

# Implementing KNIME Analytical Platform for visualizing data in educational context

Gjergji Make



<b>Author(s)</b> Gjergji Make	
<b>Degree programme</b> Business Information Technology	
<b>Report/thesis title</b> Implementing KNIME Analytical Platform for visualizing data in educational context	<b>Number of pages and appendix pages</b> <b>50 +   1</b>
<p>Big Data is playing a big role in technology Industry nowadays. Decision making is becoming easier throughout Data mining and Machine learning. There is vast amount of data provided by human beings' behaviour and technology usage. With the right interpretation on the proper context these data are far more than useful to predict and make correct decisions for better performance.</p> <p>This thesis paper will be focusing on implementing Data Mining (DM), Machine Learning (ML) and Data Analytics to BITE Program students of Haaga-Helia University. Dataset used in this project are collected using the study of (NIEMIVIRTA, 2002) on eight scales of motivational factors. Results of this questionnaire are constructed, transformed and visualized using KNIME Analytical Platform. Using the same platform data mining will be implemented to this dataset for understanding and finding hidden truths for increasing student performance.</p> <p>CRISP-DM methodology is used to develop this project. There will be 5 main sub-chapters explaining step by step process of how the results are achieved:</p> <ol style="list-style-type: none"> <li>1. Data understanding</li> <li>2. Data preparation</li> <li>3. Modeling</li> <li>4. Evaluation</li> <li>5. Deployment</li> </ol> <p>To summarize, this thesis will give the reader a clear structure of the KNIME project, important student behaviour data visualization and simple prediction of fear of failure motivation factor.</p>	
<b>Keywords</b> CRISP/DM, Data Mining, KNIME Analytics Platform, Data Visualization, student performance.	

## Table of contents

Acknowledgements.....	3
Abbreviations.....	4
1 Introduction .....	1
2 Research question .....	2
2.1 Sub-research questions .....	3
3 Methods .....	4
3.1 Methodology .....	4
3.2 Workflow processes.....	4
4 Background study .....	9
4.1 Data Analytics .....	9
Descriptive data analytics .....	9
Predictive data analytics .....	10
Machine Learning .....	10
4.2 10	
Timeline of Machine Learning .....	11
Categorisation of Machine Learning.....	12
4.3 Data mining.....	14
Statistical Modeling .....	14
Statistical Limits on Data Mining .....	14
Data Mining tasks .....	15
Data Mining Tools .....	15
5 Design.....	19
5.1 Understanding the dataset .....	19
6 Implementation.....	21
6.1 KNIME Workflow Development .....	21
6.2 Data understanding.....	22
6.3 Data preparation .....	23
Read Data.....	23
Construct Data .....	24
Data Visualization Module.....	30
6.4 Modeling module.....	41
6.5 Evaluation .....	45
6.6 Deployment.....	47
7 Results.....	49
8 Conclusion .....	50
References .....	51

## Table of figures

Figure 1 CRISP-DM method .....	5
Figure 2 Raw Dataset .....	20
Figure 3 Niemivirta Scaling Questionnaire .....	21
Figure 4 File Reader Node configuration .....	23
Figure 5 Data Manipulating Metanode .....	25
Figure 6 Rule based row filter node configurations .....	26
Figure 7 Metanode for calculating averages .....	27
Figure 8 Math Formula node (Fear of Failure) .....	28
Figure 9 Column Filter Node configurations .....	29
Figure 10 Column Rename node configurations .....	30
Figure 11 Raw constructed data table after manipulations are implemented.....	31
Figure 12 List of Knime default views.....	31
Figure 13 JavaScript Interactive views .....	32
Figure 14 Interactive Sunburst chart with overall data example .....	32
Figure 15 Main KNIME workflow.....	34
Figure 16 Combination of three interactive nodes .....	34
Figure 17 Range Slider Filter Definition configuration .....	35
Figure 18 Interactive slider.....	35
Figure 19 Sunburst chart configuration .....	36
Figure 20 Niemivirta factors Sunburst chart .....	36
Figure 21 JavaScript table View configurations.....	37
Figure 22 JavaScript table viewing select data from Sunburst chart.....	38
Figure 23 Interactive Bar Charts wrapped metanode collection .....	39
Figure 24 Javascript Bar Chart configuration .....	39
Figure 25 Fear of Failure Comparison Bar Chart .....	40
Figure 26 Mining possibilities in KNIME Analytical Platform .....	41
Figure 27 Machine Learning module.....	42
Figure 28 Partitioning node configuration.....	42
Figure 29 Decision Tree Learner Configuration .....	44
Figure 30 Decision Tree Learner Tree Model.....	44
Figure 31 Decision Tree Learner Simple Model .....	45
Figure 32 Decision Tree Predictor configuration .....	45
Figure 33 Confusion Matrix for Tree Ensemble model Scorer .....	46
Figure 34 Confusion Matrix from Decision Tree model Scorer .....	47
Figure 35 PMML Writer .....	47
Figure 36 Test environment project with reporting feature .....	48
Figure 37 BIRT Report Viewer in sample testing environment .....	48

## **Acknowledgements**

I would like to thank my thesis advisor, Dr. Dirin, for the continues support during my thesis. Dr. Dirin has also shown his professionalism during many courses he teaches. I am so honoured to be one of his students and learn from him.

Another appreciation is for Mr. Silpiö, a senior Haaga-Helia lecture and my personal advisor. His courses are of a high expertise and provide value for working life. Mr. Silpiö had a very important role in my study planning.

Enxhi Nikolla, Haaga-Helia student and my wife, for cooperating with me in this thesis project and in many other university projects. She has been a great support for me in my studies.

Mr. Genoud, a senior lecture in Hes-So Valais, Switzerland, for his inspiring Data Mining course. Thank to him, I have good knowledge of KNIME Analytics Platform.

Also, I want to thank Mr. Välimäki, a senior teacher in Haaga-Helia, for his fundamental programming courses that helped me during my studies and finding an IT job.

My appreciation goes to all Haaga-Helia staff for the great work they have done.

## **Abbreviations**

KNIME - Konstanz Information Miner

MOA - Massive Online Analysis

KEEL - Knowledge Extraction for Evolutionary Learning

Rattle - R Analytical Tool To Learn Easily

ML – Machine Learning

UAS – University of Applied Sciences

JVM – Java Virtual Machine

IDE – Integral Development Kit

# 1 Introduction

The amount of data generated nowadays with the current level of technology is insane high. It is interesting looking on infographics done from (Josh James, 2016) which shows a general idea of the amount of data generated in one minute: ([https://web-assets.domo.com/blog/wp-content/uploads/2016/06/16\\_domo\\_data-never-sleeps-4-2.png](https://web-assets.domo.com/blog/wp-content/uploads/2016/06/16_domo_data-never-sleeps-4-2.png))

With these numbers in the board we understand that now it is much easier to make statistical studies which help us understand our goals. It is very easy to access worldwide statistics in most of life aspects related to technology. It is the reader's duty to find the best way of interpretation of these statistics and how much they influence in decision making. Misinterpretation of statistical data can lead to distraction and decision-making failure. It is crucial that the data available should be interpreted properly

The knowledge gained by these data amount that human beings are generating every minute is priceless. Through this knowledge people evolve and innovate.

"Humans today are like most smartphones and tablets" is the saying of Tomas Chamorro-Premuzic. What does he mean by this? Nowadays we can find a lot of information anywhere in the world, enough that we know how and where to search for it. We can out-source, crowdsource, cloud source. Human knowledge and storage of information is going nominal. You just need the capacity to connect hitch the right sources and find the solution. (Tomas Chamorro-Premuzic, 2013)

The previous paragraph shows that all this knowledge surrounding humans with easy access is making us incapable.

I relate these findings to the main property of the case study of this paper which are students. Students are the sponge of the knowledge. They are the ones who are meant to learn this knowledge and innovate after processing it. Of course, some of this group are not as good as the others and drop out of the process of learning or (studies) and some other succeed.

Nowadays we know that an excellent student who has great performance in university can be one who knows how to find knowledge and not really knowing it. Some other good students are the ones who have no fear of failure and courage to face new knowledge and

why not fail couple of times on learning it. The ones whose surroundings influence their performance are indicated to be good in knowing but not in showing it and many more.

To conclude, all the learning process done on universities by students somehow influence the overall society knowledge. That is why, in this thesis, we try to combine newest technology (machine learning) in helping to increase study performance by learning from 8 main factors which indicate university students nowadays shown the study of (NIEMIVIRTA, 2002):

1. Mastery-intrinsic orientation (One of the most important goals in school is acquiring new knowledge)
2. Mastery-extrinsic orientation (Most important goal is to succeed at school)
3. Performance-approach orientation (Performing better than other students is an important goal)
4. Performance-avoidance orientation (Avoiding situations where failing and making mistakes are most likely)
5. Avoidance Orientation (Trying to get away with little effort in school)
6. Academic Withdrawal (Tendency to give up in or withdraw from demanding or difficult learning or performance situations)
7. Fear of failure (Worrying that would do worse than the others in exams or classes)
8. School value (Giving value to the school and its learning processes)

Machine learning solution will be implemented to find out how all these factors are related to each other and to overall study performance with the goal of visualising benefiting statistical results using KNIME Analytical Platform.

## **2 Research question**

The work done in this thesis is aimed to answer this main research question:

How to use KNIME Analytics to develop an interactive visualisation tool for students' performance analyses?

KNIME Analytics software provide vast amount of functions which can help in answering these questions. Focusing on these functionalities a project will be developed and explained step by step using data provided.



## 2.1 Sub-research questions

To fully understand the main research question, we must go through a filter of sub questions such as:

What findings can we relieve from KNIME analytics data visualization functionalities to help students' performance understanding?

It is crucial for the reader to understand the results of the first raw findings from data analysis done in this research. KNIME Analytics Platform offers variety of algorithms which makes it easier to understand the general findings from the data provided. The goal of answering this sub question is to fully understand the given dataset.

In some scenarios, experience has shown that the biggest findings are done after pre-processing the given data and making easy to understand visualisations of them.

What are the trends and approach to increase students' academic performance?

Above we mentioned eight factors from (NIEMIVIRTA, 2002) study that influences the overall academic performance. Machine learning models will be trained in KNIME with the given data to find out how fear of failure indicates the overall students' performance.

Is the data available accurate enough to make good behaviour prediction?

Machine learning models need to have accurate data to make accurate predictions. After training the models we will evaluate the accuracy of the predictions by using the Score Node (Node to compare predicted data with the real data for evaluating model accuracy). This sub-question will let us know how good our predictions are and was the data enough accurate to do these predictions.

## **3 Methods**

### **3.1 Methodology**

Data analytics and machine learning are quite complicated to be implemented in a dataset. For this reason, having the project organized is important. We will divide this project in 3 main iterations: Data preparation, Data visualisation and Machine Learning.

The first part of the project will be focusing mainly on data preparation. Structure of the dataset given is not clear enough to implement any data visualisation nor machine learning. Giving meaning to the values provided will help us conducting proper solution on the problem. This step will be elaborated in detail on the Implementation chapter of this thesis.

When having understandable data in hand we will be ready to work on finding best ways on how to visualize them. KNIME platform offers variety of possibilities in data visualisation. There will be focus on doing interactive data visualization for better performance.

The last but not the least will be implementing machine learning for predicting 'Fear of Failure' factor. By implementing this part on the project there will be big value added to the student performance studies. This will also give a chance to the user to interact with the system by inserting own dataset and let the system predict this factor. Deeper explanation will be given later on the implementation phase.

### **3.2 Workflow processes**

To conduct a professional analytics and data mining, We will follow the cross-industry process for data mining CRISP-DM (Chapman *et al.*, 2000) methodology. There will be 6 main steps to be followed are shown in the figure (figure 1).

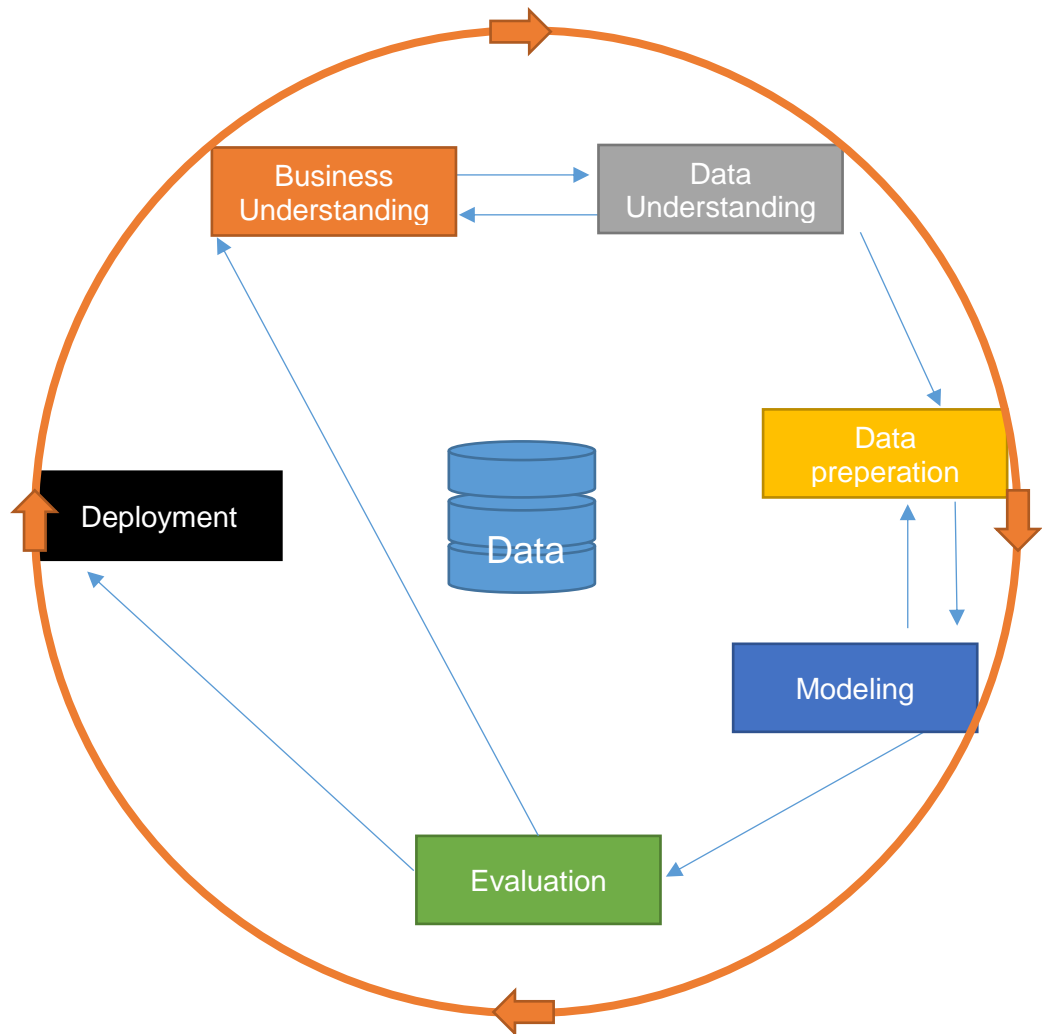


Figure 1 CRISP-DM method

### 1. Business understanding / Problem Understanding

First step of this methodology is to understand the purpose of why we are conducting this Data mining.

