CLUSTER PANEL

The configuration window allows us to select:

- Distance Function (Euclidean, Manhattan, etc.)
- Limit Max Iterations for the clustering algorithm, default is 500.

- Number of clusters to be clustered, default is 2.
- Number of seeds, default is 10.

The configuration can be saved to an ARFF file. After configuring the clustering algorithm we go back to the cluster panel.

In the cluster panel the following options are available:

- Use Training Set – this will send the entire data set to the clustering algorithm.
- Percentage split – this will split the data set in 2 accordingly the percentage set where X% will be used for training set and Y% will be used for test set.
- Ignore attributes button allows us to deny variables in the clustering algorithm
- Start button will run the clustering algorithm, the output will be presented in the Cluster output.

After hitting the start button it is possible to see the cluster output. For this dummy case, the output was split into 3 tables.

In Table 35 - **Cluster Output - Run Information** it is possible to see the Run information for this clustered data set where the scheme has the command that invokes what was configured for clustering. The option –I represents max iterations and –S represents the number of seeds. Configured parameters are in bold.

It also provides information about the data set like:

- Number of instances – 100
- Attributes/Variables – a list of the variables used for the clustering algorithm.
- Test mode – this will display if it is training set or test set.

TABLE 35 - CLUSTER OUTPUT - RUN INFORMATION

```
=== Run information ===
Scheme:weka.clusterers.SimpleKMeans -N 4 -A "weka.core.EuclideanDistance -R first-
last" -I 500 -S 10
Relation:     car-browsers
Instances:    100
Attributes:   8
              Dealership
              Showroom
              ComputerSearch
              M5
              3Series
              Z4
```

```
              Financing
              Purchase
Test mode:evaluate on training data
```

In Table 36 the model and evaluation obtained for the training set is displayed where:

- Number of iterations 8 was the number of iterations the algorithm performed.
- Sum of squared errors – this is a parameter that is often referred as the quality of the clustering.
- Table with cluster centroids for each attribute/variable. The second column of the table (Full Data) represents the centroids for the full data set.

TABLE 36 - CLUSTER OUTPUT - TRAINING SET

```
=== Model and evaluation on training set ===
kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 121.77671880091235
Missing values globally replaced with mean/mode

Cluster centroids:
                            Cluster#
Attribute         Full Data          0          1          2          3
                      (100)       (31)       (44)        (9)       (16)
=============================================================================
Dealership               0.6     0.0645       0.75          1          1
Showroom                0.72          1     0.7273     0.2222     0.4375
ComputerSearch          0.43     0.3548     0.2273          1     0.8125
M5                      0.53     0.0323     0.9773          1          0
3Series                 0.55          1     0.3409     0.6667     0.1875
Z4                      0.45     0.6774     0.2955     0.7778       0.25
Financing               0.61     0.4516     0.7727     0.4444     0.5625
Purchase                0.39     0.2903     0.5455     0.2222       0.25
```

For this dummy test 4 clusters were inserted and the output can show the centroid for each of them. It is also possible to see the number of instances that were calculated in that cluster. Finally, in Table 37 we have the clustered instances into clusters and the percentage they take in the full data set.

TABLE 37 - CLUSTER OUTPUT - CLUSTERED INSTANCES

```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       31 (31%)
1       44 (44%)
2        9 (9%)
3       16 (16%)
```

The first analysis made to clustering is performed in this output where there is a need to run it several times changing the configuration. The number of runs is accounted in the result list being easy to consult and access. This can be seen in Figure 48 where 4 runs were already performed.

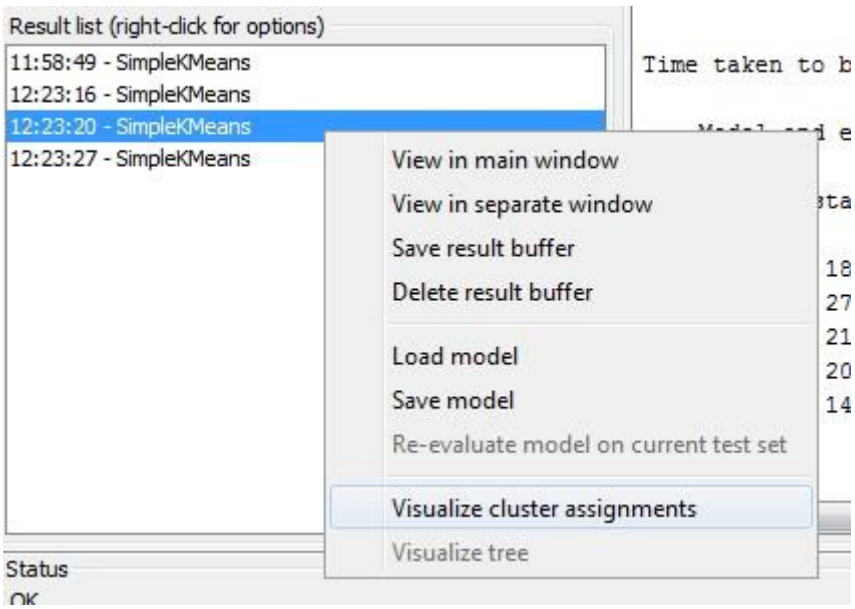

FIGURE 48 - CLUSTER PANEL - RESULT LIST

Right-clicking on the result list section of the cluster tab has an option ‘Visualize cluster assignments’.

After clicking this option a window will pop-up to see the clustering graphically. X axis, Y axis and color can be edited in order to retrieve good conclusions. In Figure 49 X axis was set to M5 variable, Y axis was set to Purchase variable and the color to cluster. This will show a chart how clusters are grouped in terms of who look at a M5 and who purchased one.

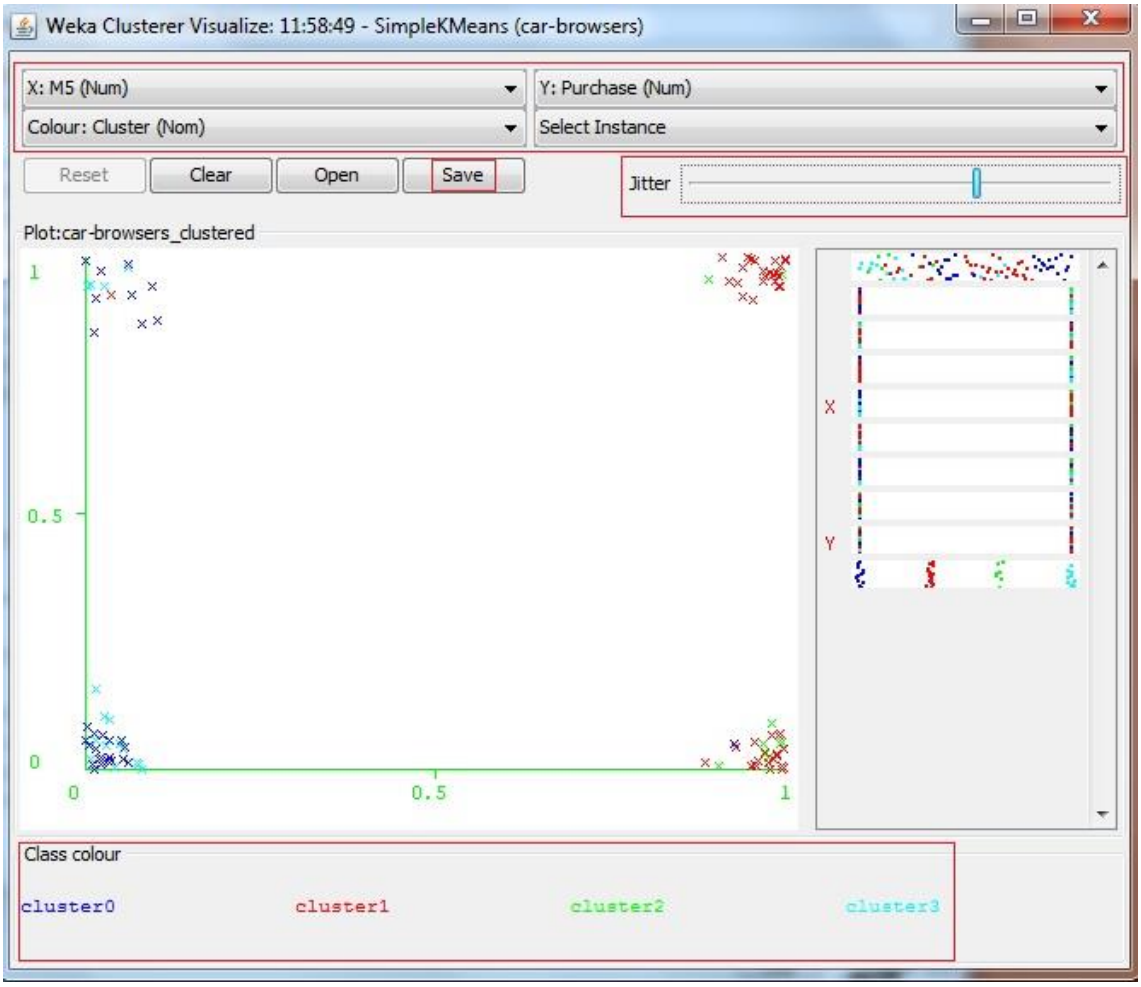

FIGURE 49 - VISUALIZE CLUSTER ASSIGNMENT

By looking at Figure 49 highlighted in red there are the options that matter for a proper analysis:

- X – Set a variable to X axis.
- Y – Set a variable to Y axis.
- Color – set a variable to be the color of the instances in the graphic.
- Jitter – will artificially scatter the plot points to allow them to be seen more easily.
- Save – save this particular set.

- Class Color (subtitle) – the color that each cluster represents in the graphic (blue, 31%, and red, 44%, have a heavy presence in this graphic however it is possible to see a few cyan and green instances).

So, analyzing the graphic, this is telling people who looked at the M5 and made a purchase. It seems that cluster 1 has a strong presence of instances that looked at the M5 and actually made a purchase so let's compare this with the cluster obtained centroids again in Table 38.

TABLE 38 - CLUSTER OUTPUT - CLUSTER 1 CENTROID DETAILED

Attribute	Cluster# 1 44)
M5	**0.9773**
Purchase	**0.5455**

In Table 38 - **Cluster Output - Cluster 1 Centroid Detailed** there is the cluster 1 centroid with the attributes we are looking at in the graphic from Figure 49 - Visualize Cluster Assignment and we can see that this distribution of instances is correct since most of the users looked at the M5 and at least half made a purchase.

Weka also provides an option to click an instance and see the details of it like in Table 39 - **Cluster Instances Details**.

TABLE 39 - CLUSTER INSTANCES DETAILS

Plot : 11:58:49 - SimpleKMeans (car-browsers) Instance: 5 Instance_number: 4.0 Dealership: 1.0 Showroom: 0.0 ComputerSearch: 1.0 M5: 1.0 3Series: 1.0 Z4: 0.0 Financing: 1.0 Purchase: 1.0 **Cluster: cluster2**	Plot : 11:58:49 - SimpleKMeans (car-browsers) Instance: 73 Instance_number: 72.0 Dealership: 0.0 Showroom: 0.0 ComputerSearch: 0.0 M5: 1.0 3Series: 0.0 Z4: 0.0 Financing: 1.0 Purchase: 1.0 **Cluster: cluster1**	Plot : 11:58:49 - SimpleKMeans (car-browsers) Instance: 75 Instance_number: 74.0 Dealership: 1.0 Showroom: 0.0 ComputerSearch: 0.0 M5: 1.0 3Series: 0.0 Z4: 1.0 Financing: 1.0 Purchase: 1.0 **Cluster: cluster1**

Table 39 provides us the details of a clicked instance at the right top corned of the graphic. We can see 2 instances that belong to cluster 1 and one that belongs to cluster 2. This is a good way to "debug" a misplaced instance.

So, from this dummy set we can take a few conclusions that will be detailed in the experiments sections, like:

- Cluster 0 is filled worth customers that visited the showroom, had a look at some of the models except the M5.
- Cluster 1 is filled with customers that made more purchases and this information can be confirmed since the values of looking at M5's and financing is high.
- Cluster 2 is filled with customers that do a lot of research about the models available but don't end up buying cars.
- Cluster 3 is a mid-term cluster where customers visited the dealership and ended up performing a few purchases. One interesting fact in this cluster is that customers didn't look at the M5 model.

Many more conclusions can be drawn from a data set depending on the judgment criteria of the user. This section will detail experiments performed for this thesis main goal in order to enrich the solution and study this data mining technique.

4.3.2. Overview

The experiments performed for clustering provide a way to study the algorithm workaround implemented for the application, in order to improve and understand how this method can be used to predict and offer valuable conclusions to the end user.

This section details the data and configurations used to perform the experiments.

The output obtained presents valuable conclusions in order to enrich the overall knowledge of data mining techniques that are applied in the developed application. For clustering, it was possible to study the algorithm results for the academic datasets used in the application.

The experiments for clustering offer different approaches that differ in small steps taken towards the final result as detailed in the next sections.

4.3.3. Data

The data used for the experiment is similar to the data used in association rules. The data available to develop the application is mainly records for students and subjects since 2008 of this academic institution.

Information is a key value for data mining, and for clustering experiments various variables were taken in consideration while some were irrelevant this experiments and therefore discarded. CSV files were generated with filtered data to be used in WEKA containing a set of variables related to:

- Subjects.
- Seasons.
- Students.
- Classes.
- Frequency, exam and final grades.
- Student entrance year in the academic institution.
- Enrollment type.
- Difference between final grade and exam grade.
- Student final result to the subject (approved or failed).

The files/tables were then created with the structure detailed above for each subject in the academic institution course. For the experiments were created around 30 files, one for each subject.

The files created are presented in Table 5 , separated by the year they are lectured for each semester.

Conclusions from experiments are then obtained by clustering them, retrieving several aspects as detailed in the next sections.

4.3.4. CONFIGURATION

The algorithm used for clustering was Simple K-Means as detailed in section 3.2.2.2.1.

Simple K-means can be configured in order to refine and improve the output obtained, by setting a number of variables:

- Distance function
- Number of Clusters
- Seeds
- Maximum iterations

For the study of clustering, Simple K-Means algorithm was used with standard configurations for almost all cases with some exceptions to provide different points of view. It is crucial to know that the more instances the bigger the sum of squared errors will be as this is understood as K-Means behavior and not a negative factor.
In order to portray different approaches to reach the final conclusions from the clustering algorithm, different configurations were used:

- Euclidean distance is standard, with Manhattan distance also used in some particular sets of data
- Minimum and standard number of clusters were 2, some cases used more.
- The number of seeds was 100 for all experiments
- The maximum number of iterations was 500 for all experiments.
- Percentage split (50% for training, 50% for test)

In Weka, there were also more configurations implemented in order to enrich the clustering method, by applying filters in the data preprocessing phase. The filter applied was the 'add cluster filter' which performs the clustering based on the data set prepared. In this mode Weka starts by ignoring the class attribute and generate the clustering. Then, during the test phase, it assigns classes to the clusters, based on the majority value of the class attribute within each cluster. Then it computes the classification error, based on this assignment and also shows the corresponding confusion matrix. To reach this output a classification to clustering action is performed which is out of scope of this document.

For each data set it requires a few steps in preprocessing to improve the data in order to help Simple K-Means to reach the most accurate result. Different approaches will be presented in the experiments presented in the next sections.
For the record, clustering graphics will be presented to offer more detailed conclusions with 3 dimensions:

- X axis
- Y axis
- Color

It will be detailed further in this document.

4.3.5. RESULTS

Several runs were performed for each experiment however experiment only show the best result run. Experiment 1 however is fully detailed with a set of 4 runs that were accomplished to reach the best output overall.

Clustering must be analyzed and runt more than one time for the same data set swapping the configurations values. This will result in better conclusions and better appliance of the clustering algorithm.

4.3.5.1. EXPERIMENT 1 – SUBJECT IARTI

Experiment 1 refers to the subject IARTI.

Clustering was processed with the following variables:

- Frequency Grade.
- Exam Grade.
- Final Grade.
- Season.
- Student entrance year.
- Number of subjects enrolled.
- Number of subjects approved.
- Percentage of subjects approved.
- Result.
- Difference between final and frequency grade.

The first run of the algorithm was performed for 263 instances with a percentage split of 50%, 5 seeds and 3 clusters.

TABLE 40 - IARTI EXPERIMENT – 1ST RUN RESULTS

	Training Set	**Test Set**
Sum Of Squared Errors	528	250
Number of Iterations	6	6

By looking at Table 40 it is possible to see that the output has improved quite a bit.

The centroids in the first run of the algorithm for the test set are presented in Table 41.

TABLE 41 - IARTI EXPERIMENT – 1ST RUN TRAINING SET CENTROIDS

Variable	Full Data	Cluster 0	Cluster 1	Cluster 2
Frequency_Grade	NR	10-11	NR	NR
Exam	NR	NR	FT	10-11
Final_Grade	10-11	10-11	NC	10-11
Season	NORMAL	NORMAL	NORMAL	APPEAL
Enrollment_Year	2010-2011	2010-2011	2010-2011	2010-2011
Res	APPROVED	APPROVED	FAILED	FAILED
Number_Subjects	8.52	8.99	6.91	9.24
Number_Approvals	5.28	7.18	1.41	6.57
Approval_Percentage	0.58	0.80	0.19	0.70
Grade_Difference	NR	0.0	NR	NR

The centroids in the first run of the algorithm for the test set are presented in Table 42.

TABLE 42 - IARTI EXPERIMENT – 1ST RUN TEST SET CENTROIDS

Variable	Full Data	Cluster 0	Cluster 1	Cluster 2
Frequency_Grade	NR	NR	NR	NR
Exam	NR	10-11	FT	7-8
Final_Grade	10-11	10-11	NC	7-8
Season	NORMAL	NORMAL	NORMAL	APPEAL
Enrollment_Year	2010-2011	2010-2011	2007-2008	2010-2011
Res	APPROVED	APPROVED	FAILED	FAILED
Number_Subjects	8.24	8.82	6.92	8.26
Number_Approvals	4.70	6.57	1.44	3.60
Approval_Percentage	0.54	0.74	0.21	0.39
Grade_Difference	NR	0.0	NR	NR

The clustered instances are then presented in Table 43.

TABLE 43 - IARTI EXPERIMENT - 1ST RUN CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage

0	97	61%
1	32	20%
2	29	18%

The cluster centroids seem to be quite different from each other. A split has occurred where one cluster ended with approved students and two clusters with students that failed the subject.

The second run of the algorithm was performed for 263 instances with a percentage split of 40% to training set and 60% to test set, 5 seeds and 3 clusters.

TABLE 44 – IARTI EXPERIMENT – 2ND RUN RESULTS

	Training Set	**Test Set**
Sum Of Squared Errors	528	217
Number of Iterations	4	7

By looking at Table 44 it is possible to see that the output has improved quite a bit and the number of iterations increased by 3.
The centroids in the first run of the algorithm for the training set are presented in Table 45.

TABLE 45 - IARTI EXPERIMENT - 2ND RUN TRAINING SET CENTROIDS

Variable	**Full Data**	**Cluster 0**	**Cluster 1**	**Cluster 2**
Frequency_Grade	NR	NR	10-11	NR
Exam	NR	FT	NR	10-11
Final_Grade	10-11	NC	10-11	10-11
Season	NORMAL	NORMAL	NORMAL	APPEAL
Enrollment_Year	2010-2011	2010-2011	2010-2011	2010-2011
Res	APPROVED	FAILED	APPROVED	APPROVED
Number_Subjects	8.52	6.91	8.98	9.24
Number_Approvals	5.28	1.41	7.18	6.56
Approval_Percentage	0.58	0.19	0.80	0.70
Grade_Difference	NR	NR	0.0	NR

The centroids in the second run of the algorithm for the test set is presented in Table 46.

TABLE 46 - IARTI EXPERIMENT - 2ND RUN TEST SET CENTROIDS

Variable	Full Data	Cluster 0	Cluster 1	Cluster 2
Frequency_Grade	NR	NR	10-11	NR
Exam	NR	10-11	NR	FT
Final_Grade	10-11	10-11	10-11	NC
Season	NORMAL	APPEAL	NORMAL	NORMAL
Enrollment_Year	2010-2011	2010-2011	2010-2011	2010-2011
Res	APPROVED	APPROVED	APPROVED	FAILED
Number_Subjects	8.24	9.31	8.68	6.84
Number_Approvals	4.70	6.34	6.48	1.74
Approval_Percentage	0.54	0.67	0.77	0.26
Grade_Difference	NR	0.0	NR	NR

The clustered instances are then presented in Table 47.

TABLE 47 - IARTI EXPERIMENT – 2ND RUN CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage
0	77	49%
1	44	28%
2	37	23%

Sum of squared errors obtained compared to the first run has decreased quite a bit. Looking at the centroids they seem to be different from each other and good conclusions can be withdrawn from it. The clusters swapped their centroids to positive results and balanced the instances percentage.

The third run of the algorithm was performed for 263 instances with a percentage split of 50%, 9 seeds and 3 clusters.

TABLE 48 – IARTI EXPERIMENT – 3RD RUN RESULTS

	Training Set	Test Set

Sum Of Squared Errors	637	211
Number of Iterations	4	6

By looking at Table 48 it is possible to see that the output has improved quite a bit. The number of iterations increased by 2. So far this is the best run.

The centroids in the third run of the algorithm for the training set are presented in Table 49.

TABLE 49 - IARTI EXPERIMENT – 3RD RUN TRAINING SET CENTROIDS

Variable	Full Data	Cluster 0	Cluster 1	Cluster 2
Frequency_Grade	NR	NR	10-11	NR
Exam	NR	NR	FT	9.0
Final_Grade	10-11	10-11	NC	9.0
Season	NORMAL	NORMAL	NORMAL	APPEAL
Enrollment_Year	2010-2011	2010-2011	2010-2011	2010-2011
Res	APPROVED	APPROVED	FAILED	FAILED
Number_Subjects	8.5209	9.0513	7.0196	8.4107
Number_Approvals	5.2814	7.0769	1.4314	3.7857
Approval_Percentage	0.5886	0.7795	0.198	0.4125
Grade_Difference	NR	NR	NR	NR

The centroids in the third run of the algorithm for the test set is presented in Table 50.

TABLE 50 - IARTI EXPERIMENT – 3RD RUN TEST SET CENTROIDS

Variable	Full Data	Cluster 0	Cluster 1	Cluster 2
Frequency_Grade	NR	NR	NR	10-11
Exam	NR	10-11	FT	NR
Final_Grade	10-11	10-11	10-11	NC
Season	NORMAL	APPEAL	APPEAL	NORMAL
Enrollment_Year	2010-2011	2010-2011	2010-2011	2010-2011
Res	APPROVED	APPROVED	FAILED	APPROVED
Number_Subjects	8.2476	9.0294	7.4348	8.68
Number_Approvals	4.7048	6.5882	2.3478	6.48
Approval_Percentage	0.5429	0.7147	0.2913	0.772

Grade_Difference	NR	0.0	NR	0.0

The clustered instances are then presented in Table 51.

TABLE 51 - IARTI EXPERIMENT – 3RD RUN CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage
0	60	38%
1	54	34%
2	44	28%

Looking at the third run results and centroids it is possible to see that the sum of squared errors has slightly improved however the centroids didn't change much. This is positive for the clustering algorithm since a pattern is starting to be accomplished.

The fourth run of the algorithm was performed for 263 instances with a percentage split of 50%, 3 seeds and 3 clusters.

TABLE 52 - IARTI EXPERIMENT - 4TH RUN RESULTS

	Training Set	Test Set
Sum Of Squared Errors	644	258
Number of Iterations	4	6

By looking at Table 52 it is possible to see that the output this run has the worst output so far. Despite the centroids may not change much it is considered the best result since the number of clustered instances in this run is higher.

The centroids in the fourth run of the algorithm for the training set are presented in Table 53.

TABLE 53 - IARTI EXPERIMENT - 4TH RUN TRAINING SET

Variable	Full Data	Cluster 0	Cluster 1	Cluster 2
Frequency_Grade	NR	NR	NR	NR
Exam	NR	10-11	NR	9.0

Final_Grade	10-11	10-11	NC	9.0
Season	NORMAL	APPEAL	NORMAL	APPEAL
Enrollment_Year	2010-2011	2010-2011	2010-2011	2010-2011
Res	APPROVED	APPROVED	FAILED	FAILED
Number_Subjects	8.5209	9.3223	7.2989	8.6909
Number_Approvals	5.2814	7.2479	3.4483	3.8545
Approval_Percentage	0.5886	0.7628	0.4644	0.4018
Grade_Difference	NR	NR	NR	NR

The centroids in the fourth run of the algorithm for the test set is presented in Table 54.

TABLE 54 - IARTI EXPERIMENT - 4TH RUN TEST SET

Variable	Full Data	Cluster 0	Cluster 1	Cluster 2
Frequency_Grade	NR	NR	NR	NR
Exam	NR	NR	10-11	7-8
Final_Grade	10-11	NC	10-11	7-8
Season	NORMAL	NORMAL	APPEAL	APPEAL
Enrollment_Year	2010-2011	2010-2011	2010-2011	2007-2008
Res	APPROVED	FAILED	APPROVED	FAILED
Number_Subjects	8.2476	7.1951	7.1951	7.35
Number_Approvals	4.7048	3.2683	3.2683	2.25
Approval_Percentage	0.5429	0.4732	0.4732	0.28
Grade_Difference	NR	NR	NR	NR

The clustered instances are then presented in Table 55.

TABLE 55 – IARTI EXPERIMENT – 4TH RUN CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage
0	51	32%
1	84	53%
2	23	15%

Again, looking at the fourth run results it is possible to conclude that the number of incorrectly clustered instances has raised, so the best value to perform the clustering algorithm probably stands between 7 and 9. There can also be higher values with similar configurations but since a pattern begins to show up, results won't change much by swapping Simple K-Means configuration.

So let's analyze the third run corresponding to the clustering configuration with 9 seeds which seems to be the best result so far.

The improvement of the SSE is an indicator that the clustering algorithm has improved its results.

- Cluster 0 is filled with students that were approved to the subject in appeal season. Their frequency grade did not reached the minimum requirement or they didn't attended to it, ending up by being approved to the subject in exam grade. This cluster can be called 'approved-appeal'. An interesting fact is that this cluster has the biggest number of subjects students are attending to.
- Cluster 1 is the 'failed' cluster, where its instances are mainly negative for grades, results and also for the number of subjects students got approval. Most of the students tried be approved in appeal season which is possible to see by looking at the cluster centroid variable season.
- Cluster 2 is once again filled with approved students but that got it at their first try. Their frequency grade and their final grade centroid are similar and the number of subjects they are enrolled to and got approved is almost 100%. This cluster is mostly populated with the best students for the subject IARTI.

By analyzing the output from the clustering algorithm, the following conclusions are presented:

- Students from cluster 0 tried to pass the subject in exam grade and ended up by accomplishing it.
- The number of subjects enrolled in cluster 0 is a little higher than students in cluster 2, which means that students with more subjects per year struggle in being approved to subjects. This information is important since the number of subjects a student is attending to seems to decrease their results.
- For the test set data, the majority of students have been approved (66%).

- Students that failed the subject tend to fail way more subjects, as they have an approval rate of 30% or the subjects they are enrolled with.
- Students with entrance year 2010 and 2011 seem to be dominant in this data set.
- Grades obtained for this subject are around 10 and 11 values. Maybe this subject is slightly hard compared to others.

By running the clustering algorithm more times more conclusions can be withdrawn.

Graphics were also generated for this data set in order to retrieve more details from the clustering performed.

The Figure 50 demonstrates this output where:

- X axis represents the entrance year.
- Y axis represents the final results.
- Color represents the season.

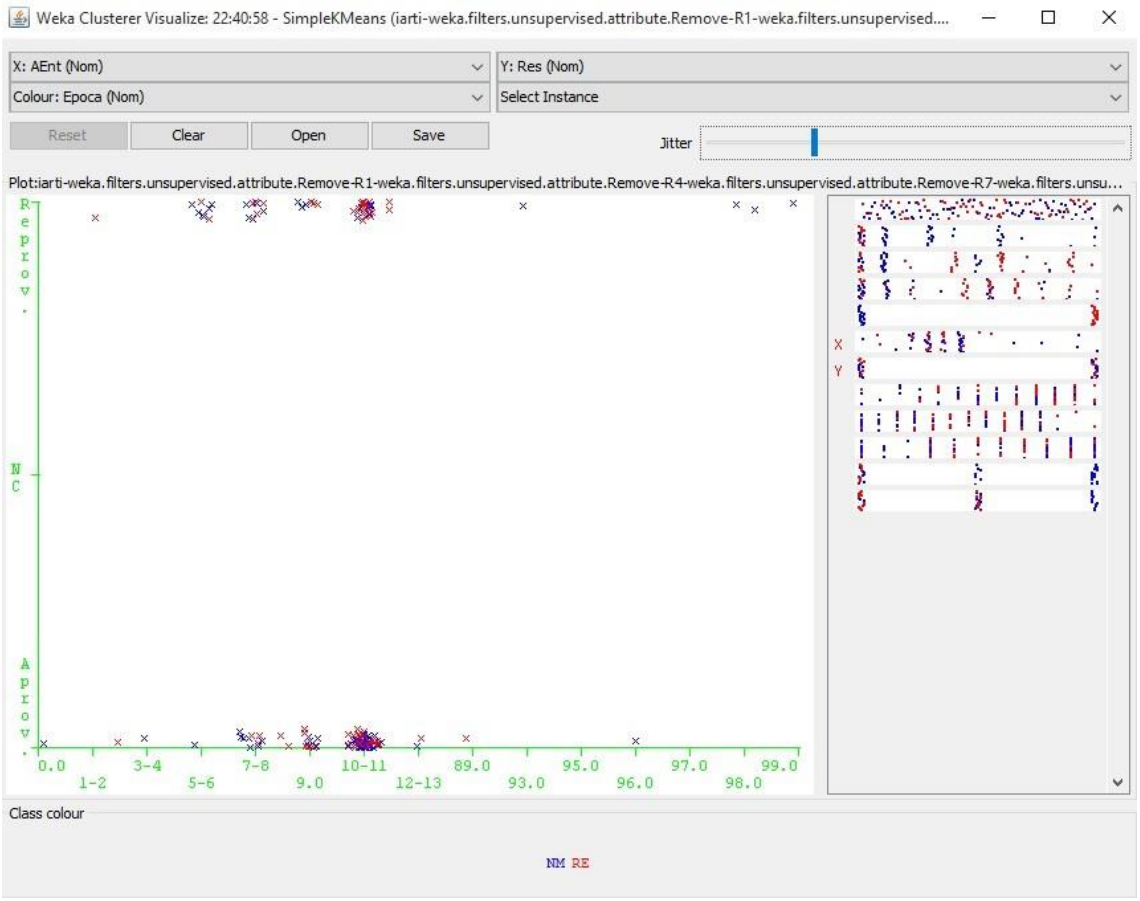

FIGURE 50 - IARTI EXPERIMENT - ENTRANCE YEAR VS RESULTS

Figure 50 shows clustered instances contrasting their entrance year and subject result with the season in which the final results were obtained.

- There is a heavy presence of students with entrance year 2010 and 2011, followed by 2005 to 2009.
- The number of students that failed the subject seem to be more dispersed accordingly their entrance year.
- There are more students that were approved in normal phase rather than appeal season.

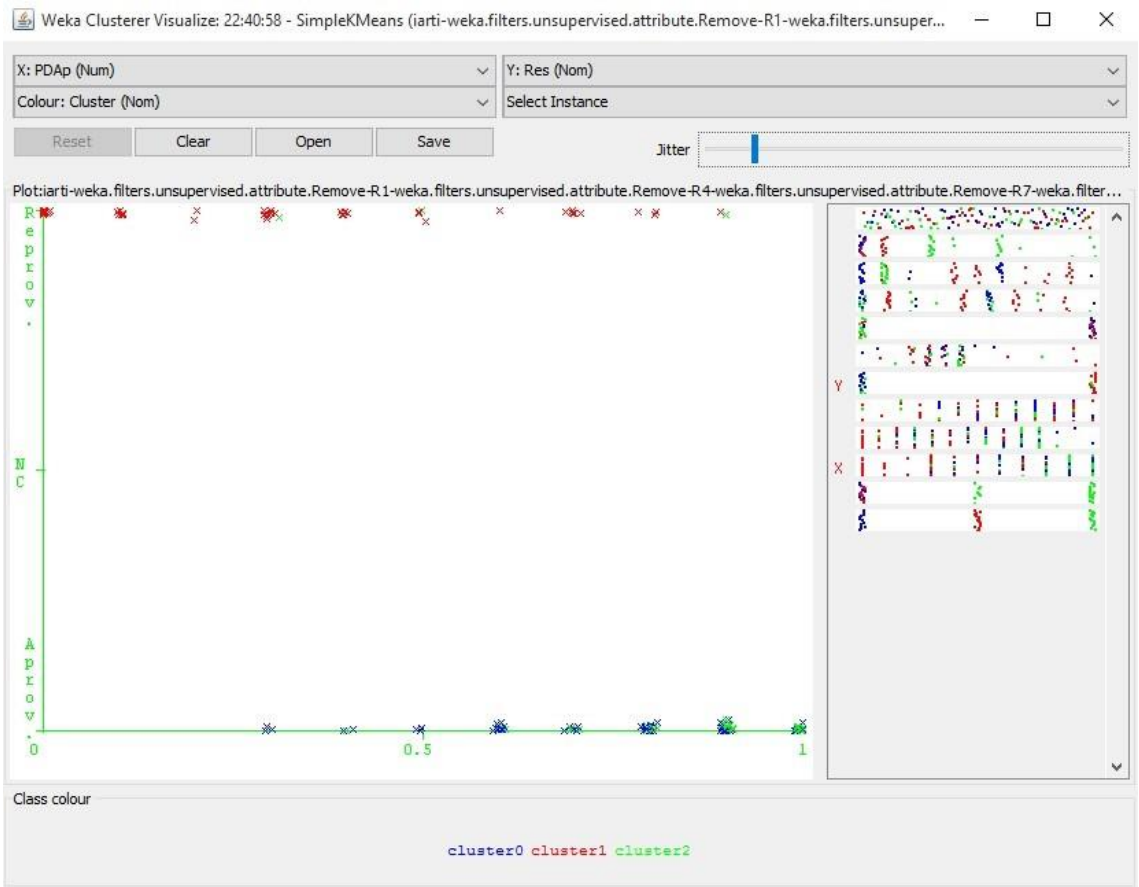

FIGURE 51 - IARTI EXPERIMENT - APPROVAL PERCENTAGE VS RESULTS

Figure 51 confirms that instances clustered are well placed according to their parameters. It is possible to see a slight distribution of the students final result according to the number of subjects they have been approved.

- Cluster 0 are approved students with approval percentage between 30% and 85%.

- Cluster 2 are approved students with approval percentage between 85% and 100%. These conclusions are aligned with the cluster centroid for each cluster.
- There is a heavy concentration of students with percentage approvals below 30% that failed this subject.
- There are a few students with approval percentage rate below 50% that have been approved to IARTI subject.

More conclusions can be retrieved from observing the outputs provided by Weka and by interacting more with the tool. This experiment is successful for proper analysis and study of clustering in Weka.

4.3.5.2. EXPERIMENT 2 – SUBJECT LAPR2

Experiment 2 refers to the subject LAPR2.

Clustering was processed with the following variables:

- Frequency Grade
- Exam Grade
- Final Grade
- Student entrance year
- Result
- Difference between final and exam grade

The first run of the algorithm was performed for 415 instances, 20 seeds and 2 clusters.

TABLE 56 – LAPR2 EXPERIMENT – RESULTS

	Training Set
Sum Of Squared Errors	820
Number of Iterations	5

By looking at Table 56 it is possible to see that the output has improved quite a bit.

The centroids for the training set are presented in Table 57.

TABLE 57 – LAPR2 EXPERIMENT – CENTROIDS

Variable	Full Data	Cluster 0	Cluster 1
Frequency_Grade	NF	12-13	NF
Exam	16-17	16-17	NR
Final_Grade	NF	12-13	NF
Enrollment_Year	2012-2013	2012-2013	2012-2013
Res	APPROVED	APPROVED	FAILED
Grade_Difference	1-2	1-2	NR

The clustered instances are then presented in Table 57.

TABLE 58 – LAPR2 EXPERIMENT – CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage
0	251	58%
1	174	42%

The clustered instances are presented in Table 58.

Analyzing the output, the centroids for cluster 0 and cluster 1 are presented in a table format compared to the full data.

- Cluster 0 is called the 'approval' cluster since the centroid contains mainly by positive results which means that the students in this cluster ended up increasing their final grade and having positive results. As we can see in the output, the centroids calculated for each variable are all positive comparing it to the full data available. It is possible to conclude that 241 students were clustered as approved being 58% of the full data set.
- Cluster 1 is called the 'failed' cluster since it is mainly composed by students that failed the subject minimum requirements or didn't attended to the subject. The instances contained in this cluster are mainly negative and represent students that failed. It is also possible to see that students in this cluster did not increase their final grade in the exam. It is possible to conclude that 174 students failed, representing 42% of the full data set.

This cluster shows that there is a significant difference between the exam and frequency grades, provoking changes to the final grades. By clustering it is possible to conclude that the exam is good for students to improve their final result at the subject, with the average exam grade for this cluster set between 16 and 17.

