10/8/2020

Analysing machine learning algorithms to automate a decision making process

UNIVERSITY OF TWENTE.

Britt Marsman

Bachelor thesis
Title: Analysing machine learning algorithms to automate a decision making process
Date: 08-10-2020
City: Enschede

Student
B. Marsman (Britt)
Industrial Engineering and Management
University of Twente

Educational institution
University of Twente
Drienerlolaan 5
7522 NB Enschede
The Netherlands

Hosting Company
K&R Consultants
Gladsaxe 26
7327 JZ Apeldoorn
The Netherlands

First supervisor University of Twente
Dr. A. Abhishta

Supervisors K&R Consultants
Robert Jan Zwama and Wiert Evers

Second supervisor University of Twente
Dr. Ir. W.J.A. van Heeswijk

Acknowledgment

Before you lies my bachelor thesis which concludes my Bachelor's program in Industrial Engineering & Management at the University of Twente. The study has been performed at K&R Consultants in Apeldoorn.

Firstly, I would like to thank my supervisors, Dr. A. Abhishta, Dr. Ir. W.J.A. van Heeswijk, Robert Jan Zwama, and Wiert Evers. Abhishta acted as my first UT supervisor and provided valuable insights regarding both the machine learning analysis and the construction of this thesis. From the beginning, he was available for support and would help me when possible and his feedback was well elaborated which helped to improve the thesis to the current level. Next to this, I would like to thank Wouter for being my second supervisor. With his feedback, he helped me to even further improve my thesis.

Secondly, I want to thank both Robert Jan Zwama and Wiert Evers for allowing me to do this assignment at K&R Consultants and for their guidance and information needed to make this research a success. They were always available when I had questions and tried to help me. Furthermore, I would also like to thank the other colleagues at K&R Consultants for their support and fun times at the team outings.

Lastly, I want to thank family and friends for their support and interest in the project.

Britt Marsman

Enschede, October 2020

Management summary

This research has been conducted at K&R Consultants. K&R Consultants is a consultancy company in the construction sector located in Apeldoorn. They provide both advice on costs, installation and management. One of the activities of K&R Consultants is to analyse the budgets of contractors.

In Chapter 1 the motivation of this thesis is explained and the problems encountered in the manual analysis of K&R Consultants are illustrated. From these problems, the main research question is formulated, which is later divided into multiple sub-questions. The main research question is stated as follows: *How can the application of machine learning algorithms help K&R Consultants to automate the labelling process and in turn decrease the lead time of the process?*

First of all, the current situation of the manual analysis at K&R Consultants is explained in Chapter 2. This process goes from exporting the open budget from PDF to an Excel file, to filling in their tool, to labelling all the elements present in the open budget, to at last comparing the prices from K&R Consultants to the contractors.

After this, a literature study on machine learning is conducted. Here various types of machine learning are described and elaborated on, as well as on the different machine learning algorithms that fit the type of machine learning of K&R Consultants, which is supervised machine learning. Next to this, ways to validate the machine learning algorithm are explained. These entail the division between training and testing of datasets, the confusion matrix, the f-score, the ROC curve, and at last k-fold cross-validation.

Next, the data of K&R Consultants is analysed in RapidMiner, a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, and predictive analytics. Before analysing the data, the data is prepared first by removing inconsistent data and by balancing the data. After this, the predictive models and validation models are built.

In Chapter 5, the results are given. Here we find that three machine learning algorithms work well on the data of K&R Consultants. These are Gradient Boosted Trees, ID3, Naive Bayes, and k-Nearest Neighbour. After this, the algorithms are analysed using k-fold cross-validation where is shown that there is very little bias in the algorithms, because the accuracy does not fluctuate too much. Next to this, we also explain why the ROC curve does not work for the data of K&R Consultants, since the error is not uniformly distributed in the matrix. In this chapter also the advantages and disadvantages of the best performing algorithms are discussed.

At last, several recommendations are given to the company. First of all, the recommendation for continuous improvement. K&R Consultants should be aware of the fact that new data should first be assessed before being added to the dataset. Next to this, it is recommended to, when implementing the algorithm into their analysis, show the three highest predictions for every prediction. It occurs sometimes that the second prediction is the right prediction when showing the three highest predictions the right prediction could be picked more efficiently by an expert. There are also some points of improvement to be given. K&R Consultants should keep in mind that in the future a more accurate machine learning algorithm could be available. Next to this, when adding new data to the dataset an imbalance can occur again, which can be favourable for another machine learning algorithm.

