Haaga-Helia
University of Applied Sciences

BIT students' performance analysis with KNIME Analytics Platform

Enxhi Nikolla

Bachelor's Thesis
Degree Programme in Business Information Technology

Haaga-Helia
University of Applied Sciences

Abstract

30.11.2018

Author(s)
Enxhi Nikolla

Degree programme
Business Information Technology

Report/thesis title

BIT students' performance analysis with KNIME Analytics Platform

Number of pages and appendix pages
48 + 1

The ability to analyse students' behaviours and make discussions, recommendation or future predictions is the core idea of Educational Data Mining. Universities are more and more interested to implement this data mining technique for better students' performance and good school reputation.

In this thesis, we are analysing students' behaviours of Business Information Technology degree program in Haaga-Helia University of Applied Sciences to assess current trends for future recommendations and improvements.

There is provided valuable information for understanding the environment we are working on such as statistical analysis, machine learning, data mining and educational data mining.

Two iterations will be implemented for reaching the desired level of results in this analysis. The first iteration will provide a brief description on the KNIME Analytics Platform project development with the data collected from the questionnaire. On the second iteration, the results provided by the first iteration will be interpreted into valuable conclusions, observations as well as suggestions on student's performance. Interpretation is done based the Niemivirta study of 8 scale factors of student performance.

This thesis is target to Haaga-Helia University of Applied Sciences' pedagogical and administrative staff, academic advisors and students. However, psychologists, data analysers and anyone else who is interested in Educational Data Mining and students' behaviours can find this thesis useful.

To summarize the value of this thesis is to give a big picture of the current level of student performance for the given dataset and analyse how this performance is affected from each scale factor mentioned in the Niemivirta study.

Keywords
Machine Learning, Data Mining, KNIME, Education, EDM

Table of Contents

Table of figures

Acknowledgements

I would like to start my thesis expressing my appreciation for all the support and time all the lecturers have put during these years I have been studying in Haaga-Helia University of Applied Science, in Helsinki.

First, I am very grateful to Dr. Amir Dirin that gave me the opportunity to be part of this thesis project. He also has been a great thesis supervisor, arranging weekly meetings with me for checking the progress. I am honoured to be his student and start my path in IT with Dr. Dirin's courses, going more professional with User Experience, Software Engineering etc.

Pr. Dr Dominique Genoud for his amazing lectures in Data Mining. He has a fundamental role to make me study more in Data Mining and Machine Learning. His knowledge in this field is just inspirational and adorable.

Gjergji Make, student at Haaga-Helia UAS for cooperating with me during the thesis and developing the KNIME project. We have had a lot of project together and a lot of learnings supporting each-other.

Mr. Kari Silpiö, senior lecturer and my personal advisor during my studies in Haaga-Helia UAS has had a great role in my study growth. Besides, he is an expert in database teaching.

Mr. Juhani Välimäki, senior lecturer and the best programming teacher in Haaga-Helia UAS. He has always been there to help me with every question I have had. Moreover, I got a clear understanding of SCRUM methodology in just few lessons by Mr. Välimäki.

After all, I would like to thank every lecturer in Haaga-Helia University of Applied science for all the support and contribution they have given to me for my professional growth. Appreciation goes also for the whole personnel of the university.

Abbreviations

KNIME - Konstanz Information Miner

Weka - Waikato Environment for Knowledge Analysis

GUI - Graphical User Interface

EDM - Educational Data Mining

ML - Machine Learning

CRISP/DM - Cross-industry standard process for data mining

UAS – University of Applied Sciences

BIT – Business Information Technology

Td-idf – Term’s Frequency and Its inverse Document Frequency

1 Introduction

Data analysis is having a huge impact in today's world. Through the years, we can notice a huge development of data analysis especially from the evolution of computers. Nowadays, we see an increasingly large importance of data analysis in all kind and size of businesses with a lot of fancy and sophisticated tools and techniques available.

This thesis is about implementation of KNIME Analytics Platform in Haaga-Helia University of Applied Science to increase BITe student's performance. By creating a KNIME workflow, a detailed statistical analysis is available for future interpretations and predictions.

KNIME Analytics Platform provides vast amount of analytics possibilities as well as prediction capabilities. For better usage of these capabilities we need to keep in mind that the dataset provided is accurate enough and has the required information. It is worth mentioning that the questionnaire conducted to Haaga-Helia UAS BITe students was based on Niemivirta study of 8 scale factors of student performance.

The questionnaire was answered by 2018 students of BITe program. The structure was based on numeric values for better analytics possibilities. There were about 100 students who were able to answer. Later in the Implementation chapter we will explain deeper the construction of the data collected from the questionnaire.

While doing this analysis, machine learning, data mining etc will be explained in detail, as well as the technical information about KNIME Analytics Platform. For better results, we must be aware for the current situation of the industry.

Data interpretation will play a big role in this thesis. Conclusions are generated based on interpreting the charts and data provided by the platform. We will focus on comparing main factors which influences the student performance the most such as fear of failure, academic withdrawal or avoidance orientation factors. This will help in making conclusions on what factors should be focused on to raise the performance

Finally, the reader will get a deep knowledge of the industry as well as how the overall BITe students' performance looks like in Haaga-Helia UAS.

2 Research Question

The main research question of this thesis is determining how we could assess the future trends on BITe students' performance through KNIME Analytics Platform.

2.1 Sub-Questions

Students' performance is a psychology related topic that depends on internal, external and natural factors. It explains how students progress or regress in an academic year and aims to increase the performance through methodologies.

Educational Data Mining has significantly bloomed after the birth of the new technologies which can control education related data. Several machine learning algorithms and data mining and statistical techniques are used during educational data mining researches. These researches aim to understand students' way of learning in order to predict students' academic performance and make improvements in the way students learn. (Roy & Garg, 2017)

Pointing to technology and considering the student as the main asset of the university, we are now utilizing KNIME Analytics Platform to analyse students' performance at Haaga-Helia University of Applied Sciences in BITe degree program based on avoidance orientation, fear of failure, school value and academic withdrawal.

3 Methods

In this chapter, we are going to get an understanding of the methodology followed for writing this thesis, mixed methods and workflow processes.

3.1 Methodology

To fulfil the thesis and to achieve the best results wanted, we will divide the project in two iterations. During the first iteration, we are going to develop the project in KNIME Analytics Platform for proof of concept.

The second iteration is focused in interpretation of KNIME main workflow and discussion of results.

3.2 Mixed Methods:

The research methodology that involves collection, analysing and integration quantitative, through surveys, experiments, and qualitative, through interviews, focused groups etc, are known as mixed methods research. The mixed methods research is used when the integration provides more effective understanding of the research problem than each of the methods alone can provide. (Johnson, Onwuegbuzie, & Turner, 2007)

In this thesis we have collected data statistically, but we will have qualitative analysis. The statistical data or quantitative data is collected through surveys to BITe students early in 2018. This includes all the close-ended information such as behaviours, students' performance, fear of failure etc. The qualitative data is aggregated into categories of data and provides useful results.

3.3 Workflow Processes

It has a high importance the understanding of the process from the beginning until the deployment of it. In this thesis, we are going to use CRISP/DM method because it provides a structured approach for planning a data mining project.

CRISP/DM is a data mining method that provides an overview of the data mining project's life cycle. It contains all the phases of a data mining project, with their particular tasks. In description level we can uncover some of the relationships between the tasks, while in all

the other levels, all the relationships between the tasks are clear and intelligible. The relationships depends on the goals, interests of the user, background as well as the data provided, so they can be between any data mining tasks. (Ncr et al., 2000)

The phases of CRISP/DM method are as shown in the figure (figure 1).

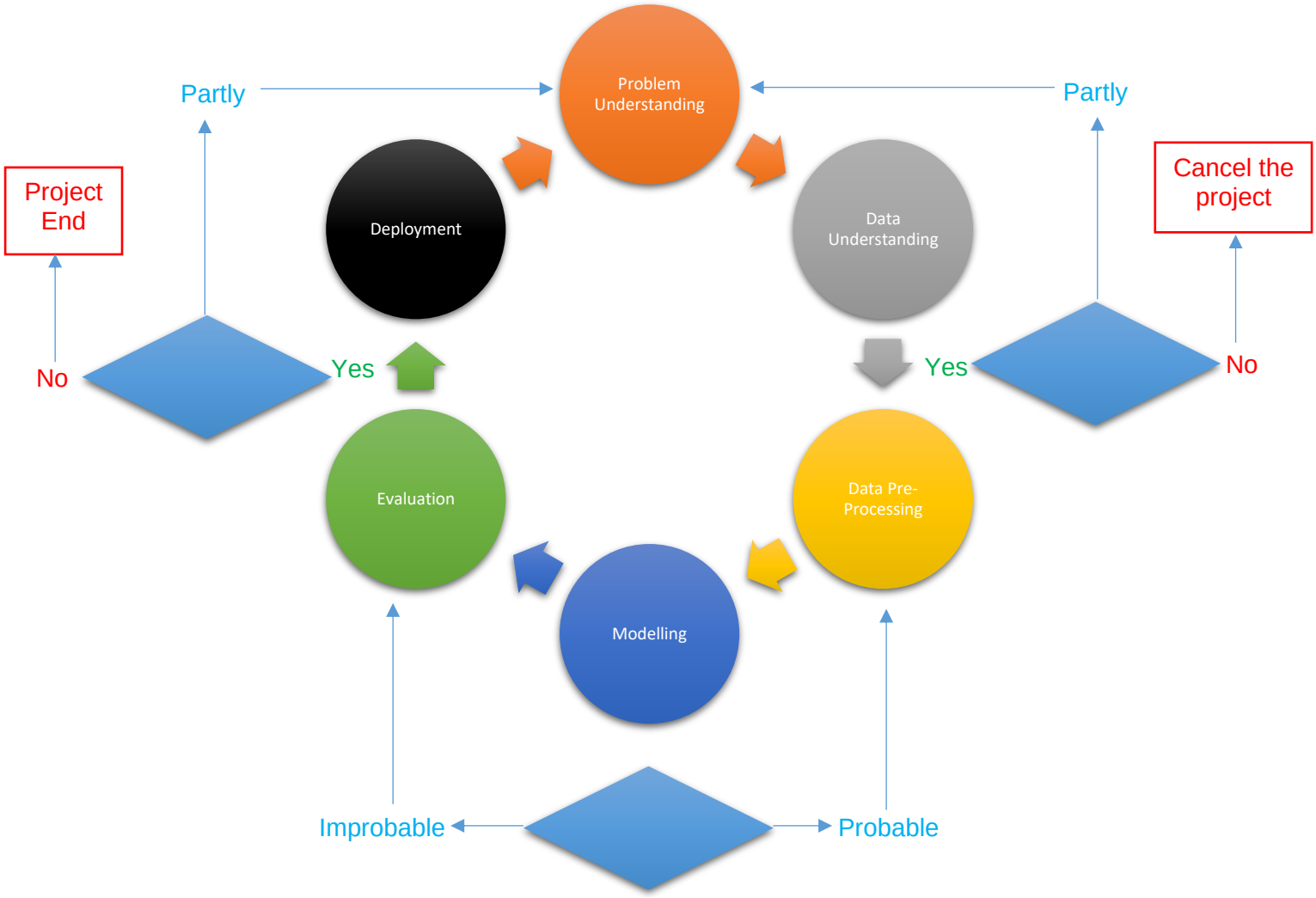

Figure 1 CRISP/DM method: Intelligent data analysis processing

Problem understanding

During the first phase we need to clarify what exactly is the problem we need to solve, how should the solution look like and what are the knowns of the problem.

The focus of this phase is the understanding of project objectives and requirements from a business prospective. Then is important to convert all the knowledge into a data mining problem so we can design a plan to successfully achieve this phase. (Chapman et al., 2000)

Data understanding

It takes into consideration what available data we have, which is relevant, validity, quality and sufficiency of the data related to the problem.

But how do we go through data understanding? First step we have to do is data collection. Then we have to proceed with different activities for getting familiar with data provided. Some of this activities are: identifying data quality problems, discovering first insights into the data and detecting interesting subsets for making hypothesis regarding hidden information. (Chapman et al., 2000)

Data pre-processing

We must remove the unnecessary data such as empty rows and rows with no meaning, transform the data in the best way for modelling and increase data quality.

Data pre-processing phase is based on all the required activities that help to convert all the initial raw data into a constructed final dataset. There are few tasks performed during data pre-processing, such as table, record, attribute selection, as well as data transformation and cleaning for modelling tools. All these tasks can be performed several times without following any order. (Chapman et al., 2000)

Modelling

We need to clarify the best model architecture that suits the problem, find out the techniques for getting the desired model and best prediction accuracy.

Modelling phase includes a lot of selected and applied techniques. Their parameters are calibrated to optimal value. For resolving the same data mining problem, we can use different modelling techniques. Some of the techniques have specific requirements on the form of data, so it is necessary to go back in the data pre-processing phase. (Chapman et al., 2000)

Evaluation

It is the phase we evaluate how good does the model meets the project's requirements and what have been our learnings. In case the goal is not met, we must end the project.

During modelling phase, we have already built the model that seems to be of a high quality from a data analysis perspective. Still, the model is not ready yet for deployment. That is why we have evaluation phase in the middle. It is important to do a deep evaluation of the model, review all the steps executed for creating it and make sure that the model reaches the business objectives exactly as require. Also, we should ascertain that we have adequately contemplated all the significant business issues that might raise. At the end, we need to take a decision on the use of the data mining results that should be reached. (Chapman et al., 2000)

Deployment

We need to understand how to deploy the project and the validity of the model for real cases.

After we have created the model and evaluated it, we have the deployment phase. But what is the deployment phase and what do we have to do? We already have the model that has increased the knowledge of the data. This knowledge is easily understood by developers, but for it is highly important that the customer can understand and use it. The most common way to visualize the data is on a real-time web page designed according to customer's needs.