We intend to acquire the required business objectives, as seen from a business prospective, during the business understanding phase. We should generate a data mining problem definition based on the business needs. Afterwards, we can make the project plan. (M.P. Bloothoofd, A. Francken, 2018)

### 2. Data Understanding

Raw data most of the time are not that straight forward and you need to make sure to understand them properly before moving forward. Source of the data and form of data collection helps on understanding dataset. If you don't get familiar with the dataset then you must go back to the main problem we need to solve and see if the data provided is accu-

rate enough for solving the problem decided on the previous phase. If you end up deciding that the problem can't be solved with the available resources, then you either stop working on the project or change the scope to a profitable lane.

In data understanding phase, we have a lot of available data that has been collected from the project resources. Despite that, we cannot just start our procedures if we do not understand the data, so in this phase we check the data carefully to get familiar with it. Moreover, we investigate the data to point out the main characteristics of it. When we are able to distinguish the data, we reconnoitre for particular data mining questions and we test its quality. (M.P. Bloothoofd, A. Francken, 2018)

### **3. Data preparation**

The third phase of CRISP/DM method is data preparation. All the required activities for converting the preliminary raw data into the final dataset are included in this phase. It is very important that the data is clean, which means there are no empty rows or data with no meaning. Regardless this, we still might need to work with dataset for things like: derivation of attributes by errors in calculation or generation of new records. Furthermore, we have data integration and combination of information from different tables. All these small things can bring the need to transform the data, so it is proper for the modelling tools. (M.P. Bloothoofd, A. Francken, 2018)

Easier said, after fully understanding the resources we have in hand we need to prepare them for data mining. Usually these resources are collected in different ways such as numeric type, text type, voice type etc. as well as mixed of all. It is our duty to clean, adjust and prepare the final set that is going to be used for making the calculations needed. Most of the time if the data are provided as numeric there is a need for numeric translations to provide meaning.

### **4. Modeling**

Using the provided data mining tools, we must build models out of the final dataset generated in the data preparation phase. Various tools offer different ways of how to build these models. Best approach at this point of the project is trying many available models and reveal more than one DM modelling results for comparing reasons. This will widen the possibilities and raise our accuracy in the DM.

Essentially, when we have done data pre-processing, we can move on to the modelling phase. There are many accessible data mining techniques that help us to achieve the

data mining goal. Still, before we go to the actual model, we create a test model to evaluate the quality and validity of it. Then, the data mining engineer measures the accuracy of the model based on the given results. (M.P. Bloothoofd, A. Francken, 2018)

### **5. Evaluation (compute standard error to compare models)**

Performing evaluation to our solutions is very important. In this phase we will define if we managed to solve the problem. Evaluations should be done from different perspectives. This means that there should be involved not only technical parties but also other business-related individuals. It might be so that even though we did our best to conduct best modelling provided by the DM tool but still did not manage to solve the solution fully or not at all. At this point the whole project must be reconsidered from the beginning. This check point on the overall process is the most important part because it puts together every aspect of the project for evaluating the results.

In contrary to modelling phase, where we evaluate the model from the technical data mining viewpoint, during evaluation phase we need to evaluate the model from the business prospective, where we check if we reached the business objectives. The moment, the business needs are reached, we must go through all the process to check for any missing aspects. Finally, the researcher has to conclude with a settlement for the upcoming course of action. (M.P. Bloothoofd, A. Francken, 2018)

### **6. Deployment (Visualise results)**

Evaluation is the most difficult phase of CRISP/DM method. When you manage to successfully pass that phase, you need to think how to deploy the results so they are accessible for further usage.

This can be another separate project since nowadays there is plenty of different way on how to do this. It can be internal server on which should be taken action for deployment or some other outsourced provider who is using totally different environment and you need to do many more modifications for deploying purposes.

Deployment can be done by a data analyst or the customer. When the data analyst does the deployment, he need to make sure that the enterprise can understand all the activity that should be done for using the model and results properly. A final report and presentation should be carried out to the customer or enterprise.

We will follow all the phases of CRISP-DM methodology in our project step by step. The main reason for following this path in our development is because this methodology is still on top for conducting Data Mining projects.

In one of the statistic conducted by (KDnuggets, 2014) CRISP-DM methodology was used by 43% from 200 people working the industry. Historically this hasn't changed much from year 2007 when same methodology was used by 42% of the people.

## **4 Background study**

Before going to deep on the study it good to have some background study on the main topics. This will help the reader to understand much better the research paper and get enlightened with the terminology used.

### **4.1 Data Analytics**

Statistics is the field which gives a general idea of a collection of data. It calculates necessary numbers to understand the data set.

“Collection, examination, summarization, manipulation, and interpretation of quantitative data to discover its underlying causes, patterns, relationships, and trends.” (Web Finance Inc., 2018)

This field is widely used where ever we have a dataset, it doesn't matter the purpose. It can be numerical statistics, graph statistics, diagram etc.

We are going to do statistical analysis to the dataset given for the project to really understand the dataset but also make meaningful findings out of numbers, diagrams and graphs that KNIME Platform will generate for us.

It is worth mentioning that in Machine learning processes, Data Analysis plays a great role. It helps you understand the given data but also compare predictions and actions taken. Machine Learning is all about statistics.

There are two main types of Data Analytics: descriptive and predictive statistical analysis.

#### **Descriptive data analytics**

This kind of data analysis describes historical data of a company, environment or anything to give an overview of the current status. It can be precise on certain factors of the target or it can give an overall picture. We are going to implement descriptive data analysis in our solution. Of course, you can make conclusion out of descriptive data analysis and come up with ideas on the future trend, but this will not be as accurate as the Predictive data analysis.

Descriptive data analysis is beneficial to companies because it can be used for describing various business aspects and providing a comprehensive view of the overall level of the

company. Important discoveries can be made because the reader has the opportunity to get familiar with the current state of the problem. (Halo Business IntelligenceHalo Business Intelligence, 2018)

### **Predictive data analytics**

The name reveals itself on this type of DA. Predictive analytics are able to forecast the future behaviour of the data and reveal truths of what might happen in the future. Nowadays there is many ways on how humans are able to do these DA using computing power of computers via Software solutions. In business environment this type of DA can make forecast on demand so that the business knows the supply capabilities, if they are able to face the future or not and if not take actions based on predictions. Of course, people might think that this is not 100% sure but nowadays prediction algorithms are becoming complex and smart enough to predict close to 100% truth accuracy.

Each time we are requiring some figures related with the future, we can refer to predictive analysis. What is more, it can help us to fill the missing information. (Halo Business IntelligenceHalo Business Intelligence, 2018)

## **4.2 Machine Learning**

Machine Learning can be interpreted differently by different people. Below you can find few definitions given by different sources in almost the same timeline.

Artificial Intelligence is a huge field with many subfields on it. One of those is Machine learning. So, what is machine learning? It is a science that learns from data and information provided using different algorithms. The algorithms can be improved and changed by themselves, so there is no need for programming. (Marr, 2016)

“Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.” (Daniel Faggella, 2017)

Machine learning is an example of human and animals' behaviours that learn from the experience. This is adapted to computers for future improvements based on past experience. (MathWorks, 2016)



In other words, Machine Learning is the process of training computers with historical data and use the learning for acting in the same way. This varies on how the reader defines the concepts but overall the idea of machine learning is understandable.

The influences that Machine Learning have in technology innovations are enormous starting from simple task to very complicated calculation that even humans can't accomplish with success. Computers have the capacity of storing historical information and easily access them with speed. Such a thing that humans don't have.

### **Timeline of Machine Learning**

For a better understanding of machine learning, it is important to go back in years and get a timeline how it developed from beginning till now.

The history of machine learning dates back in 1943. In this year is modelled the first human neural network.

In 1950, Alan Turing wanted to challenge computers' intelligence, so he created a test, known as "Turing's Test". The test was given to a human and should make the human believe that he was talking with another human not with a computer. (Marr, 2016)

In 1952, we have the first computer learning program done by Arthur Samuel. IBM implemented this learning program to make improvements in moves and get some winning strategies. Later on, those moves were absorbed into IBM's program.

In 1959, the phone calls became clearer due to implementation of adaptive filter. (Marr, 2016)

Later, in 1968, the film director, Stanley Kubrick, described the intelligent computer of 2001 and he wondered if it would really exist. It seems like a relevant assumption. (Marr, 2016)

In 1982, Blade Runner science-fiction movie develop the idea of smart machines giving emotions. In 1985 was the invention of NETtalk that was a self-tough program that learns how to pronounce about 20,000 words per week. (Marr, 2016)

A revolutionary date in Machine Learning history is 1997 when IBM created Deep Blue. Garry Kasparov, the chess expert was defeated by a computer and this was the first time in history. (Marr, 2016)

In 1999, University of Chicago developed CAD Prototype Intelligent Workstation for detecting cancer through mammograms. The accuracy was 52% better than radiologists. (Marr, 2016)

After 2000s, Machine Learning became more popular and got usage in many aspects of life and technology. Still we have many definitions about ML and I would choose some of them. (Marr, 2016)

The timeline above makes us understand how new the topic is and how powerful has become in human life nowadays. Only for about 7 decades machine learning has evolved together with technology rapidly and has influenced Technology considerably. Some will say that machine learning is a modern concept, but the timeline shows that it is not that modern considering that it is about seven decades old.

### **Categorisation of Machine Learning**

“One bit of advice: it is important to view knowledge as sort of a semantic tree—make sure you understand the fundamental principles, ie the trunk and big branches, before you get into the leaves/details or there is nothing for them to hang on to”.—Elon Musk

Machine Learning is divided in three main sub-categories: supervised learning, unsupervised learning and reinforcement learning. There is also semi-supervised learning, that stands between supervised and unsupervised learning.

### **Supervised Learning**

The supervised learning is based on practical machine learning. There we have one input variable (X) and one output variable (Y). We can find the relation between these two variables, we can use algorithm. The mapping function is as shown in the formula below: (Brownlee, 2016)

$$Y = f(X)$$

It is important to imprecise, as good as possible, the mapping function. This way we can always predict the output variable (Y) for the data, when new input variable (X) is added. In this case, the algorithm is learning all the time from the training dataset. It is likely as a teacher supervising the learning process. Constantly, the teacher knows the right answer

and she gives continuous feedback to students. In supervised learning we have predictions on the training data made by the algorithm and a teacher to correct all the time. This process goes on until the algorithm achieves a bearable level of performance. (Brownlee, 2016)

Using computers to find the right patterns on which the supervised machine learning will be based on, makes this process smoother. The reason behind this is the computing ability of computers that human biological brain would never have. Goal of the supervised machine learning is to learn by reusing system which makes computers smarter by time and increase prediction accuracy. Results are accomplished on same manners both from computers and humans at supervised machine learning. (Maini, 2017)

Supervised learning uses classification and regression.

### **Classification:**

In all kind of dataset exists classification. Data with similarities are classified into classes that then are used for making the prediction. This classification can be improved with the supervised learning by inputting more data that will make the accuracy of the prediction higher. This accuracy is affected by the algorithm we choose to classify as well as the way of applying it to the dataset. Since classification requires vast amount of data to be implemented properly, training data on which we work on should be very much useful and accurate. (Maini, 2017)

### **Regression:**

In difference with classification, regression splits the data in training and testing data. The training dataset is labelled so that allows the model to learn from the labelled training examples. Labels also allows tracking so that the model is trained. On the other hand, the test dataset has no labels so the value we would like to predict using this model is unknown. Most important feature of regression is that the model should be able to generalize cases that hasn't faced before to increase the performance on the test data. (Maini, 2017)

### **Unsupervised learning**

In difference from supervised learning, in unsupervised learning the algorithm is provided only with the input variable ( $X$ ). There is no output variable ( $Y$ ), no correct answer and no teacher to supervise the learning process. The algorithms should learn on their own device by discovering rudimentary distributions in the data so that the algorithms can discover more about it. After, they should come up with engrossing structure in the data. This structure should also be presented. (Brownlee, 2016)

Unsupervised learning uses clustering, association problems, K-means etc. Below there is a brief description of clustering.

### **Clustering:**

When the problem to be solved is finding certain groupings in the dataset based on any factor then clustering is the way to proceed. Examples of clustering can be finding groups of buyers based on products they bought. (Brownlee, 2016)

Hierarchical clustering: It is used to create a hierarchy of clusters.

### **Reinforcement learning**

This way of learning the dataset has no answer key but lets the agent to make the learning decisions. There is no training dataset, so agent learns on the way and becomes experienced on the go. While passing the data to be analysed the agent learns and make decisions such as good and bad actions. This brings the agent into a trial and error way of accomplishing predictions so that it can maximize the long-term rewards. (Maini, 2017)

### **4.3 Data mining**

Data mining can be defined as finding models in a set of data that can be used for future predictions. These models can be different depending on problems intended to be solved. They are generated and trained by the available data set. (Leskovec, Rajaraman and Ullman, 2014)

### **Statistical Modeling**

The term “data mining” was firstly use by statisticians to define the process of extracting some information that was not supported by the given data. By this means data mining has gone through evolution and now helps the statistical process. These days statisticians consider this term as the construction of statistical models. It helps on pulling visible data as an underlying distribution. (Leskovec, Rajaraman and Ullman, 2014)

### **Statistical Limits on Data Mining**

Statistical limits are a data mining problem that try to uncover all the abnormal phenomenon occurring to vast amount of data. It also treats the “Bonferroni’s Principal” which is a caution averse to fanatic data mining usage. (Leskovec, Rajaraman and Ullman, 2014)

## **Data Mining tasks**

For a smooth workflow of the project we need to follow some tasks. All the tasks needed are described briefly in the following paragraphs.

### **Pre-processing:**

Data mining pre-processing contains all the steps we need to take in order to start working with the data we have. Differently said, pre-processing includes removing of abnormalities from the data, checking for null values and taking care of those. Moreover, we can use aggregation and generalization for compressing the data. (Shravan, 2017)

### **Outlier analysis:**

Outlier analysis is used for catching all the irregularities that are present in a dataset. This unusual data includes all the eccentric and outlying elements. (Shravan, 2017)

### **Associative analysis:**

In a large dataset it is always difficult to uncover all the relationships of the data items. That is why we have associative analysis to unwrap all these hidden correlations. This is pretty obliging in making accurate predictions for particular items. (Shravan, 2017)

### **Regression:**

Another data mining task is regression. We have a deep explanation in chapter 4.2.2.1, but just to remember, regression is used for making predictions of dependent variables through a model created from independent variables.

### **Summarisation:**

We use summarization for concluding a compress illustration of the complete dataset. An assortment of techniques such as statistics, machine learning and pattern recognition create the definition of data mining. As we can notice from chapter 4.2.2, data mining and machine learning algorithms go hand in hand with each-other. For this reason, there are a lot of crossing points between data mining and machine learning. (Shravan, 2017)

## **Data Mining Tools**

Nowadays there are many data mining tools, but we are going to have a brief introduction to some of them and we will focus on KNIME Analytics Platform because that is the tool we are using for achieving the goal of this thesis.

