NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Data Mining Application for Healthcare Sector

Predictive Analysis of Heart Attacks

Maria Inês Resende da Lomba Fernandes

Project Work presented as partial requirement for obtaining the Master's degree in Information Management with major in Knowledge Management and Business Intelligence

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Data Mining Application for Healthcare Sector

Predictive Analysis of Heart Attacks

by

Maria Inês Fernandes

Project Work presented as partial requirement for obtaining the Master's degree in Information Management with a specialization in Knowledge Management and Business Intelligence

Advisor: Rui Gonçalves

December 2020

DEDICATION

For my parents, of whom I am so proud, and who have always supported and motivated me.

For uncle Carlos Lomba, aunt Maria José, and grandfather Serafim Lomba, *in memoriam,*
who have always accompanied me and whom I miss so much.

ACKNOWLEDGEMENTS

Developing this final master's project revealed to be a challenging path marked by small and large steps, but mainly, it gave me the feeling of overcoming a goal that culminated on the accomplishment of another important chapter of my life, the master's degree.

Thereby, it is indispensable for me to give my deepest thanks and appreciation for all the people that gave me their unconditional support, namely, the guidance, motivation, patience, comprehension, positivity, and mental energy. Thank you for being there and for accompanying me on this journey, for the fantastic support, and for the strength that has always motivated me to keep going and that contributed for the achievement of this project.

Nevertheless, I would also like to express some special gratitude:

To my parents, who are an example for me and have always given me their support and motivation;

To my family, who although distant, always follows me closely;

To my advisor, Professor Rui Gonçalves, for his support, availability and guidance;

To Daniel Almeida, for his incentive, comprehension and recommendations;

To my friends Filipa Pedro, Mariana Nunes, Margarida Pereira, Joana Aço, and Inês Antunes for their support and for being strong pillars during this process.

My sincere thanks to you all.

ABSTRACT

Cardiovascular diseases are the main cause of the number of deaths in the world, being the heart disease the most killing one affecting more than 75% of individuals living in countries of low and middle earnings. Considering all the consequences, firstly for the individual's health, but also for the health system and the cost of healthcare (for instance, treatments and medication), specifically for cardiovascular diseases treatment, it has become extremely important the provision of quality services by making use of preventive medicine, whose focus is identifying the disease risk, and then, applying the right action in case of early signs. Therefore, by resorting to DM (Data Mining) and its techniques, there is the ability to uncover patterns and relationships amongst the objects in healthcare data, giving the potential to use it more efficiently, and to produce business intelligence and extract knowledge that will be crucial for future answers about possible diseases and treatments on patients. Nowadays, the concept of DM is already applied in medical information systems for clinical purposes such as diagnosis and treatments, that by making use of predictive models can diagnose some group of diseases, in this case, heart attacks.

The focus of this project consists on applying machine learning techniques to develop a predictive model based on a real dataset, in order to detect through the analysis of patient's data whether a person can have a heart attack or not. At the end, the best model is found by comparing the different algorithms used and assessing its results, and then, selecting the one with the best measures.

The correct identification of early cardiovascular problems signs through the analysis of patient data can lead to the possible prevention of heart attacks, to the consequent reduction of complications and secondary effects that the disease may bring, and most importantly, to the decrease on the number of deaths in the future. Making use of Data Mining and analytics in healthcare will allow the analysis of high volumes of data, the development of new predictive models, and the understanding of the factors and variables that have the most influence and contribution for this disease, which people should pay attention. Hence, this practical approach is an example of how predictive analytics can have an important impact in the healthcare sector: through the collection of patient's data, models learn from it so that in the future they can predict new unknown cases of heart attacks with better accuracies. In this way, it contributes to the creation of new models, to the tracking of patient's health data, to the improvement of medical decisions, to efficient and faster responses, and to the wellbeing of the population that can be improved if diseases like this can be predicted and avoided.

To conclude, this project aims to present and show how Data Mining techniques are applied in healthcare and medicine, and how they contribute for the better knowledge of cardiovascular diseases and for the support of important decisions that will influence the patient's quality of life.

KEYWORDS

Cardiovascular Disease, Heart Attack, Data Mining, Machine Learning, Predictive Analysis, Neural Networks, Logistic Regression, Decision Trees

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI - Artificial Intelligence;

ANN - Artificial Neural Networks;

AUC - Area Under the Curve;

BI - Business Intelligence;

CAD - Coronary Artery Disease;

CARE - Collaborative Assessment and Recommendation Engine;

CART - Classification and Regression Trees;

CHD - Coronary Heart Disease;

CRISP-DM - Cross Industry Process for Data Mining;

CVD - Cardiovascular Heart Disease;

DM - Data Mining;

DT - Decision Tree;

EU - European Union;

ID3 - Iterative Dichotomiser 3;

KNN - Key Nearest Neighbor;

LAD - Left Anterior Descending vessel;

ML - Machine Learning;

MLP - Multi Layer Perceptron;

PA - Predictive Analytics;

REP - Reduced Error Pruning;

RELU - Rectified Linear Unit:

RF - Random Forests;

RIPPER - Repeated Incremental Pruning to Produce Error Reduction;

ROC - Receiver Operating Characteristics Curve;

SAS - Software SAS Enterprise Miner;

SEMMA© - Sample, Explore, Modify, Model, Assess;

SLP - Single Layer Perceptron;

SVM - Support Vector Machine;

WHO - World Health Organization;

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Cardiovascular Diseases

CVD (Cardiovascular diseases) belong to the group of disorders related to the heart and blood vessels, which affect the circulation of the blood throughout the body till the most important organ, the heart muscle. This group includes coronary and rheumatic heart disease, cerebrovascular disease, and other similar cardiac situations. These conditions may be related and triggered by different aspects such as severe illness, disability, but most likely, different types of angina pains, levels of blood pressure and cholesterol, blood sugar values, maximum heart rate and genetics, which are the most significant symptoms. Nevertheless, other outside routines that could negatively influence it include stress, overweight, smoking, alcohol intake and less exercise (Alzahani, Althopity, Alghamdi, Alshehri, & Aljuaid, 2014). These factors affect the way the blood is pushed and spread throughout the body, which can result in a heart attack, severe illness, disability, and most likely, death.

A heart attack occurs when the circulation of the blood through the body to the heart is interrupted due to low levels of oxygen and nutrients, causing serious damage to this indispensable organ. One of the main reasons is the existence of fat accumulations inside the walls of the vessels, which are responsible to supply the blood to the heart and the brain. When the obstruction of the blood is immediate it can happen a heart attack, but if it is restricted and the blood levels to the heart are reduced, it can be developed a symptom called angina, which is considered a chest pain. The angina may not origin injury to the heart muscle, but it is a warning signal that the individual has a chance to develop a heart attack (*Avoiding heart attacks and strokes*, 2005).

1.2. STUDY MOTIVATION AND RELEVANCE

Bearing in mind the WHO (World Health Organization), cardiovascular diseases are the main cause of the number of deaths in the world, being responsible for 17.9 million deaths per year, which represents 31% of the worldwide deaths. Heart disease is the most killing one affecting more than 75% of individuals living in countries of low and middle earnings, being 85% of all cardiovascular diseases caused by strokes and heart attacks (*Cardiovascular Diseases*, 2020). Regarding the distribution around the world, CVD are the primary cause of mortality in Europe, leading to 3.9 million deaths per year. The past 25 years have shown that the absolute number of these cases have raised in the Europe

continent and in the EU (European Union), where most of the countries registered new numbers of CVD cases (figure 1).

It was concluded that the rates from stroke and CHD (Coronary Heart Disease) are usually higher in central Europe and eastern Europe, rather than in other regions like northern, southern, and western Europe (Wilkins, Wilson, Wickramasinghe, Bhatnagar, Leal, Luengo-Fernandez, ... & Townsend, 2017).

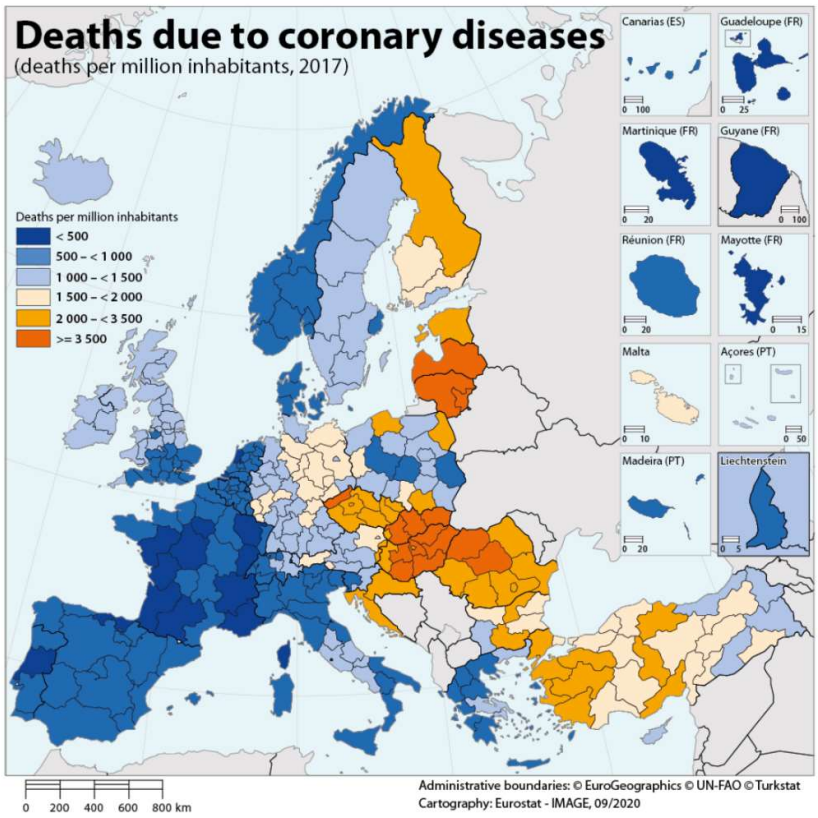


Figure 1 - Deaths caused by CHD in EU. Adapted from *Eurostat - Deaths due to coronary heart diseases in the EU*, by Eurostat, 2020, Retrieved from: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20200928-1. Copyright 2020 by Eurostat.

Overall, it is estimated that every year the EU economy spends 210€ billion on CVD (Wilkins et al., 2017). In the American continent, specifically in the United States, an individual dies every 36 seconds from CVD, and every 40 seconds someone has a heart attack. Each year, about 655.000 Americans die from heart problems, which represents 1 in every 4 deaths, and about 805.000 suffer a heart attack (*Heart Disease Statistics and Maps*, 2020).

CVD is a development issue in many different countries: people with low and medium earnings usually do not own the basic and primary healthcare benefits for early disease detection and its treatments, and the ones who may suffer from cardiovascular or other types of diseases may not have all the access to functional and good healthcare services to meet their needs (*Cardiovascular Diseases (CVDs)-Key Facts*, 2020).

The lack of efficient medical decisions can also have impact, leading to serious consequences which are thereby intolerable (Srinivas, Rani, Govrdhan, 2010), and as a result, some countries show lack of responses to the diseases when compared to high income ones (*Cardiovascular Diseases (CVDs)-Key Facts*, 2020).

The main and critical reason for the high number of heart attack cases and deaths is the fact that the disease is not identified at an early stage, and in majority, people do not pay attention to their health and body signs, such as chest pains and other health indicators, that may warn for the occurrence of dangerous diseases and serious consequences (*Cardiovascular Health for Everyone*, 2020). Hence, this disease reflects into serious issues for the general population and the health system, and in many cases, it could be easily avoided if it was predicted and identified earlier, preventing future complications and consequently, a lower number of patient's deaths (Singhal, Kumar & Passricha, 2018).

The total price of healthcare, specifically for cardiovascular diseases, is rapidly becoming hard to manage, and the distribution of good services at comfortable costs represents a vital issue across healthcare organizations (Singhal et al., 2018). This influenced the seek for preventive medicine, where the main focus is identifying the disease risk, and then, applying the right measures in case of early signs. In consequence, the role of healthcare has become more dynamic while recognizing the specific disease and its risks at an early stage (Davis, Chawla, Blumm, Christakis & Barabasi, 2008), as the best precaution for this illness is to make use of an efficient system that can forecast as soon as possible the symptoms, which will be able to save more lives (Khan, Nawi, Shahzad, Ullah, Mushtaq, Mir, Aamir, 2017).

In order to identify and recognize the signs and patterns that can give indications of a person having a possible heart disease in the future, it is used PA (Predictive Analytics). PA makes use of technology such as AI (Artificial Intelligence) and statistical methods, called learning models, that search through high volumes of data and examine it, leading to the extraction of useful information and the prediction of individual patient outcomes. This information includes data about past treatments and historical patient records, where PA can disclose astonish associations that the human brain would not be able to conjecture, by creating a profile that will be able to make predictions based on data from past individuals. This model is hereafter applied to new cases, so that a new prediction about a specific person's health and possible occurrence of diseases can be made. (Winters-Miner, 2014). There are already some studied methodologies that support this approach, such as CARE, which stands for Collaborative Assessment and Recommendation Engine. This approach makes predictions about future disease risks based on the patient's history and combines clustering with filtering methods, which contributes for the support of medical decision making.

These methods or algorithms are related with the correction of errors in a group of data, and they are based on its evaluation metrics, such as F-Score and correlation coefficients, so that the errors of input data can be minimized in the analysis. (Davis et al., 2008).

Making use of predictive analytics in the health sector brings many advantages, because it will positively influence on preventive medicine and public health. Some examples are the accuracy of diagnoses, the search for the right answer or treatment of individual patients, the development of models that require few cases and can be precise over time, the understanding of the best public medication needs by pharmaceutical companies, and the potential benefits of the future patient's life. Not only they will get more life quality and potentially receive the treatments and medication that best suits them, but they will also become more aware of the probability of health risks due to their genetic examination, to the analysis of predictive models transmitted by the experts, to the use of health applications and medical services, and to the better accuracy of the available information required for precise and correct predictions (Winters-Miner, 2014).

The study and prediction of CHD and CVD has been a challenging research issue to the medical field and to our day-to-day society, and some algorithms have been developed to predict the durability of CHD in patients (Xing, Wang, Zhao, 2007). These studies applied a distinct number of approaches regarding the related problem, and they have reached high classification accuracy values of 77% or higher (Kumari & Godara, 2011). It is of extreme importance to make use of technology advancements for the benefits of the society, mainly, to contribute to the development of medicine, to predict serious diseases at early stages, to treat patients with the best suitable treatments, and to contribute for the health and wellbeing of the populations (Winters-Miner, 2014).

1.3. STUDY OBJECTIVES

In sum, bearing in mind the researched topic and its relevance, the main goal of this project is to contribute for this global health issue by learning about the problem, and by using tools to apply machine learning algorithms and build predictive models that predict whether a person diagnoses a heart attack or not, so that in a real case the disease can be identified earlier, prevented, and treated as soon as possible. It is also aimed to understand if Data Mining applications are a reliable choice in the health sector, and if they can be efficiently used to alert for the prevention, as well as to determine the model that has the best results for the prediction of heart attacks.

Thus, to achieve this goal, there were defined specific objectives:

- Classify people that may develop a heart attack and build a predictive model about it;
- Identify the variables that most influence the occurrence of a heart attack;
- Apply predictive models - decision trees, logistic regression, artificial neural networks and an ensemble – to predict whether a person develops an heart attack in the future;
- Evaluate the complementarity used during the practical methodology for the support of the final results;
- Understand in which way this methodology is useful and shows a scientific contribute for the prevention of future problems;
- Analyse the importance of the information extracted from the model, as well as the respective selection of variables for the predictive modelling, whose goal is to attribute to a specific person his/her possibility of having the disease in the future;

1.4. DOCUMENT STRUCTURE

The project is organized in different phases which are divided and described in this document, structured as follows:

Chapter 2: Literature Review - a chapter focused on the literature review, containing all the theoretical survey regarding data mining meaning and techniques such as machine learning and predictive models, and also, research about heart diseases;

Chapter 3: Research Methodology - contains the structure of the methodology applied in this project which starts by the definition of the methodology, followed by the practical steps, and the description of the methodological procedures to be done;

Chapter 4: Results and Discussion - the chapter where the final results are described and the best model is selected after the implementation of the predictive models;

Chapter 5: Conclusions - includes the conclusions and exploratory analysis of the output data that is more favourable and give more value to the study goal;

Chapter 6: Limitations and Recommendations for Future Work - the final chapter where there are covered the limitations encountered during the project and recommendations for future work;

2. LITERATURE REVIEW

This second chapter presents the literature review that is most important for the work project. By making a main framework of the topics, it allows a better intuitive understanding of the referred concepts that will be focused and divided along the chapters. This review is done accordingly with the related subjects that are considered for the preparation of this project, which covers in majority data mining, machine learning, predictive algorithms, and cardiovascular diseases.

2.1. INTRODUCTION TO DATA MINING AND MACHINE LEARNING

Data Mining

Through the last years our society has been facing a huge increase in the use of information technology and information systems, and consequently, the production and storage of large volumes of data, which is in constant growth. It is estimated that more than 90% of the knowledge we have nowadays started to be acquired in 1950 (Nisbet, Elder, & Miner, 2009). This applies to data that flows in our personal life and to the world we live in, regarding education and health sectors, commerce, and industries. It is evaluated that the volume of data stored in different databases around the world doubles every 18 (Maheshwari, 2015) or 20 months. As this volume is continuously growing, there is an important subject to be referred: the understanding of this data. A critical success factor for companies is their capacity to handle all this information, which turns out to be a difficult challenge as the more information volume they have, the less will be the data proportion a human being can analyse (Witten & Frank, 2002).

Under all the layers of data that is collected, stored, and managed through the different devices, systems and others, there are useful patterns and hidden information that if transformed and analysed, can be crucial and lead to important results, which afterwards will have an impact on the day-to-day businesses decisions. This is all possible because of Data Mining, a buzz term that aims to analyse data present in databases and solve problems with it, by automatically or semi-automatically process data, discover patterns, and understand their relations (Witten & Frank, 2002).

DM can be seen as a synonym for “knowledge discovery from data” as it is an important step for the knowledge extraction (Han, Kamber & Pei, 2012), and in its simplest definition is the capacity to detect patterns in the available data (Srivastava & Han, 2016).

In sum, DM is characterized by a process that generates a pipeline composed by different phases, that must be followed to achieve the desired results.

A typical data mining application workflow contains a first phase called data collection, followed by data preprocessing, and a final analytical processing/algorithm phase. This pipeline can be viewed in figure 2, where the building blocks represent a particular design for an application solution (Aggarwal, 2015).

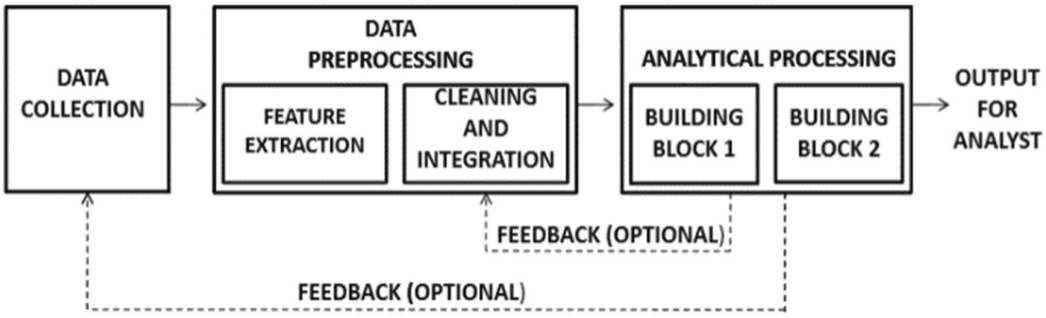


Figure 2 - Data Mining Flow. Adapted from *Data Mining: The Textbook,* (p. 4), by C. Aggarwal, 2015, New York, NY: Springer. Copyright 2015 by Springer International Publishing.

The first phase, as the name suggests, consists of collecting the data through the use of software and hardware tools, and then, storing it in a database or data warehouse for the preprocessing (Aggarwal, 2015). This is a critical stage because in order to learn from data, there is the need to collect quality data so good choices made may significantly impact the data mining process (Maheshwari, 2015).

