Universidad Carlos III de Madrid
www.uc3m.es

APPLICATION OF DATA SCIENCE TO REDUCE EMPLOYEE ATTRITION

Bachelor Thesis

Degree in Telematics Engineering

Author: Clara Cabañas Pujadas

Tutor: Julio Villena Román

Leganés, October 2016

Keywords

Data Science, Data Mining, Analytics, Employee Attrition, Employee Turnover, People Analytics, Staffing

Abstract

Retaining valuable employees and preventing their resignation is a matter that can make a company save a considerable amount of time and money. Traditionally, this task had been carried out by the Human Resources department of the companies, who would regularly conduct interviews among the employees in order to subsequently analyse them and try to extract conclusions and patterns that could help them understand the reasons why employees leave and thus, prevent the resignation of other employees in the future.

Nowadays, with the existence of Data Science and prediction techniques, this task can be automatically done, which allows the managers of the companies to obtain the information they require from the employees in a much faster and efficient way than it was obtained in the past when the task was done manually by the Human Resources department. This results in a significant decrease of the costs associated with employee attrition, maximizing the revenue of the company.

Index

List of Figures

List of Tables

1 Introduction

This project intends to provide, by means of Data Science, an answer to a problem that all company managers have to deal with: how to retain talent and avoid attrition in the organizations. In a company, talent refers to the implemented capacity of a committed professional or group of professionals that achieve superior results [1] and thus, it is paramount for companies to retain talented employees.

1.1 Motivation

The reason why employers are putting so much effort into preventing employee turnover, i.e. employees leaving the company and having to be replaced, is that it has an adverse impact for the company, including a huge waste of time and money.

Employee attrition leads to losses in all these areas:

- Overworked remaining employees: The rest of the employees have to take over the job that the person who has left was in charge of, which translates into unsatisfied employees, who are more likely to leave the company.
- Lost experience and knowledge: All the experience and knowledge that the former employee had acquired through the years are no longer available for the company.
- Recruiting costs: The selection process to replace the employee costs the Human Resources (HR) department both money and time.
- Training costs: In addition to the training courses the new employee might need, further costs must be taken into account. This worker will take some time to be fully able to do his/her job and he/she will need help from other co-workers, which leads to a significant drop in productivity.

The actual costs have been estimated and published in a study by the Center of American Progress, which reveals that, in average, replacing a lost employee costs businesses one fifth of the employee's yearly salary [2].

All this information shows the importance of retaining talented employees in the company in order to reduce costs and increase productivity. However, avoiding employee attrition is not an easy task, since the factors that drive their voluntary exit might not always be foreseeable and thus, preventable. Such task had traditionally been carried out by the HR department of the companies, who would manually go over exit interviews of former employees trying to extract conclusions or patterns that could explain their voluntary exit, which implies a significant waste of time and money for the company.

This is where Data Science plays a vital role. Even though the term "Data Science" is an evolving concept and does not have a clear and unique definition, it could be defined as *the exploration and quantitative analysis of all available structured and unstructured data to develop understanding, extract knowledge, and formulate actionable results* [3]. By means of Data Science, the analysis of the available exit interviews at a company can be done systematically, and predictions for the future can be obtained based on the data from previous employees, minimizing costs and increasing the chances of retaining good employees and preventing their attrition.

1.2 Project Objectives

The context exposed in the previous section reveals the importance of retaining talent at the organizations and preventing employee attrition. Thus, the aim of this project is to provide a solution to this problem by means of Data Science and Data Analytics.

To achieve this goal, several datasets have been collected from open data sources in order to be processed with data analytics technologies to extract insights that can help understand the data and can model the profile of the employees that abandon the company. In addition, data mining techniques will be used with the goal of obtaining a prediction that can allow managers to anticipate the employees' attrition in order to prevent it.

The objective is to apply some Data Science techniques to analyze employee attrition in two different scenarios.

1.3 Document Structure

This document has six main chapters. Current chapter 1 describes the motivation and objectives of the project, along with the document structure and the regulatory framework that affects the project. In chapter 2, the State of the Art of the concepts and tools covered in this project are explained. Next, in chapter 3, the CRISP-DM model is defined in detail, with all the phases and all the tasks and outputs for every phase. Chapters 4 and 5 present two case studies, in which the CRISP-DM methodology will be applied in order to obtain a prediction of the attrition in two specific organizations. The next chapter, number 6, contains the conclusions and future work lines for this project. To conclude with this report, in chapter 7 the project planning is explained in detail, including the time schedule for the project and the costs projection.

1.4 Regulatory Framework

Vast amounts of data are increasingly being generated by sensors, people or mobile apps every second. These data, if stored and correctly analyzed, represent a source of huge economic and social value. But the Big Data revolution brings both benefits

and threats for society, since it could result in privacy infringement on a massive scale [4].

In order to set a unified regulatory framework to address the Big Data threats in Europe, the European Union released the General Data Protection Regulation (GDPR) [5] in April 2016. This Regulation has the objective of strengthening and unifying data protection for individuals within the European Union, and addresses export of personal data outside the EU.

The Spanish law on the protection of data, i.e. Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal (LOPD) [6] has the objective of protecting, in terms of personal data, the freedom and fundamental rights of every person. This project is not affected by this law, since all the data have been anonymized and none of the datasets contain any personal information.

2 State of the Art

2.1 Data Science

Since Data Science is an evolving subject, there is not a single definition for the term. However, it could be defined as the study of the generalizable extraction of knowledge from data [7]. The term "Data Science" refers to an area of work that involves the collection, preparation, analysis, visualization, management and preservation of large collections of information.

Data Science is closely related to widely-known terms such as Big Data, Data Mining or Business Intelligence. However, those terms are neither equivalent nor mutually exclusive. In order to explain why those concepts are related and yet do not refer to the same things, all these concepts will be defined and described.

Big Data is a collection of very huge data sets with a great diversity of types so that it becomes difficult to process by using state-of-the-art data processing approaches or traditional data processing platforms. It is characterized by what has been called The 3Vs: Volume, which is the size of the data set, Velocity, which indicates the speed of data in and out, and Variety, which describes the range of data types and sources [8]. Data Science involves the use of Big Data technologies to be able to work with huge datasets, but they are not the same concept.

Data Mining involves the inferring of algorithms that explore the data contained in large and complex datasets, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction [9]. It represents the more technical part in the whole Data Science process, but it is not fully equivalent to it. Although, in practice both terms are commonly used indistinctly.

Business Intelligence (often referred to as BI) is a process that includes two primary activities: getting data in, i.e. moving data from a set of source systems into an integrated data warehouse, and getting data out, i.e. the access to data from the data warehouse done by business users and applications to perform enterprise reporting, OLAP (On-Line Analytical Processing), querying and predictive analytics [10]. The results of Business Intelligence processes help explain the performance of a particular business in the past, but do not represent predictions for the future, which is the aim of Data Science processes.

The Data Science process can be divided into well-defined phases shown in Figure 1 [11].

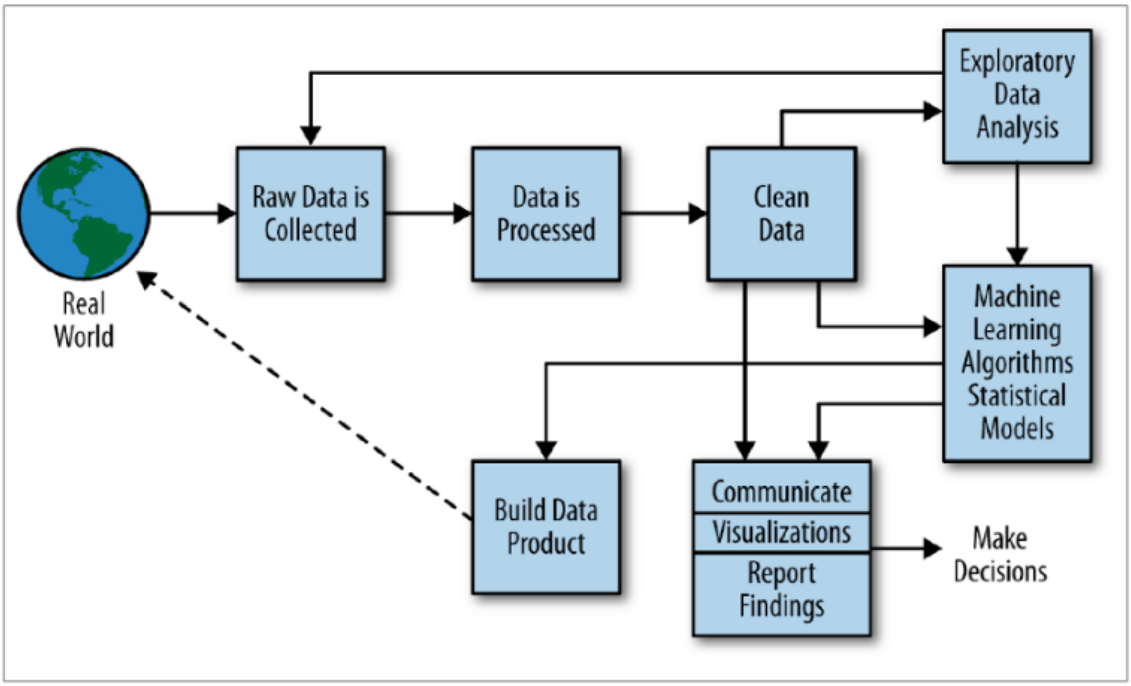

Figure 1 - Data Science Process

2.2 People Analytics

People Analytics, sometimes also referred to as HR Analytics, is the use of data and analytic tools to inform decisions about how to manage people. It represents a data-driven approach to managing people at work, instead of using traditional methods

of personal relationships, decision making based on experience, and risk avoidance [12].

Staffing is a very difficult task for managers that involves selecting and training individuals for specific job functions, and assigning them associated responsibilities.

Figure 2 - The staffing cycle

The staffing cycle is composed of three parts that can be seen in Figure 2 [12].

- The first part is Hiring, which means making sure that the right people are appointed for the job.
- The second part is Internal Mobility and Career Development, which means assessing if the employees that are already working for the organization are in the right job and deciding how to allow promotions through different jobs over time.
- The third and last part is Attrition, which means making sure that the talented employees stay in their jobs.

The last part of the cycle, employee attrition, is the main focus of this project. The importance of reducing high employee turnover rates is based on the high business losses that it leads to, which represent sunk costs for the company. The final goal of this project is to reduce the attrition rate at the organizations.

2.3 Tools

In this section, the data mining tools that will be used for the execution of the project will be covered.

2.3.1 IBM Watson Analytics

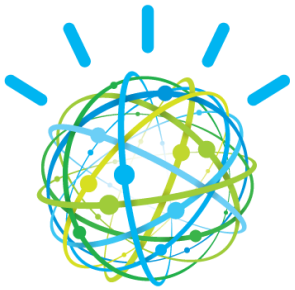

Figure 3 - IBM Watson Analytics logo

IBM Watson Analytics is a web application developed by IBM that offers data exploration tools, along with predictive analytics and dashboard and infographic creation on the cloud. The tool offers four environments for Data Science: Explore, for performing an exploratory analysis on the data, Predict, for obtaining predictions, Assemble, for creating dashboards to present the results, and Refine, for performing dataset cleansing operations [13].

For the execution of this project, the free version will be used, since there are no academic licenses available for students.

A screenshot showing the basic interface of the Prediction environment is shown in Figure 4.

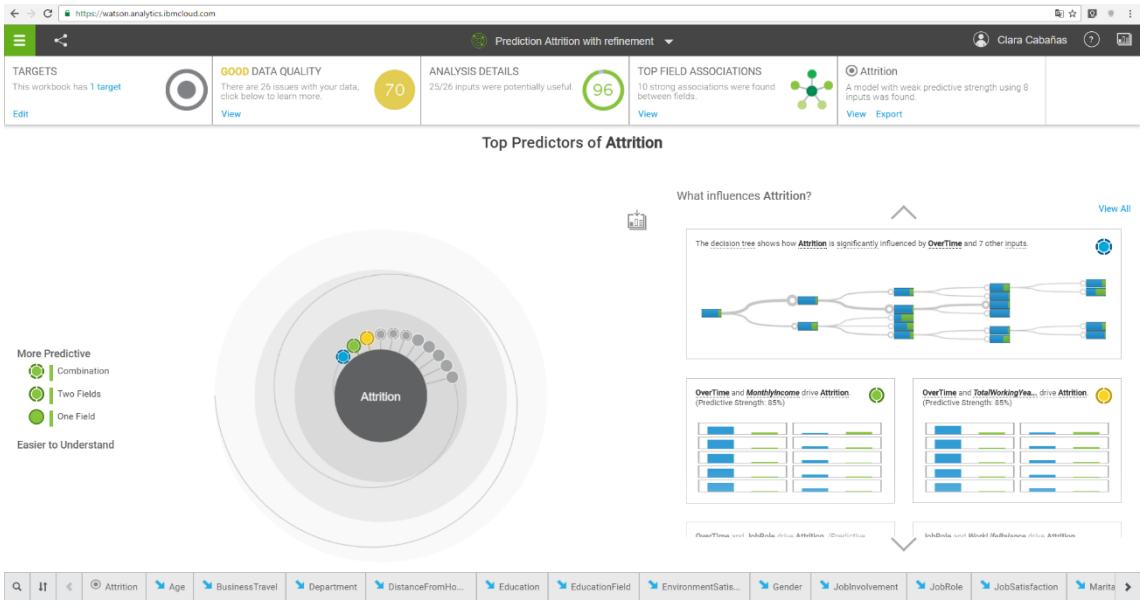

Figure 4 - Screenshot of the Watson Analytics interface

2.3.2 Rapidminer

Figure 5 - Rapidminer logo

Rapidminer Studio is a cross-platform software that provides a graphical environment for data mining and analytics. It provides operators that allow data access (any data source or format can be imported), data exploration, data blending, data cleansing, modeling and validation. In addition, operators for the execution of scripts written in Python or R are included. A poll carried out by KDnuggets in 2010 revealed that Rapidminer was the most popular tool for data mining [14].

For the execution of this project, an educational license will be used in order to have full access to all the features in the environment.

A screenshot showing the basic interface of the Rapidminer software can be seen in Figure 6.

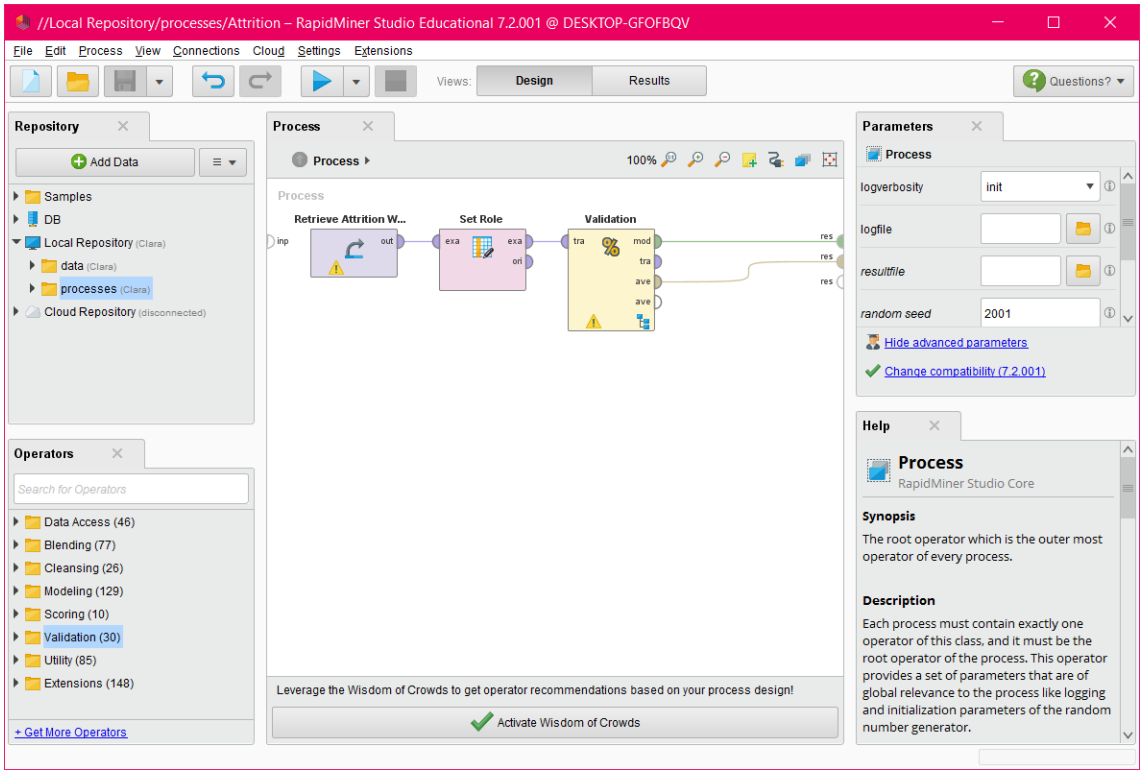

Figure 6 - Screenshot of the Rapidminer interface

2.3.3 Weka

Figure 7 - Weka logo

Weka is an open source software in Java that provides a graphical environment for machine learning and data mining. It offers several panels for different uses:

preprocessing utilities to import the data and perform cleansing operations on it, classifying and regression algorithms and methods to calculate their accuracy, association rules learners, clustering techniques, and visualization tools to create scatter plots.