Another way to look at the results is to visualize the cluster graphically. Figure 52 demonstrates this output where:

- X axis represents the final grades
- Y axis represents the final result

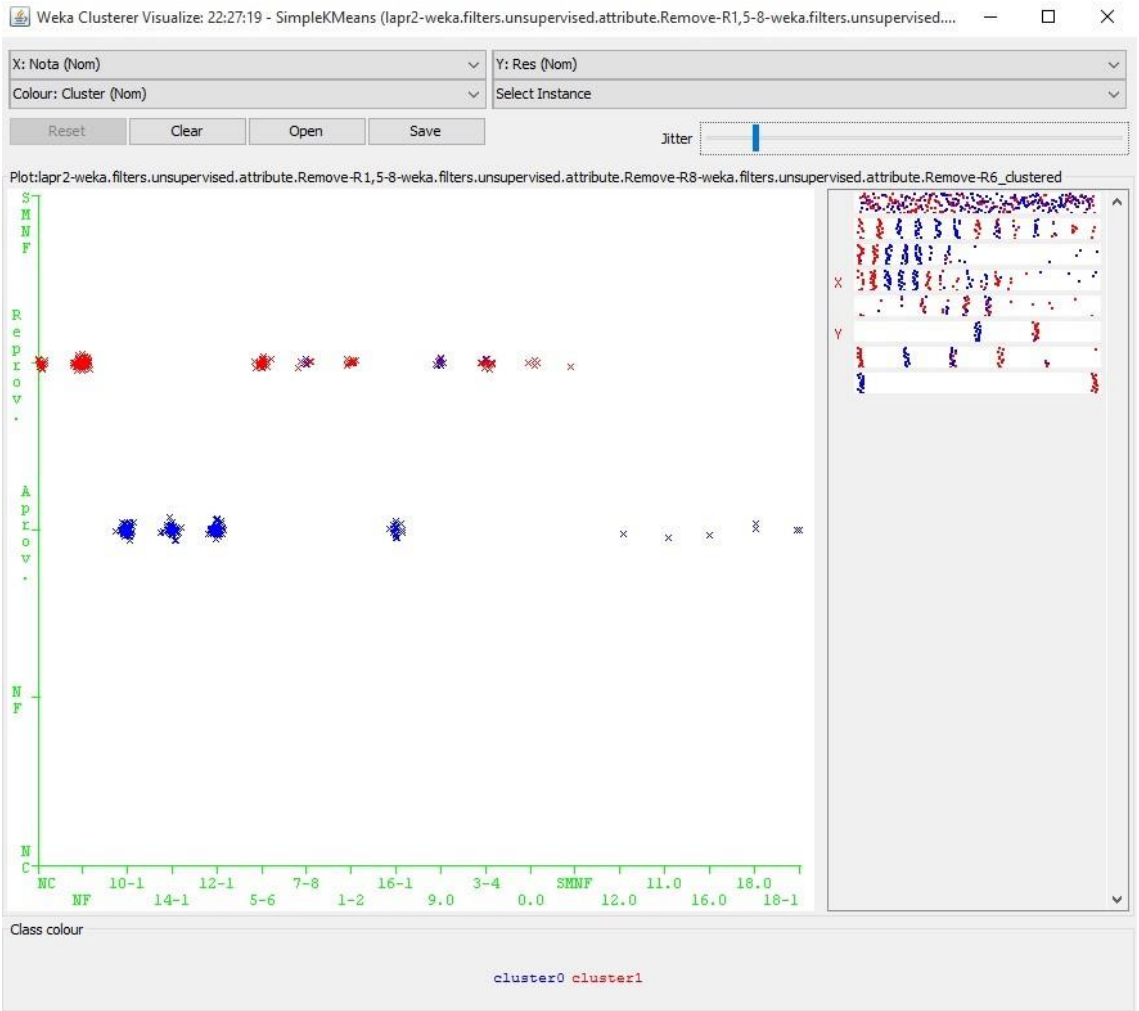

FIGURE 52 - LAPR2 EXPERIMENT – FINAL GRADE VS RESULTS

It is possible to see that cluster 1 results are correctly composed with instances of failed students as seen in the Y axis, and no records were found with positives final grades, X axis. The

graphic output also shows some mistakes performed by the clustering algorithm for this data set, where X axis with value 9.0 have some misplaced students.

Weka also has an option to click in each record and access a detailed view of the variables for the selected student presented in Table 59.

TABLE 59 - LAPR2 EXPERIMENT - INSTANCE DETAILS

```
Instance_number: 190.0
          Frequency_Grade: 7-8
          Exam: 16-17
           Final_Grade: 9.0
           Enrollment_Year: 12-13
            Res: Failed
 Grade_Difference: 1-2
        Cluster: cluster0

Instance_number: 95.0
          Frequency_Grade: NR
          Exam: 12-13
           Final_Grade: 9.0
           Enrollment_Year: 12-13
            Res: Failed
 Grade_Difference: NR
        Cluster: cluster1
```

In Table 58 there are presented 2 students, referred as instances, that were placed in X axis with final grade value of 9.0 and Y axis with the subject's result value of 'failed'. It is possible to conclude that despite instance 190 belongs to cluster 0, the student failed the subject with a high exam grade. The exam grade was not good enough to even the frequency grade. Some variables like the entrance year also promoted instances wrong placement in the cluster due to the similarity due to this variable's value being similar in both clusters.

4.3.5.3. EXPERIMENT 3 – SUBJECT LAPR3

Experiment 3 refers to the subject LAPR3.

For the record, clustering was processed with the following variables:

- Frequency Grade
- Exam Grade
- Final Grade
- Student entrance year
- Result
- Difference between final and exam grade

The algorithm was performed for 253 instances, 20 seeds and 2 clusters.

TABLE 60 – LAPR3 EXPERIMENT – RESULTS

	Training Set
Sum Of Squared Errors	970
Number of Iterations	4

By looking at Table 60 it is possible that the SSE value is quite high.

The centroids for the training set are presented in Table 61.

TABLE 61 – LAPR2 EXPERIMENT – CENTROIDS

Variable	**Full Data**	**Cluster 0**	**Cluster 1**
Frequency_Grade	12-13	12-13	NF
Exam	NR	NR	NR
Final_Grade	12-13	12-13	SMNF
Enrollment_Year	2012-2013	2012-2013	2012-2013
Res	APPROVED	APPROVED	FAILED
Grade_Difference	0-0	0-0	NR

The clustered instances are then presented In Table 62.

TABLE 62 – LAPR3 EXPERIMENT – CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage
0	213	84%
1	40	16%

Analyzing the output, the centroids for cluster 0 and cluster 1 are presented in a table format compared to the full data.

- Cluster 0 is called the 'approval' cluster since the centroid contains only positive results which means that the students in this cluster ended up by being approved to the subject. As we can see in the output, the centroids calculated for each variable are all positive comparing it to the full data available except for the exam grade which was not a requirement for this subject.

  We can conclude that 213 students were clustered as approved being 84% of the full data set. The average frequency and final grade is between 12 and 13 values and the difference between frequency and final grade is 0, which is true since the exam grade was not a requirement for this subject as mentioned above.
- Cluster 1 is called the 'failed' cluster since it is composed by students that failed the subject minimum requirements or didn't attended to the subject. The centroids presented to this cluster are mainly negative and their result is failed.

We can conclude that 40 students were clustered as failed being 16% of the full data set. Students that failed the frequency grade did not had the chance to improve their final grade by attending the exam grade.

This cluster has no improvements in frequency difference has no exams were performed for this particular subject.

A good way to prove this assertion, a graphic was built for this subject with the following values for the axis:

- X axis is the final grade
- Y axis is the frequency difference
- Color identifies clusters.

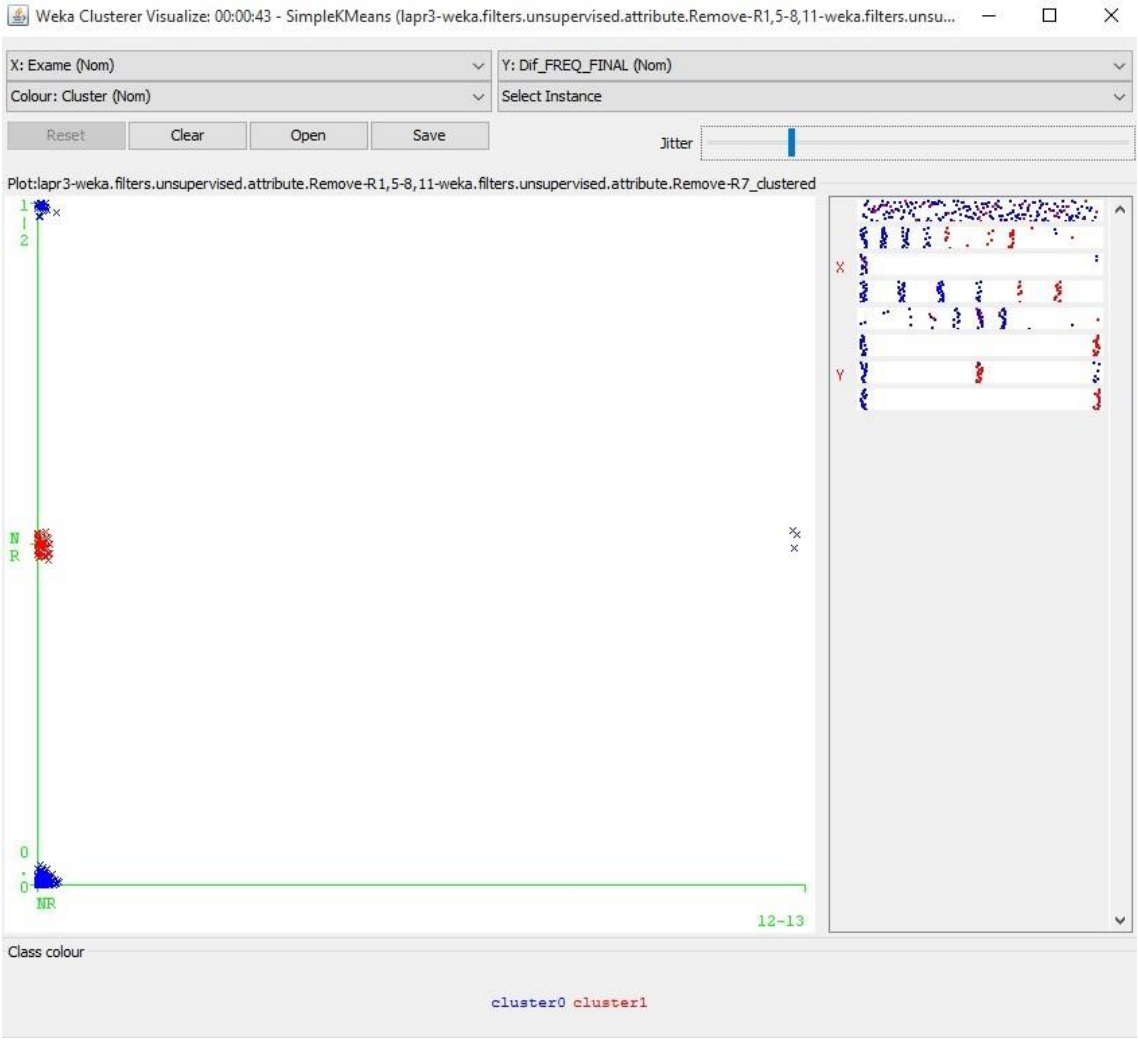

FIGURE 53 - LAPR3 EXPERIMENT - EXAM GRADE VS FREQUENCY GRADE DIFFERENCE

The graphic in Figure 53 can be easily read where only 3 students performed the exam grade, probably due to special status or needs and ended up with values between 12 and 13. The remaining students were placed in a NR value which means that frequency grade didn't suffered changes. It is possible to see a few bad calculated students were placed in the frequency difference with value 1-2. This errors occur since this 2 variables are not the only fields that are used to cluster students.

4.3.5.4. EXPERIMENT 4 – SUBJECT LAPR4

Experiment 4 refers to the subject LAPR4.

For the record, clustering was processed with the following variables:

- Frequency Grade

- Exam Grade
- Final Grade
- Student entrance year
- Result

The algorithm was performed for 297 instances, 20 seeds and 4 clusters.

TABLE 63 – LAPR4 EXPERIMENT – RESULTS

	Training Set
Sum Of Squared Errors	640
Number of Iterations	6

By looking at Table 63 it is possible that the SSE value is quite high.

The centroids for the training set are presented in Table 64.

TABLE 64 – LAPR4 EXPERIMENT – CENTROIDS

Variable	**Full Data**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Frequency_Grade	14-15	0.0	NF	12-13	16-17
Exam	NR	NR	NR	NR	NR
Final_Grade	14-15	SMNF	12-13	16-17	14-15
Enrollment_Year	2012-2013	2010-2011	2010-2011	2012-2013	2012-2013
Res	APPROVED	SMNF	APPROVED	APPROVED	APPROVED

The clustered instances are then presented In Table 65.

TABLE 65 – LAPR4 EXPERIMENT – CLUSTERED INSTANCES

Clustered Instances	**Instances**	**Percentage**
0	45	15%
1	104	35%
2	55	19%
3	93	31%

By looking at the cluster centroids in Table 64 there is a small difference in values between them like the experiments in the previous sections of this document. With more clusters, Weka struggles in finding different centroids by exploring more alternatives of calculating and building them. Centroids values for each variable become more close to each other.

For this subject, exam grade was not a requirement and following conclusions can be retrieved:

- Cluster 0 represents students that have no frequency grade and therefore failed the subject. It has 45 students which is 15% of the full data.
- Cluster 1 has 104 students and can be called 'average-grade'. The values in the centroid that differ them from the other clusters is the frequency and final grade values between 12 and 13.
- Cluster 2 has 55 students (19%) with frequency and final grades above 16 and can be called as 'high-grade'. This cluster has approved students to the subject with grades around 16 and 17.
- Cluster 3 is composed by approved 93 students (31%) to the subject and can be called 'mid-grade' with frequency and final grade between 14 and 15 values.

If more clusters were calculated for this data set, Weka would have trouble placing students accordingly the centroids calculated for each cluster. Results could start repeating themselves and ending up in different clusters.
A graphic was generated for this case in order to see the placement of students in 4 clusters.

The axis variables for the Figure 54 are the following:

- X axis is the final grade.
- Y axis is the result.

This figure provides a good look at the clusters placement accordingly the student's final grade and their result to the subject LAPR4.

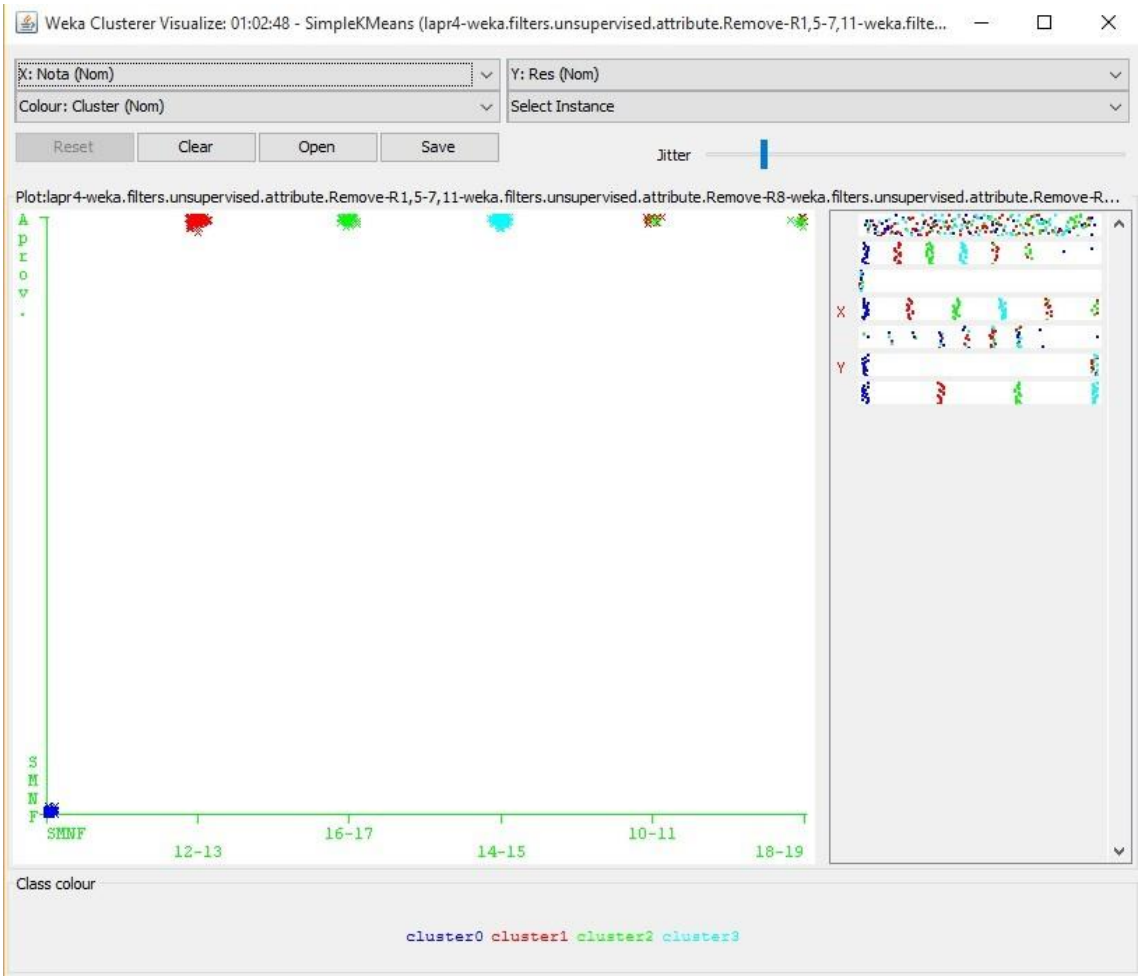

FIGURE 54 - LAPR4 - FINAL GRADE VS RESULTS

Cluster 0 is well separated from the rest of the clusters as it stands for students that failed the subject. The following clusters can be well separated from each other accordingly to the final grade.

There is a clear separation between students that were approved with final grades difference of 2 values.

Cluster 2 belongs to the 'high-grade' students and we can see that it has records for final grades of value 16 to 19.

Cluster 3 also provides a strong presence in final grades with value 14 and 15.

4.3.5.5. EXPERIMENT 5 – SUBJECT LPROG

Experiment 5 refers to the subject LPROG.

Clustering was processed with the following variables:

- Frequency Grade
- Exam Grade
- Final Grade
- Season, in this data set season can bring important information that was preprocessed before clustering.
- Student entrance year
- Result
- Difference between final and exam grade

The algorithm was performed for 338 instances, 16 seeds and 4 clusters.

TABLE 66 – LPROG EXPERIMENT – RESULTS

	Training Set
Sum Of Squared Errors	900
Number of Iterations	4

By looking at Table 66 it is possible that the SSE value is quite high.

The centroids for the training set are presented in Table 67.

TABLE 67 – LPROG EXPERIMENT – CENTROIDS

Variable	**Full Data**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Frequency_Grade	NR	NF	NR	NC	NR
Exam	10-11	10-11	7-8	FT	7-8
Final_Grade	10-11	12-13	SMS	SMNF	10-11
Season	NORMAL	NORMAL	APPEAL	NORMAL	APPEAL
Enrollment_Year	2012-2013	2012-2013	2010-2011	2012-2013	2012-2013
Res	APPROVED	APPROVED	FAILED	FAILED	APPROVED

The clustered instances are then presented In Table 68.

TABLE 68 – LPROG EXPERIMENT – CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage
0	165	49%
1	79	23%
2	42	12%
3	52	15%

Analyzing the output, we can reach conclusions for each cluster. The configuration changed in order to retrieve different approaches to the output. Four clusters were calculated instead of two and the number of seeds used was 10.

- Cluster 0 can be seen as 'normal approved' students that ended up with a final grade around 12 and 13 values. The season variable can tell us that for this cluster centroid, students that were placed in the cluster ended up by being approved to the subject in the normal season. It is possible to conclude that 165 students were clustered as approved being 49% of the full data set.
- Cluster 1 is called the 'appeal failed' cluster since it is mainly composed by students that failed the subject in the appeal season. Students placed in this cluster most likely couldn't reach the subject requirements ended up with SMS as final grade. It is possible to conclude that 79 students failed, representing 23% of the full data set.
- Cluster 2 can be called 'normal failed' cluster since it is composed by failed students like cluster 1 with a particular difference, the season. In this cluster the students failed the subject in the normal phase and did not tried to be approved in the appeal season. This particular information can be reached due to the preprocessing that was performed for this data set. It is possible to conclude that 42 students failed, representing 12% of the full data set.
- Cluster 3 is the last cluster and can be called 'appeal approved' once again due to the season variable. This cluster has students that were approved to the subject but only in appeal season. These students attended to the subject in the normal season phase but did not reach the minimum requirements. Once again, preprocessing done to the data set avoided mixing those students in the clustering algorithm.

Another way to look at the results is to visualize the cluster graphically where the axis values are:

- X axis represents the season.
- Y axis represents the final result.

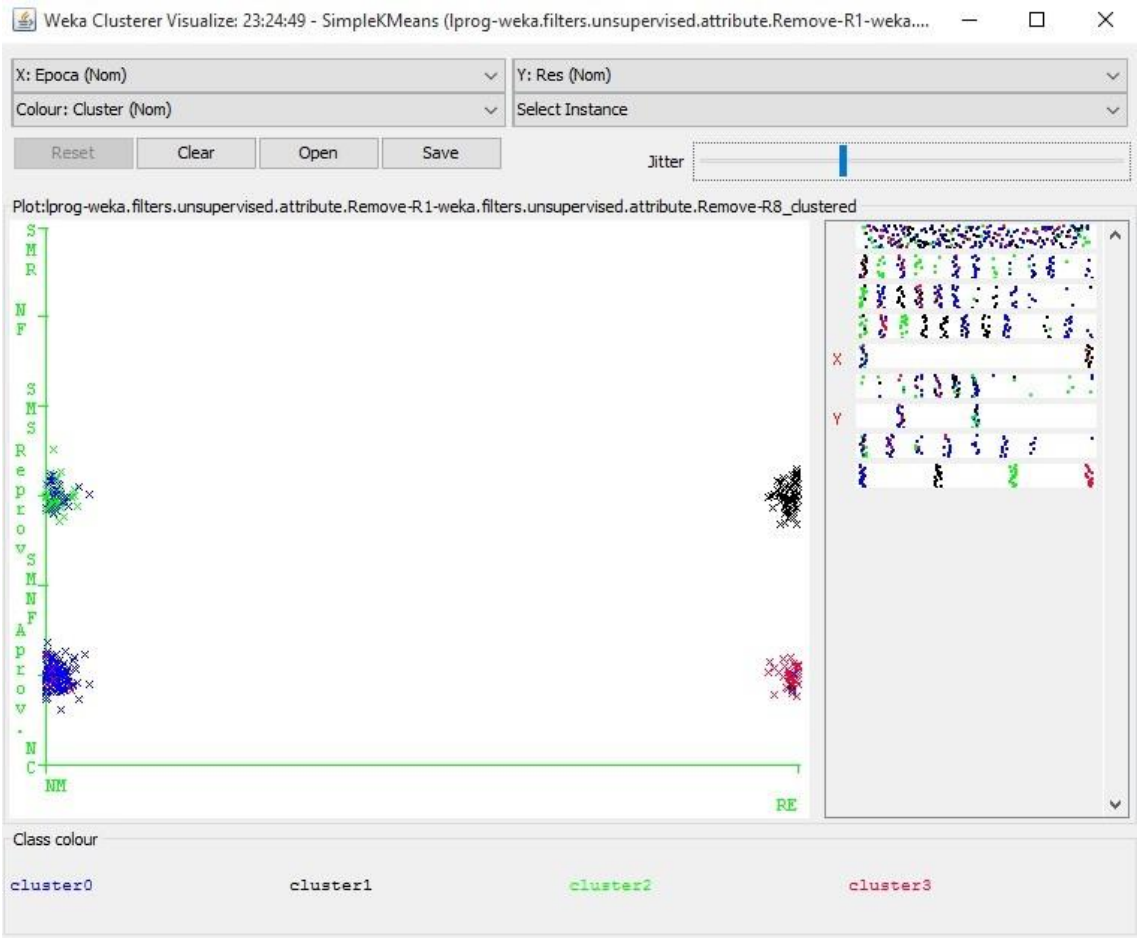

FIGURE 55 - LPROG EXPERIMENT - SEASON VS RESULTS

By looking at the clustering algorithm output it was clear that the 4 clusters were separated mainly by the variables final result, season and final grade.

Figure 55 can offer that same information with a strong impact. A few misplaced students from cluster 0 were placed in cluster 2. This misplacement can be analyzed by retrieving a few instances details presented in Table 69.

TABLE 69 – LPROG EXPERIMENT - INSTANCE DETAILS

Instance_number:	**330.0**
Frequency_Grade:	**NF**
Exam:	**NR**

```
            Final_Grade: NF
           Season: NM
            Enrollment_Year: 96.0
             Res: Failed
  Grade_Difference: NR
         Cluster: cluster0

 Instance_number: 332.0
           Frequency_Grade: 5-6
           Exam: FT
            Final_Grade: SMNF
           Season: NM
            Enrollment_Year: 98.0
             Res: Failed
  Grade_Difference: NR
         Cluster: cluster2
```

Table 69 has 2 students that are in different clusters but ended up both by failing the subject. However at least 4 variables are different for both of them which is the root cause of ending close to each other but in different clusters.

4.3.5.6. EXPERIMENT 6 – SUBJECT FSIAP

Experiment 6 refers to the subject FSIAP. This particular subject has interesting conclusions. Clustering was processed with the following variables:

- Frequency Grade
- Exam Grade
- Final Grade
- Season
- Student entrance year
- Result
- Class
- Difference between final and exam grade

The algorithm was performed for 441 instances, 13 seeds and 3 clusters.

TABLE 70 – FSIAP EXPERIMENT – RESULTS

	Training Set
Sum Of Squared Errors	860
Number of Iterations	6

By looking at Table 70 it is possible that the SSE value is quite high.

The centroids for the training set are presented in Table 71.

TABLE 71 – FSIAP EXPERIMENT – CENTROIDS

Variable	Full Data	Cluster 0	Cluster 1	Cluster 2
Frequency_Grade	NR	14-15	NR	NR
Exam	NR	NR	3-4	FT
Final_Grade	10-11	10-11	7-8	NC
Season	NORMAL	NORMAL	APPEAL	NORMAL
Enrollment_Year	2012-2013	2012-2013	2010-2011	2010-2011
Res	APPROVED	APPROVED	FAILED	FAILED
Class	ST	2DG	ST	ST
Grade_Difference	NR	NR	NR	NR

The clustered instances are then presented In Table 72.

TABLE 72 – FSIAP EXPERIMENT – CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage
0	238	54%
1	104	24%
2	99	22%

Analyzing the output, we can reach conclusions for each cluster. The configuration changed in order to retrieve different approaches to the output. Three clusters were calculated instead of two and the number of seeds used was 2.

This particular subject is very interesting since it suffered a lot of changes during its lifecycle. Many students often fail this subject or decrease their final grade compared to the frequency grade obtained.

- Cluster 0 can be seen as 'approved' students. Most of the students in this cluster ended up by being approved to the subject despite their frequency grade is inferior to their final grade. An interesting fact is that class 2DG and entrance year 2012 and 2013 are mainly composed by approved students. It is possible to conclude that 238 students failed, representing 54% of the full data set.
- Cluster 1 is called the 'appeal failed' cluster since it is mainly composed by students that failed the subject in the appeal season. Students placed in this cluster most likely couldn't reach the subject exam grade minimum requirements and ended up by failing the subject. However, it is very important to see that this cluster is filled with students that tried to attend the subject but couldn't pass it after all. It is possible to conclude that 104 students failed, representing 24% of the full data set.
- Cluster 2 can be called 'normal failed' cluster since it is composed by failed students like cluster 1 with a particular difference, they did not attended the subject. It is possible to conclude that 99 students failed, representing 22% of the full data set.

The class centroid for cluster 1 and 2 is ST which means that most of the students do not own a class. This kind of information is also valuable to see how many students don't even try to attend to the subject.

Another way to look at the results is to visualize the cluster graphically. For this option a variable must be assigned to the X axis and another variable to the Y axis:

- X axis represents the entrance year.
- Y axis represents the class.

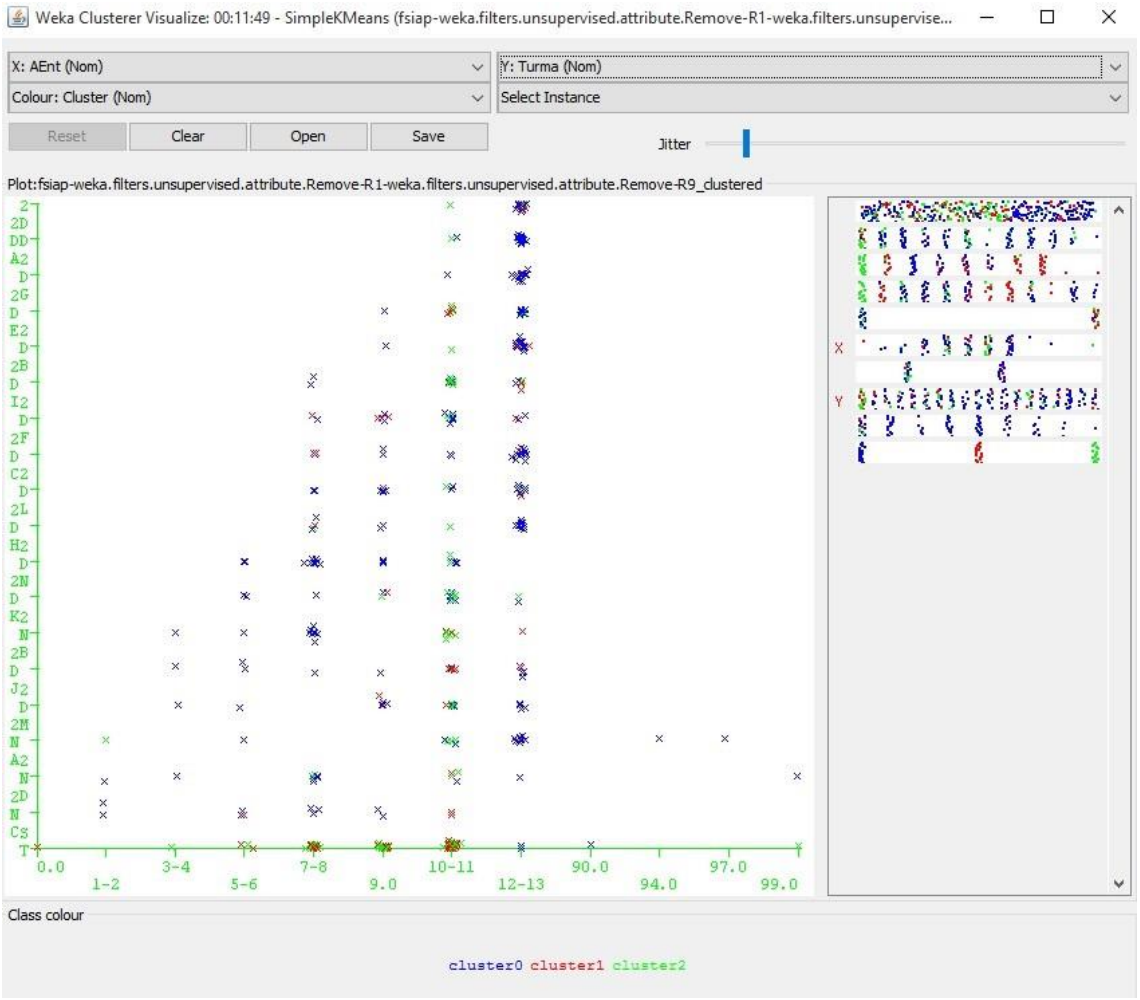

FIGURE 56 - FSIAP EXPERIMENT - ENTRANCE YEAR VS CLASS

As stated before, the entrance years with most number of approved students was 2012 and 2013. The Figure 56 contains that information clearly and it offers a very good output. Compared to Figure 57 obtained from the preprocessing window, it possible to see that years 2012 and 2013 are 35,37% of the full data.

No.	Label	Count
1	0.0	1
2	1-2	4
3	3-4	5
4	5-6	15
5	7-8	60
6	9.0	51
7	10-11	144
8	12-13	156
9	90.0	1
10	94.0	1
11	97.0	1
12	99.0	2

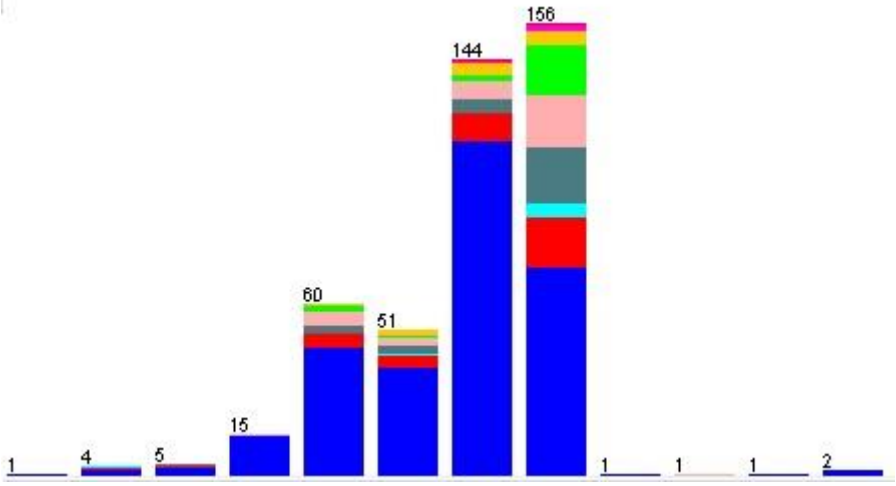

FIGURE 57 – FSIAP EXPERIMENT - PREPROCESSING DETAILS

Despite the entrance years 2012 and 2013 are 35% of the full data it is possible to conclude that students from this years had average grades and ended up by being approved more often that students from other years.

Another graphic was generated with another kind of conclusion that we may retrieve from the same clustering iteration with the following configuration:

- X axis represents the final grade.
- Y axis represents the frequency grade.

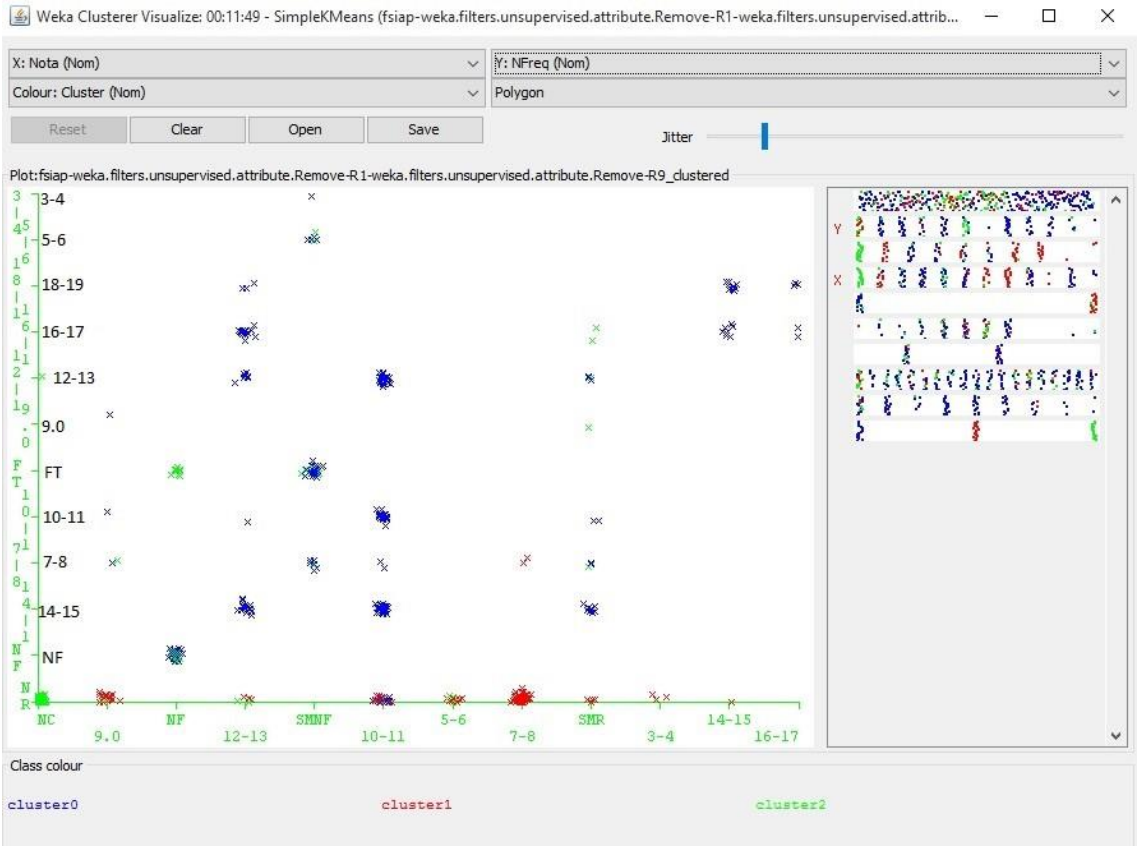

FIGURE 58 - FSIAP EXPERIMENT - FINAL GRADE VS FREQUENCY GRADE

Figure 58 demonstrates a clear and distinct placement of students accordingly to their clusters centroids. This figure had to be manipulated in order retrieve conclusions.