On top of that, we can represent the knowledge of the data in a simple report. If required by the customer, we can also implement a repeatable data mining process across the enterprise, which is a pretty complex implementation. The responsible person for deployment is, in most of the cases, the customer. When the data analyst takes care of deployment, he needs to make sure that the customer understand all the actions needed to be carried out so he can profit from the models created. (Chapman et al., 2000)

4 Background Studies

Data Mining, Machine Learning, data science, etc are common words we see quite often on news and they are considered “the hot topic” of technology. But what do all these words mean and are they related with each other?

They are all related and subfields of one-another as shown in figure (figure 2). Deep Learning is a subfield of Machine Learning, which itself is a subfield of Artificial Intelligence. Computer science covers all these fields. A new term is Data Science that includes ML and some computer science.

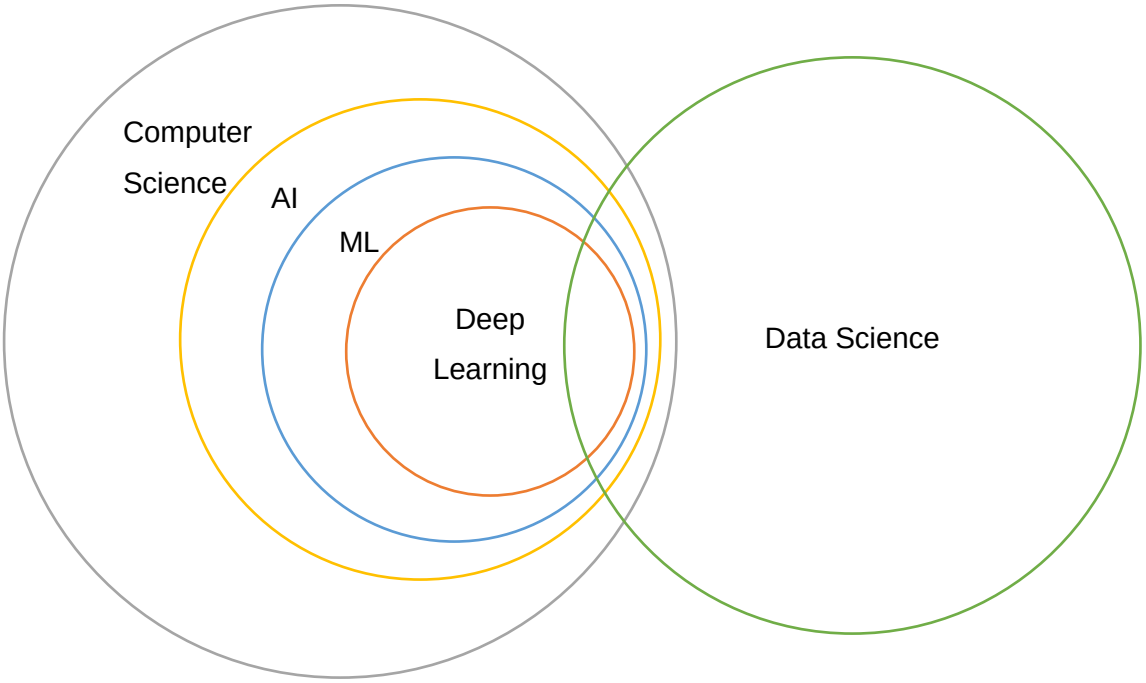

Figure 2 Related fields

A better understanding and explanation of the main terms we need in this thesis is given below.

4.1 Statistical Analysis

Statistical analysis is a collection, examination, summarization, manipulation and interpretation of quantitative data. It is used for uncovering trends, patterns, relationships and causes that are inherent in the first sight. (WebFinance Inc., 2018)

Statistics are applied every day from the basic research, to companies and government. From a vast amount of data, we try to manipulate it, so the data will have meaning. We

can summarize the data, calculate mean value and spread of data, make future predictions and test hypothesis.

Statistical Analysis is a branch of mathematics dealing with data. It operates in two main methods for data analysis: descriptive analysis and inferential statistics.

Descriptive analysis summarize data from a sample using parameters such as mean, mode, standard deviation and median. It allows us to describe the data and it uses all the data we are interested in or population to give us results.

Inferential statistics use sample data or only the part of the population we are interested to investigate to draw conclusions and determine if the data can give us relevant prediction. Population is the whole data we are interested in and sample is just the part of data we need.

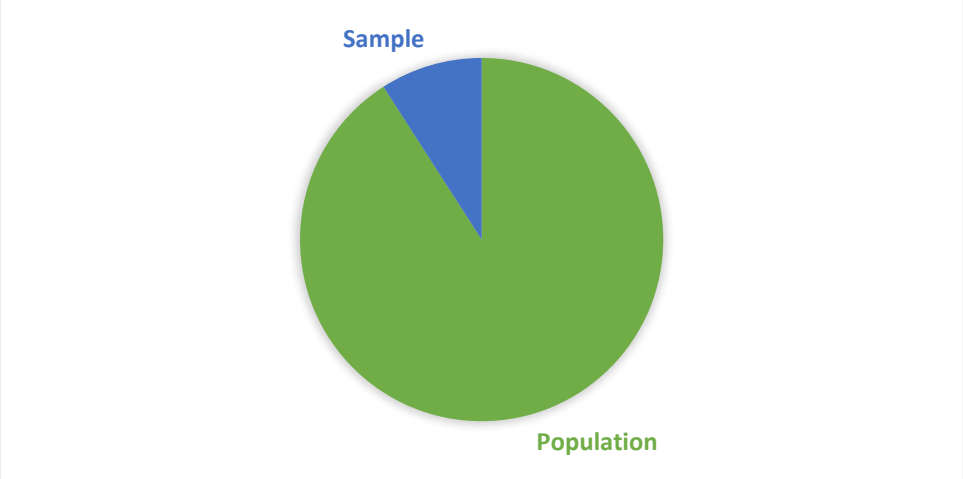

Figure 3 Population vs. Sample

Statistical analysis methods

Once we have collected the data, we need to give meaning to them. In statistical data analysis there are few methods that help us interpret the data collected.

Mean

Mean is widely known as the average. But why we use mean and how do we calculate it? Mean is used to determine the overall trend, so we can have a snapshot of the whole data provided. Calculating mean value is pretty easy and quick. We have to divide the sum of the whole list of numbers with the number of the items in the list. (Pong, 2017)

Formula to calculate mean value is: $\bar{x} = (\Sigma xi) / n$, where $\bar{x}$ is the mean, xi are all the x values and n is the number of participants on the sample.

Standard deviation

Standard deviation evaluates the spread of data around the mean value. Sigma, the Greek letter, stands for standard deviation. There are two types of standard deviation: low and high. A low standard deviation means that more data is lined up with the mean value. A high standard deviation foretokens that data is spread more widely from the mean value. Besides, standard deviation is used for determining dispersion of data pints in some data analysis methods. (Pong, 2017)

Formula to calculate standard deviation is: $\sigma = \sqrt{(\sum | x - \bar{x} |^2 / N)}$, *where σ stands for standard deviation, x stands for a value in the data set,* $\bar{x}$ for mean value and n for number of data in population.

Regression

Regression is a statistical process that shows the relationship between dependent variables and explanatory variables or independent ones which are used for prediction. There are quite many techniques for making the models and analysis between variables and the most common one is scatterplot. There we can check if we have strong relations between the variables

Formula to calculate regression is: $Y = a + bX$, where Y stands for regression, a for the intercept, b for slope of the line and X is explanatory variable.

Sample Size Determination

Sample size determination is used in all the cases we want to make some measures in a population. For example, if we need to get some predictions about the elections in Helsinki, we do not need to ask every inhabitant about their opinion, but we make assumptions and predictions based on the answers of a group of people. To make right predictions, we need to define the right amount of people involved in the sample, and the correct size of the sample is calculated by using standard deviation methods and proportion.

Hypothesis Testing

Hypothesis testing is all about making assumptions in a defined parameter of a population and it finds a widely usage in researches, science, businesses etc. Experimental data is used for making statistical decisions. Hypothesis testing is considered to be relevant when the results we get are not happening by chance.

Statistical analysis techniques

There are few techniques that help us to make statistical analysis, such as: summarizing data, measuring the location of data, measuring the spread of the data and skew. A description of each technique is given below.

Summarizing Data: Grouping and Visualizing

Data summarizing is a combination of grouping and visualization. First thing to start with is grouping the raw data into categories and then visualize it. For example, taking this thesis into consideration, we are interested into differences by gender with the eight Niemivirta categories (appendixes 1). This way we have two groups: male and female for making comparison. To visualize the results we get, we use can bar charts, graphs etc.

Measures of Location: Averages

Average is a calculated central value of a group of numbers. It gives us information of the size of the effect we are trying to test. In other words, it is the division of the sum of all numbers we have with the amount of the numbers in the list. There are three measure of average: mean, mode and median. Habitually, when people want to express something about average, they refer to mean value. Taking in consideration all these three types of average, they all differ from one another, so we cannot consider all them the same. It is very important to clear out for which one we are referring. (Skillsyouneed, 2017)

Average uses all the acquired data values for statistical analysis but can be skewed by extreme values. To avoid this situation, researchers utilize median instead. Median is the mid-point of all the data and it does not get skewed by outliers. Regardless, median is not very accurate for further statistical analysis. Mode present the most common value in a data set, but it is not unerring enough for statistical analysis. (Skillsyouneed, 2017)

Measures of Spread: Range, Variance and Standard Deviation

Researchers often want to look at the spread of the data, that is, how widely the data are spread across the whole possible measurement scale.

To measure the spread of the data we can use range, variance and standard deviation. A brief description of these three terms is given below.

Range is the difference between the largest values with the smallest ones. Ofttimes we can hear the interquartile range used by researchers. But what is it? Interquartile range is

the range of the middle half of the data, from 25% of the lower quartile to the upper quartile up to 75%, of the values. To find the value of the quartiles, we follow the same process as we do with median. Median is all the time 50% of the values. Upper quartile is 75% of the values and lower quartile is 25% of the values. (Skillsyouneed, 2017)

For a easier understanding of the interquartile range, we can use figure (figure 4).

Figure 4 Interquartile range

Standard deviation measures the average spread of the data. A more detailed description can be found in chapter 4.1.1.2

The variance is the square of the standard deviation. They are calculated by:

1. Calculating the difference of each value from the mean
2. Squaring each one (to eliminate any difference between those above and below the mean)
3. Summing the squared differences
4. Dividing by the number of items minus one.

This gives the variance.

Standard deviation is the square root of the variance.

Skew

Skew is used to evaluate how symmetric is the data set visualized into a graph, which means it demonstrates if we have more high or low values. The skew can be positive, negative or we have no skew at all. When we have high values, then the skew is positive and when we have low values, the skew is negative. There is no skew when the distribution between high and low values is the same. A high skew indicates a low mean, mode and median.

Let's understand the negative, neutral and positive skew.
In the figure (figure 5), we have a positive skew. In a positive skew, the tallest line is always on the left side of the graph, as well as the mean value is on the left.

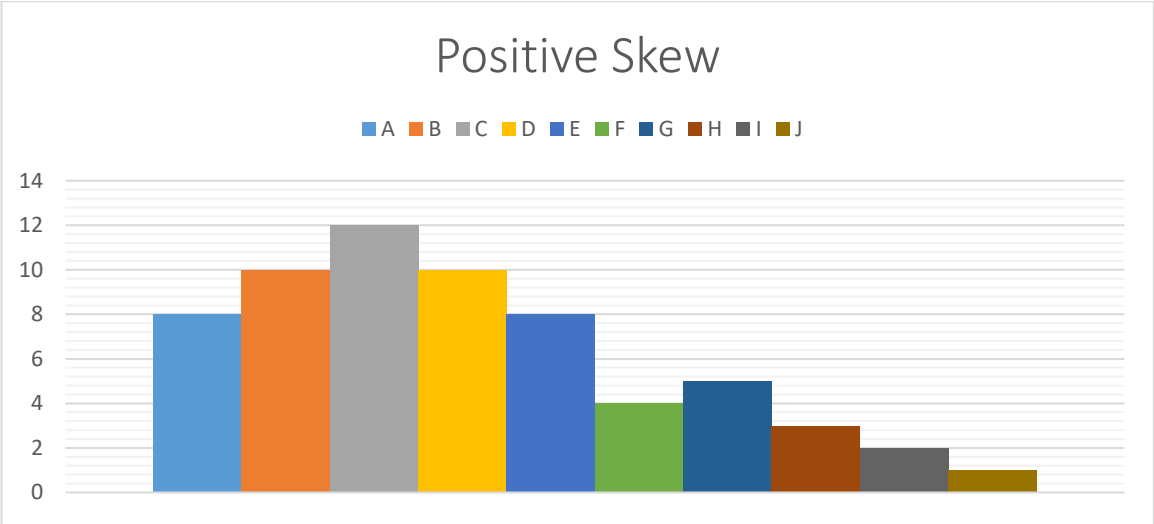

Figure 5 Positive Skew

There is no skew when there is a normal distribution which creates a symmetrical and identical graph in both sides. The mean, mode and median are exactly at the peak of the graph, as shown at the figure (figure 6).