## **DataMelt**

DataMelt or DMelt as is widely known, is not just a data mining software but a computational platform that provides statistics, numeric, scientific visualisation and symbolic computations. In this paragraph we will touch only the data mining function of DataMelt software. As a data mining tool, DMelt provides numerous features such as clustering, regression, curve fitting, neural networks, cluster analysing, fuzzy algorithms, analytics calculations, interactive visualization through histograms and 2D/3D plots. Everyone can use the IDE of DataMelt and its functions can be requested from the application using Java API. DMelt is available in two versions: community and commercial. Both versions can be accessed from Windows, Mac OS, Linux and Android platforms. jHelpWork and SCaVis programs are the basic of DMelt, which is the better version of those two together. Datamelt is a great software for scientists, engineers and students. (Shravan, 2017)

## **Apache Mahout**

Apache Mahout is a very powerful machine learning tool that finds usage in huge technologic companies such as Twitter, Drupal, Adobe and AOL. Moreover, many academics and researchers have found this machine learning tool quite impactful. But how does Apache Mahout function? Mahout is like an algorithms' library that supports classification, clustering and frequent pattern mining. Also, it has a simple integration with Hadoop in a distribution mode. This is helpful for integration and massive volume of data mining. (Shravan, 2017)

## **ELKI**

ELKI is a licensed open source software under AGPLv3. It is developed in Java and has the central focus on outlier discernment and cluster analysis, where it uses a selection of quite a few algorithms. ELKI is approached through a GUI that manifest the outcome when the chosen algorithm is run. This software has few designing goals that are related with scalability, extensibility, performance, the design used for receiving contribution and completeness. However, this software has not yet a professional support, which gives it limitations in usage. That is why the best utilization of ELKI is in research work and some scientist work. (Shravan, 2017)

## **MOA**

MOA stands for Massive Online Analysis. As we can get from the name, MOA has as the principal function the data stream mining. This is a quite handy software that can deal with massive real-time data streams at a very high speed. MOA can be used through GUI, Java API or even command line, since it has been given out by GNU GPL. This powerful

software is a good alternative for designing real-time application since it includes a collection of machine learning algorithms. A swift computation should be done in the stream mining algorithms because the time is limited and we need to go through the dataset without sorting it. As a solution to this data mining issue, MOA and Weka would be the perfect solutions because both of them support this matter. In spite of that, MOA would be the best fit since it can provide more analysis and mining information for real-time data.

(Shravan, 2017)

### **KEEL**

KEEL stands for Knowledge Extraction for Evolutionary Learning. It is an open source tool that functions on Java. It is an unionized tool supplied by GUI. In KEEL we can both manage and experiment with data. For managing the data with different file formats, there are four functions we can use: import, export, edit and visualize. Statistical libraries and standard data mining is utilised for data experimenting. To sustenance the data experimenting, it is important that a Java Virtual Machine (JVM) is installed in the system. KEEL provides quite many algorithms, and it is pretty clear understanding. That is why academics and researchers try to implement KEEL. We can find an intelligible guide and support at the official web-page: <http://keel.es/>. (Shravan, 2017)

### **Rattle**

Rattle stands for R Analytical Tool To Learn Easily. As we can understand from the name, this data mining tool is developed in R statistical programming language. It has a wide range of features such as clustering, modelling, visualization. Rattle is supported in Windows, Mac OS and Linux. It is a good tool for teaching purposes that is why many American and Australian universities are using this tool. Moreover, commercial enterprises and businesses make use of Rattle. (Shravan, 2017)

### **KNIME Analytical Platform**

KNIME is an affluential open source analytics platform that sustains analytics, integration and reporting functions in one main software. It is a Java developed platform build upon Eclipse. KNIME is available in two versions: free software and commercial version, that make it good for individual and enterprise usage. It can be accessed through a GUI that provides different solutions for creating data flow. On top of that, KNIME manages data pre-processing, collection, analysis, modelling and reporting. Moreover, it is a very flexible platform that provides a smooth integration of Weka and R data mining software. Also, we should mention that KNIME use some extension mechanisms from Eclipse, such as text and image mining. That is why KNIME is on top used data mining platforms. (Shravan, 2017)

We will use KNIME Analytics Platform, which is an open source modular environment of KNIME. This platform gives flexibility for any data analysing, machine learning and data mining. It is open source powerful platform that enables everyone perform accurate prediction.



## 5 Design

The design of this case study will be in accordance with KNIME Workflow. What is a KNIME Workflow?

All the project created in KNIME are created by using the KNIME Workflow. It is the process of connecting and configuring all Nodes needed for receiving the result. It is a workflow because each node has one job to do and the overall task flows through nodes from reading, processing, mining to displaying.

In this way we will be able to understand how the results are concluded.

### 5.1 Understanding the dataset

Before starting the development of the workflow, we need to understand the data given to work with. In our case the dataset given are the answers gathered from the (NIEMIVIRTA, 2002) questionnaire handled to the 2018 BITe students of Haaga-Helia UAS. In total the questionnaire has 30 questions and all of them can be answered numerically only. In the first 6 question there is gathered some basic information about the student such as:

1. Gender
2. Age
3. Nationality
4. Which study semester is the student in?
5. Does the student have already accomplished a higher education?
6. Was BIT program the first choice?

All this information is shown in numerical values in our dataset such as:

1. Gender
  - a. 1 = Female
  - b. 0 = Male
2. Age
  - a. 1 = 18-21
  - b. 2 = 22-25
  - c. 3 = 26-29
  - d. 4 = 30-35
  - e. 5 = >35

3. Nationality

- a. 1 = Finland
- b. 2 = Europe
- c. 3 = South American
- d. 4 = North American
- e. 5 = Asian
- f. 6 = African

4. Study semester defined by the number

5. Does the student have already accomplished a higher education?

- a. 1 = YES
- b. 0 = NO

6. Was BIT program the first choice?

- a. 1 = YES
- b. 0 = NO

The real dataset provided to conduct this study is shown in figure (Figure 2)

Q1	Q2	Q3	Q4	Q5 & Q6
Female	age 18-21	1 Finland	1	Yes
Male	age 22-25	2 Europe	2	No
	age 26-29	3 South American	3	
	age 30-35	4 North American	4	
	age > 35	5 Asian	5	
		6 Africa	6	

Question	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
P2	1	2	5	3	1	1	1	3	7	6	4	7	1	5	7	1
P6	1	2	4	7	1	1	4	6	7	6	7	7	7	6	7	2
P13	0	2	2	3	0	1	4	6	6	7	5	5	6	7	5	4
P40	0	3	6	2	1	0	1	5	7	6	5	7	6	6	7	2
P44	0	2	2	1	0	1	4	6	7	7	6	7	3	5	7	3
P52	0	1	2	1	0	1	1	5	3	5	4	7	5	4	4	1
P53	1	3	1	1	0	1	6	7	7	7	7	7	6	6	6	3
P55	1	3	2	1	0	1	6	6	7	5	5	3	3	5	6	2
P56	1	4	2	1	1	1	6	5	6	6	4	7	3	5	4	1
P64	0	4	1	1	0	1	4	4	4	4	4	7	7	5	7	3
P66	1	2	1	1	0	1	2	6	7	7	6	5	5	6	6	5
P70	0	4	2	1	0	1	4	5	7	5	4	6	0	5	4	1
P75	0	2	5	1	0	0	7	3	7	4	7	7	7	6	1	1
P77	1	3	3	1	1	1	3	3	7	7	7	2	2	3	5	2
P78	1	1	4	1	0	1	2	4	7	4	3	7	4	4	5	1
P79	1	1	5	1	1	1	3	6	6	5	6	5	5	5	6	3
P80	1	2	1	1	1	1	2	1	7	2	6	7	2	7	4	1

Figure 2 Raw Dataset

The rest of the columns are rating based questions. Students are rating 24 questions from 1 to 7 or 1= Not true at all to 7= Very true.

In relation to (NIEMIVIRTA, 2002) study all these 24 rating questions are grouped accordingly in 8 main scale which in reality are 8 main study factors influencing study performance.

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.

Figure 3 Niemivirta Scaling Questionnaire

There are about 100 rows collected for this dataset. After implementing the project, we will see if this is enough rows to work with. At this point the reader can understand the dataset given.

## 6 Implementation

### 6.1 KNIME Workflow Development

This thesis is narrowed into a certain group of students that are Business Information Technology (BITE) students of Haaga-Helia University of Applied Sciences in Helsinki Finland. The study, statistics and predictions handled in this work are focused on this group as well as giving the opportunity of generalisation.

The goal is to provide an interactive user-friendly statistical view on which a user can see current situation of students' behaviour as well as Fear of Failure predictions.

In details we will explain how we can create a KNIME workflow. Step by step this paper will show how the research is done and how the prediction is handled.

Even though the paper will be narrowed to the pre-defined dataset provided from the above-mentioned group of people, this paper will give general understanding on the power of Machine Learning and Data Analytics. With a little more research, the reader will be able to develop own KNIME project on his/her own requirements.

## 6.2 Data understanding

The first step of this process is to know in which area are we working in and what kind of supply information is available for the process. Relating to our case this section will clarify the reader in general the start points of the project by answering the following questions:

- What problem should we solve exactly? What are the expectations?

The main problems we are solving in this paper are:

- a) How (NIEMIVIRTA, 2002) 8 study factors relate to each other (in the given dataset)?
- b) What happens to Fear of Failure if we increase/lower the coefficient for another factor?
- c) By visualizing the pre-processed dataset, find hidden results for students' performance.

The hypothesis:

- a) Visualise in the best way the dataset so that it is understandable
  - b) User can upload another dataset collected for one or many persons and interact with the results
- How should the solutions look like?

There will be dynamic data visualization such as graphs, tables, Scatterplots etc. The solution will allow the user to interact with data as well as test his/her own dataset in the system.

- What are the hypothesis from (NIEMIVIRTA, 2002) study about the problem?

We know that the 8 factors influence accordingly to the overall study performance. Logically we can conclude some finding from the first observation of the (NIEMIVIRTA, 2002) 8 factors results:

- a) Mastery-intrinsic orientation > should be high
- b) Mastery-extrinsic orientation > should be high
- c) Performance-approach orientation > should be high
- d) Performance-avoidance orientation > should be low

- e) Avoidance orientation > should be low
- f) Fear of Failure > should be low
- g) Academic withdrawal > should be low
- h) School value > should be high

We also know the corresponding questions for each factor shown in Table3.

### 6.3 Data preparation

#### Read Data

Every KNIME workflow must start with some dataset, which can be provided by a server or a simple dataset file such as excel file, csv file or any other format. The software offers quite many nodes which support reading any type of file as well as able to connect to databases. In this project we will be using the 'File Reader' node to read our dataset provided.

We have converted the file to .csv (comma separated values) for an easier reading purpose. After locating the file in the local machine and configure the node to read headers and values properly we end up having the dataset part of the workflow.

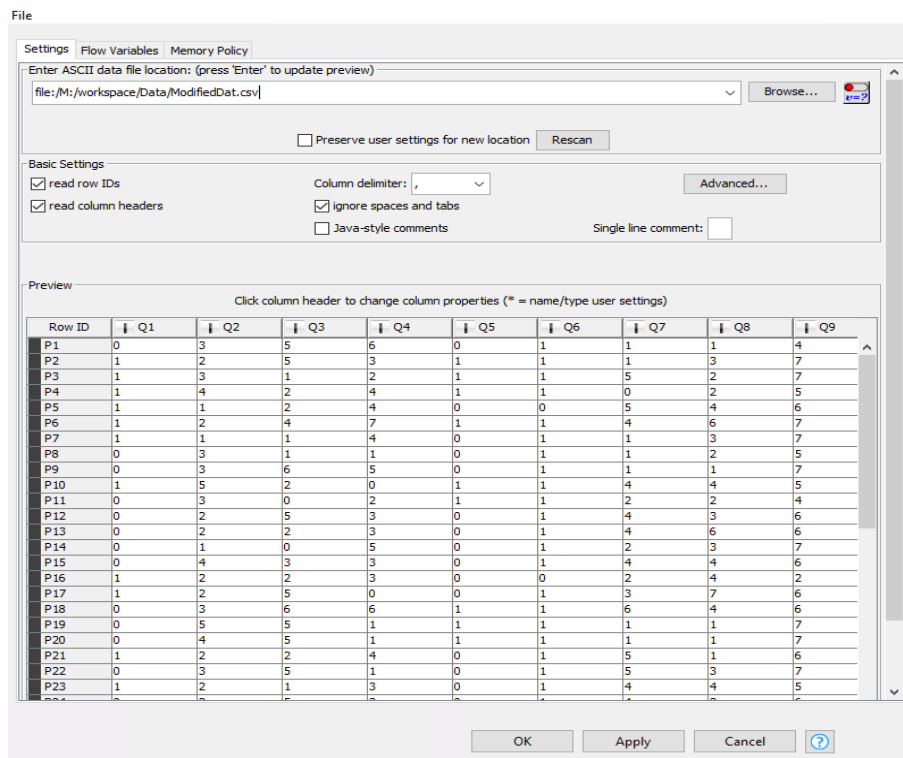


Figure 4 File Reader Node configuration

As we can see on the figure (Figure 4), the data is quite encrypted with numerical values and it is non-understandable and hard to work with. The next steps will show how to give meaning to these numerical values.

### **Construct Data**

Most of data collections are done in numeric values for easier export from any platform in which the data are collected. There is always data construction taken place to work with the dataset. In the 4.1 Data Understanding chapter dataset value meaning is explained in detail. We will use Nodes provided by KNIME to reach the goal.

List of nodes needed to construct our dataset are:

- 1) Math Formula (used to conduct data calculations throughout individual lines)
- 2) Column filter (used to filter unnecessary columns)
- 3) Column Rename (used to give meaning to the column name)
- 4) Numeric Binner (used to binn numeric values into meaningful text using value boundaries)
- 5) Rule based Row-Filter (to remove rows with null values for better data accuracy)

Every single node has its own purpose in this workflow. Below is an overall picture of the data construction flow.

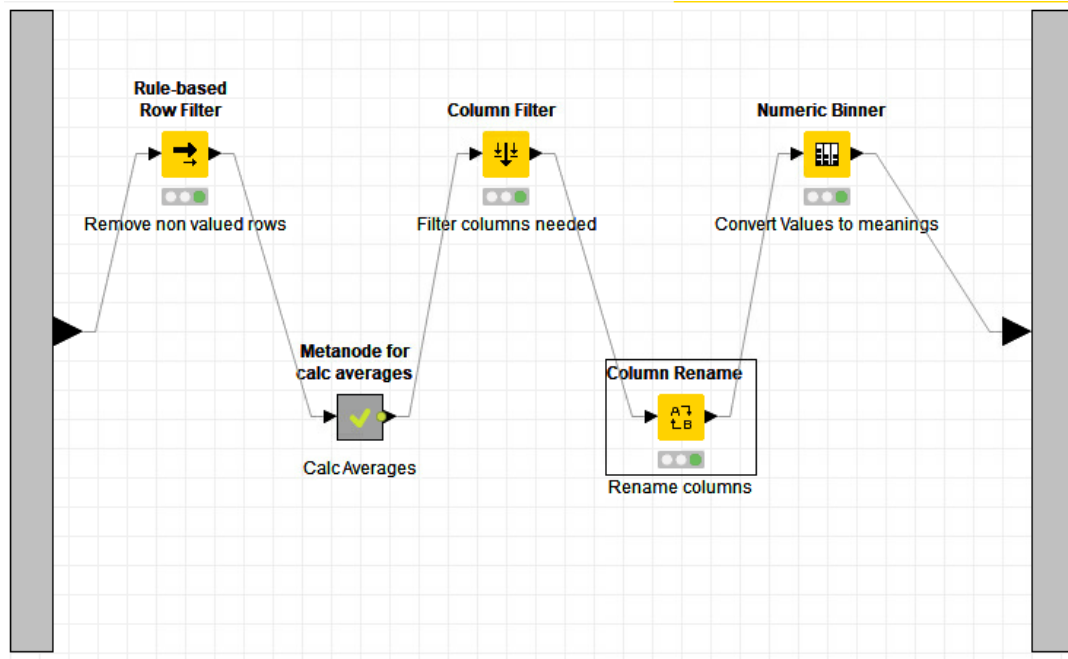


Figure 5 Data Manipulating Metanode

The end goal of this flow is to have a much more understandable dataset table with which we will be working to achieve the bigger goals of this topic. This process is the most important and it must be accurate enough to trust on building a whole project out of it. The results will be our key resource in conducting data analytics as well as model training for our machine learning prediction models' purposes.