Continuously, there is the preprocessing phase which englobes the extraction of features and data cleaning. The purpose is to put data in its suitable form to be preprocessed, for instance, transform it into a format that DM algorithms can understand, such as multidimensional, semi-structured format or time series. By doing this, the most relevant data features of an application are extracted from various and distinct sources and grouped in a combined format (Aggarwal, 2015). The quality of the data is critical for the success, so data cleaning is done in parallel because it usually contains inconsistencies to be removed, and missing values and errors that should be corrected and estimated. There are many ways to solve this, such as dealing with missing values, analyzing the outliers' effects, and transforming fields or variables (Maheshwari, 2015).

Finally, the last phase is the analytical processing which is responsible for the design of an effective analytical method, that is dependent on the experience of the analyst that is working on it (Aggarwal, 2015).

There are two main types of data mining processes – supervised learning and unsupervised learning – which will be covered hereafter, but for both types it is necessary to validate their results. For that, a common approach is making use of the confusion matrix (figure 3) which gives the precision capacity and generalization of the created models.

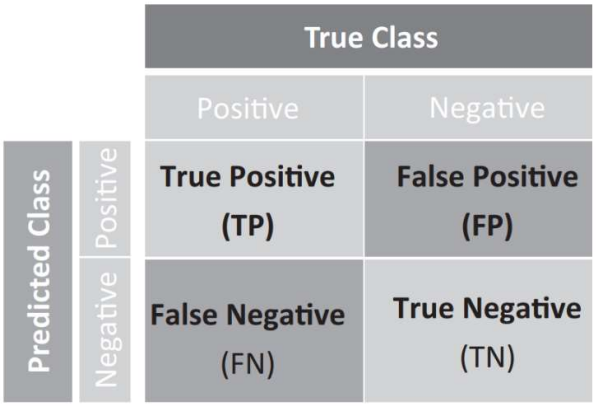


Figure 3 - Confusion Matrix. Adapted from *Business Intelligence and Data Mining*, (p.51), by A. Maheshwari, 2015, New York, NY: Business Expert Press. Copyright 2015 by Business Expert Press.

Some of the common metrics used are:

Predictive Accuracy = (TP+ TN) / Nr of Total Predictions;

Error = (FP + FN) / Nr of Total Predictions;

Precision= TP / (TP + FP);

A True positive -TP- is a true prediction when it is classified as true positive. In the same way, TN - True Negative – happens when it is truly classified as negative. On the opposite meaning, there is FN - False Negative – that appears when a true positive point is labelled as negative but is not a correct prediction of the model. Similarly, FP – False Positive – happens when a true negative value is estimated as positive one. This is how a confusion matrix should be interpreted.

Usually, every classification technique used for prediction has an accuracy value related to a specific predictive model. The maximum value that it can reach is 100 percent, but in fact, the models that show 70 percent or more of accuracy are trusted and considered to be used in different business fields, depending on their business nature (Maheshwari, 2015).

Data Mining Techniques

Data Mining covers a variety of applications and techniques that vary depending on the goals, assumptions, problem knowledge and the data being used. Those techniques constitute keys elements for the modeling process, and as Berry and Linoff (2004) say, the most common can be grouped in two categories named (1) descriptive and (2) predictive modeling. The first one is a type of modeling that aims to obtain summary descriptions and patterns of the data that is being used, so the analysts can understand it and increase their knowledge about the database. Usually, it includes different methods such as segmentations, cluster analysis, and association rules.

The second technique is different from the previous one, as it can be divided into classification or regression, depending on what we want to predict. In general, it works by learning from the existent data, and then choosing a decision criterion to classify new unknown examples. For instance, predict correlations between products and direct them for marketing campaigns (Sondwale, 2015).

DM consists of merging different areas of interest, such as data analysis, artificial intelligence and learning models (Nisbet et al., 2009). Making use of these theories and tools will help humans on the extraction of useful information from large amounts of data, and also, on making the base for the knowledge discovery process in databases (Lavalle, Hopkins, Lesser, Shockley, & Kruschwitz, 2010).

Machine Learning

ML (Machine Learning) is another important term related to DM. ML differs from DM as it represents the techniques, more specifically, the algorithms used during the DM process to acquire the structural description from the raw data (Patterson & Gibson, 2017). ML algorithms represent methods used by data scientists to uncover patterns in big data, and those can be categorized in two main groups depending on the methodology used to learn data: supervised learning or unsupervised learning. The first one, being the most used, requires that the training data - data used for the algorithm learning - is labeled with corrected and known answers, meaning that it has the correct values of the target variable it needs to predict. The algorithm will learn from its training dataset and make predictions on new data. Some examples of this type of learning are linear and logistic regression, classification, and decision trees (Castle, 2017).

On the other hand, the unsupervised learning refers to the idea that computers can learn, describe data, identify patterns and processes without a human to provide guidance, which means there is no need to learn from predefined labels.

In this type of learning, the algorithms learn to identify the patterns without a specified target variable, but with the behavior of the independent variables, so they can apply that knowledge and try to predict the actions or groups of the new elements (Chitra & Subashini, 2013). Some unsupervised learning examples are PCA (Principal Component Analysis), K-means Clustering, and Association rules (Castle, 2017). A summary of data mining techniques and algorithms is described in figure 4.

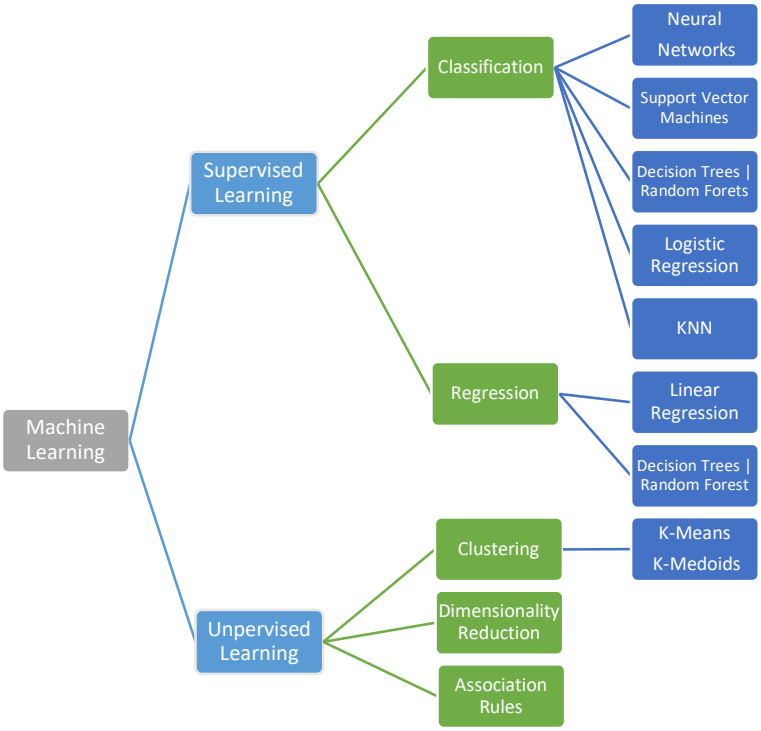


Figure 4 – Models and Algorithms used in Machine Learning (Source: Author Based)

2.2. ALGORITHMS FOR PREDICTIVE MODELING

Predictive analysis and modeling consist of using data, algorithms, and ML techniques to identify and forecast future outcomes, taking into account available historical data. With predictive analytical tools and models, organizations and businesses can transform their data and generate useful future insights with a good degree of precision, which will be crucial to forecast trends and behaviors, and lead to better decisions (Edwards, 2015).

As previously described, predictive models can be classified in two types: Regression and Classification.

The first one can be divided into linear and logistic regression, and the goal is to use input variables to approximate a mapping function to a continuous dependent variable, the output. This process demands the prediction of a certain quantity, where the final output may be a real value such as a floating point or an integer.

In classification, on the other hand, the goal is to use input variables to approximate a mapping function to classes, meaning the dependent variable is a category or label, and the observations are classified into one of two or more classes. For instance, an email can be assigned as "spam" or "not spam" (Brownlee, 2017). Classification is one of the most common tasks used in DM field, and the hole process is seen as a classification model development of the reality. In fact, in order to learn and understand the world we live in, we are making constant classifications which is how we can distinguish, for example, people from animals, objects and facts. The general flow of a classification process can be seen in figure 5, where the learning and classification phases are distinct.

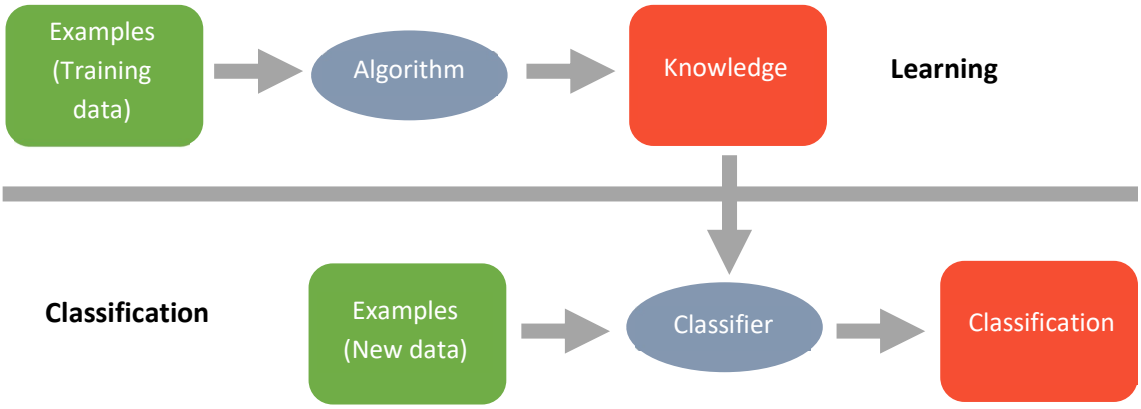


Figure 5 - General View of Modulation of the Prediction Process – it starts by a pre-classified training dataset (the examples), where through the use of an algorithm (for instance, decision tree, regression or neural network) is extracted knowledge that will be consequently applied for the classification of new unknow elements (F. Bação, personal communication, June 15, 2019).

To develop a predictive model, different algorithms can be used. For this effect, and to better distinct the algorithms used in the project, in this chapter a revision about each of them will be made. This part will be followed by a theoretical analysis of the problem in question, focusing on the cardiovascular disease topic, and on the use of technologies and DM techniques in medicine.

2.2.1. Artificial Neural Networks

Introduction to Artificial Neural Networks

The first model to be referred, and one of the main classification methods used in ML are ANN - Artificial Neural Networks.

These networks were already a topic of study and research in the middle of 20th century (Deng & Yu, 2014), and like the "neural" part of the name indicates, they are networks inspired by the way our brain works and processes the natural signs, and they intend to replicate its behavior, specifically, the method of human's learning. The ANN term took shape during the middle of XX century, when McCulloch and Pits presented in their paper a simplified model of a neuron (McCulloch & Pits, 1943).

ANN are composed by neurons, the principal unit or processing element, which represent analogically the neuron in the biological brain, and are modelled by them. Neurons make a dense and complex network by connecting with each other using signals, which is the way they learn and train themselves. Between the units there are weights, that are the primary means of long-term information storage in neural networks, and their update is the primary way neural networks learn new information (Patterson & Gibson, 2017). A neural network is seen as an artificial intelligence method which can solve different problems (Wang & Raj, 2017), and it is an optimal choice when the problem is difficult to specify or there is a lack of knowledge about it, but there is sufficient data (Zhang, Patuwo, & Hu, 1998).

There are different architectures to represent neural networks, which can go from one SLP (Single Layer Perceptron) to more complex structures - the MLP (Multilayer perceptron). These architectures display the connections between the nodes, and they are the key for the effectiveness of the neural network used (Aggarwal, 2015).

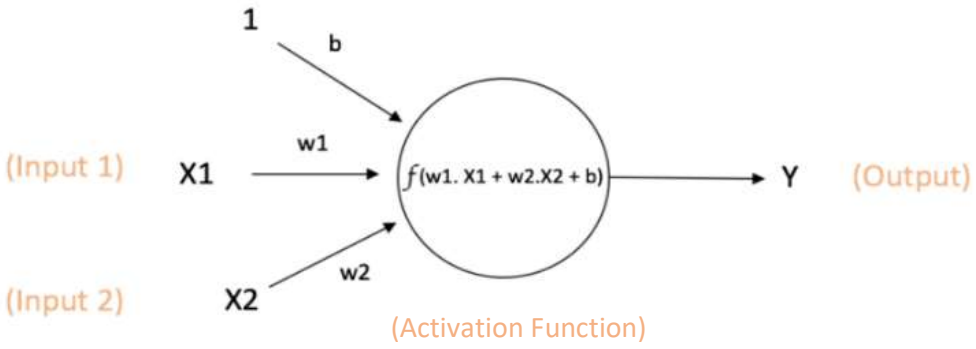
Single Layer Perceptron

The SLP, name given by Rosenblatt, is the simplest form of a neural network. He defended that its design represented some of the basic functionalities of intelligent systems, as well as its analogy to biological systems, such as the human brain. This fundamental unit is also called neuron or node, and it is intended to receive input from external sources or further neurons to compute an output. The SLP is composed by a weight **w** that is associated with each input, and which is assigned taking in consideration its importance with the other remaining inputs.

This representation is seen bellow in figure 6, where **X1** and **X2** are numerical inputs, and **W1** and **W2** are their respective weights. Each signal **X** is integrated in a neuron and multiplied by a weight **W,** being the final value the sum of the remaining product. The node will then use a function - the activation function **f -** to the previous summed value, which contains another input represented by **b**, named bias. The aim of the bias is to add a constant value to each node, which will give more flexibility to the model to adapt and fit the data.

Finally, the output represented by **Y** is calculated through the activation function, which takes the previous summed value and applies a mathematical operation on it, squashing the final value to a specific number rage, and determining the output, accuracy, and efficiency of the training model. The activation function is linked to each neuron, making it possible to define the activation or non-activation of the neuron, considering the relevance of each node's input for the prediction of the model (Sharma, 2017).

As most of the data used nowadays is non-linear, the aim of this calculation is to introduce non-linearity to the neuron's output, so they can master and learn the representations that are not linear (Ujjwalkarn, 2016). The activation function, also known as transfer function, is used to understand the final result of an ANN, such as a "Yes" or "No", by mapping the result in ranges from -1 to 1, or 0 to 1, based on the function that is being used (Sharma, 2017).



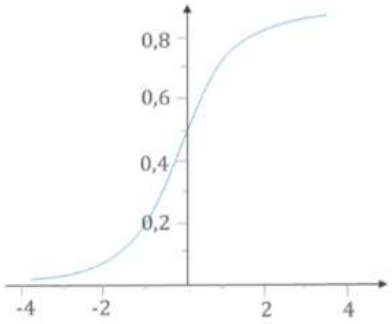
Output of neuron = f (w1 * x1 + w2 * x2 + b)

Figure 6 - Single Layer Perceptron. Adapted from *A Quick Introduction to Neural Networks,* by Ujjwalkarn, 2016, Retrieved from https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/. Copyright 2016 by Ujjwalkarn.

The activation function is a non-linear function, as adding non-linearity to the network avoids the restrictions with linear functions and allows the learning of compound networks. In ANN field there are distinct activation functions that may be applied, being the most common described below:

Sigmoid: the function takes a real input value and compresses it into an interval ranging from 0 to 1. The Sigmoid is mostly used in feedforward networks that need positive outputs, introduces non-linearity, and is a simple derivative function (figure 7);

$$f(x) = sigmoid(x)$$

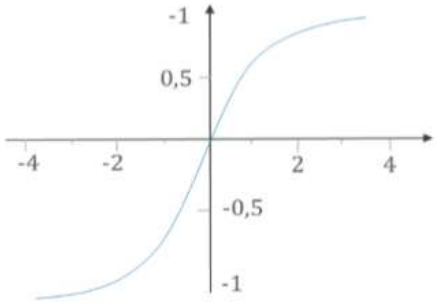


$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Figure 7 - Sigmoid Activation Function. Adapted from *Understanding Activation Functions in Neural Networks*, by A. Sharma, 2017, Retrieved from https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0. Copyright 2017 by A. Sharma.

Hyperbolic Tangent: the function is also called Tanh, and it takes a real input value and compresses it into an interval ranging from -1 to 1. Unlike the first function, tanh's output is zero-centered, and the gradient is stronger (figure 8);

$$f(x) = \tanh(x)$$



$$\tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

Figure 8 - Hyperbolic tangent Activation Function. Adapted from *Understanding Activation Functions in Neural Networks*, by A. Sharma, 2017, Retrieved from https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0. Copyright 2017 by A. Sharma.

Rectified Linear Unit (ReLU): This function is the most used around the world, and it is faster to train networks with larger structures (Krizhevsky, Sutskever & Hinton, 2012). ReLU takes real input values and replaces the negative ones with 0, meaning the activation is generated at zero (figure 9);

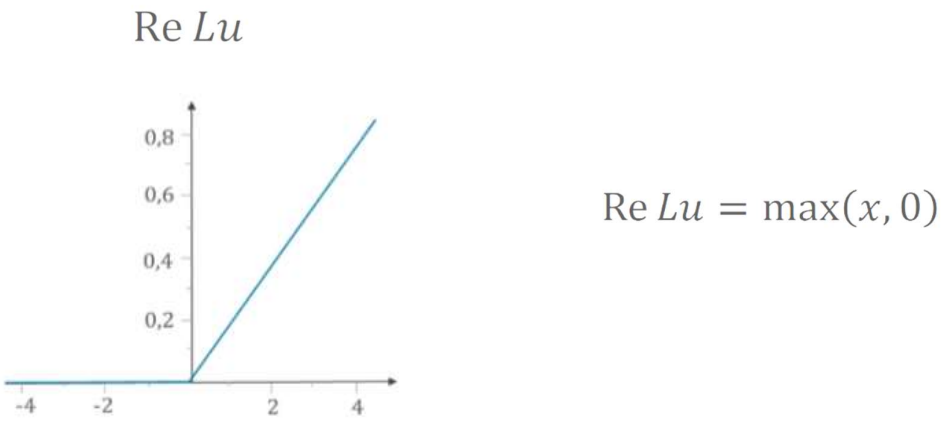


Figure 9 – Rectified Linear Unit Activation Function. Adapted from *Understanding Activation Functions in Neural Networks*, by A. Sharma, 2017, Retrieved from https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0. Copyright 2017 by A. Sharma.

The SLP is the simplest feedforward network as the information flows only into a single direction, starting from the input to the output nodes, and it does not have any hidden layers (Ujjwalkarn, 2016).

Multi Layer Perceptron

The MLP is a system consisting of simple interconnected neurons and represents the structure between an input and output vector without linearity. MLP has proved to be an efficient substitute to traditional statistical based techniques, and it is also an attractive option for the development of numerical models, as it can map non-linear functions and be trained to induce in an accurate way when presented with new data. MLP has been used in different tasks, being categorized in prediction, pattern classification, and approximation of functions. The first one, refers to the forecast of new and future patterns in a group of data. The second, classification, is related with the process of labeling data into classes, meaning it is a categorical value, and the approximation of functions is a technique used to map inputs to outputs, reflecting the variable's relationship (Gardner & Dorling, 1998).

The MLP is a feedforward network, which represents the simplest type of ANN, and is made of multiple neurons arranged in three different layers of the network. The neurons have connections between each other, that have respectively weights associated with them. Figure 10 shows an example of a MLP composed by the input, hidden and output layers.