Cluster 0, blue, is mainly populated by approved students that are placed all around the graphic with different kind of final grade values. There are several portions of clustered students that can offer good conclusions, like:

- X axis with value 10-11
- Y axis with value 14-15

Students located in this coordinates have lowered their frequency grade by 4-5 values compared to their final grade.

- X axis with value 7-8
- Y axis with value NR

For cluster 1, red, a portion of students did not attended the subject in its normal phase and tried to pass it by attending only to the exam grade. For this particular case the final grade is equal to the exam grade accordingly to the subject requirements for that academic year.

Another graphic was generated with another kind of conclusion that we may retrieve from the same clustering iteration, with the following axis variables:

- X axis represents the frequency difference.
- Y axis represents the number of students.

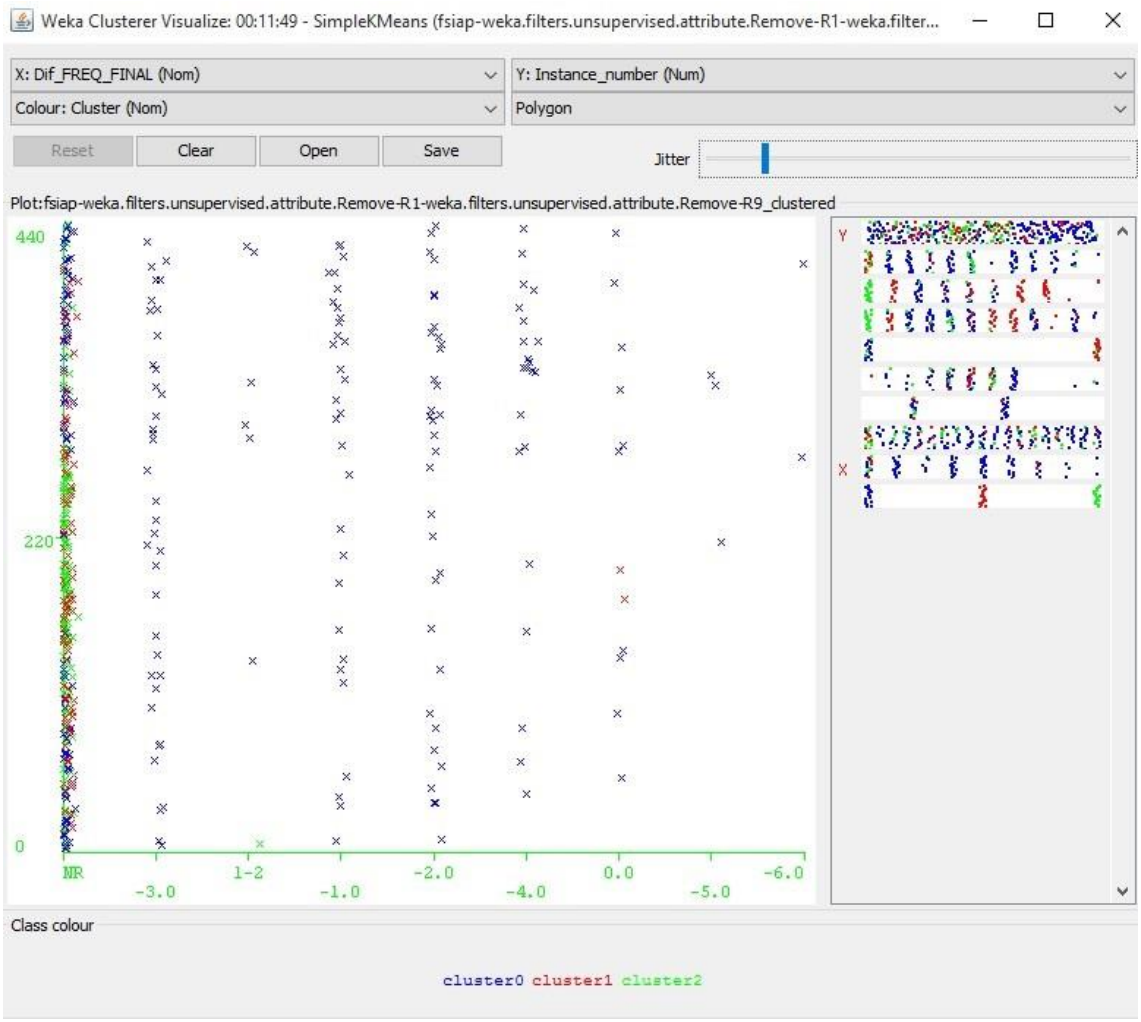

FIGURE 59 - FSIAP EXPERIMENT - FREQUENCY DIFFERENCE VS INSTANCE NUMBER

By looking at Figure 59 it is possible to see that cluster 1 and cluster 2 are mainly placed in the X axis NR which means that these students didn't attended to the subject. A few students from

cluster 1 are placed in a frequency difference of 0, despite they attended the subject it was not possible to reach the required grade to be approved.

Many conclusions can be obtained by looking at one of the possible data comparison that can be retrieved from clustering. There are more variables that can be added to the recipe to retrieve different outputs.

4.3.5.7. EXPERIMENT 7 – SUBJECT ALGAN

Experiment 7 refers to the subject ALGAN.
Clustering was processed with the following variables:

- Frequency Grade
- Exam Grade.
- Final Grade.
- Season.
- Regime.
- Schedule.
- Student entrance year.
- Result.
- Class.
- Difference between final and exam grade.

The algorithm was performed for 369 instances, 9 seeds and 3 clusters.

TABLE 73 – ALGAN EXPERIMENT – RESULTS

	Training Set
Sum Of Squared Errors	978
Number of Iterations	3

By looking at Table 73 it is possible that the SSE value is quite high.

The centroids for the training set are presented in Table 74.

TABLE 74 – ALGAN EXPERIMENT – CENTROIDS

Variable	Full Data	Cluster 0	Cluster 1	Cluster 2
Frequency_Grade	NR	NR	10-11	16-17
Exam	NR	FT	NR	NR
Final_Grade	10-11	NC	10-11	16-17
Season	NORMAL	APPEAL	NORMAL	NORMAL
Regime	INTEGRAL	INTEGRAL	INTEGRAL	INTEGRAL
Schedule	DIURNAL	DIURNAL	DIURNAL	DIURNAL
Enrollment_Year	2012-2013	2012-2013	2012-2013	2012-2013
Res	APPROVED	FAILED	APPROVED	APPROVED
Class	1DQ	1DP	1NB	1DG
Grade_Difference	0.0	NR	0.0	0.0

The clustered instances are then presented In Table 72.

TABLE 75 – ALGAN EXPERIMENT – CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage
0	126	34%
1	203	55%
2	40	11%

Analyzing the output, the centroids for cluster 0, 1 and 2 are presented in a table format compared to the full data. The following conclusions may be retrieved:

- Cluster 0 is called ‘failed’ and students placed here didn’t attended to the subject requirements or most of them tried to complete it in appeal season. The cluster centroid is mainly composed by negative or failed results in each subject grade. It is possible to conclude that 126 students were clustered as approved being 34% of the full data set.
- Cluster 1 is called the ‘mid-grade approved’ cluster since it is mainly composed by students that passed the subject with an average grade of 10-11 values. The instances contained in this cluster are mainly positive. It is also possible to see that students in this cluster did not increase their final grade in the exam accordingly to the cluster

centroid. It is possible to conclude that 203 students failed, representing 55% of the full data set.

- Cluster 2 can be seen as an approved students cluster but with high grades. This cluster can be called ‘high-grade approved’ since most of the students ended up with grade above 16 values. It is possible to conclude that 40 students failed, representing 11% of the full data set. Since ending a subject with a grade above 16 is way more difficult to accomplish, this cluster only represents 11% of the full data set which is true.

For this particular data set, more variables were introduced in the clustering algorithm in order to offer more conclusions and see how Simple K-Means reacts with more information. Introducing more variables will result in higher misplaced instances however it brings useful conclusions.

Another way to look at the results is to visualize the cluster graphically. The following axis variables are applied to the graphic:

- X axis represents the final grades
- Y axis represents the final result

FIGURE 60 - ALGAN EXPERIMENT - FINAL GRADE VS RESULTS

Analyzing the Figure 60 it is possible to check that cluster 2 is mainly populated in final grades around 16 and 17 values for students that have been approved, cluster 1 is populated in final grades from 10 to 15 values that have been approved and finally cluster 0 is populated in failed Y axis for grades from 1 to 9 values and also for students that didn't reached the minimum final grade requirements to the subject. We can also check this behavior using the third dimension attribute present in cluster graphics.

Like mentioned in the experiments configurations, graphics have 3 dimensions. By changing the color dimension to other attributes, it is possible to see their distribution within each of the clusters.

FIGURE 61 - ALGAN EXPERIMENT – INSTANCE NUMBER VS CLUSTER COLORED FREQUENCY GRADE

Figure 61 has frequency grades as colors, while the X's axis is the number of instances and the Y's axis represents the clusters. This results in a visualization of the distribution of the frequency grades in each cluster.

It is also possible to change the class color for each variable. In this experiment, the data was prepared as:

- Black stands for negative grades.
- Green and pink for grades around 9 and 12.
- Blue for midterm grades from 12 to 16.
- Yellow for grades above 16.

4.3.5.8. EXPERIMENT 8 – SUBJECT CORGA

Experiment 8 refers to the subject CORGA and percentage split was applied again.

Clustering was processed with the following variables:

- Frequency Grade
- Exam Grade
- Final Grade
- Student entrance year
- Result
- Difference between final and exam grade

The algorithm was performed for 164 instances with a percentage split of 50%, 8 seeds and 2 clusters.

TABLE 76 - CORGA EXPERIMENT – RESULTS

	Training Set	**Test Set**
Sum Of Squared Errors	430	190
Number of Iterations	3	2

Table 76 provides a good sum of squared errors value obtained for both sets, however the test set has shown accurate results. This was obtained because the number of instances is lower than the usual experiments referred in this document section. The number of variables is also lower which helps providing the algorithm efficiency.

The centroids for the test set are presented in Table 77.

TABLE 77 – CORGA EXPERIMENT – TEST SET CENTROIDS

Variable	**Full Data**	**Cluster 0**	**Cluster 1**
Frequency_Grade	14-15	14-15	14-15
Exam	14-15	14-15	12-13
Final_Grade	14-15	14-15	14-15
Enrollment_Year	2010-2011	2010-2011	2010-2011

Res	APPROVED	APPROVED	APPROVED
Grade_Difference	-1.0	0.0	-1.0

The centroids for the training set are presented in Table 78.

TABLE 78 – CORGA EXPERIMENT – TRAINING SET CENTROIDS

Variable	Full Data	Cluster 0	Cluster 1
Frequency_Grade	14-15	14-15	10-11
Exam	14-15	FT14-15	NR
Final_Grade	14-15	14-15	10-11
Enrollment_Year	2010-2011	2010-2011	2005-2006
Res	APPROVED	APPROVED	NF
Grade_Difference	0.0	0.0	NR

The clustered instances are then presented In Table 79.

TABLE 79 – CORGA EXPERIMENT – CLUSTERED INSTANCES

Clustered Instances	Instances	Percentage
0	73	89%
1	9	11%

By looking at the number of clustered instances it seems that every student with this subject shares the same values of the full data centroid. The percentage split reduced the number of students calculated in the training set, however the calculated prediction is very accurate.

According to previous tables, the output from a percentage split is slight different. It is possible to see that the centroids have changed from one clustering to another. The clustering of the training set ended up with the following conclusions:

- Cluster 0 is mainly populated by students with grades above 14 values, approved to the subject and with no differences in their frequency and final grade.
- Cluster 1 is very similar to cluster 0, however the exam grade lowered by 2 values and the difference between frequency and final grade for this case is -1. Students in this cluster had worse grades than students in cluster 0, but they ended up being approved to the subject.

The training set didn't offered a very good output for all the instances of the data set, however looking at the output from the test set final conclusions can be obtained.

- Cluster 0 didn't suffered any change, and stands for the 'approved' students cluster, being 89% of the full data set.
- Cluster 1 has changed completely ending up with 'failed' students that couldn't be approved to the subject. This cluster represents only 11% of the full data set.

The full data set is 50% of the full data set and has a very low number of not approved students, however, the training set aided simple K-means to calculate and cluster students in their correct position. This is a clear example that despite splitting the data, the output is much better and offers better conclusions to the user for this subject. Also notice in Table 76 that the sum squared error has lowered due to the fact that less instances are calculated and the algorithm's final result has been enhanced.

Another way to look at the results is to visualize the cluster graphically. Figure 62 graphic has the following variables for the axis values:

- X axis represents the final grades
- Y axis represents the clusters.
- Color represents the final results.

FIGURE 62 - CORGA EXPERIMENT - FINAL GRADE VS CLUSTER COLORED RESULTS

Figure 62 is very simple and easy to read, blue are students that ended up in the ‘approved’ cluster and they are populated with final grades from 10 to 17 and red are students that didn’t reached the subject minimum requirement and ended up with NF. They are also placed in the correct clusters.

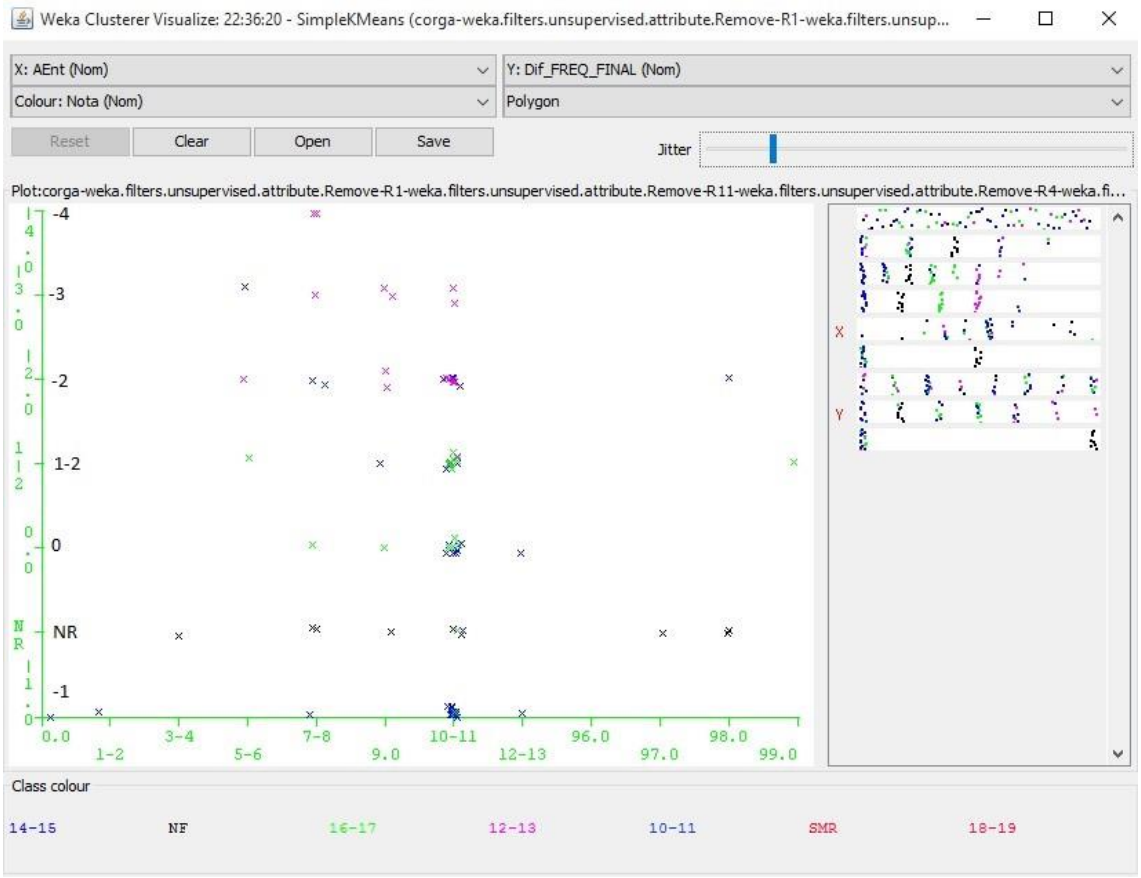

FIGURE 63 - CORGA EXPERIMENT - ENTRANCE YEAR VS FREQUENCY DIFFERENCE COLORED FINAL GRADE

Figure 63 was slightly manipulated in order to see the values for the Y axis. Students in this graphic are distributed accordingly their entrance year and final and frequency grade difference and colored with the grades they had. Some conclusions can be taken from looking at this figure, like:

- Years 2010 and 2011 seem to be more populated than others as they have 109 instances of the full data set. The rest of the years summed have 55 instances.
- Students from 1997 and 1998 were not approved to the subject.
- Students from 2000, 2001 and 2002 were approved but ended up by decreasing 1 value in their final grade.
- Students from 2007 and 2008 had the biggest decrease in frequency grade difference. They ended up with final grades of 12 and 13, which means that their frequency grade was around 16 and 17 values.
- Students with final grades of 16 and 17 from 2010 and 2011 improved their final grade by 1 or 2 values while some maintained it.

- A lot of students from year 2010 and 2011 that had final grades of 10 and 11 values ended up by losing 1 value.

4.3.6. CONCLUSIONS

After presenting and studying the experiments presented in the above sections it is possible to conclude that Weka offers many options to apply one of the available clustering algorithms in many ways. Accordingly to the user requirements we can reach different conclusions by changing standard parameters in clustering configuration. When trying to reach a conclusion for clustering, the user has the need to understand what can be accomplished to help simple K-means the best route to reach it.

It requires several tests in the same data set in order to provide a better output. Errors are part of the equation, misplaced instances cannot be seen as a threat. In order to reduce the error while applying clustering algorithms, data requires preprocessing, configuration and testing. Using probabilistic values also boosts the final output.

5. Application Development

This chapter describes the application's engineering requirements, as well as its architecture and implementation techniques.

5.1. Engineering Requirements

5.1.1. Use Cases

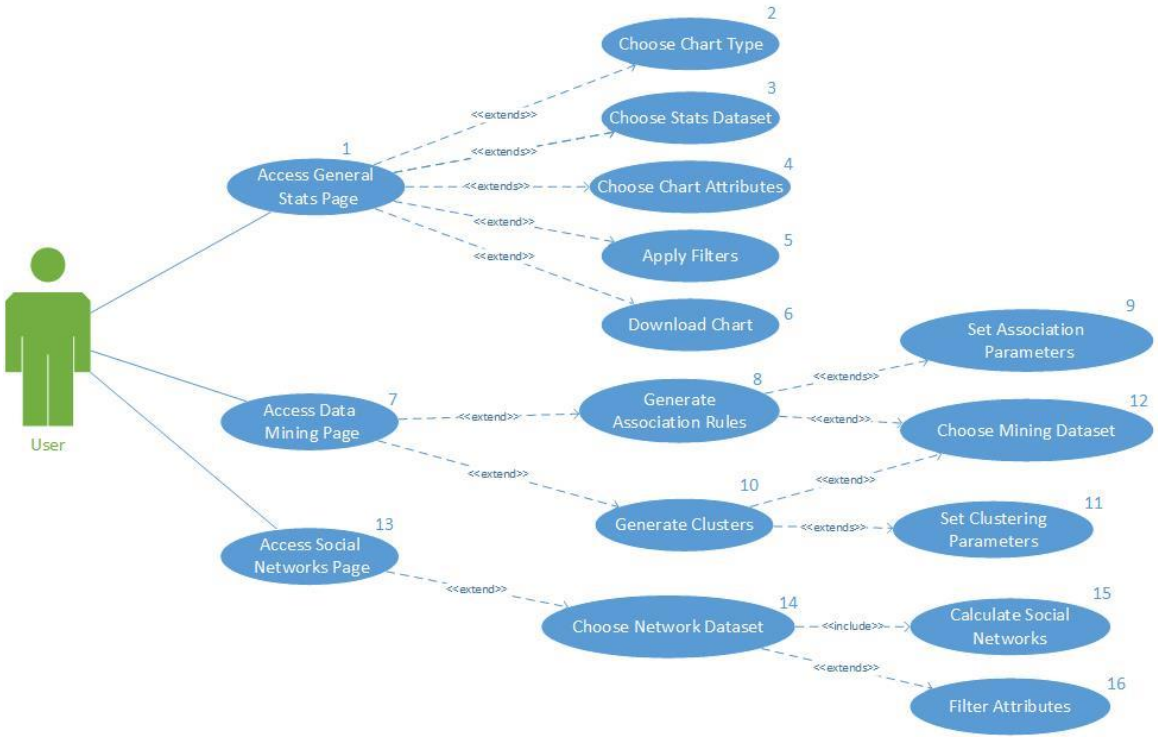

FIGURE 64 - USE CASES DIAGRAM

TABLE 80 - USE CASES - ACCESS GENERAL STATS

Use Case 1 – Access General Stats Page

Use Case ID	1
Use Case Name	Access General Stats Page
Description	Access the general statistics page

Actor	User
Pre-conditions	N.A.
Post-conditions	N.A.

Event Flow

Actor Actions	System Actions
1. Selects the general statistics page	Presents general statistics page

TABLE 81 - USE CASES - CHOOSE CHART TYPE

Use Case 2 – Choose Chart Type

Use Case ID	2
Use Case Name	Choose Chart Type
Description	User selects a chart representation
Actor	User
Pre-conditions	Dataset is sucessfully loaded
Post-conditions	N.A.

Event Flow

Actor Actions	System Actions
1. Selects a new representation	Saves selected representation
	Updates available representations
	Provides updated representation

TABLE 82 - USE CASES – CHOOSE STATS DATASET

Use Case 3 – Choose Stats Dataset

Use Case ID	3
Use Case Name	Choose Stats Dataset
Description	User selects a dataset to load
Actor	User
Pre-conditions	N.A
Post-conditions	Dataset is sucessfully loaded

Event Flow

Actor Actions	System Actions
1. Selects a new dataset to load	Saves chosen dataset Updates dataset list Loads dataset Updates representation list for dataset

TABLE 83 - USE CASES – CHOOSE CHART ATTRIBUTES

Use Case 4 – Choose Chart Attributes

Use Case ID	4
Use Case Name	Choose Chart Attributes
Description	User selects the dataset attributes the chart should consider

Actor	User
Pre-conditions	Dataset is sucessfully loaded
Post-conditions	N.A

Event Flow

Actor Actions	System Actions
1. Selects two attributes for the chart to consider	Saves chosen attributes
	Updates attribute list
	Provides updated representation

TABLE 84 - USE CASES – APPLY FILTERS

Use Case 5 – Apply Filters

Use Case ID	5
Use Case Name	Apply filters
Description	User applies a filter in the data
Actor	User
Pre-conditions	Dataset is successfully loaded Data is segmented Chart representation is drawn
Post-conditions	N.A

Event Flow

Actor Actions	System Actions

1. Changes the values considered by the filter provided	Applies chosen values
	Provides updated representation

TABLE 85 - USE CASES – DOWNLOAD CHART

Use Case 6 – Download Chart

Use Case ID	6
Use Case Name	Download Chart
Description	User downloads the chart being visualized
Actor	User
Pre-conditions	Dataset is successfully loaded Chart representation is drawn
Post-conditions	Chart's image is saved

Event Flow

Actor Actions	System Actions
1. Requests the visualization to be downloaded	Converts chart into image
2. Chooses image name	Prompts user to save image
3. Chooses image directory	

TABLE 86 - USE CASES – ACCESS DATA MINING

Use Case 7 – Access Data Mining Page

Use Case ID	7

Use Case Name	Access Data Mining Page
Description	User accesses the data mining page
Actor	User
Pre-conditions	N.A
Post-conditions	User selects mining technique to apply

Event Flow

Actor Actions	System Actions
1. Accesses the data mining page	Presents data mining page
2. Chooses a data mining technique to apply	Prompts user to choose a data mining technique

TABLE 87 - USE CASES – GENERATE ASSOCIATION RULES

Use Case 8 – Generate Association Rules

Use Case ID	8
Use Case Name	Generate Association Rules
Description	User activates the associations rules option
Actor	User
Pre-conditions	Dataset is successfully loaded
Post-conditions	N.A

Event Flow

Actor Actions	System Actions

1. Requests association rules over selected dataset	Calculates association rules Builds chart with association rules Presents chart to the user

TABLE 88- USE CASES – SET ASSOCIATION PARAMETERS

Use Case 9 – Set Association Parameters

Use Case ID	9
Use Case Name	Set Association Parameters
Description	User selects the parameters for association rule mining
Actor	User
Pre-conditions	Dataset is successfully loaded
Post-conditions	Parameters are valid

Event Flow

Actor Actions	System Actions
1. Inserts new parameters for association rule mining	Validates the inserted parameters Calculates association rules for new parameters Builds chart with association rules Presents chart to the user

TABLE 89- USE CASES – GENERATE CLUSTERS

Use Case 10 – Generate Clusters

Use Case ID	10

Use Case Name	Generate Clusters
Description	User activates the clustering option
Actor	User
Pre-conditions	Dataset is successfully loaded
Post-conditions	N.A.

Event Flow

Actor Actions	System Actions
1. Requests clustering over selected dataset	Calculates clusters Builds chart with calculated clusters Presents chart to the user

TABLE 90 - USE CASES – SET CLUSTERING PARAMETERS

Use Case 11 – Set Clustering Parameters

Use Case ID	11
Use Case Name	Set Clustering Parameters
Description	User selects the parameters for clustering
Actor	User
Pre-conditions	Dataset is successfully loaded
Post-conditions	Parameters are valid

Event Flow

Actor Actions	System Actions
1. Inserts new parameters for clusters mining	Validates the inserted parameters Calculates clusters for new parameters Builds chart with calculated clusters Presents chart to the user

TABLE 91- USE CASES – CHOOSE MINING DATASET

Use Case 12 – Choose Mining Dataset

Use Case ID	12
Use Case Name	Choose Mining Dataset
Description	User selects the dataset for mining purposes
Actor	User
Pre-conditions	N.A.
Post-conditions	Dataset is successfully loaded

Event Flow

Actor Actions	System Actions
1. Selects a new dataset to load	Saves chosen dataset Updates dataset list Loads dataset Updates representation list for dataset

TABLE 92- USE CASES – ACCESS SOCIAL NETWORKS PAGE

Use Case 13 – Access Social Networks page

Use Case ID	13
Use Case Name	Access Social Networks page
Description	User selects the Social networks page
Actor	User
Pre-conditions	N.A.
Post-conditions	N.A.

Event Flow

Actor Actions	System Actions
1. Accesses the Social networks page	Presents data Social networks page

TABLE 93- USE CASES – CHOOSE NETWORKS DATASET

Use Case 14 – Choose Network Dataset

Use Case ID	14
Use Case Name	Choose Network Dataset
Description	User chooses a dataset to calculate Social networks from
Actor	User
Pre-conditions	N.A.

Post-conditions	Dataset is successfully loaded

Event Flow

Actor Actions	System Actions
1. Selects a new dataset to load	Saves chosen dataset Updates dataset list Loads dataset Updates representation list for dataset

TABLE 94 - USE CASES - CALCULATE SOCIAL NETWORK

Use Case 15 – Calculate Social Network

Use Case ID	15
Use Case Name	Calculate Social Network
Description	User requests social network to be calculated and drawn
Actor	User
Pre-conditions	Dataset is chosen Dataset was successfully loaded
Post-conditions	N.A

Event Flow

Actor Actions	System Actions
1. Requests Social Network to be calculated	Calculates dataset correlations Generate graphic visualization Shows network

TABLE 95 - USE CASES - FILTER ATTRIBUTES

Use Case 16 – Filter Attributes	
Use Case ID	16
Use Case Name	Filter Attributes
Description	User filters the attributes being considered in the social networks correlation calculation
Actor	User
Pre-conditions	Dataset is chosen Dataset was successfully loaded
Post-conditions	N.A

Event Flow

Actor Actions	System Actions
1. Indicates dataset attributes to filter	Validates the user has selected the minimum number of attributes Filters the attributes selected

5.2. Analysis and Architecture

5.2.1. Overview

As shown in Figure 65 the system's divided into 4 tiers:

- **Client tier**: This component allows the client to use his browser to access the web tier, developed in Asp.net which then fires the requests to Google Charts or Vis.js when necessary.

- **Web Tier**: This Tier will serve as a bridge between the users and the application treating and answering their requests by providing and processing the necessary UI options that allow them to interact with and consult the results given by the underlying tier, as well as customize the available dashboard views for a more specific analysis of the data and graphics extracted from the database and Weka's mining capabilities.
- **Application Tier:** The application's divided into a Data Analysis and predictive modeling tool (Weka), and the Data Access Layer (DAL). DAL is responsible for receiving requests, from both Weka and the Web tier, to provide data for analysis. Every time DAL receives a request, it sends a query to the Database Tier for the SQL Server to handle. Weka is a java tool responsible for the extensive data mining techniques and cross-data comparisons to apply, dynamically searching for the best possible answers to the given scenario, identifying patterns, clusters and correlations when requested.
- **Data Tier**: This tier contains the MySQL Server that holds all of the available data and is responsible for replying to the queries received from the DAL.

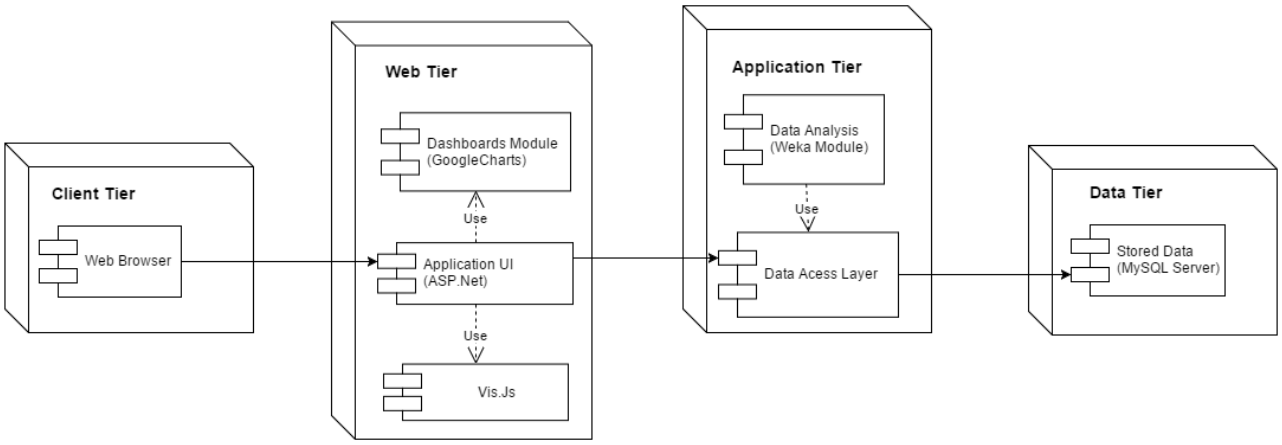

FIGURE 65- SYSTEM OVERVIEW

5.2.2. MAIN FEATURES

The main features provided by this applications are:

- Consult stats over the raw database data by the means of predefined queries provided by the interface.

- Make use of data mining tools to further analysis and draw knowledge from the results.
- Choose from an array of charts and graphs the best suited visualization for the user's own needs.
- Reach conclusions about very different aspects of educational data.

5.2.3. DATABASE

The database has to store a lot of data because there is a need to decentralize the data of all the entities or the external systems, such as ISEP's portal, which at the moment holds all of project's necessary information. For this particular project, data from 2008 onward was gathered.
Educational institutes like ISEP have the need to continuously save and update the data of both their students and staff, in order to be able to manage the community. Each individual student can amount to dozens of records and there are thousands of students enrolled every year, greatly enhancing the amount of available data.

5.2.3.1. DOMAIN MODEL

The domain model shown in Figure 66 provides a better overview of the overall database. Each table is associated with the entity they belong to. For instance, student and class data belong to student data category.

As seen in the Figure 66, there are 3 major entities:

- Student Data
- Course Related Data
- General Data

This project has access to the data ISEP has gathered over the years however, not all of this data should be extracted for the correct implementation of this tool. The multitude of available data led to a need for a selective approach as to what should and shouldn't be considered valuable for the future analyses. For instance, having the student's contacts is important for the university to communicate with them, but it's irrelevant as a means to study their academic history.

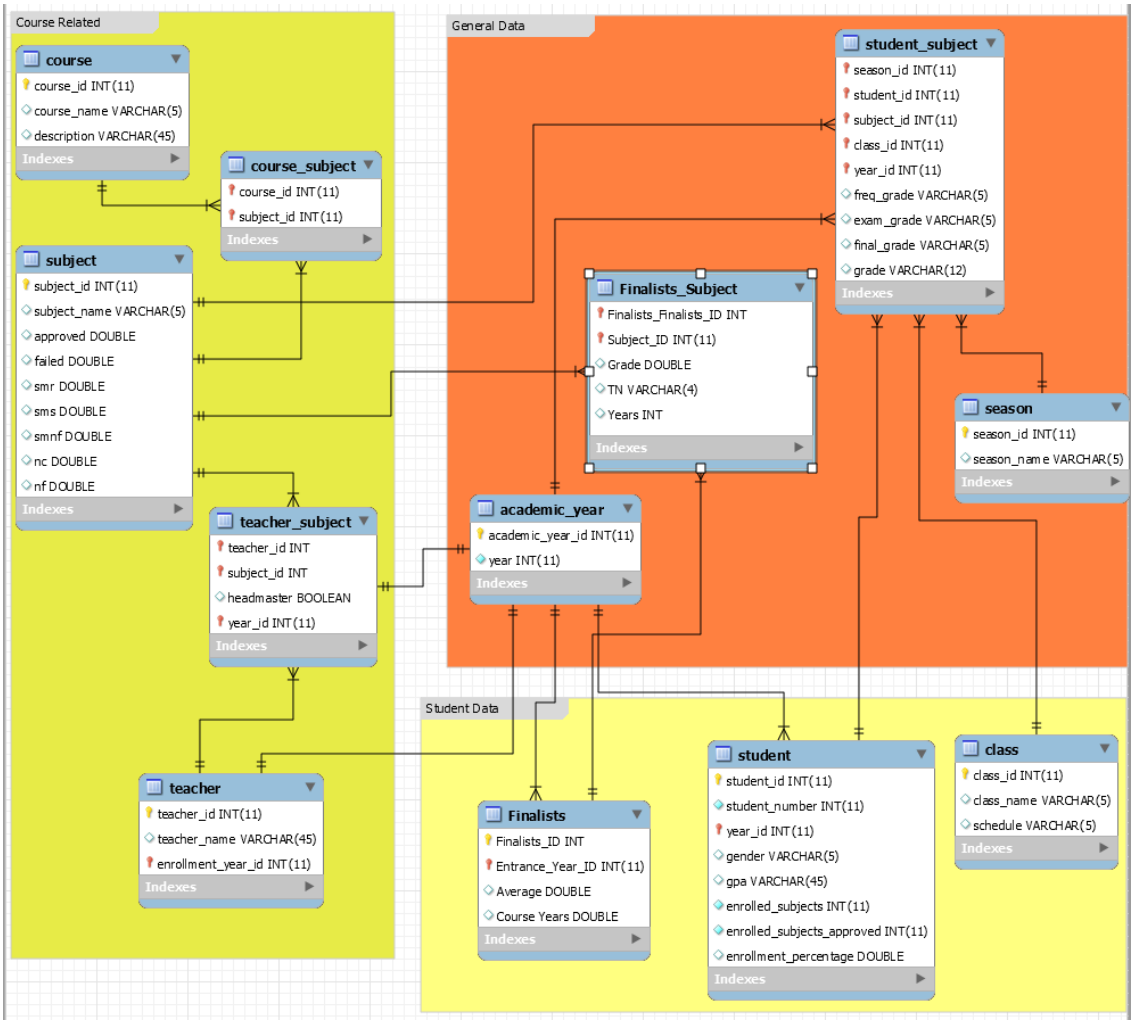

FIGURE 66 - DOMAIN MODEL

The tables and variables of the database for ViDi Stat are presented in the next list.

Academic Year Table:

- Year_ID – Year unique identifier as primary key.
- Year – Correspondent year.

Table academic year has a unique identifier for each academic year in the institution.

Student Table:

- Student_ID – Student's unique identifier as primary key.
- Student_Number – Student's number.

- Enrollment_Year – ID of Year of student enrollment in the institution as primary and foreign for table academic year.
- Gender – Student's Gender
- Average Grade – Average of student frequency and final grade.
- Enrolled_Subjects – Number of subjects the student is enrolled.
- Enrolled_Subjects_Approved – Number of subjects the student has been approved.
- Enrollment_Percentage – Percentage of approval subjects.

Table student as the name says, holds all the personal details that are valuable for ViDi Stat in terms of students variables. The students grades, gender and entrance year are presented in this table in order to provide statistics about his progression.

Student_Subject Table:

- Season_ID – The season unique identifier as primary keys and foreign key for table season.
- Student_ID - The student unique identifier as primary and foreign key for table student.
- Subject_ID - The subject unique identifier as primary and foreign key for table subject.
- Class_ID - The class unique identifier as primary and foreign key for table class.
- Year_ID – The year unique identifier as primary and foreign key for table academic year.
- Freq_Grade - The class obtained average frequency grade.
- Exam_Grade - The class exam obtained grade.
- Final_Grade - The frequency and exam grades combined.
- Grade – The result the student obtained.

This table holds the basis of the database. With the major aspects being accounted for, it is possible to study the student's progression over the years, as well as compare it to others.