Let's dive deep on the process to understand more what every single step is helping data construction flow:

### Remove non-valued rows by implementing rules

Part of the data construction process is also to make sure that the data we will be working with, is as well-aimed as possible. We need to make sure that the calculation of results are accurate ones. Rows with empty values can mislead the calculations and provide wrong results to the calculation. The best practice is to remove the rows with empty or null values. This can be done with Rule-Based Row filter. We configure this node by assigning some rules for removing rows with empty values.

The Rule-Based Row filter peers the user-defined rules with one row in the input table. The peering can have a TRUE or FALSE outcome. Presuming that we have a TRUE outcome, then the row is chosen for inclusion. Or else, if the outcome is FALSE, the outcome would be excluded. At the point we have no peering between the rules and the row, then

the row would be excluded. Both, inclusion and exclusion can be inverted. (KNIME User+Developer, 2018)

After this step we will be doing average calculations from the values provided. To make sure that the averages are calculated properly we have to remove the rows where the value on the columns needed for averages that have value = 0. If the value of the column is 0 then the average will be wrongly calculated and will be misleading us to wrong results.

In figure (figure 6) is displayed how to configure the Rule-Based row filter node according to our needs.

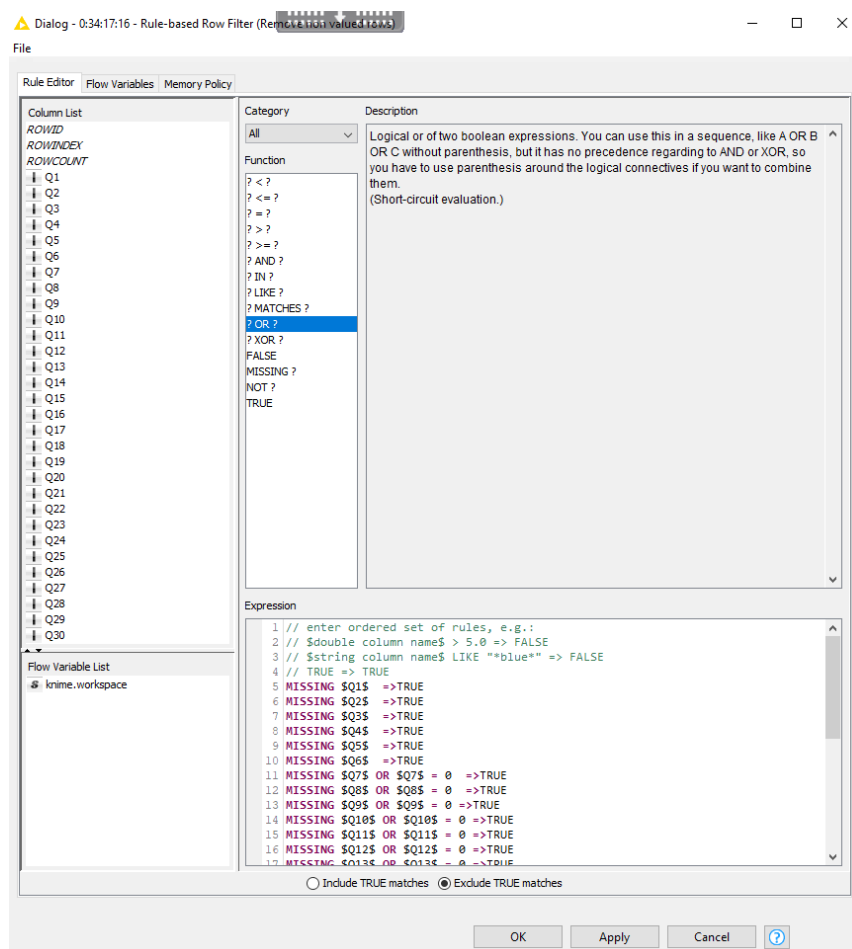


Figure 6 Rule based row filter node configurations

### Metanode for calculating the averages.

Before on the earlier chapters we mentioned that we will be using averages of the 8 factors of (NIEMIVIRTA, 2002) study as well as the table shows which questions corresponds to each factor. In the given dataset we have columns named by question number. This will help us to calculate the average results for each factor. The calculations are made inside a Metanode for a better project flow management and easy to understand



data construction. A metanode in KNIME Software is a group of nodes which get input and give the processed output.

Our Metanode for calculating averages are using several Math Formula nodes to make computation for every single one of eight factors we are interested in studying in this thesis topic. From inside this metanode looks as follows:

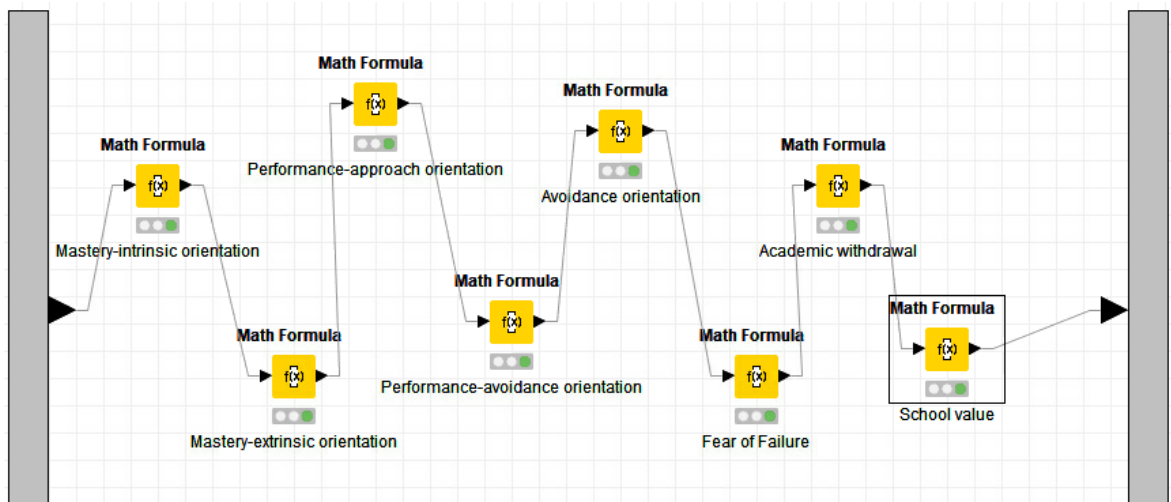


Figure 7 Metanode for calculating averages

First let's understand what a Math formula node is in KNIME environment and what it does.

This node implements mathematical calculations to desired values in the dataset row. After performing, computed results can be displayed either as a certain column replacement or as totally new one. It come with many built in mathematical functions that help the calculating process. Some of these functions are: pi factor, total number of rows in the table as well as many more column-based constants. (KNIME User+Developer, 2018)

This node uses JEP, the Java Math Expression Parser. Configuration example about this node can be found on the figure (Figure 8) which shows how we calculate the 'Fear of failure' factor average and add it to a new column named as such:

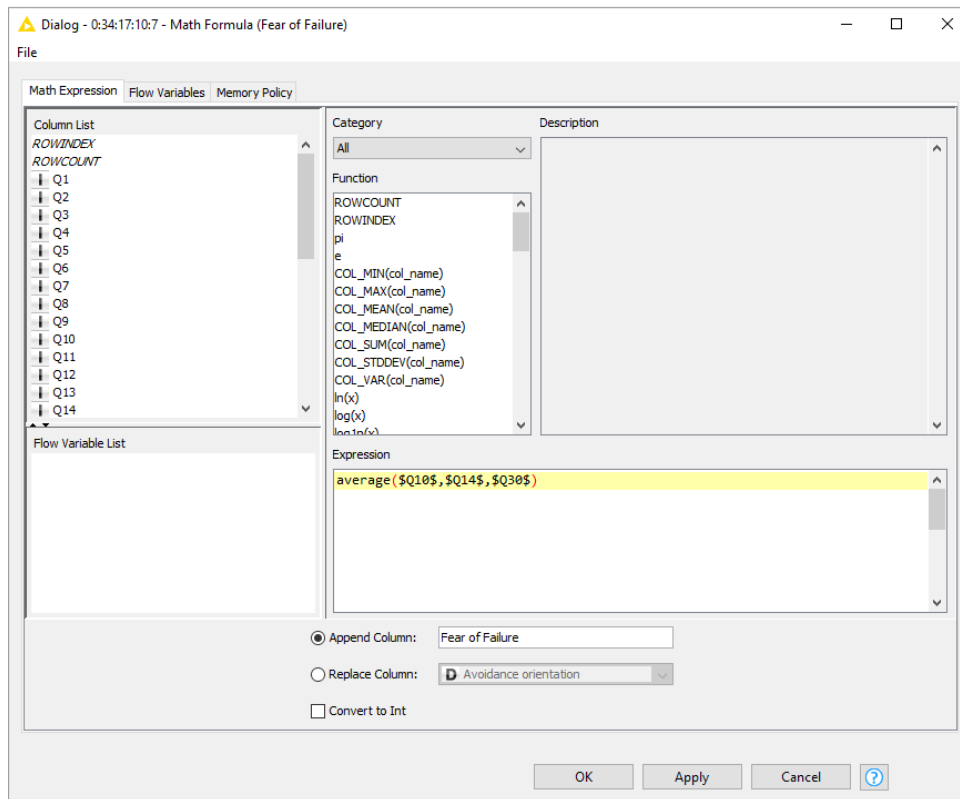


Figure 8 Math Formula node (Fear of Failure)

According to the appendix (appendix 1) attached with the dataset provided we know that numeric values of Q10, Q14 and Q30 columns correspond to this factor. Using this node, we calculate the average of these three columns for each line (record) and display it to a newly created column with a much more descriptive name as 'Fear of Failure' that will help us for all analytics and prediction requirements.

We are using the same way to calculate all 8 factors independently and generate separate columns accordingly. All these custom columns will be added to the dataset.

### Filter the unnecessary columns

In the first step we added 8 more columns to the dataset by calculating the averages for every single factor. In this case we need to filter the columns from which we got the average results from. This filtering will clean the dataset for the next step. To achieve this KNIME offers the Column filter node which allows us to decide which columns to remove from the listing and which ones to display in the output. This node is straightforward and easy to configure as in figure (Figure 9).

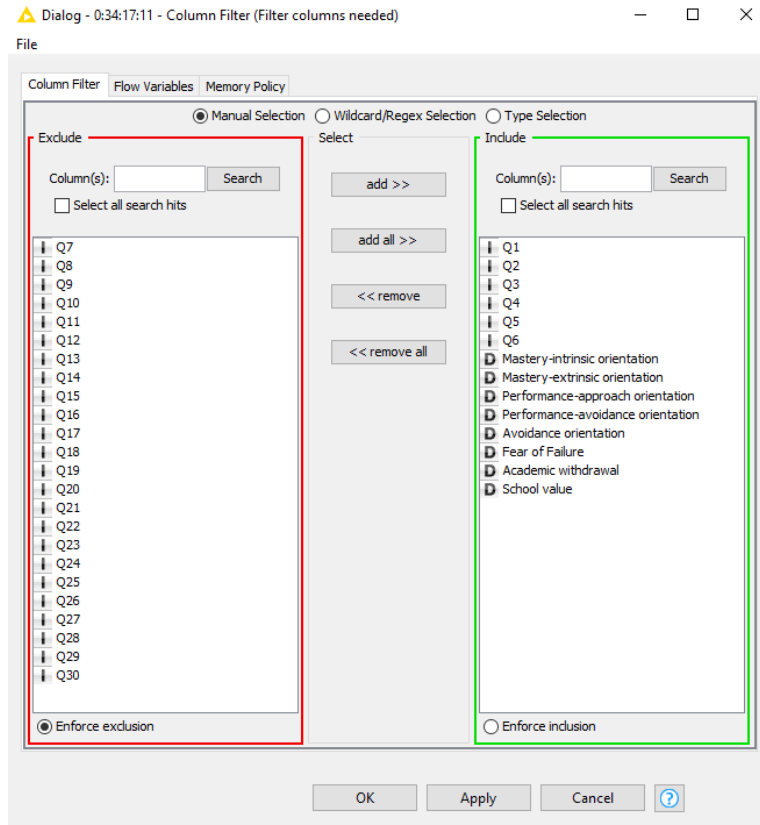


Figure 9 Column Filter Node configurations

By using this node, the user can filter part of the data set columns. This node is easy to be configured since it has left part which shows the excluded columns and right part the included columns. After performing all the right part selected will be shown as part of the table which will exclude the left column ones. (KNIME User+Developer, 2018)

### **Rename the generic columns accordingly**

After filtering the columns, we can see that some of the columns are not named accordingly and they have no meaning for us to understand the values given for those columns. In this case we use Column rename node to rename this column to a more understandable way. Again, for accomplishing this we refer to the data explanation appendix (appendix 1) provided with the dataset. Configurations of the node are shown in figure (Figure 10)

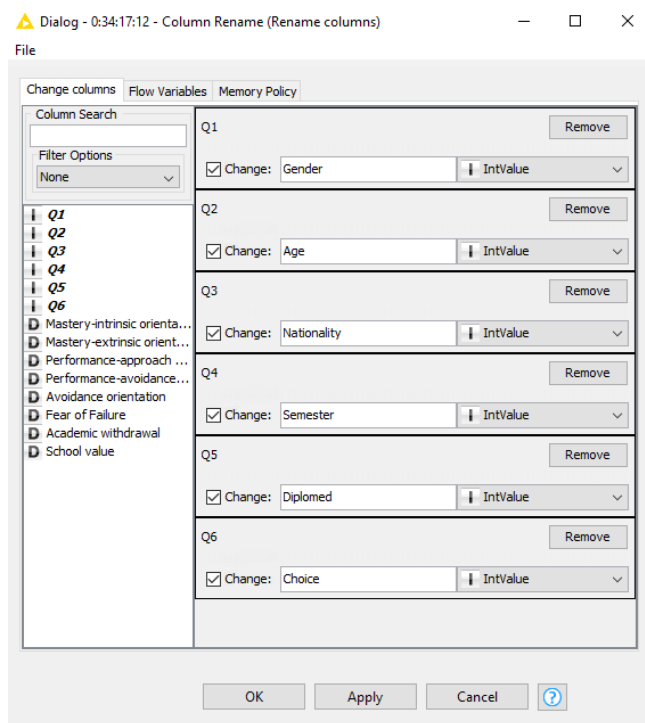


Figure 10 Column Rename node configurations

Column rename node performs the renaming of multiple columns names as well as their type. The type can be changed only if the cells of the selected column can be either safely cast or transformed to. There can be multiple renaming as well as type transformations in one single node. If the line of desired changed is marked in red it means that the type transformation is not allowed due to the values of the cells of that particular columns cannot be transformed for the selected type. (KNIME User+Developer, 2018)

## Data Visualization Module

In this Module we will go through KNIME possibilities of visualizing our dataset on best understandable way. Main purpose of this project is to analyse 8 factors that influence students' performance and visualize them in an easy to understand way. There are many possibilities on how we can achieve this goal.

Before moving to the actual process of data visualization we need to make sure that our data is ready for these analytics. Current stage of our dataset is that every value has a meaning and we can read clearly through them as shown in the figure (Figure 11).