It is free software licensed under the GNU General Public License.

A screenshot showing the basic interface of the Explorer environment is shown in Figure 8.

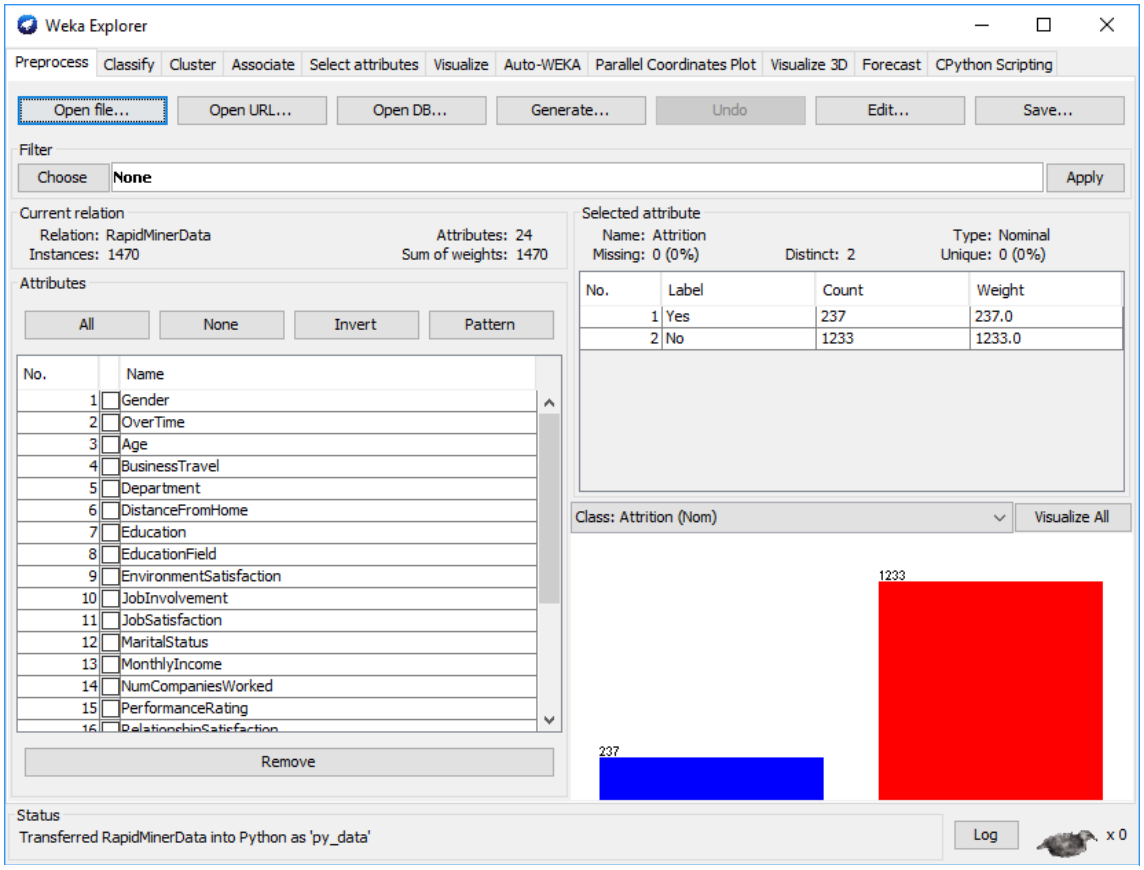

Figure 8 - Screenshot of the Weka interface

3 Standard Process for People Analytics: CRISP-DM

In this section, the standard process model for a People Analytics project will be covered in detail. This process is actually the *de facto* standard methodology for data mining: CRISP-DM.

CRISP-DM stands for **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining [15], and is a data mining process model that describes the most commonly used approaches that data mining experts use to tackle problems. This model encourages best practices and offers organizations the structure needed to achieve better, faster results from data mining. Polls carried out by KDNuggets[1] (leading site that covers the news in the field of Business Analytics, Big Data, Data Mining and Data Science) in 2002 [16], 2004 [17], 2007 [18] and 2014 [19] revealed that CRISP-DM was the most used methodology among the users.

The process is made up of six phases, as it is shown in Figure 9 [20]:

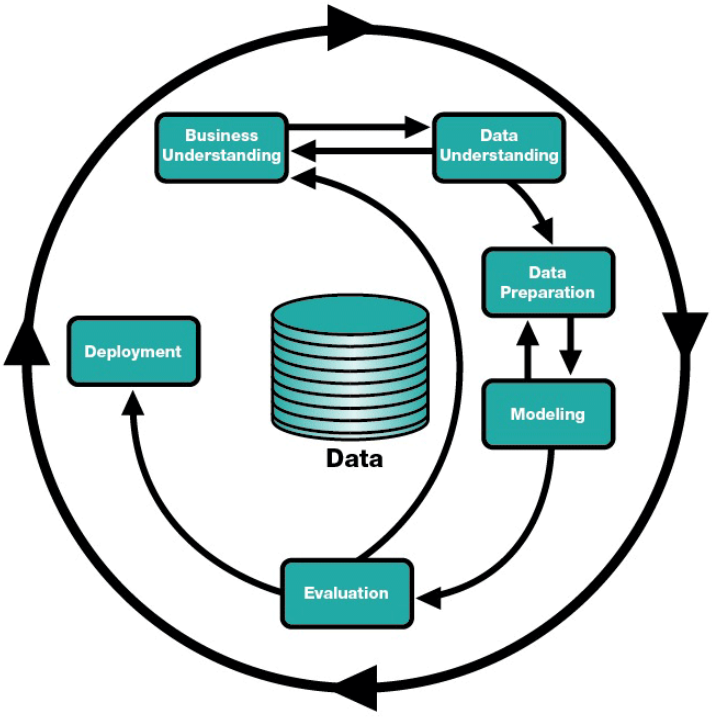

Figure 9 - Phases of the CRISP-DM model

[1] http://www.kdnuggets.com/

The arrows inside the diagram represent the connections and dependencies between phases, while the outer circle symbolizes the cyclic nature of data mining.

3.1 Business Understanding

The first phase in the CRISP-DM process model is Business Understanding. This phase focuses on understanding the project objectives and requirements from a business perspective and gathering all the details about the resources, assumptions and constraints. In addition to that, the specific data mining goals and data mining success criteria are specified. Finally, a preliminary project plan is developed.

The tasks and outputs corresponding to each task of this phase can be seen in detail in Figure 10 [21].

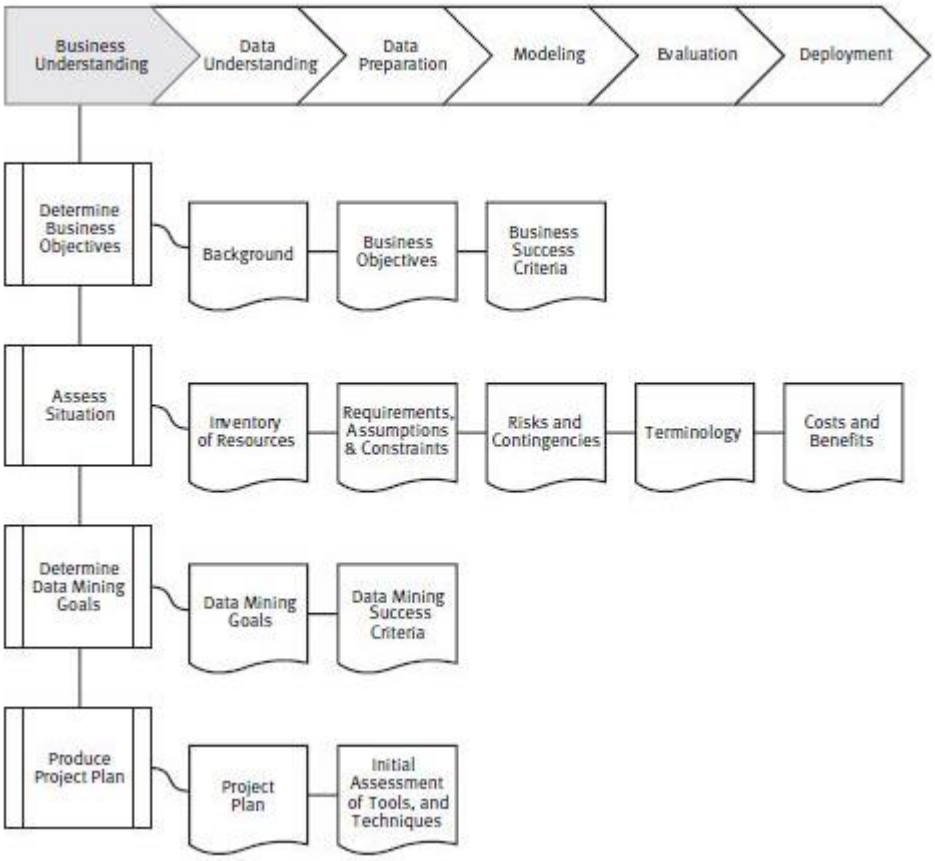

Figure 10 - Bussiness Understanding phase

3.2 Data Understanding

The second phase in the CRISP-DM process model is Data Understanding. This phase starts with the data collection, followed by the initial description and exploration of the data, where the distribution of the variables or relationships between pairs of them may be included. Finally, a data quality report is generated, listing the results of the data quality verification, and explaining if the data is complete, if it contains errors, or if any missing values are present.

The tasks and outputs corresponding to each task of this phase can be seen in detail in Figure 11 [22].

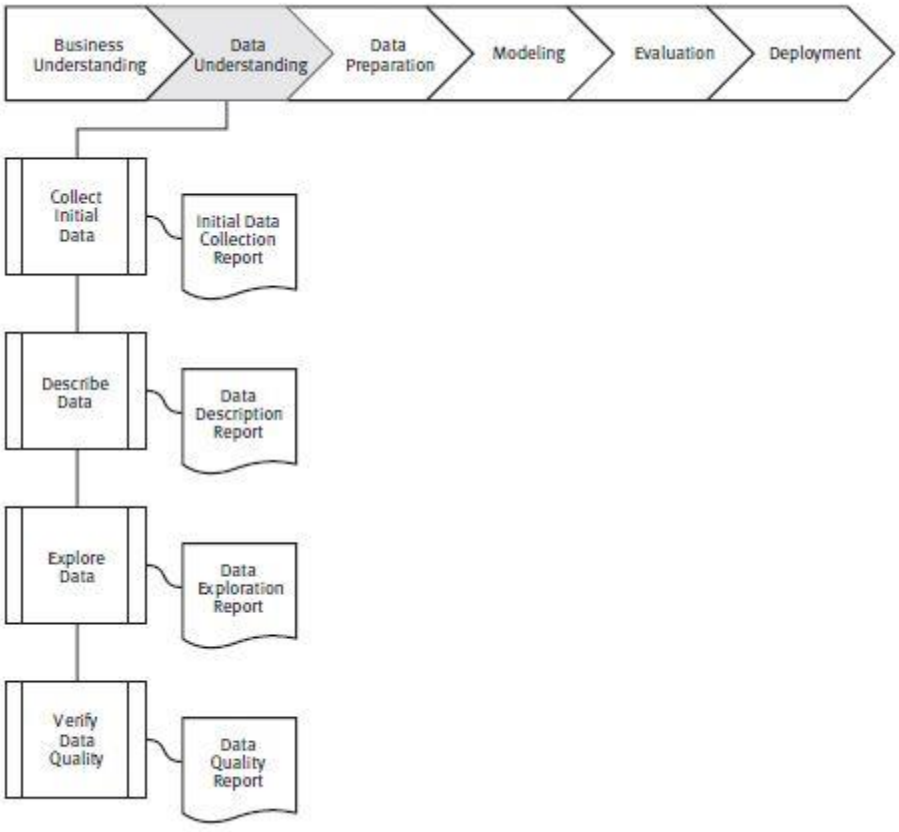

Figure 11- Data Understanding phase

3.3 Data Preparation

The third phase in the CRISP-DM process model is Data Preparation. The main goal of this phase is to construct the final dataset that will be fed into the modelling algorithms. In order to achieve that, a data cleansing process must be carried out, along with constructive data preparation operations to obtain derived attributes or create generated records. Besides, the data has to be formatted to meet every algorithm's requirements.

The tasks and outputs corresponding to each task of this phase can be seen in detail in Figure 12 [23].

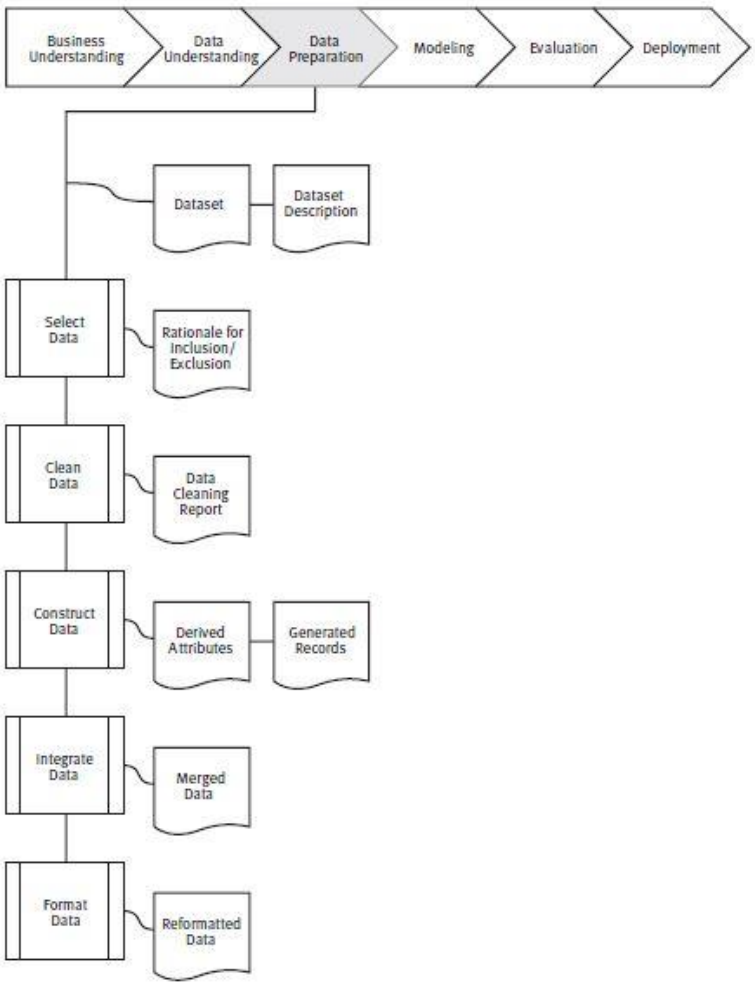

Figure 12 - Data Preparation phase

3.4 Modeling

The fourth phase in the CRISP-DM process model is Modeling. In this phase, the modeling techniques are selected and applied, calibrating their parameters to the optimal values. Finally, the models are assessed in terms of accuracy and the results are collected and ranked.

The tasks and outputs corresponding to each task of this phase can be seen in detail in Figure 13 [24].