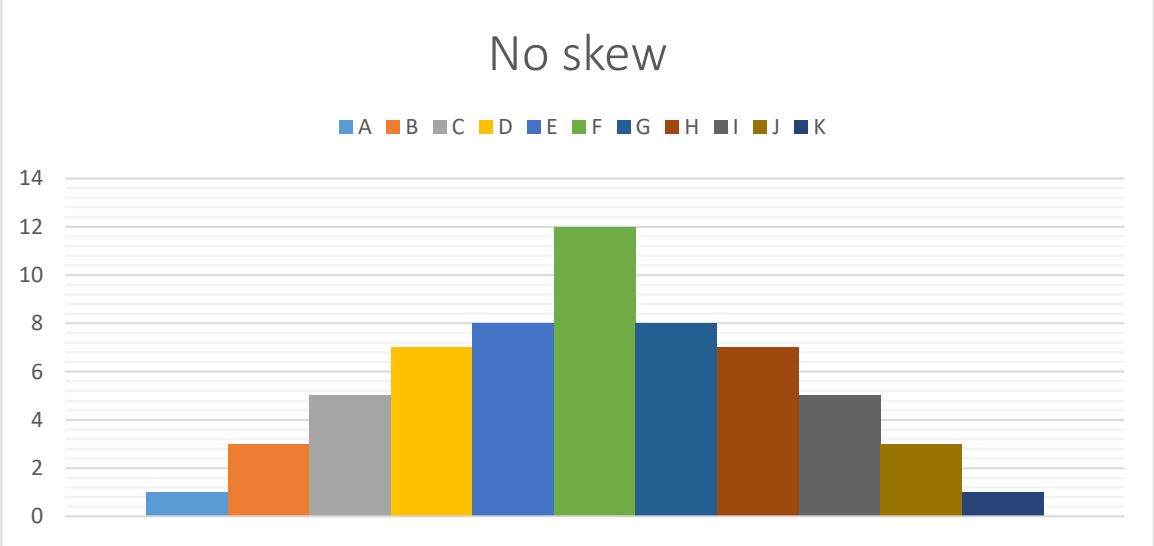

Figure 6 Normal distribution

There is a negative skew, when the tallest line of the graph is on the right side it as shown in the figure (figure 7). Also, the mean value can be found on the right side of the graph.

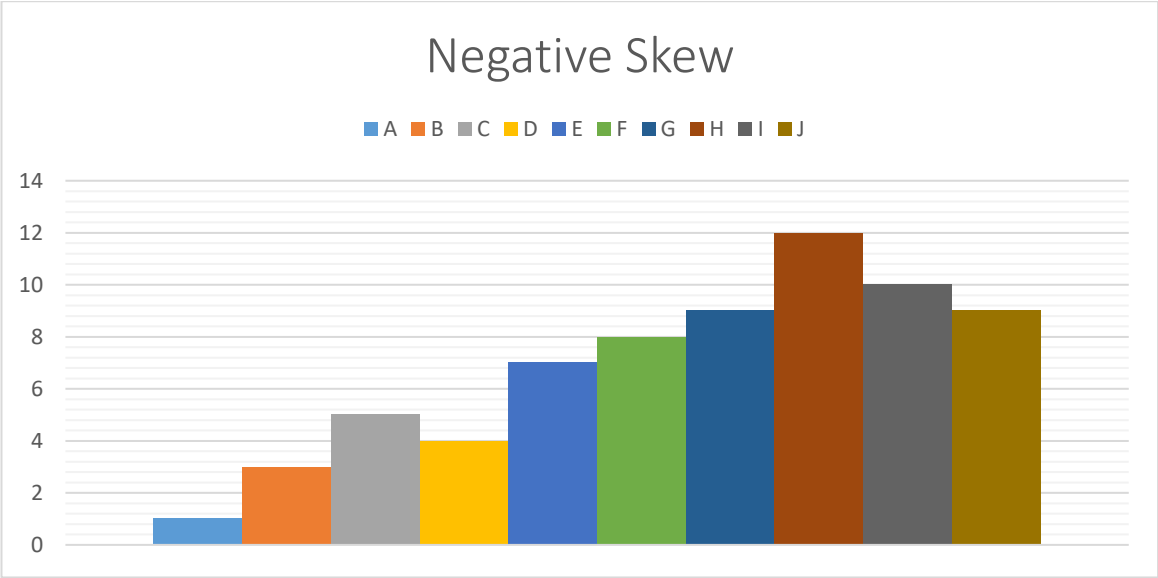

Figure 7 Negative skew

4.2 Data Mining

Recently we can notice a tremendous number of digital data due to the advancing technologies. In this case, it was required a scientific way to extract the useful information from the huge data repositories. There are few explanations for data mining. Below you will find two definitions from Margaret Rouse and SAS Institute.

In enterprises, we need to solve problems through data analysis and we have in front of us a large data set where we need to sort the data for recognizing patterns and setting up relationships. All this process is done through data mining. Tools used in data mining give the opportunity to enterprises to make future trends predictions. (Margaret Rouse, 2017)

In a large data set where we need to predict outcomes, we use data mining process. That allows us to find anomalies, patterns and correlations. The wide variety of different techniques, allows us to use the information for increasing revenues, cutting costs, reducing risks, improving customer relationships etc. (SAS Institute, 2018)

Data mining in years and now

“Knowledge discovery in database” is the term used decades ago for referring data mining. This new term, “data mining” came in use in 1960s and now is a key word in technology. Data mining is the process of uncovering unknown relations in the data and making future trends predictions. It is always evolving as the amount of data is growing so fast. (SAS Institute, 2018)

Data mining is an entwined settlement of three scientific disciplines: statistics, artificial intelligence and machine learning. Statistics is used for making numeric study of data relationships, artificial intelligence shows human-like behaviours displayed by the software and machine learning uses algorithms that learn from current data for making future predictions. (SAS Institute, 2018)

In the last decade, especially in the recent years, we can notice a radical enhancement of technology. Many machines and processes have been automated. Within the same time limit, we can be more productive than before. Besides, the work and energy required for the same job is relatively less than it used to be.

Another important thing about this upgraded technology is related with data complexity. A complex data set gives us a wider view of insights. That is why businesses, banks, manufactures, retailers etc use data mining to lighten up relationships from social media, risk, revenue, pricing etc.

Importance of Data Mining

The amount of data is growing rapidly and it is doubling every other year. The greatest part of the digital universe is formed by unstructured data. Be that as it may, more data does not indicate more knowledge. So, why is data mining important?

Data mining allows you to:

1. Sift through all the chaotic and repetitive noise in your data.
2. Understand what information is relevant and later use it for making likely outcomes.
3. Speed up the rhythm of taking decisions

(SAS Institute, 2018)

Data Mining Methods

There is a huge number of Data Mining Methods, but they are categorized in two major ones. We have Descriptive methods and Predictive methods. In this thesis we are going to use descriptive methods.

The Descriptive methods, analyse the data and check for interpretable patterns to describe the data but don’t do any mathematical analysis. The Predictive methods use variable for predicting future outcomes.

How does Data Mining work?

To understand the process, we start from raw data to obtain the knowledge. As shown in figure (figure 8), the steps to be followed are:

Data → Selection → Pre-processing → Transformation → Data Mining → Knowledge

During the Selection phase, we go from a bunch of raw data to the target data, by removing non-useful data or data that don't fit our main format. From the target data we go to pre-processed data by detecting outliers and missing values. After pre-processed data, we go to transformed data by normalizing it or finding correlated variables. The transformed data helps us to create patterns by using several data mining algorithms. Then, the user can create the knowledge by interpreting the patterns.

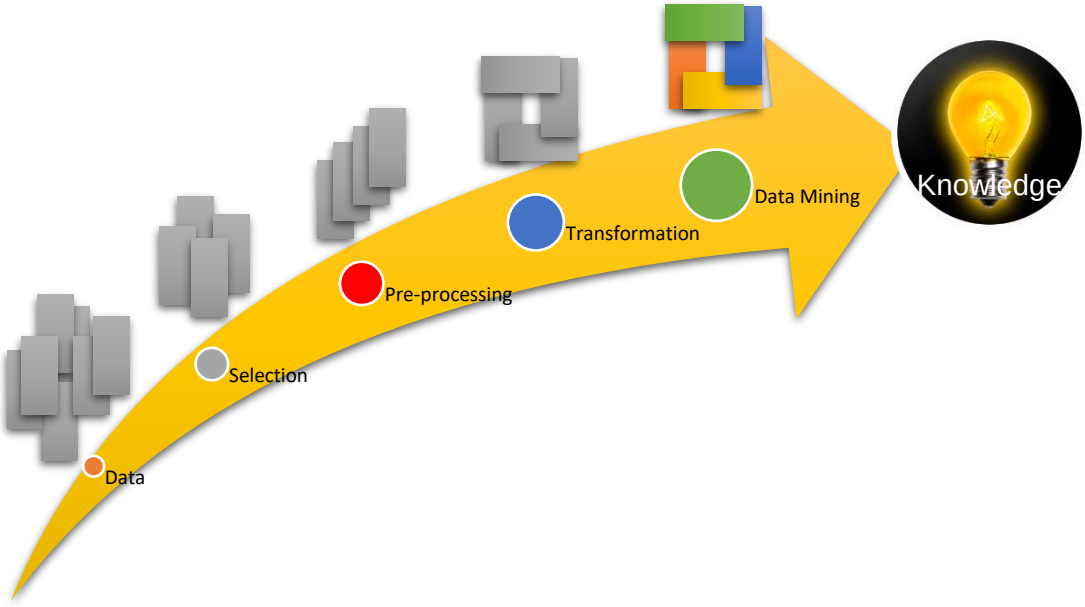

Figure 8 From data to knowledge

4.3 Machine Learning

Machine Learning is a system that is continuously improving and becoming more unerring in predictions with the more data we annex.

It is a very wide topic with lots of discussions and definitions. It is considered as a new topic in technology, but it lies years back. Roots of Machine Learning are in statistics, the art of extracting knowledge from data.

History of Machine Learning

Machine leaning's history foredates the computers, but it is still considered as an excrescence of statistics.

A brief history of machine learning is given below.

1943 - 1959 Neural Network

In 1943 where a human "neural network" is modeled with electrical circuits. This was used by computer scientists in 1950s. This is the point where Machine Learning started to become reality. In 1959, the phone calls became clearer due to implementation of adaptive-filter.(Orcibal & Jerphagnon, 2018)

Predictive analysis:

Anderson is one of the first ones used predictive analysis. He observed three different types of Irises and documented all the measurements. By using those measurements, Anderson developed a simple algorithm that could predict the species of the flower. The dataset for this algorithm was pretty small and could haven been solved effortlessly by hand. But the idea of all this was to use a wide range of methods. Algorithm based on neural networks are different from the methods that are used for discovering potential combinations of rules that execute the classification. (Houle, 2016)

1957 Perceptron:

The most straightforward neural network is Perceptron. Perceptron was a vogue in 1960s. It acquired knowledge to a separating knowledge between two different categories. It was operating as a parallel computer and could perform much better than conventional computers at visual recognition tasks. In 1969, Minsky and Papert published a book that proved the limitations of Perceptron. This is the year, Perceptron vanished hastily. (Houle, 2016)

1971 Full-text search:

In 1971, we have the development of the algorithm for full-text search. This is known as Term Frequency (TF) and Inverse Document Frequency (IDF) or tf-idf algorithm. This algorithm is pretty similar with Perceptron. Tf-idf is a computation of a dot product in a high-dimensional space. The document we are using is considered a list of scores based on how the word is used. Then we can compute the similarity metric between two documents or one document and one query.

How to understand this? We have a document with 1000 words and we have to find Tf and Idf for the term “algorithm” which has been used 50 times. Also, the term “algorithm has been used Z amount of times in a 200 000 000 principal and in 100 000 documents we have involved the term “algorithm”. Tf will show us the frequency of the word “algorithm”. Idf will show how noteworthy the term “algorithm” is in the whole principal.

To calculate the tf we have the formula:

$$TF_{term} = N_{term} / N_{words}$$

In our case we have:

$$TF_{algorithm} = N_{algorithm} / N_{words} = 50 / 1000 = 0.05$$

To calculate idf we have the formula:

$$IDF_{term} = \log (N_{DF} / N_{words})$$

In our case we have:

$$IDF_{term} = \log (200\ 000\ 000 / 100\ 000) = \log (200) = 2.3$$

1992 US post office: Handwritten digits:

Neural networks had continuous progress from the beginning till 1992. It was put a lot of importance to visual processing in the human brain. The researchers tracked down that human brain is built from several layers. These layers encode all the multiplex features of the images. This result had a fundamental role for further development, even thought it took time to come up with the ideal method for training the neural networks. However, when this ideal method was developed, the multiple-layer neural networks had an essential and effective role in visual recognition tasks. In 1992 the US Post Office adapted an automatic ZIP code reading machine which was developed as a result of the multiple-layer neural networks.

Neural networks are trained to memorize different cases and examples and this is the way they operate. Presuming that the examples we use for training the neural networks are just basic ones, this will not be efficient. A neural network with a single hidden layer can learn any function and if we add more layers we can form “information bottlenecks” that help in sorting the information. (Houle, 2016)

Data driven competitions:

To understand the data driven competition we can take as an example the Fisher-Anderson Iris dataset. There are many scientists and programmers who have been working hard to ameliorate the performance of a task averse to standardized data, indeed it can

never be an overvaluation of it. Quite many improvements in image recognition has came as a result of competitions around ImageNet. (Houle, 2016)

2006 Deep learning:

Deep learning is a recent concept in the world of technology, that was initiated by Geoff Hinton and his students. This concept is related with multiple-layer neural networks and the way of training these neural networks. To simplify the process, Geoff and the students decided to divide it in two phases. During the first phase, the network is handled the same way as a Bayesian Belief Network. This network was instructed to uncover statistical regularities in the data. To make the input data more accessible for further usage, they chose not to label it. In this way, if the model was shown a face, it develops a “theory of faces”. This would give model a chance to be more open to important discoveries related with face’s features, and running the network backwards, enables the model to draw faces and face-like-shapes. Indeed, first phase creates a model of inputs that can be used during the second phase. In the second phase all the information is used for training multiple hidden layers. The data is now labeled, that can help for classifying the images based on the features. (Houle, 2016)

You can find more about deep learning in chapter 4.4.

Present: Explosion of Neural Network architectures:

Nowadays, machine learning has been ameliorated quite a lot and neural networks are just a small part of it. Still and all, they are very active. A lot work has been done to organize deep networks for a more proficient recognition of image and speech. One area of neural network is recurrent neural network. This finds usage in text understanding and areas where a stream of symbol should be transformed into an object. It can deal with several inputs happening as a series in time. (Houle, 2016)

What is Machine Learning?