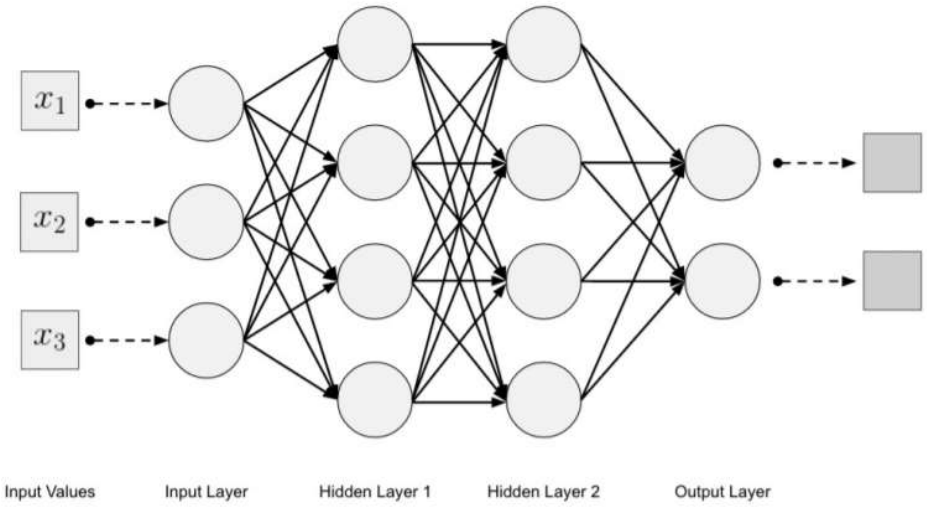


Figure 10 - Multi-layer Perceptron. Adapted from *Deep Learning: A Practitioner's Approach,* (p.79), by J. Patterson and A. Gibson, 2017, Sebastopol: O'Reilly Media. Copyright 2017 by J. Patterson and A. Gibson.

The first layer is responsible for introducing the values in the nodes, providing information received from the external environment to the network, passing it to the hidden layer without the need of having any computation. On this second layer, the nodes need to execute calculations and send that information from the input nodes to the output ones, as they do not contact with the outside (Ujjwalkarn, 2016). The biggest question arises in the hidden layer, specifically on the number of layers and nodes that should be used. There is no exact method to obtain the ideal number beyond trial and error (Sheela & Deepa, 2013), nevertheless, one layer is enough for most of the problems because the situations in which performance improves with at least two or more layers are not so frequent (figure 11). The existence of a larger number of units in the hidden layer might cause overfitting, which is not good for the model as the network learns to well the training dataset and may not be able to correctly classify new cases (Macukow, 2016).

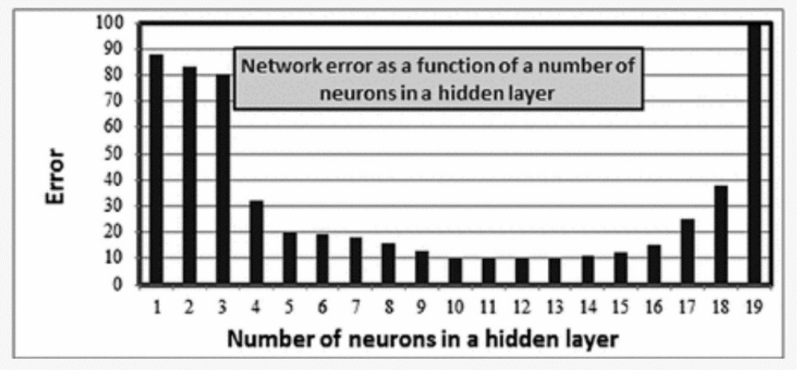


Figure 11 – Evolution of the Network Error based on the Number of Neurons in the Hidden Layer. Adapted from *Neural Networks- State of Art, Brief History, Basic Models and Architecture*, (p.11), by B. Macukow, 2016, Vilnius, Lithuania. Copyright 2016 by B. Macukow.

The third and final one is the output layer, which oversees the passing of information from the inside of the network to the outside. MLP is described as a feedforward network because the information moves only in a unique direction, starting in the input layer, passing through the hidden layer, finishing in the output nodes, meaning that there are no cycles (Ujjwalkarn, 2016). In a feedfoward network there is no backpropagation, however, in some networks the backpropagation makes the signal to move in both directions, allowing the neuron's output values to feed others from the same layer or previous ones. This methodoloy pretends to undersand which weights contribute more for the generated error, and adjust them in order to imporve the model (Rojas, 2013).

Training and Backpropagation Algorithm

The backpropagation algorithm is responsible for the training and learning of a MLP. The backpropagation of errors is one of the ways ANN can be learn, which means that is a supervised method, as the neurons learn from the training and labeled data. As it was referred, ANN are composed by nodes, which are connected between each other, having associated weights. The learning goal is to correctly assign these weights, so given an input value the weights will define the output (Ujjwalkarn, 2016). This procedure works due to the continuous adjust of the connection weights, which will decrease the difference between the expected output of the network and the real output value (Rumelhart, Hinton, & Williams, 1986). This algorithm is the most common method used for training, and it was appraised as the beginning of the contemporary deep learning theory (Wang & Raj, 2017).

The backpropagation is divided in two phases, the forward and backwards phase. In the first one, the inputs used for the training instances get into the ANN, which will lead to calculations using the existing set of weights in the layers. The final prediction is then compared to the training instance label in order to verify if the new label prediction is an error or not. In the backward phase the main goal is to learn the weights in the opposite direction, the backwards, by estimating the error of the nodes' output in the previous layers taking in consideration the errors of the subsequent layers (Aggarwal, 2015).

Backpropagation is an algorithm that improves its results regarding the loss function (J), also named cost function, which represents the difference between the estimated target value (y') and the true value (y). At the end, it shows how different are the mappings that we compared from the real values. The representation of the estimated and true values is clarified as follows:

$$J = y - y'$$

2.2.2. Decision Trees

DT (Decision Trees) are a type of supervised learning and one of the most popular predictive algorithms due to their structure and interpretability (Quinlan, 1986). DT are considered a classification approach, where the process of classifying is made by using a group of hierarchical choices, which are based on the variables' features displayed in a tree format (Aggarwal, 2015). At the end, this process produces a tree composed by nodes which are connected through edges that allow the extraction of decision rules expressed in English, so it can be understandable with an easy interpretation for different people (Berry & Linoff, 2004). The goal is to establish these rules solutions, so they can be well interpreted and allow the prediction of a value assumed by a target variable.

DT are composed by different nodes linked with each other: a root node that represents the parent of all the nodes located at the topmost of the tree, and the other nodes named *split criterion*, which is where happens a specific decision. Each node represents an attribute, the feature, each link represents a decision, a new rule, and each leaf represents a final result, for instance a class (Patel & Prajapati, 2018). The objective of each node criteria is to increase the difference of the distinct classes amongst the children nodes. Typically, it is a condition where the variables' features of the training dataset will be divided into one or more parts (Aggarwal, 2015). This model will split a sizeable set of data into smaller parts of records, that represent segments with the respect of the target. At the end, the process will originate a tree, usually with a hierarchical structure that partitions the training data in a top-down approach, having a root node at the top, and the respective decisions at the nodes being constructed under it. The group of rules are applied until the tree is pruned or there is no possible split, and it becomes a leaf node, with a completed tree. Once a decision rule solution or decision tree is generated from data, it can be used for predicting or estimating a class or value for a new case (Apté & Weiss, 1997).

There are two kinds of DT: the (1) classification tree, where the outcome is a categorical variable such as "fit person" or "unfit person", and the (2) regression tree, where the outcome is a continuous variable, for instance, a number like "123". DT algorithms use a group of training data that is rich and varied, to primarily train the tree and extract a set of rules that express what is known about the problem in question. After training, it is necessary to assure if the algorithm is classifying correctly new data, and for that a test set is used. Usually, DT are used to answer simple binary questions (Kulkarni, 2017).

The use of decision trees can be reflected in good advantages, such as the existence of understandable knowledge structures that are easily traceable, the efficiency of computational runtime (not very high),

the fast construction process when compared to other classification models, and the handling of both categorical and real values. This model uses rule-based classifiers, a group of *if – then* rules, in order to match antecedents to consequents. These rules are usually structured as follows:

IF Condition THEN Conclusion

The antecedent is the statement on the left side of the rule and may contain different logical operators such as "<, ≤, >, =,⊆ or ∈", that are implemented to the variables' features. The right side of the rule, the consequent, contains the class variable. The rule will cover a training example if the condition on the left matches the conclusion on the right (Aggarwal, 2015). These logical rules combine a sequence of tests and make decision trees that are intuitive and easy to interpret.

A practical example of this is to predict on whether to decide on playing golf or not, considering three variables: the outlook, humidity, and wind. The respective decision tree is shown below in figure 12 (Kulkarni, 2017):

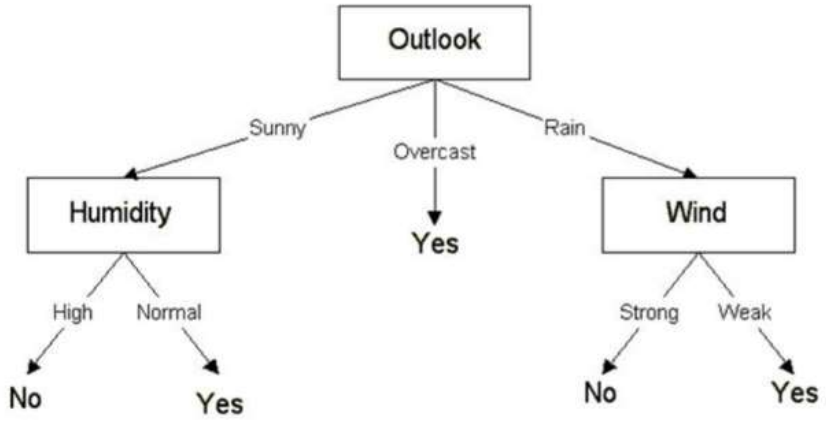


Figure 12 - Decision Tree Example. Adapted from *Decision Trees for Classification: A Machine Learning Algorithm*, by M. Kulkarni, 2017, Retrieved from https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html. Copyright 2017 by M. Kulkarni.

2.2.2.1. Decision Trees Algorithms

In order to construct a DT, it is necessary a training dataset that allows the algorithm to learn what are the features of the observations that can assume a predictive function. The training set is composed by observations characterized by the same variables of interest presented in the validation set, and it is used by the algorithm to define the decision rules that are going to be placed in the DT. The specific algorithm aims to generalize and discover patterns by figuring out which question is the best to separate the instances into distinct classes, resulting in a hierarchical structure of a tree.

The created rules will be applied to the observations of the validation dataset, and after, will allow the prediction of future behaviors regarding the values assumed by the target variable. Over the years, numerous DT algorithms were developed, for instance, CART (Breiman, Fiedman, Olshen & Stone, 1984), ID3 (Quinlan, 1986), SPRINT (Shafer, Agrawal, & Mehta, 1996) and CHAID (Milanović & Stamenković, 2016), but for the purpose of this project the focus will be placed on the behavior of ID3 and CART.

ID3 Algorithm

ID3 algorithm stands for Iterative Dichotomiser 3, as it iteratively divides at each step the features in different groups. The algorithm was invented by Ross Quinlan and it employs a top-down approach, meaning that trees are constructed from the top, root node, and at each repetition the best feature is selected, and a node is created. ID3 is applied for cases with a high number of attributes whose training set contains several objects, and for reasonable DT that do not require too much computation. This algorithm has an iterative structure, as a subpart of the training set is randomly chosen, and a DT is constructed from it. The other part of the training set is hereafter classified using the constructed tree (Quinlan, 1986).

Typically, ID3 is used for nominal classification problems (Sakkaf, 2020), and it can be synthetized through the following rules (Rita, 2018):

Split Criterion (node, (examples):

A ← best attribute for splitting the (examples)

Decision attribute for this node ← A

For each new child node

Split training (examples) to child nodes

For each child node / subset:

If subset is pure: STOP

Else: Split (child_node, (subset))

Following these rules will result in the construction of a new DT, that in the presence of new observations may obtain a predicted value for the target by following the branches of the tree.

In order to choose the most important feature the ID3 algorithm makes use of Information Gain. This nomenclature calculates the reduction in the Entropy (Wang & Suen, 1984), which estimates the level of order or disorder of the target dataset, and determines if the given feature separates the target classes well. The feature with the highest Information Gain value is then selected as the right choice.

For instance, if we designate an example dataset as S, the Entropy may be determined through the following expression (Sakkaf, 2020):

$$\text{Entropy}(S) = -\sum p_i * \log_2(p_i);\ i = 1 \text{ to } n$$

where,

n represents the total number of existing classes in the column of the target (for example, n=2 if the target class has two possibilities: “Yes” or “No”);

$\mathbf{p_i}$ is the ratio of the number of rows with a specific class **i** in the target column **n**, in the total number of rows of the dataset **S** (Sakkaf, 2020).

Assuming that dataset **S** has a total of 14 rows**,** with 3 feature columns: “Fever”, “Cough” and “Breathing Issues”, the representation will be as seen in table 1:

ID	Fever	Cough	Breathing issues	Infected
1	NO	NO	NO	NO
2	YES	YES	YES	YES
3	YES	YES	NO	NO
4	YES	NO	YES	YES
5	YES	YES	YES	YES
6	NO	YES	NO	NO
7	YES	NO	YES	YES
8	YES	NO	YES	YES
9	NO	YES	YES	YES
10	YES	YES	NO	YES
11	NO	YES	NO	NO
12	NO	YES	YES	YES
13	NO	YES	YES	NO
14	YES	YES	NO	NO

Table 1 - *Dataset S Example.* Adapted from *Decision Trees: ID3 Algorithm Explained, by Y. Sakkaf, 2020, Retrieved from https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df. Copyright 2020 by Y. Sakkaf.*

The table has two different target values (Infected column): 8 rows for “Yes” and 6 rows for “No”, hence, the entropy may be determined by the following expression (Sakkaf, 2020):

$$\text{Entropy (S)} = — (8/14) * \log_2(8/14) — (6/14) * \log_2(6/14) = 0.99$$

The highest the Entropy, the higher the measure of disorder will be, which means more randomness. On the other hand, the lowest the Entropy, the less the measure of disorder will be, which turns out to be the goal of machine learning models, to reduce uncertainty (figure 13).

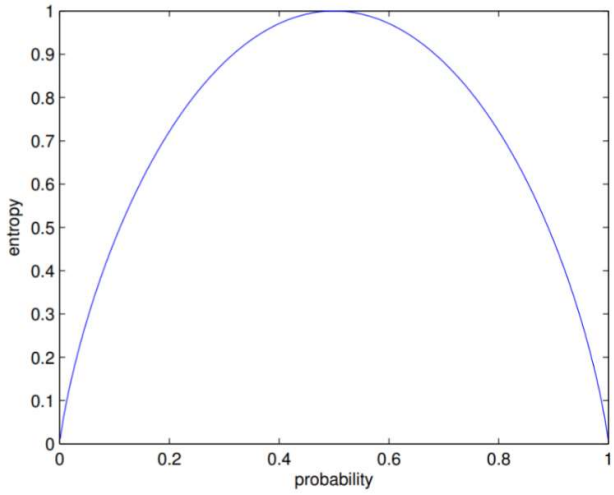


Figure 13 - Representation of the Entropy Relation. Adapted from *Model Selection for Data Analysis Based on the MDL Principle*, (p.122), by B. S. Nonchev, 2015. Copyright 2015 by B. S. Nonchev.

Consequently, by having the Entropy value it is possible to calculate the Information Gain**,** which measures how much information has been gained through a given choice at a given node. It is defined as the influence of the knowledge of a feature at a specific node. This influence is determined by the Entropy, previously calculated on the first place, so then the Information Gain can be applied by the following expression (Sakkaf, 2020):

$$IG\ (S, A) = \text{Entropy (S)} - \sum((|S_v| / |S|) * \text{Entropy } (S_v))$$

where,

v are the classes the target can assume, for instance, v=”yes” or v=”no”;

$\mathbf{S_v}$ is the number of rows in the dataset **S** with a specific feature **A** of value **v**;

$\mathbf{|S_v|}$ is the absolute value of the number of rows in $\mathbf{S_v}$**;**

|S| is the absolute value of the total number of rows in **S.**

Continuing with the example of dataset S, the calculation of the Information Gain will be made. The process should be applied to all the feature columns in order to calculate the different Information Gain values, but for the purpose of this practical explanation, it will only be explained for "Fever" feature as it is detailed bellow (Sakkaf, 2020):

$$IG(S, Fever) = Entropy(S) - (|S_{YES}| / |S|) * Entropy(S_{YES}) - (|S_{NO}| / |S|) * Entropy(S_{NO})$$

$|S| = 14$ For v = YES, $|S_v| = 8$ $Entropy\ (S_v) = - (6/8) * \log_2(6/8) - (2/8) * \log_2(2/8) = 0.81$

For v = NO, $|S_v| = 6$ $Entropy\ (S_v) = - (2/6) * \log_2(2/6) - (4/6) * \log_2(4/6) = 0.91$

Entropy (S) = 0,99

IG(S, Fever) = 0.99 - (8/14) * 0.81 - (6/14) * 0.91 = 0.13

By applying the same calculations to the other features, the process will end with the following Information Gain values:

- IG (S, Fever): 0,13
- IG (S, Cough): 0,04
- IG (S, Breathing Issues): **0,40**

Since the calculation of the Information Gain values is done, the DT can be started with the choice of the root node, which is the feature with the highest IG value. By looking at the results, the root node will be "Breathing Issues", and the initial step would look like this (figure 14):

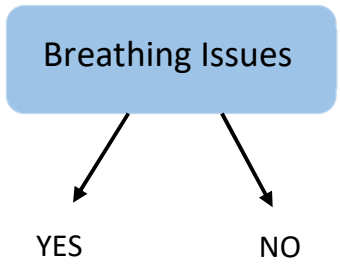


Figure 14 - Root Node of Decision Tree (Source: Author-based)

In recap, by having the Information Gain it is possible to select at each step the best feature, the one that has the highest value, and create a decision tree node. This process continues recursively, and if all the rows are considered to be in the same class, the actual node is set to a leaf node where the label will be the current class. When all the features are seen, or if all the decision tree nodes become leaf nodes, it means the tree is finished (Sakaff, 2020).

Overfitting and Pruning

ID3 is a recursive algorithm, so it divides the training dataset into smaller subsets till they are pure. This means there may exist parts that contain only one observation, which is not necessarily good, as we may be in the presence of overfitting. In due course, each leaf will represent a small group of feature combinations that are included in the training set, and the tree will posteriorly be unable to classify specific feature values that were not visualized in that data. In this way, the overfitting problem happens when training data is memorized very well by the model, and it learns noise and random features on instances belonging to the training data (Hoare, 2017).

Looking at figure 15, it is seen that the highest the number of decision nodes in the training set, the better the tree precision level will be. However, in the validation set this precision level decreases on a certain point, as the size of the tree increases. This is due to the fact the algorithm becomes too specific for the training set, being incapable of generalizing and leading to DT structures with more complexity than the necessary.

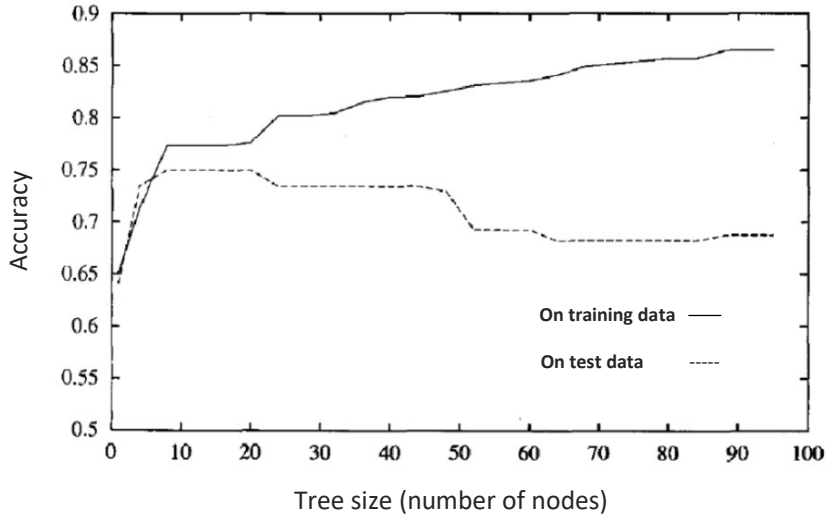


Figure 15 – Precision Level of a Decision Tree. Adapted from *A Comprehensive Guide to Decision Tree Learning,* by *A.* Chavan, 2019, Retrieved from https://www.aitimejournal.com/@akshay.chavan/a-comprehensive-guide-to-decision-tree-learning. Copyright 2019 by A. Chavan.