File HiLite Navigation View

Table "default" - Rows: 30 Spec - Columns: 14 Properties Flow Variables

Row ID	Gender	Age	Nationa...	Semester	Diplomed	Choice	D Master...	D Master...	D Perfor...	D Perfor...	D Avoida...	D Fear of...	D Acade...	D School ...
P1	Male	Age 26-29	Asian	6	Not Diplomed	First Choice	7	4	3	3.333	6	1	1	3
P2	Female	Age 22-25	Asian	3	Diplomed	First Choice	7	7	4.333	4	1.333	6	3.667	1.333
P3	Female	Age 26-29	Finland	2	Diplomed	First Choice	6.667	6.667	5.667	2.667	2	3.333	2.333	1.333
P5	Female	Age 18-21	Europe	4	Not Diplomed	Not First Ch...	7	5.667	4.333	2.333	3	2.333	2.667	2
P6	Female	Age 22-25	North Ameri...	7	Diplomed	First Choice	7	5.667	5	6	4	6	4	1.667
P7	Female	Age 18-21	Finland	4	Not Diplomed	First Choice	5.667	6.667	2.667	1.667	2	1.667	2.333	1
P8	Male	Age 26-29	Finland	1	Not Diplomed	First Choice	6.667	5.667	3.333	2.667	3	4.333	3	3.333

Figure 11 Raw constructed data table after manipulations are implemented

To begin with we must define what exactly do we need to visualize and how. First there will be an overall data visualization where we can see all available factors including the person information such as gender, age, nationality etc. as well performance factors such as fear of failure, school value, academic withdrawal etc.

Let's have a look at the possibilities that KNIME Platform provides and decide on which fits our need in this situation.

On figure (figure X) is the list of default ways of data visualizations nodes provided. These are not interactive but static visualizations. With static visualization we understand that user is not able to interact with data in real time. There must be changes on the node configurations so that changes are applied. The list has self-explanatory names as well as icons, so it is easy to understand.

But this is not the only options. KNIME labs have developed awesome JavaScript views for visualizing interactive charts and graphs. On figure (Figure 12) is the list of available interactive data visualization nodes.

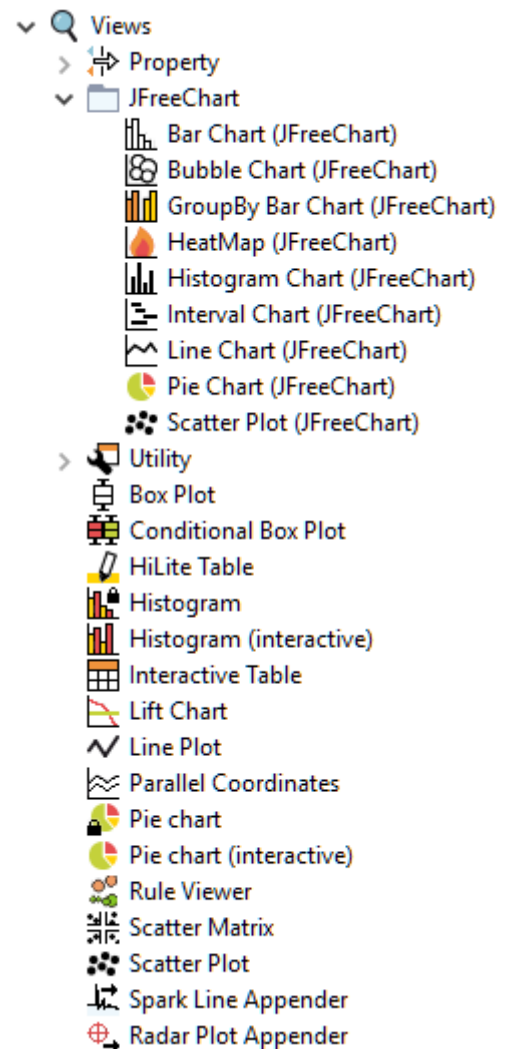
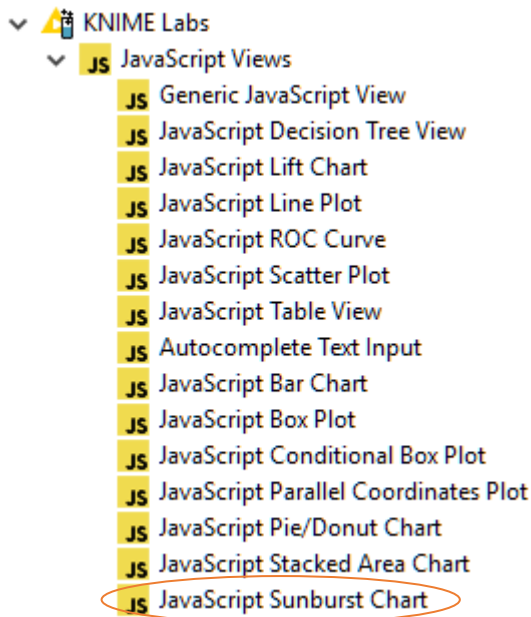


Figure 12 List of Knime default views

## Interactive JavaScript Sunburst chart



For achieving our goal on visualizing whole dataset give as understandable as possible we decided to go with the JavaScript Sunburst Chart (Figure 14).

Sunburst Chart is a type of visualisation that is built in data hierarchy. Hierarchy levels are being visualised as rings with the root in the centre and the rest of the hierarchy moving outwards from it. These series of rings are sliced for each category of that hierarchy level. (Severino Ribecca, no date)

Figure 13 JavaScript Interactive views

This complex chart is autogenerated by the JavaScript node. In our case it creates a hierarchy of all the columns starting from gender all the way to avoidance orientation.

Overall Student Data Sunburst Chart

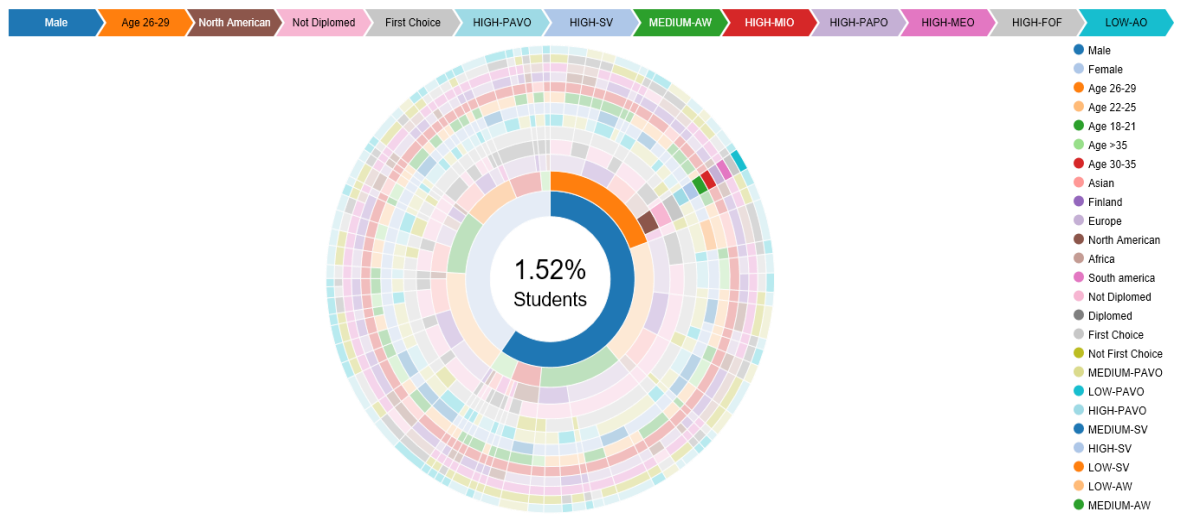
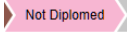
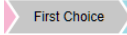
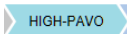
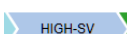
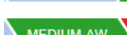


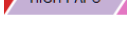
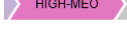
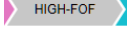


Figure 14 Interactive Sunburst chart with overall data example

The line on top displays the hierarchy details of the selected data. It uses the values provided for each column. Let's start by reading the sample selected data in the figure (Figure 13):

- 1.52 % of students are

- Gender: Male
- Age: 26-29
- Nationality: North America

-  o Graduated before: No
-  o Was BITE as first choice: Yes
-  o Performance-approach orientation: high (average more than 5)
-  o School value: medium (average between 3 and 5)
-  o Academic withdrawal: medium (average between 3 and 5)
-  o Mastery-intrinsic orientation: high (average more than 5)
-  o Performance-approach orientation: high (average more than 5)
-  o Mastery-extrinsic orientation: high (average more than 5)
-  o Fear of Failure: high (average more than 5)
-  o Avoidance Orientation: high (average more than 5)

Since the number of columns to be calculated analysed is 13, Sunburst chart looks quite complex to understand and it gives very precise details. From the general observation we can't conclude beside the fact all the data are visualized. After you start hovering on top of the graph it starts to make sense. We will leave the reader to use this interactive chart and make own interpretation of the data.

The flexibility of data visualization we can conduct with is interactive JavaScript view is considerable. In another part of this data visualisation module we combined other interactive features provided by KNIME Analytic Platform. In this part there are combined 3 different interactive visualization nodes Range Slider Filter Definition, JavaScript Sunburst Chart and JavaScript Table View.

All these nodes are gathered in a wrapped metanode so that they can displayed and interact with each in a single view. After we configured all nodes that we want to be viewed in a single page we wrap them into a wrapped metanote. This metanode generates a list of JavaScript views of the selected nodes into a single HTML page. We have used about four of these wrapped metanodes to separate different data visualization possibilities in different pages. Main KNIME workflow looks like in figure (Figure 15). Everything is gathered in separate metanodes and wrapped metanodes. The difference between these two is that a simple metanode just gathers selected nodes while wrapped metanodes are able to generate page views out of the view nodes inside. In the beginning of this chapter we explained Interactive Sunburst Chart with overall information wrapped metanode, while now we will talk in detail how did we managed to generate a single page view with a list of interactive data visualization nodes. A print preview of this single page is attached to this thesis (Appendix 1).

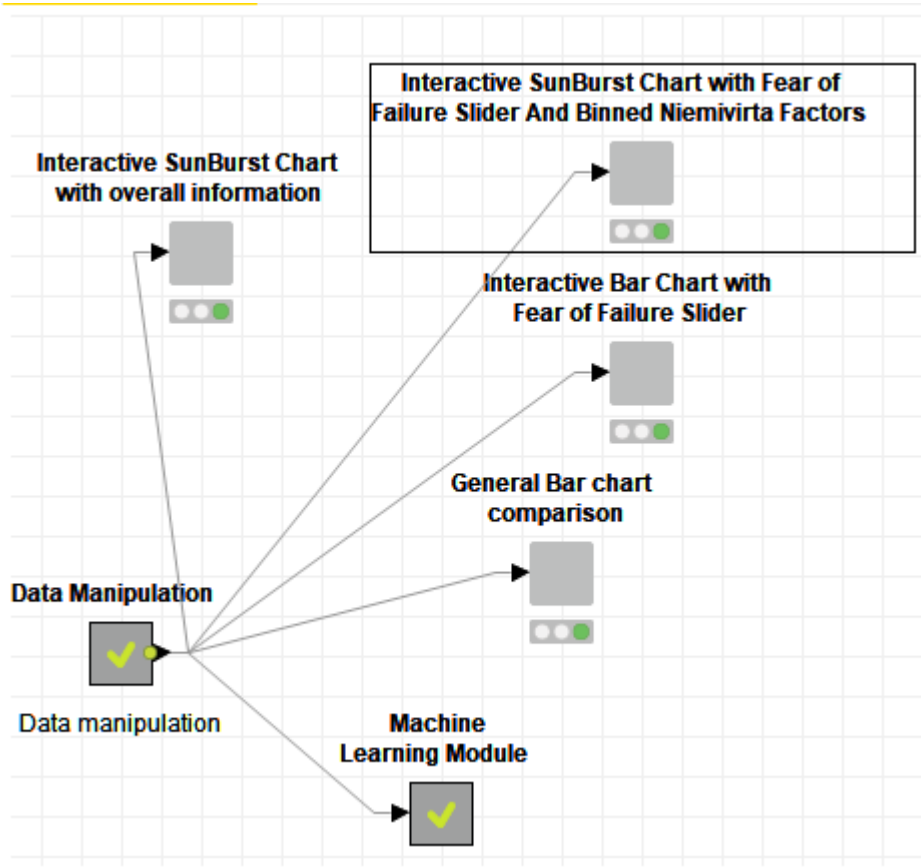


Figure 15 Main KNIME workflow

Inside the selected wrapped metanode there are three main interactive ones with data nodes as shown in figure (Figure 16)

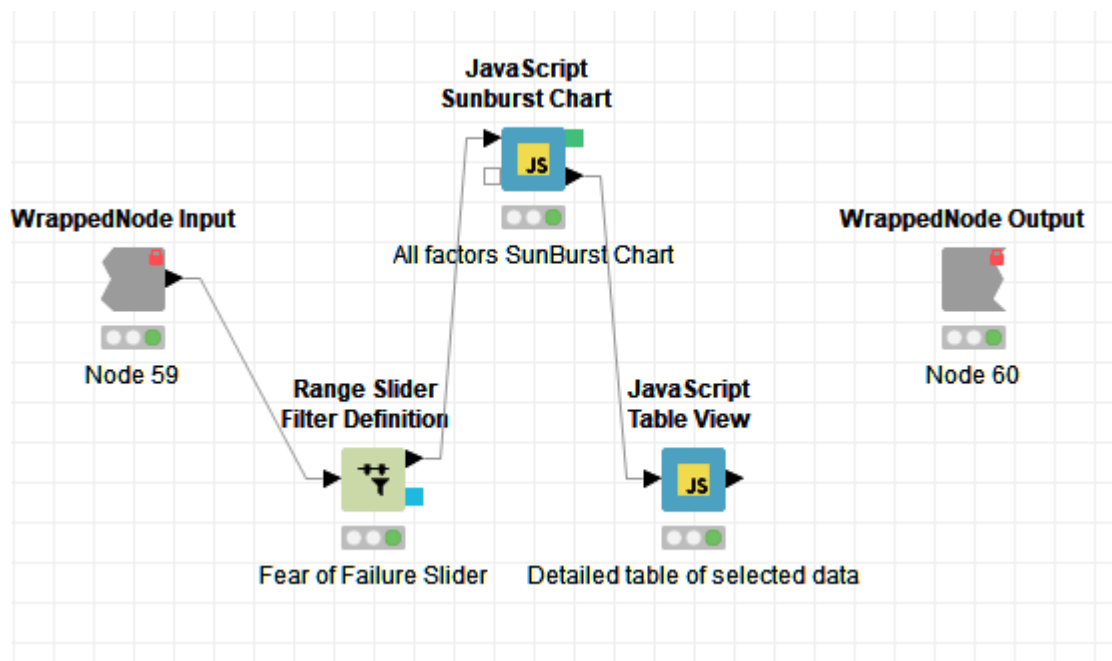


Figure 16 Combination of three interactive nodes



We want to give the user the possibility to filter the data according to Fear of Failure factor. After completing data manipulation in our constructed dataset, we currently have a column name Fear of Failure which has double values calculates as averages of the corresponding questions answers.