Teacher Table:

- Teacher_ID - Teacher's unique identifier as primary key.
- Teacher_Name – Teacher's name.
- Enrollment_Year_ID – Teacher's ID Year of enrollment in the institution as foreign key for table academic_year.

Table teacher holds information about the teacher which identified by a unique ID. It holds the teacher name and ID of the year that started lecturing at the referred institution.

Teacher_Subject Table:

- Subject_ID – Subject's unique identifier as primary and foreign key for table subject.
- Teacher_ID – Teacher's unique identifier as primary and foreign key for table teacher.
- Academic_Year_ID – Academic Year of teacher history in the institutions as primary and foreign key for table academic year.
- Headmaster – Boolean attribute as true or false if teacher is headmaster of the subject id.

By saving the teachers data, we can relate them to the classes they've been assigned to and have a better understanding of their effect on their disciples. It's also important to note that a teacher can also be a headmaster in a determined year.

Subjects Table:

- Subject_ID - The subject unique identifier as primary key.
- Subject_Name - The subjects name.
- Academic_Year – Academic Year ID as primary and foreign key for table academic year.
- Approved – Subject's student's approval percentage.
- Failed – Subject's student's failure percentage.
- SMR – Subject's student's SMR percentage.
- SMS - Subject's student's SMS percentage.
- SMNF - Subject's student's SMNF percentage.
- NC - Subject's student's NC percentage.
- NF - Subject's student's NF percentage.

The subject's progression over the years, tied with the rates it produced, make up for a more detailed view of the learning process. It also makes possible to compare the correlation between subjects that are supposed to go hand to hand and assess how trustworthy those assumptions are, while possible finding new patterns we didn't know existed.

Course Table:

- Course_ID - The course unique identifier as primary key.
- Course_Name – The course name abbreviated.
- Description – The course description and full name.

This table contains a unique identifier for each course of the academic institution.

Course_Subject Table:

- Course_ID - The course unique identifier as primary and foreign key for table course.
- Subject_ID - The subject unique identifier as primary and foreign key for table subject.

This is an intermediate table between courses and subjects with complementary information that will make it possible to separate the different courses being analyzed.

Class Table:

- Class_ID – Class's unique identifier as primary key.
- Class_Name - Class's name.
- Schedule - Class's schedule.

The class table will offer valuable and complementary information about each class for every student.

Season Table:

- Season_ID – Season unique identifier as primary key.
- Season_Name – Season name, can be Normal (NM) or Appeal (RE).

This table will hold an ID per season that is available in the academic institution.

Finalists Table:

- Finalist ID – Finalist unique identifier as primary key.
- Subject ID – Subjects unique identifier as foreign key for table subjects.
- Grade – Final grade to the subject.
- TN – Academic process for which the student completed the subject.
- Entrance Year ID - Year unique identifier as foreign key for table academic year.
- Average – The student average grade.
- Course Years – The number of years that the student took to finish the course.
- Year – Number of years it took to finish the subject.

This table is filled with data concerning students that already ended the course with the required approval value.

5.2.3.2. AVAILABLE INFORMATION

Given its sensitive/confidential nature, not all of the information gathered will be available to the end user. The user will however, be able to observe generic and basic facts such as:

- Approval rates per Subject
- Approval rates per Year
- Average grade per Subject
- Average GPA per Year

As well as more complex points extrapolated from the data, such as:

- Different clusters of students
- Possible grade patterns
- Social networks
- Possible correlation between subjects

By eliminating ambiguity and redundancy where possible and filtering unnecessary, overwhelming information, anonymity can be achieved where needed, while assuring that only pertinent data will be available to the user, making for smoother and clearer consults.
With the needed data structure lined out and implemented, it's possible to access a visual overview of the whole system, generate overall statistics over particular datasets and apply the data mining techniques that can greatly enhance the value over data.

5.2.3.3. VIEWS

Some of the information stored in the database, such as confidential or sensible data, should not be shown to the user or used to generate conclusions. There was also a need to join or aggregate multiple records so that the application retrieves them in a format that can be used with the implemented features. In order to solve these problems, the records made available for the users do not correspond to the original tables, but rather subsets of them in the form of views.
Three type of views were implemented:

- **General Views –** Used to generate the charts in the general statistics page, these views aggregates, averages and counts it's records, providing an overall notion of what exists without getting into too much details.

- **Mining Views –** Used to provide the records used by the mining techniques, these views provide individual records while concealing attributes that are sensible and / or wouldn't bring reliable information to the results, such as the student number.
- **Network Views –** Only contains attributes that can generate correlations between different nodes, ignoring all other parameters that wouldn't make sense for this implementation.

These views were only created where the target data required it, with the general views having the 'view' word next to the table's name, the mining views with the word 'miningView' and the network views with the word 'networkView'. Example: Subject**View**, Subject**MiningView**, Subject**NetworkView**.

5.2.4. Association Rules

The option to use generate association rules in the application is provided by the imported Weka libraries.

After treating the request sent by the user, these association rules are presented via word trees, with a single antecedent being shown at a time, with the consequents that are related to it as seen in Figure 67.

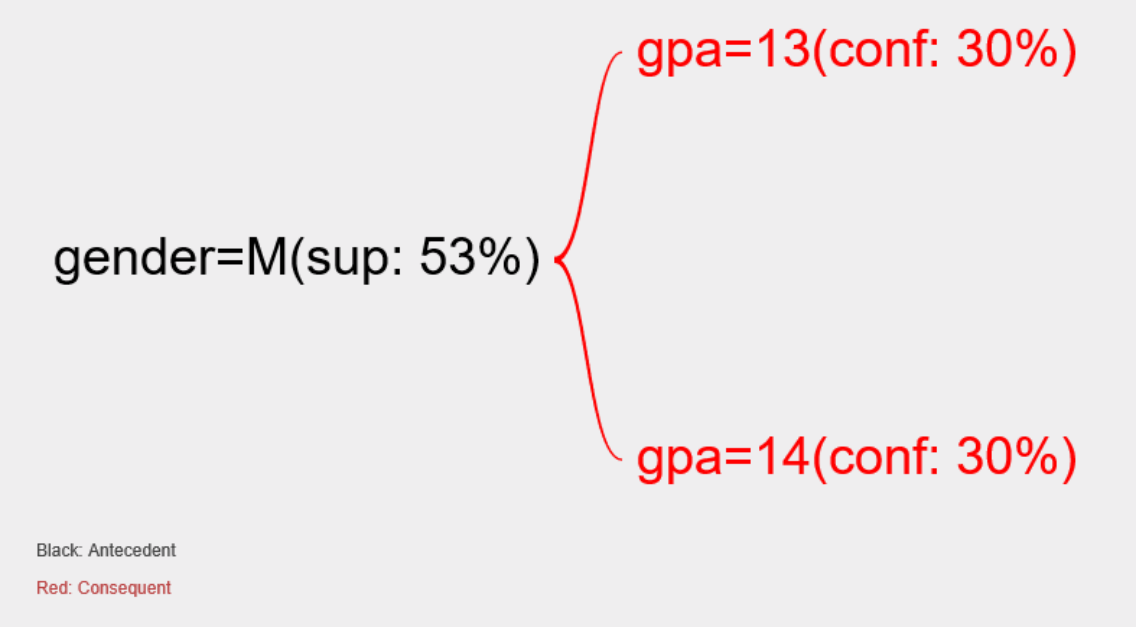

FIGURE 67 - WORD TREE

5.2.5. CLUSTERING

The option to use generate clusters in the application is provided by the imported Weka libraries.

After treating the request sent by the user, these clusters are presented via tree maps, with each cluster being represented as a rectangle as show in Figure 68, a table with structured text describing the cluster's properties as seen in Figure 70 and Figure 69, or by a social network representing cluster's attributes as nodes and connecting the properties of the same cluster via edges as seen in Figure 71 and Figure 72.

FIGURE 68 - TREE MAP

The tree map represents each cluster with a square and a color, the gradient above indicates the colors used, from less relevant to most relevant. This is also observable by noticing the sizes of the rectangles that hold the cluster information: the more instances it holds, the bigger it is. The user is able to click any of the clusters in order to expand it and observing the attributes it holds.

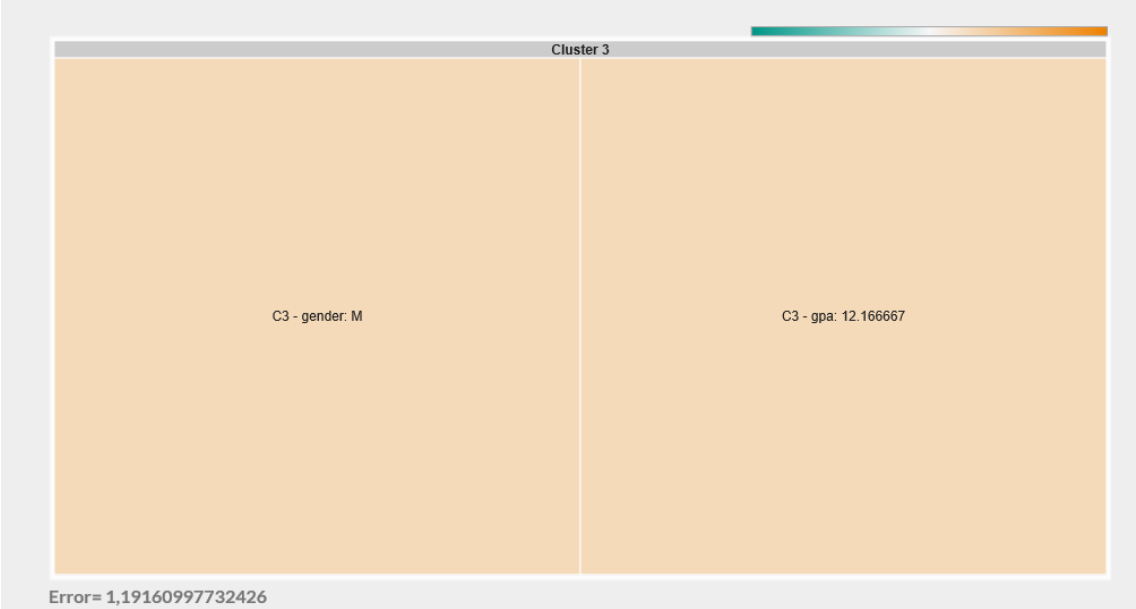

FIGURE 69 - TREE MAP EXTENDED

Figure 69 shows the attributes held by cluster number 3: 'M' gender and '12.166667' GPA, meaning that the instances contained in it fit into those values.

When represented as a table, the clusters are described by their cluster ID in the first column, the attributes into the following and the number of instances associated with the given cluster in the last column.

Cluster ID	gender	gpa	Number Instances
Cluster 0	F	12.777778	9
Cluster 1	M	16	1
Cluster 2	M	14	3
Cluster 3	M	12.166667	6

Error= 1,19160997732426

FIGURE 70 – GOOGLECHARTS TABLE

The social networks representation of clusters starts by joining all instances of every cluster into a single node which is named after it's ID, with the 'First Cluster' found being cluster 0, the second being 'Cluster 1' and so forth.

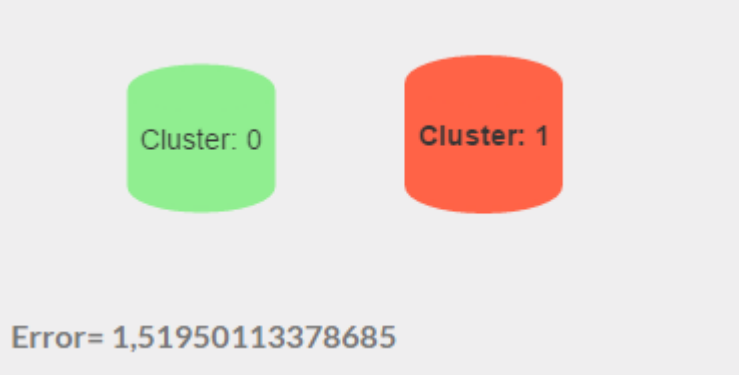

FIGURE 71 - CLUSTERS NETWORK

When the user clicks the nodes the represent the clusters, they are decomposed and their instances shown, with a connection being drawn between the attributes of the same clusters. The cluster's color is also unique and as seen in Figure 72

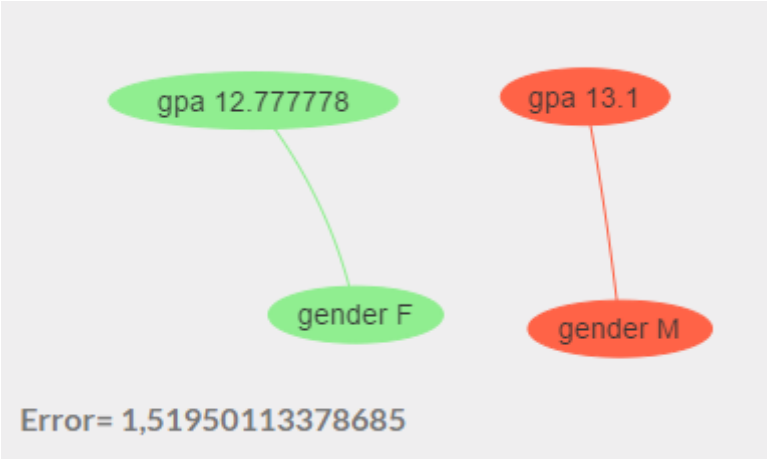

FIGURE 72 - CLUSTERIZED INSTANCES NETWORK

Below each of the graphs described, an error label is displayed, indicating the squared sum of errors within the represented clusters.

5.2.6. Social Networks

According to the user settings in the webpage, academic data stored in the database is retrieved and presented in a fully interactive Social network graphic in order to provide conclusions.

A Social network example can be seen in Figure 73, which shows a basic sample from the application's Social network model where subjects are represented as nodes and the correlation between them as edges.

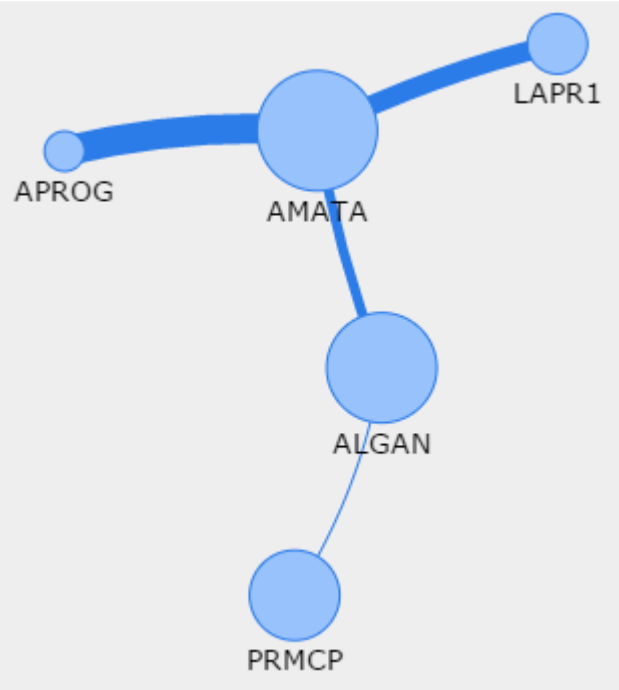

FIGURE 73 - SOCIAL NETWORK EXAMPLE

The Social network architecture is simple and easily extensible. Nodes are created in runtime as edges, thus providing a fully scalable network to the user.

5.2.7. Web Application

The web application was developed in Asp.net has 4 major components:

1. **Weka –** Weka was integrated in this platform by the means of Ikvm, which is a java implementation for the Microsoft .net framework, thus providing java and .net interoperability: Java code can be compiled and executed directly on Microsoft.net.
   This component is responsible for all of the data mining operations: It collects the data made available by the DAL and, when prompted by the user, mines it in a timely manner and returns the warranted results. Given the amount of possibilities the user has, Weka is ready to preprocess the dataset so that it fits the criteria of a wide range of available algorithms of association, classification, clustering and decision making.

2. **DAL/BLL –** DAL/BALL are the Data Access Layer and Business Access Layer, granting database access and its separation from the business logic:

   DAL provides services which communicate directly with the database, being responsible for all interactions with it – no outside method is to be allowed to communicate with the database. This layer has objects mapped to the appropriate database entities.

   BLL is where the business logic services are included, successfully isolating business logic into classes that can be reused by the entire application, or recycled/modified at a further point without compromising its underlying layer, DAL If a request is to be made, BLL handles it accordingly, conveys it to DAL and then proceeds to map the results to their appropriate business entities.

3. **Google Charts –** Google Charts is a platform developed by Google which offers many dashboard features. This component is responsible for the presentation of the data in friendly, easy-to-read and interactive charts.

   These charts are interactive, allowing the users to navigate through the data when possible, while using a mix of symbols and colors to better grasp the user's attention, making for a seamless graphical overview of the requested information. Although most visual representations have simplicity as a core, these charts can become increasingly complex if necessary implementing additional event handling on the chart, making future adjustments feasible.

   It is also possible to dynamically change the data source or the type of displayed chart by the means of dropdown lists.

   Since the charts are rendered using HTML5/SVG, the charts offer cross-browser and cross portability compatibility to iPhones, iPads and Android so that the only real requirement for this technology is for the user's platform to have a web browser.

4. **Vis.Js -** offers a set of operations that the user can use and interact with, developing the visual representations that the application manipulates as it sees fit manipulated. Only two of the available components were used: a network composed by nodes and edges, which ultimately represents this project's Social networks and a flexible key/value based dataset with add, update, and remove operations to manipulate and build the data as required.

5.2.7.1. USER INTERFACE

This is a crucial point of the project since it'll have to be able to present complex and detailed data in a visual, easy to read, environment. This alone requires the platform's interface to be as intuitive, easy to learn and use, as possible, using graphics and charts to portray meaningful data in a friendly way.

The interface will offer the user to use a set of options, not only to go through the raw and analyzed data's dashboards, but to query and mine the dataset for themselves in order to obtain facts and knowledge that may prove interesting or pertinent.

The 5 key principles to apply to the interface were:

1) **Structure –** The UI should be organized in meaningful and useful matter, based on clear and consistent models the user can recognize, joining related things and differentiating dissimilar ones.
2) **Simplicity –** The Interface should make tasks easy, communicate clearly and simply in the user's own language, and provide adequate shortcuts where necessary.
3) **Visibility –** Only the needed options and materials should be available to the user at a given task to avoid any possible confusion or overwhelming information.
4) **Feedback –**The feedback given should keep the users informed of the outcome of their actions. Eventual errors or exceptions, of interest to the user, should be presented in a clear and unambiguous language adequate to the user.
5) **Tolerance –** The system has to be tolerant and flexible to the user's possible mistakes and misuse, preventing errors wherever possible.

With this principles in mind, the UI was developed using Microsoft .net visual capabilities, as well as an intrinsic implementation of Google Charts.

The webpages generated by the .net platform act as an intermediary between the user's needs, the underlying data contained by the application, and the Google Charts component which handles the visual illustrations. When a user loads a page, he is confronted with multiple options concerning his needs, which fire an appropriate answer when triggered:

- If the user wants to verify the data made available by the platform, he can do so by consulting a wide range of charts provided by the Google Charts component and the Vis library. Both the chart types and the selected data can be dynamically changed.
- If the user wants to get additional answers regarding a specific topic, he will provided with the data mining tools to do so. These tools can be dynamically parameterized by the users to best fit their needs, the .net platform handles the request and ultimately illustrates the results in a more intuitive manner.

All of the parameterization and request handling is adapted so that errors and user misuse are treated accordingly and instant feedback is provided concerning what exactly went wrong, thus acting as an instructive tool while making the application more consistent, robust and tolerant.

5.3. Application Design

5.3.1. Class Diagram

The class diagram for ViDi Stat is presented in Figure 108.

5.3.2. Sequence Diagrams

5.3.2.1. Data Mining

The Data Mining webpage starts by offering the user options regarding the dataset to be analyzed. With the dataset chosen, it's possible to choose the procedure to perform, whether it is association rule mining, or clustering.

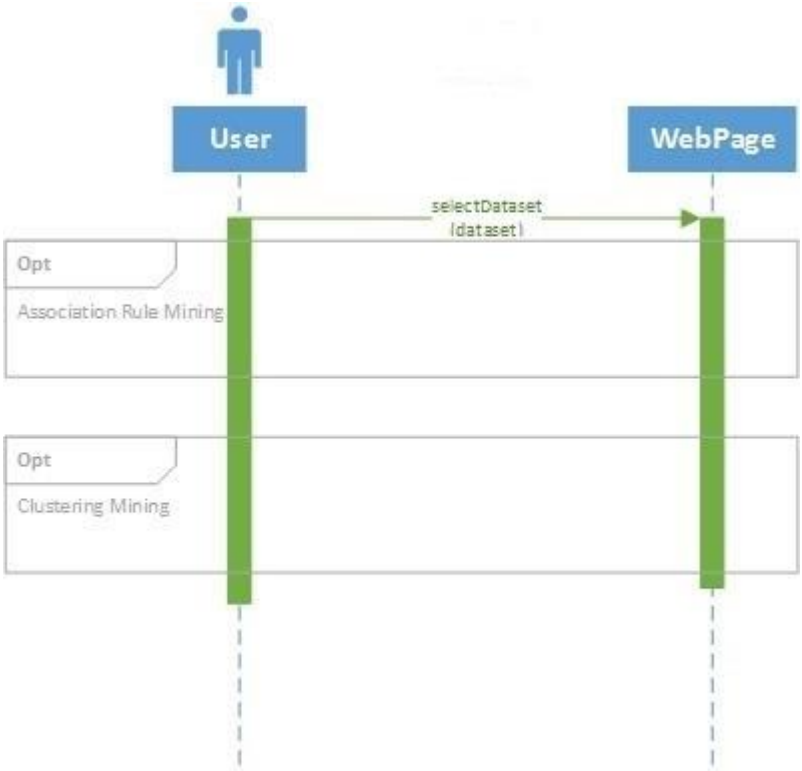

FIGURE 74 - MINING OPTIONS

5.3.2.1.1. ASSOCIATION RULES

The only association rules algorithm implemented was the Apriori one. After activating this method, the user needs to indicate the values intended for the key parameters:

- **Support –** Minimum rule support
- **Confidence –** Minimum rule confidence
- **Delta –** Factor through which the support is iteratively decreased until minimum support is reached or the maximum number of rules has been generated.
- **Number of Rules –** Maximum number of rules to get.

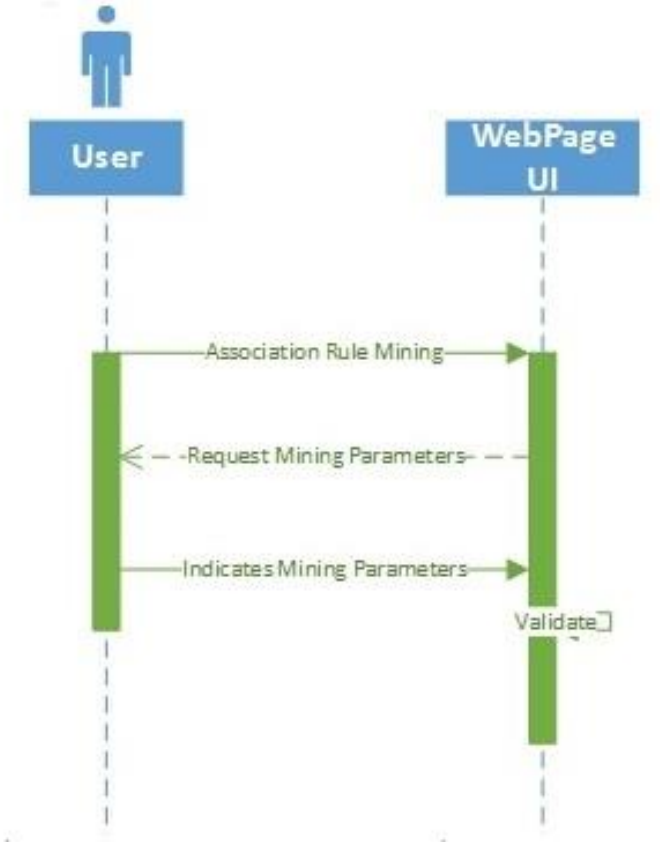

FIGURE 75 - INSERTING RULES PARAMETERS

As detailed in Figure 76, after collecting and validating the parameters, the web tier requests the Weka library to perform the association rule mining.

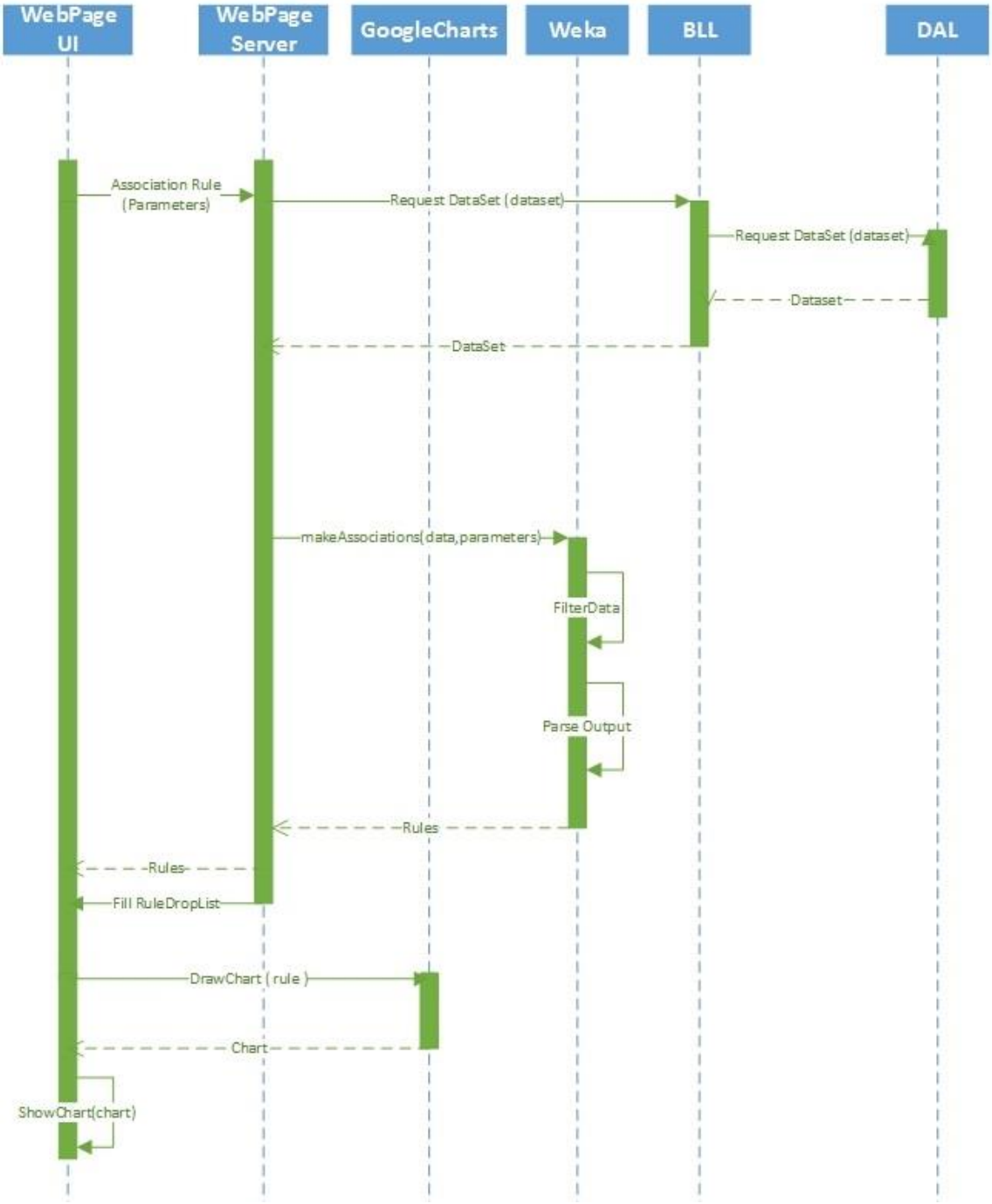

FIGURE 76- DRAWING ASSOCIATION RULE CHART

The chart is first drawn with the first rule found. The user is then able to iterate through the existing rule antecedents and choose a new rule to be represented as seen in Figure 77. A new request is sent to the google charts platform and a new representation is returned.

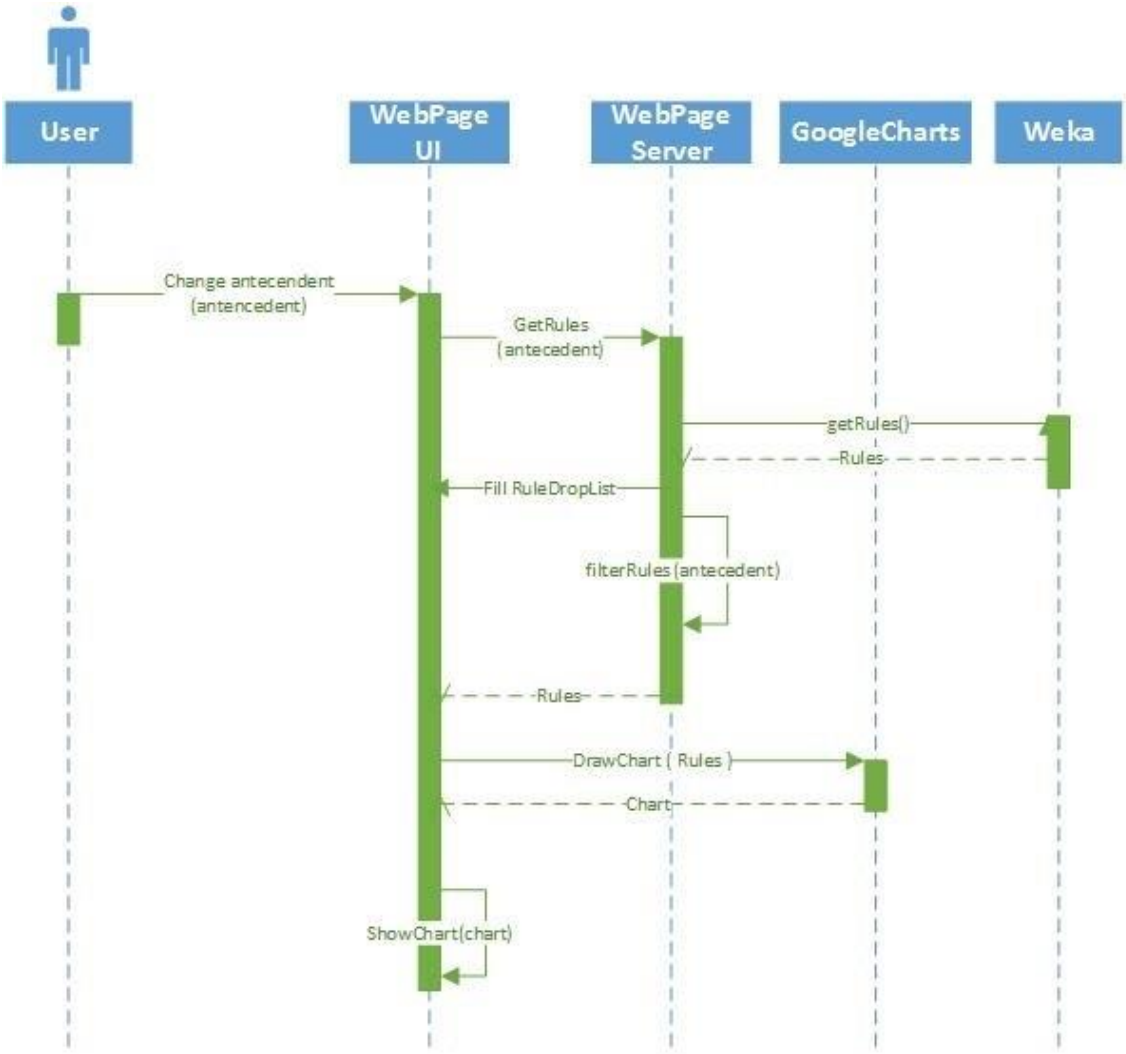

FIGURE 77 - CHANGING RULES

The user is free to change the rules observed or to request a mining product of another dataset / other parameters at any time.

5.3.2.1.2. CLUSTERS

The only clustering algorithm implemented is the Simple K-Means one. After activating this method, the user needs to indicate the values intended for the key parameters:

- **Number of Clusters –** Number of clusters intended.
- **Number of Seeds –** Number of random starting points to build the clusters from.

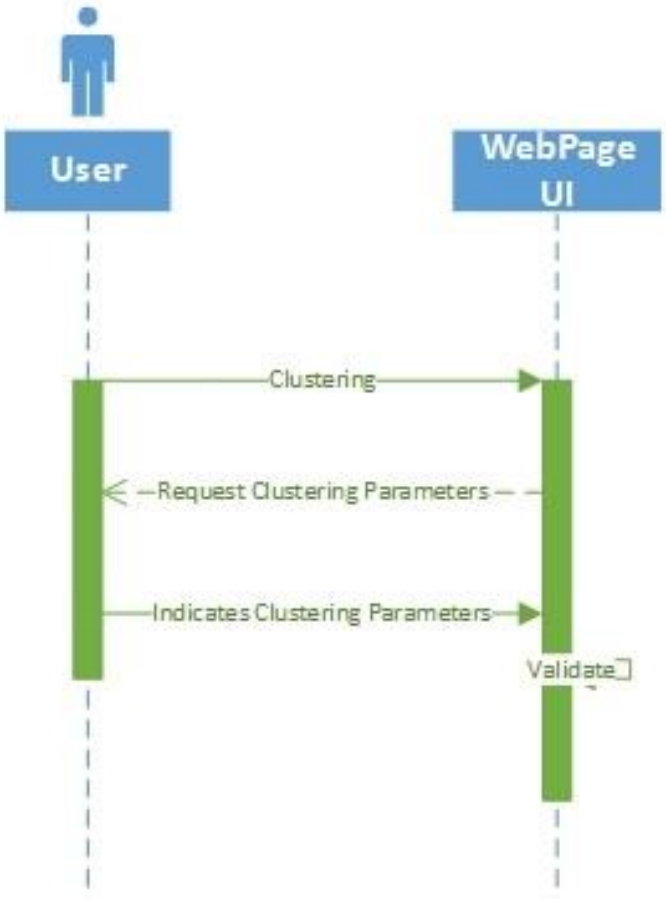

FIGURE 78 - INSERTING CLUSTERING PARAMETERS

After collecting and validating the parameters, the web tier requests the Weka library to perform the clustering as detailed in Figure 79. These parameters can be changed at any time.

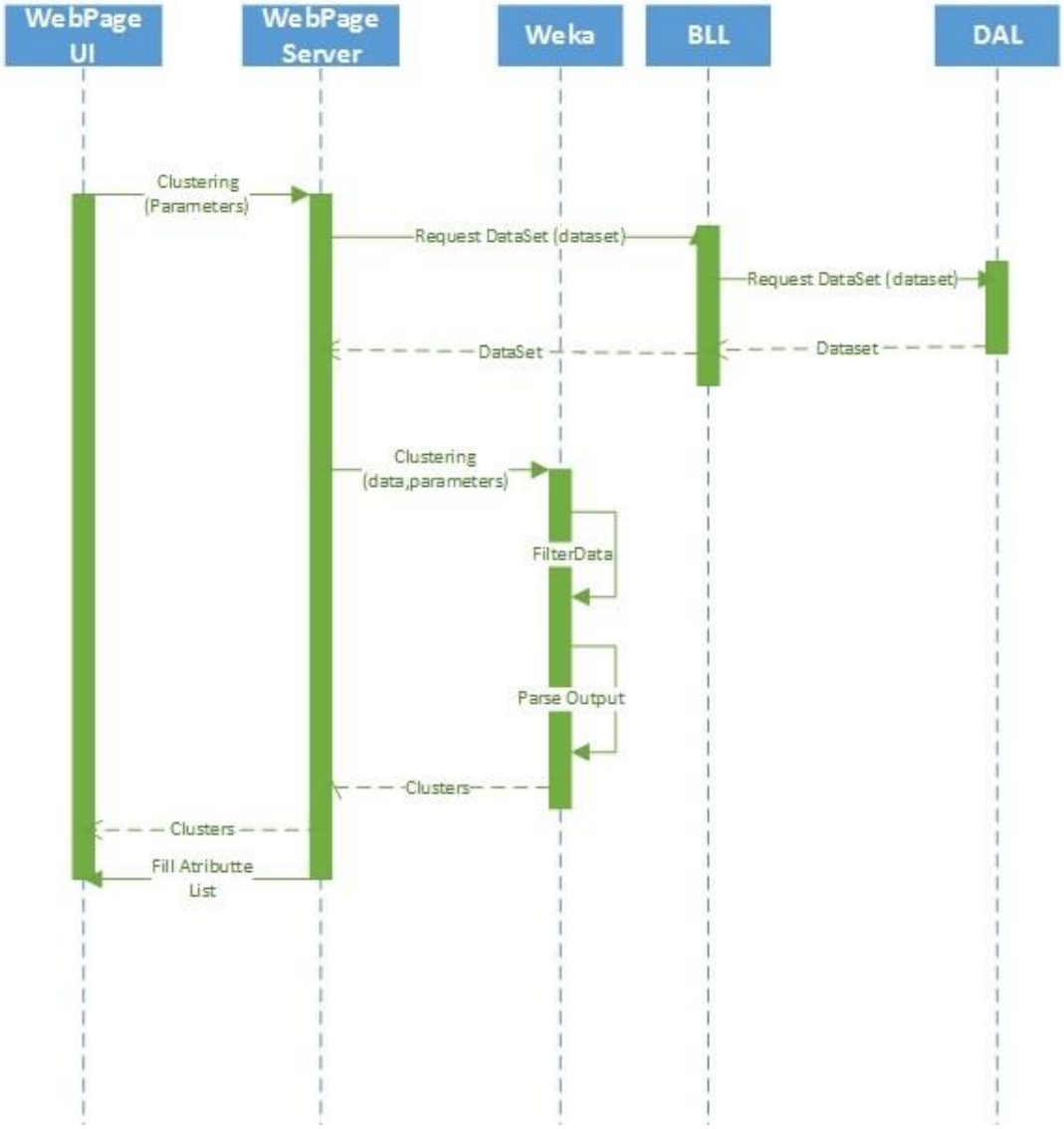

FIGURE 79 - PERFORMING CLUSTERING

The initial clusters will consider all of the dataset attributes, but the UI offers the possibility to filter these attributes in order to get new instances that only consider the filtered attributes.