One way of avoiding this problem is by pruning the tree. Pruning is an approach used in DT and aims to reduce their size by removing the lower leaves of the structure that do not add valuable information – the capacity to classify new cases. As DT are likely to suffer from overfitting, an effective pruning or an early stop can reduce this likelihood (Hoare, 2017), but at the same time, maintaining the predictive accuracy as measured by a cross-validation set (Mangale, 2017).

Pruning can be named post-pruning when the tree is created in the first place, and in consequence, the non-significant branches are removed. However, it is also called pre-pruning when during the construction of the tree it is decided to stop developing some parts and terminate them, which turns out to be an interesting approach as it would reduce the probability of spending time and work on creating subtrees that afterwards would be discarded and not used (Witten & Frank, 2002).

There are two main techniques of doing post pruning: the Reduced Error Pruning, and the Cost Complexity Pruning. The first one, was proposed by Quinlan and is the simplest and the most understandable method while doing pruning. It considers that each of the decision nodes are candidates for pruning by removing a subtree rote node into a leaf node. REP will perform this operation if it does not increase the total number of classification errors. The process begins with the full structure of the tree and will go through each node to compare the number of classification errors made when the subtree is remained, with the number of errors made when the current node is transformed into a leaf node linked to a labelled class. Pruning is advisable if the simplified tree shows greater results than the one in the beginning. The process is iterative and will chose the nodes that enlarge the DT accuracy on the pruning set when removed (Patel & Upadhyay, 2012).

The second method, Cost Complexity Pruning also known as Weakest Link Pruning, is focused on estimating the error cost of a node and works by calculating the tree score. This method considers a function of the number of leaves in the DT and its error rate, which represents the percentage of the misclassified set. This cost will be computed for each node starting at the lowest leaves of the subtree, and it can be calculated in two ways: if the subtree was meant to be pruned, and if it was not. These two values are then compared, and if the first one shows a smaller cost pruning is done in the subtree, otherwise the structure remains the same (Han et al., 2012).

CART Algorithm

CART, which stands for Classification and Regression Trees, is on the other hand a binary DT containing specifically two branches for each decision node. This method was suggested by Breiman et al. (1984), and it recursively partitions the records of the training data into subsets, two child nodes, with similar values for the target variable (Larose, 2015). The methodology of this algorithm is based on three parts (Timofeev, 2004):

1. Construction of a DT with a maximum size – In the first phase the algorithm will split the training dataset till the final nodes hold observations of only one class. For the splitting process it is used a function called Gini Index or Gini Impurity, represented by the following expression:

$$\text{Gini (t)} = 1 - \sum_i p_i^2$$

Where ***pi*** is the instance's probability of being classified into a specific class.

The Gini Index measures the purity (Silipo & Melcher, 2019), the degree of a particular variable not being correctly classified when it is chosen in a random way. It varies on a scale from 0 to 1, where 0 indicates that all instances are part of a certain class, and 1 indicates that all instances are distributed randomly amongst different classes. After the Gini Index is calculated for all the feature classes, the lowest value (maximum reduction of impurity) will be select as the root node for the new DT. The goal is to choose decision variables that create subsets that are most close to a pure value, which is 0.

2. Definition of the best DT size - The second phase aims to optimize the trees before they are used for new data classification, because their maximum size may reflect too many levels and a high complexity. For this, pruning algorithms are used in order to cut of insignificant nodes and subtrees, so the accuracy of the DT increases.

3. Classification of a new dataset based on the built DT – The DT that is being developed will have two utilities: classification or regression. The result of this phase is a label class or value to each of the new instances, that by answering the tree questions will lead to one of the node terminals.

2.2.3. Linear and Logistic Regression

Regression is a concept that characterizes functions with the goal of predicting real values outputs, that by knowing the independent variables estimate the dependent one, meaning that it describes the linear dependence of the outcome variable from one or more predictor variables. Regression can be categorized in two groups: linear and logistic regression (Tripepi, Jager, Dekker & Zoccali, 2008).

Linear regression is the most common class of regression, and attempts to create a function that can describe the relationship between the variable X and Y, where by having the values of X, it can predict the values of Y in an accurate way. This is used to describe the linear dependence of the outcome variable - the dependent one - based on the independent variables, called the predictors. This is a commonly used technique, because it is appropriate to evaluate the relationship's strength between two variables (Bakar, Mohemad, Ahmad & Deris, 2006). Linear Regression is modelled by using the equation bellow:

$$Y = a + B_x$$

Where **Y** is the dependent variable, **X** the parameter vector, **a** is what intercepts **Y,** and **B** represents the input features. The previous equation can be expanded to the following one:

$$Y = a + b_0 * x_0 + b_1 * x_1 + \ldots + b_n * x_n$$

In a visual way, linear regression is designed as a straight line in a bi-dimensional plan, and its goal is to find the line that best fits the points, being close to the most numbers of points as it is shown in figure 16.

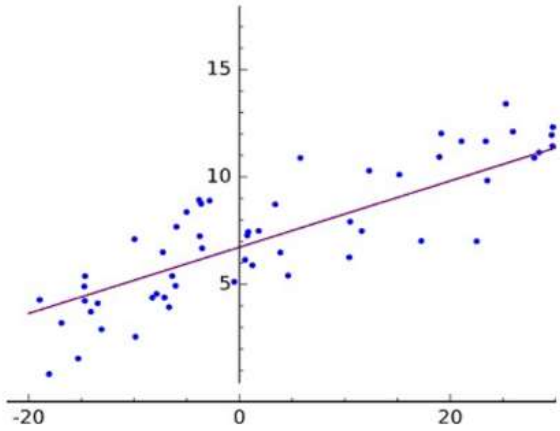
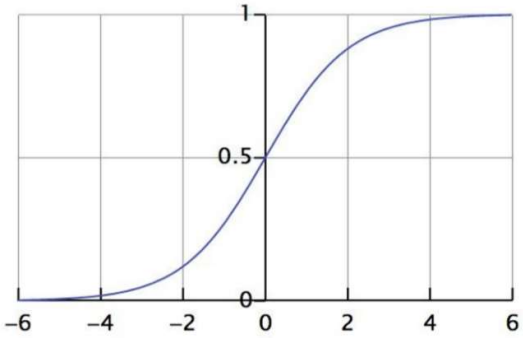


Figure 16 - Linear Regression Plotted. Adapted from *Deep Learning: A Practitioner's Approach,* (p.46), by J. Patterson and A. Gibson, 2017, Sebastopol: O'Reilly Media.
Copyright 2017 by J. Patterson and A. Gibson

While linear regression deals with continuous dependent variables, logistic regression on the other hand is a statistical method that represents the relationship between continuous or discrete independent variables, and the target, which is a categorical variable (Tripepi et al., 2008). It is also called a binary logistic model because by assuming the input variables, it aims to evaluate the probability of a binary result, where the output will be a probability of a specific category based on the predictors. In logistic regression the logistic function can be defined as follows in figure 17:



$$f(x) = \frac{1}{1 + e^{-\theta x}}$$

Figure 17 - Logistic Regression Function and Plot. Adapted from *Deep Learning: A Practitioner's Approach,* (p.66), by J. Patterson and A. Gibson, 2017, Sebastopol: O'Reilly Media. Copyright 2017 by J. Patterson and A. Gibson

The logistic regression function, being a continuous log-sigmoid function, is extremely useful as it takes different inputs from negative to positive infinity ranges and put them in a range between 0 and 1, allowing the interpretation of the output value as a probability with only two possible values, a dummy variable. In this function, ***f(x)*** is the probability of ***Y*** equals 1 given each ***X*** input. For instance, if our goal is to determine the probability of an email being spam and if **f(x)** equals to 0.6, we could formulate that **Y** has 60 percent of probability of being 1, which means that given the input, the email shows 60 percent of possibility of being spam.

2.2.4. Ensemble Models

An ensemble model is a classification and a composite methodology, that by combining independent and different classifiers known as base learners (Garg, 2018), increases the classification accuracy, resulting in a more powerful model (Han et al., 2012). Ensemble helps on reducing the noise, variance, and bias, which are weaknesses of independent models (Garg, 2018), creating at the end a better solution (Augusty & Izudheen, 2013).

Mostly, ensemble techniques can be classified in Bagging and Bosting. Bagging, or Bootstrap Aggregating, has the primary advantage of reducing the model variance (Bühlmann, 2012), and is a simple and efficient method for the creation of samples (bootstrap samples) from the original dataset, that will bring diversity to the model (Flach, 2012). This algorithm starts by repeatedly taking these samples with replacement and equal size from the training dataset, so that every record has the same probability of being chose. Afterwards, an estimation or classification model is trained on each bootstrap, so that each model is trained with different observations and a prediction is done for all. Bagging builds many independent predictors and combines them using model averaging techniques such as majority vote, weighted average, or normal average. An example of bagging ensemble is Random Forest Model (Grover, 2017).

The second technique – Boosting – is similar to the first one, but has a more sophisticated manner to create a variety of group samples, giving more prominence on choosing the points with incorrect predictions, so the accuracy can be improved (Bühlmann, 2012). Boosting is primarily focused on reducing the bias and differs from bagging because it is an adaptative algorithm, where the predictors are not made independently in parallel, but sequentially. This logic means that the subsequent predictors learn from the mistakes of the previous ones, and by subsequently adding models, it reduces the classifier's error.

In this technique each training instance is assigned with a weight that will be used for the training of the various classifiers (Aggarwal, 2015). The weights are then modified at each iteration and the same classification model is applied to the training sample, however, at each repetition the algorithm will apply a higher weight to the misclassified instances, in order to emphasize the most difficult cases. By doing this, a new classifier will be able to correct the bias on these specific data (Larose, 2015).

In this way, the algorithm works by fitting new models to the errors of the previous ones, focusing on improving the overall prediction. The future models will be dependent on the previous results and will end in the construction of a new classifier with lower overall bias (Aggarwal, 2015).

A visual comparison between bagging, boosting, and a single classifier can be explained in figure 18, where it is emphasized the distinction of parallel bagging and sequential boosting.

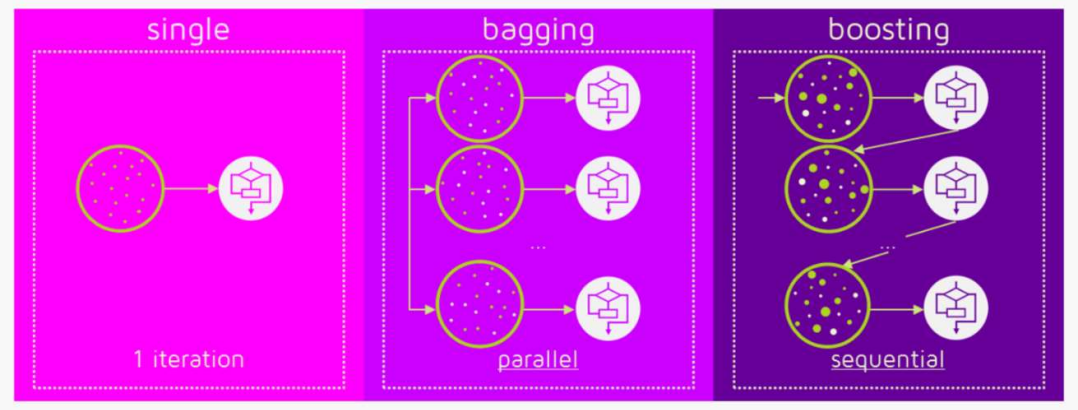


Figure 18 - Bagging and Boosting Examples. Adapted from *What is the difference between Bagging and Boosting*?, by A. P. *Garrido, 2016, Retrieved from: https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/. Copyright 2016 by A. P. Garrido.*

Random Forests

As it was covered on the previous point, ensemble combines different predictive models. Besides Bagging and Boosting, another famous technique is RF, Random Forests (Breiman, 2001), which is one the most prone supervised classification bagging techniques in ensemble learning, and works by combining different trained DT in order to merge them together and obtain a more robust model (Silipo & Melcher, 2019), as it is seen in figure 19. A RF is an extension of a classification or regression tree, being flexible, fast, and representing a strong method to extract high-dimensional data (Ziegler & Konig, 2014).

This algorithm was designed by Leo Breiman (2001) and differs from DT because it creates trees with more depth and less leaves. They have a good performance when confronted with many features and a low number of observations, and they can deal with continuous and categorical outcomes, creating different versions of RF such as classification, probability estimation, and regression (Ziegler & Konig, 2014). Not only random forests use a subgroup of data randomly chose, but they also take a group of random selected features to construct trees, rather than selecting of all of them (Napgal, 2017).

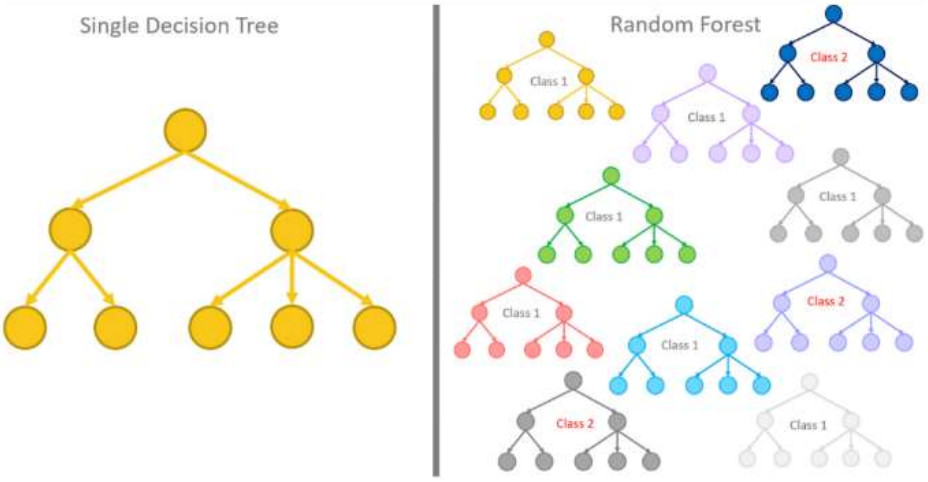


Figure 19 – Creation of a Random Forest: RF is created from a group of DT that learn in distinct ways, being the final classification based on all DT. Adapted from *From a Single Decision Tree to a Random Forest,* by R. Silipo & K. Melcher, 2019, Retrieved from: https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147. Copyright 2019 by R. Silipo & K. Melcher.

This model, by combining multiple DT, brings different advantages such as the fact that it is highly accurate and robust, it handles well data with high dimensionality, it is unlikely to overfit because it takes the average prediction of the individual constructed trees, and it is used for the two methods: classification and regression (Navlani, 2018).

In a classification problem the result is based on the majority vote where the most favoured class is selected as the output, whereas in a regression problem it is considered the average of all the trees' output as the final output. An illustration of a random forest ensemble algorithm is shown in figure 20.

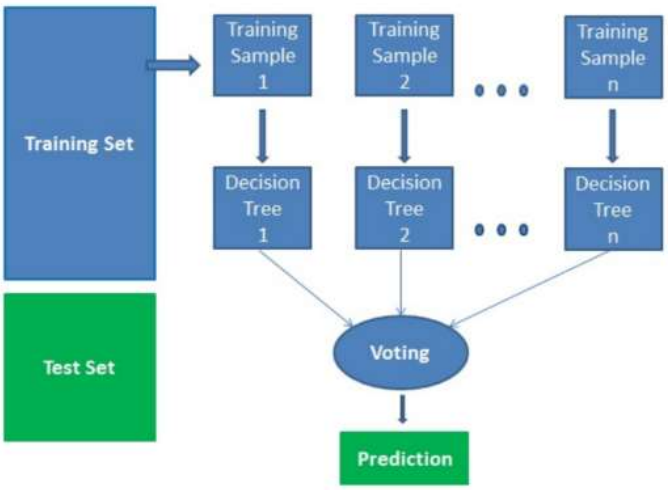


Figure 20 – Random Forest Classification Model. Adapted from *Understanding Random Forests Classifiers in Python,* by A. Navlani, 2018, Retrieved from: https://www.datacamp.com/community/tutorials/random-forests-classifier-python. Copyright 2018 by A. Navlani.

2.3. DATA MINING TECHNIQUES APPLIED TO HEART DISEASE

Data Mining and Medicine

Nowadays, the majority of recent hospitals provide essential equipment such as monitoring or observation tools and other important devices that gather all types of data and keep it saved in the hospitals' information systems. The role of Information Technology in healthcare is well established, and the extensive use and need for medical information systems, as well as the large rise of databases in healthcare, required that long-established analysis methods should be combined with recent ones to increase the efficiency of computer analysis (Lavrač, 1999). The continuous development of Information Technology applications in healthcare has brought society desires for greater services and lesser costs, and the modernization of the hospitals' systems has allowed a much simpler access to relevant information. This type of systems covers a variety of functions, such as patient's acceptance, treatment records, prescriptions, accounting and payment, but it can also increase the pace and quality of the tasks in reference to medical documentation, pharmacies, and laboratories (Wasan, Bhatnagar, & Kaur, 2006).

The high volumes of data combined in medical databases contain huge records that reflect the individuals' history, such as diseases, diagnosis, treatment approaches, historical information, and also, other clinical administration data. Such volume of data makes it difficult on the extraction of useful information for decision support, meaning there are required specific tools and tasks for the gathering and visualization of data, but also, methods for efficient computer-based analysis that will help on the posterior effective use of the data, and will ensure a flow of helpful, accurate, and meaningful information that can support important medical decision-making (Lavrač, 1999). To satisfy these needs, DM is again introduced and used in medical information systems, where it is normally applied for clinical objectives like diagnosis, therapy, and treatment. The concept of DM is so far successful and has been applied in different fields, and in the last twenty years it has been expanded to other areas of medicine along with artificial intelligence. Some examples of applications are diagnostic studies, health insurance fraud detection, therapeutic decisions, molecular biology, and medicine application (Lee, Liao & Embrechts, 2000).

DM methodologies are capable to uncover patterns and discover relationships amongst the instances in healthcare data, and an unlimited capacity to use it in a precise and efficient way. Therefore, DM contributes to the business intelligence process, being relevant for medical approaches like monitoring and diagnosing different kinds of diseases (Patel & Patel, 2016), being an important method to find

behaviours and extract and spread medical knowledge, providing the best patient care needs and effective diagnosis (Wasan et al., 2006).

Medical informatics make use of the existent tools, technologies and methodologies already developed and used in the context of data management and databases, which englobe different approaches like statistical analysis methods, identification of relationships and patterns, the use of machine learning algorithms, the use of tools that interactively allow the visualization, analysis and organization of data, and the discovery of uniformities in the data. Such methodologies and others help the intervention of humans in the process on analyzing and discovering data (Lee et al., 2000).

The early identification of diseases is arduous, however, it has a crucial role in medicine. The high volumes of data in this sector brought as a requisite the use of modern DM techniques to uncover difficult patterns which are hard to find with more traditional methods, and that can give support in decision making and in doing predictions in healthcare, so that the identification of diseases can be made easily and on time (Wasan et al., 2006). For this, DM approaches have been growing considerably in the medical industry, specifically with relation to different diseases like heart disease, diabetes, different types of cancer, hepatitis, and others, and they have been applied to these, showing an important advance regarding medical predictions and decisions. Table 2 summarizes some medical DM methods applied for different diseases (Patel & Patel, 2016).

DISEASE NAME	METHODOLOGY USED
CHD - Coronary Heart Disease	Predictive Model using DT Algorithms: ID3, C4.5, C5 and CART (Shalvi & DeClaris, 1998), (Xing et al., 2007).
Diabetes	Medical Data Classification with Genetic Algorithm (Brameier & Banzhaf, 2001).
Diabetes, Cancer	Disease Classification with KNN (Tang & Tseng, 2009).
Chest Disease	Model Construction with ANN (Yumusak & Temurtas, 2010).
CHD - Coronary Heart Disease	Accuracy Improvement with Naïve Bayesian (Xue, Sun, & Lu, 2006).