Range Slider Filter is a node that can perform filtering of the data set rows for the desired range of values of a certain column. This filtering is visualized as a slider and triggers filtering interactively for visualization purposes. (KNIME User+Developer, 2018)

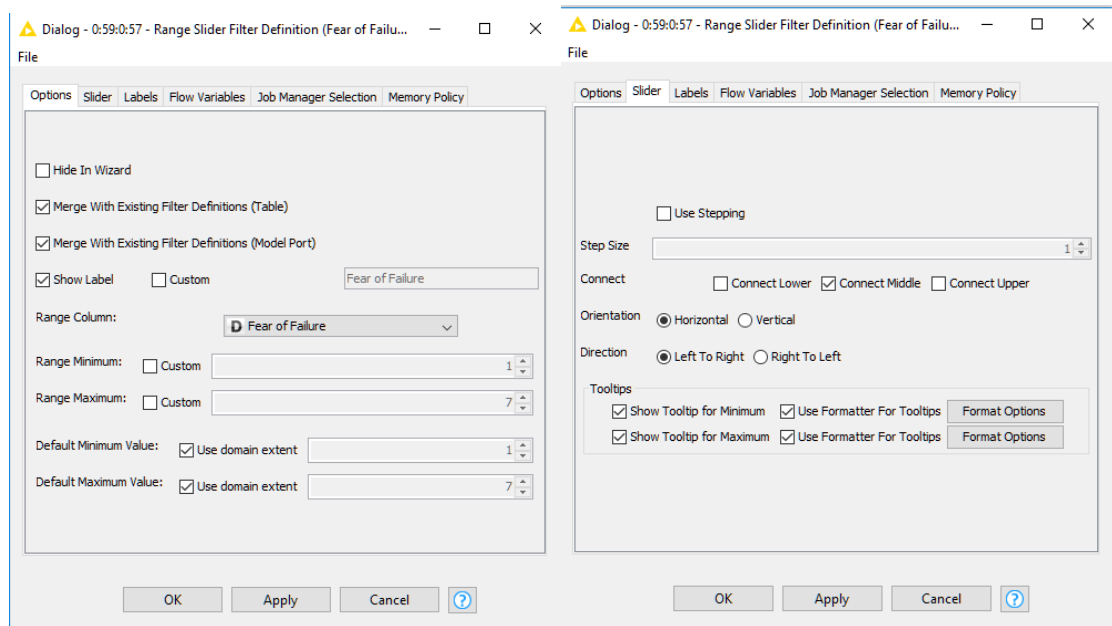


Figure 17 Range Slider Filter Definition configuration

In figure (Figure 17) are shown the configurations of this node. First, we need to define on which column we want to implement filtering and change make changes for the cosmetics of the slider. With these configurations the slider looks as in figure (Figure 18).

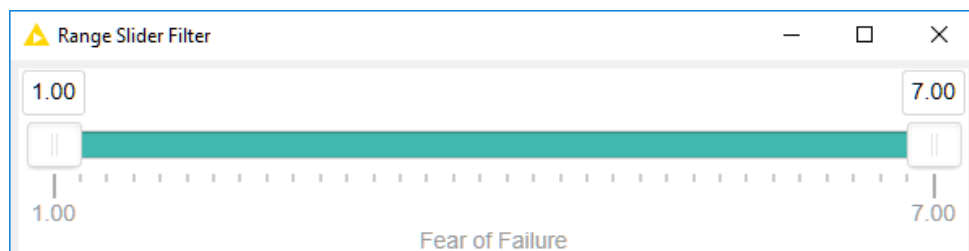


Figure 18 Interactive slider

After finished with the slider we move on by connecting it with the Sunburst chart. In this chart we want to create a hierarchy of the 8 Niemivirta factors. Configuration of this node is shown in figure (Figure 19).

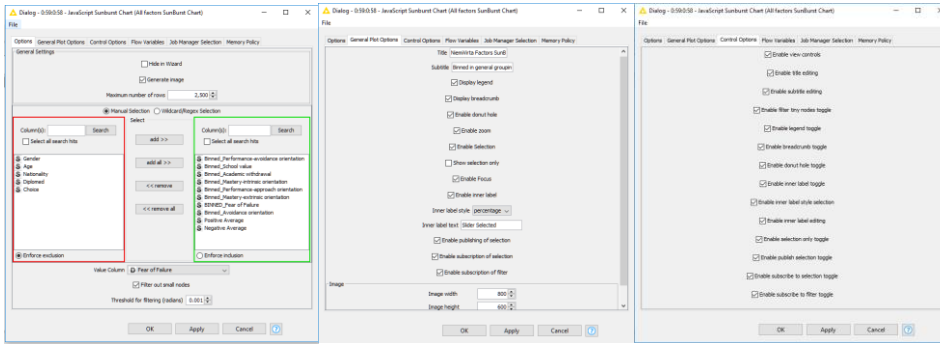


Figure 19 Sunburst chart configuration

First step for configuring this is to select columns on which to build the hierarchy for displaying the data. In our case we have selected columns with the binned values for the 8 Niemivirta factors. In the value column Fear of failure numeric column is selected to be as our value column.

On the next tab 'General Plot Options' visuals can be configured such as Title of the chart and which interactive features should be allowed. In the inner label style, we decide to use percentage instead of sum because the sum of the rows is not that descriptive for our visualisation.

In the 'Control Options' tab we can allow the user to interact with the chart by enabling the control options. In our case we have allowed all control possibilities, so it provides good variety of chart configuration options.

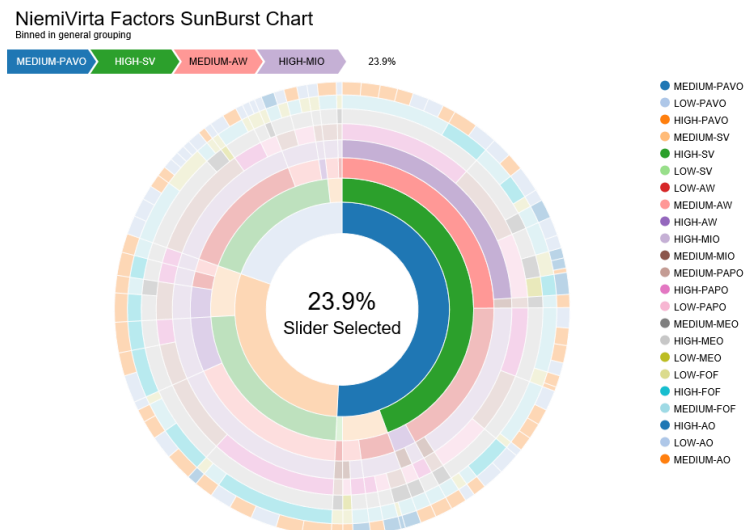


Figure 20 Niemivirta factors Sunburst chart

The results of the node we configured earlier are displayed in the figure (Figure 20). There has been selected some data to show that the chart is interactive as we configured it to be.

Data analytics can be conducted on this chart by playing around with the interactivity on the chart

The final step for completing our page view with all three nodes is adding and configuring the JavaScript Table View. The purpose of this node is to display the detailed values of the selected dataset from the Sunburst chart. Since this is a JavaScript Table view, it allows us to interact with this table itself as well.

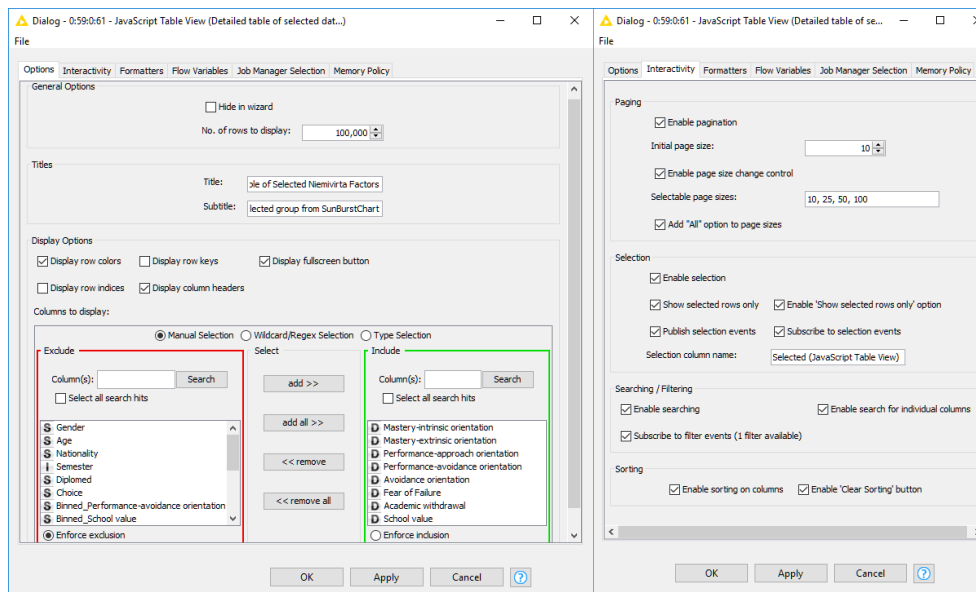


Figure 21 JavaScript table View configurations

As always, the first step of a node configuration is selecting the columns on which the view will be built on. In this case we select the columns which have actual factor average calculated for each factor. In total there will be 8 columns to be displayed. At this tab the title and subtitle of the table can be written as well as the number of rows to display as well as some simple display options.

In the 'Interactivity' tab we will configure all interactivity option for the reader. Important to notice at this point is that we need to have the 'Show selected rows only' active so that the table will display only selected rows.

In figure (Figure 22) is displayed an example of Sunburst chart data selection and Table auto populated with the selected dataset. In the chart we have no values but only binned values which means that the chart find the values for the selected data and displays them accordingly. This gives a deeper data transparency by interacting while making analytics.



### Table of Selected Niemivirta Factors

More detailed information on the selected group from SunBurstChart

Search:

<input type="checkbox"/>		Mastery-intrinsic orientation	Mastery-extrinsic orientation	Performance-approach orientation	Performance-avoidance orientation	Avoidance orientation	Fear of Failure	Academic withdrawal	School value
<input checked="" type="checkbox"/>	■	7.00	6.67	2.33	3.00	2.33	5.00	2.00	1.00
<input checked="" type="checkbox"/>	■	7.00	6.67	2.67	3.67	2.33	5.00	1.00	1.00

Showing 1 to 2 of 2 entries (filtered from 90 total entries)

Figure 22 JavaScript table viewing select data from Sunburst chart

To sum up this whole page is now ready to be used. In this page are three different nodes influencing each other's data availability. The flow goes so that first you decide to filter by Fear of Failure factor using the slider which will interactively filter the data for displaying in the chart. In the chart you can explore by hovering on top of it but then if you desire to see the real raw numbers use the selector for selecting sections of the chart (example shown in (Figure 22) and marked in red narrow). After selecting the section JavaScript table will be auto filled with the selected data so you can dive deep in actual numbers.

### Interactive bar charts

In our main KNIME workflow there are at least two wrapped metanodes which have variety amount of Interactive bar charts. In this part we will show how these bar charts are configured as well as what valuable findings came out of them.

What is a bar chart?

A bar chart is a graph with rectangular bars. The graph usually compares different categories. Although the graphs can be plotted vertically (bars standing up) or horizontally (bars laying flat from left to right), the most usual type of bar graph is vertical.

“The horizontal (x) axis represents the categories; The vertical (y) axis represents a value for those categories.” (To, 2018)

In one of the wrapped metanodes named 'Interactive Bar Charts' we have grouped 4 main JavaScript Bar charts to be displayed in a single page view as shown in figure (Figure 23)

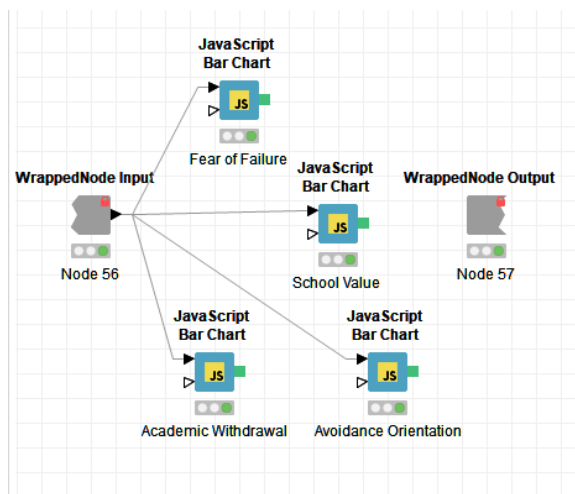


Figure 23 Interactive Bar Charts wrapped metanode collection

As we can see we don't do anymore any data manipulation since all of them were done in the beginning of the process.

4 Niemivirta factors are chosen to be compared with the rest 4 other factors. The goal of this data visualization is to find out how these factors influences each other. By interpreting on right way very good findings can be revealed by studying these specific bar charts.

At this part there will be explained one of these 4 bar charts how it is configured and what is the purpose and meaning of it. Configuration windows are shown in figure (Figure 25)

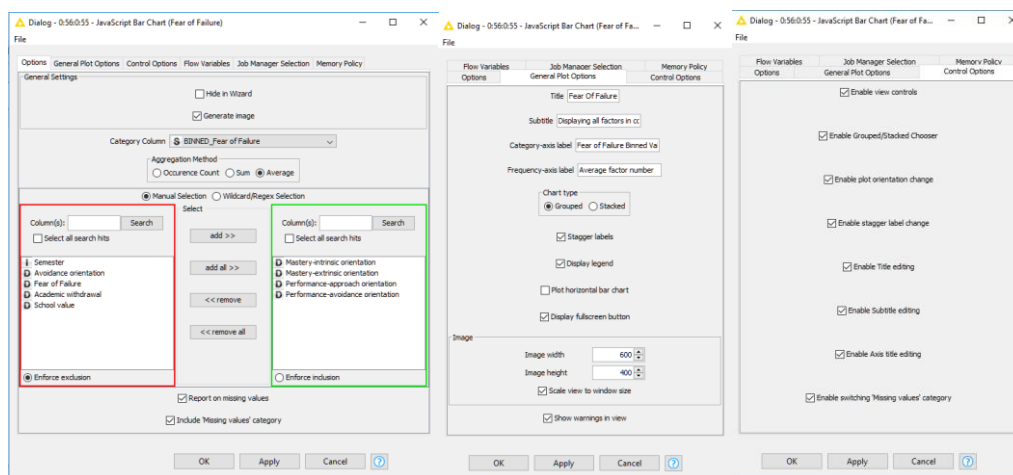


Figure 24 Javascript Bar Chart configuration

To configure this bar chart, we start by choosing our main category column. We have binned the values of Fear of Failure into three main bins low, medium and high. These are going to be our categories. Then we select the columns on which the categories will be implemented in. Because we want to work with numbers now we select the columns which have only numeric values. If we select Average on the Aggregation Method, node will to the following: for each category of fear of failure this node will calculate the average of all data for the corresponding column.

In the General Plot Option tab, we can configure the cosmetics of the bar chart such as Title and Subtitle as well as labels for the axes. It is worth mentioning here that we select Grouped as Chart type because in this way we will understand more the differences on each factor. The control rights for the chart interactivity are set in the Control Options tab.