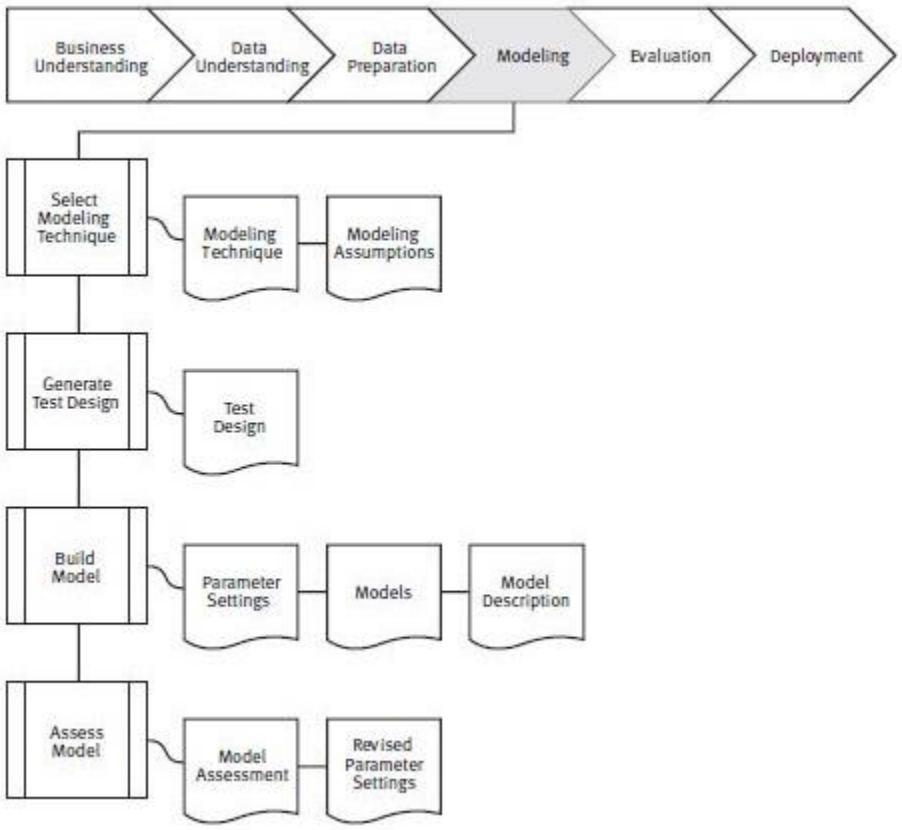

Figure 13 - Modeling phase

3.5 Evaluation

The fifth phase in the CRISP-DM process model is Evaluation. In this phase, the model that has been built and evaluated in the Modeling phase is assessed with respect to business success criteria. Besides, the whole process is reviewed to

determine if any important factors or tasks have been missed and, finally, the following steps are determined.

The tasks and outputs corresponding to each task of this phase can be seen in detail in Figure 14 [25].

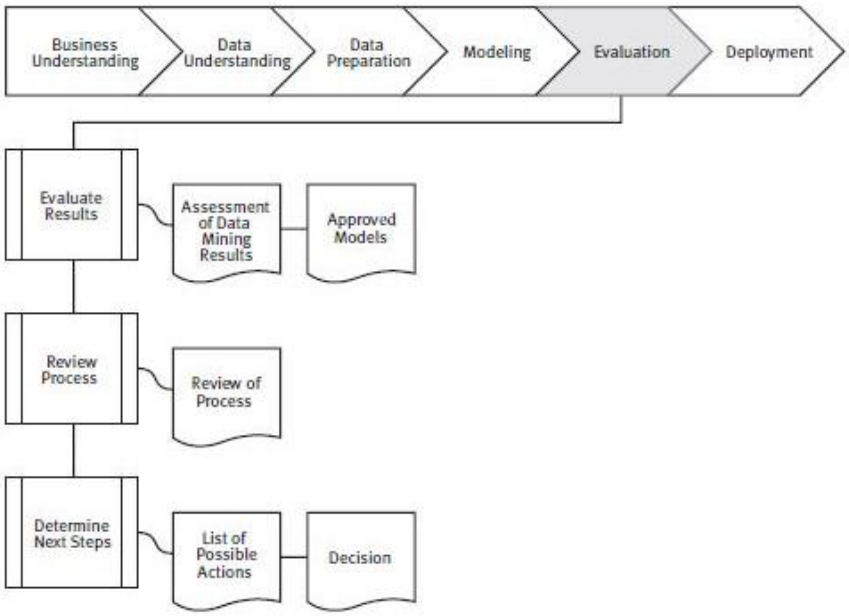

Figure 14 - Evaluation phase

3.6 Deployment

The sixth and last phase in the CRISP-DM process model is Deployment. In this phase, the deployment strategy is carried out. This strategy could be very simple, like generating a report, or very complex, like implementing a repeatable data mining process.

The tasks and outputs corresponding to each task of this phase can be seen in detail in Figure 15 [26].

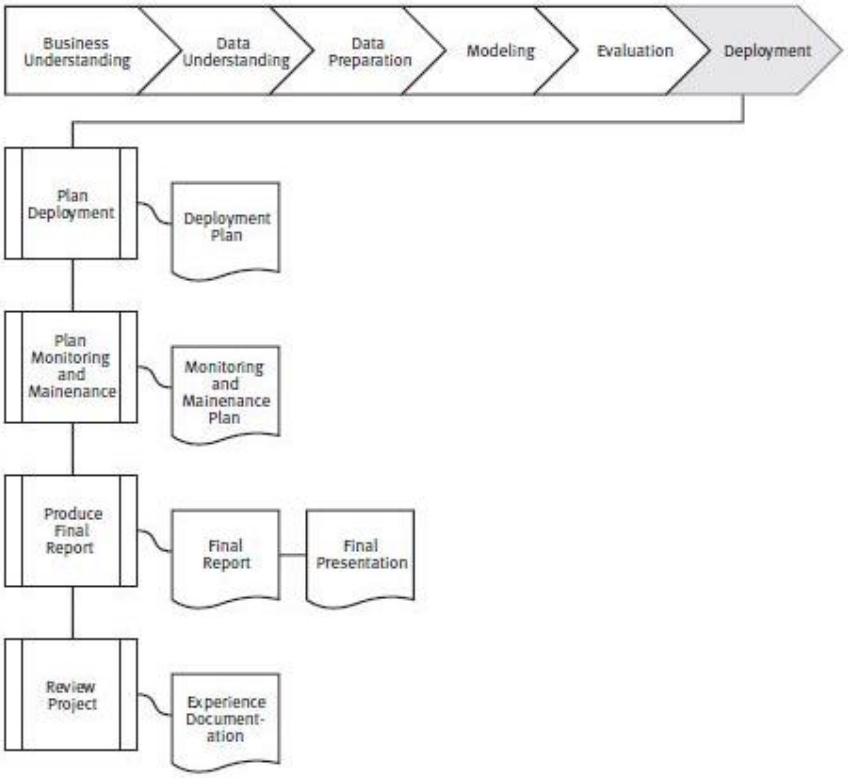

Figure 15 - Deployment phase

4 CASE STUDY 1: Pilot Scenario

The goal of this study is to determine using Data Science the reasons that drive employee attrition in a specific company. In this case, a dataset obtained from the Watson Analytics Forum will be used, which contains real data with all personal identifiers removed. Besides, the data has been tweaked so that it performs better in this illustrative example of the standard Data Science process carried out for People Analytics.

For this purpose, the IBM Watson Analytics tool will be used, along with the Rapidminer Studio software and the Weka software.

The CRISP-DM process model has been the reference model for the execution of this project and the phases will be covered in detail.

4.1 Business Understanding

The business goal of this study is to predict whether an employee is going to voluntarily leave the company or stay. The benefits of this project will be minimizing employee turnover costs for the company and retaining valuable workers.

The hardware available for the execution of this project is a Lenovo Ideapad Z510 laptop, along with an Asus desktop computer with Windows 10, an i5-6600 microprocessor and 32GB RAM.

The data that will be analyzed has been obtained from the Watson Analytics website and is static, meaning that no further collection of data will take place. Metadata for the variables in the dataset is available.

The data mining goal of this study is to extract insights from the data that can explain which factors drive attrition in the company, and to predict attrition for a particular employee, given some personal and professional information. For this purpose, several prediction algorithms will be tested in order to obtain the attrition prediction.

The output model of the data mining process will be required to have a minimum value of accuracy of 70% for the project to be successful. If such value is not achieved, the whole process must be repeated and corrected in order to obtain the desired accuracy.

4.2 Data Understanding

The data that will be used for this study has been collected from the Watson Analytics Community forum. It contains information that has been gathered from the answers of a survey that an unidentified company performed on their employees. All the information related to the company has been anonymized.

The dataset consists of 1470 records and 35 variables, one of which is the target variable: `Attrition`. The variables are:

- `Age`: numerical variable indicating the age of the employee. It takes values from 18 to 60. It contains no missing values.
- `Attrition`: binary variable indicating if the employee has voluntarily left the company or not. It can take values "Yes" or "No". It contains no missing values.
- `BusinessTravel`: categorical variable indicating the business travel activity of the employee. It can take the values "Non-travel", "Travel frequently" or "Travel rarely". It contains no missing values.
- `DailyRate`: numerical variable indicating the value of the employee. It takes values from 102 to 1499. It contains no missing values.

- `Department`: categorical variable indicating the department to which the employee belongs. It can take the values "Human Resources", "Research & Development" or "Sales". It contains no missing values.
- `DistanceFromHome`: numerical variable indicating the distance from the workplace to home. It takes values from 1 to 29. It contains no missing values.
- `Education`: categorical variable indicating the education level reached by the employee. It can take the values 1 (which corresponds to "Below College"), 2 (which corresponds to "College"), 3 (which corresponds to "Bachelor"), 4 (which corresponds to "Master") or 5 (which corresponds to "Doctor"). It contains no missing values.
- `EducationField`: categorical variable indicating the education field of the employee. It can take the values "Human Resources", "Life Sciences", "Marketing", "Medical", "Technical Degree" or "Other". It contains no missing values.
- `EmployeeCount`: numerical variable indicating the count of the employee. It takes the value 1 for all the records. It contains no missing values.
- `EmployeeNumber`: id variable indicating the employee's id number in the dataset. It takes values from 1 to 2068. It contains no missing values.
- `EnvironmentSatisfaction`: categorical variable indicating the level of satisfaction of the employee with the working environment. It can take the values 1 (which corresponds to "Low"), 2 (which corresponds to "Medium"), 3 (which corresponds to "High") or 4 (which corresponds to "Very High"). It contains no missing values.
- `Gender`: categorical variable indicating the gender of the employee. It can take the values "Female" or "Male". It contains no missing values.
- `HourlyRate`: numerical variable indicating the hourly rate of the employee. It takes values from 30 to 100. It contains no missing values.
- `JobInvolvement`: categorical value indicating the level of involvement of the employee with his job. It can take the values 1 (which corresponds to "Low"), 2

(which corresponds to “Medium”), 3 (which corresponds to “High”) or 4 (which corresponds to “Very High”). It contains no missing values.

- `JobLevel`: categorical variable indicating the job level of the employee. It can take the values 1, 2, 3, 4 or 5. It contains no missing values.
- `JobRole`: categorical variable indicating the role of the employee in the company. It can take the values “Healthcare Representative”, “Human Resources”, “Laboratory Technician”, “Manager”, “Manufacturing Director”, “Research Director”, “Research Scientist”, “Sales Executive” or “Sales Representative”. It contains no missing values.
- `JobSatisfaction`: categorical variable indicating the level of satisfaction of the employee with his job. It can take the values 1 (which corresponds to “Low”), 2 (which corresponds to “Medium”), 3 (which corresponds to “High”) or 4 (which corresponds to “Very High”). It contains no missing values.
- `MaritalStatus`: categorical variable indicating the marital status of the employee. It can take the values “Divorced”, “Married” or “Single”. It contains no missing values.
- `MonthlyIncome`: numerical variable indicating the monthly income of the employee. It takes values from 1009 to 19999. It contains no missing values.
- `MonthlyRate`: numerical variable indicating the monthly rate of the employee. It takes values from 2094 to 26999. It contains no missing values.
- `NumCompaniesWorked`: numerical variable indicating the number of companies in which the employee has worked before the current one. It takes values from 0 to 9. It contains no missing values.
- `Over18`: binary variable indicating if the employee is older than 18. It takes the value “Yes” for all the records. It contains no missing values.
- `OverTime`: binary variable indicating if the employee works overtime or not. It can take the values “Yes” or “No”. It contains no missing values.
- `PercentSalaryHike`: numerical variable indicating the percentage of salary raise of the employee. It takes values from 11 to 25. It contains no missing values.

- `PerformanceRating`: categorical variable indicating the performance rating of the employee. It can take the values 1 (which corresponds to "Low"), 2 (which corresponds to "Good"), 3 (which corresponds to "Excellent") or 4 (which corresponds to "Outstanding"). It only takes values 3 and 4. It contains no missing values.
- `RelationshipSatisfaction`: categorical variable indicating the level of satisfaction with the work relationships of the employee. It can take the values 1 (which corresponds to "Low"), 2 (which corresponds to "Medium"), 3 (which corresponds to "High") or 4 (which corresponds to "Very High"). It contains no missing values.
- `StandardHours`: numerical variable indicating the number of standard hours that the employee works. It takes the value 80 for all the records. It contains no missing values.
- `StockOptionLevel`: categorical variable whose meaning is unknown. It can take values 0, 1, 2, or 3. It contains no missing values.
- `TotalWorkingYears`: numerical variable indicating the total amount of years that the worker has been employed. It takes values from 0 to 40. It contains no missing values.
- `TrainingTimesLastYear`: numerical variable indicating the amount of times that the employee undertook training programs last year. It takes values from 0 to 6. It contains no missing values.
- `WorkLifeBalance`: numerical variable indicating the level of satisfaction of the employee with his work-life balance. It can take the values 1 (which corresponds to "Bad"), 2 (which corresponds to "Good"), 3 (which corresponds to "Better") or 4 (which corresponds to "Best"). It contains no missing values.
- `YearsAtCompany`: numerical variable indicating the years that the employee has been working for the company. It takes values from 0 to 40. It contains no missing values.

- `YearsInCurrentRole`: numerical variable indicating the years that the employee has had his current role at the company. It takes values from 0 to 18. It contains no missing values.
- `YearsSinceLastPromotion`: numerical variable indicating the number of years since the employee was last promoted. It takes values from 0 to 15. It contains no missing values.
- `YearsWithCurrManager`: numerical variable indicating the number of years that the employee has been working with his current manager. It takes values from 0 to 17. It contains no missing values.

The quality of the variables can be assessed with Watson Analytics. The results of such assessment reveal that three of the variables in the dataset have constant values: `EmployeeCount` has always the value 1, `Over18` has always the value Yes, and `StandardHours` has always the value 80.

The histogram of all the continuous variables is obtained in order to know the distribution of the values in every variable, along with the bar chart for all the categorical variables representing the frequencies of every value. A histogram is used to represent the characteristics of a variable. A graph is constructed by subdividing the axis of measurement into intervals of equal width, and constructing a rectangle over each interval, such that the height of the rectangle is equal to the total number of height of measurements falling in each cell [27].

The histograms of the variables `Age`, `MonthlyIncome` and `YearsAtCompany` are shown in figures 16, 17 and 18, respectively.

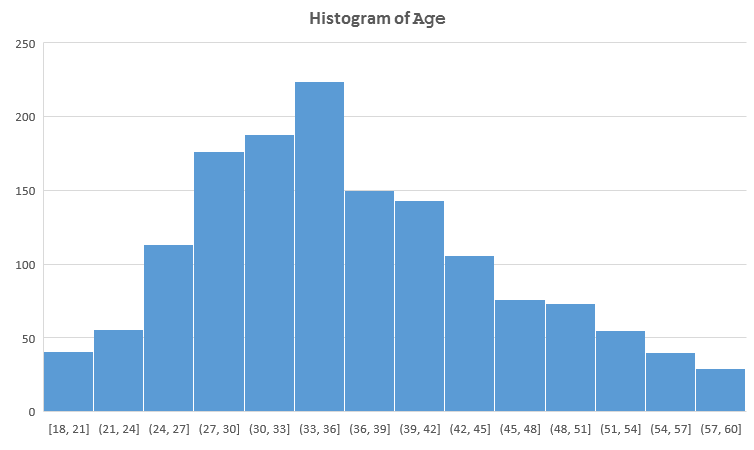

Figure 16 – Histogram of `Age`

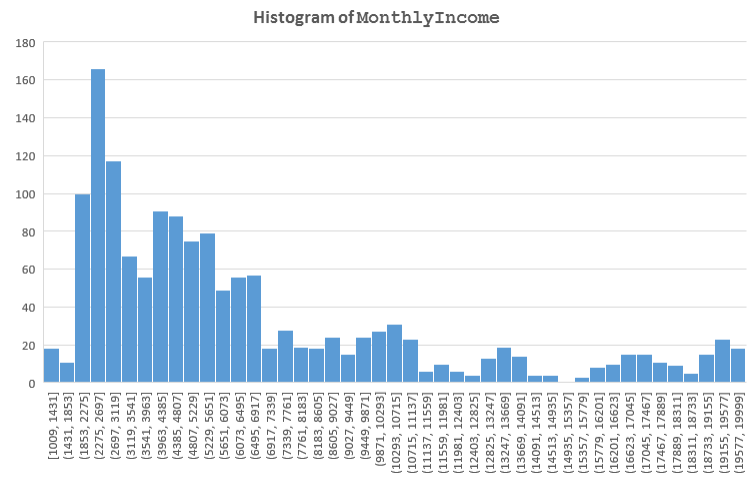

Figure 17 – Histogram of `MonthlyIncome`

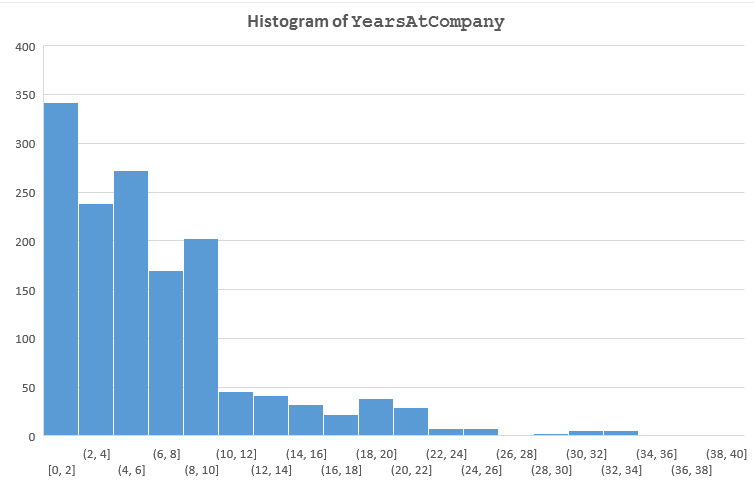

Figure 18 – Histogram of `YearsAtCompany`

The bar charts showing the distribution of values of the variables `OverTime` and `Attrition` are shown in figures 19 and 20, respectively.