Table 2 - Summary of Medical Data Mining Techniques. Adapted from *Survey of Data Mining Techniques used in Healthcare Domain,* by S. Patel and H. Patel, 2016, *International Journal of Information Sciences and Techniques*, Vol.6, p.57. Copyright 2016 by S. Patel and H. Patel.

Data Mining and Heart Disease

DM is proved to be part of the development of medicine applications and a contribution for the identification and treatment of different health problems, mainly heart diseases, which are considered one the main cause of deaths in the world (Khan et al., 2017). One type of heart disease is cardiovascular disease, a term that denotes a group of conditions that influence how the body vessels and the heart work, affecting the way the blood circulates throughout the body, which can result in a serious illness, disfunctions, and most likely, death. A lot of information and studies about the diagnose of heart attacks and its symptoms have been made, where doctors and physicians list the most significant ones as being the individual's age and sex, the existence of different types of chest pains, the level of blood pressure and cholesterol, the values of blood sugar, the maximum heart rate achieved, and genetics. There are also other habits that might negatively influence it, which include lifestyle, levels of stress, smoking, overweight, frequency of alcohol consumption and levels of exercise (Alzahani et al., 2014).

The concern and prevention of heart disease has shown to have a great impact around the world, and studying its relevance, impact, and prediction, is the focus of this project. This disease is one example of how DM techniques are applied in medicine, which includes visualisation of data, analysis of variable correlations, supervised methods like ANN and how they are used in medical databases for the purpose of handling this type of data, and to successfully determine patients with an elevated risk, to recognize the most relevant factors for the occurrence of heart disease, and to set up models that reflect the relationship between two or more variables in an understandable way. These models are examples of how and where current medical knowledge is extracted (Lee et al., 2000).

Several medical DM techniques have been used for diagnosing and predicting a few groups of heart diseases, which are focused mainly on three types named, CAD (Coronary Artery Disease), CVD (Cardiovascular Disease) and CHD (Coronary Heart Disease). These techniques are proved by many literature studies on the field, where different classification algorithms have been implemented. CAD, named the most usual type of heart disease, was covered in some specific studies that used three features of the stenosis vessels, which are named Right Coronary Artery, Left Anterior Descending and Left Circumflex. In one study ANN was applied to predict the stenosis of each individual vessel, and a MLP was employed for the classification showing results of 69.39%, 73%, and 64.85% of accuracy for each respective vessel. On another study, other classification methods were used such as Naïve Bayes, as well as the C4.5 DT algorithm, and the KNN. In this study, the best accuracy was reached by C4.5 algorithm with accuracy values of 68.33%, 74.20%, and 63.76% respectively, and its use revealed to be the best one to diagnose CAD through Left Anterior Descending stenosis when compared to the values of the other studies.

For the remaining categories of heart attacks, CVD and CHD, different studies and predictions were also explored and discussed, showing more promising results. The following table 3 displays a summary of the effectiveness of the DM approaches applied on these different types of heart diseases, where the highest accuracy value is shown for the respective disease (Alzahani et al., 2014).

Data Mining Approaches	Study Objective	Maximum Accuracy	Reference
ANN	Diagnose the presence of CHD	91.0%	Xing et al., 2007
Naïve Bayes Classifier and GA Feature Reduction	Diagnose the presence of CVD	96.5%	Anbarasi, Anupriya, & Iyengar, 2010
DT and GA Feature Reduction	Diagnose the presence of CVD	99.2%	Anbarasi, et al., 2010
RIPPER Classifier	Diagnose the presence of CVD	81.08%	Kumari & Godara, 2011
C4.5 DT Classifier	Diagnose CAD through stenosis of LAD vessel	74.20%	Alizadehsani, Habibi, Bahadorian, Mashayekhi, Ghandeharioun, Boghrati, & Sani, 2012
	Diagnose the presence of CHD	82.5%	Srinivas, Rao, & Govardhan, 2010
C5 DT Classifier	Diagnose the presence of CHD	89.6%	Xing et al., 2007
SVM	Diagnose the presence of CHD	92.1%	Xing et al., 2007
Clustering	Diagnose the presence of CVD	88.3%	Anbarasi, et al., 2010
KNN	Diagnose CAD through stenosis of Left Circumflex vessel	61.39%	Alizadehsani et al., 2012
Hybrid Genetic Neural Network	Diagnose the presence of CVD	89%	Amin, Agarwal, & Beg, 2013

Table 3 - Data Mining Techniques used for the Prediction of Heart Disease. Adapted from *An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction,* by Alzahani et al., 2014, *Lecture Notes on Information Theory Vol. 2*, p. 313. Copyright 2014 by Alzahani et al.

The use of DM approaches in CAD, CVD, and CHD showed to be encouraging, but although the heart disease predictions do not have an accuracy of 100%, and hence must not be used alone for the process, it is seen that these results are promising to be used and would assist the professionals on making early diagnosis, on decision making, on preventing the diseases, and on saving more humans from heart attacks (Alzahani et al., 2014). Because of the prosperous application of DM methodologies for heart diagnosis, a few prediction systems that already use the previous techniques have been proposed (Palaniappan & Awang, 2008).

There are also other types of studies on cardiovascular diseases and their diagnosis, and research has provided different methodologies for the treatment of such diseases (Khan et al., 2017). For instance, Milan Kumari designed a system named RIPPER that stands for Repeated Incremental Pruning to Produce Error Reduction, and it is a rule-based classification algorithm that generates rules matching the performance of DT. Babaoglu et al. (2009) studied the use of ANN to determine the existence of artery disease based on the exercise stress testing, and other techniques like DT and SVM were also applied to explore datasets of coronary diseases, and different accuracy results were obtained (Parthiban, Rajesh & Srivatsa, 2011).

Anbarasi et al. (2010) assessed the use of CART and ANN with the intent of predicting heart diseases in individuals, and Detrano et al. (1989) did different investigations to predict the heart disease on a particularly useful dataset. On this case, the results showed that DT performance had highest accuracy, however, it was found that Bayesian classification had related truthfulness as that of decision tree method. Genetic Algorithm was also used by Anbarasi et al. (2010) to determine the attributes that have more influence on contributing to the diagnosis of the cardiac disease, and Detrano et al. (1989) performed experimental results that exhibited precise classification of heart diseases, having an accuracy of nearly 77% by using logistic regression (Srinivas, Rani, & Govrdhan, 2010).

Different results were discussed regarding the prediction of any type of heart disease by applying DM techniques with their classifiers and extension of the classifiers, showing surprising results. Also, different research papers used distinct classifiers algorithms or techniques, such as SVM, ANN, and Naïve Bayes, as well as their extensions which are different types of DT, KNN, ANN, MLP, Genetic Algorithms, and Feature Subset Selections.

3. RESEARCH METHODOLOGY

This section describes how the project's methodology was conducted and the different phases that were followed to accomplish the final goal. The chapter will cover the practical methodology used, the description of the dataset, the necessary steps followed since the beginning till the choice of the best model, as well as the software chosen for the evolution of this project.

To achieve a coherent structure, it was defined at the beginning a proposal with concrete steps that should be followed:

1st - Preparation: Definition of the project's topic with the support of the project coordinator, as well as the goals to be achieved, the respective work structure, the dataset and the practical methodology to be used;

2nd - Literature Review: Include the theoretical investigation and the approach contextualized with the field being study - heart attack disease, predictive models, and the use of machine learning algorithms;

3rd - Development of the Model: Specification of the methodology steps followed for the construction of the model and their application and development within the chosen software;

4th - Presentation of Results: Presentation of the main results and discussion;

5th - Conclusion: Presentation of the conclusions giving answers to the investigation questions and to the defined goals;

6th - Limitations and Future Work: Specification of suggestions and recommendations to be done in future studies.

3.1. DEVELOPMENT OF THE MODEL - METHODOLOGY STEPS

In DM field, conceptual models should be followed to provide guidance when planning and developing a project, in order to achieve the needs of any particular industry or company. The conceptual base model used in this project is named CRIPS-DM, which was conceived in 1996 by a consortium, and stands for "Cross-Industry Standard Process for Data Mining". This approach is aimed to be a tool for industries and is composed by six phases that are adaptive, meaning the next phase often depends on the outcomes of the previous one. Figure 21 represents these phases connected by the arrows (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer & Wirth, 2000). This conceptual sequence registers the life cycle of a DM project, and one of its main advantages is the fact that it is independent from the sector and from the analysis tool being used (Larose, 2015).

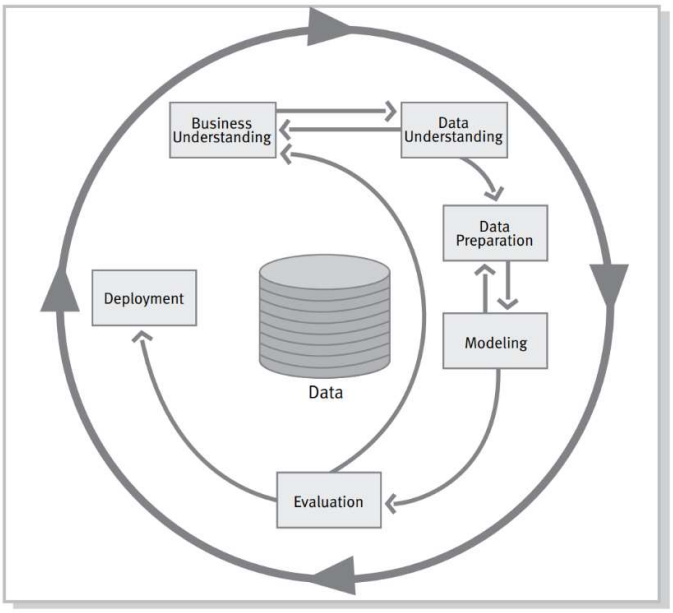


Figure 21 – Methodology Steps of CRISP-DM. Adapted from *CRISP-DM 1.0: Step-by-step Data Mining Guide*, (p.10), by Chapman et al., 2000, SPSS Inc. Copyright 2000 by the Members of CRISP-DM Consortium.

The model is not rigid, is complete, organized, and structured (Azevedo & Santos, 2008). The first stage is Business Understanding, which focus on looking for the project objectives and requirements from a business side, and it is considered the most important step as an incorrect analysis may influence the whole project, and consequently, its conclusions. It involves several steps, including business and data mining goals, as well as the production of the project plan (Shearer, 2000). Right after comes Data Understanding, where the collection and initial data analysis, insights, and quality is done. The third stage is Data Preparation, which groups all the activities required for the construction of the final dataset since its original raw form (Azevedo & Santos, 2008). This includes different tasks such as dimensionality reduction, data transformation methods, cleaning, and formatting. Modelling is the fourth phase where different model techniques are applied, and it is divided into four substages: the choice of modelling approaches, the test design, and the creation and evaluating of the model. On the fifth stage, the Evaluation, the model(s) are built and obtained. Before proceeding into the final model deployment, they are evaluated and their steps reviewed, to make sure they meet the business goals. Having this phase finished, a decision about the use of DM results may be reached. The final stage - Deployment - is where happens the process of planification and implementation of the model. As it is the last task, it should be made a review about the whole project, the identification of the positive and negative points, as well as the improvements to make in the future and the respective final report (Shearer, 2000).

SEMMA©

Having the conceptual base for the development of the project, the practical operationalization of the work is done by applying SEMMA©, which stands for **S**ample, **E**xploration, **M**odify, **M**odel and **A**ssess (figure 22).

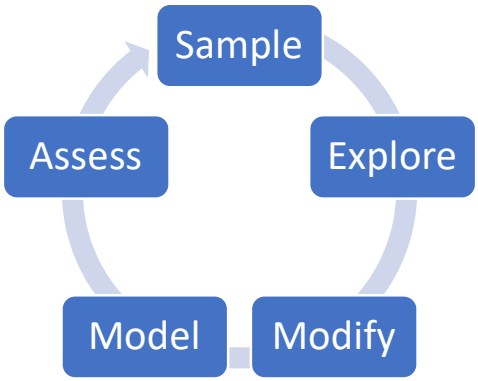


Figure 22 - Structure of SEMMA© Methodology (Source: Author Based)

SEMMA© was developed by SAS Institute, a company of Business Analytics software that aims to subserve and facilitate the data exploration process and their treatment, and the software that will support this project is *SAS Enterprise Miner* 15.1, whose goal is to clarify the data treatment process so that predictive and descriptive models can be created with high precision based on the analysis of high dimensions of data, and allowing the construction of interactive process flow diagrams. As the name suggests, SEMMA© is composed by five stages with different meanings:

Sample: creation of the input tables, partition, categorization, and data sampling;

Explore: data exploration and variable selection, used to graphically and statistically explore the dataset to understand its clarity and organization. In this phase, it is also performed the elimination of outliers and the handling of missing values;

Modify: Modification and variable transformation. In this step new variables are created from the initial ones in order to give value to the exploration and to the extracted knowledge. The existence of incoherencies and incorrected values are also checked;

Model: Data modulation phase. At this point the dataset is divided in two parts: the training set representing 70%, and the validation set representing 30%. For the modulation process some variables are selected based on their correlation and worth.

Assess: the final stage consists of the evaluation of the obtained results and model comparison;

SEMMA© is an intuitive approach applied in a sequential way, with a user-friendly application allowed by SAS Miner tools and tasks. This process is conducted by a flow chart that can be modified and saved which allows the analyst of the business to understand the flow and the meaning of the new data. SAS Miner not only contains a group of tasks that can be combined so multiple models can be created and compared, but it also allows the detection of isolation points, variable transformation, sampling, data partition, and advanced visualization tools that provide a quick and easy analysis of the dataset.

3.2. DATASET

The chosen dataset was provided originally from UCI Machine Learning Repository that englobes four databases linked with the related source. Having a total of 76 attributes, it was initially donated by David Aha, however, all the accessed experiments were done with a subset of 14 variables, with details of 303 patient's information. This specific subset was extracted from Kaggle public website (*Healthcare: Dataset on Heart attack possibility, 2020*) and contains health cardiac indicators of this total 303 patients, which diagnoses the occurrence of a heart attack in a specific individual. The independent variables refer to health indicators, such as chest pain, blood pressure and cholesterol, and the dependent variable - target - characterizes the occurrence of a heart attack in an individual. This variable has a value of 0 if there is no/less chance of heart attack, and assumes the value of 1 if there is more chance of having the problem.

3.3. DATA ANALYSIS AND EXPLORATION

Variable analysis and Classification

Having the dataset defined and extracted, the initial action is to import it to SAS Enterprise Miner with the use of File Import Node. Then, understanding the variables of the dataset is important, which includes the register of their name, meaning, and level. There are a total of 14 Input variables, where 4 are binary and 10 are interval, and 303 records. Table 4 summarizes this information:

Variable Name	Variable Description	Variable Level
Age	individual's age in years	Interval
Sex	individual's gender - 1: male | 0: female	Binary
Chest Pain Type (cp)	chest pain classified in values: - 0: typical angina - 1: atypical angina - 2: non-anginal pain - 3: asymptomatic pain	Interval
Blood Pressure (trestbps)	resting blood pressure in mm Hg. (Above 130-140 is a cause for concern)	Interval
Cholesterol (chol)	measurement of serum cholesterol in mg/dl. (Above 200 is a cause for concern)	Interval
Fasting Blood Sugar (fbs)	Measurement of how a person's body is managing blood sugar. (FBS> 120 mg/dl signals diabetes) value 1: > 120 mg/dl | value 0: < 120 mg/dl	Binary
Electrocardiographic results (restecg)	Results of the electrical activity and rhythm of the heart - Value 0: normal - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) - Value 2: probable or definite left ventricular hypertrophy	Interval
Maximum Heart Rate (thalach)	Highest value of heart rate achieved in min/beat achieved during thalium stress test	Interval
Exercise Induced Angina (exang)	angina induced by exercise value 1: Yes | 0: No	Binary
Oldpeak	st depression (electrocardiogram term) induced by exercise relative to rest	Interval
Slope	the slope of the peak exercise in st segment 0: upsloping |1: flatsloping | 2: downsloping	Interval
Number of Major Vessels (ca)	number of major vessels colored by fluoroscopy (0 - 4)	Interval
Thalassemia (THAL)	Thalium Stress test Results: 0: Normal | 1: Fixed Defect | 2: Reversable Defect	Interval
Target Variable (DepVar)	Have or don't have the disease 0: no |1: yes	Binary

Table 4 -Variable Description. Adapted from *Healthcare: Dataset on Heart attack possibility*, by R. Rahman, 2020, Retrieved from https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset). Copyright 2020 by Kaggle.

Data Exploration

Secondly, the analysis of interval and class variables is done in more detail using StatExplore Node, which allows to make a statistical analysis and to understand better the different values, as well as the patients' features (table 5 and table 6).

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
age	INPUT	54.36634	9.082101	303	0	29	55	77	-0.20246	-0.54217
chest_pain	INPUT	0.966997	1.032052	303	0	0	1	3	0.484732	-1.19307
cholestoral	INPUT	246.264	51.83075	303	0	126	240	564	1.143401	4.505423
electrocardiographic_results	INPUT	0.528053	0.52586	303	0	0	1	2	0.162522	-1.36267
max_heart_rate	INPUT	149.6469	22.90516	303	0	71	153	202	-0.53741	-0.06197
nr_major_vessels	INPUT	0.729373	1.022606	303	0	0	0	4	1.310422	0.839253
oldpeak	INPUT	1.039604	1.161075	303	0	0	0.8	6.2	1.26972	1.575813
resting_blood_pressure	INPUT	131.6238	17.53814	303	0	94	130	200	0.713768	0.929054
slope	INPUT	1.39934	0.616226	303	0	0	1	2	-0.50832	-0.62752
thal	INPUT	2.313531	0.612277	303	0	0	2	3	-0.47672	0.297915

Table 5 - Interval Variables Summary Statistics (Source: SAS Miner Output, 2020)

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	blood_sugar	INPUT	2	0	0	85.15	1	14.85
TRAIN	exe_induced_angina	INPUT	2	0	0	67.33	1	32.67
TRAIN	sex	INPUT	2	0	1	68.32	0	31.68
TRAIN	target	TARGET	2	0	1	54.46	0	45.54

Table 6 - Class Variables Summary Statistics (Source: SAS Miner Output, 2020)

At a first glance it is concluded that both interval and class variables do not have missing values, and the individual's age varies between 29 and 77 years, which makes this sample of people with a mean of 55 years. In this group of patients, the dependent variable tells us that 54.46% had a heart attack (165 people) and 45.54% (138 people) did not have.

3.4. DATA PREPROCESSING

After analyzing and exploring the dataset, it is important to prepare the data as it is a large part of the Data Mining process. This phase allows to find impurities, such as outliers and incoherencies that should be eliminated or cleaned. For this, data pre-processing techniques are used to guarantee the proper condition of data, so it can be correctly used by the future models.

3.4.1. Outlier Detection and Removal

Outliers are observations that are located on the extreme boundaries of the dataset. The goal is to look for these values and handle them, because they can negatively impact the performance and quality of the model. To do so, the Multiplot node was used to check the existence of outliers, and the Filter Node used to exclude them by using a manual filtering method that allows the elimination of the values that represent obvious extreme values, in this case, on three variables: “cholesterol”, “max_ heart_rate” and “oldpeak”. After this process is made, the output shows the exclusion of only 4 outliers, resulting in 299 records which represents 1.3% of the dataset (Appendix 2).

3.4.2. Missing Values

Missing values within data can be problematic for some algorithms, such as regression and neural networks, because it can lead to the weakness of the predictive power. In order to deal with missing values it is used the Impute Node that fills the values that were missing, but as the dataset did not show this problem, there was no need to apply this for this case.

3.4.3. Coherence Checking

In this step, the consistency of the data is checked to make sure the program is dealing with corrected and meaningful values. For this purpose, SAS Code node is used allowing the introduction of new expressions to validate data (table 7).