The results of this JavaScript Bar Chart are shown in figure (Figure 25)

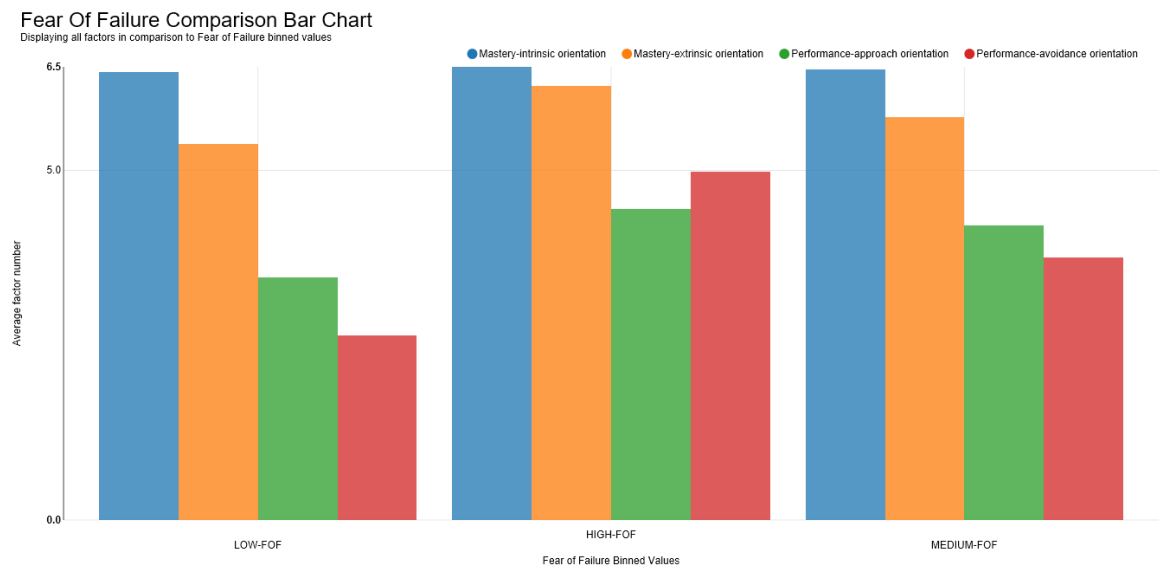


Figure 25 Fear of Failure Comparison Bar Chart

From the first quick observation of this generated chart we can see that performance-avoidance orientation factor is influenced the most by fear of failure. If there is high fear of failure, then there is high performance avoidance as well as for the other levels it changes in the same way.

How can interpret these results from the bar chart?

We know from (NIEMIVIRTA, 2002) that fear of failure factor is the one which shows that student is afraid to fail in tests and classes. On the other hand, performance avoidance related to the ability of avoiding failing situations. Keeping the above-mentioned statements, we can say that student who are afraid in failings try to avoid situations where they can fail. As well students who are not afraid in failing are more daring and don't avoid situations where the might fail.

Many more interpretations can be implemented at this graph, as well as all the other ones generated in this project.

## 6.4 Modeling module

In this module we will be talking about building models from the data provided using the features of the KNIME Analytics Platform. Modeling on other words is implementing machine learning to the dataset. This platform provides us many possibilities on modeling as well. In KNIME modeling can be performed by the learner nodes. These nodes are the ones which we train for creating models on which we make predictions and calculations on.

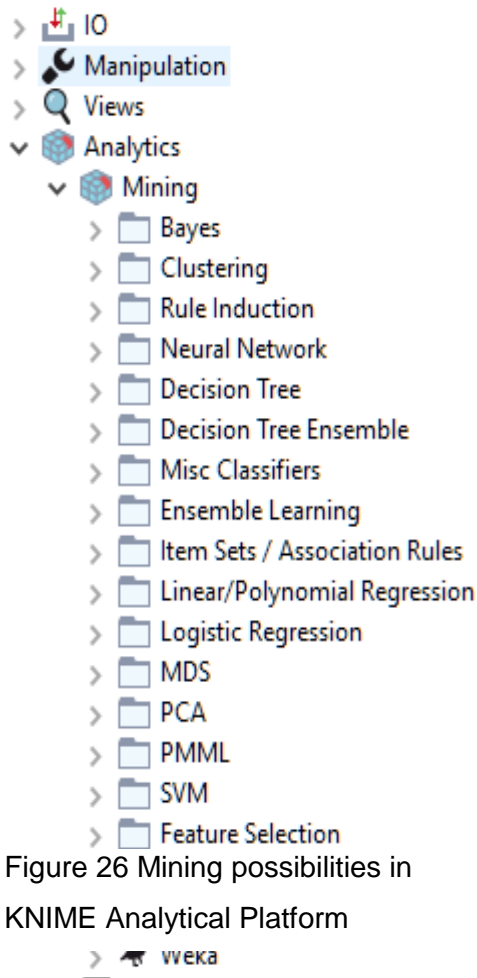


Figure 26 Mining possibilities in KNIME Analytical Platform

As shown in the figure (Figure 26) KNIME has many options on conducting modeling phase. For our case we tried many on the testing iteration but decided to go with Decision Tree and Tree Ensemble because they made highest accurate predictions in our most wanted factor 'fear of failure'

The factor on which the prediction is going to be made is 'Fear of Failure' as stated beforehand. This factor is assumed to be affecting most of the students' performance and through predictions by implementing machine learning we will try to prove this assumption. In the same time, we will use the model created to predict fear of failure if other data is inserted in the platform from another source.

In KNIME environment there are some steps to be followed to be able to train proper models. We have followed these steps in our project as shown in figure (Figure 27)

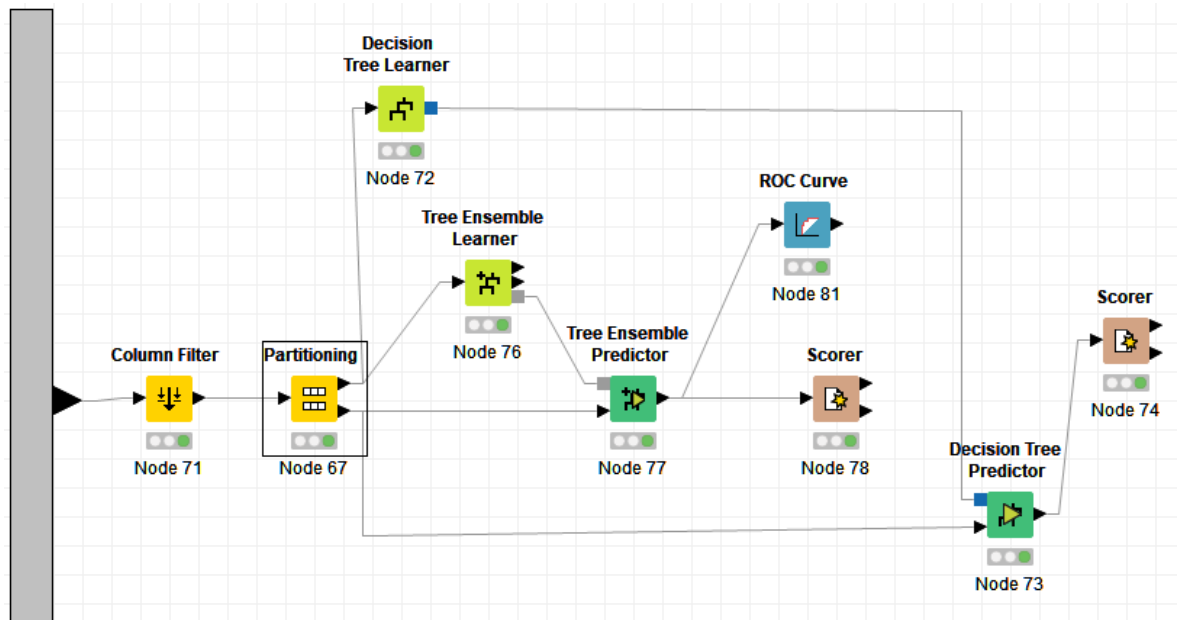


Figure 27 Machine Learning module

As a first step of this process is data partitioning (dividing the dataset into two chunks). This allows us to have two separate datasets, one for training the model and the other one for making comparisons for the predicted field. This can be achieved by Partitioning node which in our case is configured as in figure (Figure 28).

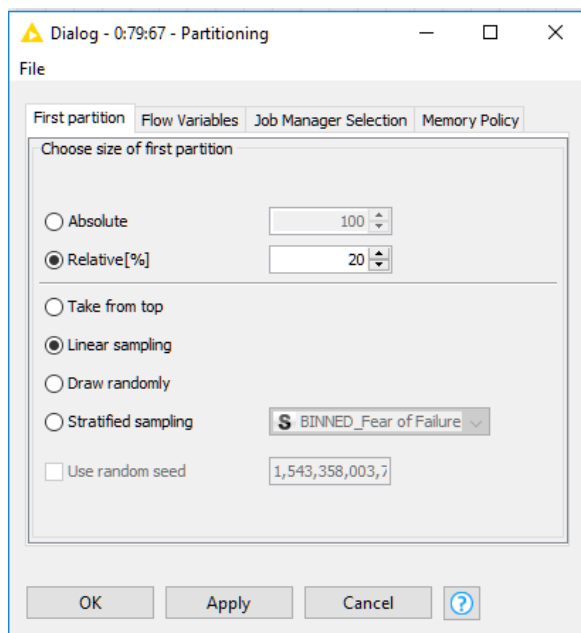


Figure 28 Partitioning node configuration

Partition node is configured to split 20 per cent of the input data using Linear Sampling.

Linear Sampling mode divides the dataset by including the first and last rows, as well as selects the remaining rows linearly from all the rows available. One example of this linear selection is by selecting every second row or every third row of the dataset. (KNIME User+Developer, 2018)

In our project we decided to go with only two Learners (nodes that get trained by the inputted data).



## **1. Decision Tree Learner**

This learner node is based on making decisions using data classification in a tree like model. As in all learners we need to have a target column which in our case must be nominal to perform this kind of learning model. The rest of the data set inputted for learning purposes can be either numerical or nominal. In the background logic of this node there are pre-formed numeric splits as well as nominal splits for classifying the data set. Numeric splits are always handled as binary ones with two outcomes which divides the domain in two separate parts based on a split point. On the other hand, the nominal splits can be binary split as well but also classified or split in as many nominal values as they contain. Then the algorithm comes up with two quality measures for calculating the classifications: the gain index which comes from the numerical splitting and gain ratio from nominal splits. (KNIME User+Developer, 2018)

## **2. Tree Ensemble Learner**

Different from the decision tree learner, tree ensemble learner follows a different approach. It uses regression method for learning from the dataset. There are multiple regression trees generated from learning different dataset and/or columns that are used as rulesets for making predictions. These predictions are based on the tree models by applying the tree ensemble predictor node to generate predictions. (KNIME User+Developer, 2018)

The reason why we decided to have two learners is because we can compare accuracy of the prediction conducted and choose the best for deployment. We are not going to explain in detail how these nodes are configured and what each configuration means because this is not the purpose of this thesis. Even though below we will briefly show how to setup this learner basic configurations.

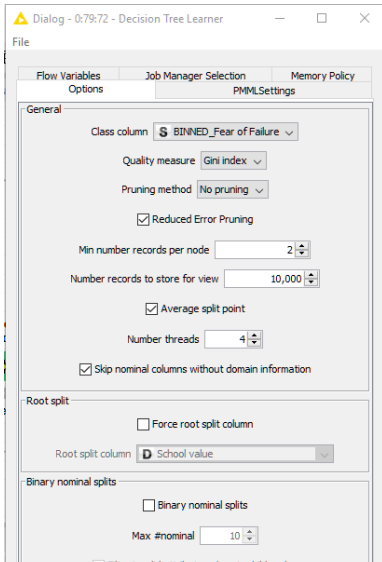


Figure 29 Decision Tree Learner Configuration

In figure (Figure 29) we can see a list of available configurations for this node. Usually these type nodes are auto-configured to perform the normal process of learning from the data provided. The most important change we need to make is define the Class Column to the one we are aiming to train the data for to make predictions on the next steps.

Configuration of Tree Ensemble learner is handled in the same way.

After executing these nodes, they conduct learning processes of the part of data we input for learning. Since both selected Learners are based on tree rule learning, pair of tree rulesets models can be found on the execution properties of these nodes shown in figure (Figure 30) and figure (Figure 31).

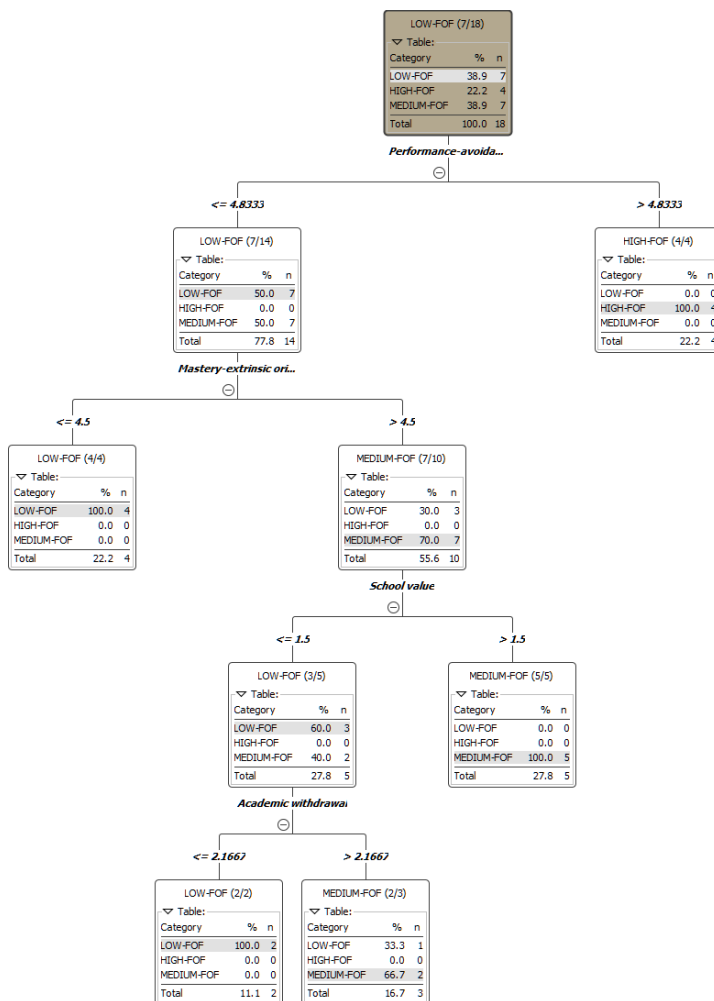


Figure 30 Decision Tree Learner Tree Model

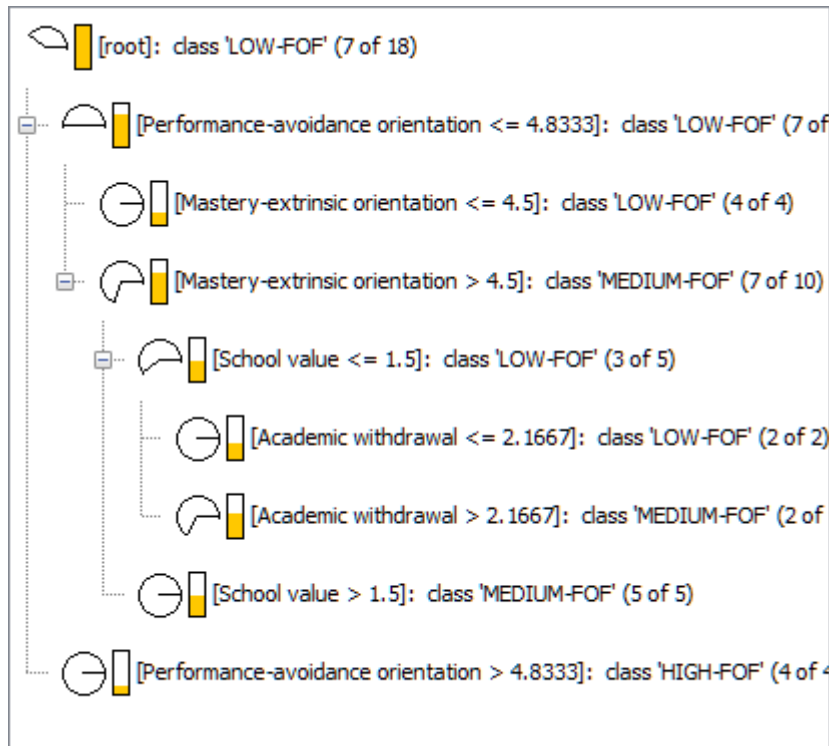


Figure 31 Decision Tree Learner Simple Model

The next step in the process is to use the other part of the data which we divided earlier using Partitioning node and make predictions using Decision Tree Predictor node.