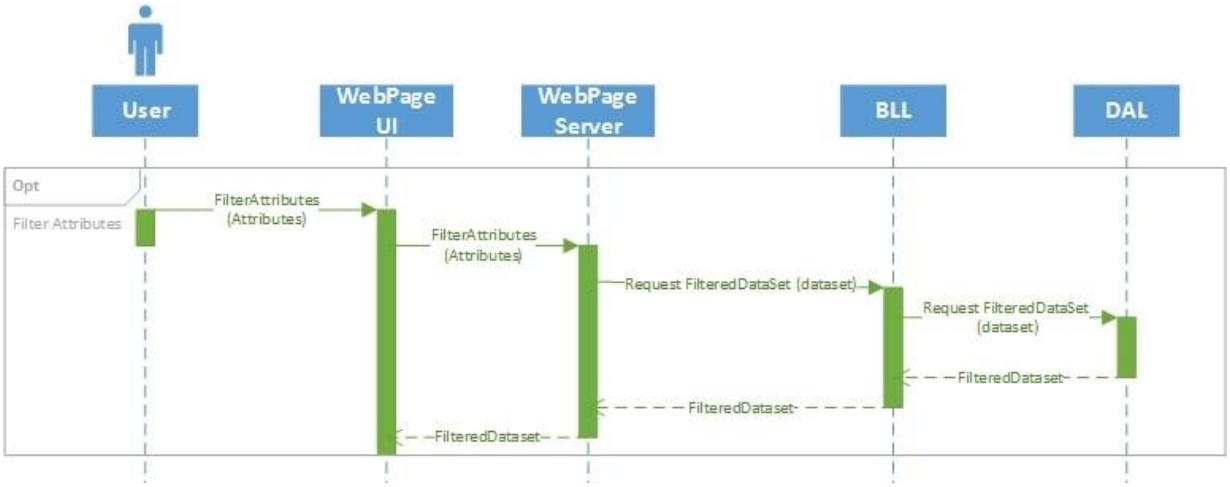

FIGURE 80 - FILTERING ATTRIBUTES

The user is prompted with the default representation after defining both the parameters and dataset and is offered the possibility to later the representation of the clustered instances, with the UI performing requests to different platforms accordingly as seen in Figure 81 and Figure 82

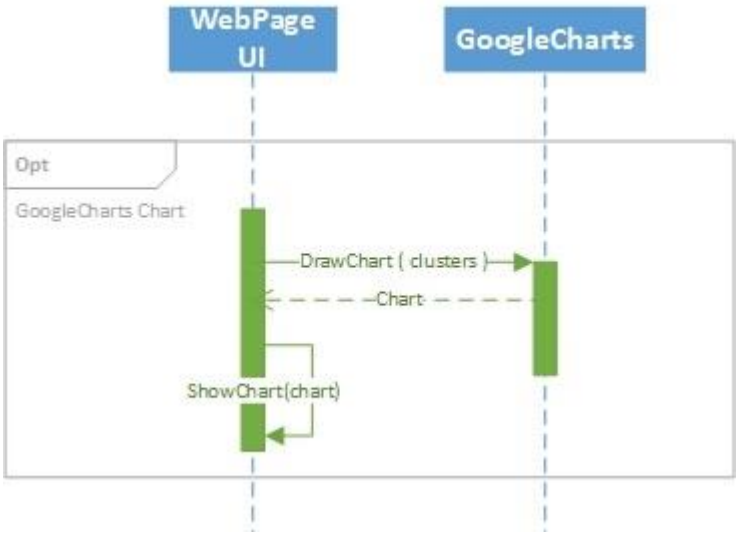

FIGURE 81 - GOOGLE REPRESENTATION

If a representation that uses GoogleCharts is chosen, the request is made to the GoogleCharts platform, and the returned chart is presented as seen in Figure 82.

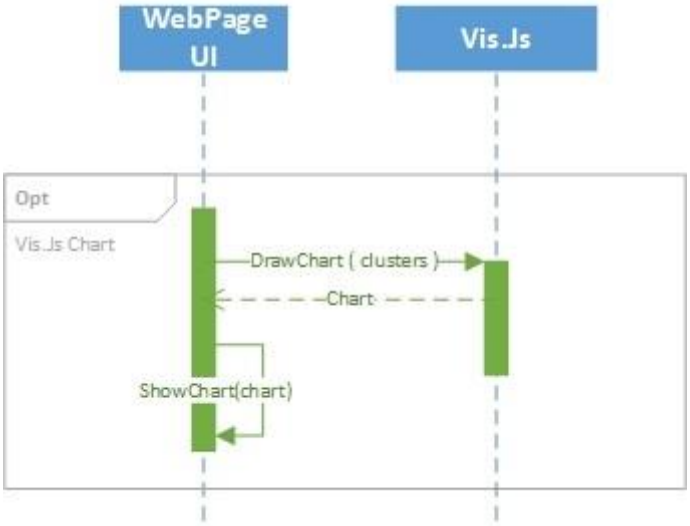

FIGURE 82 - VIS.JS REPRESENTATION

If the representation method selected requires Vis.JS to be developed, the request will be made to the Vis.Js component and the returned chart presented.

5.3.2.2. GOOGLE CHARTS

In the General Statistics page, the GoogleCharts platform is responsible for elaborating all forms of representation provided. When accessing the page, the user is able to choose the dataset to be represented. This process is detailed in Figure 83.

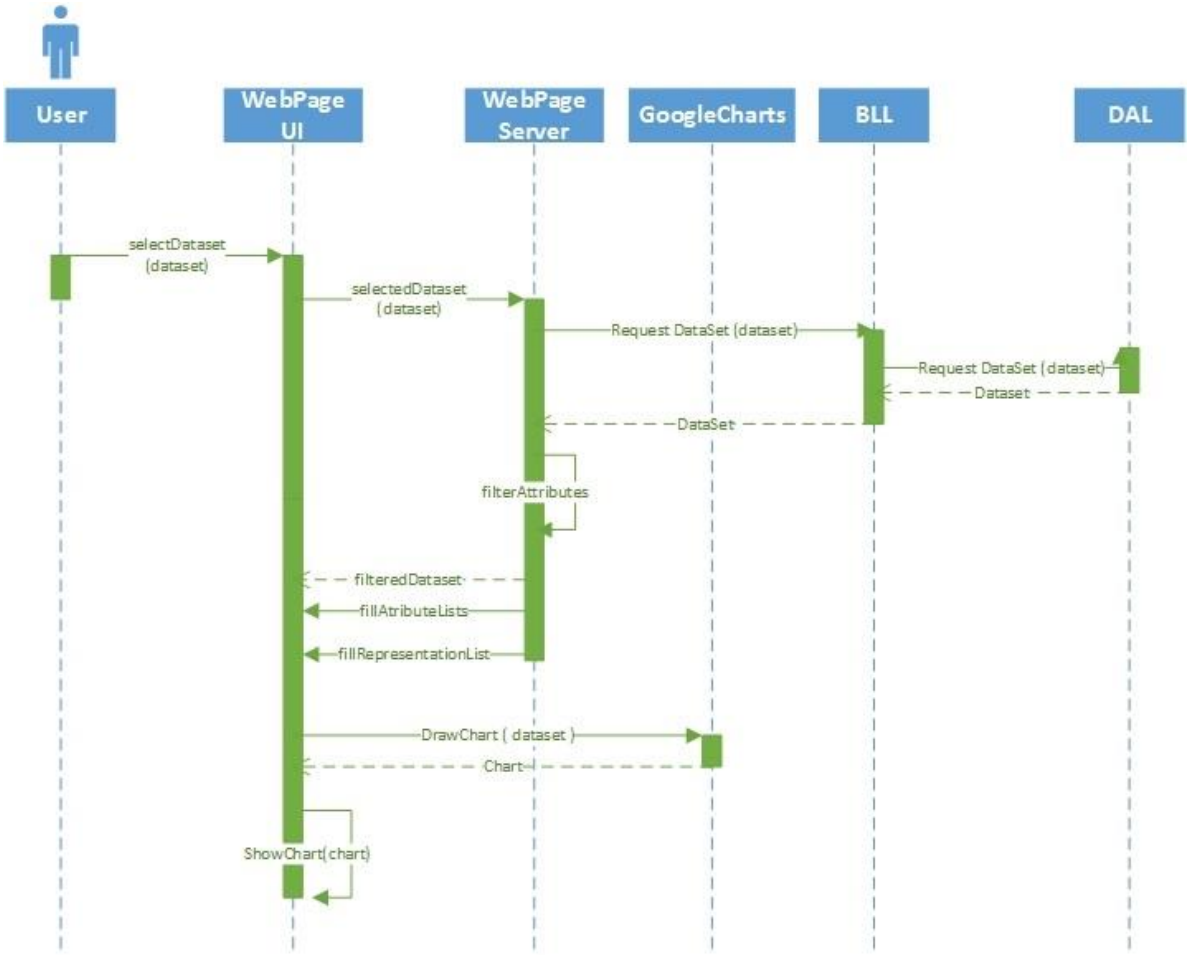

FIGURE 83 – GENERAL STATISTICS SELECTING DATASET

With the dataset selected, the designed charts take into consideration two attributes at a time. These attributes can be dynamically selected by the user as seen in Figure 84.

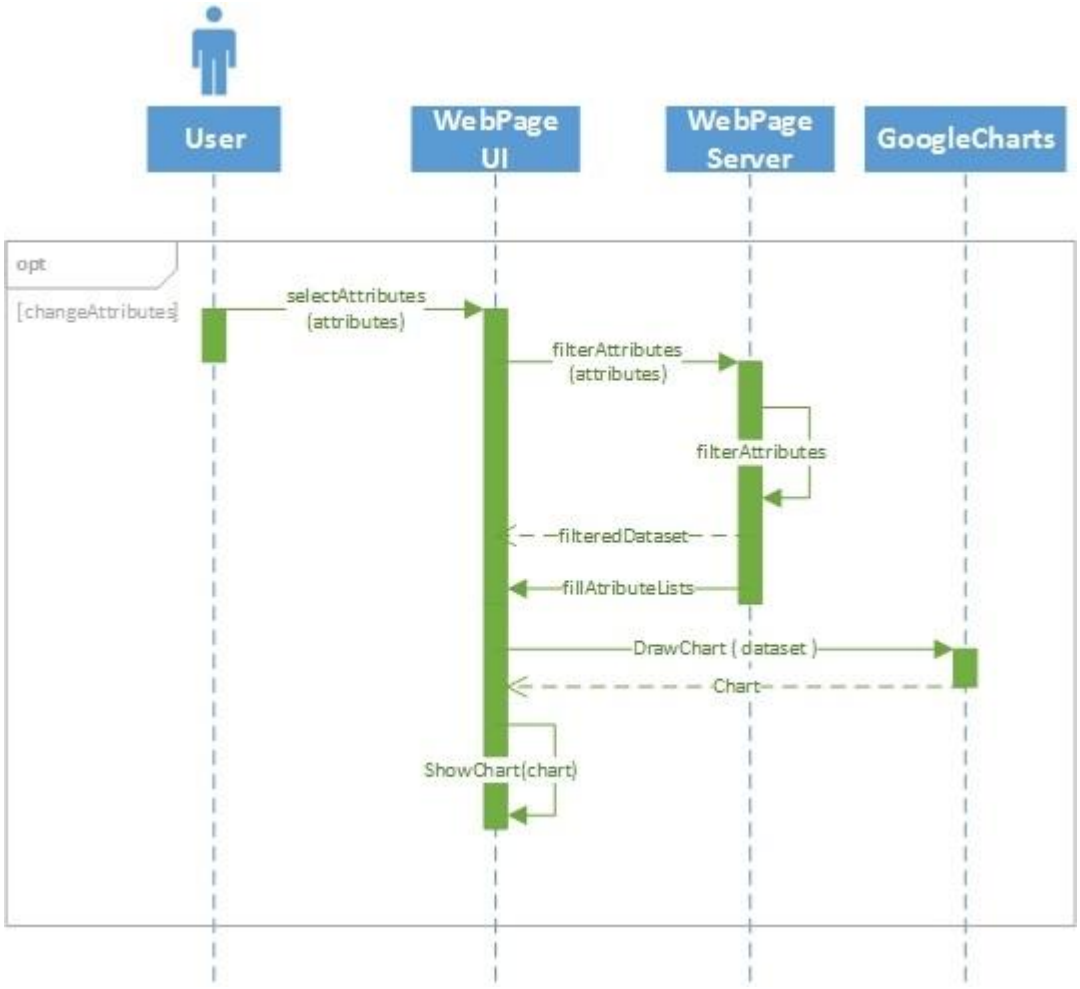

FIGURE 84 - CHANGING CHART ATTRIBUTES

After selecting a new set of attributes, the server filters out the necessary data and the UI is able to request an updated chart which is then shown to the user. This process is very similar to that of changing the attributes displayed in the cart as displayed in Figure 85

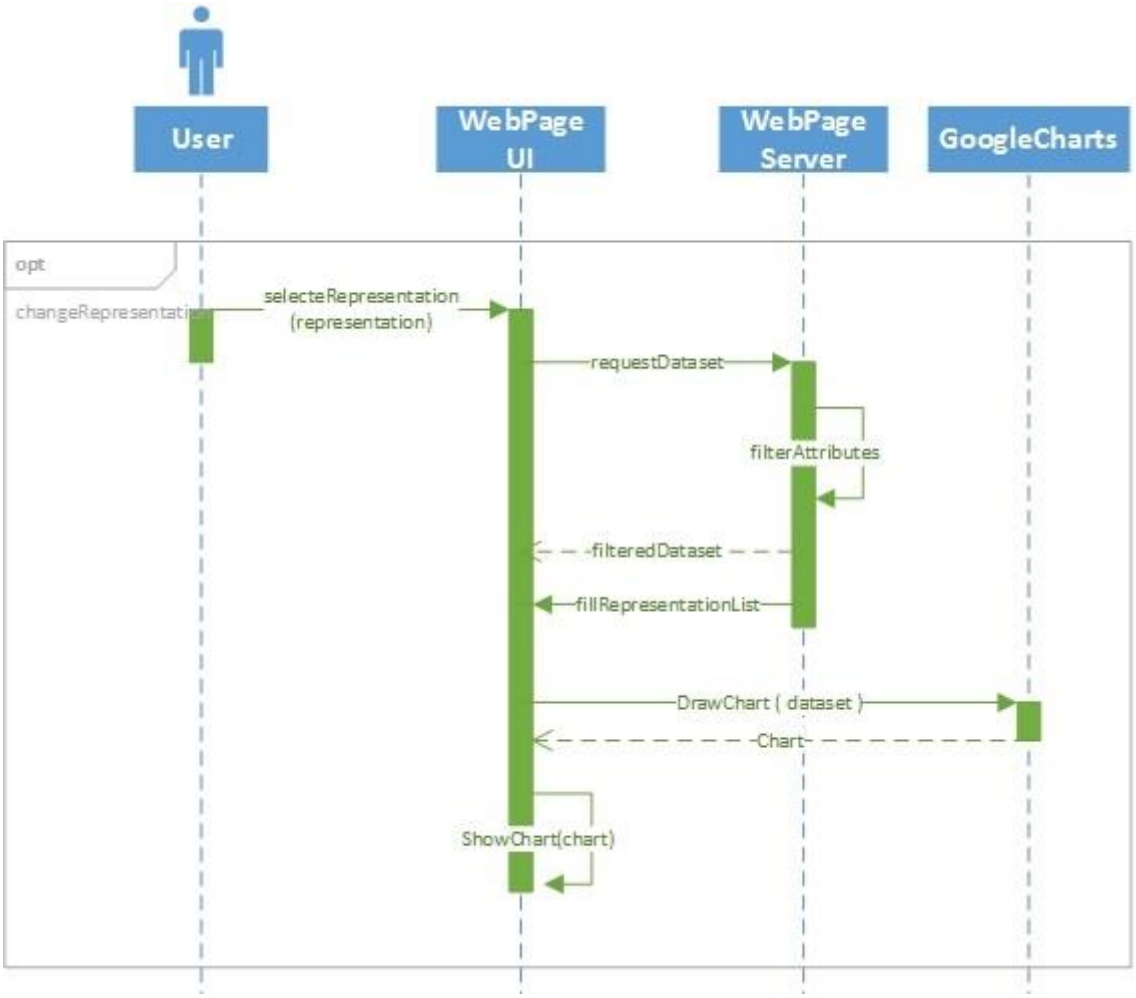

FIGURE 85 - CHANGING REPRESENTATIONS

After selecting a new type of representation, the server processes the data accordingly, updates the list of available representations and delivers the data necessary to request a new representation, which is then shown to the user.

As shown in Figure 86 when the data can be segmented by years, the server will take that into account when filtering the attributes so that the data is segmented and the UI can apply a level of interactivity to the resulting chart, allowing the user to dynamically filter the dataset.

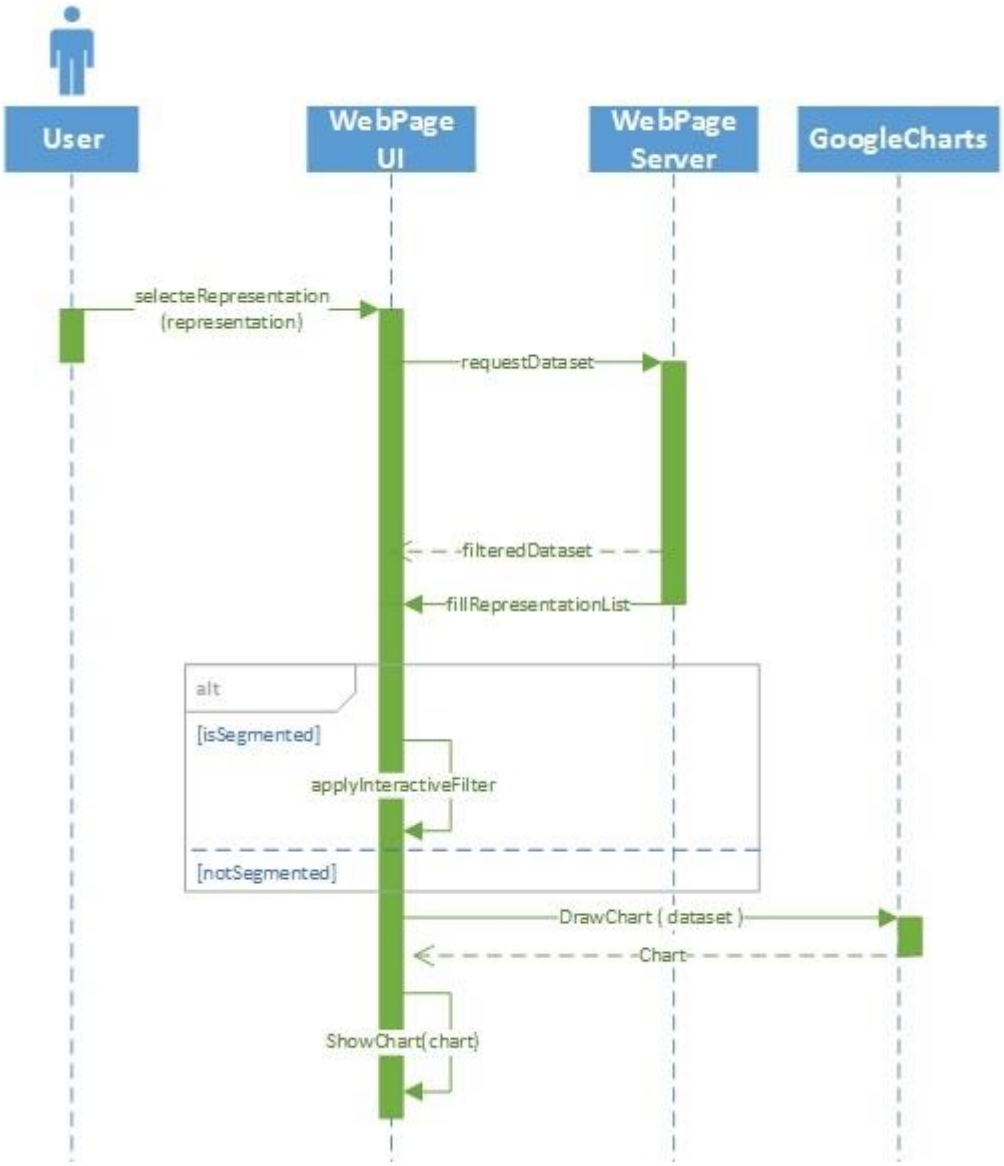

FIGURE 86 - SEGMENTED DATA CHART

The interactive filter consists of a horizontal bar holding the absolute minimum and maximum year values retrieved from the dataset its been given, allowing the user to then change these parameters at will and check the results in real time as the chart dynamically modifies itself to fit the new requirements. This bar is similar to the one in Figure 87.

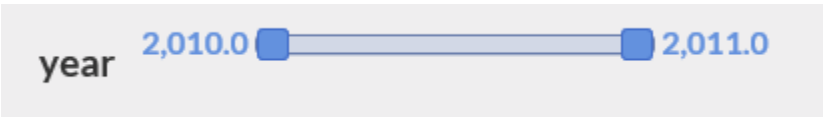

FIGURE 87 – CHART'S INTERACTIVE FILTER

5.3.2.3. SOCIAL NETWORKS

When the users access the social networks page, it is possible to choose a dataset and have the social network of the entire dataset drawn or filter the attributes of a dataset and have the social network only consider the attributes chosen.

As shown in Figure 123, after selecting a dataset, the user is able to filter the attributes being considered, with the application ignoring those attributes before request the correlations calculation.

As shown in Figure 124, when the user requests the social networks of a dataset to be drawn, it is retrieved from the DAL and passed over to Weka, which proceeds to process the necessary calculations to determine the correlations, parses the output and returns the raw structure of a network which is given to the Vis.js libraries that return a visual representation of the network.

5.3.2.4. HTLM2CANVAS

When observing a chart, a download button is visible under it, and by pressing it, the user is prompted to save the chart as a JPG image. This workflow is shown in Figure 125.

5.3.2.5. DATABASE

The database requests are always made through the BLL layer that transmits the appropriate request to the DAL, which then accesses the database and returns the required dataset.

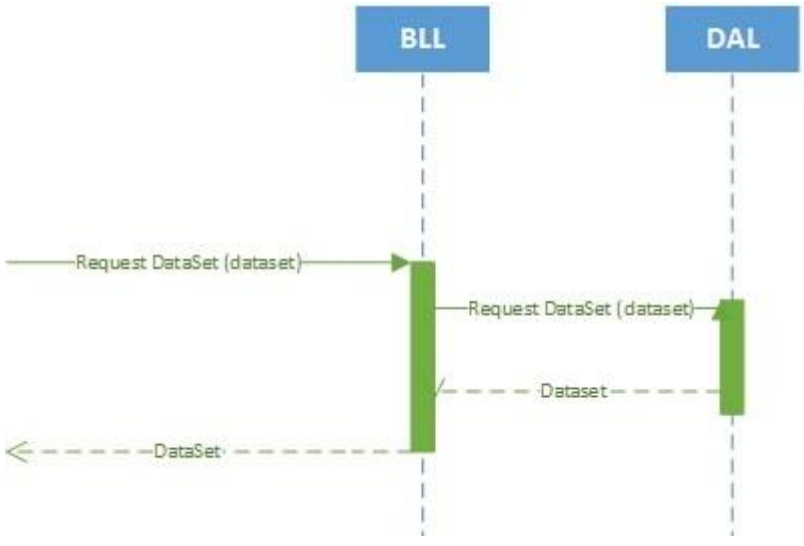

FIGURE 88 - DATABASE ACCESS

The requests to the BLL can be made by the general stat, data mining or social networks pages.

5.4. IMPLEMENTATION

5.4.1. TECHNOLOGIES USED

5.4.1.1. WEKA

A Weka library was generated for the Microsoft .Net Framework via IKVM (Anon., s.d.), a tool that enables java and .Net interoperability.
With the Weka library integrated in the project, the web tier is responsible for handling the data mining requests made by the user and transmit them to this module, which performs the required operations and delivers the results.

5.4.1.2. ASSOCIATION RULES

The association rules are extracted from the data using the functionalities offered by Weka, however, the developed output has to be treated so that it can be used as input data for the forthcoming charts.

TABLE 96 - APRIORI RAW OUTPUT

```
Apriori
=======

Minimum support: 0.1 (2 instances)
Minimum metric <confidence>: 0.1
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 6

Best rules found:
```

1. gpa=12 2 ==> gender=M 2 <conf:(1)> lift:(1.9) lev:(0.05) conv:(0.95)

Table 96 shows an example raw output of the Apriori algorithm. The important information to be retrieved from this output and shown to the user is:

- Antecedent
- Consequent
- Support
- Confidence

The first step taken into the treatment of this output is to trim everything above the "Best Rules Found:" sentence. The rest of the output represents the rules obtained from the dataset; these are then split into words:

1) The first word is the number of the rule, irrelevant for the intended purposes and is therefore discarded.
2) The second and forthcoming words, until the characters **'==>'** appear, are the rule's antecedent.
3) The number after the rule's antecedent and before the **'==>'** indicates how many times this particular set of antecedents was found in the data. This information is used to later calculate the rule support.
4) The **'==>'** characters separates the rule's antecedents from its consequents.
5) The following words, until the characters **'<conf:('** appear, are the rule's consequents
6) The number after the rule's consequent's and before the **'conf:('** characters is the number of times this rule's consequent was found.
7) The **'<conf:('** characters indicate that the following number is the rule's confidence, and its value is stored. This number varies between 0 and 1, 0 being 0 % and 1 being 100%.
8) The **'lift:('** characters indicate that the following number is the rule's lift. This information is not relevant to the user and is therefore discarded.
9) The **'lev:('** characters indicate that the following number is the rule's leverage. This information is not relevant to the user and is therefore discarded.
10) The **'conv:('** characters indicate that the following number is the rule's conviction. This information is not relevant to the user and is therefore discarded.

For example, for the rule:

1. gpa=12 2 ==> gender=M 2 <conf:(1)> lift:(1.9) lev:(0.05) conv:(0.95)

- **'1.'** Is the rule number, indicating that this is the first rule found.
- **'Gpa=12'** is the rule's antecedent fraction
- **'2'** Is the number of times the itemset **'Gpa=12'** was found as an antecedent in the given dataset.
- **'gender=M'** represents the rule's consequent
- **'2'** indicates how many times **'gender=M'** occurred as a consequent of **'GPA=12'**
- **'conf:(1)'** indicates this rule's confidence is 1 (100%).
- **'lift:(1.9)** indicates this rule's lift is 1.9.
- **'lev:(0.05)** indicates this's rule's leverage is 0.05
- **'conv:(0.95)'** indicates this rule's conviction is 0.95

Each rule's antecedent is given a unique id to be identified with. When a rule has a composed antecedent (more than one itemset), the algorithm checks if there's already a rule with the same partial antecedent and, if so, a parental ID is given to the different set of the antecedent, relating it to the pre-existing one. The consequents are also related to their respective antecedents by the means of the parental ID.

Example:

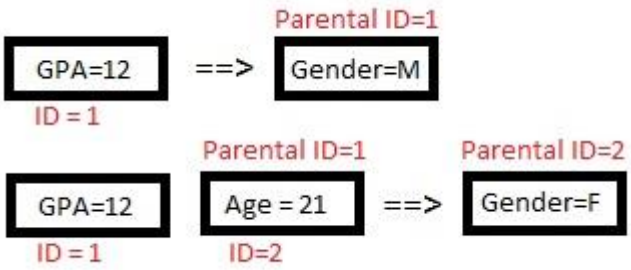

FIGURE 89 - RULE'S IDS

After treating the output, the rules are stored as objects of a 'Rule' class and can be easily converted into the formats required by the UI.

5.4.1.3. CLUSTERING

The clusters are extracted from the data using the functionalities offered by Weka, however, the developed output has to be treated so that it can be used as input data for the forthcoming charts.

TABLE 97 - K-MEANS RAW OUTPUT

```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 1.5195011337868483

Initial starting points (random):

Cluster 0: F,16
Cluster 1: M,16

Missing values globally replaced with mean/mode

Final cluster centroids:
                              Cluster#
Attribute    Full Data        0        1
                (19.0)       (9.0)    (10.0)
=============================================
gender             M        F        M
gpa          12.9474    12.7778     13.1
```

Table 97 shows an example raw output of the Simple K Means algorithm. The important information to be retrieved from this output and shown to the user is:

- Sum of squared errors
- Number of instances per cluster
- Cluster Centroids

The output will always contain the amount of clusters requested by the user. The information gathered from the output is stored in a 'Cluster' class.

This output is treated with the help of the Weka component:

1) The first step is to retrieve the number that follows the "Within cluster sum of squared errors:" line. This number represents the sum of squared errors and is important to the user.

2) The second step is to retrieve the size of each cluster. This number is shown below the cluster number.
3) Third step is to trim everything above the attributes definition. With this step completed, every line represents a new cluster attribute, with the first value after it representing the most common value for that attribute in the whole dataset, and the second onwards representing the value that attribute held in the following clusters.

5.4.1.4. GOOGLECHARTS

Google Charts is a tool developed by Google. This tool was chosen as a means to provide visual representations because it is powerful, simple and free to use.

The GoogleCharts platform was integrated into a C# project by the means of imported libraries and Javascript code. This platform requires the end-user machine to be connected to the internet, but that wasn't considered a limitation because it is assumed that such a need will always be fulfilled, especially since the database is remote and has the same prerequisite.

The google library loads the required packages at the start of the pages that use the platform, with different types of charts belonging to different packages as shown in Figure 90.

```
google.load("visualization", "1",{packages:["corechart","wordtree","treemap"]});
google.load("visualization", "1.1", { packages: ["table"] });
```

FIGURE 90 – LOADING GOOGLE PACKAGES

The charts are rendered using HTML5/SVG technology, providing browser compatibility and cross compatibility to iPhones, iPads and Android, so that a user will never have to have additional work to successfully interact with the application, making it available on almost any machine, as long as it has a web browser. Every time a representation is necessary, the application send outs a request with the data for the online Google Charts, which compiles a graph and delivers it back to the application that proceeds to show it.

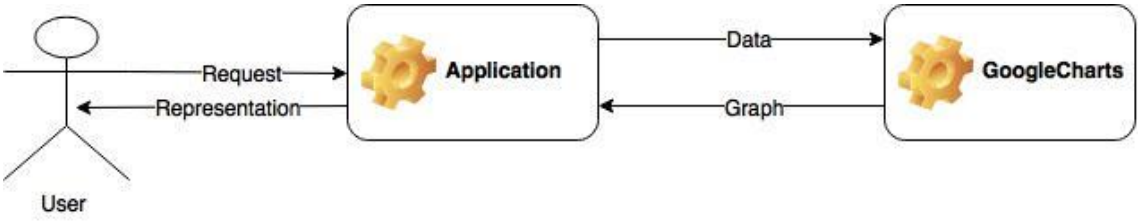

FIGURE 91 - GOOGLECHARTS INTERACTION

The application allows the users to dynamically choose between different graphs according to their needs. Changing the observed dataset will also trigger a request for an updated visualization; Every time the application is ready to draw, the google library's given the chosen dataset and the chart's specifics.

```
var chart = new
google.visualization.PieChart(document.getElementById('chart_div'));
options = {
          title: title,
          is3D: true,
          backgroundColor: "transparent",
          legend: 'right'
          };
chart.draw(data, options);
```

FIGURE 92 - GOOGLECHARTS REQUEST EXAMPLE

There's a wide array of available charts, each with some level of interactivity, with symbols and colors used often to better illustrate relationships.

The Google Charts API is used to represent both the knowledge mined by the user and the general statistics retrieved from the data.

The kind of representations available for the user varies according to the types of attributes chosen.

If they're both numerical attributes, then the available representations are:

- Column chart, with vertical bars as seen in Figure 109.
- Bar chart, with horizontal bars as seen in Figure 110.
- Line chart, which draws a line through the points where the two attributes are related as seen in Figure 111
- Scatter chart shows a point for each value where the two attributes are related as seen in Figure 112

If one of the attributes is numerical and the other one is a string, then the available representations are:

- The bar chart with a string attribute draws horizontal bars with the string attribute always placed in the y's axis, as seen in Figure 113.
- The column chart with a string attribute draws vertical bars, with the string attribute always placed on the x's axis, as seen in Figure 114.
- Pie chart is a circle that holds the percentages of instances by each of the possibilities as seen in Figure 115
- The 3D pie chart is a 3 dimensional circle that holds the percentages of instances held by each of the possibilities. Figure 116shows the gender distribution of a training dataset of students between 2010 and 2011. The additional information is shown by dragging the cursor across one of the components of chart as seen in Figure 117
- The Donut chart is much like a pie chart, only it has a hole in it and isn't 3 dimensional as seen in Figure 118
- The histogram draws vertical bars, each measuring the amount of instances that fit a particular condition as seen in Figure 119
- The line chart with a string attribute draws a line between the points where the two attributes are related, with the string attribute always placed in the x's axis, as seen in Figure 120
- If both the attributes are strings, then an organizational chart is displayed, as seen in Figure 121. This particular chart allows the user to select the top node value, and then proceeds to draw the bottom nodes accordingly.

There's also a particular set of datasets used to detail the performances at each subject, indicating each of their variables, as well as the average approval rate and how it compares to other subjects. These datasets are uniquely assigned to combo charts, which draws vertical columns portraying a range of numerical attributes, and then draws the line that represents the average value of one of them, as seen in Figure 122.

5.4.1.4.1. INPUT STRUCTURES

The charts used by this application have a predefined structure that has to be respected for them to work. For that reason, the data retrieved from the MySQL database has to be treated and transformed into a data table before being provided to the google charts platform.

For scatter charts, both the first and second columns of the data table have to be numerical attributes as shown in Figure **93**.

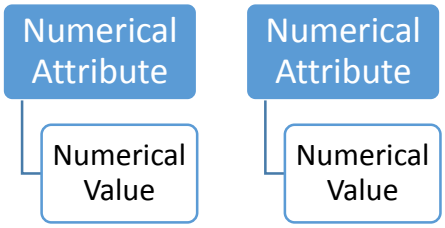

FIGURE 93 - SCATTER CHART STRUCTURE

For bar, column and line charts the data table's first column can be either a numerical or a string attribute, the second has to be a numerical attribute as seen in Figure **94**.

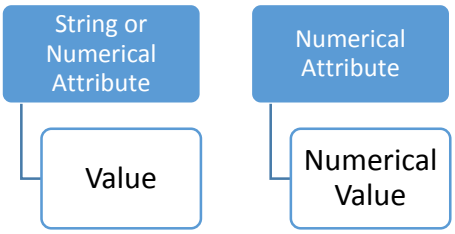

FIGURE 94 - BAR, COLUMN AND LINE CHART STRUCTURE

For histograms, donut and 3 dimensional pie charts, the data table's first column has to be a string attribute and the second column has to be a numerical attribute as shown in Figure 95.

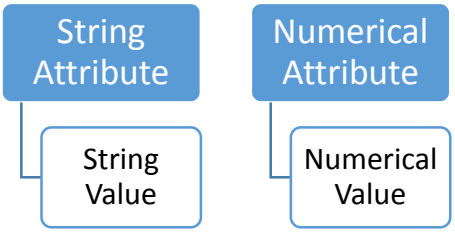

FIGURE 95 - HISTOGRAM, DONUT AND PIE CHART STRUCTURE

For the combo chart, the data table's first column has to be a string attribute and all of the following have to be numerical as described in Figure 96.

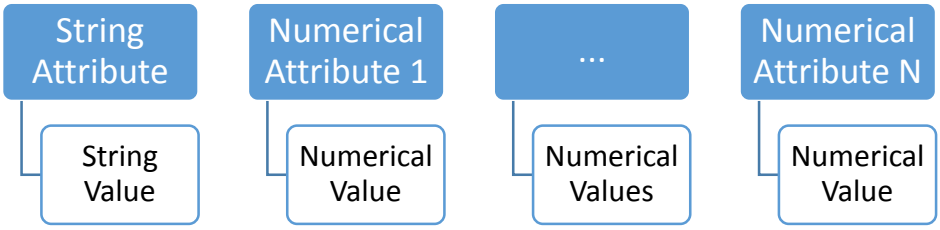

FIGURE 96 - COMBO CHART STRUCTURE

For the org chart, both the first and second columns of the data table have to be string attributes, the first one representing the node value, and the second one indicating a parent node, if any. This structure can be seen in Figure **97**.