INCOHERENCE	CODE	DESCRIPTION
Negative Age	Incoherent_age= 0; If (age) <0 then do; Incoherent_age= 1; */delete/* End;	The patient’s age needs to be a positive number

Table 7 - Variable Coherence Checking Description (Source: Author based)

3.4.4. Data Transformation

At this step, SAS Code node is used again to transform variables with the intent of creating new ones that can facilitate and improve the modulation process. For this project, it was decided to turn 3 variables into categories. Table 8 summarizes these transformations:

VARIABLE	CODE	MEANING
age_cod	age_cod=""; IF age>=29 and age<=38 THEN age_cod="1"; ELSE IF age>=39 and age<=48 THEN age_cod="2"; ELSE IF age>=49 and age<=57 THEN age_cod="3"; ELSE IF age>=58 and age<=67 THEN age_cod="4"; ELSE age_cod="5";	Transformation of variable Age into classes of ages
cholesterol_cod	cholesterol_cod=""; IF cholesterol>=126 and cholesterol<=180 THEN cholesterol _cod="1"; IF cholesterol>=181 and cholesterol<=234 THEN cholesterol_cod="2"; IF cholesterol>=235 and cholesterol<=288 THEN cholesterol _cod="3"; IF cholesterol>=289 and cholesterol<=342 THEN cholesterol="4"; ELSE cholesterol="5";	Transformation of variable Cholesterol into classes of cholesterol
max_heart_rate_cod	max_hear_rate_cod=""; IF max_heart_rate>=71 and max_heart_rate<=107 THEN max_hear_rate_cod= "1"; IF max_heart_rate>=108 and max_heart_rate<=131 THEN max_hear_rate_cod= "2"; IF max_heart_rate>=132 and max_heart_rate<=155 THEN max_hear_rate_cod= "3"; IF max_heart_rate>=156 and max_heart_rate<=179 THEN max_hear_rate_cod= "4"; ELSE max_hear_rate_cod="5";	Transformation of variable Maximum Heart Rate into classes of Maximum Heart Rate

Table 8 - Variable Transformation (Source: Author Based)

3.4.5. Data Partition

As part of the data preparation process, it is important to avoid the model overfitting. For that, the primary dataset is divided into a training and validation datasets. The first one covers 70% of the data, and the second one covers 30%. Since there are only 299 observations after the outlier removal, it was decided to not include any data in a test dataset. To divide the dataset, the Data Partition node is used generating the following observations (table 9):

DATASET	PERCENTAGE	NUMBER OF OBSERVATIONS
TRAINING	70%	209
VALIDATION	30%	90
TEST	0%	0

Table 9 - Data Partition Values Description (Source: Author based)

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
target	0	0	135	45.1505	target
target	1	1	164	54.8495	target

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
target	0	0	95	45.4545	target
target	1	1	114	54.5455	target

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
target	0	0	40	44.4444	target
target	1	1	50	55.5556	target

Figure 23 - Summary Statistics for Class Targets (Source: SAS Miner Output, 2020)

By looking at the results of Data Partition node (figure 23), it is assured that the original dataset is split into a training and validation datasets. The training data is used towards the learning and fitting of the model, so within SAS Miner it is defined a percentage for each target value that is in proportion with the original dataset, and where the stratified method is used for the partitioning. The same will be applied to the validation dataset, that will be used to evaluate the model's appropriacy in the Model Comparison node, and will allow in a future step to compare the accomplishment of the different models, and to choose the best one.

3.5. ANALYSIS OF VARIABLES

At this step, the original dataset is already split in two datasets, the training and validation. After partitioning it, the analysis and exploration of this data is done again with the nodes already used - Multiplot, GraphExplore, and StatExplore – which allow an in-depth and ad-hoc exploration of the data and their variables, for instance, their worth. In this section, it is also described the use of the Variable Selection and Correlation Matrix nodes to understand the worthiness and importance of the variables.

3.5.1. Variable Worth

The StatExplore node is important for the analysis of each variable's worth. From the node results, it is seen in the histogram (figure 24) the variables organized by their worth in a descendent order, which allows to understand the worth that each variable assume on the definition of the target. In the case of this project, the ones that most contribute for it and have the highest values are "thal", "chest_pain", "nr_major_vessels" and "oldpeak", and the ones with the lowest contribution are "blood_sugar", "cholesterol_cod", and "electrocardiographic_results", showing the lowest worths. The values of these independent variables are listed below in table 10 by importance and worth.

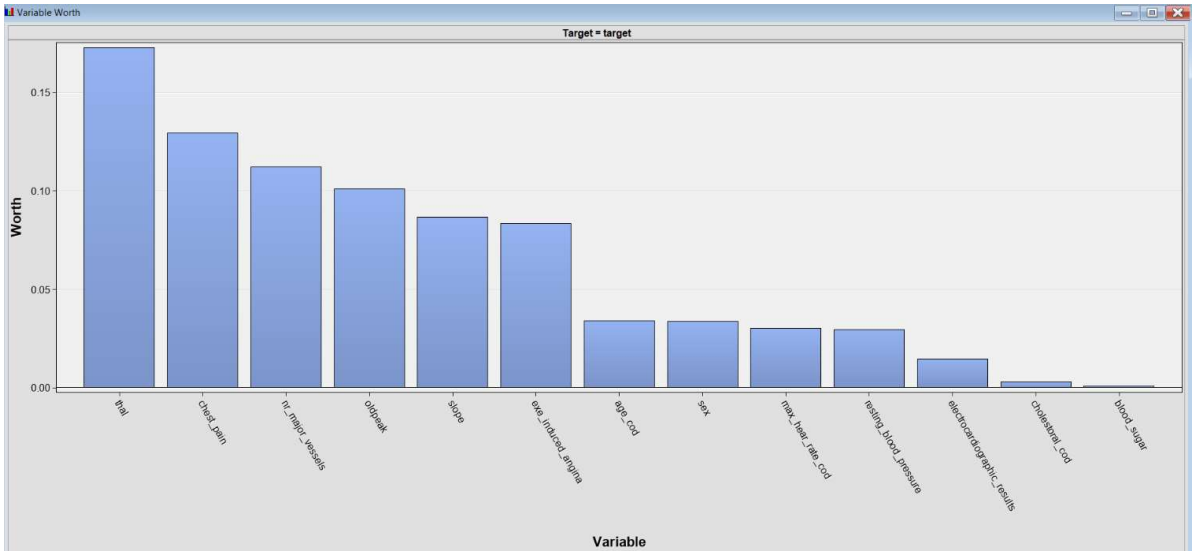


Figure 24 - Variable Worth (Source: SAS Miner Output, 2020)

VARIABLE	IMPORTANCE	WORTH
thal	1	0.1730
chest_pain	2	0.1295
nr_major_vessels	3	0.1123
oldpeak	4	0.1012
slope	5	0.0867
exe_induced_angina	6	0.0833
age_cod	7	0.0341
sex	8	0.0338
max_heart_rate_cod	9	0.0303
resting_blood_pressure	10	0.0295
electrocardiographic_results	11	0.0146
cholestoral_cod	12	0.0030
blood_sugar	13	0.0007

Table 10 - Variable Importance (Source: SAS Miner Output, 2020)

3.5.2. Variable Selection Node

The Variable Selection Node selects variables by looking for the Sequential R-Square. In this case, by using the default parameters of the program, the selection node chooses ten variables and rejects three because of the small values of R-Square. The rejected variables are “blood_sugar”, “max_heart_rate_cod”, and “resting_blood_pressure”.

This is a similar approach to another that will be covered hereafter – stepwise regression – because the R-Square uses a stepwise method of selecting variables as well, stopping when the improvement of adding a variable becomes less than 0.0005. In this case, the method stops at the variable “age_cod” with a sequential R-Square of 0.0005694. The variable selection node output is seen in the histogram (Figure 25), and the Sequential R-Square values are listed in table 11.

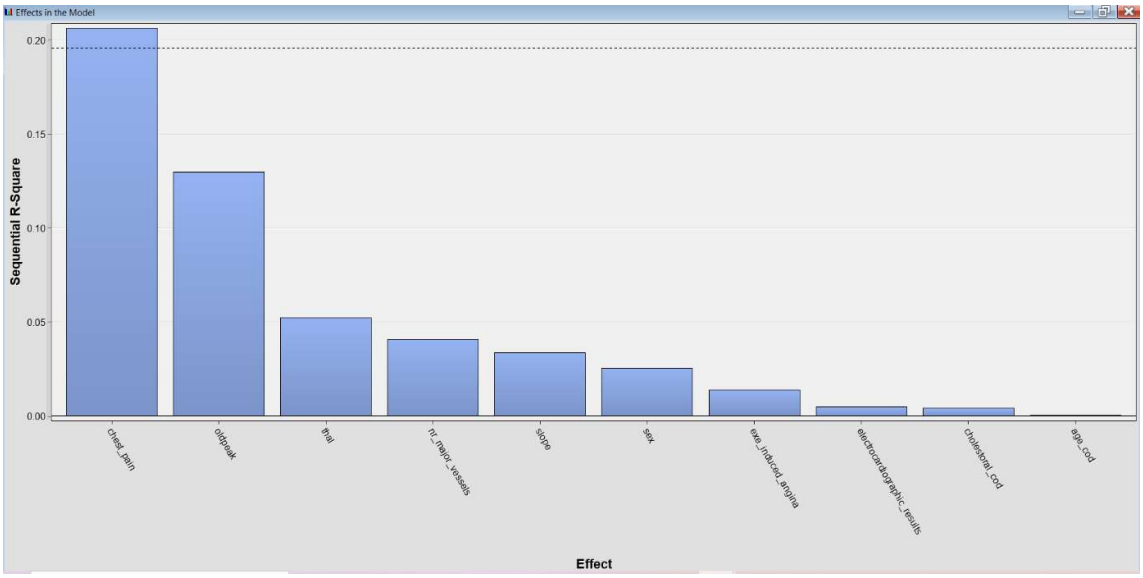


Figure 25 - Variable Selection Node Results (Source: SAS Miner Output, 2020)

VARIABLE	R-SQUARE
chest_pain	0.206529
oldpeak	0.129893
thal	0.052286
nr_major_vessels	0.040793
slope	0.033941
sex	0.025431
exe_induced_angina	0.014031
electocardiographic_results	0.00494
cholestoral_cod	0.004206
age_cod	0.000569

Table 11 - Sequential R-Square Values (Source: SAS Miner Output, 2020)

3.5.3. Correlation Matrix Analysis

A dimensionality analysis is done to measure the individual value of each variable and its impact in the context of the problem. As covered before, variables with high correlation give the same type of information, being unnecessary to use both in the model. Usually, it is opted to keep the variable with the highest worth, wherefore, in the presence of two variables highly correlated, the one with the lowest worth will be excluded.

In order to look for these values, it is analysed the Correlation Matrix where it is visualized in a cross-tabular format different correlations, specifically, the Pearson correlation. This value differs from -1 to +1 and measures the robustness of a linear association between two interval variables, where the correlation goal is to create a line that represents the best fit of the data variables. If the correlation value is 0, it means that the two variables do not have an association, if the value is higher than 0 it means that the variables have a positive association, and if the correlation value is lower than 0 it means that the variables are in the presence of a negative association. The strength of the variables' association will be higher when the Pearson coefficient is closer to +1 in case of a positive relationship, or closer to -1 for a negative relationship.

The Correlation Matrix (Appendix 3) shows the Pearson correlation values of the variables. Table 12 summarizes the Pearson correlations that are higher than 0.1, which is already a method to exclude the correlations with the lowest values.

VARIABLE 1	VARIABLE 2	PERSON CORRELATION
nr_major_vessels	Thal	0.1025
nr_major_vessels	Resting_blood_pressure	0.1067
cholesterol_cod	Chest_pain	0.1179
max_heart_rate_cod	age_cod	0.1592
oldpeak	Resting_blood_pressure	0.1666
thal	Oldpeak	0.2232
oldpeak	nr_major_vessels	0.2331
oldpeak	age_cod	0.2423
nr_major_vessels	age_cod	0.2620
resting_blood_pressure	age_cod	0.2765
max_heart_rate_cod	oldpeak	0.3471

Table 12 – Variables Pearson Correlation (Source: SAS Miner Output, 2020)

The most relevant correlation is between “max_heart_rate_cod” and “oldpeak” (0.35). The “oldpeak” variable has a higher worth (0.1012) than the “max_heart_rate_cod” (0.0303), meaning it is the most relevant for the definition of the target variable. Nevertheless, “max_heart_rate_cod” was not excluded as the Pearson correlation is less than 0.8 and there is no redundancy between both variables.

3.6. MODELING

After the pre-processing, data partition, and variable analysis, there is finally the phase where the modulation process is done. The modeling phase is where different models are created and applied, so given a group of predictors - independent variables - it can predict the value of the target and understand which one has the best behavior. Thus, several classifications are done regarding a set of observations to select the best model that can predict, in this case, patients that will develop heart attacks. To achieve the results, different algorithms with different ways of inputting data and calculating the output are used to test the problem with different complexities. It was applied a total of twelve models: seven different configurations of Artificial Neural Networks, three types of Logistic Regression, one Decision Tree, and an Ensemble Model (Table 13):

MODEL	OBSERVATIONS
ANN 1	MLP – 1 Hidden Unit
ANN 2	MLP – 2 Hidden Units
ANN 3	MLP – 3 Hidden Units
ANN 4	MLP – 4 Hidden Units
ANN 5	MLP – 5 Hidden Units
ANN 6	MLP – 6 Hidden Units
ANN 7	MLP – 7 Hidden Units
Regression (Forward)	Type – Logistic / Model Selection - Forward
Regression (Stepwise)	Type – Logistic / Model Selection - Stepwise
Regression (Backwards)	Type – Logistic / Model Selection - Backwards
Decision Tree	Ordinal Criterion – Entropy
Ensemble	Class Target – Voting

Table 13 - Predictive Algorithms Used (Source: Author based)

3.6.1. Artificial Neural Network

ANN is the first method because of its robustness and good results achieved throughout the history in different scenarios. They are settled on the MLP model, having each one a hidden layer with a different number of neurons. As covered before, the idea of ANN is that they can learn through the provided data so they can develop an automatic learning and reach conclusions in the future when faced with new cases.

MLP can have one or more units in the hidden layer, so the complexity of the neural network will be dependent on it: the higher the number of neurons, the more complex it will be. Accordingly, it makes sense to analyze the networks with different neurons starting by only 1 neuron in the hidden layer, and after, expand it to evaluate the network's performance. As the number of neurons increases, it is expected that the model gives worst results because of its complexity, so it was decided to test 7 different networks, starting with 1 neuron to a maximum of 7 neurons in the hidden layer.

3.6.2. Regression

As covered before, logistic regression is a useful algorithm applied to predict dependent variables that are limited to a binary response (1 or 0). There will be tested three different types of regression named, forward, backward, and stepwise. The forward regression starts with zero variables and proceeds adding variables at a time according to their worth; the backward regression considers each variable to be eliminated, starting with a multiple regression model, and then eliminating non-important variables along the way; and the stepwise regression is a mixture of the two, it starts with zero variables and every time it adds one variable, it can delete another, meaning it adds and deletes at each step of the process, reinforcing the regression node in comparison with the others (Shtatland, Kleinman, & Cain, 2008).

The results of the forward, stepwise, and backward regression nodes are presented in tables 14, 15 and 16, respectively.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	chest_pain	1	1	43.1646	<.0001
2	oldpeak	1	2	32.0881	<.0001
3	nr_major_vessels	1	3	16.8661	<.0001
4	thal	1	4	17.1161	<.0001
5	slope	1	5	11.2089	0.0008
6	sex	1	6	8.1440	0.0043
7	exe_induced_angina	1	7	4.0061	0.0453

Table 14 - Forward Regression Results (Source: SAS Miner Output, 2020)

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	chest_pain		1	1	43.1646		<.0001
2	oldpeak		1	2	32.0881		<.0001
3	nr_major_vessels		1	3	16.8661		<.0001
4	thal		1	4	17.1161		<.0001
5	slope		1	5	11.2089		0.0008
6	sex		1	6	8.1440		0.0043
7	exe_induced_angina		1	7	4.0061		0.0453
8		oldpeak	1	6		3.6297	0.0568

Table 15 - Stepwise Regression Results (Source: SAS Miner Output, 2020)

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	max_hear_rate_cod	1	12	0.0494	0.8242
2	blood_sugar	1	11	0.2035	0.6519
3	resting_blood_pressure	1	10	0.2535	0.6146
4	age_cod	1	9	0.7027	0.4019
5	cholestoral_cod	1	8	2.0853	0.1487
6	electrocardiographic_results	1	7	2.9070	0.0882
7	oldpeak	1	6	3.6297	0.0568

Table 16 - Backward Regression Results (Source: SAS Miner Output, 2020)

The variables selected by the forward and stepwise regression nodes were almost the same, except for one variable, “oldpeak”, that was added and then removed in the stepwise node. For the backward regression node, there were eliminated seven variables: “max_heart_rate_cod”, “blood_sugar”, “resting_blood_pressure”, “age_cod”, “cholesterol_cod”, “electrocardiographic_results”, “oldpeak”.

3.6.3. Decision Tree

The third predictive model used is the Decision Tree, which is based on a group of Boolean branch rules that test different facts and construct different link paths in a hierarchical way to determine conclusions, such as the class of a specific instance. The DT used has a combination of (2,6) that represents the maximum branch and depth, respectively. The maximum branch represents the split between the branches, and the maximum depth refers to the maximum growth of the tree in the vertical. This type of model has an easy interpretation, and the most important variables appear at the most top of the tree (Appendix 4).

3.6.4. Ensemble

The final applied model is the Ensemble. This methodology, as referred before, is used to combine different classifiers, at least 2 or more, so that a new model with a more robust prediction and accuracy is created (Maldonado, Dean, Czika & Haller, 2014). In this case, it was used a selection method named "average value" for interval variables, and another one named "voting" for class variables. The models that were chosen for the Ensemble node were based on two metrics: the ROC Index, which should have the highest value, and the Misclassification Rate, which should have the lowest value. By looking for the validation ROC Index values, the best top two models were the ANN1 and Forward Regression, with the values 0.893 and 0.89, respectively. For the valid Misclassification Rate the best top two were ANN5 and ANN1, with the values of 0.14 and 0.16, respectively. After selecting them, this approach is concluded by selecting the three models for the ensemble node, which are ANN1, ANN5 and Forward Regression.

4. ASSESSMENT: RESULTS AND DISCUSSION

After the definition of the necessary steps to develop a project about predictive modelling and its successful application, it is necessary to look at the results. In this chapter, there are presented the results of the practical part of the project, the predictive power of the algorithms used, and the comparison and final choice of the best model, which aims to meet the initial problem and show a good performance on predicting whether a person develops a heart attack or not.

4.1. MODEL COMPARISON

Having the different algorithms applied, Model Comparison is the final step where the purpose is to determine which model has the best performance while predicting, by applying the Model Comparison node. The analysis and final choice may be based on different metrics such as the ROC curve, Misclassification Rate, Lift, Gain, Akaike Criteria, Bayesian Criteria and Kolmogorov-Smirnov (Dean, 2014). For this project, two metrics were used for the model selection: the (1) ROC Curve and the (2) Misclassification Rate. The first one stands for Receiver Operating Characteristics Curve and displays graphically two axes that range from 0 to 1: the Y axis, that represents the sensitivity or TPR, and the X axis that represents the 1 – specificity or FPR. Those delimit the dispersion of test value points in a bi-dimensional chart, along with the representation of the ROC curve between the X and Y axes. The AUC, Area Under the Curve, is considered an efficient method to measure the sensitivity and specificity, and to assess the test validity (Indrayan & Kumar, 2011). The AUC resumes the ROC curve position, and hence the best model, being used for binary classification problems. The curve that is most located to the upper left side will be the one with the best capacity to differentiate and diagnose tests. For instance, in figure 26 it is shown that test B has a better capacity to discriminate when compared to test A, as it is the closest to the upper left corner (Tilaki, 2013).