“Machine learning is the science of getting computers to act without being explicitly programmed.” (Pyle & San Jose, 2015)

The basic of machine learning starts with the practice of using algorithms to parse data, learn from that data and later start making predictions about different things in the world. (Copeland, 2016)

On the other side, machine learning is an answer for the question “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” (Mitchell, 2006)

In other words, Machine Learning is the science of teaching computers to act like humans but getting more productive and predictive than us. We offer them data from real world, so the processes of algorithm are improved.

Nowadays, many application use Machine Learning to personalize members’ data for future personalized suggestions. Some of the most well-known applications that use Machine Learning are: Facebook News Feed, Netflix, Snapchat, Google Maps, Tinder etc.

But how does Machine learning works? The processes that need to be followed are shown in the figure (figure 9).

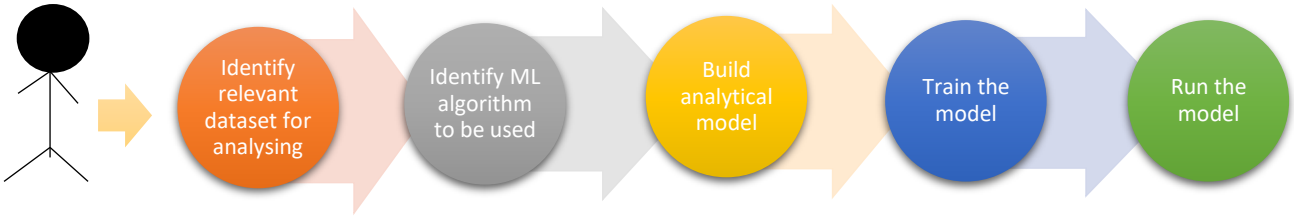

Figure 9 Machine Learning Processes

Types of Machine Learning

Machine Learning functions in three different algorithm types: supervised learning, unsupervised learning and reinforcement learning.

Supervised learning

The supervised learning consists in a given input predicting the correct output. The training of this algorithm continues until we get the desired result.

In simpler words, if we have a picture with a number on it, the machine learning must predict what is the right number on it. Human does not write down rules for classification, but it trains the learner to recognize the correct output automatically, by providing correct answer.

Decision tree, regression etc are part of supervised learning.

How it works?

We have an outcome variable, or differently known as the dependent variable, that is used to be predicted by a given set of predictors or independent variable. These creates the algorithm. The set of these variables is used for generating a specific function for mapping inputs with the desired outputs. This is the training process and it continues until the model achieves the desired level of accuracy on the training data. (Sunil, 2016)

Unsupervised learning

In unsupervised learning humans do not provide the correct answer, so the process is a bit more complicated. The learner tries to get some data from visualization, clustering or generative modelling.

How it works?

During unsupervised learning there is not provided an outcome variable to predict. This prediction is used for clustering population in different groups. That is why it finds a wide usage in segmentation of customers in different groups for specific intervention. (Sunil, 2016)

Reinforcement learning

Reinforcement learning is used in more sensitive cases, such as self-driving cars. Continues feedback is given to the learner, so the result is more accurate.

How it works

In reinforcement learning algorithm, the machine is trained to make distinct decisions. The machine learns continuously from the previous experience. First, we expose the machine to an environment. Then, the machine should train itself using trials and errors. Replication of the process, gives the machine the possibility to improve ceaselessly until it captures the best knowledge for making business decisions. (Sunil, 2016)

Automatic Tools for Machine Learning and Data Mining

A variety of tools are used in analysis statistical data and data mining. Data mining tools are divided in three categories: free software, free software with paid service and enterprise software. A list of most used software is presented in the figure (figure 10).

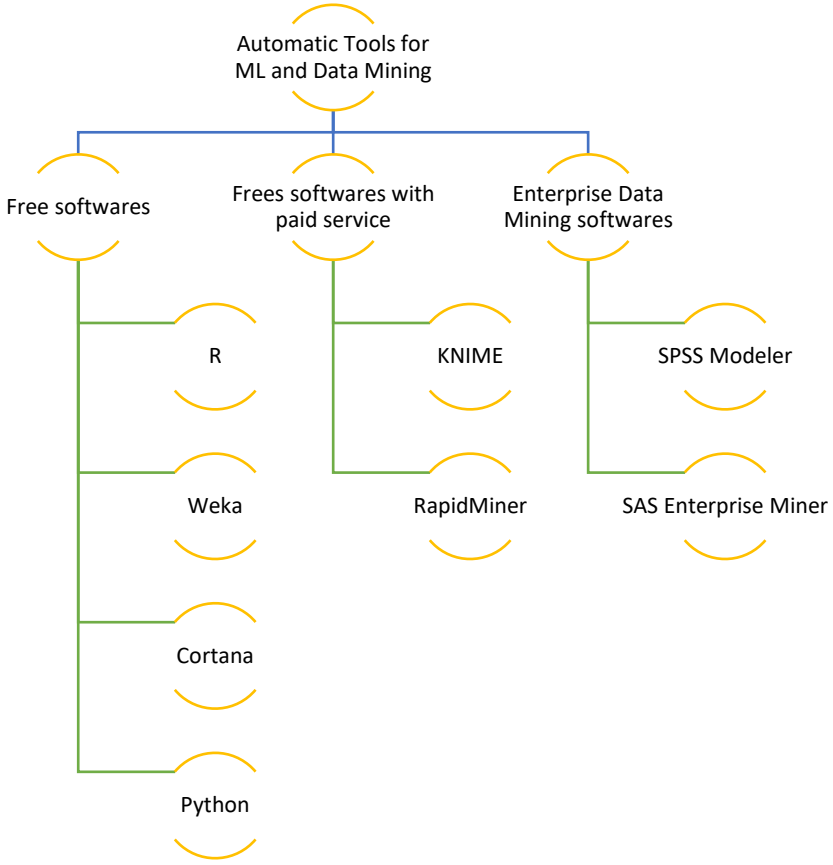

Figure 10 Automatic Tools for Machine Learning & Data Mining

Machine Learning and Data Mining main tools in years

1993 - WEKA:

Weka is a data mining and machine learning toolkit with an extensive usage nowadays. It was initially developed at the University of Waikato in New Zealand and now it is quite well-known to professors, universities and industrial and academic researchers. Weka uses Java for writing data mining algorithms and it has a huge collection of state-of-the-art machine learning algorithms. Moreover, regression, clustering, classification, visualization, association rules and data pre-processing are supported by Weka. That is why Weka is a very powerful tool. (Naik & Samant, 2016)

Weka is an open source Machine Learning software that does not require any programming skills since it has a GUI. There are three ways Weka operates:

1. Explorer – allows people to play with data and transform it according to desired algorithms
2. Experimenter – allows people to analyse the results of the algorithm chosen
3. Knowledge Flow – allows people to create the graphic design of the process and run it

1997 - Orange*:*

Orange is another open source data mining and machine learning platform, developed and maintained by the Bioinformatics Laboratory of the Faculty of Computer and Information Science at University of Ljubljana. It is written in Python and it provides visual programming front-end for visualization and investigative data analysis. (Naik & Samant, 2016)

Orange has a canvas interface with widgets that vary from the simplest data visualization to predictive modelling. Workflows can be created through user-designed widgets or Python libraries.

2001 - RapidMiner: is a user interactive environment for machine learning and data mining processes. It is opensource, free project implemented in Java. It represents a modular approach to design even very complex problems - a modular operator concept which allows the design of complex nested operator chains for a huge number of learning problems.

2003 - Tanagra:

Tangara is a free suite of machine learning software. It was developed at the Lumiere University Lyon 2, in France by Ricco Rakotomalala. It gets usage in researches and academic programs. Similar to Weka, quite many data mining tasks are supported by Tangara such as classification, clustering, regression, visualization, factor analysis, instance selection, feature selection, feature construction, descriptive statistics and association rule learning. (Naik & Samant, 2016)

Tanagra is a simplified data mining tool that works with diagrams and nodes. The models are shown in a tree diagram and the results are displayed in HTML format.

2004 - KNIME:

KNIME is an open source data mining and visualization platform. It is used worldwide by more than 3000 organizations. It was developed at the University of Konstanz. It is based on Eclipse IDE platform, making it a quite powerful development and data mining platform. KNIME desktop in the entry version of KNIME. (Naik & Samant, 2016)

KNIME is a powerful and simple platform that allows everybody to create machine learning solutions on a GUI based workflow. No coding skills are required. This thesis is based on KNIME Analytics Platform, so we will get to know a lot about it.

KNIME Analytics Platform

According to chapter 4.5, KNIME is a powerful platform for Data Mining and Machine Learning solutions. In this section, we will have a step-by-step introduction to KNIME Analytics Platform, so at the end the user will know how to build a project.

Installation

First, we download KNIME from the official web page: http://knime.com
When the installation is ready, we set up our workspace, as shown in the figure (figure 11).

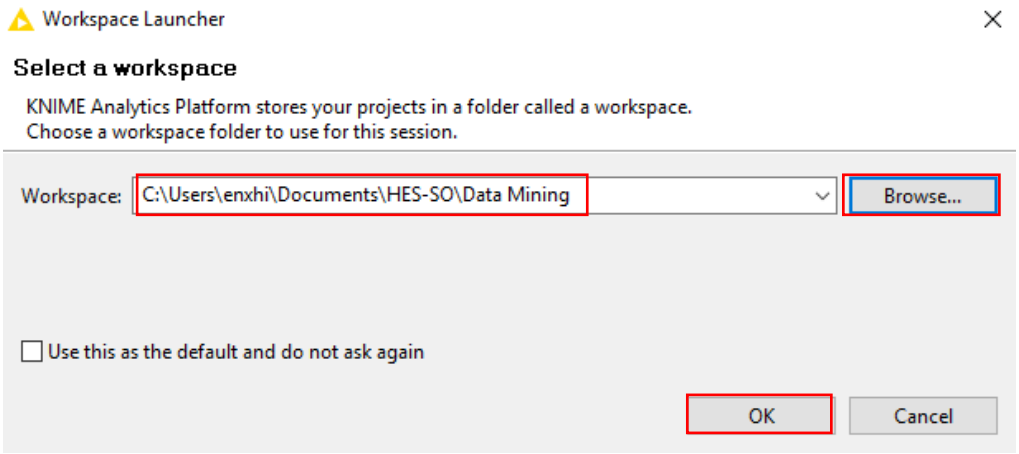

Figure 11 KNIME Workspace Set Up

First Introduction with KNIME

After the installation, we are in contact with the view of the platform as shown in figure (figure 12).

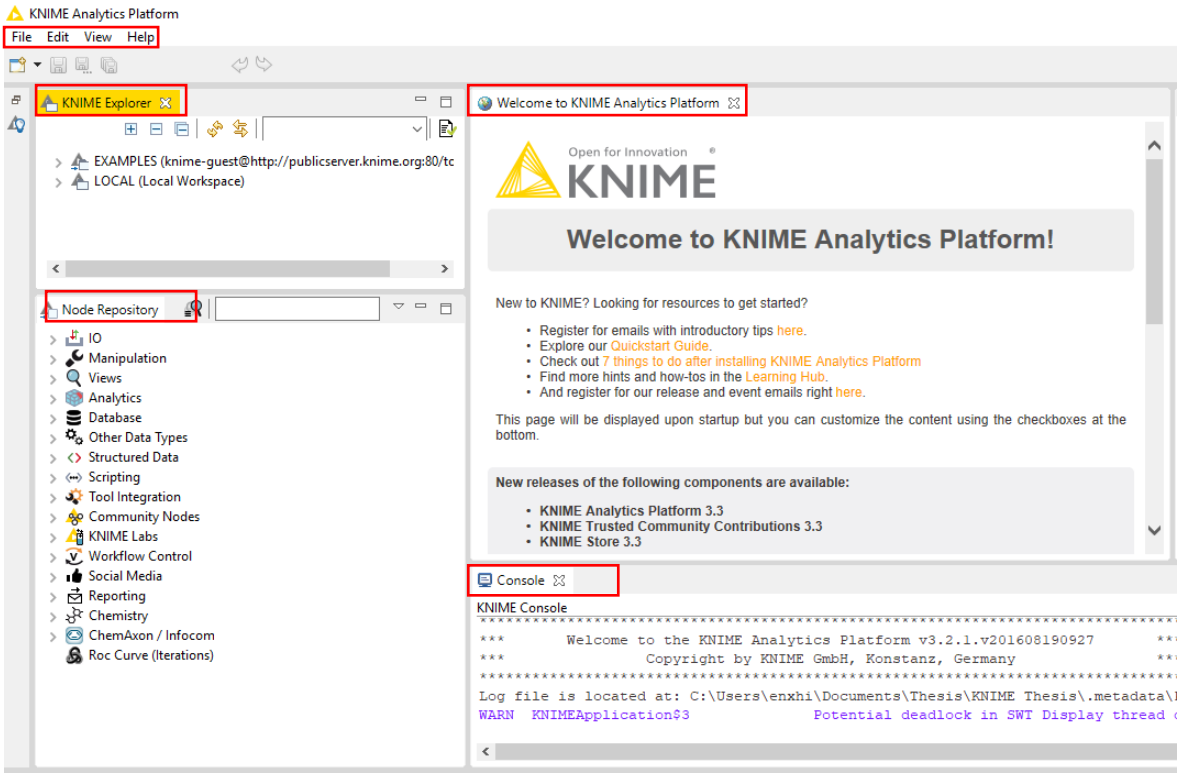

Figure 12 KNIME Introduction

We have the menu bar on top of the page: File, Edit, View, Help. We can update the platform under File -> Update KNIME and then Restart.

In KNIME Explorer, it is the workspace repository that can be in server or local.

In KNIME Repository, there are located all the nodes available in KNIME platform, so it is a node repository.

Welcome to KNIME Analytic Platform is the workflow development space where we can create our project.

KNIME Console is an error and information console.

Introduction to software

KNIME is a very simple platform that works on a GUI based workflow, functioning with nodes. Nodes are divided into colors, where each color represents a function. Orange color is used to nodes that read information and manipulate it. Writing nodes are in red, visualizing nodes are blue, mining nodes are green etc.