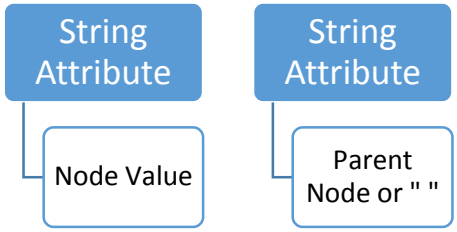

FIGURE 97 - ORG CHART STRUCTURE

5.4.1.4.2. GOOGLECHARTS IN DATA MINING

In the Data Mining page, the GoogleCharts platform is responsible for elaborating word trees to represent association rules, and tree maps and tables to represent clusters. When accessing the page, the user is able to choose the dataset to be represented, as well as the kind of data mining to be applied.

The only available type of representation when performing association rule mining, is a word tree as shown in Figure 67.

This word tree is in an explicit format and suffix type, meaning it takes a rule's antecedent component and draws its consequent components, with a subtitle indicating that the antecedent is written in 'black' and the consequent in 'red'. The antecedent being treated is dynamically chosen by the user.

When clustering data, the type of representations offered by GoogleCharts is either tree maps or tables:

5.4.1.4.3. INPUT STRUCTURES

The charts used by this application have a predefined structure that has to be respected for them to work. For that reason, the data mined by the user has to be treated and transformed into a data table before being provided to the google charts platform.

The word tree data table has five columns: the first column contains the node's ID, which is incremental, the second contains the node's unique label, the third its parent node's ID if any, the fourth its size and the fifth its color as seen in Figure 98.

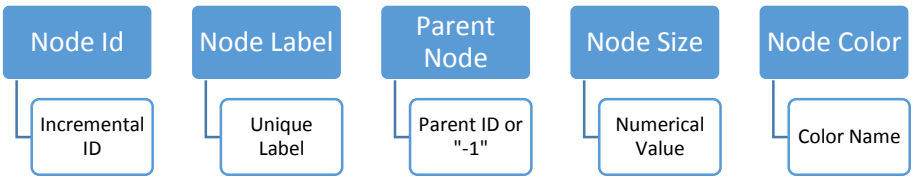

FIGURE 98 – WORD TREE STRUCTURE

For the tree map, the data table has three columns: the first column contains the node's unique label, the second its parent node, if any, and the third its size as shown in Figure **99**.

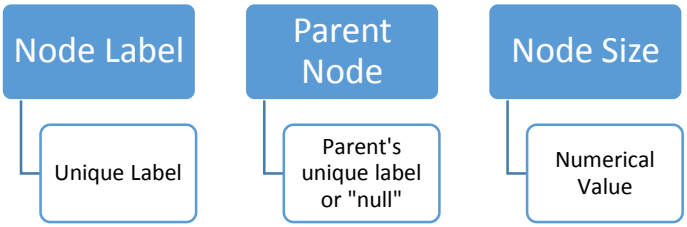

FIGURE 99 - TREE MAP STRUCTURE

Since the table chart represents its data in text, it doesn't have a fixed structure and any data table can be portrayed. The structure implemented has the node's unique label in the first column, the following columns with the cluster's attributes and the last columns with the number of instances in the cluster as represented in Figure **100**.

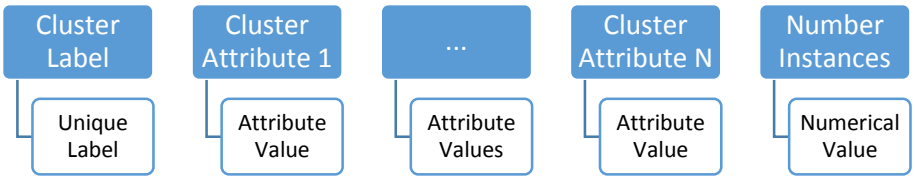

FIGURE 100 - TABLE CHART STRUCTURE

5.4.2. Social Networks

Social networks were implemented by importing a JavaScript library called Vis.js, transmuting a dataset into a fully scalable and easy to read and interact network for the user to analyze and collect valuable information from.

As previously referred, Vis.js is used to represent the Social networks in order to obtain a fully scalable and easy to read interactive network that the users can consult in their process to analyze a given dataset.

To load the library the code shown in Figure 101:

```
<script src="components/vis/vis.js"></script>
  <link href="components/vis/vis.css" rel="stylesheet" type="text/css" />
```

FIGURE 101 - IMPORTING VIS.JS

Both JavaScript and CSS files are loaded locally in the projects folder, eliminating the need to fire a request to its internet repository in runtime.

To make use of the Vis.js library, a div container was created to hold the required graphic for the final output. The dataset is treated in backend in order to assume a structure that fulfils the user's expectations throughout their network exploring experience.

The declaration of an example dataset can be seen in Figure 102.

```
var container = document.getElementById('visualization');
  // Create a DataSet (allows two way data-binding)
  var data = new vis.DataSet([
    {id: 1, content: 'item 1', start: '2013-04-20'},
    {id: 2, content: 'item 2', start: '2013-04-14'},
  ]);
<body>
<div id="visualization"></div>
</body>
```

FIGURE 102 – NETWORK EXAMPLE

5.4.2.1. INPUT STRUCTURES

In order to provide data to the social network libraries a predefined structure has to be respected. For that reason, the dataset chosen by the user has to be treated and transformed before the request to draw the network is fired.

A network is composed by: the first column contains the node's ID, which is incremental, the second contains the node's unique label, the third it's the node's color which is also unique, the fourth it's connected nodes and the fifth the weight of those connections. This structure is shown in Figure 103

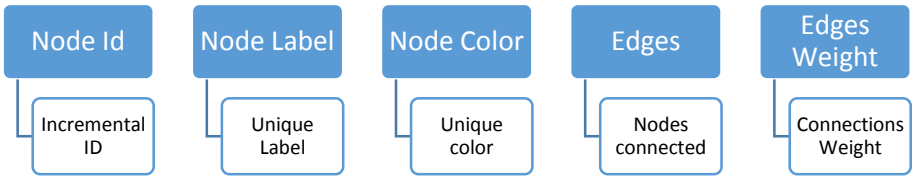

FIGURE 103 – SOCIAL NETWORK STRUCTURE

5.4.3. HTML2CANVAS

Html2Canvas is a script distributed under the MIT license that allows to save a webpage or a part of it as an image, acting as a sort of "screenshot". The script doesn't actually take a screenshot but rather builds the screenshot based on the DOM components present in the area to be "screenshot".

This implementation allows the user to download the charts visualized so that they can be checked offline or shared with others.

```
protected void downloadImage(object sender, EventArgs e)
        {
            string base64 =
Request.Form[hiddenChartImage.UniqueID].Split(',')[1];
            byte[] bytes = Convert.FromBase64String(base64);
            Response.Clear();
            Response.ContentType = "image/jpg";
            Response.AddHeader("Content-Disposition", "attachment;
filename=chart.jpg");
            Response.Buffer = true;
            Response.Cache.SetCacheability(HttpCacheability.NoCache);
            Response.BinaryWrite(bytes);
            Response.End();
        }
```

FIGURE 104 - HTML2CANVAS SNIPPET

As seen in Figure 104, when the download order is sent, the webpage takes the div where the chart is drawn and converts it to base64 data via the HTML2Canvas script. The base 64 data is then sent to the server which constructs the Figure contained in the data and returns it.

5.4.4. DATABASE

5.4.4.1. MYSQL WORKBENCH

MySQL workbench provides great a way to look at the database in an overall view. It allows to fully manipulate table's structure and data, build and interact with a domain model and perform changes to the database.

5.4.4.2. MYSQL FOR EXCEL

This plugin was valuable to the application development by providing the means to create and insert data to each table.

Both the raw and processed anonymous data were previously available in excel sheets, with this plugin being responsible for mapping it their respective tables in the MySQL database created for the application.

Each table was mapped to new sheets respecting the table's format.

Each table had a corresponding sheet with a table structure:

- Columns
- Rows

The first row contained the columns names while the following held the respective records.

An example of how tables are created in excel can be found in the following Figure:

	A	B	C	D	E	F	G	H	I
1	subject	subject	approved	failed	smr	sms	smnf	nc	nf
2	1	ALGAN	0,684782609	0,217391304	0	0	0	0,059782609	0,038043478
3	2	ALGAV	0,526315789	0,013245033	0,218543046	0	0	0,026490066	0
4	3	AMATA	0,441364606	0,234541578	0	0	0	0,159914712	0,164179104
5	4	APROG	0,60968661	0,071225071	0	0	0,202279202	0,011396011	0,105413105
6	5	ARQCP	0,398753894	0,009345794	0,271028037	0	0,183800623	0	0,137071651
7	6	ARQSI	0,574561404	0,105263158	0,078947368	0	0	0,096491228	0
8	7	ASIST	0,793103448	0,013888889	0,055555556	0	0	0	0
9	8	BDDAD	0,540166205	0,03601108	0,216066482	0	0,03601108	0,088642659	0,083102493

FIGURE 105- CREATING TABLES IN EXCEL

Figure 105 represents a small portion of a table. It is possible to observe the table's structure and conclude that it respects that of a regular table for a MySQL database.

After creating and mapping each table, using MySQL for Excel plugin, the data was inserted in the database tables to be accessed by the application backend that creates and enriches the representations available in frontend.

In the tables navigation pane it is possible to export data and create a table at the same time and also to import data from a previously created table to excel and manipulate it as desired.

This tool offers great flexibility in manipulating data and provides an easy way to interact with the database.

5.4.4.3. BLL

As exemplified in Figure 106 the business logic layer receives the requests from the underlying application and conveys the necessary commands to the DAL layer, which processes the requests and returns the required datasets.

```
internal static List<List<object>> getAllData(string table)
 {
        return DAL.getAllData(table);
 }
```

FIGURE 106 - BLL REQUEST EXAMPLE

5.4.4.4. DAL

As exemplified in Figure 107, when the DAL layer receives a request, it connects to the database via an ODBC connection, retrieves the information requested to a 'data' variable, and returns the output to the underlying layer, BLL.

```
        internal static List<List<object>> getAllData(string table)
        {
                using (OdbcConnection connection = new
OdbcConnection(ConfigurationManager.ConnectionStrings["****"].ConnectionString))
                {
                    connection.Open();
                    using (OdbcCommand command = new OdbcCommand(query, connection))
                    {
                        using (OdbcDataReader reader = command.ExecuteReader())
                        {
                            while (reader.Read())
                            {
                            }
                            reader.Close();
                        }
                    }
                    connection.Close();
                }
            }
            return data;
        }
```

FIGURE 107 - DAL REQUEST EXAMPLE

5.5. Application Demonstration

This section provides a look at the developed application for each of its main roles.

5.5.1. Homepage

The homepage is simple and acts as a starting point with a navigation menu for the web application pages.
This page is illustrated in Figure 130.

5.5.2. General Statistics

The general statistics page provides several type of charts and lets the user pick which data he wants to see.
This page is illustrated in Figure 131.

5.5.3. Association Rules and Clusters

The association rules and clustering page allows the user to pick one of the data mining techniques. After selecting one, a set of possible datasets are presented and a simple configuration panel pops up to allow the user to input his preferred values.

This page is illustrated in Figure 132 and Figure 133.

5.5.4. Social Networks

Finally the social networks page is a prototype and allows the user to see a network of the subjects available in the system. The subjects are presented as nodes and are connected by their correlation, represented as edges.

This page is illustrated in Figure 134.

6. WORKPLAN

As referred in chapter 2, this project adopted a RUP methodology, meaning the development and tasks associated with it were not a part of a concrete process, but rather one that progressively and dynamically changed as tasks were defined or completed. The work plan originally created is shown in Figure 126 and Figure 127 in the appendix.

Since not all of the requirements and technologies were known in the early stages of the project, the processes described in these figures suffered changes with each iteration. There were several occasions where the work plan did not result as intended, with some tasks taking longer than previously predicted and new tasks being defined as the project developed.
The final work plan is shown in Figure 128 and Figure 129

7. CONCLUSIONS

Data mining and analysis has gained great momentum in recent years. Accurately applied, it's possible to save time and money, promoting a more effective and economic environment, and to create new business opportunities.

The interface and integrated dashboard make it easier to interact, visualize and study the system's dynamics, making for a smoother and friendlier analysis of the educational program.

When applied to the Polytechnic of Porto – School of Engineering's Software Engineering course, the solution presented will be able to extract and treat pertinent data so that the user can study patterns, clusters and networks that would hardly be detected otherwise.

This tool can change the way school's approach and prepare students by actively providing feedback to the implemented measures, not only pointing out when things aren't working as intended and can be improved, but also by discovering what's right and allowing us to learn from it. The possibilities are endless and the project has the potential to possibly become a national 'standard' in the future.

7.1. THESIS CONTRIBUTIONS

This thesis contributes to the educational data mining field of expertise. Specifically, it helps improving the quality of service of education institutions by analyzing all the available data about students until today with computer software, providing the necessary means to reach important conclusions that can be presented to the academic institution collaborators and promote changes that can greatly impact the overall performance.

The primary objectives of this thesis are:

- Develop a data mining application capable of offering solutions to nowadays academic institutions
- Study, experiment and apply association rules algorithms like Apriori in academic data
- Study, experiment and apply clustering algorithms like Simple K-Means in academic data
- Build a social network prototype regarding the correlation between different academic entities.

The thesis focused on studying, analyzing and developing an application in the scope of extracting association rules, clusters and Social networks from a centralized database filled with data describing the academic route of each student, subject lifecycle, etc.

Providing conclusions and predictions about several aspects of the school system, offers the tools that promote well informed decision making, which in its turn can greatly improve the academic quality of service and greatly impact the current educational model.

Chapter 1 is an introduction of the project problem where several aspects regarding the need to study the school data in order to improve education quality of service are documented.

Chapter 2 contains the state of art where many different software tools and approaches are presented, compiling an overview of the current state of technological progress regarding the theme approached by this thesis. There is a brief description and illustration of several data mining tools that are currently used, some of them too for educational purposes.

Chapter 3 is where all the tools, methodologies, frameworks and study performed for this thesis are introduced. It is possible to find very important information that is the "floor" of the development of this thesis document and application. An educational data mining tool is both powerful and important, and this section describes why.

Chapter 4 refers to a set of experiments used to analyze the dataset, show how useful the implemented methodologies are and the value that can be drawn from their usage. The experiments brought several conclusions that were applied in the development phase of the application. The background and information provided by experimenting was very positive in order to enrich the knowledge for EDM.

Chapter 5 is where the engineering requirements of this thesis's application are described.

Chapter 6 starts with the application architecture and all the software tools, frameworks and patterns that were applied to develop the final application. In order to enrich the final output several implementations were performed and tested accordingly the study performed for the thesis. Since this was an iterative and incremental process many of the technologies introduced were subject to improvement or change until the date this document was written.

Chapter 7 explains the application design and details its workflow, describing the interactions of between systems and the messages being traded.

Chapter 8 refers to the implementation with technical details and code snippets providing the overview over what was used and how it was used.

Chapter 9 contains a demonstration sample of the developed application showing its main features.

Chapter 10 compares the initial and final work plans developed and discusses the changes verified, as well as the reasons behind them.
Chapter 11 is where the final conclusions about the thesis are presented, as well as its contributions, limitations and future work.
Chapter 12 refers to the document references and citations.
Chapter 13 contains the attachments.
As a final remark, it is very important to acknowledge that much of the inspiration and motivation for this work derived from the vision of the future of information applications in educational data mining. It's also important to note that the technologies applied in this thesis can be used for other scenarios.

7.2. LIMITATIONS

Limitations are mainly found in the application developed and the room left for improvement. Being in technology branch, applications need continuous delivery of content as they struggle in reaching a "perfect" phase. However, this application may lack a few topics.

7.2.1. GENERAL STATISTICS

General statistics are only available through the GoogleCharts platform, having an implicit dependency on a network connection. All of the graphics provided are also related to this platform and thus have the same limitations it has, with a limited set of visual representations and the attribute types and structure that go with them.
Although the number of charts with only numerical or both numerical and text variables is reasonable, there is not a lot to work with when it comes to multiple text variables, thus limiting what the application could offer in such situations.

7.2.2. ASSOCIATION RULES

The association rules could only be generated with some parameters of the Apriori algorithm, thus limiting the extent of discovery that can be done. The parameters and algorithm implemented are the same that were used in the experimentations chapter and therefore, the conclusions that can be drawn are the same.

There's also the limitation of not being able to dynamically change, normalize, trim or delete values of the dataset, but rather use what's been given to the user.

7.2.3. CLUSTERING

Clustering's main limitation is the lack of options, with only one of the algorithms being studied and applied.
In order to provide more conclusions from different backgrounds more data should be added to the equation and probabilistic data should also be prepared and set. The application's clustering option is limited to the Simple K-Means algorithm and to the training sets previously prepared and provided by the database. The data available also does not show differences between euclidean and manhattan distances.
There could also be more graphical options for the use to choose from, with more details over the instances being shown.

7.2.4. SOCIAL NETWORKS

The Social networks developed for the application can be explored further in order to offer the user more options or perform more actions dynamically, responding to the user's requests
Developing and implementing a social network can be very useful yet it requires a good background to do so. To offer more social network algorithm, more academic knowledge required.

7.3. FUTURE WORK

Future work mainly focus in studying more data mining techniques and tools in order to enrich and offer more options to the academic world.
It's important to gather more and better data so that the results are more accurate and reliable.
It is possible to offer more association rules and clustering algorithms and provide additional configurable parameters in the application. Integrating the system with more data mining tools can also provide different backgrounds and present succinct conclusions to be compared with similar results.

There’s a need to implement preprocessing options so that the user can edit and normalize a given dataset at will, empowering them to fully explore the quantity and quality of results.

Clustering can be performed in many different and its output can be improved before running the algorithms which was not detailed and fully applied in this thesis. Implementing more algorithms, provide more configuration options and have more datasets prepared would greatly improve the application’s overall performance.

A mechanism of login and roles for each user is also a very important aspect for future work, allowing for different hierarchical entities to access different layers of data, with different output treatments, datasets and configuration options being offered to different users.

8. REFERENCES

Abernethy, M., 2010. *Data mining with WEKA, Part 1: Introduction and regression.* [Online] Available at: http://www.ibm.com/developerworks/library/os-weka1/ [Acedido em 30 May 2015].

Albion Research, s.d. *What Is Data Mining?.* [Online] Available at: http://www.albionresearch.com/data_mining/ [Acedido em 15 Mar 2015].

Andale, 2015. *Pearson Correlation: Definition and Easy Steps for Use.* [Online] Available at: http://www.statisticshowto.com/what-is-the-pearson-correlation-coefficient/ [Acedido em 1 Sept 2015].

Anon., 2006. *Pentaho Open Source Business Intelligence Platform Technical White Paper.* [Online] Available at: http://sourceforge.net/projects/pentaho/files/White%20Papers/

Anon., 2015. *Cross Industry Standard Process for Data Mining.* [Online] Available at: http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Anon., 2015. *What is QlikView.* [Online] Available at: http://www.visualintelligence.co.nz/qlikview/ [Acedido em 29 May 2015].

Anon., s.d. *BrightBytes.* [Online] Available at: http://brightbytes.net/ [Acedido em 17 Mar 2015].

Anon., s.d. *IKVM.NET.* [Online] Available at: http://www.ikvm.net/ [Acedido em 17 Jul 2015].

Anon., s.d. *Introduction to Weka.* [Online] Available at: https://www.cs.auckland.ac.nz/courses/compsci367s1c/tutorials/IntroductionToWeka.pdf

Anon., s.d. *Item-based collaborative filtering.* [Online] Available at: http://www.cs.carleton.edu/cs_comps/0607/recommend/recommender/itembased.html [Acedido em 1 Aug 2015].

Anon., s.d. *Jaspersoft.* [Online] Available at: http://www.jaspersoft.com/ [Acedido em 10 May 2015].

Anon., s.d. *Jaspersoft.* [Online] Available at: http://www.sodexis.com/services/what-is-jaspersoft-longwood-orlando-florida.html [Acedido em 10 May 2015].

Anon., s.d. *Pentaho BI Platform FAQ.* [Online] Available at: http://community.pentaho.com/faq/bi_platform.php

Anon., s.d. *R.* [Online]
Available at: http://www.R-project.org/
[Acedido em 16 Apr 2015].

Anon., s.d. *Rational Unified Process.* [Online]
Available at: https://www.ibm.com/developerworks/rational/library/content/03July/1000/1251/1251_bestpractices_TP026B.pdf
[Acedido em 19 Mar 2015].

Anon., s.d. *RUP Fundamentals Presentation.* [Online]
Available at: http://era.nih.gov/docs/rup_fundamentals.htm
[Acedido em Mar 19 2015].

Anon., s.d. *SAS® Enterprise Analytics for Education.* [Online]
Available at: https://www.sas.com/en_us/industry/higher-education/enterprise-analytics-for-education.html
[Acedido em 6 Jun 2015].

Anon., s.d. *Tableau Software - Visão Geral.* [Online]
Available at: http://www.xpand-it.com/pt/tecnologias/tableau
[Acedido em 11 Jul 2015].

Anon., s.d. *Using Google Charts.* [Online]
Available at: https://developers.google.com/chart/interactive/docs/
[Acedido em 22 Jun 2015].

Arthur, S. V. a. D., s.d. *K-means++: The advantages of careful seeding.* [Online]
Available at: http://theory.stanford.edu/~sergei/slides/BATS-Means.pdf
[Acedido em 28 May 2015].

Behrouz Minaei-Bidgoli, D. A. K. G. K. W. F. P., 2003. *PREDICTING STUDENT PERFORMANCE: AN APPLICATION OF DATA MINING METHODS WITH THE EDUCATIONAL WEB-BASED SYSTEM LON-CAPA,* Boulder, Colorado: s.n.

Borah, R. J. a. M. D., 2013. *A Survey on Educational Data Mining And Research Trends.* [Online]
Available at: http://airccse.org/journal/ijdms/papers/5313ijdms04.pdf
[Acedido em 15 Jul 2015].

Business Week, 2015. *Company Overview of Tableau Software, Inc..* [Online]
Available at: http://investing.businessweek.com/businessweek/research/stocks/private/snapshot.asp?privcapId=11421199
[Acedido em 10 Oct 2015].

Butts, C. T., 2008. Social network analysis: A methodological introduction. *Asian Journal of Social Psychology,* p. 29.

Chamatkar, A. J., 2014. *Importance of Data Mining with Different Types of Data Applications and Challenging Areas.* [Online]
Available at:

http://www.academia.edu/7675788/Importance_of_Data_Mining_with_Different_Types_of_Data_Applications_and_Challenging_Areas
[Acedido em 1 Jun 2015].

Council, C., 2014. *Yellowfin.* [Online]
Available at: http://technologyadvice.com/products/yellowfin-reviews/
[Acedido em May 6 2015].

Cronstrom, H., 2014. *QlikView and Qlik Sense.* [Online]
Available at: https://community.qlik.com/blogs/qlikviewdesignblog/2014/07/29/view-or-sense
[Acedido em 6 Apr 2015].

David Ronka, M. A. L. R. S. a. J. M., 2009. *Educational Leadership - Answering the Questions That Count.* [Online]
Available at: http://www.ascd.org/publications/educational-leadership/dec08/vol66/num04/Answering-the-Questions-That-Count.aspx
[Acedido em 22 Jul 2015].

Dell, 2015. *Data Mining Techniques.* [Online]
Available at: http://documents.software.dell.com/Statistics/Textbook/Data-Mining-Techniques

Dell, s.d. *How To Group Objects Into Similar Categories, Cluster Analysis.* [Online]
Available at: http://www.statsoft.com/Textbook/Cluster-Analysis
[Acedido em 17 Jun 2015].

Editechreview, 2013. *What is Educational Data Mining (EDM)?.* [Online]
Available at: http://edtechreview.in/dictionary/394-what-is-educational-data-mining
[Acedido em 10 Jun 2015].

Educational Data Mining, 2015. *EDM 2015 - The 8th International Conference.* [Online]
Available at: http://www.educationaldatamining.org/EDM2015/
[Acedido em 1 July 2015].

Estivill-Castro, V., 2002. *Why so many clustering algorithms — A Position Paper.* [Online]
Available at: http://dl.acm.org/citation.cfm?doid=568574.568575
[Acedido em 30 Jul 2015].

Ferreira, C., 2006. *Designing Neural Networks Using.* [Online]
Available at: http://www.gene-expression-programming.com/webpapers/Ferreira-ASCT2006.pdf

Flexidash, s.d. *Flexidash.* [Online]
Available at: https://flexidash.com/
[Acedido em 1 Apr 2015].

Frand, J., s.d. *Data Mining: What is Data Mining?.* [Online]
Available at: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[Acedido em 22 Feb 2015].

Garg, R. L. a. K., July 2012. *Effect of Distance Functions on K-Means Clustering.* s.l.:International Journal of Computer Applications (0975 – 8887).

Géryk, J., 2015. *Using Visual Analytics Tool for Improving Data.* [Online] Available at: http://www.educationaldatamining.org/EDM2015/uploads/papers/paper_65.pdf [Acedido em 8 Jul 2015].

Infoescolas, 2015. *Estatísticas do Ensino Secundário.* [Online] Available at: http://infoescolas.mec.pt/ [Acedido em 19 Jun 2015].

Jaspersoft, 2011. *JASPERSOFT SCHOLARS PROGRAM PROVIDES FREE ENTERPRISE BI SOFTWARE TO HIGHER EDUCATION.* [Online] Available at: http://www.jaspersoft.com/press/scholars-program [Acedido em 9 Apr 2015].

John Scott, P. J. C., 2011. *The SAGE Handbook of Social Network Analysis.* s.l.:SAGE Publications.

Karl Rexer, H. A. &. P. G., 2011. *2011 Data Miner Survey Summary.* [Online] Available at: http://www.rexeranalytics.com/Data-Miner-Survey-Results-2011.html [Acedido em 1 Oct 2015].

Kiri Wagstaf, C. C. S. R. S. S., 2001. *Constrained K-means Clustering with Background Knowledge.* [Online] Available at: https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf [Acedido em 17 Jun 2015].

Lane, D. M., s.d. *Values of the Pearson Correlation.* [Online] Available at: http://onlinestatbook.com/2/describing_bivariate_data/pearson.html [Acedido em 6 Sept 2015].

Mark Hall, E. F. G. H. B. P. P. R. I. H. W., 2009. *The WEKA Data Mining Software: An Update.* s.l.:SIGKDD Explorations.

MateuCC, s.d. *A Tutorial on Clustering Algorithms - K-Means Clustering.* [Online] Available at: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html [Acedido em 6 May 2015].

Matheson, R., 2015. *High "return-on-learning".* [Online] Available at: http://newsoffice.mit.edu/2015/brightbytes-business-intelligence-data-analytics-schools-0302 [Acedido em 30 Mar 2015].

McCullock, J., 2012. *Clustering k means example 1.* [Online] Available at: http://mnemstudio.org/clustering-k-means-example-1.htm [Acedido em 6 Mar 2015].

McGee, M. K., 2009. *Ohio Schools Use Business Intelligence To Improve Student Performance.* [Online] Available at: http://www.informationweek.com/ohio-schools-use-business-intelligence-to-improve-student-performance/d/d-id/1080159? [Acedido em 13 Jul 2015].

MING-CHUAN HUNG, J. W., 2005. *Department of Information Engineering and Computer Science.* [Online] Available at: http://www.iis.sinica.edu.tw/JISE/2005/200511_04.pdf [Acedido em 3 Jun 2015].

Minitab® 17 Support, 2015. *What is a cluster centroid?.* [Online] Available at: http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/multivariate/item-and-cluster-analyses/what-is-a-cluster-centroid/

Moody, G., 2010. *How Do You Make a Pentaho?.* [Online] Available at: http://www.computerworlduk.com/blogs/open-enterprise/how-do-you-make-a-pentaho-3568902/
[Acedido em 25 May 2015].

Moore, A., 2005. The case for approximate Distance Transforms. Em: s.l.:University of Otago. Dunedin, New Zealand.

Morgner, T., 2010. *What is Pentaho Reporting.* [Online] Available at: http://wiki.pentaho.com/display/Reporting/What+is+Pentaho+Reporting
[Acedido em 25 May 2015].

Naik, A., 2014. *Data Clustering Algorithms - Simple K-Means.* [Online] Available at: https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
[Acedido em 1 May 2015].

O'Connor, B., 2012. *Cosine similarity, Pearson correlation, and OLS coefficients.* [Online] Available at: http://brenocon.com/blog/2012/03/cosine-similarity-pearson-correlation-and-ols-coefficients/
[Acedido em 1 Sept 2015].

Pang-Ning Tan, M. S. &. V. K., 2006. Association Analysis: Basic Concepts and Algorithms. Em: *Introduction to Data Mining.* s.l.:s.n.

Pang-Ning Tan, M. S. &. V. K., 2006. Cluster Analysis: Basic Concepts and Algorithms. Em: *Introduction to Data Mining.* s.l.:s.n.

Pedro Strecht, J. M.-M. &. C. S., s.d. *Merging Decision Trees: A Case Study in Predicting Student Performance,* INESC TEC/Faculdade de Engenharia: s.n.

Pentaho, s.d. *Pentaho, A Hitachi Data Systems Company.* [Online] Available at: http://www.pentaho.com/

Pete Chapman, J. C. R. K. T. K. T. R. C. S. &. R. W., s.d. *Crisp-DM 1.0.* [Online] Available at: http://the-modeling-agency.com/crisp-dm.pdf
[Acedido em 13 May 2015].

Popelínský, J. G. &. L., s.d. *Analysis of Student Retention and Drop-out using Visual Analytics.* [Online] Available at: http://educationaldatamining.org/EDM2014/uploads/procs2014/posters/7_EDM-2014-Poster.pdf
[Acedido em 17 Mar 2015].

Qlik, 2015. *Business Discovery for Education.* [Online] Available at: http://www.qlik.com/us/explore/solutions/industries/public-sector/education
[Acedido em 21 Jun 2015].

Qlik, 2015. *Qlik Academic Program.* [Online] Available at: http://global.qlik.com/dk/company/academic-program [Acedido em 9 Aug 2015].

Qlik, 2015. *Qlik Sense Desktop.* [Online] Available at: http://www.qlik.com/us/explore/products/sense

Rajaraman, J. a. A., s.d. Clustering Algorithms. Em: *CS345a:DataMining.* s.l.:Stanford University, p. 46.

Rakesh Agrawal, T. I. &. A. S., 1993. *Mining association rules between sets of items in large databases.* [Online] Available at: http://dl.acm.org/citation.cfm?doid=170035.170072 [Acedido em 30 Apr 2015].

Rouse, M., 2011. *association rules (in data mining).* [Online] Available at: http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining [Acedido em 1 Mar 2015].

Rouse, M., s.d. *Rational Unified Process (RUP).* [Online] Available at: http://searchsoftwarequality.techtarget.com/definition/Rational-Unified-Process [Acedido em 28 Feb 2015].

Russell, I., 1996. *Neural Networks Module.* [Online] Available at: http://uhaweb.hartford.edu/compsci/neural-networks-definition.html

Sang Su Lee, D. W. a. D. M., s.d. Discovering Relationships among Tags and Geotags. Em: Los Angeles, CA 90089 : Computer Science Department University of Southern California.

SAS, 2015. *SAS Visual Analytics and Teradata University Network.* [Online] Available at: http://www.sas.com/en_us/offers/14q3/teradata-university-network-tun/overview.html [Acedido em 6 Jun 2015].

SAS, s.d. *Data Mining What it is and why it matters.* [Online] Available at: http://www.sas.com/en_lu/insights/analytics/data-mining.html

SAS, s.d. *SAS Enterprise Miner.* [Online] Available at: http://www.sas.com/en_id/software/analytics/enterprise-miner.html [Acedido em 6 Jun 2015].

SAS, s.d. *SAS OnDemand for Academics.* [Online] Available at: http://www.sas.com/en_us/industry/higher-education/on-demand-for-academics.html [Acedido em 16 Jun 2015].

Sayad, D. S., 2010-2015. *An Introduction To Data Mining - K-Means Clustering.* [Online] Available at: http://www.saedsayad.com/clustering_kmeans.htm [Acedido em 26 Mar 2015].

Sayad, D. S., s.d. *Clustering.* [Online] Available at: http://www.saedsayad.com/clustering.htm [Acedido em 26 Mar 2015].

Senior Technology Advisor, O. o. T. A. V. P. f. F. a. U. C., 2009. *Leading institution of higher education optimizes financial & procurement processes with QlikView.* [Online] Available at: http://www.perfect-image.co.uk/upload/downloads/case-studies/University-Success-Story-EN.pdf
[Acedido em 5 Aug 2015].

Silva, P. C. a. A., n.d. *USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE,* Guimarães: s.n.

Smith, D., 2012. *R Tops Data Mining Software Poll.* [Online] Available at: http://java.sys-con.com/node/2288420
[Acedido em 20 Apr 2015].

Stephen P. Borgatti, A. M. D. J. B. G. L., 2009. *Network Analysis in the Social Sciences.* [Online] Available at: http://www.sciencemag.org/content/323/5916/892
[Acedido em 28 Mar 2015].

Strangroom, J., 2015. *Pearson Correlation Coefficient Calculator.* [Online] Available at: http://www.socscistatistics.com/tests/pearson/
[Acedido em 10 Sep 2015].

Tableau, 2015. *Tableau Education Reporting.* [Online] Available at: http://www.tableau.com/solutions/education-analytics
[Acedido em 16 May 2015].

Tableau, 2015. *Tableau for Students.* [Online] Available at: http://www.tableau.com/academic/students
[Acedido em 16 May 2015].

Tableau, s.d. *Tableau Business Intelligence.* [Online] Available at: http://www.tableau.com/business-intelligence
[Acedido em 11 Jul 2015].

Tapas Kanungo, D. M. M. ,. N. S. N. ,. C. P. ,. R. S. ,. A. Y. W., s.d. *The Analysis of a Simple k-Means Clustering Algorithm (2000).* [Online] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.341&rep=rep1&type=pdf
[Acedido em 1 Mar 2015].

Taylor, J., 2011. *First Look – SAS Enterprise Miner 7.1.* [Online] Available at: http://jtonedm.com/2011/11/11/first-look-sas-enterprise-miner-7-1/
[Acedido em 8 May 2015].

Traxion Consulting, 2011. *News Detail » What is QlikView? Good question.....* [Online] Available at: http://www.traxionconsulting.com/news_detail.php?id=17

Variar, G., 2005. *Sophisticated Reporting Bolsters An Enterprise-Ready BI Suite.* [Online] Available at: http://www.informationweek.com/software/information-management/sophisticated-reporting-bolsters-an-enterprise-ready-bi-suite/d/d-id/1034140?

Wan Aezwani Wan Abu Bakar, M. A. J. a. M. Y. M. S., s.d. *Mining Educational Data : A Perspective Review on Data Mining Suites.* [Online]

Available at: http://www.academia.edu/7036598/Mining_Educational_Data_A_Perspective_Review_on_Data_Mining_Suites
[Acedido em 15 Mar 2015].

Whitehorn, M., 2006. *The parable of the beer and diapers.* [Online] Available at: http://www.theregister.co.uk/2006/08/15/beer_diapers/
[Acedido em 5 May 2015].

White, T., s.d. *Webfocus.* [Online] Available at: http://www.informationbuilders.com/products/webfocus
[Acedido em 19 May 2015].

Wikibooks, 2015. *Data Mining Algorithms In R/Frequent Pattern Mining/The Apriori Algorithm.* [Online] Available at: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_Apriori_Algorithm

Wikipedia, 16 July 2015. *Lloyd's Algorihtm.* s.l.:Wikipedia.

Wikipedia, 2015. *K-Means Clustering.* [Online] Available at: https://en.wikipedia.org/wiki/K-means_clustering

Wikipedia, 2015. *Voronoi Diagram.* [Online] Available at: https://en.wikipedia.org/wiki/Voronoi_diagram

Wizdee, s.d. *Wizdee.* [Online] Available at: http://wizdee.com
[Acedido em 29 May 2015].

Yacef, B. &., 2009. The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining.*

Yellowfin, 2014. *Education Analytics: Why schools should care about Business Intelligence.* [Online] Available at: http://www.yellowfinbi.com/YFCommunityNews-Education-Analytics-Why-schools-should-care-about-Business-Intelligence-154884
[Acedido em 30 Jul 2015].

Yellowfin, s.d. *An Analytics Platform That Is Incredibly Easy To Use.* [Online] Available at: http://www.yellowfinbi.com/YFWebsite-Business-Intelligence-and-Analytics-Platform-24427
[Acedido em 30 Mar 2015].

Zhao, Y., 2015. *RDM.* [Online] Available at: http://www.rdatamining.com/resources/courses
[Acedido em 19 Apr 2015].