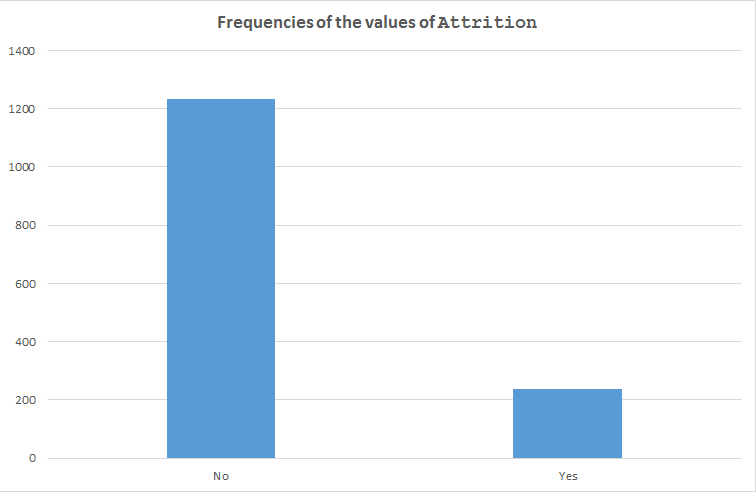

Figure 19 - Frequencies of the values of `Attrition`

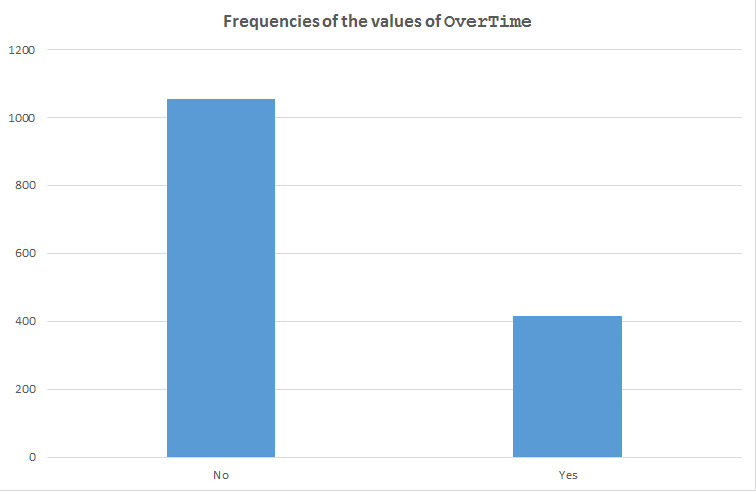

Figure 20 - Frequencies of the values of OverTime

The correlation matrix of all the variables is obtained in order to know how strongly every pair of variables is related. A correlation is a number between -1 and +1 that measures the degree of association between two variables. A positive value for the correlation implies a positive association, i.e. large values of one of the variables tend to be associated with large values of the other variable, and small values of the former tend to be associated with small values of the latter. A negative value of the correlation implies a negative or inverse association [28].

Attributes	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	Hour
Age	1	0.159	-0.025	0.011	0.032	-0.002	0.208	-0.007	?	-0.010	0.010	-0.036	0.02
Attrition	0.159	1	0.000	0.057	0.064	-0.078	0.031	-0.075	?	0.011	0.103	-0.029	0.00
BusinessTravel	-0.025	0.000	1	0.004	-0.009	0.024	-0.001	-0.018	?	0.016	-0.004	0.033	-0.02
DailyRate	0.011	0.057	0.004	1	-0.007	-0.005	-0.017	-0.016	?	-0.051	0.018	-0.012	0.02
Department	0.032	0.064	-0.009	-0.007	1	-0.017	-0.008	-0.057	?	0.011	0.019	0.042	0.00
DistanceFromHome	-0.002	-0.078	0.024	-0.005	-0.017	1	0.021	0.020	?	0.033	-0.016	-0.002	0.03
Education	0.208	0.031	-0.001	-0.017	-0.008	0.021	1	-0.003	?	0.042	-0.027	-0.017	0.01
EducationField	-0.007	-0.075	-0.018	-0.016	-0.057	0.020	-0.003	1	?	0.008	0.016	-0.004	-0.02
EmployeeCount	?	?	?	?	?	?	?	?	?	?	?	?	?
EmployeeNumber	-0.010	0.011	0.016	-0.051	0.011	0.033	0.042	0.008	?	1	0.018	0.023	0.03
EnvironmentSatisfaction	0.010	0.103	-0.004	0.018	0.019	-0.016	-0.027	0.016	?	0.018	1	0.001	-0.05
Gender	-0.036	-0.029	0.033	-0.012	0.042	-0.002	-0.017	-0.004	?	0.023	0.001	1	-0.00
HourlyRate	0.024	0.007	-0.027	0.023	0.004	0.031	0.017	-0.024	?	0.035	-0.050	-0.000	1
JobInvolvement	0.030	0.130	-0.039	0.046	0.025	0.009	0.042	-0.006	?	-0.007	-0.008	0.018	0.04
JobLevel	0.510	0.169	-0.019	0.003	-0.102	0.005	0.102	0.010	?	-0.019	0.001	-0.039	-0.02
JobRole	0.160	0.028	-0.029	-0.003	0.425	-0.044	-0.019	0.045	?	0.015	-0.010	-0.015	-0.02
JobSatisfaction	-0.005	0.103	0.034	0.031	-0.021	-0.004	-0.011	-0.055	?	-0.046	-0.007	0.033	-0.07
MaritalStatus	0.095	0.162	0.024	0.070	0.056	0.014	-0.004	0.018	?	0.008	0.004	0.047	0.01
MonthlyIncome	0.498	0.160	-0.034	0.008	-0.053	-0.017	0.095	0.008	?	-0.015	-0.006	-0.032	-0.01
MonthlyRate	0.028	-0.015	0.014	-0.032	-0.024	0.027	-0.026	-0.013	?	0.013	0.038	-0.041	-0.01
NumCompaniesWorked	0.300	-0.043	-0.021	0.038	0.036	-0.029	0.126	0.005	?	-0.001	0.013	-0.039	0.02
Over18	?	?	?	?	?	?	?	?	?	?	?	?	?
OverTime	-0.028	0.246	0.017	-0.009	0.007	-0.026	0.020	-0.003	?	0.024	-0.070	0.042	0.00
PercentSalaryHike	0.004	0.013	0.029	0.023	0.008	0.040	-0.011	-0.038	?	-0.013	-0.032	0.003	-0.00

Table 1 - Correlation matrix

The strongest correlations (greater than ± 0.7) between every pair of variables can be seen in Table 2.

First Attribute	Second Attribute	Correlation ↓
JobLevel	MonthlyIncome	0.950
JobLevel	TotalWorkingYears	0.782
PercentSalaryHike	PerformanceRating	0.774
MonthlyIncome	TotalWorkingYears	0.773
YearsAtCompany	YearsWithCurrManager	0.769
YearsAtCompany	YearsInCurrentRole	0.759
YearsInCurrentRole	YearsWithCurrManager	0.714
Age	TotalWorkingYears	0.680
MaritalStatus	StockOptionLevel	0.663
TotalWorkingYears	YearsAtCompany	0.628
YearsAtCompany	YearsSinceLastPromotion	0.618
YearsInCurrentRole	YearsSinceLastPromotion	0.548
JobLevel	YearsAtCompany	0.535
MonthlyIncome	YearsAtCompany	0.514
YearsSinceLastPromotion	YearsWithCurrManager	0.510
Age	JobLevel	0.510

Table 2 - Greatest correlations among all the variables

The highest correlations are between the variable `JobLevel` and variables `MonthlyIncome` and `TotalWorkingYears`. Thus, a high number of working years and a high monthly income are associated with a high job level.

4.3 Data Preparation

The results of the data exploration carried out in the Data Understanding phase are applied now to the data in order to clean the dataset and prepare it for the Modeling phase.

The variables `EmployeeCount`, `Over18` and `StandardHours` have constant values for all the rows in the dataset, so they are removed from the dataset because they do not provide any information.

The variable `EmployeeNumber` represents the employee's id, and is removed from the dataset since it provides irrelevant information.

None of the variables contain missing values, so no records are excluded from the dataset.

4.4 Modeling

In the Modeling phase, several analytical techniques and prediction algorithms will be applied to the dataset using Watson Analytics, Rapidminer and Weka in order to select the one that gives the highest accuracy as the final model.

4.4.1 Word Cloud

Using the Prediction tool from the Watson Analytics environment a word cloud can be obtained, where the size of the variable's name represents its predictor importance.

Figure 21 – Word Cloud for Case Study 1

Thus, the three variables that have a higher predictor influence on `Attrition` are `OverTime`, `YearsAtCompany` and `MonthlyIncome`.

4.4.2 Decision Rules

In order to understand the data and identify the variables that have the strongest prediction importance, the decision rules for the dataset are obtained using the Prediction tool from the Watson Analytics environment. A decision rule is a set of conditions that classify records, and it predicts an outcome in the target field [29], in this case, `Attrition`.

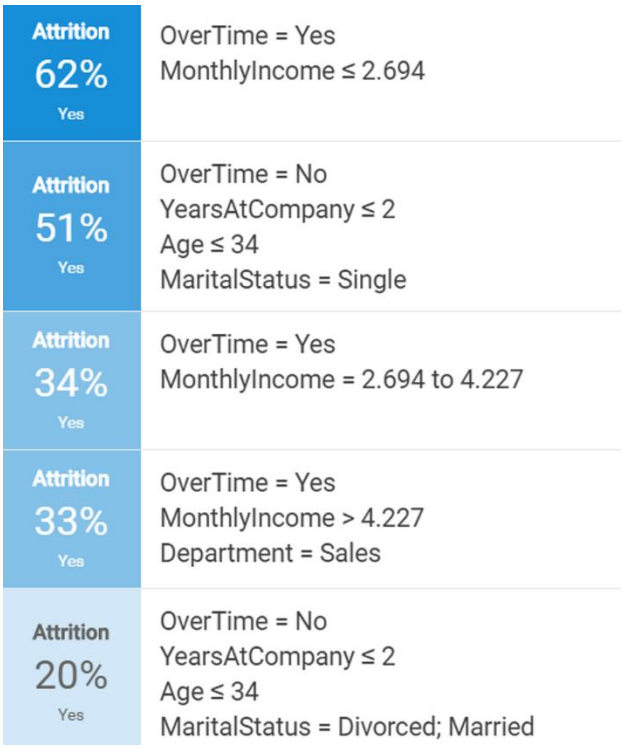

Figure 22 - Decision rules for Case Study 1

Figure 22 shows the set of rules that have been obtained. Very valuable information can be extracted. Among the employees that work overtime and have an income lower than $2,694 per month, 62% of them left the company voluntarily. In addition, among the employees that do not work overtime, have been working less than 2 years at the company, are younger than 34 years old and are not married, 51% of them left the company voluntarily.

In both cases, more than 50% of the employees belonging to each group left the company. This is an indication that shows which are the main reasons that trigger attrition in the company.

4.4.3 CHAID Classification Tree (Watson Analytics)

With regards to prediction algorithms, a CHAID classification tree [30] is obtained using the prediction tools from the Watson Analytics environment. The parameters for the algorithm cannot be modified, and they are automatically set to their optimal values.

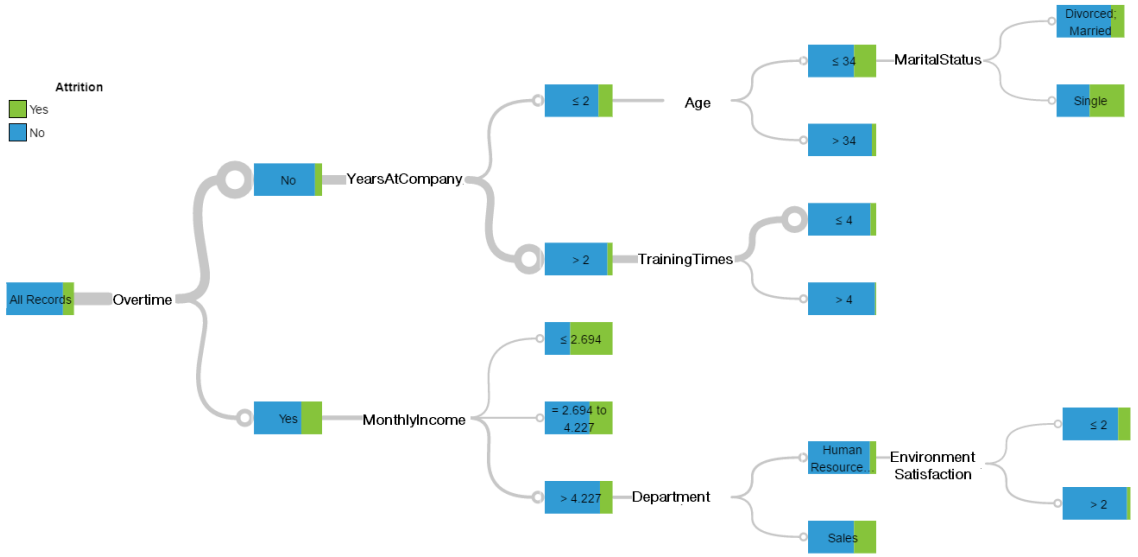

Figure 23 - Decision Tree obtained with Watson Analytics for Case Study 1

The output of the decision tree algorithm is highly visual and easy to interpret. From the tree, the groups that are more likely to leave the company are the employees who work overtime and have a monthly income lower than $2,694, and the employees who do not work overtime, have been less than or 2 years working for the company, are younger than 34 and are single.

The accuracy of the applied algorithm is obtained in order to know how it performed on the dataset. The result is an overall percentage of accuracy of 85%. The precision and recall of every class can be seen in Table 3.

Predicted Attrition	Observed Attrition		
	No	**Yes**	**Class precision**
No	1055	97	78%
Yes	178	140	22%
Class recall	86%	59%	81%

Table 3 - Accuracy of the decision tree obtained with Watson Analytics for Case Study 1

4.4.4 Decision Tree (Rapidminer)

A Decision Tree is obtained using Rapidminer's Decision Tree operator. The parameters are tuned in order to obtain the best possible model. The whole Rapidminer process is shown in Figure 24.

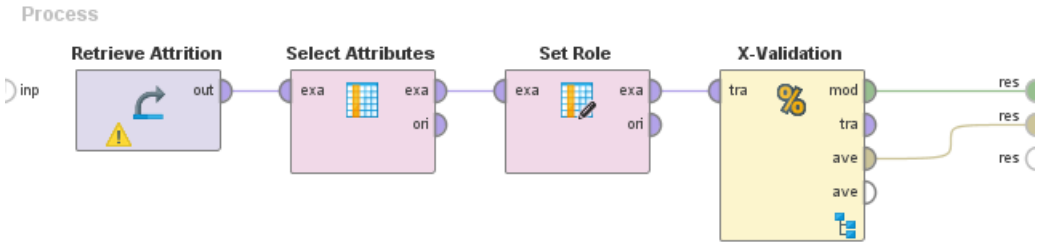

Figure 24 - Rapidminer process for Case Study 1

The Retrieve Attrition operator loads the Attrition dataset to be used. The next operator, Select Attributes, selects all the variables that were discarded in the Data Preparation phase and excludes them from the dataset. Next, the Set Role operator assigns the role "label" to the variable `Attrition`, in order to specify that it is the target attribute. Finally, the X-Validation operator performs a cross-validation with 10 folds to estimate the accuracy of the decision tree. In k-fold cross-validation, the dataset is split into k mutually exclusive subsets (the folds) of approximately equal size. Then, the model is trained and tested k times, each time with a different subset. The cross-validation estimate of accuracy is then the average of the k results from the k iterations [31].

The X-Validation operator is a nested operator. The operators that are inside the X-Validation operator can be seen in Figure 25.