This node must have as input the decision tree generated by the learner node in order to predict the decided class value. It uses patterns found by the learner node to make as accurate prediction. (KNIME User+Developer, 2018)

Configurations of this node are shown in figure (Figure 32). Most of these are cosmetic configurations for the new generated column which has the predicted data.

This node has the ability of using the Learner node ruleset by applying it to the test part of the data to make predictions.

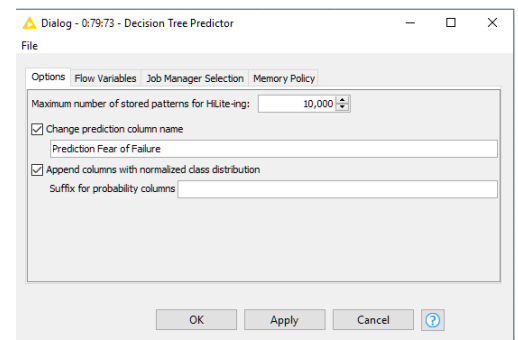


Figure 32 Decision Tree Predictor configuration

## 6.5 Evaluation

The most important step of machine learning is evaluating the results. The goal of this evaluation is to find the model out of the one we created in the previous steps and decide

how to use it for beneficial purposes. KNIME Analytics Platform offers good variety of evaluation nodes that can calculate and evaluate how accurate the predictions are. One of the most used nodes for this purpose is the Scorer node.

Scorer node provides as an output the calculated confusion matrix by comparing two specific columns (most of the time the predicted column with the original one). It compares in details the matches and mismatches of the selected columns. User has to choose these two columns manually for comparison, the node displays the first selected column values on the rows of the confusion matrix and the second one on the columns. The rest of the cells of this table are filled by the match and mismatch numbers accordingly. Another output of this node is a more detailed analytical table. This table has many other accuracy metrics such as: True-positives, True-Negatives, False-Positives, True-Negatives, False-Negatives, Recall, Precision, Sensitivity, Specificity, F-measure, as well as the overall accuracy and Cohen's kappa. (KNIME User+Developer, 2018)

In figure (Figure 27), the overall flow of machine learning is displayed. It is easy to notice that both flows end in this Scorer node. Every time we conduct a prediction of any kind there must be an evaluation for choosing the best model as final. In this project we will do evaluation manually by running the Scorer node and viewing accuracy results.

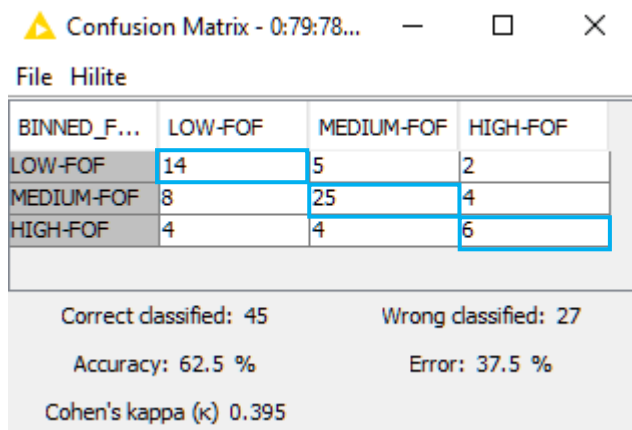


Figure 33 Confusion Matrix for Tree Ensemble model Scorer

In figure (Figure 33) is displayed the confusion matrix for the Tree ensemble predictions. This table is generated after executing the Scorer node. There is displayed most important numbers we need for evaluating the results.

We can see that Tree Ensemble model managed to make 62.5 % of the predictions correctly and the rest predicted wrongly.

Marked with blue are the data which were predicted correctly. Let's have a look at the other model results shown in figure (Figure 34).

It seems like our Decision Tree model has a higher accuracy rate and more correct predictions.

After comparing the two models we decided to go with Decision Tree model for the deployment sample workflow developed for testing the environment.

BINNED_F...	LOW-FOF	MEDIUM-FOF	HIGH-FOF
LOW-FOF	12	7	2
MEDIUM-FOF	4	28	5
HIGH-FOF	3	3	8

Correct classified: 48      Wrong classified: 24  
 Accuracy: 66.667 %      Error: 33.333 %  
 Cohen's kappa ( $\kappa$ ) 0.455

Figure 34 Confusion Matrix from Decision Tree model Scorer

## 6.6 Deployment

Deployment of this project will not be performed on a real KNIME server as such, but we will create a separate project to stimulate the server and make tests when applying new modified dataset in the solution.

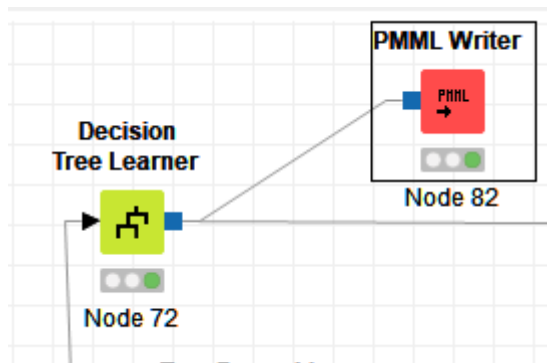


Figure 35 PMML Writer

First step is writing the selected model somewhere accessible because we will read it later in the project. The PMML Writer node is used to write models locally.

This node allows the user to define the location where the model is saved. After running this node, a local model file will be created

that can be used only by the PMML reader node which reads these files as models. (KNIME User+Developer, 2018)

After saving the model locally we try to generate new data for testing the environment. In our solution we pulled one person's answers from the dataset provided and used it as input to the testing project shown in figure (Figure 36)

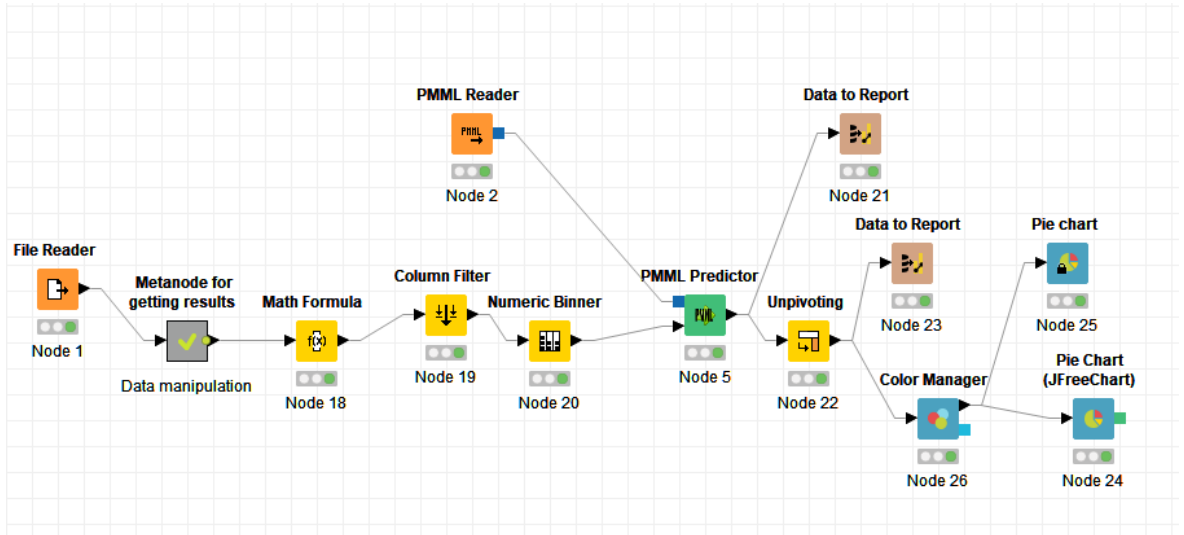


Figure 36 Test environment project with reporting feature

As mentioned before this simple project is developed to stimulate some environment for performing single row predictions of fear of failure as well as give a grade to the overall students' performance by calculating averages of all factors. For running this project, it is used BIRT Report Viewer which is running KNIME reports locally.

BIRT (Basic Intelligence Reporting Tool) is reporting software which belongs to open source software. It allows the users to generate reports out of the input data and these reports can be exported in PDF, PowerPoint, HTML etc. formats. In our case Knime has this tool already implemented. User just need to report the results of the flows to this report and configure BIRT user interface for better reporting understanding. (KNIME, 2018)

Results of this testing reporting as shown in figure (Figure 37)

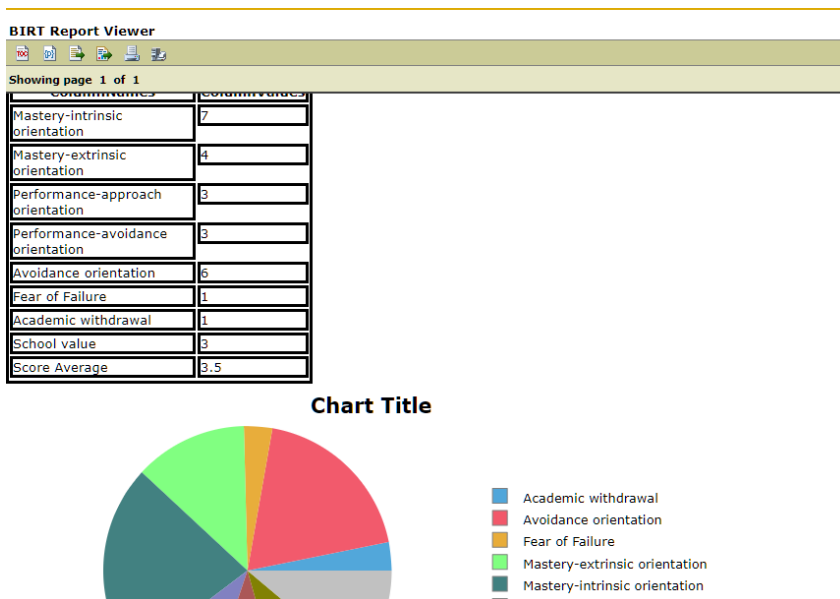


Figure 37 BIRT Report Viewer in sample testing environment

## 7 Results

We currently have a working KNIME project which has a stable data manipulation as well as variety of interactive data visualizations. Basics of machine learning for predicting fear of failure factor are implemented with a Decision Tree learner model of an accuracy more than 65%. In the same time some server data testing stimulations are handled on the other project submitted with this study.

Beside the technical results we also provided meaningful data visualisations for being interpreted by other student colleges involved in the same project such as Enxhi Nikolla. The graphs provided helped on finding hidden truths about current students' performance as well as making assumptions for future trends.

A detailed KNIME project workflow has been explained in this paper which includes almost all aspect of working with data and implementing basic machine learning. The reader will be able to understand the basics of KNIME Analytic Platform and its capabilities.

## **8 Conclusion**

The objectives of this project were to implement KNIME Analytics Platform in to studying and predicting Niemivirta's 8 scale factor which indicate the students' performance. The data given in use were generated by real students studying in the same degree.

There were many challenging situations during this thesis due to the limitation of amount of data we were provided, but this was a step forward in understanding KNIME in deep and going forward with test-projects during this whole time.

To conclude, capabilities of the platform decided to be used as a tool for developing this project are considerable and flexible. With KNIME Platform we managed to create many useful interactive charts that can be used in studying status of students' performance of BITe degree programme in Haaga Helia.



## References

- Brownlee, J. (2016) *Supervised and Unsupervised Machine Learning Algorithms, Machine Learning Mastery*. doi: 10.1287/inte.1070.0314.
- Chapman, P. et al. (2000) *CRISP-DM 1.0 (Step-by-step data mining guide), CRISP-DM Consortium*. doi: 10.1056/NEJMoa1108524.
- Daniel Faggella (2017) 'What is Machine Learning?', *What is Machine Learning?* Available at: <https://www.techemergence.com/what-is-machine-learning/>.
- Dietz, C. and Berthold, M. R. (2016) 'KNIME for open-source bioimage analysis: A tutorial', *Advances in Anatomy Embryology and Cell Biology*. doi: 10.1007/978-3-319-28549-8\_7.
- Halo Business Intelligence Halo Business Intelligence (2018) *Descriptive, Predictive, and Prescriptive Analytics Explained The two-minute guide to understanding and selecting the right Descriptive, Predictive, and Prescriptive Analytics*. Available at: <https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/>.
- Josh James (2016) *Data Never Sleeps 4.0 | Domo, DOMO*.
- KDnuggets (2014) *CRISP-DM, still the top methodology for analytics, data mining, or data science projects, 2014*.
- Knime (2018) *About Knime*. Available at: <https://www.knime.com/about>.
- KNIME User+Developer (2018) *NodePit*. Available at: <https://nodepit.com/>.
- Leskovec, J., Rajaraman, A. and Ullman, J. D. (2014) *Mining of massive datasets: Second edition, Mining of Massive Datasets: Second Edition*. doi: 10.1017/CBO9781139924801.

M.P. Bloothoofd, A. Francken, R. G. (2018) 'CRISP-DM METHODOLOGY'. Available at:

[http://www.luchtvaartfeiten.nl/uploads/thema/file\\_nl/5a8d67dc70726f300f010000/Factsheet\\_CRISP-DM.pdf](http://www.luchtvaartfeiten.nl/uploads/thema/file_nl/5a8d67dc70726f300f010000/Factsheet_CRISP-DM.pdf).

Maini, V. (2017) *A Beginner's Guide to AI/ML – Machine Learning for Humans – Medium, Medium*.

Marr, B. (2016) *A Short History of Machine Learning -- Every Manager Should Read, Forbes*. Available at:

<https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#5b9b941515e7>.

MathWorks (2016) 'What is Machine Learning?', *Machine Learning with MATLAB*. doi: 10.1111/j.2041-210X.2010.00056.x.

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*. doi: 10.2117/psysoc.2002.250.

Severino Rebecca (no date) *The data visualization catalogue*. Available at: <https://datavizcatalogue.com/about.html>.

Shravan, I. . (2017) 'Top 10 open source data mining tools'. Available at: <https://opensourceforu.com/2017/03/top-10-open-source-data-mining-tools/>.

To, S. H. (2018) *Bar Chart / Bar Graph: Examples, Excel Steps & Stacked Graphs*. Available at: <https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/bar-chart-bar-graph-examples/>.

Tomas Chamorro-Premuzic (2013) 'Is Technology Making Us Stupid (and

Smarter)?', *Is Technology Making Us Stupid (and Smarter)?* Available at:  
<https://www.psychologytoday.com/us/blog/mr-personality/201305/is-technology-making-us-stupid-and-smarter>.

Web Finance Inc. (2018) *Business Dictionary, Statistical Analysis*. Available at:  
<http://www.businessdictionary.com/definition/statistical-analysis.html>.