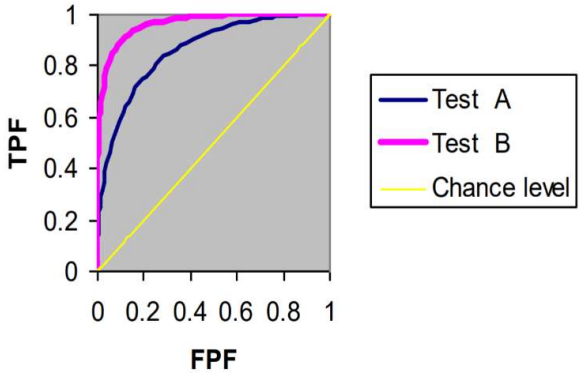


Figure 26 – Example of two diagnostic tests: ROC Curves from Test A and Test B. Adapted from *Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation,* by K.H. Tilaki, 2013, Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824. Copyright 2013 by K.H. Tilaki.

A maximum AUC, meaning it is equal to 1, indicates the diagnostic test is ideal when differentiating some subjects, such as disease and non-disease topics, as a value adjacent to 1 reflects a greater test performance. On the other hand, the minimum AUC value is pondered 0.5 because a value of 0 means it tests incorrectly all objects: the ones with disease will be classified as negative, and the non-disease will be classified as positive. The final goal is that the chosen model reflects the greatest area under the ROC curve (Indrayan & Kumar, 2011).

During the last forty years, the analysis of the ROC curve became a well-liked approach to examine and evaluate the accuracy of medical systems used in healthcare, which is considered an appropriate metric to be used in this project. The estimation of a specific test accuracy, like the AUC, influences its capacity to differentiate specific cases, such as diseased and non-disease citizens (Tilaki, 2013), as it estimates the quality of the model's predictions within different limits. The general meaning of AUC refers to the probability of a true positive being precisely classified, as well as with a true negative, being correctly classified as negative (Mysiak, 2020).

From the output of Model Comparison node, it is seen the ROC Curves of the different tested models in the ROC Chart, where each curve represents a different classifier (figure 27).

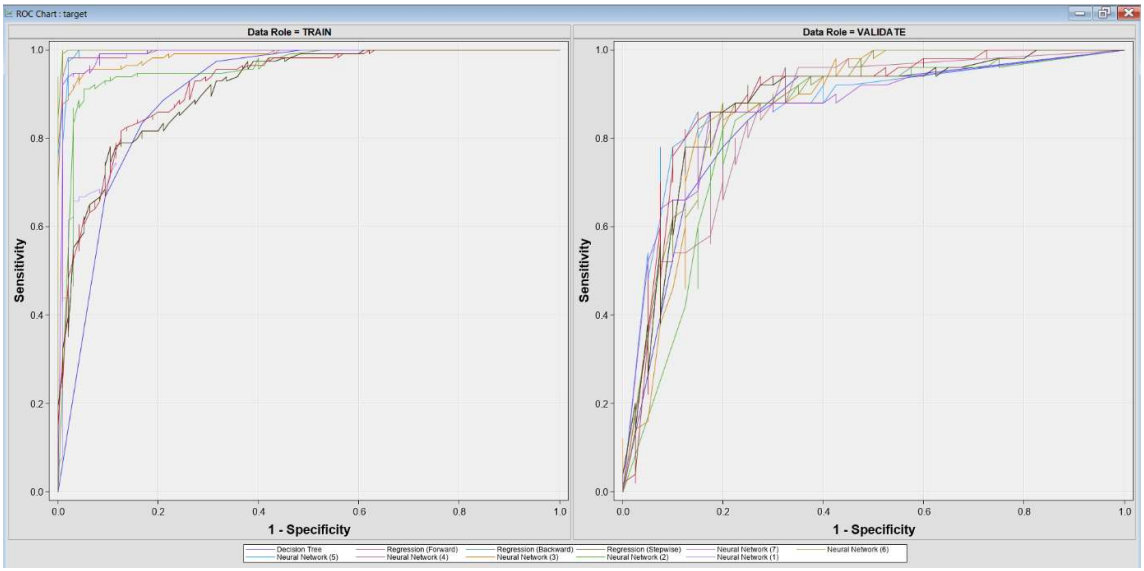


Figure 27 - ROC Chart Model's Output (Source: SAS Miner Output, 2020)

Once there are used different predictive models, a graphical analysis may become too complex. An alternative consists of using directly the ROC Index which represents the AUC. For this, the model with the highest index will be the one with the best performance, as curves that are on the most top left side corner represent better models. The higher the degree of concavity, the higher the AUC will be, and the better the model performance is expected to be.

Table 17 summarizes the ROC Index values of the used models, both for training and validation:

MODEL	TRAIN: ROC INDEX	VALID: ROC INDEX
ANN 5	0.996	0.887
ANN 1	0.923	0.893
Ensemble	0.98	0.888
Regression - Stepwise	0.912	0.873
Regression - Backward	0.912	0.873
ANN 3	0.987	0.871
ANN 7	0.989	0.862
Regression – Forward	0.917	0.89
Decision Tree	0.903	0.849
ANN 6	0.999	0.877
ANN 2	0.955	0.84
ANN 4	0.995	0.858

Table 17 – ROC Index Performance of Used Models (Source: SAS Miner Output, 2020)

ROC Index was the first metric used, but it is also important to accompany it and complement the analysis with another one. The second metric referred and used for the choice of the best model is the Misclassification Rate, which represents the total of false positives and negatives classified in the total of instances. This metric is used to measure the effectiveness of a DM model, and it is interpreted as the percentage of training and testing examples misclassified from a given dataset (Baughman & Liu, 2014). The goal is that the best model achieves the minimum possible value for this statistic, as it will indicate that the obtained model is accurate (Khoshgoftaar, Yuan, & Allen, 2000). In fact, if the value of this statistic is high, in general the generated model is not useful, and the predictions will be inefficient. The Misclassification Rate values for the training and validation sets are detailed in table 18:

Model	Train: Misclassification rate	Valid: Misclassification rate
ANN 5	0.0191	0.1444
ANN 1	0.1340	0.1556
Ensemble	0.0813	0.1667
Regression - Stepwise	0.1914	0.1667
Regression - Backward	0.1914	0.1667
ANN 3	0.0526	0.1667
ANN 7	0.0478	0.1667
Regression - Forward	0.1770	0.1778
Decision Tree	0.1579	0.1889
ANN 6	0.0048	0.1889
ANN 2	0.0813	0.1889
ANN 4	0.0191	0.2

Table 18 – Misclassification Values of Used Models (Source: SAS Miner Output, 2020)

4.2. Choice of the Best Model

After the modulation process is done, and having the selected metrics defined and their values compared, it is required the choice of the model with the best results. As covered before, a good model will have a high ROC Index value, which corresponds to a bigger area under the curve, and a low value for the Misclassification Rate, representing low misclassified training and testing examples. Having this in mind, the best model chose is the Artificial Neural Network with five hidden units (ANN5), with a ROC Index value of 0.996 for the train and 0.887 for the validation (table 17). For the Misclassification Rate, ANN5 shows a value of 0.0191 for the train and 0.1444 for the validation, the lowest of all models (table 18). Hence, within all the tested models, ANN5 becomes the model that will lead to more satisfactory results and with a better predictive performance for the problem in question.

For the development of the practical part of the project and to reach the final results, different steps were followed, starting by the dataset analysis, continuing with data pre-processing steps, and then, applying the modelling phase with a total of twelve predictive models that were tested: seven different types of Artificial Neural Networks, one Decision Tree, three different types of Logistic Regression and one Ensemble. At the modulation step, it was decided to use for each model all the total thirteen independent variables of the transformed dataset: “thal”, “chest_pain”, “nr_major_vessels”, “oldpeak”, “slope”, “exe_induced_angina”, “age_cod”, “sex”, “max_heart_rate_cod”, “resting_blood_pressure”, “electrocardiographic_results”, “cholestoral_cod” and “blood_sugar”.

After the variable analysis and the application of the models, it is possible to understand and conclude which variables most contribute for the prediction of heart attacks. From a group of thirteen predictors, six of them stood out, which are “thal”, “chest_pain”, “nr_major_vessels”, “oldpeak”, “slope”, and “exe_induced_agina”. This choice is due to their high worth from the analysis of StatExplore node, which translates on the meaning and contribution they have on defining the target; to their R-Square values from the Variable Selection node where some variables are rejected and others accepted; to the exploration of the Pearson Correlation values, and finally, due to the variable selection in the Forward, Stepwise and Backward Regression nodes. Despite the identified correlations, some variables were opted to keep as their correlation values were under 0.8 and they did not add redundancy to the model.

Regarding the results of the remaining applied models, it was also analysed their contribution: it is shown that ANN2 and ANN4 are the models with the lowest performance: ANN2 has the lowest valid ROC Index, 0.84, and the second highest valid Misclassification rate, 0.1889. The ANN4 has the highest valid Misclassification Rate, 0.2, and a valid ROC Index of 0.858, one of the lowest. After ANN5, the best second model is ANN1 with a valid Misclassification Rate of 0.1556 and a valid ROC Index of 0.893.

The Stepwise, Backward and Forward Regression showed to give better results than the Decision Tree: the first two have the same values of 0.1667 for the valid Misclassification Rate and 0.873 for valid ROC Index, and the Forward Regression has 0.1778 for the valid Misclassification Rate and 0.89 for valid ROC Index. For the decision tree, the model shows a value of 0.1889 and 0.849, respectively.

Considering the ensemble that was formed with ANN1, ANN5, and Forward Regression, it showed to give better results (valid Misclassification Rate of 0.1667 and a valid ROC Index of 0.888) than the remaining models, which include most of the Artificial Neural Networks, the Decision Tree, and the three Regressions.

In general, most of the models showed satisfactory results having valid ROC Indexes equal or higher than 0.84, and Misclassification Rates mostly lower than 0.2.

5. CONCLUSIONS

The use of Data Mining is valuable and extremely important on different fields, from business management, to healthcare, and preventive medicine, as it gives a great contribution for the analysis of high volumes of data, for the discovery of important relationships and patterns, for the understanding of the impact that objects and variables have, and for the support of important decisions that may change people's lives. Considering the numbers of heart attacks happening nowadays and worldwide deaths, it is of great interest and importance to make use of Information Technology, Data Mining, Machine Learning techniques, and methodologies, to improve the knowledge that doctors and medical staff have about these diseases, in order to detect at early stages symptoms of possible complications or diseases that might happen on patients, so future problems can be warned and prevented, and in more serious cases, new deaths.

The present work is based on a real dataset retrieved from UCI Machine Learning Repository, being composed by 303 records that represent significant health indicators, the predictors, used for the analysis of the target, which is the occurrence of heart attacks. The project has different parts and consists of analysing the dataset and using DM algorithms to create and develop predictive models that can predict whose patients may develop heart attacks in the future, and posteriorly, helping on the prevention of this serious problem in real life situations.

In the literature review it is possible to understand the general overview about the DM meaning, its applications, and its types of algorithms, as well as its use in the healthcare sector, medicine, and on the project's topic – occurrence of heart attacks. In this topic and within the chapter, there are explored some studies that already support the developments done with DM and technologies over the years in healthcare and medicine field, mainly on the prediction of serious diseases, such as cancer, diabetes, and on the project's case, cardiovascular diseases.

The development of the theoretical part is proceeded by the practical one that was done in different phases: the understanding of the problem, mainly done with different types of literature and the understanding of the heart disease topic, followed by data analysis, data pre-processing techniques, implementation of the predictive models, the analysis of their results, and final conclusions. These steps were done with the help of software *SAS Enterprise Miner 15.1*, which allowed the transformation of variables and the dataset cleaning, and in a next phase, the implementation of the modulation process where different algorithms were applied for the prediction of heart attacks. For this project, it was opted to use seven types of Artificial Neural Networks, one Decision Tree, three different types of Logistic Regression, and one Ensemble, so that different models, complexities, and performances could be tested.

Lastly, the models' results were assessed and compared, mainly using two metrics: the ROC Curve (ROC Index) and the Misclassification Rate. From a technical point of view, the model with the best metric values, and hence with the best predictive performance was chose, being the Artificial Neural Network with 5 hidden units the final choice. Having the variables analysed, the models compared, and the final decision, it was also possible to identify and conclude which variables most contribute for the occurrence of heart attacks, due to the statistical analysis. The final group is composed by six variables that represent important values or indicator levels of a patient's health, and those are: (1) Thalassemia (Thal), which is a genetic blood disorder responsible for the lower levels of haemoglobin in the body, influencing the way the red cells carry oxygen throughout the body and the functioning of the system, which may cause fatigue and weakness; (2) Chest pain, which is labelled with distinct values regarding the level of pain. Those are 0- Typical angina, 1- Atypical angina, 2- Non-anginal pain and 3- Asymptomatic pain; (3) the Number of Major Vessels coloured by the method of fluoroscopy, which is a procedure done in cardiac catheterization that aims to see the flow of blood through the coronary arteries, and to check for eventual arterial blockages; (4) the Oldpeak, which is the continuous value for ST depression in the treadmill electrocardiogram, caused by exercise related to rest; (5) Slope, meaning the peak exercise in ST segment and that can be classified in upsloping, flat sloping or down sloping; and finally, (6) if the exercise induced angina, which is a typical pain in the chest originated by the low circulation of the blood to the heart.

Different studies have shown a number of factors that may enlarge the risk of developing heart attacks, mainly the family history of cardiovascular diseases, the existence of a high fat diet, obesity, lack of exercise, hypertension, cholesterol, (Kumari & Godara 2011), chest pain type, feature of the body vessels, blood pressure and other severe illness (Alzahani et al., 2014). These and other external factors are important for the analysis and influence of cardiovascular diseases, and they are also supported in the literature review. It is assured that the prevention of these diseases starts in the self-conscious of each individual by modifying one's lifestyle to a healthier one, mainly with a better stress management and physical activity, which will eventually contribute to the reduction of obesity and hypertension, and consequently, to lower the risk of cardiovascular diseases (Mohsenipouya, Majlessi, Shojaeizadeh, Foroushani, Ghafari, Habibi, & Makrani, 2016).

Nevertheless, the use and study of the specific group of variables applied in this project allowed to understand which ones most contribute for heart attacks despite the other external factors. As referred before, these variables are important predictors that reflect the status of a patient's health, and consequently, by understanding their worth and impact for the problem in question, it is possible to make easier predictions about the occurrence of the problem based on their values. From a total of thirteen variables of the dataset, six of them showed to have more importance, thereby, looking for symptoms of thalassemia, possible chest pains, several major vessels coloured by fluoroscopy method,

possible angina pains caused by exercise, and examining the values of ST depression and slope, are ways of preventing and predicting heart attacks. This final group of predictors are previously described, and they are also supported and explored by other studies on the field. For instance, in one case it was proposed an interpretable fuzzy rule-based system that could predict the CAD only based on the age, the levels of angina caused by exercise, the group of major vessels coloured by the fluoroscopy method, the thallium stress results and the slope, which showed to be promising to diagnose CAD and a credible system to be administered as a support for diagnostic decisions. Another diagnosis for CAD was made by examining the stenosis of each vessel separately, where chest pain, angina and ST elevation/depression were some of the biggest factors for the stenosis of the LAD vessel (Alizadehsani et al., 2012). It is also registered that the coronary artery stenosis has a remarkable relationship with abnormal levels of cholesterol, hypertension, and stress caused by exercise (Mohsenipouya et al., 2016).

Not only the awareness for these variables and factors is important, but also, alerting for the problem with educational interventions is also critical to change the consciousness of the individuals with risk of having cardiovascular diseases. It is indispensable that each one gains responsibility for his/her health, and that people focus on the benefits of physical activity, on a healthy lifestyle, and on the importance of stress levels management. Therefore, it is a top priority to develop and implement educational programs focusing on health, in order to raise awareness for the literacy and knowledge that people have on diseases and their risks, and to influence and inspire them on the adoption of a healthy lifestyle (Mohsenipouya et al., 2016).

Although the results of the models covered in the literature review studies and in this project do not have 100% of performance, they have shown that DM techniques are strongly encouraging and achieve promising results (showing high classification accuracies of at least 77% or higher) that can be positively used to complement and assist people while using heart disease prediction systems, which will enable more efficient and corrected medical decisions, early diagnosis, and more suitable treatments.

The number of records of the dataset is small, nevertheless, resorting to these DM techniques on the sample of people used in this project has shown to have a good predictive capacity, which enables the establishment of patterns and profiles on patients, for instance, the long-term evaluation of their chest pain, blood pressure or other indicators, that should be under control. This capacity applied to higher volumes of medicine data and the use of DM methods will have a direct influence on the analysis of the patients' health, which may indicate at an early stage if a person is at risk of developing an heart attack. These approaches would allow medical innovation, early action and quicker answers, more life quality for patients, and the reduction on the number of deaths due to heart attacks.

This work project pretends to be a contribution for the Data Mining field and their researchers. It is also aimed to give the importance and show the impact that Information Technology and Data Mining techniques can have on the medical field, mainly to support the development of new medical techniques, to easily keep up with the patients' health, to do early diagnose of diseases, to define the best suitable treatments on the right time, and most importantly, to save lives.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORK

DM is a complex and important analysis process for huge numbers of data applied in distinct sectors, which allows the detection and discovery of important patterns that were previously unknown, contributing for the necessary decision making. However, it has some limitations starting by the quality and veracity of the data being used, as well as the techniques and tools. Data should always be treated, cleaned, and in proper conditions so it can be used successfully, as its preparation represents about 75% of the necessary resources of a DM project.

Regarding the improvements of the work, some aspects should be taken into consideration. It could be included new transformed variables and other ones that were not analysed, and that may add more information and improve the predictive power. For instance, the number of hours that people exercise during the week, alcohol intake, tobacco consumption, and diet type, which have a high influence on people's health and consequent development of diseases and illnesses. The number of records could also be increased to have more distinct samples of people and their respective data about health cardiac values, which would enrich the dataset.

Regarding the performance of the model, it can also be increased and better explored by introducing and testing other machine learning and deep learning algorithms, for example, Support Vector Machine, K-Means, KNN, or Convolutional Neural Networks.

The focus of the project was limited to the data of UCI Machine Learning Repository because of the availability and quantity of data, but nevertheless, it was sufficient to analyze, understand and put in practice the prediction of the occurrence of heart attacks.

7. BIBLIOGRAPHY

Aggarwal, C. C. (2015). *Data mining: The Textbook*. New York: Springer International Publishing.

Alizadehsani, R., Habibi, J., Bahadorian, B., Mashayekhi, H., Ghandeharioun, A., Boghrati, R., & Sani, Z. A. (2012). Diagnosis of coronary arteries stenosis using data mining. *Journal of medical signals and sensors, 2(3),* 153.

Alzahani, S. M., Althopity, A., Alghamdi, A., Alshehri, B., & Aljuaid, S. (2014). An overview of data mining techniques applied for heart disease diagnosis and prediction. *Lecture Notes on Information Theory, 2*(4).

Amin, S. U., Agarwal, K., & Beg, R. (2013). Genetic neural network-based data mining in prediction of heart disease using risk factors. In *2013 IEEE Conference on Information & Communication Technologies* (pp. 1227-1231). IEEE.

Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, *2*(10), 5370-5376.

Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future generation computer systems*, *13*(2-3), 197-210.

Augusty, S. M., & Izudheen, S. (2013). A survey: evaluation of ensemble classifiers and data level methods to deal with imbalanced data problem in protein-protein interactions. *Review of Bioinformatics and Biometrics*, *2*(1), 1-9.

Avoiding heart attacks and strokes: don't be a victim-protect yourself. World Health Organization (WHO), World Self Medication Industry (WSMI), World Heart Federation (WHF), and International Stroke Society (ISS), 2005. Retrieved January 11, 2020, from: http://library.health.go.ug/publications/heart-disease/avoiding-heart-attacks-and-strokes-don%E2%80%99t-be-victim-protect-yourself.

Azevedo, A. I. R. L. & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Babaoglu, I., Baykan, O. K., Aygul, N., Ozdemir, K., & Bayrak, M. (2009). Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization. Expert Systems with Applications, 36(2), 2562-2566.

Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M. (2006, June). A comparative study for outlier detection techniques in data mining. In *2006 IEEE conference on cybernetics and intelligent systems* (pp. 1-6). IEEE.

Baughman, D. R., & Liu, Y. A. (2014). Neural networks in bioprocessing and chemical engineering. Academic press.