9. APPENDIX

9.1. MULTIPLE SUBJECTS SCENARIO EXPERIMENTS

9.1.1. MATH RELATED SUBJECTS

TABLE 98 A.R EXPERIMENT – MATH RELATED SUBJECTS SYNERGY

Support	**Confidence**
0,4	0,6
Result	**Description**
ALGAN='(10-11]' 272 ==> FSIAP='(10-11]' 206 conf:(0.76)	Scores between 10-11 at ALGAN usually indicate scores between 10-11 at FSIAP
AMATA='(10-11]' 246 ==> FSIAP='(10-11]' 184 conf:(0.75)	Scores between 10-11 at AMATA usually indicate scores between 10-11 at FSIAP
FSIAP='(10-11]' 319 ==> ALGAN='(10-11]' 206 conf:(0.65)	Scores between 10-11 at FSIAP usually indicate scores between 10-11 at ALGAN
FSIAP='(10-11]' 319 ==> AMATA='(10-11]' 184 conf:(0.58)	Scores between 10-11 at FSIAP usually indicate scores between 10-11 at AMATA

TABLE 99 A.R EXPERIMENT – MATH RELATED SUBJECTS SYNERGY

Support	**Confidence**
0,35	0.4
Result	**Description**
ALGAN='(10-11]' 272 ==> FSIAP='(10-11]' 206 conf:(0.76)	Scores between 10-11 at ALGAN usually indicate scores between 10-11 at FSIAP
AMATA='(10-11]' 246 ==> FSIAP='(10-11]' 184 conf:(0.75)	Scores between 10-11 at AMATA usually indicate scores between 10-11 at FSIAP

AMATA='(10-11]' 246 ==> ALGAN='(10-11]' 167 conf:(0.68)	Scores between 10-11 at AMATA usually indicate scores between 10-11 at ALGAN
FSIAP='(10-11]' 319 ==> ALGAN='(10-11]' 206 conf:(0.65)	Scores between 10-11 at FSIAP usually indicate scores between 10-11 at ALGAN
ALGAN='(10-11]' 272 ==> AMATA='(10-11]' 167 conf:(0.61)	Scores between 10-11 at ALGAN usually indicate scores between 10-11 at AMATA
FSIAP='(10-11]' 319 ==> AMATA='(10-11]' 184 conf:(0.58)	Scores between 10-11 at FSIAP usually indicate scores between 10-11 at AMATA

TABLE 100 A.R EXPERIMENT – MATH RELATED SUBJECTS SYNERGY

Support	**Confidence**
0,25	0.35
Result	**Description**
MDISC='(10-11]' 172 ==> FSIAP='(10-11]' 132 conf:(0.77)	Scores between 10-11 at MDISC usually indicate scores between 10-11 at FSIAP
MATCP='(10-11]' 169 ==> FSIAP='(10-11]' 118 conf:(0.7)	Scores between 10-11 at MATCP usually indicate scores between 10-11 at FSIAP
MATCP='(10-11]' 169 ==> ALGAN='(10-11]' 117 conf:(0.69)	Scores between 10-11 at MATCP usually indicate scores between 10-11 at FSIAP
MDISC='(10-11]' 172 ==> ALGAN='(10-11]' 117 conf:(0.68)/	Scores between 10-11 at MDISC usually indicate scores between 10-11 at ALGAN
ALGAN='(10-11]' 272 ==> MATCP='(10-11]' 117 conf:(0.43)	Scores between 10-11 at ALGAN usually indicate scores between 10-11 at MATCP
ALGAN='(10-11]' 272 ==> MDISC='(10-11]' 117 conf:(0.43)	Scores between 10-11 at ALGAN usually indicate scores between 10-11 at MDISC
FSIAP='(10-11]' 319 ==> MDISC='(10-11]' 132 conf:(0.41)	Scores between 10-11 at FSIAP usually indicate scores between 10-11 at MDISC

9.1.2. LAPR SUBJECTS

TABLE 101 A.R EXPERIMENT – LAPR SUBJECTS SYNERGY

Support	Confidence
0.2	0.4
Result	**Description**
LAPR2='(12-13]' 178 ==> LAPR3='(12-13]' 98 conf:(0.55)	Scores between 12 and 13 at LAPR2 usually indicates scores between 12 and 13 at LAPR3
LAPR3='(12-13]' 217 ==> LAPR2='(12-13]' 98 conf:(0.45)	Scores between 12 and 13 at LAPR3 usually indicates scores between 12 and 13 at LAPR2

TABLE 102 A.R EXPERIMENT – LAPR SUBJECTS SYNERGY

Support	Confidence
0.15	0.35
Result	**Description**
LAPR2='(12-13]' 178 ==> LAPR3='(12-13]' 98 conf:(0.55)	Scores between 12 and 13 at LAPR2 usually indicates scores between 12 and 13 at LAPR3
LAPR1='(12-13]' 168 ==> LAPR3='(12-13]' 91 conf:(0.54)	Scores between 12 and 13 at LAPR1 usually indicates scores between 12 and 13 at LAPR3
LAPR5='(14-15]' 137 ==> LAPR3='(12-13]' 72 conf:(0.53)	Scores between 14-15 at LAPR5 usually indicates scores between 12 and 13 at LAPR3
LAPR3='(12-13]' 217 ==> LAPR2='(12-13]' 98 conf:(0.45)	Scores between 12 and 13 at LAPR3 usually indicates scores between 12 and 13 at LAPR2
LAPR4='(14-15]' 159 ==> LAPR3='(12-13]' 70 conf:(0.44)	Scores between 14-15 at LAPR4 usually indicates scores between 12 and 13 at LAPR3
LAPR4='(14-15]' 159 ==> LAPR2='(12-13]' 69 conf:(0.43)	Scores between 14-15 at LAPR4 usually indicates scores between 12 and 13 at LAPR2
LAPR1='(12-13]' 168 ==> LAPR2='(12-13]' 71 conf:(0.42)	Scores between 12 and 13 at LAPR1 usually indicates scores between 12 and 13 at LAPR2

LAPR3='(12-13]' 217 ==> LAPR1='(12-13]' 91 conf:(0.42)	Scores between 12 and 13 at LAPR3 usually indicates scores between 12 and 13 at LAPR1
LAPR2='(12-13]' 178 ==> LAPR1='(12-13]' 71 conf:(0.4)	Scores between 12 and 13 at LAPR2 usually indicates scores between 12 and 13 at LAPR1
LAPR2='(12-13]' 178 ==> LAPR4='(14-15]' 69 conf:(0.39)	Scores between 12 and 13 at LAPR2 usually indicates scores between 14-15 at LAPR3

9.1.3. STRUCTURE AND PLANNING SUBJECTS

TABLE 103 A.R EXPERIMENT – STRUCTURE AND PLANNING SUBJECTS SYNERGY

Support	**Confidence**
0.4	0.5
Result	**Description**
BDDAD='(10-11]' 259 ==> ESOFT='(10-11]' 193 conf:(0.75)	Scores between 10-11 at BDDAD usually related to scores between 10-11 at ESOFT
ESOFT='(10-11]' 316 ==> BDDAD='(10-11]' 193 conf:(0.61)	Scores between 10-11 at ESOFT usually related to scores between 10-11 at BDDAD

TABLE 104 A.R EXPERIMENT – STRUCTURE AND PLANNING SUBJECTS SYNERGY

Support	**Confidence**
0.3	0.4
Result	**Description**
BDDAD='(10-11]' 259 ==> ESOFT='(10-11]' 193 conf:(0.75)	Scores between 10-11 at BDDAD usually related to scores between 10-11 at ESOFT
LAPR3='(12-13]' 217 ==> ESOFT='(10-11]' 150 conf:(0.69)	Scores between 12 and 13 at LAPR3 usually related to scores between 10-11 at ESOFT
PESTI='(16-17]' 224 ==> ESOFT='(10-11]' 152 conf:(0.68)	Scores between 16-17 at PESTI usually related to scores between 10-11 at ESOFT

ESOFT='(10-11]' 316 ==> BDDAD='(10-11]' 193 conf:(0.61)	Scores between 10-11 at ESOFT usually related to scores between 10-11 at BDDAD
ESOFT='(10-11]' 316 ==> PESTI='(16-17]' 152 conf:(0.48)	Scores between 10-11 at ESOFT usually related to scores between 16-17 at PESTI
ESOFT='(10-11]' 316 ==> LAPR3='(12-13]' 150 conf:(0.47)	Scores between 10-11 at ESOFT usually related to scores between 12 and 13 at LAPR3

TABLE 105 A.R EXPERIMENT – STRUCTURE AND PLANNING SUBJECTS RESULTS SYNERGY

Support	**Confidence**
0.25	0.4
Result	**Description**
BDDAD='(10-11]' 259 ==> ESOFT='(10-11]' 193 conf:(0.75)	Scores between 10-11 at BDDAD usually related to scores between 10-11 at ESOFT
EAPLI='(10-11]' 162 ==> ESOFT='(10-11]' 120 conf:(0.74)	Scores between 10-11 at EAPLI usually related to scores between 10-11 at ESOFT
LAPR4='(14-15]' 159 ==> ESOFT='(10-11]' 115 conf:(0.72)	Scores between 14-15 at LAPR4 usually related to scores between 10-11 at ESOFT
EAPLI='(12-13]' 183 ==> ESOFT='(10-11]' 132 conf:(0.72)	Scores between 12 and 13 at EAPLI usually related to scores between 10-11 at ESOFT
LAPR3='(12-13]' 217 ==> ESOFT='(10-11]' 150 conf:(0.69)	Scores between 12 and 13 at LAPR3 usually related to scores between 10-11 at ESOFT
PESTI='(16-17]' 224 ==> ESOFT='(10-11]' 152 conf:(0.68)	Scores between 16-17 at PESTI usually related to scores between 10-11 at ESOFT
ESOFT='(10-11]' 316 ==> BDDAD='(10-11]' 193 conf:(0.61)	Scores between 10-11 at ESOFT usually related to scores between 10-11 at BDDAD
LAPR3='(12-13]' 217 ==> BDDAD='(10-11]' 132 conf:(0.61)	Scores between 12 and 13 at LAPR3 usually related to scores between 10-11 at BDDAD
PESTI='(16-17]' 224 ==> BDDAD='(10-11]' 126 conf:(0.56)	Scores between 16-17 at PESTI usually related to scores between 10-11 at BDDAD
LAPR3='(12-13]' 217 ==> PESTI='(16-17]' 117 conf:(0.54)	Scores between 12 and 13 at LAPR3 usually related to scores between 16-17 at PESTI

PESTI='(16-17]' 224 ==> LAPR3='(12-13]' 117 conf:(0.52)	Scores between 16-17 at PESTI usually related to scores between 12 and 13 at LAPR3
BDDAD='(10-11]' 259 ==> LAPR3='(12-13]' 132 conf:(0.51)	Scores between 10-11 at BDDAD usually related to scores between 12 and 13 at LAPR3
BDDAD='(10-11]' 259 ==> PESTI='(16-17]' 126 conf:(0.49)	Scores between 10-11 at BDDAD usually related to scores between 16-17 at PESTI
ESOFT='(10-11]' 316 ==> PESTI='(16-17]' 152 conf:(0.48)	Scores between 10-11 at ESOFT usually related to scores between 16-17 at PESTI
ESOFT='(10-11]' 316 ==> LAPR3='(12-13]' 150 conf:(0.47)	Scores between 10-11 at ESOFT usually related to scores between 12 and 13 at LAPR3
ESOFT='(10-11]' 316 ==> EAPLI='(12-13]' 132 conf:(0.42)	Scores between 10-11 at ESOFT usually related to scores between 12 and 13 at EAPLI

9.1.4. PROGRAMMING SUBJECTS

TABLE 106 A.R EXPERIMENT – PROGRAMMING SUBJECTS SYNERGY

Support	**Confidence**
0.2	0.5
Result	**Description**
SCOMP='(10-11]' 176 ==> LPROG='(10-11]' 100 conf:(0.57)	Scores between 10-11 at SCOMP usually related to scores between 10-11 at LPROG

LPROG='(10-11]' 182 ==> SCOMP='(10-11]' 100 conf:(0.55)	Scores between 10-11 at LPROG usually related to scores between 10-11 at SCOMP

TABLE 107 A.R EXPERIMENT – PROGRAMMING SUBJECTS SYNERGY

Support	Confidence
0.15	0.5
Result	**Description**
PPROG='(10-11]' 130 ==> LPROG='(10-11]' 78 conf:(0.6)	Scores between 10-11 at PPROG usually related to scores between 10-11 at LPROG
ARQCP='(10-11]' 151 ==> SCOMP='(10-11]' 87 conf:(0.58)	Scores between 10-11 at ARQCP usually related to scores between 10-11 at SCOMP
ESINF='(10-11]' 139 ==> LPROG='(10-11]' 79 conf:(0.57)	Scores between 10-11 at ESINF usually related to scores between 10-11 at LPROG
ESINF='(10-11]' 139 ==> SCOMP='(10-11]' 79 conf:(0.57)	Scores between 10-11 at ESINF usually related to scores between 10-11 at SCOMP
SCOMP='(10-11]' 176 ==> LPROG='(10-11]' 100 conf:(0.57)	Scores between 10-11 at SCOMP usually related to scores between 10-11 at LPROG
ARQCP='(10-11]' 151 ==> LPROG='(10-11]' 85 conf:(0.56)	Scores between 10-11 at ARQCP usually related to scores between 10-11 at LPROG
LPROG='(10-11]' 182 ==> SCOMP='(10-11]' 100 conf:(0.55)	Scores between 10-11 at LPROG usually related to scores between 10-11 at SCOMP
COMPA='(10-11]' 169 ==> LPROG='(10-11]' 90 conf:(0.53)	Scores between 10-11 at COMPA usually related to scores between 10-11 at LPROG
PPROG='(12-13]' 131 ==> SCOMP='(10-11]' 69 conf:(0.53)	Scores between 12 and 13 at PPROG usually related to scores between 10-11 at SCOMP
COMPA='(10-11]' 169 ==> SCOMP='(10-11]' 89 conf:(0.53)	Scores between 10-11 at COMPA usually related to scores between 10-11 at SCOMP
APROG='(10-11]' 137 ==> LPROG='(10-11]' 70 conf:(0.51)	Scores between 10-11 at APROG usually related to scores between 10-11 at LPROG

COMPA='(10-11]' 169 ==> SGRAI='(10-11]' 86 conf:(0.51)	Scores between 10-11 at COMPA usually related to scores between 10-11 at SGRAI
SCOMP='(10-11]' 176 ==> COMPA='(10-11]' 89 conf:(0.51)	Scores between 10-11 at SCOMP usually related to scores between 10-11 at COMPA
ESINF='(10-11]' 139 ==> COMPA='(10-11]' 70 conf:(0.5)	Scores between 10-11 at ESINF usually related to scores between 10-11 at COMPA

9.2. INDIVIDUAL SUBJECTS SCENARIO EXPERIMENTS

9.2.1. BDDAD

TABLE 108 A.R EXPERIMENT – BDDAD PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Result=Aprov 195 ==>Enrollment Year=11-12 150 conf:(0.77)	77% of approved students enrolled between 2011 and 2012
Season=Appeal 166 ==>Enrollment Year=11-12 109 conf:(0.66)	66% of students who got their grades in appeal season enrolled between 2011 and 2012
Res=Failed 165 ==> Frequency Grade=NR 105 conf:(0.64)	64% of failed students don't have frequency grades
Result=Aprov 195 ==> Season=Normal 123 conf:(0.63)	63% of approved students got their results in normal season
Frequency_Grade=NR 218 ==>Enrollment Year=11-12 137 conf:(0.63)	63% of students who didn't get a frequency grades enrolled between 2011 and 2012
Enrollment Year=11-12 247 ==> Res=Aprov. 150 conf:(0.61)	61% of students who enrolled between 2011 and 2012 were approved
Res=Failed 165 ==> Enrollment Year=11-12 97 conf:(0.59)	59% of failed students enrolled between 2011 and 2012

Res=Failed 165 ==> Season=Appeal 94 conf:(0.57)	57% of failed students got their final results in appeal season
Season=Appeal 166 ==> Res=Failed 94 conf:(0.57)	57% of students who got their results in appeal season, failed
Enrollment Year=11-12 247 ==> Season=Normal 138 conf:(0.56)	56% of students who enrolled between 2011 and 2012 got their results in normal season
Frequency Grade=NR 192 ==> Res=Failed 105 conf:(0.55)	55% of students without frequency grades failed
Enrollment Year=11-12 247 ==> Frequency Grade=NR 123 conf:(0.5)	50% of students who enrolled between 2011 and 2012, didn't have frequency grades
Res=Aprov. 195 ==> Final Grade=11-12 87 conf:(0.45)	45% of approvals were obtained with grades between 11 and 12
Res=Aprov. 195 ==> Frequency Grade=NR 87 conf:(0.45)	45% of approved students don't have frequency grades
Frequency Grade=NR 192 ==> Res=Aprov. 87 conf:(0.45)	45% of students without frequency grade were approved
Enrollment Year=11-12 247 ==> Season=Appeal 109 conf:(0.44)	44% of students who enrolled between 2011 and 2012, got their final grades in appeal season

9.2.2. COMPA

TABLE 109 A.R EXPERIMENT – COMPA PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Enrollment Year=11-12 81 ==> Res=Aprov. 66 conf:(0.81)	81% of students who enrolled between 2011 and 2012, were approved

Enrollment Year=9-10 83 ==> Frequency Grade=NR Season=Appeal 63 conf:(0.76)	77% of students who enrolled between 2009 and 2010, didn't have frequency grade and go their results in appeal season
Season=Normal 101 ==> Res=Aprov. 67 conf:(0.66)	66% of students who got their final results in normal season, were approved
Res=Failed 98 ==> Season=Appeal 64 conf:(0.65)	65% of failed students got their final results in appeal season
Res=Aprov. 134 ==> Final Grade=10 73 conf:(0.54)	54% of approvals were obtained with minimum grade
Season=Appeal 131 ==> Res=Aprov. 67 conf:(0.51)	51% of appeal season students were approved
Res=Aprov. 134 ==> Frequency Grade=NR 68 conf:(0.51)	51% of approved students didn't have frequency grades
Res=Aprov. 134 ==> Season=Normal 67 conf:(0.5)	50% of approved students got their results in normal season
Frequency Grade=NR 137 ==> Res=Aprov. 68 conf:(0.5)	50% of students without frequency grades, were approved
Res=Aprov. 134 ==> Enrollment Year=11-12 66 conf:(0.49)	49% of approved students enrolled between 2011 and 2012

9.2.3. IARTI

TABLE 110 A.R EXPERIMENT – IARTI PERFORMANCE

Support	**Confidence**
0.2	0.4
Result	**Description**
Enrollment Year=11-12 86 ==> Res=Aprov. 67 conf:(0.78)	78% of students who enrolled between 2011 and 2012 were approved
Enrollment Year=9-10 96 ==> Frequency Grade=NR 74 conf:(0.77)	77% of students who enrolled between 2009 and 2010, didn't have frequency grades

Res=Failed 106 ==> Frequency Grade=NR 82 conf:(0.77)	77% of failed students didn't have frequency grades
Season=Normal 134 ==> Res=Aprov. 86 conf:(0.64)	64% of students who got their results in normal season, were approved
Result=Aprov 158 ==>Frequency_Grade=NR 94 conf:(0.59)	59% of students who were approved, didn't have a frequency grade
Frequency_Grade=NR 214 ==> Result=Failed 120 conf:(0.56)	56% of students who didn't have a frequency grade, failed
Enrollment Year=9-10 96 ==> Res=Aprov. 53 conf:(0.55)	55% of students who enrolled between 2009 and 2010 were approved
Res=Failed 106 ==> Season=Appeal 58 conf:(0.55)	55% of failed students got their results in appeal season
Result=Aprov 158 ==> Season=Normal 86 conf:(0.54)	54% of approved students, got their results in normal season
Frequency Grade=NR 176 ==> Res=Aprov. 94 conf:(0.53)	53% of students without frequency grades were approved
Season=Appeal 130 ==> Res=Failed 58 conf:(0.45)	45% of students who got their results in appeal season, failed
Res=Aprov. 158 ==> Final Grade=11-12 68 conf:(0.43)	43% of approved students had final grades between 11 and 12
Frequency Grade=NR 176 ==> Enrollment Year=9-10 74 conf:(0.42)	42% of students without frequency grade enrolled between 2009 and 2010
Res=Aprov. 158 ==> Exam Grade=NR 64 conf:(0.41)	41% of approved students didn't complete the exam

9.2.4. LPROG

TABLE 111 A.R EXPERIMENT – LPROG PERFORMANCE

Support	**Confidence**
0.2	0.4

Result	Description
Result=Aprov 178 ==>Enrollment Year=11-12 141 conf:(0.79)	79% of approved students enrolled between 2011 and 2012
Result=Aprov 178 ==> Season=Normal 134 conf:(0.75)	75% of approved students got their results in normal season
Enrollment Year=11-12 224 ==> Season=Normal 147 conf:(0.66)	66% of students who enrolled between 2011 and 2012 got their final results in normal season
Enrollment Year=11-12 224 ==> Res=Aprov. 141 conf:(0.63)	63% of students who enrolled between 2011 and 2012 were approved
Res=Failed 160 ==> Season=Normal 86 conf:(0.54)	54% of failed students got their final results in normal season
Res=Failed 160 ==> Frequency Grade=NR 76 conf:(0.48)	48% of failed students didn't have frequency grades
Res=Aprov. 178 ==> Exam Grade=9-10 82 conf:(0.46)	46% of approved students had grades between 9 and 10 in the exam
Res=Aprov. 178 ==> Final Grade=11-12 79 conf:(0.44)	44% of approved students had final grades between 11 and 12
Res=Aprov. 178 ==> Final Grade Difference=-1-2 72 conf:(0.4)	40% of approved students dropped their grades between 1 and 2 in the exam

9.2.5. FSIAP

TABLE 112 A.R EXPERIMENT – FSIAP PERFORMANCE

Support	Confidence
0.2	0.4
Result	**Description**
Res=Aprov 173 ==> Season=Normal 147 conf:(0.85)	85% of approvals were obtained in normal season

Enrollment Year=11-12 Res=Aprov 108 ==> Exam Grade=NR 90 conf:(0.83)	83% of students who enrolled between 2011 and 2012 and were approved, didn't complete the exam
Frequency Grade=NR 167 ==> Res=Failed 136 conf:(0.81)	81% of students without frequency grade failed
Res=Aprov 173 ==> Exam Grade=NR 119 conf:(0.69)	69% of approved students didn't do the exam
Enrollment Year=11-12 236 ==> Exam Grade=NR 148 conf:(0.63)	63% of students who enrolled between 2011/212 didn't complete the exam
Frequency Grade=NR 167 ==> Class=ST 104 conf:(0.62)	62% of students without frequency grades had no class
Enrollment Year=11-12 236 ==> Res=Failed 128 conf:(0.54)	54% of students who enrolled between 2012/2012 failed
Res=Failed 266 ==> Frequency Grade=NR 136 conf:(0.51)	51% of failed students didn't have frequency grade
Exam Grade=NR 231 ==> Res=Failed 112 conf:(0.48)	48% of students who didn't complete the exam, failed
Enrollment Year=11-12 236 ==> Res=Aprov 108 conf:(0.46)	46% of students who enrolled between 2011 and 2012 were approved
Exam=NR 276 ==> Result=Aprov 120 conf:(0.43)	43% of students who didn't do exam were approved
Res=Failed 266 ==> Exam Grade=NR 112 conf:(0.42)	42% of failed students didn't complete the exam