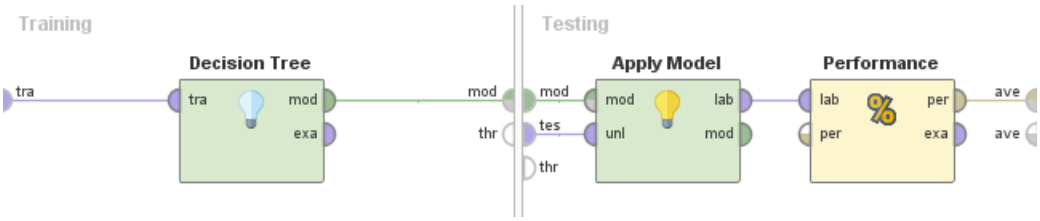

Figure 25 - Rapidminer process for Case Study 1: X-Validation

The training subprocess is used for training the Decision Tree model. The trained model is then applied in the testing subprocess, where the performance of the Decision Tree is also measured.

The depth of the tree is set to 5 in order to avoid overfitting. Besides, the pruning and prepruning parameters are set to true. The resulting tree can be seen in Figure 26.

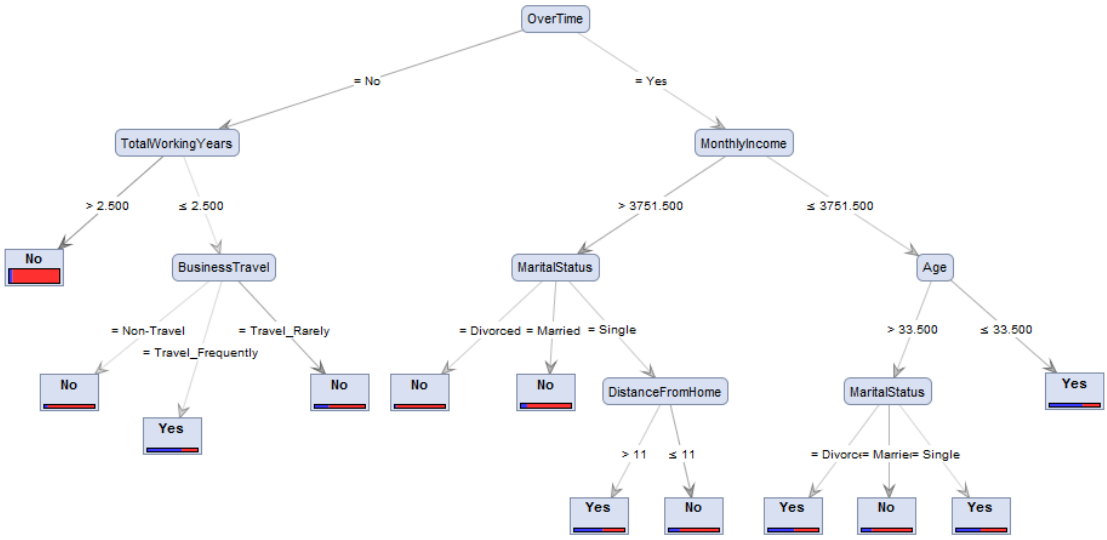

Figure 26 - Decision Tree obtained with Rapidminer for Case Study 1

The accuracy of the applied algorithm is obtained by means of cross-validation with 10 folds. The result is an overall accuracy of 83.61%. The precision and recall of every class can be seen in Table 4.

Predicted Attrition	Observed Attrition		
	No	**Yes**	**Class Precision**
No	1163	171	87.18%
Yes	70	66	48.53%
Class recall	94.32%	27.85%	83.61%

Table 4 - Accuracy of the decision tree obtained with Rapidminer for Case Study 1

4.4.5 J48 Tree (Weka)

A C4.5 tree is obtained using the J48 class from the Weka software. The parameters are tuned in order to obtain the best possible model. The reduced error pruning parameter is set to true and the minimum number of instances per node is set to 10.

A graph visualization software will be used together with Weka in order to obtain an *elegant* decision tree, since the output tree of the Weka software is not customizable and the nodes in the tree can visually overlap, making the result difficult to interpret. In this case, Weka will be used from the command line, along with the Graphviz software, also used from the command line. The two commands can be seen in Figure 27.

Figure 27 - Weka and Graphviz commands to obtain the J48 tree for Case Study 1

The output tree can be seen in Figure 28.

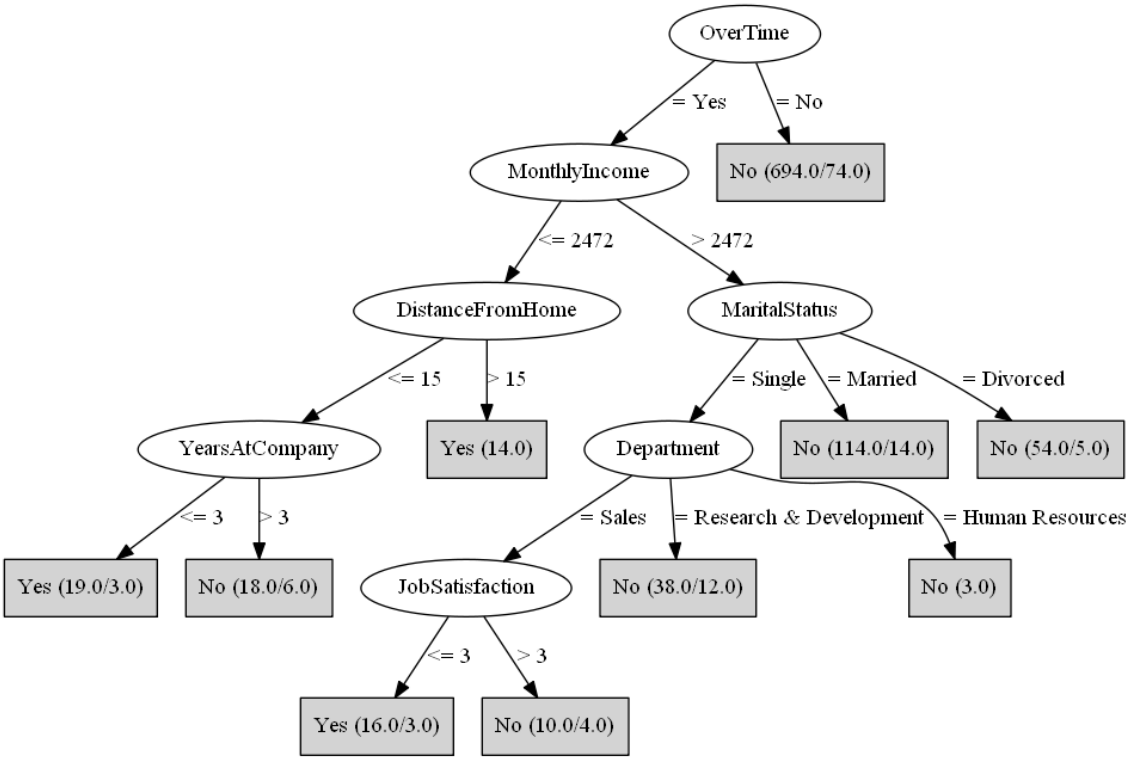

Figure 28 - J48 tree obtained with Weka for case Study 1

The accuracy of the applied algorithm is obtained using cross-validation with 10 folds. The result is an overall accuracy of 84.01%. The precision and recall of every class can be seen in Table 5.

Predicted Attrition	Observed Attrition		
	No	**Yes**	**Class precision**
No	1204	206	85.39%
Yes	29	31	51.67%
Class recall	97.65%	13.08%	84.01%

Table 5 - Accuracy of the J48 tree obtained with Weka for Case Study 1

4.5 Evaluation

In the Modeling phase three models have been obtained and the accuracy for each of the models is known:

- The Watson Analytics CHAID tree had an accuracy of 81%.
- The Rapidminer decision tree had an accuracy of 83.61%.
- The Weka J48 tree had an accuracy of 84.01%.

The highest overall accuracy was provided by the Weka model, with a level of accuracy of 84.01%. Thus, the Weka J48 tree will be the model that will be used for prediction.

The selected model fulfils the condition set for the data mining process to be considered successful, i.e. it has an accuracy level of 84.01%, which is higher than the minimum value (70%).

The model meets the business objectives, since it provides both a visual explanation and a prediction for attrition at the organization.

4.6 Deployment

In the Deployment phase, the deployment strategy must be defined. For this project, the deployment strategy will involve the formulation of actionable conclusions and the writing of the final report, which is, in fact, this very document.

Besides, the whole execution of the project is revised in order to confirm that all the steps have been carried out correctly.

From the final model, actionable conclusions can be obtained. Actionable conclusions are very valuable, since the application of an action can be directly derived from the conclusion itself.

The actionable conclusions extracted from this project are:

- Employees who don't work overtime are likely to stay in the company.
- Employees who work overtime, earn $2472 or less and live further than 15km from work are very likely to leave the company.
- Employees who work overtime, earn $2472 or less, live closer than 15km from work and have been less than 3 years working for the company, are likely to leave voluntarily.
- In addition, employees who work overtime, earn more than $2472, are single and work for the Sales department are highly likely to leave if their job satisfaction is equal or less than 3, which is equivalent to their satisfaction being low, medium or high, according to the description of the possible values of the `JobSatisfaction` variable explained in the Data Understanding phase.

5 CASE STUDY 2: Predicting employee attrition at TAFE

The goal of this study is to determine, using Data Science, the reasons that drive employee attrition in the TAFE Employment Department in Australia.

TAFE (Technical and Further Education) [32] is a training provider that offers practical and industry-relevant education in a great variety of fields for the Queensland state, which is the second largest state in Australia.

The data object of this study has been gathered from exit interviews conducted with departing employees. The information contained in these surveys can reveal important insights that help TAFE develop efficient strategies to retain talent and prevent attrition of valuable employees.

The exit interviews include information related to:

- Employee: current age and gender.
- Job position: employment type, classification and work area.
- Workplace: years of service in that place and specific questions about agreement or disagreement on topics related to the workplace.
- Institute: years of service in that institute and specific questions about the agreement or disagreement on topics related to the institute.
- Induction program: questions about the induction program carried out by the employee, in case he undertook one.
- Separation: reasons for ceasing employment, cessation year, length of service, total years working for TAFE and contributing factors for the cessation.

For this purpose, the Microsoft Excel 2016 software will be used for the EDA (Exploratory Data Analysis) part and the Rapidminer Studio software will be used for the prediction part, along with the Weka software.

The CRISP-DM process model has been the reference model for the execution of this study and the phases will be covered in detail.

5.1 Business Understanding

The business goal of this study is to determine the reasons why employees abandon the TAFE institution and predict whether an employee is going to leave the institution or not, based on available information about the employee. The results of this project will allow TAFE to retain valuable employees by anticipating their attrition.

The hardware available for the execution of this project is a Lenovo Ideapad Z510 laptop, along with an Asus desktop computer with Windows 10, an i5-6600 microprocessor and 32GB RAM.

The data that will be analyzed for this study has been obtained from the Open Data Repository of the Australian Government [33]. It is static data, since the format is a CSV (Comma Separated Values) file and no further data will be collected for the dataset. Metadata for the dataset is not available.

The data mining goal of this study is to extract insights in order to model the reasons why employees leave the TAFE institution voluntarily, and to predict whether an employee is going to abandon the institution based on information available about the worker. To do so, an exploratory data analysis will be carried out for the insight extraction and several prediction methods will be used to obtain a prediction of attrition.

The data mining process will be considered to be successful if the prediction model provides a minimum value of accuracy of 70%.

5.2 Data Understanding

The information in the dataset under study contains the answers to an exit interview carried out on former employees of the TAFE institution.

The dataset contains 72 variables and 702 records. The variables are:

- `Cessation year`: numerical variable indicating the year in which the employee left the institution. It takes values from 2009 to 2013. It contains 7 missing values.
- `Classification`: categorical variable indicating the role of the employee. It can take the values "Administration (AO)", "Apprentice", "Executive (SES/SO)", "Operational (OO)", "Professional Officer (PO)", "Teacher (including LVT)", "Technical Officer (TO)", "Tutor" or "Workplace Training Officer". It contains 106 missing values.
- Contributing factors: series of boolean variables indicating whether that was a contributing factor for separation with the institution. They can take the values True or False.
  - `Career move – private sector`: Contains 265 missing values.
  - `Career move – public sector`: Contains 265 missing values.
  - `Career move – self-employment`: Contains 265 missing values.
  - `Dissatisfaction`: Contains 265 missing values.
  - `Ill health`: Contains 265 missing values.
  - `Interpersonal conflict`: Contains 265 missing values.
  - `Job dissatisfaction`: Contains 265 missing values.
  - `Maternity/Family`: Contains 265 missing values.
  - `Study`: Contains 265 missing values.
  - `Travel`: Contains 265 missing values.
  - `Other`: Contains 265 missing values.
  - `None`: Contains 265 missing values.

- `Current age`: categorical variable indicating the current age of the employee. It can take the values "20 or younger", "21 – 25", "26 – 30", "31 – 35", "36 – 40", "41 – 45", "46 – 50", "51-55" and "56 or older". It contains 106 missing values.
- `Employment type`: categorical variable indicating if the contract is permanent or temporary and full-time or part-time. It can take the values "Contract/casual", "Permanent Full-time", "Permanent Part-time", "Temporary Full-time", or "Temporary Part-time". It contains 106 missing values.
- `Gender`: categorical variable indicating the gender of the employee. It can take the values "Female" or "Male". It contains 106 missing values.
- `Induction`: boolean variable indicating if the employee undertook an induction program or not. It can take the values True or False. It contains 83 missing values.
- `Induction Info`: series of Boolean variables indicating the type of induction program undertaken by the employee. They can take the values True or False.
  - `Corporate Induction`: Contains 270 missing values.
  - `Institute Induction`: Contains 219 missing values.
  - `Team Induction`: Contains 262 missing values.
- `Institute`: categorical variable indicating the Institute where the employee worked. It can take the values "Barrier Reef Institute of TAFE", "Brisbane North Institute of TAFE", "Central Queensland Institute of TAFE", "Metropolitan South Institute of TAFE", "Mount Isa Institute of TAFE", "SkillsTech Australia", "Southbank Institute of Technology", "Southern Queensland Institute of TAFE", "Sunshine Coast Institute of TAFE", "The Bremer Institute of TAFE", "Tropical North Institute of TAFE" or "Wide Bay Institute of TAFE". It contains no missing values.
- Institute views: series of categorical variables indicating the level of agreement or disagreement of the employee with statements about the Institute. They can take the values "Agree", "Disagree", "Neutral", "Not Applicable", "Strongly Agree" or "Strongly Disagree".
  - `I feel the senior leadership had a clear vision and direction`: Contains 94 missing values.

- I was given access to skills training to help me do my job better: Contains 89 missing values.
- I was given adequate opportunities for personal development: Contains 92 missing values.
- I was given adequate opportunities for promotion within the Institute: Contains 94 missing values.
- I felt the salary for the job was right for the responsibilities I had: Contains 87 missing values.
- The organization recognized when staff did good work: Contains 95 missing values.
- Management was generally supportive of me: Contains 88 missing values.
- Management was generally supportive of my team: Contains 94 missing values.
- I was kept informed of the changes in the organization which would affect me: Contains 92 missing values.
- Staff morale was positive within the Institute: Contains 100 missing values.
- If I had a workplace issue it was dealt with quickly: Contains 101 missing values.
- If I had a workplace issue it was dealt with efficiently: Contains 105 missing values.
- If I had a workplace issue it was dealt with discreetly: Contains 101 missing values.

- Length of service current: categorical variable indicating the length of service at the current workplace (in years). It can take the values "Less than 1 year", "1-2", "3-4", "5-6", "7-10", "11-20" or "More than 20 years". It contains 106 missing values.
- Length of service overall: categorical variable indicating the length of service at the Institute (in years). It can take the values "Less than 1 year", "1-2",

"3-4", "5-6", "7-10", "11-20" or "More than 20 years". It contains 106 missing values.

- `Main factor`: categorical variable indicating the main factor for leaving. It can take the values "Career Move - Private Sector", "Career Move - Public Sector", "Career Move - Self-employment", "Dissatisfaction with Institute", "Ill Health", "Interpersonal Conflict", "Job Dissatisfaction", "Maternity/Family", "Study", "Travel" or "Other". It contains 589 missing values.
- `Reason for ceasing employment`: categorical variable indicating the reason for separation. It can take the values "Contract Expired", "Resignation", "Retirement", "Retrenchment/ Redundancy", "Termination", or "Transfer". It contains 1 missing value.
- `Record ID`: numerical variable indicating the employee's id number in the dataset. It takes values from 634133009996094000 to 635073030973791000. It contains no missing values.
- `Work area`: categorical variable indicating the work area of the employee. It can take the values "Delivery (teaching)" or "Non-Delivery (corporate)". It contains no missing values.
- Work unit views: series of categorical variables indicating the level of agreement or disagreement of the employee with statements about the work unit. They can take the values "Agree", "Disagree", "Neutral", "Not Applicable", "Strongly Agree" or "Strongly Disagree".
  - `I was satisfied with the quality of the management and supervision within my work unit`: Contains 93 missing values.
  - `I worked well with my colleagues`: Contains 97 missing values.
  - `My job was challenging and interesting`: Contains 95 missing values.
  - `I was encouraged to use my initiative in the course of my work`: Contains 92 missing values.
  - `I had sufficient contact with other people in my job`: Contains 89 missing values.