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Brameier, M., & Banzhaf, W. (2001). A comparison of linear genetic programming and neural networks in medical data mining. *IEEE Transactions on Evolutionary Computation*, *5*(1), 17-26.

Breiman, L., Friedman J. H., Olshen R. A., Stone C.J. (1984). *Classification and regression trees. Belmont, California: Wadsworth International Group*.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brownlee, J. (2017). Difference between classification and regression in machine learning. *Machine Learning Mastery*, *25*.

Bühlmann, P. (2012). Bagging, boosting and ensemble methods. In *Handbook of Computational Statistics* (pp. 985-1 022). Springer, Berlin, Heidelberg.

Cardiovascular Diseases. (n.d.) World Health Organization. [online] Retrieved January 11, 2020, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

Cardiovascular Diseases. (CVDs)-Key Facts (n.d.) World Health Organization. [online] Retrieved January 11, 2020, from https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

Cardiovascular Health for Everyone. (n.d.) World Heart Federation. [online] Retrieved January 11, 2020, from https://www.world-heart-federation.org/.

Castle, N. (2017). Supervised vs. Unsupervised Machine Learning. [online] In *Oracle AI & Data Science Blog.* Retrieved February 12, 2020, from: https://blogs.oracle.com/datascience/supervised-vs-unsupervised-machine-learning.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, *9*, 13.

Chavan, A. (2019). A Comprehensive Guide to Decision Tree Learning. [online] In *AI Time Journal*. Retrieved February 12, 2020, from: https://www.aitimejournal.com/@akshay.chavan/a-comprehensive-guide-to-decision-tree-learning.

Chitra, K., & Subashini, B. (2013). Data mining techniques and its applications in banking sector. *International Journal of Emerging Technology and Advanced Engineering*, *3*(8), 219-226.

Davis, D. A., Chawla, N. V., Blumm, N., Christakis, N., & Barabasi, A. L. (2008, October). Predicting individual disease risk based on medical history. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 769-778).

Dean, J. (2014). Big data, data mining, and machine learning: value creation for business leaders and practitioners. John Wiley & Sons.

Deaths due to CHD in the EU. (2020) [online] In *Eurostat*. Retrieved January 11, 2020, from Eurostat: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20200928-1.

Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends in signal processing*, *7*(3–4), 197-387.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, 64(5), 304-310.

Edwards, J. (2015*).* What is predictive analytics? Transforming data into useful insights. [online] In *CIO.* Retrieved February 12, 2020, from: https://www.cio.com/article/3273114/what-is-predictive-analytics-transforming-data-into-future-insights.html.

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press.

Gardner, M. W. & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627-2636.

Garg, R. (2018). A Primer to Ensemble Learning–Bagging and Boosting. [online] In *Analytics India Magazine*. Retrieved February 12, 2020, from: https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/.

Garrido, A. P. (2016). What is the difference between Bagging and Boosting? [online] In *QuantDare.* Retrieved February 12, 2020, from: https://quantdare.com/what-is-the-difference-between-bagging-and-boosting.

Grover, P. (2017). Gradient Boosting from scratch. [online] In *Medium*. Retrieved February 12, 2020, from: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques, Waltham, MA. *Morgan Kaufman Publishers*, *10*, 978-1.

Heart Disease Statistics and Maps (n.d.) Centers for Disease Control and Prevention. [online] Retrieved January 11, 2020, from https://www.cdc.gov/heartdisease/facts.htm.

Hoare, J. (2017). Machine Learning: Pruning Decision Trees. [online] In *Display Rblog.* Retrieved February 12, 2020, from: https://www.displayr.com/machine-learning-pruning-decision-trees/.

*Healthcare: Dataset on Heart attack possibility (*2020). [online] In *Kaggle.* Retrieved October 14, 2019, from: https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility.

Indrayan, A. & Kumar, R. (2011). *Receiver operating characteristic (ROC) curve for medical researchers.* [online] In *Springer Link.* Retrieved February 5, 2020, from https://link.springer.com/article/10.1007/s13312-011-0055-4.

Khan, S. N., Nawi, N. M., Shahzad, A., Ullah, A., Mushtaq, M. F., Mir, J., & Aamir, M. (2017). Comparative analysis for heart disease prediction. JOIV: International Journal on Informatics Visualization, 1(4-2), 227-231.

Khoshgoftaar, T. M., Yuan, X., & Allen, E. B. (2000). Balancing misclassification rates in classification-tree models of software quality. Empirical Software Engineering, 5(4), 313-330.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). *ImageNet Classification with Deep Convolutional Neural Networks.*

Kulkarni, M. (2017). *Decision trees for classification: A machine learning algorithm*. [online] In *Xoriant Blog*. Retrieved February 12, 2020, from: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html.

Kumari, M., & Godara, S. (2011). Comparative study of data mining classification methods in cardiovascular disease prediction.

Larose, D. T. (2015). *Data mining and predictive analytics*. John Wiley & Sons.

LaValle, S., Hopkins, M. S., Lesser, E., Shockley, R., & Kruschwitz, N. (2010). Analytics: The new path to value. *MIT sloan management review*, *52*(1), 1-25.

Lavrač, N. (1999). Selected techniques for data mining in medicine. Artificial intelligence in medicine, 16(1), 3-23.

Lee, I. N., Liao, S. C., & Embrechts, M. (2000). Data mining techniques applied to medical information. *Medical informatics and the Internet in medicine*, *25*(2), 81-102.

Macukow, B. (2016). Neural networks–state of art, brief history, basic models and architecture. In *IFIP international conference on computer information systems and industrial management* (pp. 3-14). Springer, Cham.

Maheshwari, A. (2015). *Business intelligence and Data Mining*. New York: Business Expert Press.

Maldonado, M., Dean, J., Czika, W., & Haller, S. (2014). Leveraging ensemble models in SAS® Enterprise Miner™. In *Proceedings of the SAS Global Forum 2014 Conference*.

Mangale, S. (2017). *Decision Tree-Pruning-Cost Complexity Method.* [online] In *Medium*. Retrieved February 12, 2020, from: https://medium.com/@sanchitamangale12/decision-tree-pruning-cost-complexity-method-194666a5dd2f.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

Milanović, M., & Stamenković, M. (2016). CHAID decision tree: Methodological frame and application. *Economic Themes*, *54*(4), 563-586.

Mohsenipouya, H., Majlessi, F., Shojaeizadeh, D., Foroushani, A. R., Ghafari, R., Habibi, V., & Makrani, A. S. (2016). Health-related variables and predictors of Health-promoting Lifestyle in cardiovascular disease patients. Electronic physician, 8(4), 2274.

Mysiak, K. (2020). Classification Metrics & Thresholds Explained. *Demystifying commonly used classification metrics.* [online] In *Towards Data science.* Retrieved March 13, 2020, from: https://towardsdatascience.com/classification-metrics-thresholds-explained-caff18ad2747.

Napgal, A. (2017). *Decision Tree Ensembles- Bagging and Boosting.* [online] In *Towards Data science*. Retrieved February 12, 2020, from: https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9.

Navlani, A. (2018). Understanding random forests classifiers in Python. [online] In *Datacamp.* Retrieved February 17, 2020, from: https://www.datacamp.com/community/tutorials/random-forests-classifier-python.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

Nonchev, B. S. (2015). Model Selection for Data Analysis Based on the MDL Principle.

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE.

Parthiban, G., Rajesh, A., & Srivatsa, S. K. (2011). Diagnosis of heart disease for diabetic patients using naive bayes method. International Journal of Computer Applications, 24(3), 7-11.

Patel, H. H., & Prajapati, P. (2018). Study and Analysis of Decision Tree Based Classification Algorithms.

Patel, N., & Upadhyay, S. (2012). Study of various decision tree pruning methods with their empirical comparison in WEKA. *International journal of computer applications*, *60*(12).

Patel, S., & Patel, H. (2016). Survey of data mining techniques used in healthcare domain. *International Journal of Information*, *6*(1/2), 53-60.

Patterson, J., Gibson, A. (2017). Deep Learning A Practitioner's Approach, 117-164.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Rita, P. A. A. (2018). Aplicação data mining para análise e previsão das estratégias de pricing em companhias aéreas. Estudo de caso: registos das tarifas da rota SSA-LIS (Doctoral dissertation).

Rojas, R. (2013). *Neural networks: a systematic introduction*. Springer Science & Business Media.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533-536.

Sakkaf, Y. (2020). *Decision Trees: ID3 Algorithm Explained*. [Online] In *Towards data science*. Retrieved February 17, 2020, from:https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1.

Shafer, J., Agrawal, R., & Mehta, M. (1996). SPRINT: A scalable parallel classifier for data mining. In *Vldb* (Vol. 96, pp. 544-555).

Shalvi, D., & DeClaris, N. (1998). An unsupervised neural network approach to medical data mining techniques. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)* (Vol. 1, pp. 171-176). IEEE.

Sharma, A. (2017). Understanding activation functions in neural networks. [online] In *Medium blog*. Retrieved February 17, 2020, from: https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, *5*(4), 13-22.

Sheela, K. G. & Deepa, S. N. (2013). Review on methods to fix number of hidden neurons in neural networks. *Mathematical Problems in Engineering*, *2013*.

Shtatland, E. S., Kleinman, K., & Cain, E. M. (2008). Stepwise Methods in using SAS R proc logistic and SAS R Enterpise Miner for Prediction. *SAS Institute*.

Silipo, R., Melcher, K. (2019) From a Single Decision Tree to a Random Forest. [online] In *Towards Data Science.* Retrieved February 17, 2020, from: https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147.

Singhal, S., Kumar, H., Passricha, V. (2018). Prediction of Heart Disease using CNN. [online] In *American International Journal of Research in Science, Technology, Engineering & Mathematics* – Retrieved January 12, 2020, from: http://iasir.net/AIJRSTEMpapers/AIJRSTEM18-345.pdf.

Sondwale, P. P. (2015). Overview of predictive and descriptive data mining techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(4), 262-265.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

Srinivas, K., Rao, G. R., & Govardhan, A. (2010). Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In *2010 5th International Conference on Computer Science & Education* (pp. 1344-1349).

Srivastava, A. N., & Han, J. (Eds.). (2016). *Machine learning and knowledge discovery for engineering systems health management*. CRC Press.

Tang, P. H. & Tseng, M. H. (2009, July). Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification. In *2009 International Conference on Machine Learning and Cybernetics* (Vol. 5, pp. 3070-3075). IEEE.

Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine, 4(2), 627.

Timofeev, R. (2004). Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin*, 1-40.

Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2008). Linear and logistic regression analysis. *Kidney international*, *73*(7), 806-810.

Ujjwalkarn (2016). A quick introduction to artificial neural networks. [online] In *The data science blog.* Retrieved February 17, 2020, from: https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/.

Wang, H., & Raj, B. (2017). On the origin of deep learning. *arXiv preprint arXiv:1702.07800.*

Wang, Q. R., & Suen, C. Y. (1984). Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4), 406-417.

Wasan, S. K., Bhatnagar, V., & Kaur, H. (2006). The impact of data mining techniques on medical diagnostics. *Data Science Journal*, 5, 119-126.

Wilkins, E., Wilson, L., Wickramasinghe, K., Bhatnagar, P., Leal, J., Luengo-Fernandez, R., ... & Townsend, N. (2017). European cardiovascular disease statistics 2017.

Winters-Miner, L. A. (2014). Seven ways predictive analytics can improve healthcare. [online] In *Elsevier Connect.* Retrieved February 17, 2020, from: https://www.elsevier.com/connect/seven-ways-predictive-analytics-can-improve-healthcare?aaref=.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, *31*(1), 76-77.

Xing, Y., Wang, J., & Zhao, Z. (2007). Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)* (pp. 868-872). IEEE.

Xue, W., Sun, Y., & Lu, Y. (2006). Research and application of data mining in traditional Chinese medical clinic diagnosis. In 2006 *8th international Conference on Signal Processing* (Vol. 4). IEEE.

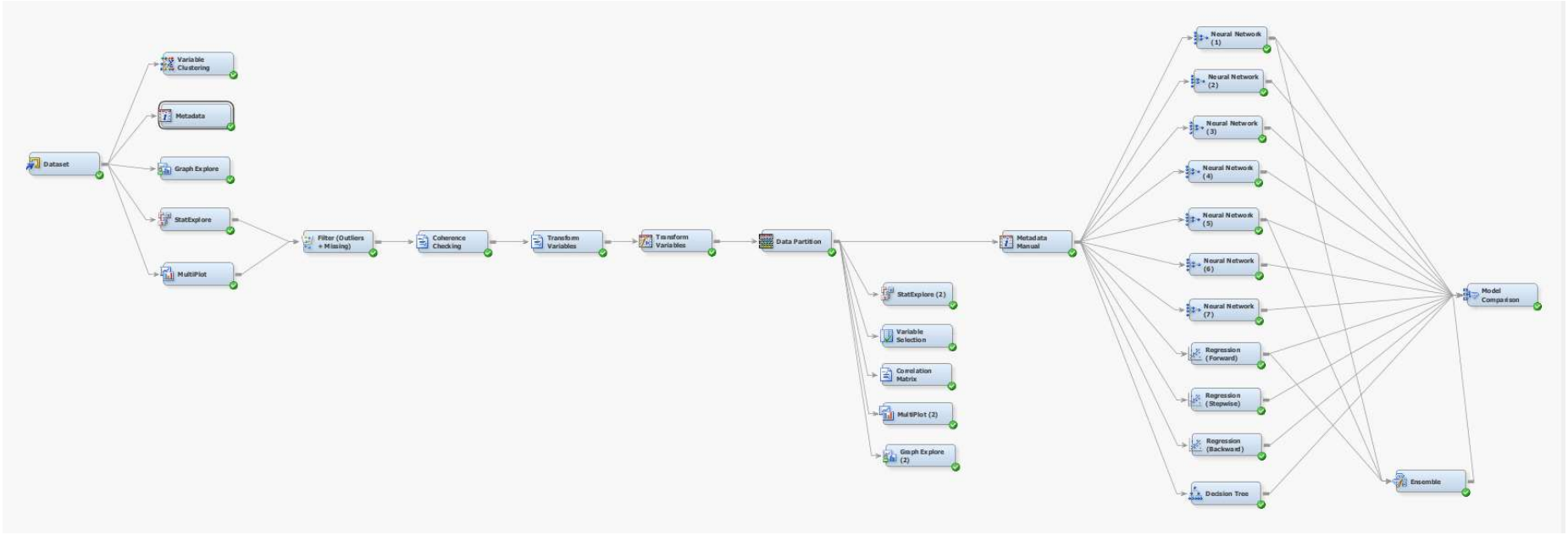
Yumusak, N. & Temurtas, F. (2010). Chest diseases diagnosis using artificial neural networks. *Expert Systems with Applications, 37*(12), 7648-7655.

Ziegler, A., & König, I. R. (2014). Mining data with random forests: current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(1), 55-63.

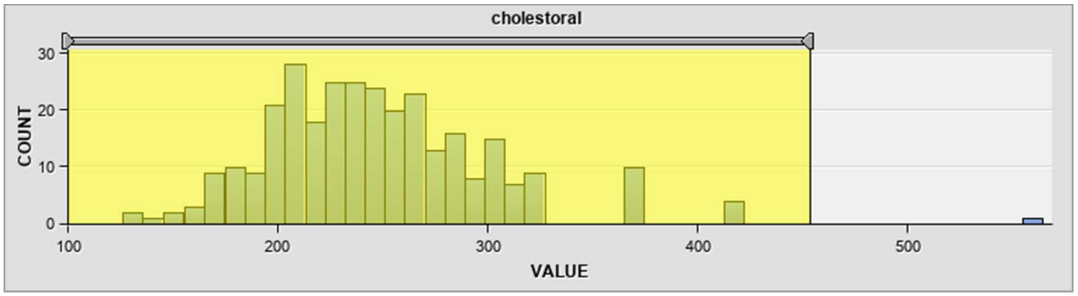
Zhang, G., Patuwo, B. E., Hu, M. Y. (1998). *Forecasting with artificial neural networks: The state of the art*. International Journal of Forecasting, 14(1), 35–62.

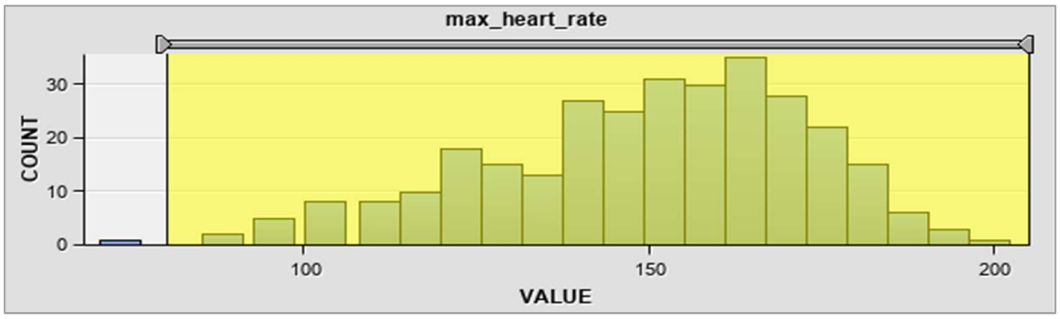
8. APPENDIX

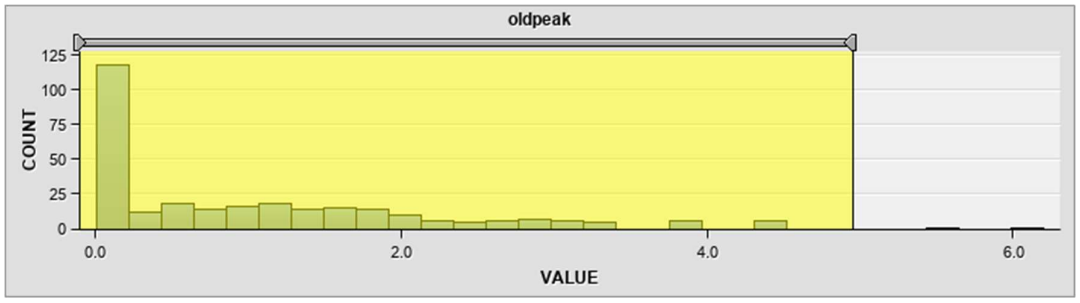
- **Appendix 1: SAS Enterprise Miner Project Diagram**



- **Appendix 2: Treatment of Outliers**







Number Of Observations

Data Role	Filtered	Excluded	DATA
TRAIN	299	4	303

- **Appendix 3: Correlation Matrix Results – Pearson Correlation**

PEARSON

NAME	age_cod	chest_pain	cholestoral_cod	electrocardiographic_results	max_hear_rate_cod	nr_major_vessels	oldpeak	resting_blood_pressure	slope	thal
age_cod	1	-0.10221	-0.07405	-0.04387	0.159228	0.262029	0.242266	0.276532	-0.13113	0.071629
chest_pain	-0.10221	1	0.117886	0.046693	-0.16101	-0.14219	-0.14846	0.061506	0.055425	-0.1814
cholestoral_cod	-0.07405	0.117886	1	-0.02344	0.005797	0.001744	-0.05828	-0.08262	-0.03399	0.010306
electrocardiographic_re...	-0.04387	0.046693	-0.02344	1	-0.16091	-0.04226	-0.03341	-0.13397	0.044343	0.002063
max_hear_rate_cod	0.159228	-0.16101	0.005797	-0.16091	1	0.040742	0.347099	0.063283	-0.38593	0.028264
nr_major_vessels	0.262029	-0.14219	0.001744	-0.04226	0.040742	1	0.233076	0.106702	-0.01997	0.102515
oldpeak	0.242266	-0.14846	-0.05828	-0.03341	0.347099	0.233076	1	0.166579	-0.51883	0.223187
resting_blood_pressure	0.276532	0.061506	-0.08262	-0.13397	0.063283	0.106702	0.166579	1	-0.11224	.0002743
slope	-0.13113	0.055425	-0.03399	0.044343	-0.38593	-0.01997	-0.51883	-0.11224	1	-0.09061
thal	0.071629	-0.1814	0.010306	0.002063	0.028264	0.102515	0.223187	.0002743	-0.09061	1

- **Appendix 4: Decision Tree Model**