For adding or replacing nodes, adding files, we can just drag and drop. Another way to add nodes, is by double-clicking on their name in the Node Repository. To check the node's configuration, we need to double-click on it.

In case of confusion about nodes' function, we can click on the node and check node description as shown in figure (figure 13).

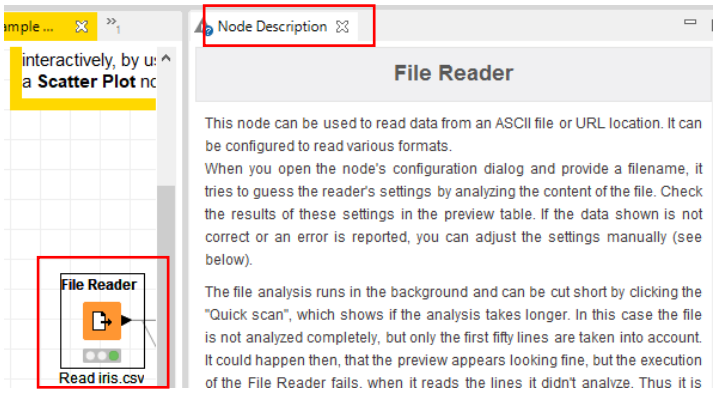

Figure 13 Node description

In workflow, there are six main functions as represented in the figure (figure 14).

1. Execute selected nodes
2. Execute all
3. Execute and open first view
4. Reset selected nodes
5. View results of selected nodes
6. View steps

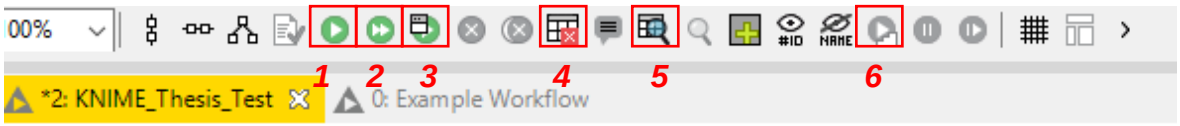

Figure 14 KNIME workflow functions

A demo of a small project is presented in the figure (figure 15). It shows 4 stages that we need to follow: Read, Transform, Analyze and Deploy.

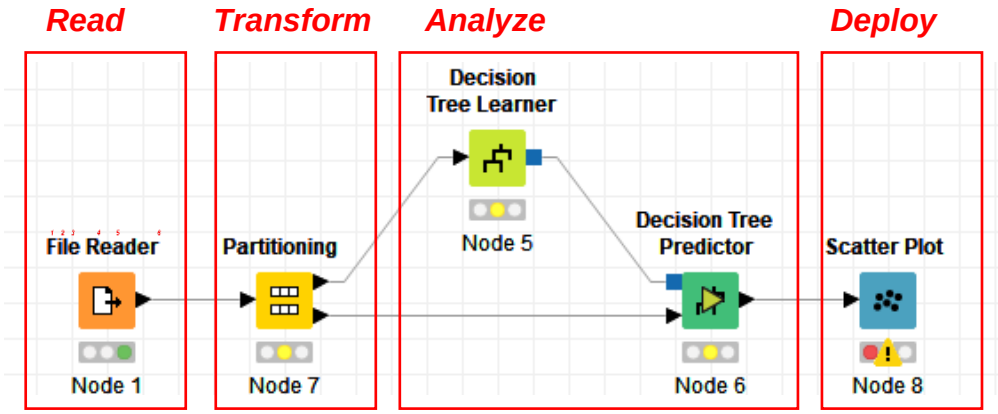

Figure 15 KNIME project demo

As shown in the figure (figure 14), we have used file reader that reads excel, csv file etc., partitioning, decision tree learner, decision tree predictor and scatter plot.

File reader: “This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats.” (KNIME.COM AG, 2016)

Partitioning: “The input table is split into two partitions (i.e. row-wise), e.g. train and test data. The two partitions are available at the two output ports.” (KNIME.COM AG, 2016)

Decision tree learner: “This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm provides two quality measures for split calculation; the gain index and the gain ratio.” (KNIME.COM AG, 2016)

Decision tree predictor: “This node uses an existing decision tree (passed in through the model port) to predict the class value for new patterns.” (KNIME.COM AG, 2016)

Scatter plot: “Creates a scatterplot of two selectable attributes. Then each datapoint is displayed as a dot at its corresponding place, dependent on it's values of the selected attributes. The dots are displayed in the color defined by the Color Manager, the size defined by the Size Manager, and the shape defined by the Shape Manager.” (KNIME.COM AG, 2016)

4.4 Deep Learning

“The field of artificial intelligence is essentially when machines can do tasks that typically require human intelligence. It encompasses machine learning, where machines can learn by experience and acquire skills without human involvement. Deep learning is a subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data. Similarly, to how we learn from experience, the deep learning algorithm would perform a task repeatedly, each time tweaking it a little to improve the outcome. We refer to ‘deep learning’ because the neural networks have various (deep) layers that enable learning. Just about any problem that requires “thought” to figure out is a problem deep learning can learn to solve. Deep learning allows machines to solve complex problems even when using a data set that is very diverse, unstructured and interconnected. The more deep learning algorithms learn, the better they perform.” (Marr, 2018)

Educational Data Mining

Applying Data Mining in educational context is an emerging topic and it is known as EDM (Educational Data Mining). EDM is the science of applying Data Mining and Machine Learning in educational institutes aiming to increase study methods for improving students’ performance by predicting future learning behaviour.

“Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.”(Team, 2013)

“Educational data mining is emerging as a research area with a suite of computational and psychological methods and research approaches for understanding how students learn. New computer-supported interactive learning methods and tools—intelligent tutoring systems, simulations, games—have opened up opportunities to collect and analyse student data, to discover patterns and trends in those data, and to make new discoveries and test hypotheses about how students learn. Data collected from online learning systems can be aggregated over large numbers of students and can contain many variables that data mining algorithms can explore for model building.” (Team, 2013)

In other words, EDM is collecting data from students and analyse those to check for risk of failing in final exams or students who want to drop university or degree program. This will

identify the students and give recommendations for future improvements. So EDM is an educational software that predicts outcome.

Educational Data Mining Processes

The processes that are followed during Educational Data Mining are shown in figure (figure 16)

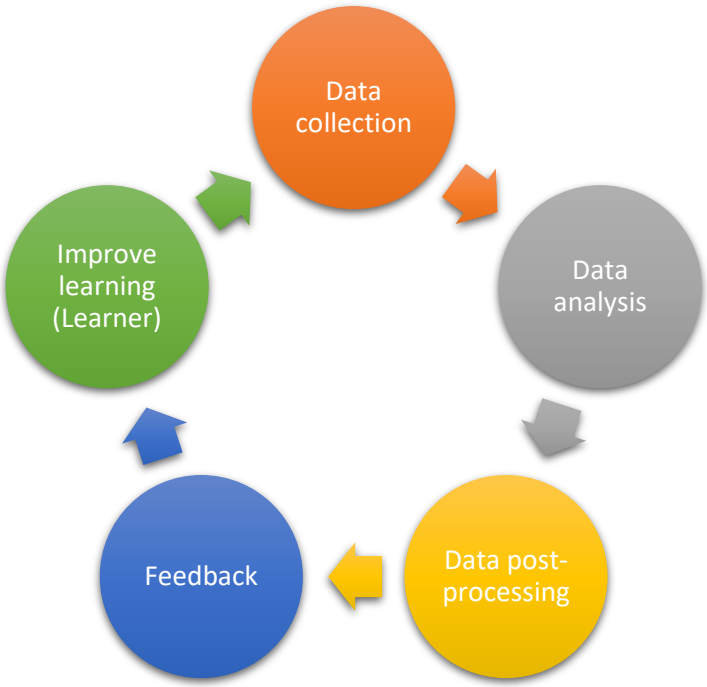

Figure 16 EDM processes

The EDM starts with data collection by conducting a survey to the students of BITe degree program in Haaga-Helia UAS. We go through all the data collected and start analyzing it. After, we must pre-process the data, feedback it and improve learning.

This is a cycle process, so every time a new data is added in the database, the process starts over again.

Educational Data Mining with KNIME

In this thesis, we are going to make students' performance analysis and some future predictions based on past behaviours. The Data Mining software we are using is KNIME Analytics Platform.

To do some data mining analysis in educational context using KNIME Analytics Platform, we need to follow a lifecycle as shown in figure (figure 17).

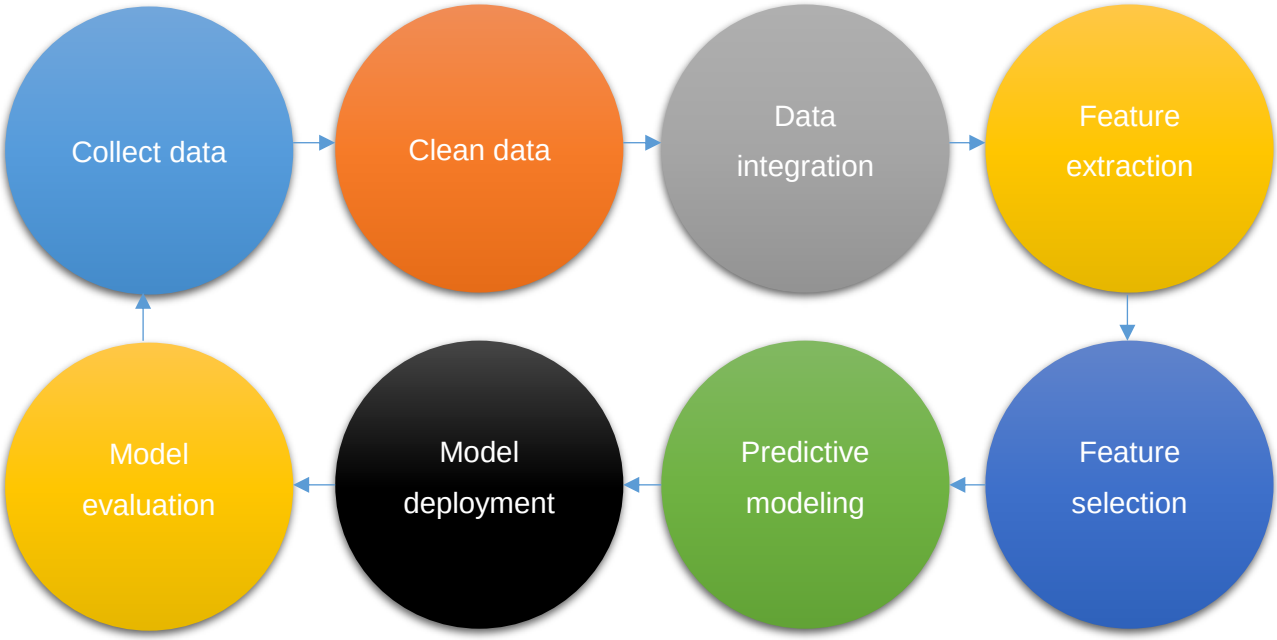

Figure 17 Lifecycle of data science

The main steps followed to get some analytic results and predictions are described in chapters 5 and 6.

Educational Data Mining in Haaga-Helia UAS

Haaga-Helia UAS is the university we are implementing educational data mining through KNIME Analytics Platform. Before we go to the implementation of KNIME Analytics Platform, there is an introduction of this university.

Haaga-Helia UAS

Haaga-Helia is a privately-run university of applied science, part of the Finnish public educational system, steered and co-founded by the Finnish Ministry of Education. (Haaga-Helia UAS, 2018b)

In figure (figure 18), there is a visual organization of Haaga-Helia UAS.

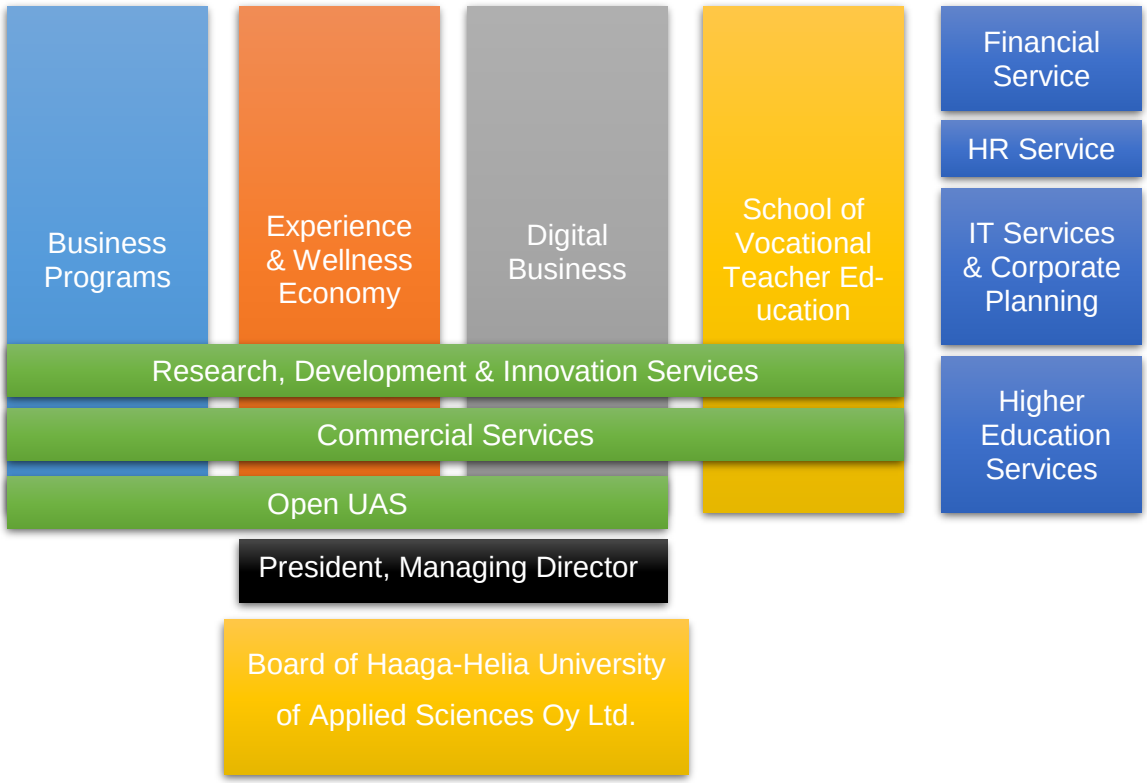

Figure 18 Organization of Haaga-Helia UAS

Haaga-Helia UAS provides quite many bachelor's degree programs, but the focus for us is the degree in Business Information Technology

Business Information Technology degree program

Business Information Technology degree program provides students with a wide range of professional skills in both business and information technology. This degree offers a combination of practical and extensive theoretical base, as well as practical application in co-operation with It or other kind of companies, both locally and internationally. This co-operation provides the students the opportunity to learn new skills and acquire the experience of working in a multicultural environment and on multinational projects. This is why, the integral part of BITe degree program is the collaboration with these companies. (Haaga-Helia UAS, 2018a)

Another important sector of this degree program is the career planning. This includes the development of entrepreneurship, sales and service abilities. What is more, there are many courses enhancing team work, communication, ability to work independently, skills required for being successful in a business environment. (Haaga-Helia UAS, 2018a)

So, as we can see, Business Information Technology degree program, enables the students everything needed for being successful in the career path.