- I was given adequate support and co-operation by my peers to enable me to do my job: Contains 93 missing values.
- I was able to use the full range of my skills in my job: Contains 93 missing values.
- I was able to use the full range of my abilities in my job: Contains 94 missing values.
- I was able to use the full range of my knowledge in my job: Contains 94 missing values.
- My job provided sufficient variety: Contains 91 missing values.
- I was able to cope with the level of stress and pressure in my job: Contains 92 missing values.
- My job allowed me to balance the demands of work and family to my satisfaction: Contains 91 missing values.
- My supervisor gave me adequate personal recognition and feedback on my performance: Contains 96 missing values.
- My working environment was satisfactory e.g. sufficient space, good lighting, suitable seating and working area: Contains 92 missing values.
- I was given the opportunity to mentor and coach others in order for me to pass on my skills and knowledge prior to my cessation date: Contains 93 missing values.
- There was adequate communication between staff in my unit: Contains 99 missing values.
- Staff morale was positive within my work unit: Contains 93 missing values: Contains 96 missing values.

- Workplace: series of categorical variables indicating the level of agreement or disagreement of the employee with statements about the workplace.
  - Did you and your Manager develop a Performance and Professional Development Plan (PPDP)? Contains 94 missing values.
  - Does your workplace promote a work culture free from all forms of unlawful discrimination? Contains 108 missing values.

- `Does your workplace promote and practice the principles of employment equity?` Contains 115 missing values.
- `Does your workplace value the diversity of its employees?` Contains 116 missing values.
- `Would you recommend the Institute as an employer to others?` Contains 121 missing values.

An exploratory data analysis suggests hypotheses about the causes of observed phenomena, and can in this study help understand the reasons that cause employee attrition at the TAFE institution.

The exploratory analysis does not pursue a specific goal, as it is the case of prediction algorithms, which expect an output prediction. Instead, the EDA is an approach to knowledge discovery in which the objectives are not fixed. However, since the relevant discoveries in this case are related to the exit of the employees, the variable `Reasons for ceasing employment` and the relationships with other relevant variables in the dataset will be studied in depth.

The distribution of values of the variable `Reasons for ceasing employment` is obtained in order to have a first understanding of the reasons why employees are leaving TAFE.

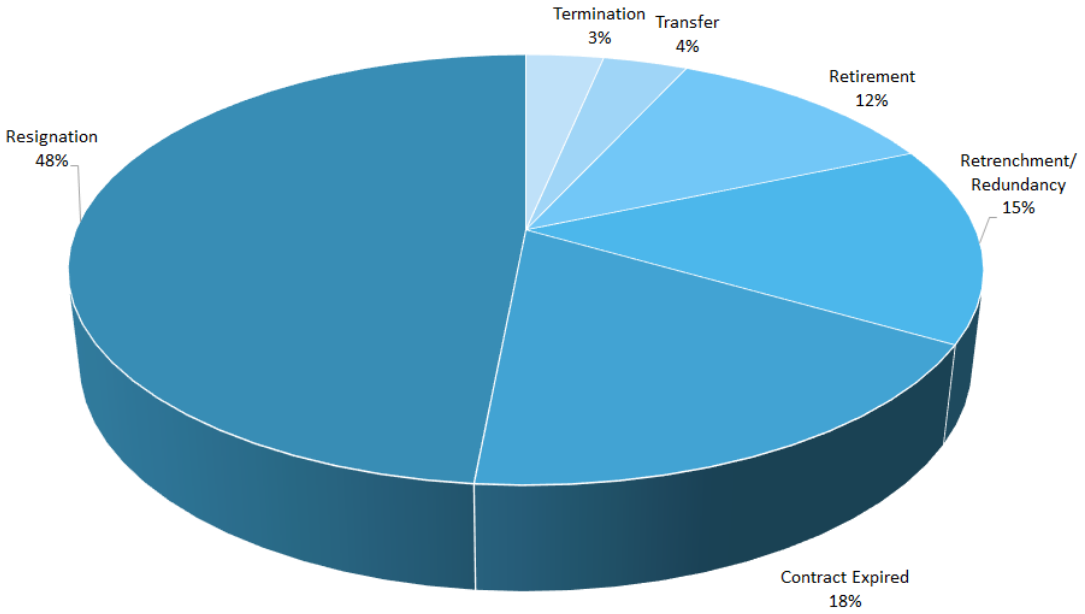

Figure 29 - Reasons for ceasing employment

This pie chart reveals that the most common reason for separation with TAFE is resignation, i.e. abandoning the institution, with 340 cases out of 702 total cases in the dataset (48%). This information is very meaningful for the HR department, so the reasons for this high rate of attrition must be studied.

In the first place, and being aware that gender inequality is a reality in every country's workforce, the relationship between the type of exit and the gender of the employees will be analyzed. For that purpose, a stacked bar chart is obtained representing the percentage of male and female employees per type of exit. Besides, a horizontal line crossing the bar chart indicates the overall percentage of women in the dataset.

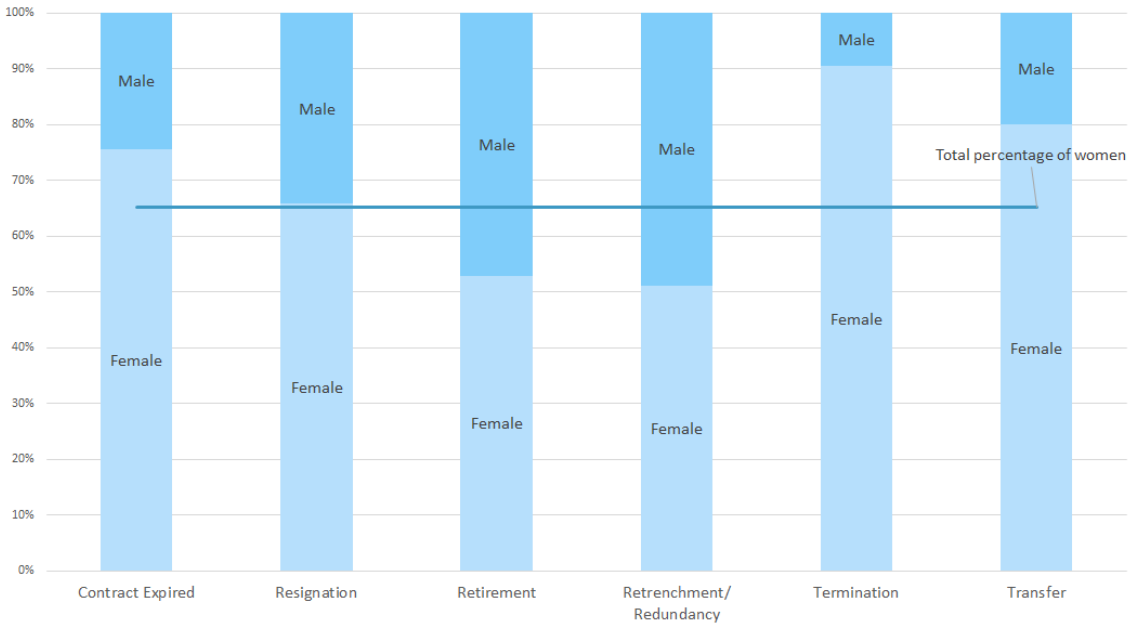

Figure 30 - Separation type per gender

The bar chart shows that the percentage of women being transferred or fired is much higher than the average. With respect to the voluntary exit, the percentage of women and men is completely balanced.

In addition, research published in Harvard Business Review shows that young talented employees are likely to leave their company and search for a new job [34]. Thus, the relationship between the separation type and the age range of the employees must be studied. To do so, a stacked bar chart is plotted representing the percentage of every type of exit per age range.

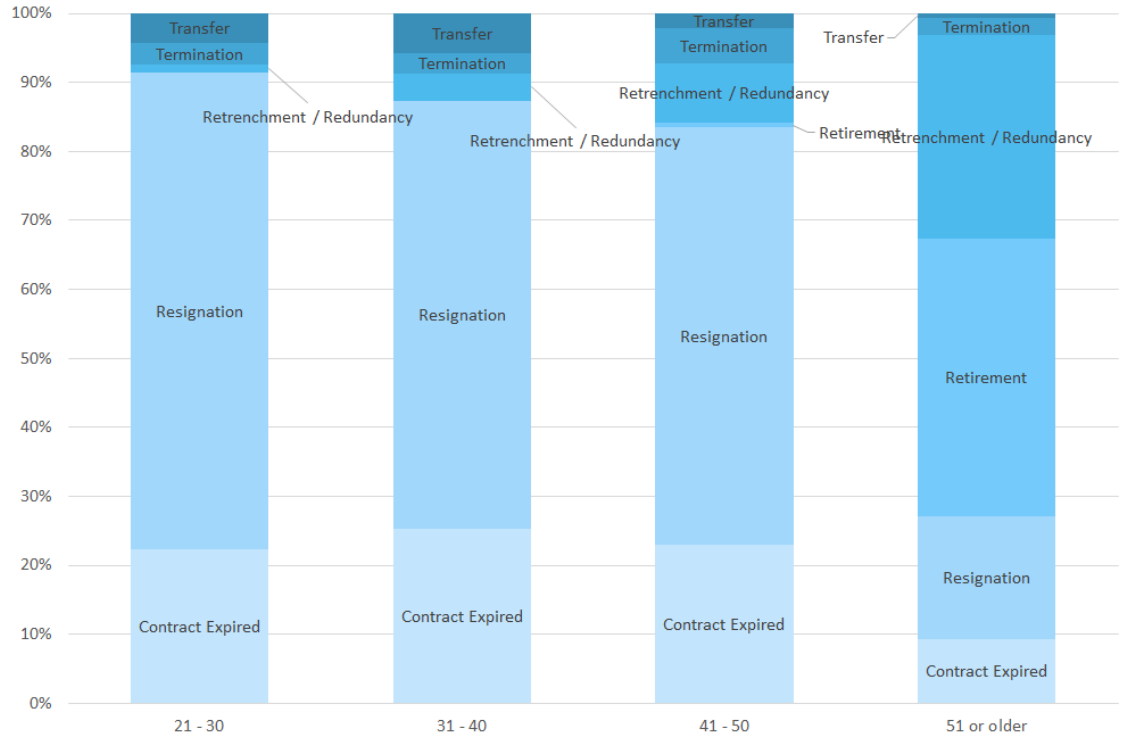

Figure 31 - Separation type per age range

The chart reveals that employees in the 21-30 age range are the most likely ones to resign, confirming the results of the research carried out by Harvard Business Review. On the other hand, employees older than 51 are more likely to retire or suffer from retrenchment plans at the institution.

To continue with the study, the values of the variable that represents the main factor that triggered employees' exit are looked at closely, filtering only those employees whose type of exit corresponded to resignation. To that end, a pie chart is obtained representing the numerical proportion of all the values of the `Main factor` variable for those employees who have voluntarily left the institution.

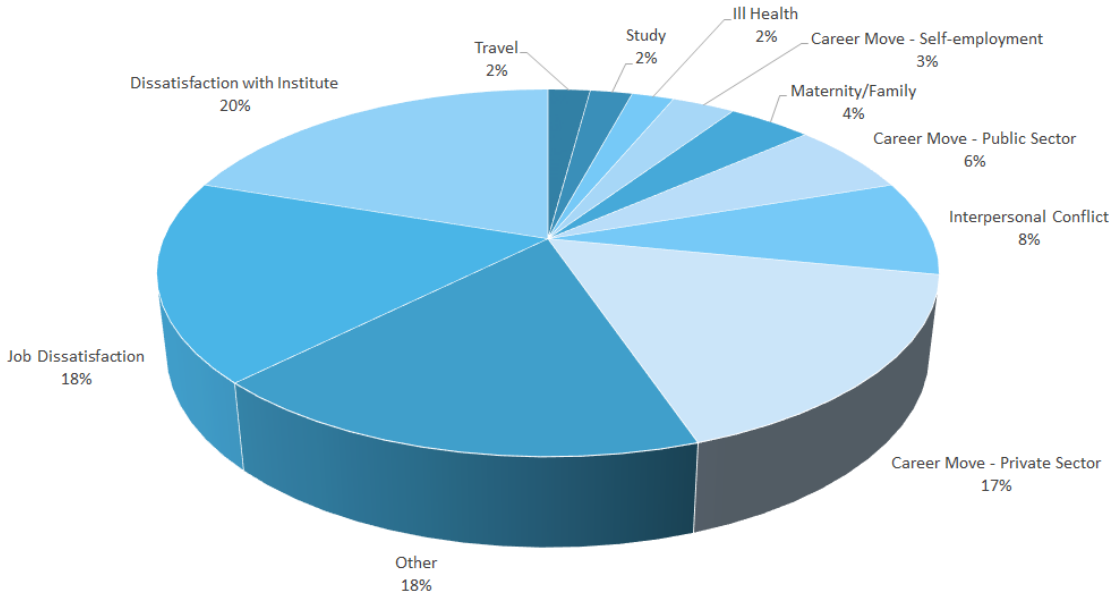

Figure 32 - Main factor for resignation

The graph clearly shows that the main reason for the exit of those employees who voluntarily left their job is the dissatisfaction with the Institute where they lecture. Thus, the situation at every institute should be analyzed to uncover more insights.

The previous plot reveals a problem of dissatisfaction of the employees with their Institute. Thus, the resignation rate per Institute will be studied to see if there is any correlation between employee attrition and any particular Institute. To do so, a bar chart is obtained representing the number of employees who resigned per Institute, along with a horizontal line that crosses the chart and indicates the average percentage rate.

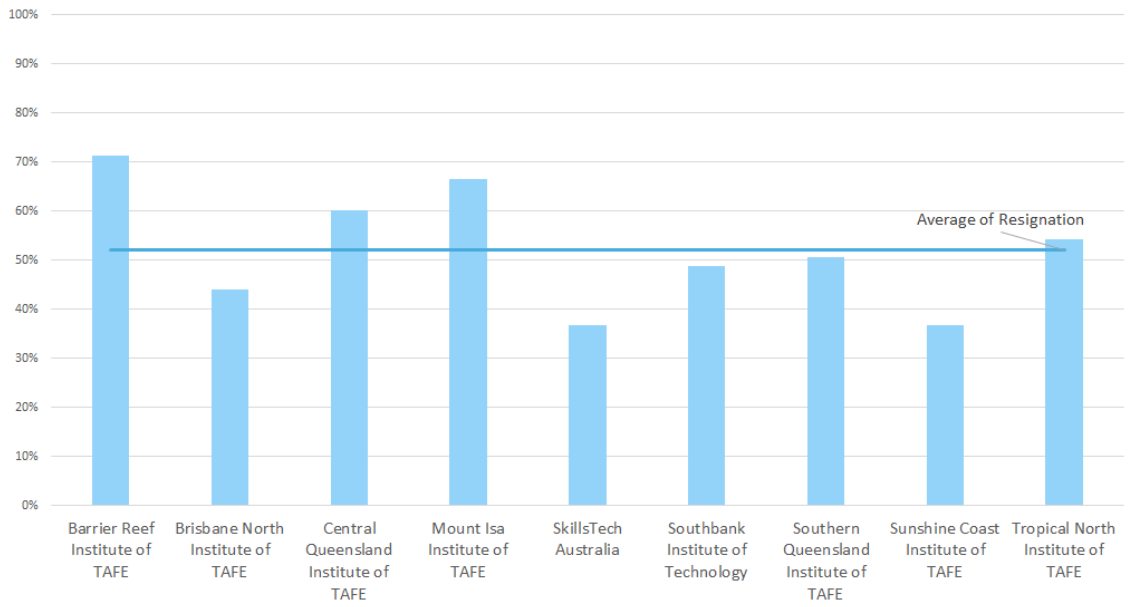

Figure 33 - Percentage of resignation per Institute

The bar chart shows a rate of resignation more than 10% higher than the average for the Barrier Reef Institute and the Mount Isa Institute. Both institutes are located in North Queensland, the northern region of the state of Queensland. Thus, the North Region of Queensland has a problem with the resignation rates at its Institutes.

Continuing with the analysis, the effect of workplace induction programmes on the employees will be studied by means of a stacked bar chart representing the percentage of employees who undertook workplace induction per type of exit. The analysis was only performed on those employees who have been working less than 2 years at the institution, since from then on the effect of the training programmes will not have an effect on their decision to stay or leave the institution.