9.3. CLASS DIAGRAM

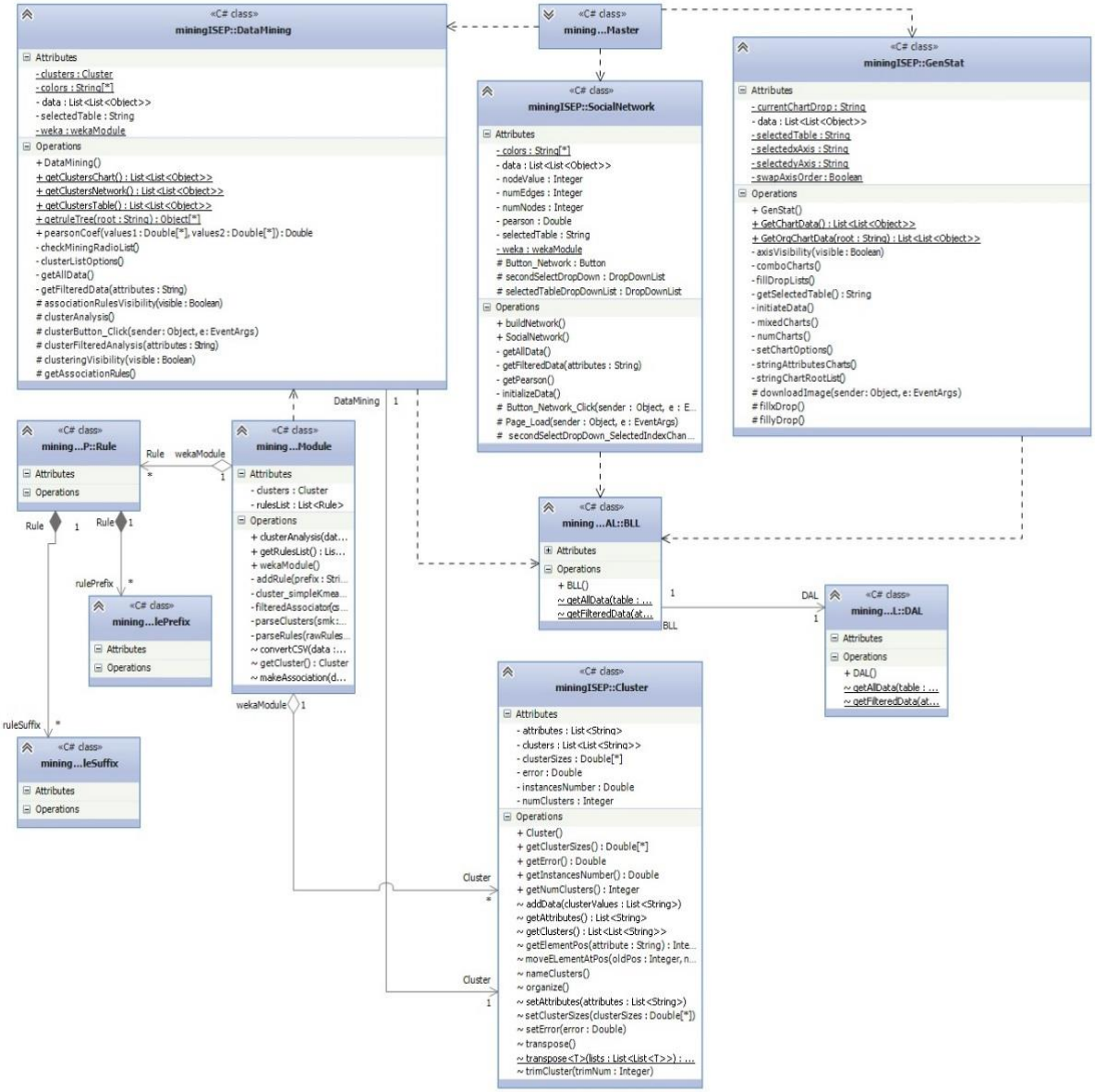

FIGURE 108 - CLASS DIAGRAM

9.4. GOOGLECHARTS AVAILABLE REPRESENTATION

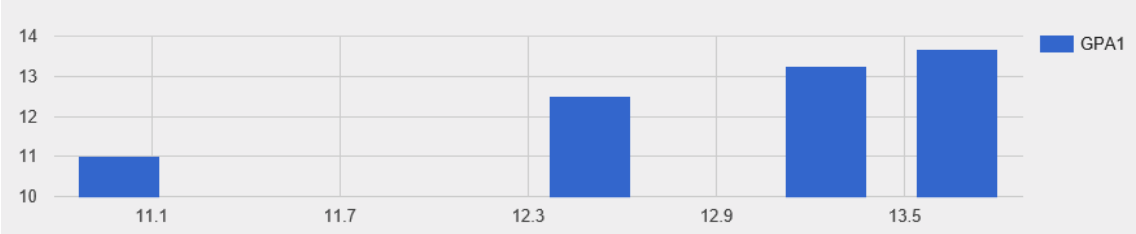

FIGURE 109 – NUMERICAL ATTRIBUTES COLUMN CHART

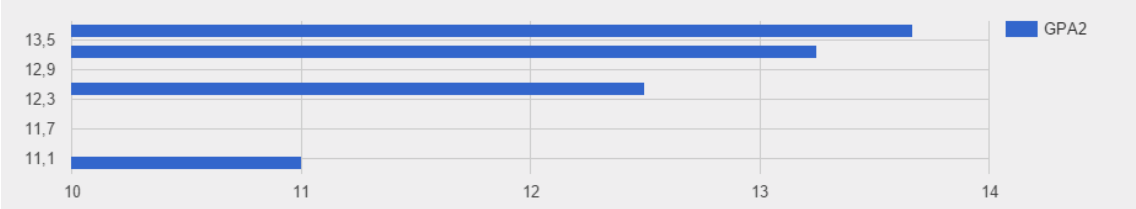

FIGURE 110 - NUMERICAL ATTRIBUTES BAR CHART

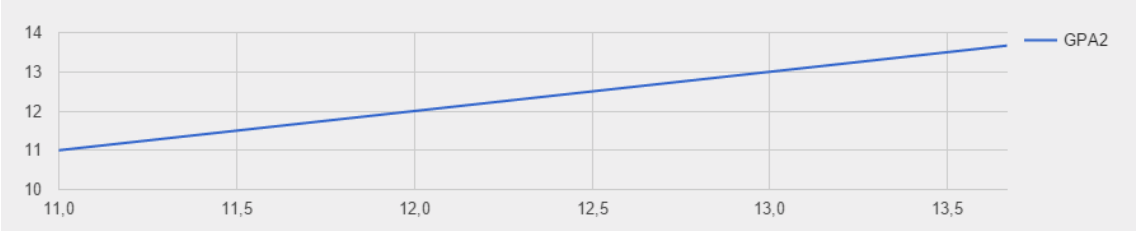

FIGURE 111 – NUMERICAL ATTRIBUTES LINE CHART

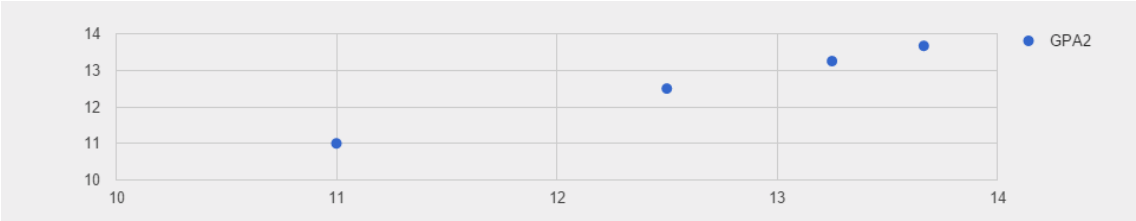

FIGURE 112 - SCATTER CHART

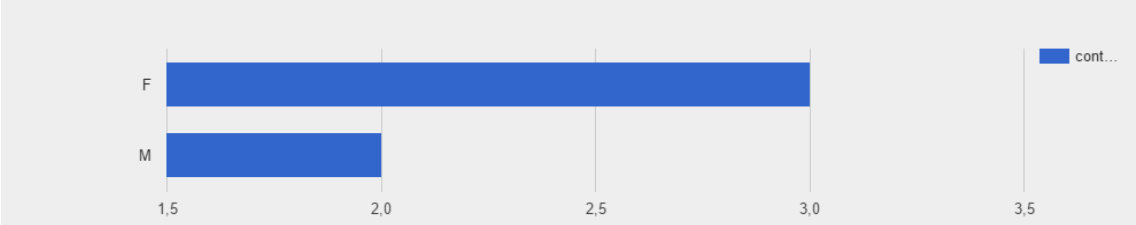

FIGURE 113 - BAR CHART WITH STRING ATTRIBUTE

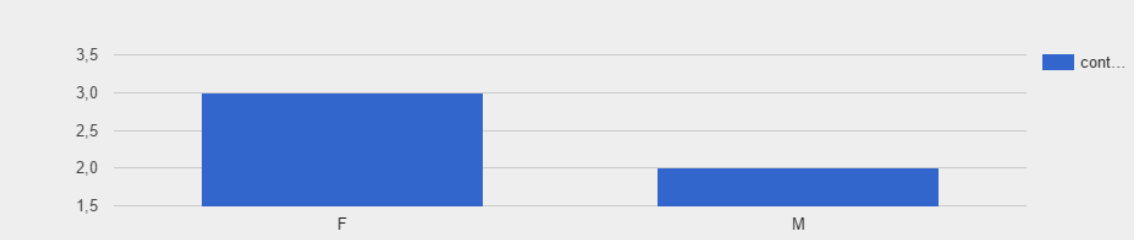

FIGURE 114 - COLUMN CHART WITH STRING ATTRIBUTE

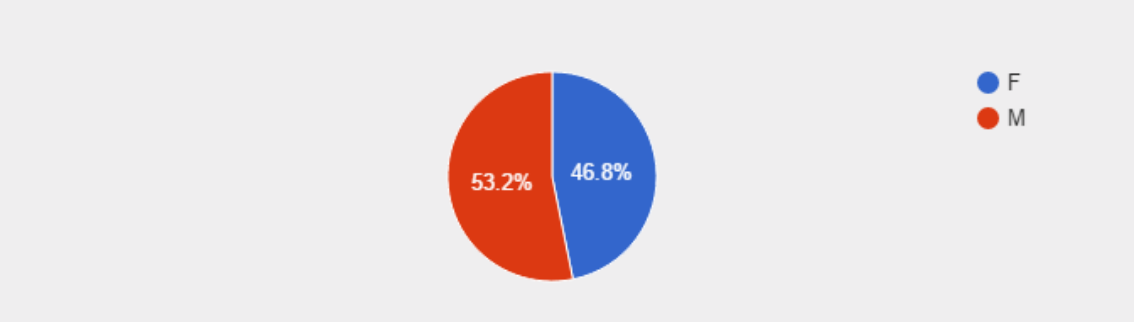

FIGURE 115 - PIE CHART

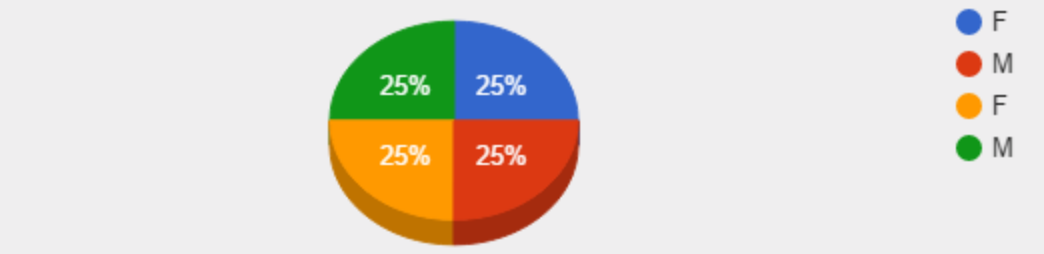

FIGURE 116 – 3D PIE CHART

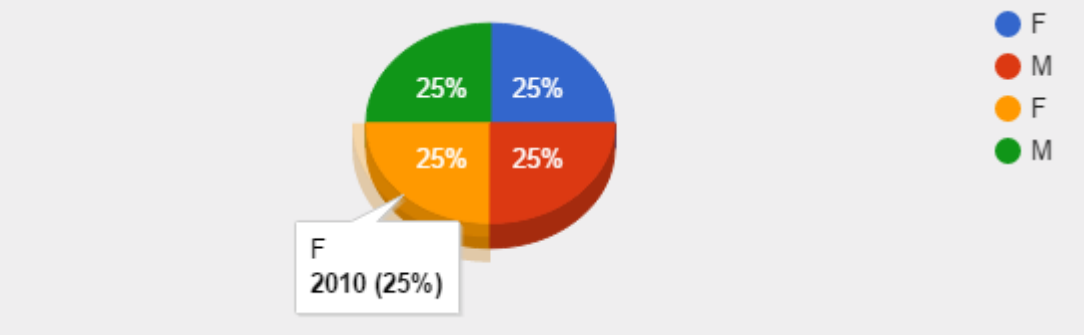

FIGURE 117 - 3D PIE CHART HIGHLIGHTED

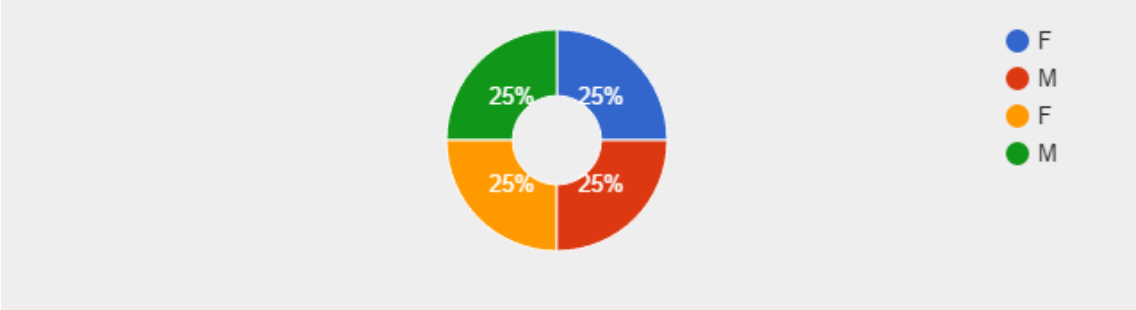

FIGURE 118 - DONUT CHART

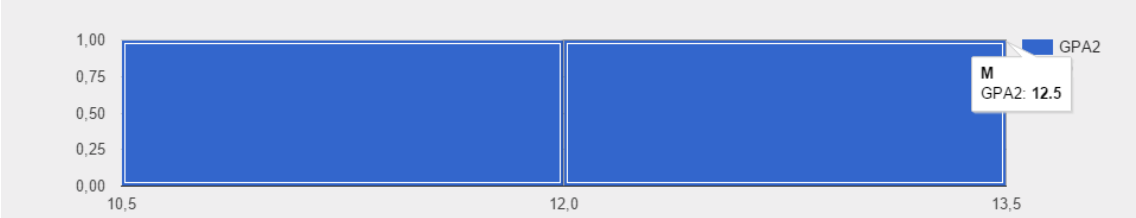

FIGURE 119 – HISTOGRAM

FIGURE 120 – LINE CHART WITH STRING ATTRIBUTE

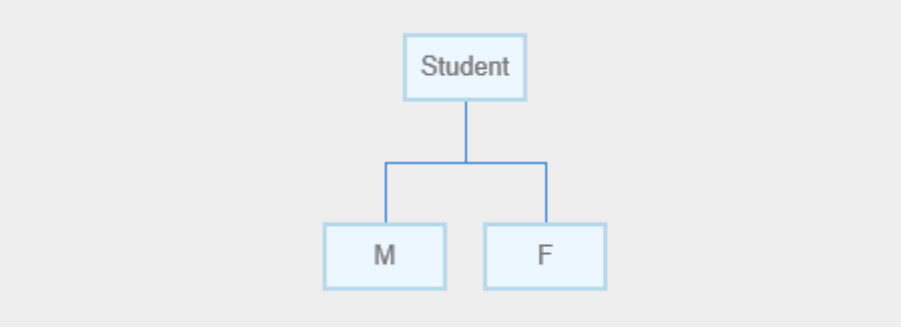

FIGURE 121 - ORG CHART

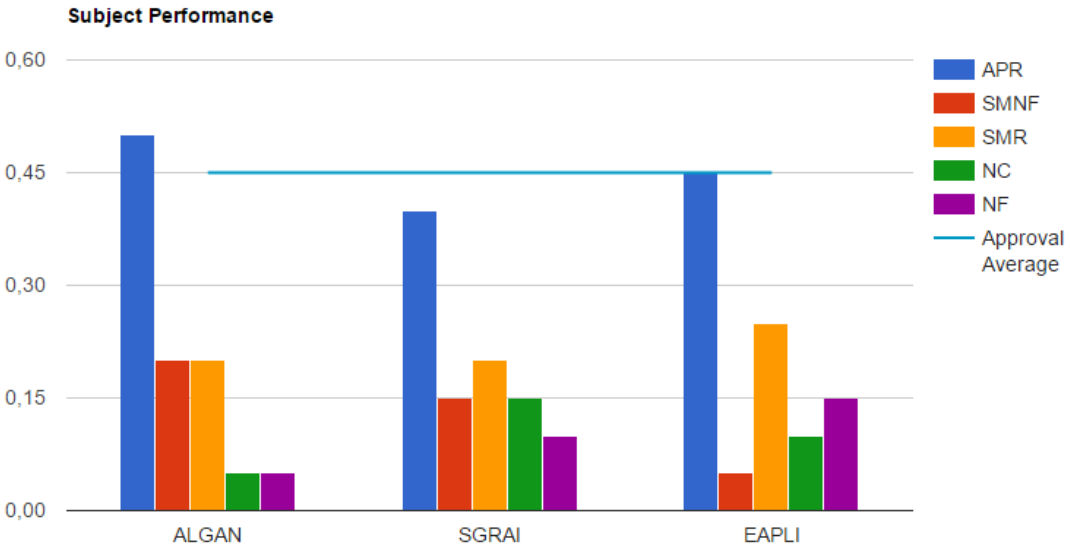

FIGURE 122 - COMBO CHART

9.5. SOCIAL NETWORK SEQUENCE DIAGRAM

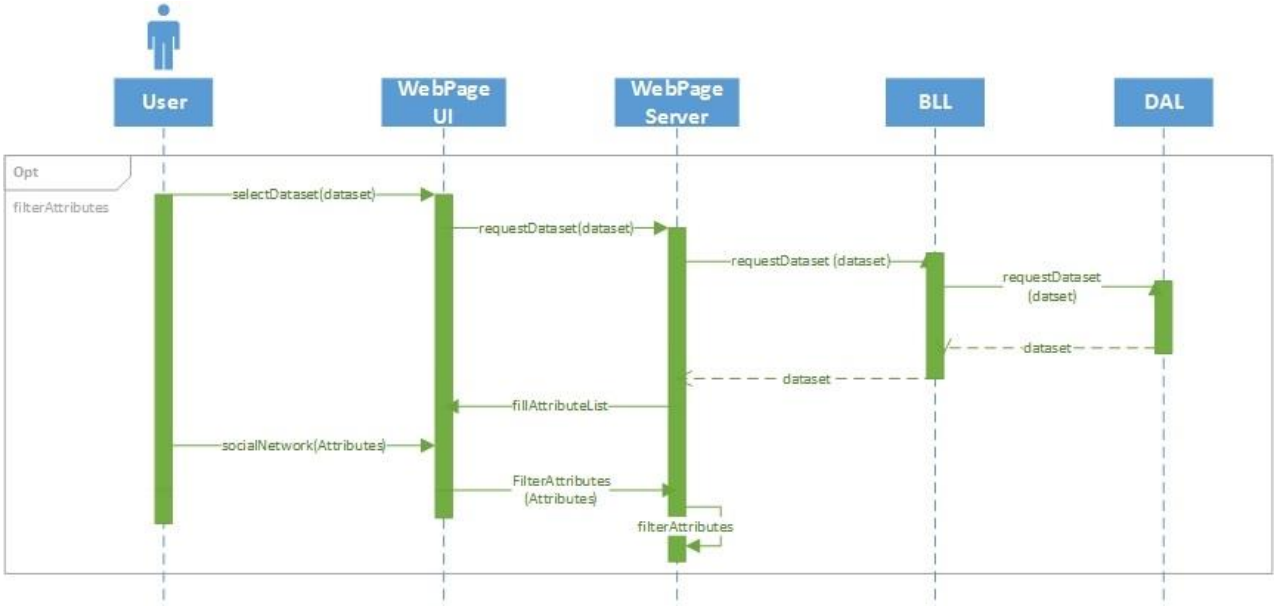

FIGURE 123 - SOCIAL NETWORK FILTER ATTRIBUTES

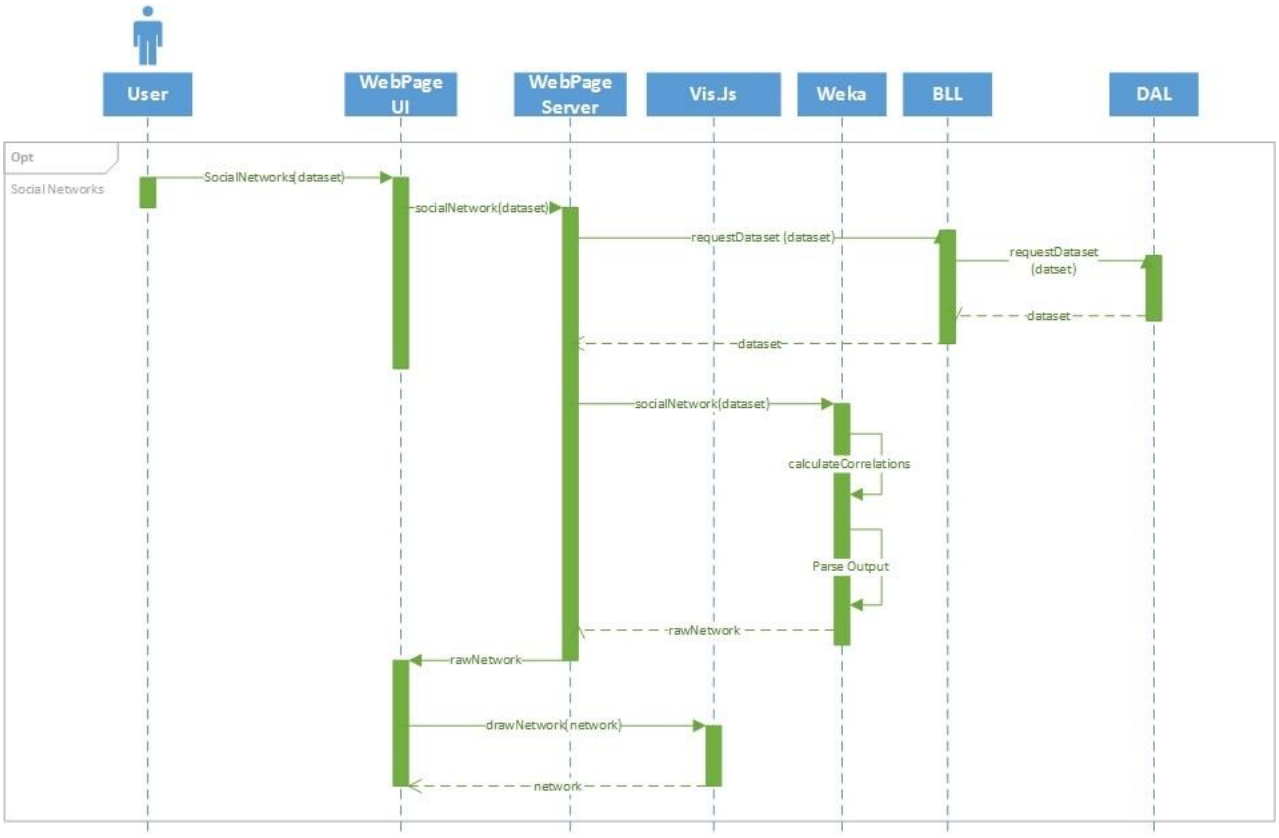

FIGURE 124 – DRAWING SOCIAL NETWORKS

9.6. HTLM2CANVAS SEQUENCE DIAGRAM

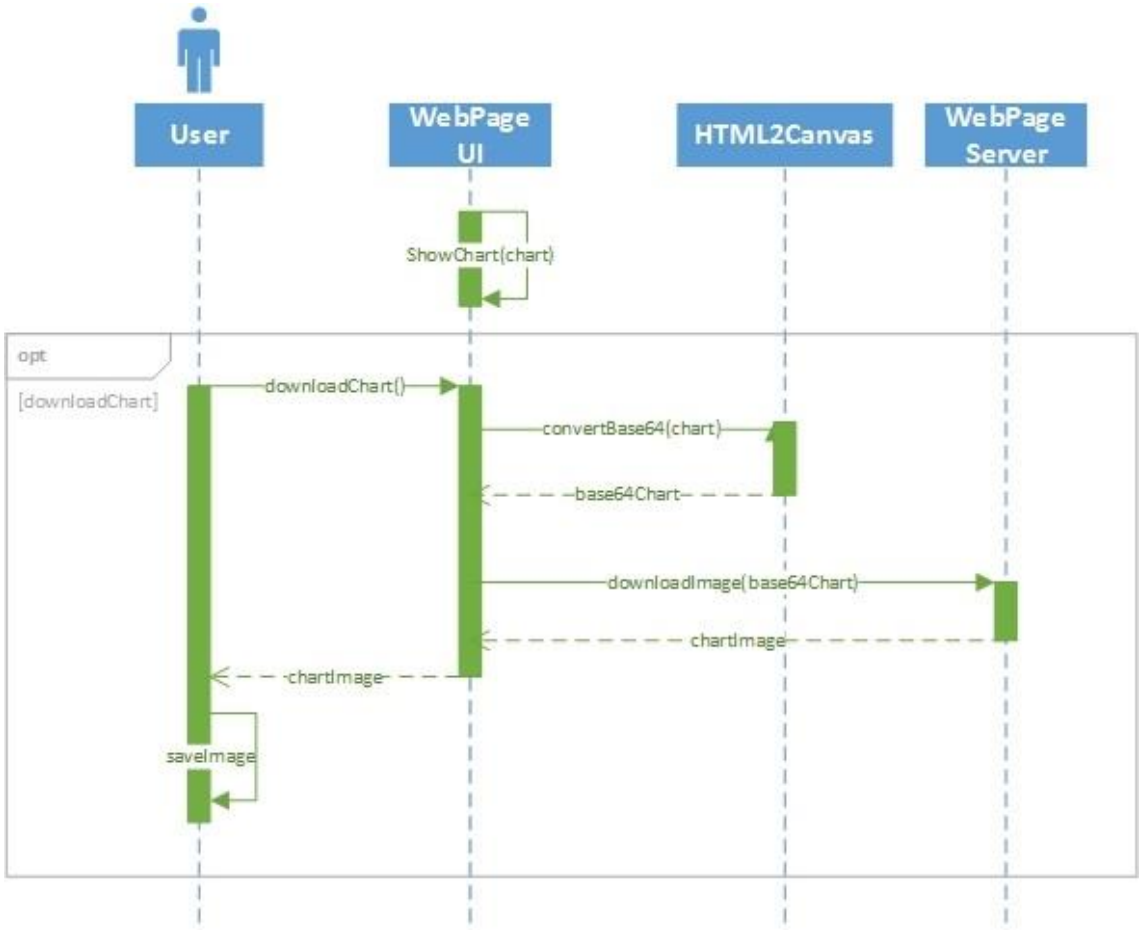

FIGURE 125 - DOWNLOAD CHART

9.7. CLUSTERING DUMMY DATA SET

@relation car-browsers

@attribute Dealership numeric
@attribute Showroom numeric
@attribute ComputerSearch numeric
@attribute M5 numeric
@attribute 3Series numeric
@attribute Z4 numeric
@attribute Financing numeric
@attribute Purchase numeric

@data

1,0,0,0,0,0,0,0
1,1,1,0,0,0,1,0
1,0,0,0,0,0,0,0
1,1,1,1,0,0,1,1
1,0,1,1,1,0,1,1
1,1,1,0,1,0,0,0
1,0,1,0,0,0,1,1
1,0,1,0,1,0,0,0
1,1,1,0,1,0,1,0
1,0,1,1,1,1,1,1
1,0,1,1,1,1,1,0
1,0,1,1,0,1,0,0
1,0,1,1,0,0,1,1
1,1,1,0,0,1,1,0
1,0,1,1,1,1,0,0
1,1,1,1,1,0,1,1
1,0,1,0,0,0,1,1
1,0,1,0,0,0,0,1,0
1,1,0,1,1,0,0,0
1,0,0,1,0,0,0,0
1,1,1,0,0,1,1,1
1,0,1,0,0,1,1,1
1,1,0,0,0,0,1,0
1,1,0,1,0,0,0,0
1,1,0,1,1,1,1,0
1,1,1,0,1,1,0,0
1,1,0,1,1,1,1,0
1,1,0,1,1,0,1,0
1,1,1,1,1,1,0,0
1,1,0,1,1,1,1,0
1,1,0,1,0,1,1,0
1,1,0,1,0,1,0,0
1,1,0,0,1,0,1,1
1,1,1,0,1,1,0,0
1,1,0,1,0,1,1,0
1,1,1,1,1,1,0,0
1,1,0,1,1,0,1,1
0,1,1,0,1,0,1,0
0,1,1,0,1,0,1,0
0,1,0,0,1,0,0,0
0,1,0,0,1,0,0,0
0,1,0,0,1,1,0,0
0,1,0,0,1,0,1,1
0,1,0,0,1,0,0,0
0,1,1,0,1,1,0,0
0,1,1,0,1,0,1,0
0,1,1,0,1,1,1,1

0,0,1,1,0,0,1,1
1,0,1,1,0,0,0,0
1,0,1,1,0,1,1,0
0,1,1,1,0,1,1,1
0,1,1,1,0,1,1,0
1,1,0,1,0,1,1,1
1,1,0,1,0,0,0,0
0,1,0,1,1,0,0,0
0,0,0,1,1,0,1,1
0,0,0,1,1,0,0,0
1,0,1,0,0,0,0,0
1,1,0,1,0,0,1,1
1,1,0,1,1,0,1,1
1,1,1,1,0,0,1,0
1,1,0,1,0,0,0,0
1,1,0,1,1,0,1,1
1,1,1,1,0,1,1,1
1,1,0,1,0,0,1,0
1,0,1,0,0,0,0,0
1,1,0,1,0,0,0,0
1,1,0,1,0,0,1,1
1,1,1,0,0,1,0,0
1,1,0,1,0,0,1,1
1,1,1,1,0,1,1,1
1,1,0,1,1,0,1,1
0,0,0,1,0,0,1,1
1,0,0,1,0,0,0,0
1,0,0,1,0,1,1,1
0,1,0,1,0,0,1,1
0,0,0,1,0,0,1,0
1,1,1,1,0,1,1,1
1,1,0,1,0,0,1,1
0,0,0,1,1,0,0,0
0,0,0,1,0,0,1,1
0,0,0,1,1,0,1,1
0,1,1,0,1,0,0,0
0,1,0,0,1,1,1,1
0,1,0,0,1,1,0,0
0,1,0,0,1,1,0,0
0,1,0,0,1,1,0,0
0,1,0,0,1,1,1,1
0,1,0,0,1,1,0,0
0,1,1,0,1,1,0,0
0,1,0,0,1,1,1,1
0,1,0,0,1,1,1,1
0,1,1,0,1,0,0,0
0,1,0,0,1,1,1,1
0,1,0,0,1,1,1,0
0,1,0,0,1,1,0,0
0,1,0,0,1,1,1,1

0,1,0,0,1,1,1,1
0,1,0,0,1,1,0,0
0,1,1,0,1,1,1,0

9.8. CLUSTERING PERCENTAGE SPLIT OUTPUT

TABLE 113 - CLUSTERING PERCENTAGE SPLIT OUTPUT ATTACHMENT

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last"
-I 500 -S 4
Relation:     car-browsers
Instances:    100
Attributes:   8
              Dealership
```

```
               Showroom
               ComputerSearch
               M5
               3Series
               Z4
               Financing
               Purchase
Test mode:split 66% train, remainder test

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 155.95000000000013
Missing values globally replaced with mean/mode

Cluster centroids:
                              Cluster#
Attribute         Full Data          0          1
                      (100)       (40)       (60)
=================================================
Dealership             0.6      0.575     0.6167
Showroom              0.72      0.675       0.75
ComputerSearch        0.43      0.375     0.4667
M5                    0.53       0.65       0.45
3Series               0.55      0.475        0.6
Z4                    0.45      0.425     0.4667
Financing             0.61          1       0.35
Purchase              0.39      0.975          0



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on test split ===

kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 94.93658536585369
Missing values globally replaced with mean/mode

Cluster centroids:
                              Cluster#
Attribute         Full Data          0          1
                      (66)       (41)       (25)
=================================================
Dealership          0.5909     0.8293        0.2
Showroom            0.7121     0.5366          1
ComputerSearch      0.4394     0.4146       0.48
M5                  0.5152     0.7805       0.08
3Series             0.5152     0.2439       0.96
Z4                  0.4091     0.2195       0.72
Financing           0.5606     0.6829       0.36
Purchase            0.3636      0.439       0.24
```

```
Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0       23 (68%)
1       11 (32%)
```

With percentage split the sum of squared errors has slightly decreased as the cluster centroids. Important changes are in bold.

9.9. PREVIOUS WORKPLAN

	Task Name	Duration	Start	Finish	Predecessors
1	**Preparation**	**43 days?**	**Wed 21-01-15**	**Fri 20-03-15**	
2	Study of tools and documentatior	43 days?	Wed 21-01-15	Fri 20-03-15	
3	State of the art	12 days	Thu 05-02-15	Fri 20-02-15	
4	Database development	14 days?	Mon 23-03-15	Thu 09-04-15	1
5	Pattern analysis	28 days?	Fri 10-04-15	Tue 19-05-15	4
6	UI Development	28 days?	Wed 20-05-15	Fri 26-06-15	5
7	Integrating Flexidash	14 days?	Mon 29-06-15	Thu 16-07-15	6
8	Final Report	171 days?	Thu 05-02-15	Thu 01-10-15	

FIGURE 126 – FIRST WORKPLAN TASKS

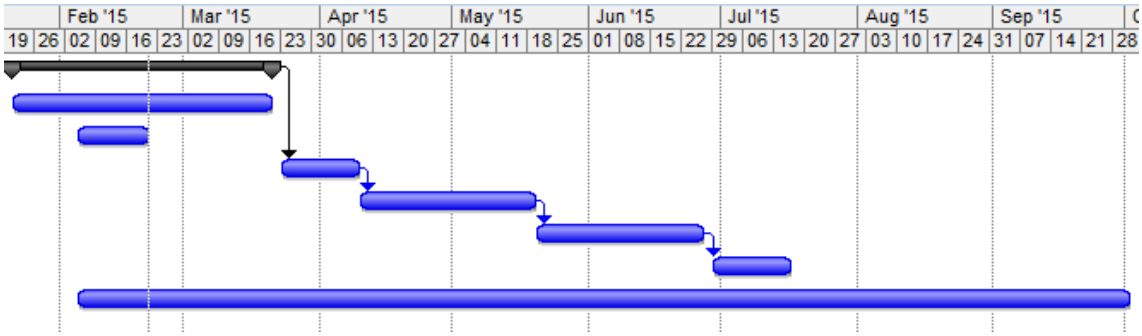

FIGURE 127 – FIRST WORKPLAN TIMELINE

9.10. FINAL WORKPLAN

	Task Name	Duration	Start	Finish	Predecessors
1	**Preparation**	**75 days**	**Mon 19-01-15**	**Fri 01-05-15**	
2	Study of technologies and documentation	45 days	Mon 19-01-15	Fri 20-03-15	
3	Study of available tools	30 days	Mon 23-03-15	Fri 01-05-15	2
4	State of the art	105 days	Thu 07-05-15	Wed 30-09-15	1
5	**Database implementation**	**32 days**	**Mon 04-05-15**	**Tue 16-06-15**	**1**
6	Database design	10 days	Mon 04-05-15	Fri 15-05-15	
7	Data treatment	15 days	Mon 18-05-15	Fri 05-06-15	6
8	Data insertion	7 days	Mon 08-06-15	Tue 16-06-15	7
9	**Data Mining**	**40 days**	**Mon 04-05-15**	**Fri 26-06-15**	**1**
10	Clustering data preprocessing	10 days	Mon 04-05-15	Fri 15-05-15	
11	Cluster Analysis	30 days	Mon 18-05-15	Fri 26-06-15	10
12	Association rules data preprocessing	10 days	Mon 04-05-15	Fri 15-05-15	
13	Association Rules Analysis	30 days	Mon 18-05-15	Fri 26-06-15	12
14	**Application Development**	**84 days**	**Mon 29-06-15**	**Thu 22-10-15**	**5;9**
15	Integrating Database	7 days	Mon 29-06-15	Tue 07-07-15	
16	Integrating Weka	35 days	Wed 08-07-15	Tue 25-08-15	15
17	Integrating GoogleCharts	15 days	Wed 26-08-15	Tue 15-09-15	16
18	Integrating Vis.js	15 days	Wed 16-09-15	Tue 06-10-15	17
19	Integrating HTML2Canvas	2 days	Mon 29-06-15	Tue 30-06-15	
20	Testing	12 days	Wed 07-10-15	Thu 22-10-15	18
21	Final Report	125 days?	Mon 04-05-15	Fri 23-10-15	1

FIGURE 128 - FINAL WORK PLAN TASKS

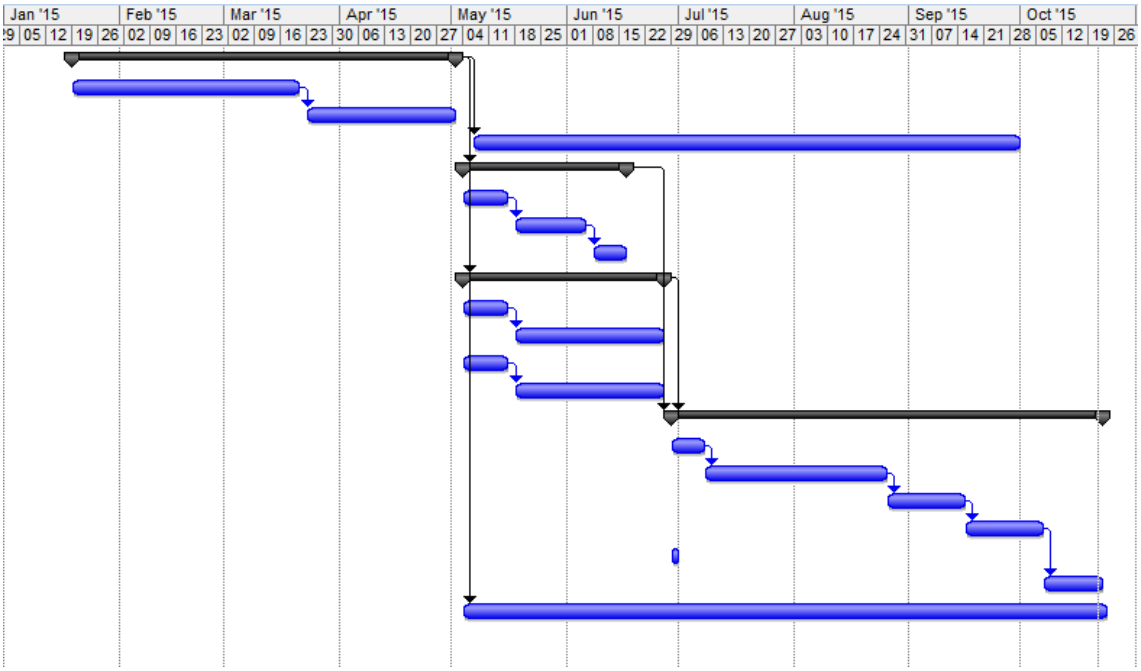

FIGURE 129 - FINAL WORK PLAN TIMELINE

9.11. HOMEPAGE

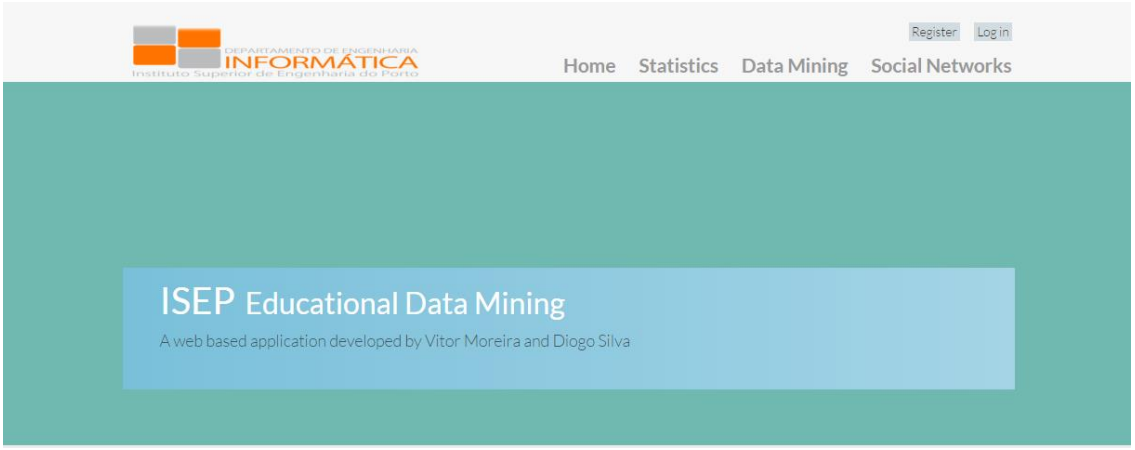

FIGURE 130- HOMEPAGE

9.12. GENERAL STATISTICS

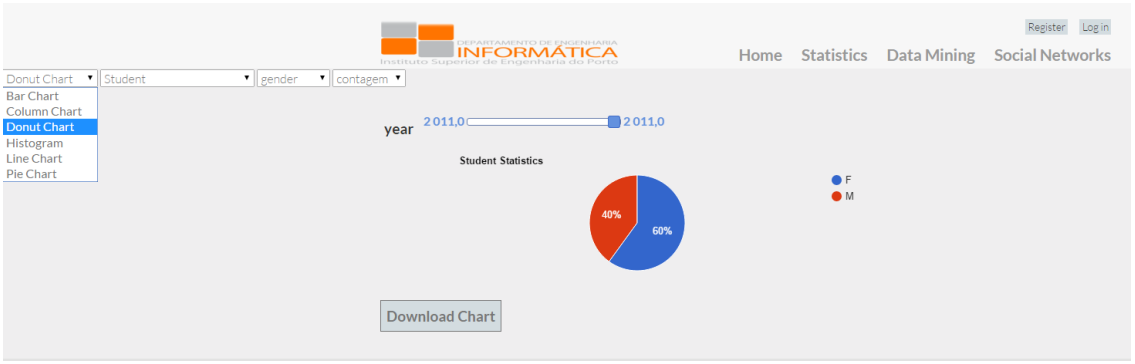

FIGURE 131 - GEN STAT

9.13. Association Rules

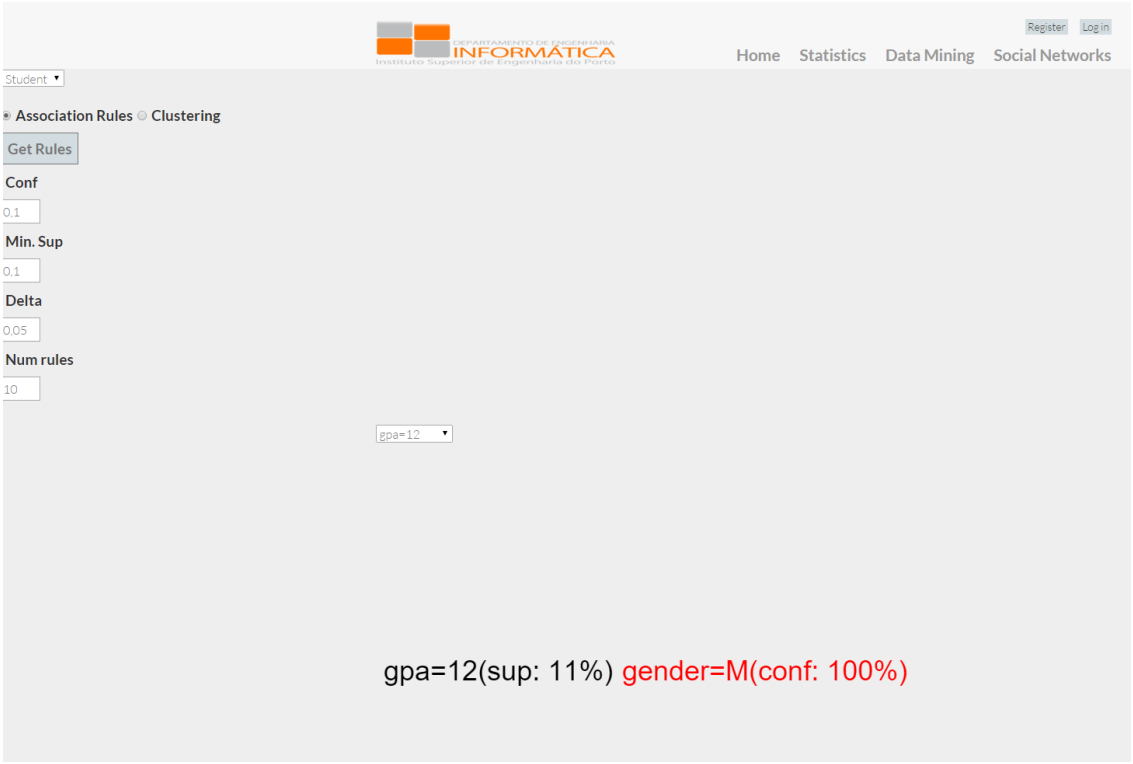

FIGURE 132 - ASSOCIATION RULES

9.14. Clustering

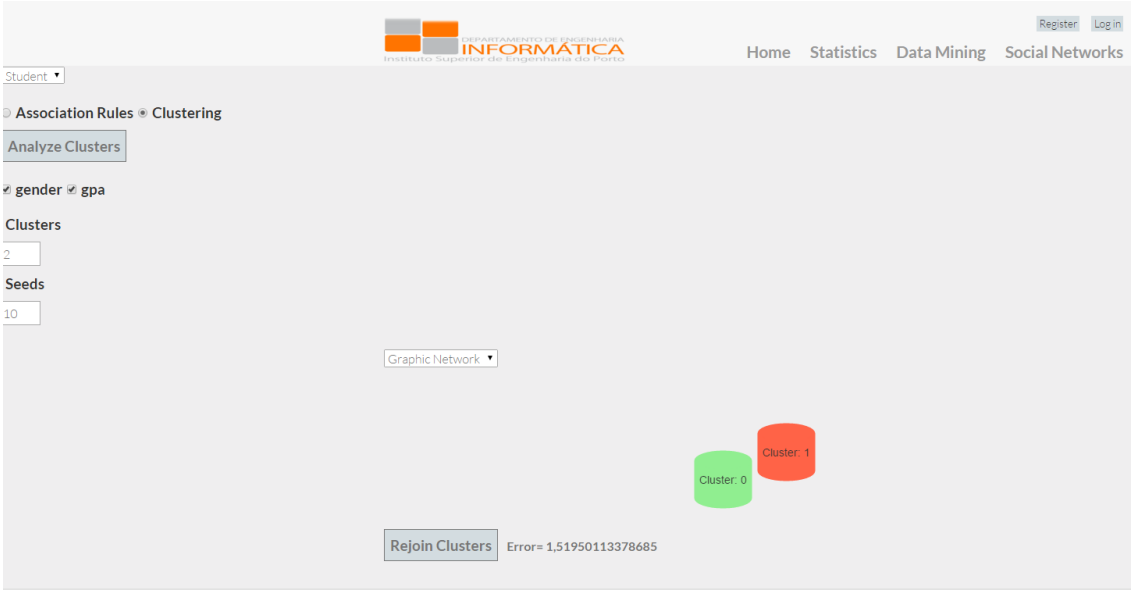

FIGURE 133 - CLUSTERING

9.15. SOCIAL NETWORKS

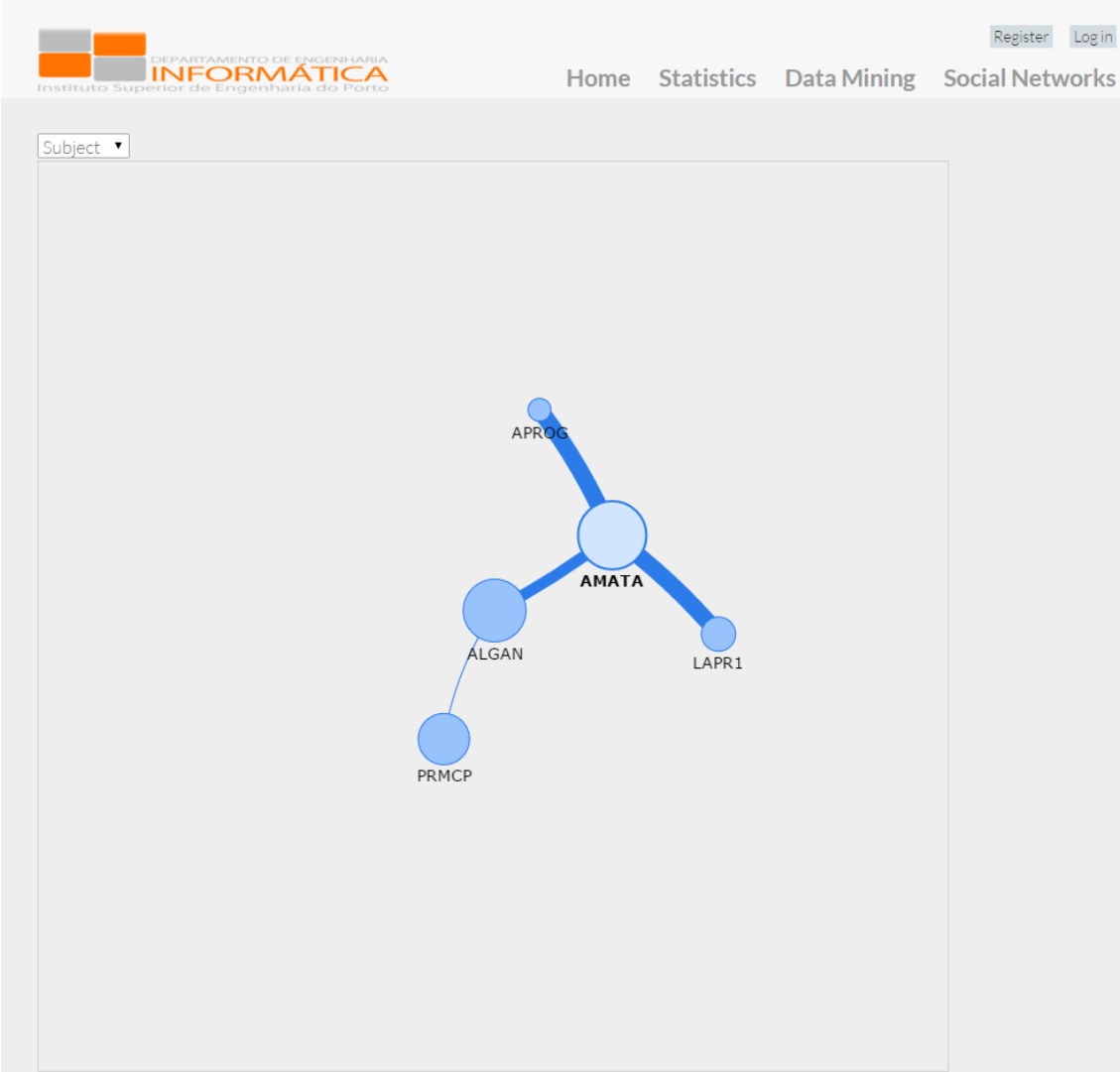

FIGURE 134 - SOCIAL NETWORK