EDM in BIT degree program

This thesis is conducted by a Haaga-Helia BITe student, so the knowledge about this degree program and experience are relevant for further research.

With implementation of educational data mining in BIT program, we aim to increase students' interest towards courses, how to minimize students' withdrawal and find and efficient way to match students' interests and strong points with the relevant courses.

Educational data mining is very beneficial for the university as well, because it will have more graduated students within three and a half years (degree program study timeline).

5 Design

This part will describe briefly the visual architecture of the whole process from data collection till visualization and the dataset of the project. More detailed information can be found in Gjergji Make's bachelor's thesis.

5.1 Visual architecture

As shown in figure (figure 19), the first step was to distribute the questionnaire to BITe Haaga-Helia students. When the questionnaire is filled, we are provided an Excel database with all the information. It is important to have data understanding and pre-processing, where we can filter the data provided and remove unnecessary data, such as wrong data.

We use this database to generate a machine learning database in KNIME Analytics Platform and start building the project. We create different meta-nodes and/or wrapped nodes for different data visualization.

The final step is data visualization and interpretation.

Figure 19 Visual architecture

5.2 Dataset

The dataset we are working during the project are the results gathered from the questionnaire (NIEMIVIRTA M. 2002) distributed to Haaga-Helia UAS students of BITe degree program in 2018. The questionnaire has 30 questions. The first 6 questions reveal basic information about the students and they are the main reference for further statistical analysis interpretations.

The dataset generated from the variables of the first 6 questions are as follow:

1. Gender:
    1. Female
    2. Male

2. Age:
    1. 18 – 21 years old
    2. 22 – 25 years old
    3. 26 – 29 years old
    4. 30 – 35 years old
    5. Older than 35 years

3. Nationality:
    1. Finnish
    2. European
    3. South American
    4. North American
    5. Asian
    6. African

4. Study Semester
    1. First semester
    2. Second semester
    3. Third semester
    4. Forth semester
    5. Fifth semester
    6. Sixth semester
    7. Seventh semester
    8. Etc

5. Accomplished a higher education before or not

6. BIT program was first choice or not

The other 24 questions are grouped in 8 categories with three questions each and scaled from 1 to 7 by the students.

The 8 categories according to (NIEMIVIRTA, 2002) are:

1. “Mastery-intrinsic orientation: “to acquire new knowledge is an important goal for me at school.
2. Mastery-“extrinsic orientation: my goal is to succeed in school.
3. Performance-approach orientation: “an important goal for me at school is to do better than other students.
4. Performance-avoidance orientation: “I try to avoid situations in which I might fail or make mistakes.
5. Avoidance orientation: “I try to get away with as little effort as possible in my school work.
6. Fear of failure: “during classes or exams I often worry that I do worse than the other students.
7. Academic withdrawal: “refers to an individual’s tendency to give up or withdraw from demanding or difficult learning or performance situation.
8. School value” (NIEMIVIRTA, 2002)

The detailed questionnaire is available in Appendixes 1.

6 Implementation

As indicated in chapter 3.1, we will work with 2 iterations. In the first iteration we will focus in the proof of concept to the deployment of the project. The second iteration we will focus on interpretation of KNIME workflow and discussion of results.

6.1 1st iteration

The first iteration is divided in six phases based on the CRISP/DM method. Business understanding is done in cooperation with Mr. Dirin and Gjergji Make, where the current status of Haaga-Helia UAS students is analysed and some alternative solutions have come into consideration. Also, the questionnaire (NIEMIVIRTA M. 2002) is handled to students of BIT degree program for data collection.

After that, together with Gjergji Make, we have followed the other five phases of CRISP/DM method, using KNIME Analytics Platform for the optimal solution. The big picture of KNIME workflow is as shown in figure (figure 20).

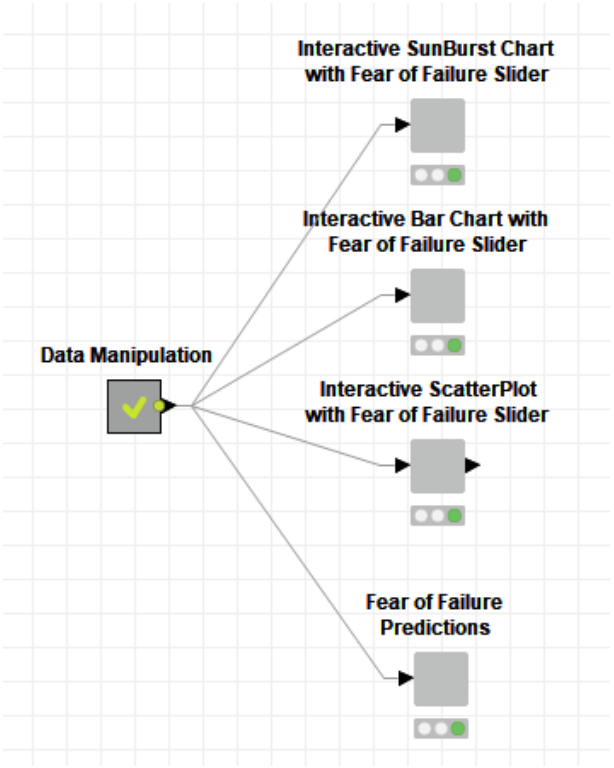

Figure 20 KNIME main workflow

In the KNIME main workflow we have Data Manipulation meta-node where all the data manipulations are handled. Also, there are wrapped meta-nodes. Each wrapped meta-

node is a group of JavaScript visualization nodes that can be displayed in a single HTML page.

Please read Gjergji Make's bachelor's thesis for a deep understanding of KNIME workflow processes that have been followed to achieve this data visualizations and interpretations of this thesis.

6.2 2nd iteration

During the second iteration of this thesis is about results, recommendations and discussions. These can be found in the upcoming chapters.

7 Results

Chapter 7 is dedicated to give straight results from the main bar charts created from all the data analysis that have been done during this thesis-projects.

In chapter 8, there is interpretation of results, so the graphs have meaning.

Some conclusions and recommendations are given on chapter 9.

7.1 Academic withdrawal

As shown in the figure (figure 21), academic withdrawal is affecting the average factor of mastery-intrinsic orientation, mastery-extrinsic orientation, performance approach-orientation and performance-avoidance orientation.

An average factor of 6.5 of mastery-intrinsic orientation is when we have a low academic withdrawal. The average factor of mastery-intrinsic orientation drops to 6.4 when there is medium academic withdrawal and 6.0 when there is high academic withdrawal.

An average factor of 5.7 of mastery-extrinsic orientation is when we have low academic withdrawal. When the academic withdrawal is medium, the average factor raises up to 5.8 and it drops to 5.4 when the academic withdrawal is high.

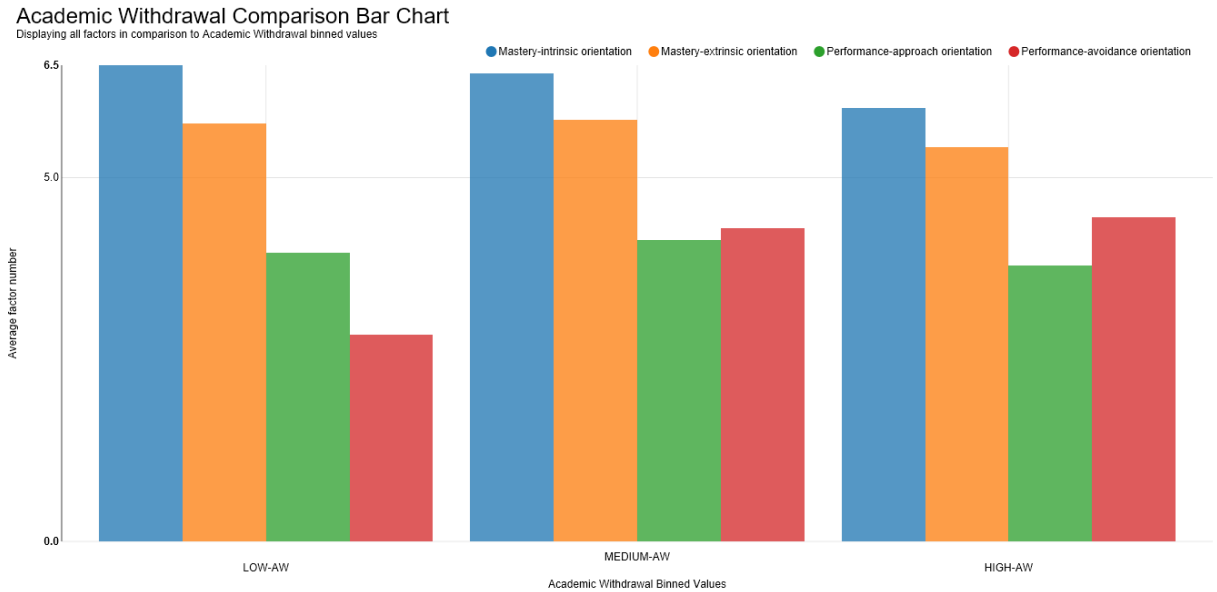

Figure 21 Academic withdrawal comparison

The performance-approach orientation has an average factor of 4.0 when the academic withdrawal is low. It raises up to 4.1 when academic withdrawal is medium and drops to 3.8 when the academic withdrawal is high. But as it is noticed from the graphic, the difference between the average factors is quite low.

An average factor of 2.8 of performance-avoidance orientation is when we have low academic withdrawal. When we have a medium academic withdrawal, the average factor of performance-avoidance orientation raises up to 4.3 and to 4.5 when the academic withdrawal is high. So, a low academic withdrawal indicates a higher average factor of performance-avoidance orientation.

7.2 Avoidance orientation

As shown in the figure (figure 22), an average avoidance orientation indicates a lower mastery-intrinsic orientation average factor. An average factor of 6.6 of mastery-intrinsic orientation is when we have a low avoidance orientation. The average factor drops to 6.2 when the avoidance orientation is medium and raises up to 6.7 when the avoidance orientation is high.

An average factor of 6.1 of mastery-extrinsic is when we have low avoidance orientation. The average factor reduces to 5.6 when the avoidance orientation is medium and to 5.0 when the avoidance orientation is high. So, the higher the avoidance orientation, the lower the average factor of mastery-extrinsic orientation is.

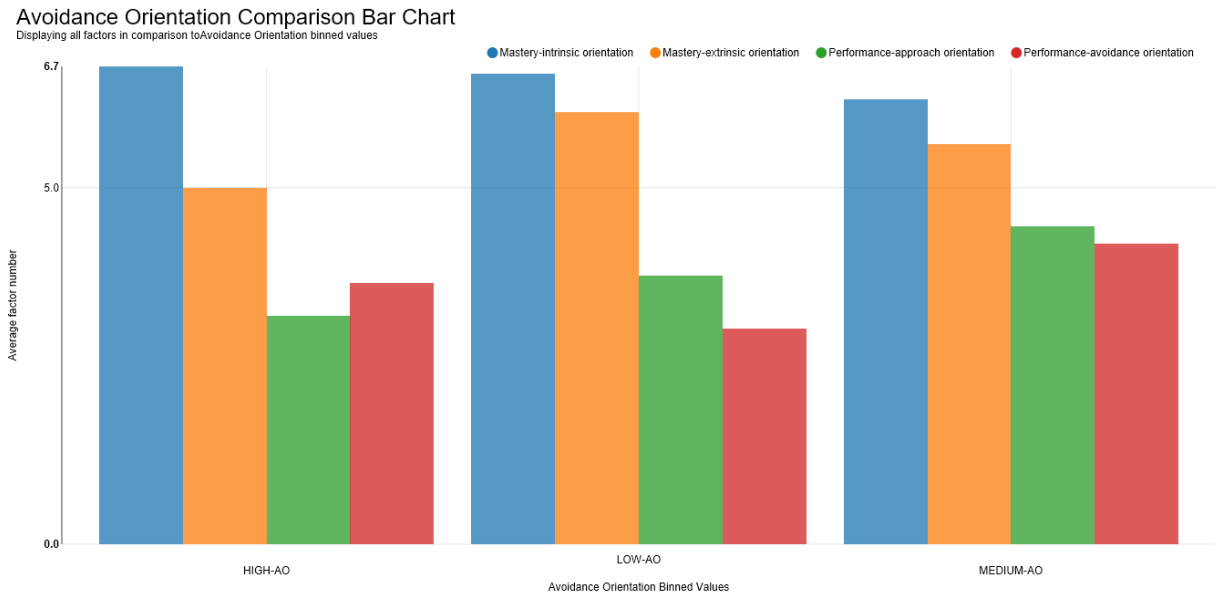

Figure 22 Avoidance orientation comparison

The performance-approach orientation has an average factor of 3.8 when the avoidance orientation is low. The average factor of it raises up to 4.5 when the avoidance orientation is medium and drops to 3.2 when the avoidance orientation is high. So, the average factor of the performance-approach orientation gets higher as the avoidance orientation is medium.