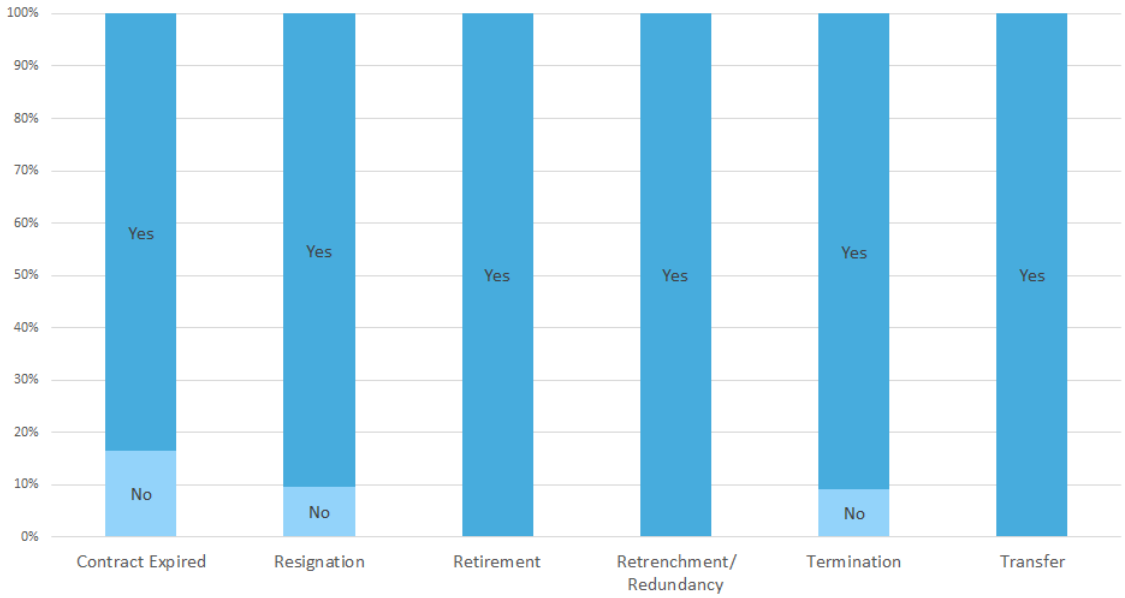

Figure 34 - Workplace induction

The plot shows high levels of participation in workplace induction programmes, with more than 90% involved in such programmes, even for those employees who voluntary left.

The next step for the analysis is the study of the relationship between the type of contract (temporary, permanent, full-time or part-time) and the separation with the institution. A stacked bar chart representing the percentage of resignation and no resignation per every contract type is obtained, along with a horizontal line that represents the average percentage of resignation.

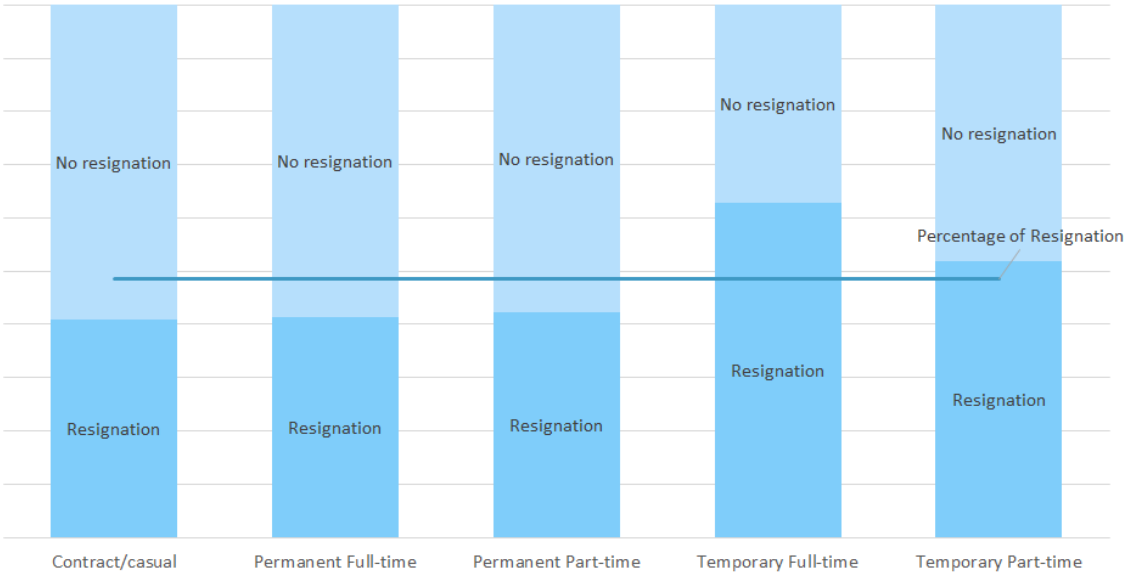

Figure 35 - Resignation per employment type

The chart clearly shows that employees with temporary contracts have a higher rate of resignation than those with permanent contracts, being more than 10% higher than the average in the case of temporary full-time contracts.

To conclude with the exploratory analysis, the relationship between the type of exit and the years of service is studied by means of a stacked bar chart representing the percentage of resignation and no resignation per years of service.

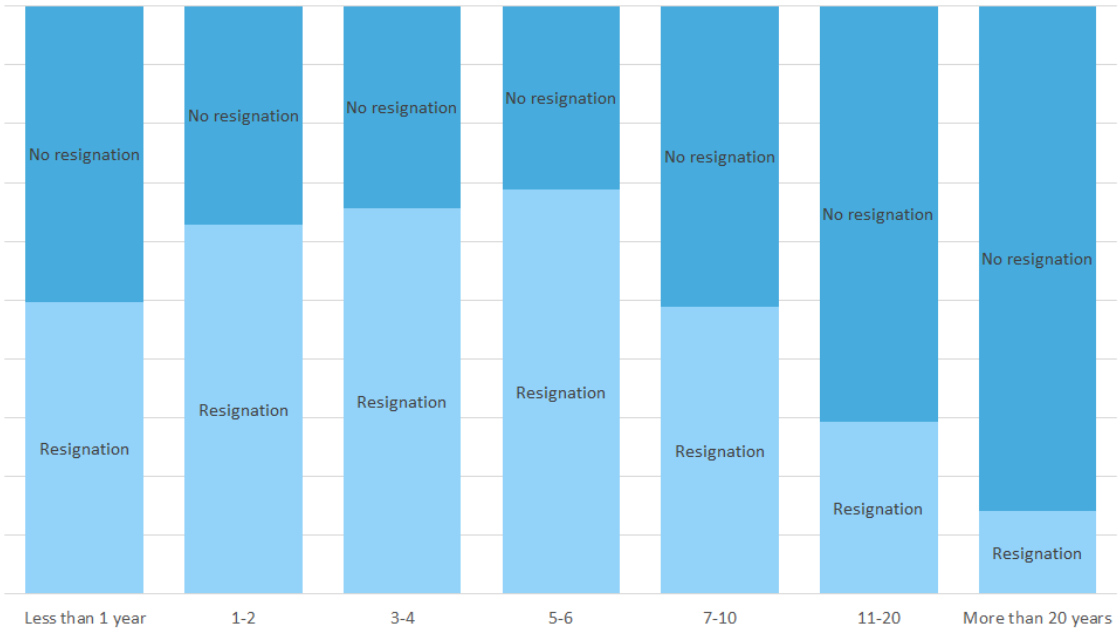

Figure 36 - Resignation per years of service

The plot reveals an increasing rate of attrition for employees that have been working for the institution less than 6 years. From then on, the rate of resignation falls with the increase of the number of the years at the institution, which is an expectable result, since the probability that an employee abandons his job descends when his tenure at the company increases.

5.3 Data Preparation

In the Data Preparation phase, a deep dataset cleansing process has taken place. All the variables have been renamed appropriately, since the variables contained in the original dataset did not have a correct name format and were not easy to understand and work with.

The variable `Record ID` represents the ID of the employee in the dataset, and thus is meaningless for the study. Consequently, it is removed from the dataset.

The values for the `Current age` variable were given in intervals of 5 years, e.g. 21-25, 26-30, and they were converted to 10-year interval format, e.g. 21-30 for simplicity in the study of the age ranges.

In the previous phase of the process, the possible values of the main variable in the dataset, `Reasons for ceasing employment`, have been shown next to their frequency in the dataset. Those values are:

- Resignation
- Contract expired
- Retrenchment/Redundancy
- Retirement
- Transfer
- Termination

All the values except Resignation represent forced exits from the institution, i.e. not voluntary. With regard to that, a synthetic variable called `Resignation` is created from the `Reasons for ceasing employment` variable. This new variable has the value Resignation when the value of the `Reasons for ceasing employment` variable is Resignation and the value No Resignation otherwise.
In addition, one of the records of the dataset has a missing value for the `Reasons for ceasing employment` variable, and thus is removed from the dataset.
Hence, the final dataset contains 701 records and 72 variables.

5.4 Data Modeling

In the Data Modeling phase machine learning algorithms for prediction will be applied in order to obtain a reliable forecast for resignation at the institution.

5.4.1 Neural Network (Rapidminer)

A neural network is obtained with Rapidminer's Neural Net operator, which learns a model by means of a feed-forward neural network trained by a propagation algorithm (multi-layer perceptron) [35]. The synthetic variable generated in the Data Preparation phase, which contains the value Resignation or No resignation for every employee, is used as target. The data to be fed to the neural net needs to be preprocessed, since it cannot deal with nominal values or missing values. Therefore, all the text values are converted to numerical values, and the missing values are replaced by the average value of the values in the column. The parameters are tuned in order to obtain the best model. The whole Rapidminer process is shown in Figure 37.

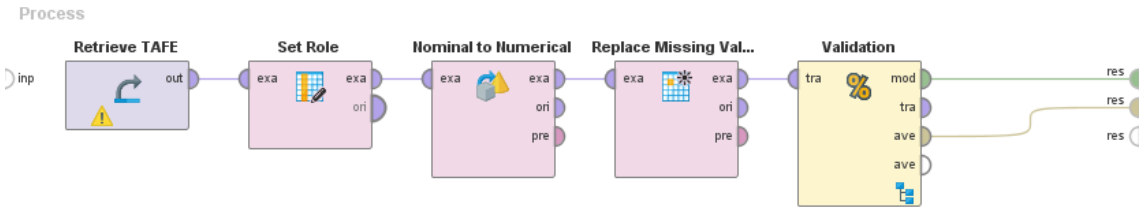

Figure 37 - Rapidminer process for the neural network of Case Study 2

The Retrieve TAFE operator loads the TAFE dataset to be used. The next operator, Set Role, sets the role "label" to the `Resignation` variable, in order to specify that it is the target attribute. The Nominal to Numerical operator converts every nominal variable into numerical, since the neural network cannot handle nominal values. Next, the Replace Missing Values operator replaces every missing value with the average of all the values in the attribute, since the neural network cannot handle missing values. Finally, the X-Validation operator performs a cross-validation with 10 folds to estimate the accuracy of the neural network.

The X-Validation operator is a nested operator. The operators that are inside the X-Validation operator can be seen in Figure 38.

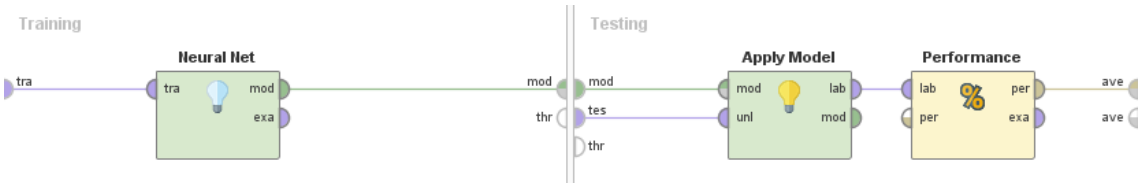

Figure 38 - Rapidminer process for the neural network of Case Study 2: X-Validation

The training subprocess is used for training the neural network. The trained model is then applied in the testing subprocess, where the performance of the neural network is also measured.

The neural network is trained in 200 cycles with a learning rate of 0.1 and a momentum of 0.2. Besides, the input data is shuffled before the training of the model, and all the attributes are normalized. The neural network is set to have 2 hidden layers, with size equal to *(number of attributes + number of classes) / 2 + 1* each.

The accuracy of the applied model is obtained by means of cross-validation with 10 folds. The result is an overall accuracy of 72.74%. The precision and recall of every class can be seen in Table 6.

Predicted Attrition	Observed Attrition		
	No resignation	**Resignation**	**Class precision**
No resignation	279	109	71.91%
Resignation	82	231	73.80%
Class recall	77.29%	67.94%	72.74 %

Table 6 - Accuracy of the neural network obtained with Rapidminer for Case Study 2

5.4.2 Logistic Regression (Rapidminer)

A logistic regression is obtained with Rapidminer's Logistic Regression operator, using as target the synthetic binominal variable with values Resignation or No resignation generated in the Data Preparation phase. The data to be fed to the

Logistic Regression operator needs to be processed, since the operator cannot handle nominal values of the target variable or any missing values. Thus, the values of the target variable are converted to numerical values, and the missing values are replaced by the average value of the values in the column. The parameters of the operator are set to their optimal values. The whole Rapidminer process is shown in Figure 39.

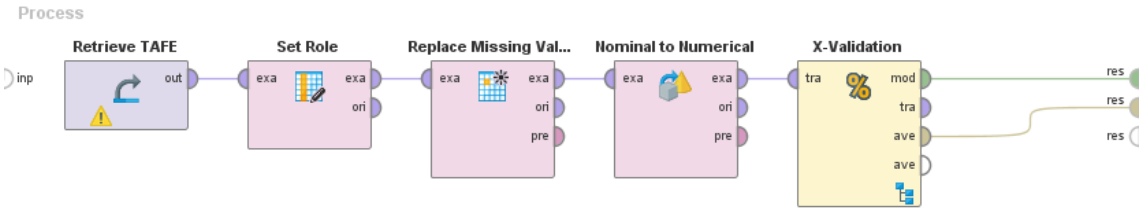

Figure 39 - Rapidminer process for the logistic regression of Case Study 2

The Retrieve TAFE operator loads the TAFE dataset to be used. The next operator, Set Role, sets the role "label" to the `Resignation` variable, in order to specify that it is the target attribute. The Replace Missing Values operator replaces every missing value with the average of all the values in the attribute, since the Logistic Regression operator cannot handle missing values. Next, the Nominal to Numerical operator converts every nominal variable into numerical, since the Logistic Regression operator cannot handle nominal values. Finally, the X-Validation operator performs a cross-validation with 10 folds to estimate the accuracy of the logistic regression.

The X-Validation operator is a nested operator. The operators that are inside the X-Validation operator can be seen in Figure 40.

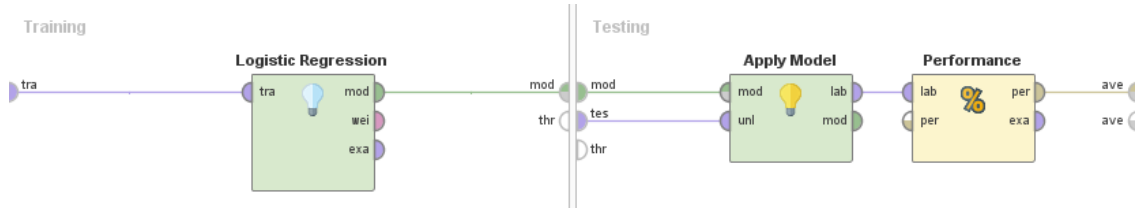

Figure 40 - Rapidminer process for the logistic regression of Case Study 2: X-Validation

The accuracy of the applied model is obtained by means of cross-validation with 10 folds. The result is an overall accuracy of 67.18%. The precision and recall of every class can be seen in Table 7.

Predicted Attrition	Observed Attrition		
	No resignation	**Resignation**	**Class precision**
No resignation	306	175	63.62%
Resignation	55	165	75.00%
Class recall	84.76%	48.53%	67.18% %

Table 7 - Accuracy of the logistic regression obtained with Rapidminer for Case Study 2

5.4.3 J48 Tree (Weka)

A C4.5 tree is obtained using the J48 class from the Weka software. The parameters are tuned to achieve the best possible model. The reduced error pruning parameter is set to true and the minimum number of instances per node is set to 10.

As in the Case Study 1, Graphviz will be used together with Weka to obtain the decision tree. The two commands used for this purpose can be seen in Figure 41.

Figure 41 - Weka and Graphviz commands to obtain the J48 tree for Case Study 2

The output tree can be seen in Figure 42.

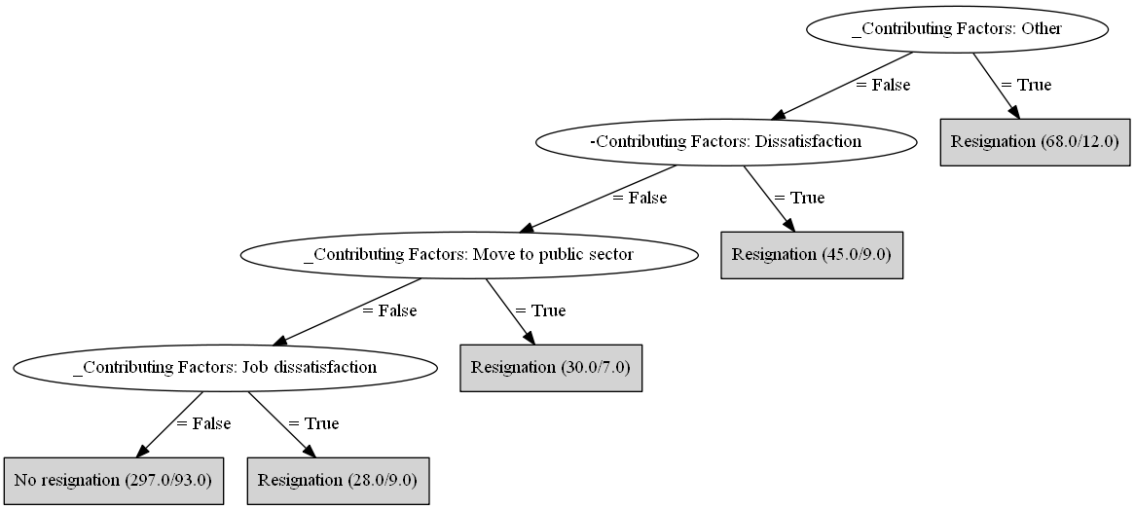

Figure 42 - J48 tree obtained with Weka for Case Study 2

The accuracy of the applied algorithm is obtained using cross validation with 10 folds. The result is an overall accuracy of 71.75%. The precision and recall of every class can be seen in Table 8.

Predicted Attrition	Observed Attrition		
	No resignation	**Resignation**	**Class precision**
No resignation	304	141	68.31%
Resignation	57	199	77.73%
Class recall	84.21%	58.53%	71.75%

Table 8 - Accuracy of the J48 tree obtained with Weka for Case Study 2

5.5 Evaluation

In the Modeling phase, three models have been applied and the accuracy for each model has been obtained:

- The neural network had an accuracy of 72.74%.
- The logistic regression had an accuracy of 67.18%.
- The Weka J48 tree had an accuracy of 71.75%.

The higher overall accuracy was provided by neural network, with a level of accuracy of 72.74%. Thus, the Neural Net operator from the Rapidminer software will be the model used for prediction.

The selected model fulfils the condition that was set in the Business Understanding phase for the data mining process to be considered successful, i.e. it has an accuracy level of 72.74%, which is higher than the minimum value (70%).

Furthermore, the model meets the business objectives, since it delivers a successful prediction for resignation in the organization, along with a series of visual explanations of the reasons why employees voluntarily leave.

5.6 Deployment

The deployment strategy for this case study will consist of the development of the final report of the project, which will contain all the actionable conclusions that have been drawn from the analysis.

The execution of all the phases of the project is reviewed with the aim of confirming that no mistakes have been made and that the methodology has been correctly applied.

From the Exploratory Data Analysis carried out in the Modeling phase, actionable conclusions can be reached. The conclusions extracted from this study are:

- Employees with less than 6 years of service at the institution have a higher probability to abandon their job than those who have overcome the 5-6 years' barrier. Therefore, TAFE should start an employee loyalty programme to retain talent and prevent the voluntary exit of those employees who have been less than 6 years in the institution.

- 90% of the employees who left during their first two years at TAFE had followed an induction programme, which reveals the low effectiveness of such programmes. Thus, the contents of the induction programmes should be revised, since they are very expensive for the institution and they are not providing a successful outcome: prevent attrition.
- Employees with a temporary contract tend to leave the company, as opposed to those with a permanent contract. That being so, the possibilities for promotion for employees with a temporary contract should be increased, offering them opportunities of upgrading to a permanent contract.
- The main cause for resignation is the dissatisfaction with the Institute, being Barrier Reef Institute and Mount Isa Institute the Institutes with the highest rates of dissatisfaction. For that reason, TAFE should open an investigation on the work conditions at those Institutes and in the North Region to determine the specific causes of generalized employee dissatisfaction.
- Employees in the 21-30 age range are more likely to resign. Thus, TAFE should focus on that age range in order to retain young talent at the institution by investing in their development and offering them growth opportunities and possibilities to learn new skills.

6 Conclusions and Future Work

6.1 Conclusions

By means of Data Science the problem of employee turnover that every company fears can be analyzed and studied in order to mitigate it, reducing costs of replacing valuable employees and maximizing profits.

An EDA (Exploratory Data Analysis) has been carried out to uncover the reasons that lead to an employee's voluntary exit, and Prediction algorithms have been used to predict when an employee is going to resign from the company.

The results of the Decision Tree applied to the dataset used in the first case study gave the managers of the company very useful information that they could use to reduce their employee turnover rate:

- Reconsider the salary of the employees who earn less than $2.694 and assess the possibility of rising their salary.
- Study the conditions of the employees working in the sales department and determine the reasons for their dissatisfaction with such department.
- Offer incentives and growth possibilities inside the company to those employees younger than 34 who have been working for the company for less than 2 years.

Equivalently, the results obtained with the analysis carried out on the dataset used in the second case study provided the managers of TAFE with valuable and reliable information that could be used to avoid attrition at the institution:

- Start and employee loyalty program to prevent the voluntary exit of the employees with less than 6 years in the institution.
- Review the contents of the induction programs to improve their effect on new employees.

- Offer opportunities for promotion to employees with temporary contracts.
- Investigate the work conditions at the Mount Isa and Reef Barrier Institutes to uncover the reasons of the dissatisfaction of the employees working in those institutes.
- Invest in the development of those employees in the 21-30 age range and offer them possibilities to learn new skills and to grow inside the company.

This illustrates the great range of possibilities that Data Science offers to help managers make data-driven decisions within the staffing process.

The general conclusions that have been extracted from the execution of the project are:

- The particular conclusions obtained for the two case studies can be generalized and extended to other similar scenarios.
- CRISP-DM is the *de-facto* standard methodology for any Data Science project, and thus represents a reference model for the execution of any project on People Analytics. It provides the methodology for the correct development of a successful Data Science project.
- The number of available Data Science tools is huge, and a big part of them are free licensed. The choice of using one or another is up to the Data Scientist, since all of them present advantages and disadvantages. Watson Analytics provides an easy way of performing analysis and prediction on data, but the possibilities of tuning any parameters is non-existent, and the choice of available algorithms is very limited. On the contrary, Rapidminer allows the user to choose among a great variety of models and algorithms, and all their parameters can be tuned. Weka has the advantage of being open source, in addition to providing a huge variety of algorithms. However, it has the drawback of the lack of customization possibilities for the output graphs.

6.2 Future Work

As future lines of work on People Analytics, some ideas are proposed:

- Extend the study to larger datasets containing real data with a higher number of variables and records.
- Re-do the analysis using programming languages instead of software tools, such as Python [36] or R [37].
- Deploy an application that automatically returns the attrition prediction for an employee, based on information given to the application.

7 Project Planning

7.1 Schedule

The schedule of a project is an inventory of the project's milestones, activities and deliverables, along with their start and finish dates. It is necessary for a correct execution of the project in terms of time and resources. One of the many methods used for project scheduling is the Gantt chart, which is a type of bar chart that illustrates the duration of the terminal elements and summary elements of a project.

The scheduling of this project is divided into four main tasks that contain several subtasks:

- Task 1: Study and research. This task involves all the prior study and research that needs to be done before the execution of the actual project starts. It can be split into three subtasks:
  - Subtask 1: Study of concepts. This subtask involves the study of all the concepts in the theoretical background needed for the execution of the project. Besides, it also involves the gathering of information about all the available tools and the decision of the tools to be used. Length: 20 days.
  - Subtask 2: Learning the tools. This subtask involves the installation and familiarization with the tools that will be used. Length: 13 days.
  - Subtask 3: Dataset search. This subtask involves the search in all available open data sources for the datasets needed for the project. Length: 2 days.
- Task 2: Case study 1. This task involves all the work related to the first case study. It cannot be started until the Study and research task has been finished. This task can be broken into three subtasks:
  - Subtask 1: Data understanding and preparation. This task corresponds to the first, second and third phases of the CRISP-DM model applied to this study. Length: 5 days.

    - Subtask 2: Modeling. This task corresponds to the fourth phase of the CRISP-DM model applied to this study. Length: 5 days.
    - Subtask 3: Evaluation. This task corresponds to the fifth phase of the CRISP-DM model applied to this study. Length: 2 days.
- Task 3: Case study 2. This task involves all the work related to the second case study. It cannot be started until the Case study 1 task has been finished, since both case studies will not be carried out simultaneously. This task can be broken into three subtasks:
    - Subtask 1: Data understanding and preparation. This task corresponds to the first, second and third phases of the CRISP-DM model applied to this study. Length: 14 days.
    - Subtask 2: Modeling. This task corresponds to the fourth phase of the CRISP-DM model applied to this study. Length: 5 days.
    - Subtask 3: Evaluation. This task corresponds to the fifth phase of the CRISP-DM model applied to this study. Length: 2 days.
- Task 4: Report. This task involves the writing of the final report. It cannot be started until all the previous tasks have been finished. It can be broken into two subtasks:
    - Subtask 1: Report writing. This subtask consists of the writing of the report. Length: 30 days.
    - Subtask 2: Report revision and corrections. This subtask involves all the revisions and corrections performed by the Senior Data Scientist that lead to the rewriting of some parts of the report in order to achieve the final version of the report and thus, the end of the execution of the project. Length: 5 days.

The Gantt chart for the project has been created with a web application [38] and can be seen in Figure 43.

Figure 43 - Gantt chart

7.2 Costs Projection

7.2.1 Hardware

The hardware used for the project consists of a Lenovo Ideapad Z510 laptop, with an estimated lifetime of 3 years, along with an Asus desktop computer with Windows 10, an i5-6600 microprocessor and 32GB RAM, with an estimated lifetime of 3 years. The total length of the project has been explained in the Schedule section, and was 93 days.

Concept	Price	Depreciation period	Days	Cost
Laptop	970€	1095 days (3 years)	93 days	82.38€
Desktop computer	1300€	1095 days (3 years)	93 days	110.41€

Table 9 - Hardware costs

Thus, the total cost of the hardware material is 192.79€.

7.2.2 Human Resources

The human resources available for the project consist of a Junior Data Scientist who will have a dedication of 4 hours/day during the 93 days of the project execution and a Senior Data Scientist who will have a dedication of 10% of the Junior Data Scientist's dedication.

Profile	Cost per hour	Hours	Cost
Junior Data Scientist	20€	372 hours (4 hours/day for 93 days)	7440€
Senior Data Scientist	50€	37.2 hours (10% of the 372 hours dedicated by the Junior Data Scientist)	1860€

Table 10 - Human resources costs

Thus, the total cost of the human resources is 9300€.

7.2.3 Total Costs

Indirect costs must be taken into account, and they represent 20% of the costs of personnel and hardware. Thus, the indirect costs will be: $\frac{20}{100} * (9300 + 192.79) =$ 1898.56€.

Concept	Total
Hardware	9300€
Human resources	192.79€
Indirect costs	1898.56€
Total	11391.35€

Table 11 - Total costs

Thus, the total cost of the project is 11391.35€.

References

[1] E. Gallardo-Gallardo, N. Dries and T. F- González-Cruz, "What is the meaning of 'talent' in the world of work?," *Human Resource Management Review,* vol. 23, no. 4, pp. 290-300, 08 2013.

[2] H. Boushey and S. J. Glynn, "Center for American Progress," 16 11 2012. [Online]. Available: https://www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf. [Accessed 27 08 2016].

[3] C. Rudin and S. Elston, "edX," 2015. [Online]. Available: https://courses.edx.org/courses/course-v1:Microsoft+DAT203x+3T2015/courseware/5bbcfcf04e6d49bda1eecb1e1c0bfc24/a57949170b154fdd87057996c87717c7/. [Accessed 27 08 2016].

[4] A. B. Munir, S. H. M. Yasin and F. Muhammad-Sukki, "Big Data: Big Challenges to Privacy and Data Protection," *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering,* vol. 9, no. 1, p. 355, 2015.

[5] "REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016," 04 05 2016. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN. [Accessed 03 09 2016].

[6] "Agencia Estatal Boletín Oficial del Estado," 13 12 1999. [Online]. Available: https://www.boe.es/boe/dias/1999/12/14/pdfs/A43088-43099.pdf. [Accessed 03 09 2016].

[7] V. Dhar, "Data Science and Prediction," *Communications of the ACM,* vol. 56, no. 12, pp. 64-73, 2013.

[8] C.-Y. Zhang and C. P. Chen, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences,* vol. 275, pp. 314-347, 2014.

[9] O. Maimon and L. Rokach, "Introduction to Knowledge Discovery in Databases," in *Data Mining and Knowledge Discovery Handbook*, New York, Springer, 2010, p. 1.

[10] H. J. Watson and B. H. Mixon, "The Current State of Business Intelligence," *Computer,* vol. 40, no. 9, pp. 96-97, 2007.

[11] R. Schutt and C. O'Neil, Doing Data Science: Straight Talk from the Frontline, Sebastopol: O'Reilly Media, 2013.

[12] C. Massey, M. Haas and M. Bidwell, "Coursera," 2016. [Online]. Available: https://www.coursera.org/learn/wharton-people-analytics. [Accessed 30 08 2016].

[13] «Watson Analytics,» [En línea]. Available: http://www-03.ibm.com/software/products/en/watson-analytics. [Último acceso: 27 08 2016].

[14] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2010. [Online]. Available: http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html. [Accessed 30 08 2016].

[15] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing,* vol. 5, no. 4, p. 13, 2000.

[16] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2002. [Online]. Available: http://www.kdnuggets.com/polls/2002/methodology.htm. [Accessed 16 08 2016].

[17] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2004. [Online]. Available: http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm. [Accessed 16 08 2016].

[18] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2016. [Online]. Available: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm. [Accessed 16 08 2016].

[19] "Data Mining, Analytics, Big Data, and Data Science," KDnuggets, 2016. [Online]. Available: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html. [Accessed 16 08 2016].

[20] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crisp-dm.eu/reference-model/. [Accessed 16 08 2016].

[21] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crisp-dm.eu/business-understanding/. [Accessed 16 08 2016].

[22] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crisp-dm.eu/data-understanding/. [Accessed 16 08 2016].

[23] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crisp-dm.eu/data-preparation/. [Accessed 16 08 2016].

[24] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crisp-dm.eu/modelling/. [Accessed 16 08 2016].

[25] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crisp-dm.eu/evaluation/. [Accessed 16 08 2016].

[26] "CRISP-DM by Smart Vision Europe," 2015. [Online]. Available: http://crisp-dm.eu/deployment/. [Accessed 16 08 2016].

[27] D. Wackerly, W. Mendenhall y R. L. Scheaffer, Mathematical Statistics with Applications, Thomson, 2008 .

[28] «Correlation Matrix - Rapidminer Documentation,» 2016. [En línea]. Available: http://docs.rapidminer.com/studio/operators/modeling/correlations/correlation_matrix.html. [Último acceso: 2016 08 27].

[29] «IBM Knowledge Center,» [En línea]. Available: https://www.ibm.com/support/knowledgecenter/SS4QC9/com.ibm.solutions.wa_an_overview.2.0.0.doc/decision_rule.html. [Último acceso: 26 08 2016].

[30] «Statistics Solutions,» [En línea]. Available: http://www.statisticssolutions.com/non-parametric-analysis-chaid/. [Último acceso: 26 08 2016].

[31] R. Kohavi, «A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,» de *International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, 1995.

[32] "TAFE Queensland," 2014. [Online]. Available: http://tafeqld.edu.au/about-us/. [Accessed 24 08 2016].

[33] "Data | Queensland Government," 2016. [Online]. Available: https://data.qld.gov.au/dataset/tafe-employee-exit-survey. [Accessed 02 07 2016].

[34] M. Hamori, J. Cao and B. Koyuncu, "Why Top Young Managers Are in a Nonstop Job Hunt," *Harvard Business Review,* Vols. July-August 2012, 2012.

[35] "Rapidminer Documentation," Rapidminer GmbH, 2016. [Online]. Available: http://docs.rapidminer.com/studio/operators/modeling/predictive/neural_nets/neural_net.html. [Accessed 29 08 2016].

[36] "Python Software Foundation," [Online]. Available: https://www.python.org/. [Accessed 27 08 2016].

[37] "The R Project for Statistical Computing," [Online]. Available: https://www.r-project.org/. [Accessed 27 08 2016].

[38] "Online Gantt chart for project planning," [Online]. Available: https://ganttpro.com/. [Accessed 10 08 2016].