An average factor of 3.0 of performance-avoidance orientation is when we have low avoidance orientation. It raises up to 4.2 when we have a medium avoidance orientation and drops to 3.7 when the avoidance orientation is high.

7.3 Fear of failure

As shown in figure (figure 23), fear of failure is not affecting the mastery-intrinsic orientation average factor, but it has affected the mastery-extrinsic orientation average factor, performance-approach orientation average factor and performance-avoidance orientation average factor.

An average factor of 6.4 of mastery-intrinsic orientation is when we have high, medium or low fear of failure.

An average factor of 5.4 of mastery-extrinsic orientation is when we have low fear of failure. When we have medium fear of failure, the average factor of mastery-extrinsic orientation raises up to 5.8, and 6.2 when the fear of failure is high. So, the higher the fear of failure is, the higher the mastery-extrinsic orientation becomes.

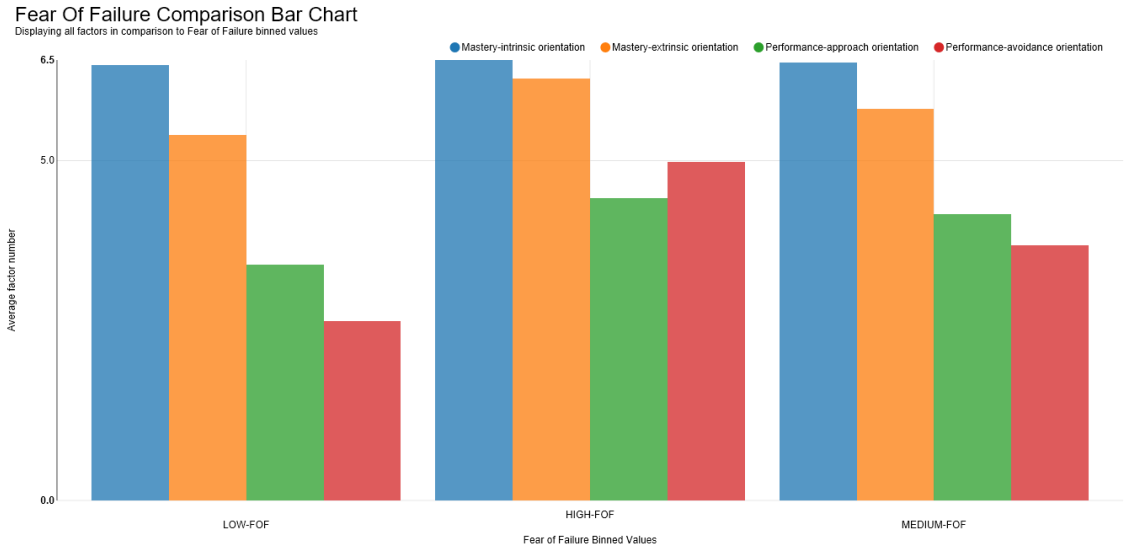

Figure 23 Fear of failure comparison

An average factor of 4.4 of performance-approach orientation is when the fear of failure is high. When we have medium fear of failure, the average factor of performance-approach orientation drops to 4.2 and to 3.5 when the fear of failure is low. As the fear of failure gets lower the performance-approach orientation drops.

An average factor of 5 of performance-avoidance orientation is when we have a high fear of failure. When we have medium fear of failure, the average-factor of performance-avoidance performance drops to 3.7 and to 2.6 when the fear of failure is low. So, the higher the fear of failure, the higher gets the performance-avoidance orientation average factor.

7.4 School value

As shown in figure (figure 24), school value has a huge impact in mastery-intrinsic orientation, mastery-extrinsic orientation and performance-avoidance orientation. The effect of school value in performance-approach orientation average factor is almost null.

An average factor of 6.5 of mastery-intrinsic orientation is when we have high school value. When we have a medium school value, the average factor of mastery-intrinsic drops to 6.2. A low school value will drop the mastery-intrinsic orientation average factor to 3.7.

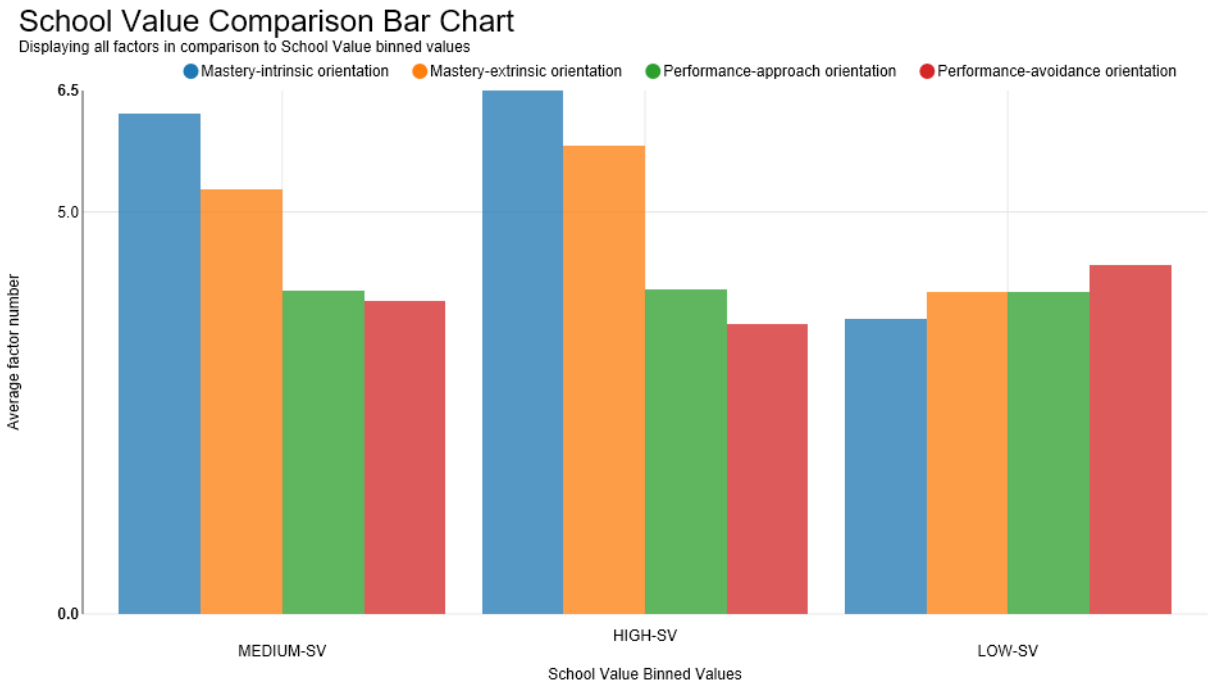

Figure 24 School value comparison

An average factor of 4.0 performance-approach orientation is in all the cases remaining the same.

An average factor of 3.6 of performance-avoidance orientation is when we have high school value. The average factor raises up to 3.9 when the school value is medium and 4.3 when the school value is low. So, as the school value is going down, the performance-avoidance orientation average factor is raising up.

7.5 Age vs avoidance orientation

As shown in figure (figure 25), students age 18-21 have an average factor of avoidance orientation of 3.4. Students age 22-25 have an avoidance orientation of 3.2 and age 26-29 of 3.4. The avoidance orientation average factor drops to 2.8 for students age 30-35 and to 2.0 for students who are older than 35 years old.

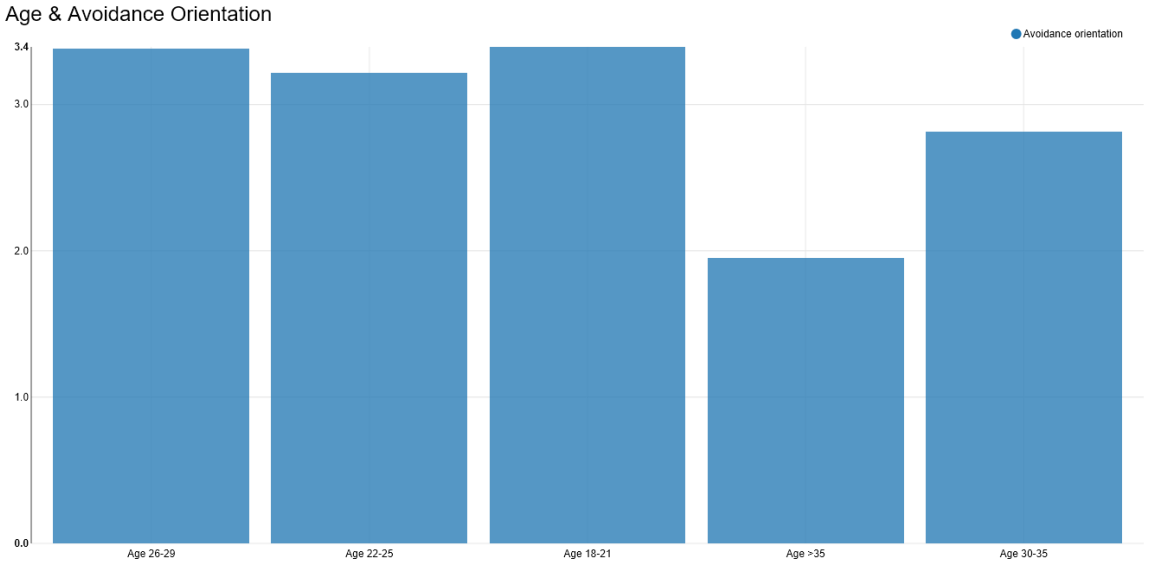

Figure 25 Age vs avoidance orientation

7.6 Nationality vs fear of failure

Nationality has some impact in fear of failure average factor. As shown in the figure (figure 26), the average factor of fear of failure for people from Asia is 3.7, 3.3 for people from Finland. Europeans have an average factor of 3.5 of fear of failure and Africans of 3.2. Americas have a higher average factor of fear of failure, where North America has 4.3 and North America 4.1.

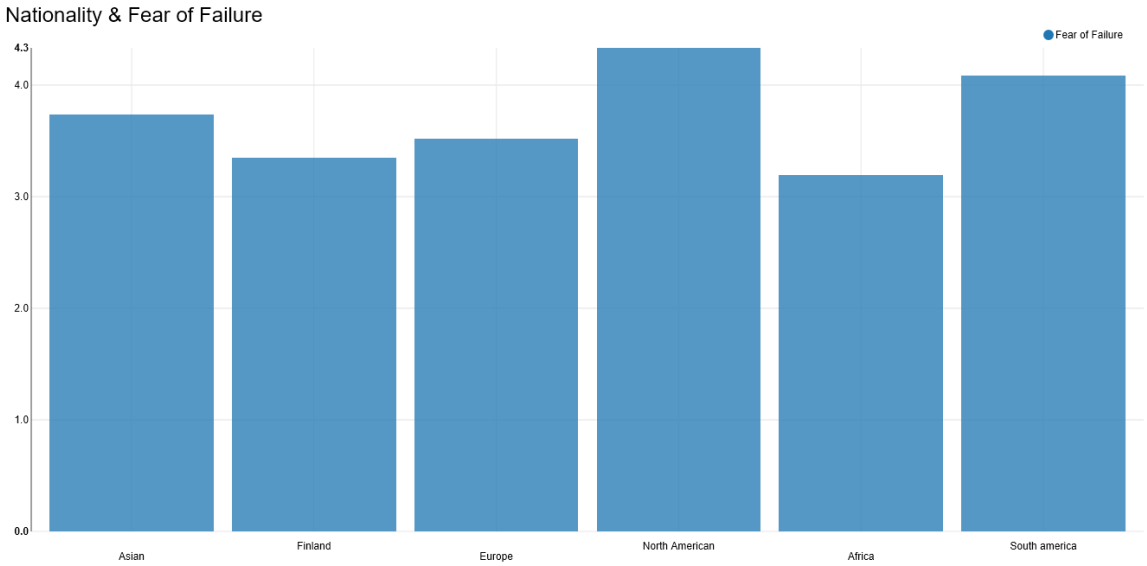

Figure 26 Nationality vs fear of failure

7.7 Gender comparison

In figure (figure 27), we have a comparison between two genders with all eight categories. From a general prospective, it is noticed that there are not many differences between males and females.

Mastery-intrinsic orientation has an average factor of 6.5 for male students and 6.4 for females.

Males have an average factor of mastery-extrinsic orientation of 5.7, while females have 5.8.

An average factor of 3.9 of performance-approach orientation is for males and 4.2 for females.

Performance-avoidance orientation has an average factor for male students of 3.7 and female of 3.6.

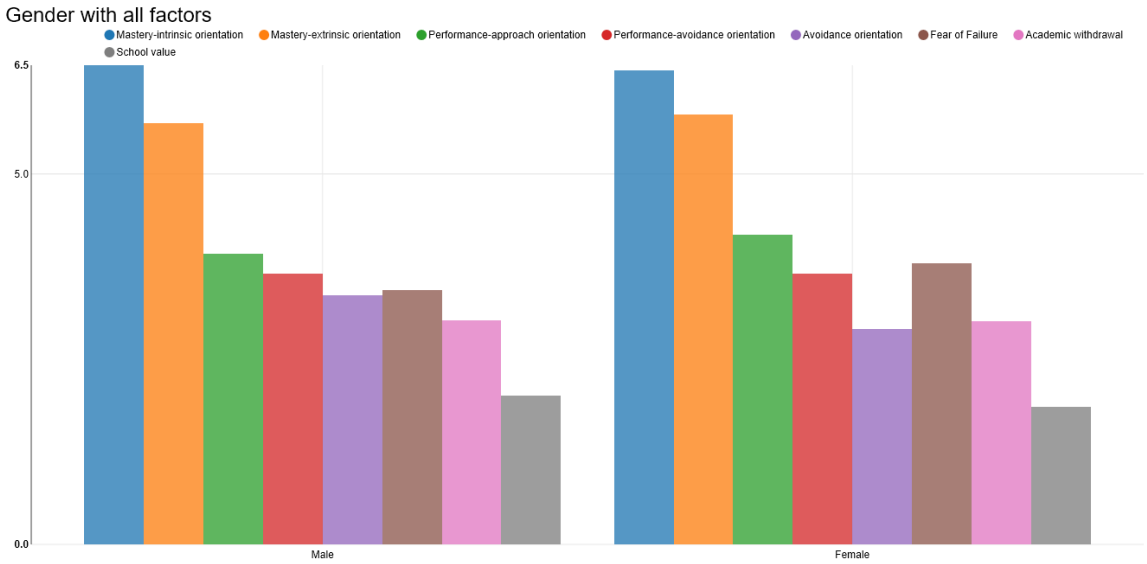

Figure 27 Gender comparison

An average factor of 3.4 of avoidance orientation is for male and 2.9 for female students.

Fear of failure has an average factor of 3.4 for males and 3.8 for females.

Academic withdrawal average factor of 3.0 both for female and male students.

School value has an average factor of 2.0 for males and 1.9 for females.

8 Discussion

In this chapter we are going to interpret the results by giving meaning to all the values of the average factors. Scaling of average factors vary from zero to seven. In seven categories of the questionnaire, mastery-intrinsic orientation, mastery-extrinsic orientation, performance-approach orientation, performance-avoidance orientation, avoidance orientation, fear of failure and academic withdrawal zero is low and seven is high. In school value, zero is high and seven is low.

Questions belonging to each category, as given by Niemivirta, are available in Appendixes 1.

8.1 Academic withdrawal

Academic withdrawal is related with uncertainty that students have in school tasks.

As shown in chapter 7.1, academic withdrawal has supremacy to performance-avoidance orientation, so students who cannot concentrate on demanding tasks during their school tasks, are more likely to stop their studies. By taking a proper action Haaga-Helia UAS may sway in students' behaviours.

Students that have a high susceptibility towards academic withdrawal, they have a lower mastery-intrinsic orientation, so they lose the motivation to learn new things.

As shown in figure (figure 21), students with a low academic withdrawal aspire to do better in their studies and succeed. That is why academic withdrawal is not influencing mastery-extrinsic orientation.

Academic withdrawal have a very low influence in performance-approach orientation, which means that students who feel that they have achieved their study goal and consider themselves capable to continue further, are not giving up easily.

8.2 Avoidance orientation

Avoidance orientation is related with students who do only the required assignments and try to complete their studies with as less effort as they can.

As shown in chapter 7.2, students who want to succeed in their studies and target high grades, are slighter likely to do only the compulsory schoolwork. So, students that have high avoidance orientation stipulate high mastery-extrinsic orientation.

Avoidance orientation has nearly no impact in mastery-intrinsic orientation. This means that students do not give up easily if they have ambition to learn new things and acquire new knowledge all the time,.

BITe students who put few efforts in school tasks and try to be done with as less work as possible have no concern about their results and others' opinions. For them is significant just to pass. That is why their performance-approach orientation flatten when the avoidance orientation is high.

Students with a high avoidance orientation, have a high performance-avoidance orientation. Differently said, all BITe students who try to be done with the minimum work required, try to refrain from unbearable situations.

8.3 Fear of failure

As shown in chapter 7.3, students who are acquiring new knowledge during their studies, is not influenced by fear of failure. This means that almost all the students of BITe degree program are open to learn new things despite their worry for failing or not doing as well as others in class.

Mastery-extrinsic orientation is higher when the fear of failure is higher. This means that students who are agonized about exams and afraid of wrong answers want more to succeed in their studies than students that have not too much concern.

Students that have a goal to do better than others and get better results have a higher fear of failure because they are scared if they know the right answer or if they will fail in some tests or exams.

Performance-avoidance orientation is the category that is more pretentious by fear of failure. In other words, students who avoid situations where they feel incompetent, often worry if they are doing worse than others. The more students get out of their comfort zone, the less they are likely to fear different school situations.

8.4 School value

School value plays a very important role in students' performance during their studies. As shown in chapter 7.4, students are more willing to learn new things and acquire knowledge when there is a high school value. When the school is not very useful in students' perspective, then they lose the interest.

There is the same situation with mastery-extrinsic orientation. Students have the goal to succeed in their studies and get good grades when the school value is high. If the school value is not high, then the mastery-extrinsic orientation will drop and students are not focused in doing well in their studies.

Performance-approach orientation is remaining almost the same despite school value. In other words, students are competing between each-other, they all want to do better in their studies and feel competent in front of others.

Performance-avoidance orientation is higher when school value is low. This means that students try to avoid situations they feel they might be wrong or failing in front of others when they see school useless and waste of time. So, a high school value is eliminating those situations.

8.5 Age vs avoidance orientation

Chapter 7.5 gives some interesting facts facing age with avoidance orientation. All the students age between 18 to 21 and 26 to 29 are the ones who are more exposed to put less effort in school tasks and do only the required assignments. As the students get older, they are less likely to be avoidance oriented. We can notice that students age 30 to 35 put more effort on school works. Avoidance orientation drops drastically for students that are more than 35 years old.

8.6 Nationality vs fear of failure

Fear of failure is an obstacle that stands between the students and their goal. This will cause worry during classes and exams.
In chapter 7.6, we can see the visualization of fear of failure with nationality and we can notice that students from Africa have less fear of failure during their studies than students from other nationalities.

Students from North America have the highest fear of failure, followed by South America. These students are afraid to do worse than others and they are afraid they might fail.

Europeans have a medium fear of failure.

Finnish students have a low fear of failure and this can be as a result they know well Finnish educational system and the culture.

8.7 Gender comparison

The bar chart in chapter 7.7 that compares mastery-intrinsic orientation, mastery-extrinsic orientation, performance-approach orientation, performance-avoidance orientation, academic withdrawal, fear of failure, avoidance orientation and school value between male and female students visualises not many differences between the two genders.

We can notice that both male and female have the goal to acquire new knowledge during their studies.

Another factor is mastery-extrinsic orientation, that is almost equally for both genders with a high importance. This means that almost all the students have the desire and goal to do well in school.

Performance-approach orientation has a difference between males and females. We can conclude that female students are more demanding in grades and results than males and they want others to think that they are competent in what they are doing.

Performance avoidance orientation is almost the same for males and females with a medium average factor. This means that all BITe students are avoiding some situations they do not give their best, but this is not a common factor.

Avoidance orientation is higher in female than male BITe students. In simpler words, female students are more likely to do only required tasks and finalize the studies with as less effort as possible.

Fear of failure is another factor that is higher in females than in males. This means that BITe female students are more afraid to fail in exams and worried for their progress than males. This can come because of the degree program, but it is still a hypothesis.

Academic withdrawal is exactly the same for males and females with a medium average factor. This way we understand that all BITe students have some uncertain situations, but they do not give up easily.

School value is highly appreciated by all BITe students who took part in the survey. This is a very important information because we understand that students find Haaga-helia UAS as a useful polytechnic that brings value towards their studies.

8.8 Recommendations

Based on the discussions, you can find in this thesis few recommendations, but they can change if more data is provided for analysing.

Recommendation 1:
Haaga-Helia UAS can accept more students from Africa in BITe degree program to low down the fear of failure during studies. Students from North America have high fear of failure in this degree, so they can be suggested to attend another degree program for more self-confidence in their studies.

Recommendation 2:
Haaga-Helia UAS can accept more female students for higher grade point average, because females are more demanding and determined to achieve better results than the others.

Recommendation 3:
Haaga-Helia UAS can put more effort to improve school services for raising school value in students. This way students have the goal to succeed in their studies and get good grades.

9 Conclusion

The primary objective of this thesis was to make students' performance analysis using KNIME Analytics Platform by taking into consideration the results of the (NIEMIVIRTA, 2002) questionnaire distributed to BITe students.

We started with proof of concept by creating a KNIME project with Gjergji Make. We were provided data from 100 students of BITe degree program taken in early 2018. To build the project was time-efficient but coming up with future trends and predictions based on the amount of data provided was time-consuming and yet not quite possible. For this reason, we changed the course of this thesis to data analysis.

Analytics conducted in this thesis are based on a small amount of data which have no time tracking on it. For this reason, we can say that making future trends predictions is not possible but making future trend strong assumptions based on the current situation is possible.

This topic was very interesting to work with and learned many new things in data analytics field as well as students' performance. Surprising findings made this work interesting as well as challenging.

To conclude, learning curve changed throughout project development from not succeeding on predicting future trends due to lack of data to achieving to analyse current situation of students' performance.

References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 (Step-by-step data mining guide)*. *CRISP-DM Consortium*. https://doi.org/10.1056/NEJMoa1108524

Copeland, M. (2016). What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning? *Nvidia*.

Haaga-Helia UAS. (2018a). DEGREE PROGRAMME IN BUSINESS INFORMATION TECHNOLOGY, HELSINKI. Retrieved from http://www.haaga-helia.fi/en/education/bachelor-degree-programmes/degree-programme-business-information-technology?userLang=en

Haaga-Helia UAS. (2018b). Organization. Retrieved from http://www.haaga-helia.fi/en/about-haaga-helia/organization?userLang=en

Houle, P. (2016). A brief history of Machine Learning. Retrieved from https://www.linkedin.com/pulse/brief-history-machine-learning-paul-houle/

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*. https://doi.org/10.1177/1558689806298224

KNIME.COM AG. (2016). KNIME Analytics Platform.

Margaret Rouse. (2017). Data Mining. Retrieved from https://searchsqlserver.techtarget.com/definition/data-mining

Marr, B. (2018). What Is Deep Learning AI? A Simple Guide With 8 Practical Examples. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/10/01/what-is-deep-learning-ai-a-simple-guide-with-8-practical-examples/#632260e18d4b

Mitchell, T. M. (2006). The Discipline of Machine Learning. *Machine Learning*. https://doi.org/10.1080/026404199365326

Naik, A., & Samant, L. (2016). Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. In *Procedia Computer Science* (Vol. 85, pp. 662–668). https://doi.org/10.1016/j.procs.2016.05.251

Ncr, P. C., Spss, J. C., Ncr, R. K., Spss, T. K., Daimlerchrysler, T. R., Spss, C. S., & Daimlerchrysler, R. W. (2000). Step-by-step data mining guide. *SPSS Inc*. https://doi.org/10.1017/CBO9781107415324.004

NIEMIVIRTA, M. (2002). MOTIVATION AND PERFORMANCE IN CONTEXT: THE

INFLUENCE OF GOAL ORIENTATIONS AND INSTRUCTIONAL SETTING ON SITUATIONAL APPRAISALS AND TASK PERFORMANCE. *PSYCHOLOGIA -An International Journal of Psychology in the Orient*. https://doi.org/10.2117/psysoc.2002.250

Orcibal, J., & Jerphagnon, L. (2018). Blaise Pascal FRENCH PHILOSOPHER AND SCIENTIST. Retrieved from https://www.britannica.com/biography/Blaise-Pascal

Pong, S. (2017). What statistical techniques are used to perform data analysis? Retrieved from https://www.quora.com/What-statistical-techniques-are-used-to-perform-data-analysis

Pyle, D., & San Jose, C. (2015). An executives guide to machine learning. *2015*. https://doi.org/10.1017/CBO9781107415324.004

Roy, S., & Garg, A. (2017). Analyzing performance of students by using data mining techniques a literature survey. *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*. https://doi.org/10.1109/UPCON.2017.8251035

SAS Institute. (2018). Data Mining What it is and why it matters. Retrieved from https://www.sas.com/en_us/insights/analytics/data-mining.html

Skillsyouneed. (2017). Simple Statistical Analysis. Retrieved from https://www.skillsyouneed.com/num/simple-statistical-analysis.html

Sunil, R. (2016). Essentials of Machine Learning Algorithms (with Python and R Codes).

Team, E. (2013). What is Educational Data Mining (EDM)? Retrieved from http://edtechreview.in/dictionary/394-what-is-educational-data-mining

WebFinance Inc. (2018). Statistical Analysis. Retrieved from http://www.businessdictionary.com/definition/statistical-analysis.html

Appendices

Appendix 1. Niemivirta questionnaire

Scale ORDER	Scale	Questions belong to the scale
A	Mastery-intrinsic orientation	Q12. I study in order to learn new things.
		Q22. An important goal for me in my studies is to learn as much as possible.
		Q28. To acquire new knowledge is an important goal for me in school.
B	Mastery-extrinsic orientation	Q9. An important goal for me is to do well in my studies.
		Q23. It is important to me that I get good grades.
		Q27. My goal is to succeed in school.
C	Performance-approach orientation	Q7. An important goal for me in school is to do better than the other students.
		Q19. I feel I have attained my goal if I get better results or grades than many other students.
		Q24. It is important to me that others think I am able and competent.
D	Performance-avoidance orientation	Q11. I try to avoid situations in which I may appear dumb or incompetent.
		Q15. I try to avoid situations in which I may fail or make mistakes.
		Q21. It is important to me that I don't fail in front of other students.
E	Avoidance orientation	Q13. I am particularly satisfied if I don't have to work much for my studies.
		Q18. I try to get away with as little effort as possible in my schoolwork.
		Q26. I always try to do nothing more than just the required schoolwork.
F	Fear of Failure	Q10. In classes I often worry that I don't understand or that I don't know the right answers.
		Q14. During classes or tests, I often worry that I do worse than the other students.
		Q30. I always worry about failing in tests and exams.
G	Academic withdrawal	Q8. I always feel very nervous and uncertain, when I should concentrate on a demanding or difficult school task.
		Q17. I have realized that I give up easily, if school tasks are difficult.
		Q20. I have realized that it's very hard for me to fully concentrate when I should work on a demanding school task.
H	School value	Q16. Studying is boring.
		Q25. I feel that studying and going to schools is useless.
		Q29. I think going to school is a waste of time.