Table of Contents

Acknowledgment 2
Management summary 3
Definitions 7
List of Tables 8
List of Figures 9
1. Introduction 10
1.1 Research aim 10
1.2 Context 10
1.2.1 K&R Consultants 10
1.2.2 Problem Identification 10
1.3 Methodological Framework 12
1.4 Structure of the Thesis 13
2. Current Situation 14
2.1 Receiving budget 14
2.2 Filling in the tool 15
2.3 Labelling 15
2.4 Comparing prices 17
2.5 Conclusion 18
3. Theoretical Framework 19
3.1 What is machine learning? 19
3.2 Types of Machine Learning 19
3.3 Machine learning algorithms 20
3.3.1 Decision trees 20
3.3.2 Random forest 20
3.3.3 Gradient boosted trees 21
3.3.4 ID3 21
3.3.5 Rule-based learning 21
3.3.6 Neural networks 21
3.3.7 Naive Bayes 21
3.3.8 K-Nearest Neighbour (k-NN) 22
3.3.9 Support Vector Machines (SVM) 22
3.4 Validation 22
3.4.1 Division of datasets 22
3.4.2 Confusion matrix 23

3.4.3 Receiver operating characteristic curve 25
3.4.4 K-fold cross-validation 25
3.5 Machine learning algorithms evolution 26
3.6 Conclusion 27
4. Design and development 28
4.1 RapidMiner 28
4.2 Data preparation 28
4.3 Balance of the dataset 29
4.4 Building predictive models 31
4.5 Testing and training data 32
4.6 K-fold cross-validation 34
4.7 Selection of Algorithms 35
4.8 Comparing Algorithms 35
4.8.1 Accuracy scores 35
4.8.2 F-score 35
4.9 ROC curve 35
4.10 Speed of process 35
4.11 Conclusion 35
5. Implementation & Demonstration 37
5.1 Results 37
5.2 K-fold cross-validation 37
5.3 ROC curve 40
5.4 Advantages and disadvantages algorithms 40
5.4.1 Naive Bayes 40
5.4.2 k-Nearest Neighbour 41
5.4.3 Gradient boosted trees 41
5.4.4 ID3 41
6. Recommendations & Conclusion 42
6.1 Conclusions 42
6.1.1 Problem 42
6.1.2 Research question 42
6.1.3 Answering the Research Question 42
6.1.4 Answering the Sub Questions 42
6.1.5 Findings 44
6.2 Recommendations 44
6.2.1 Application 44

6.2.2 Continuous improvement 44
6.2.3 Noise 44
6.2.4 Speed of process 45
6.3 Discussion 45
6.3.1 Choice of algorithms 45
6.3.2 Other algorithms 45
References 46

Definitions

AI = Artificial Intelligence

AUC = Area Under the Curve

DM = Data Mining

DMMCM = Data Mining methods conceptual map

DMMTs = Data Mining methods templates

DSRM = Design Science Research Methodology

FN = False Negative

FP = False Positive

FPR = False Positive Rate

ID3 = Iterative Dichotomiser 3

K&R = K&R Consultants

k-NN = k-Nearest Neighbour

KPI = Key Performer Indicator

ML = Machine learning

OCR = Optical character recognition

ROC = Receiver operating characteristics

SVM = Support Vector Machines

TN = True Negative

TP = True Positive

TPR = True Positive Rate

List of Tables

Table 1. Example of converted Excel file (in Dutch) 15
Table 2. Labelling process in Excel (in Dutch) 16
Table 3. Comparing prices with database, part 1 (in Dutch) 17
Table 4. Comparing prices with database, part 2 (in Dutch) 17
Table 5. Example dataset 20
Table 6. Example data used for analysis of budget 28
Table 7. Output of using Naive Bayes operator in scoring process in RapidMiner 31
Table 8. Confusion matrix using Naive Bayes operator 32
Table 9. Output of using Naive Bayes operator in testing and training process in RapidMiner 33
Table 10. Prediction of model with top three highest confidences 33
Table 11. Results of machine learning algorithms 37
Table 12. Comparison highest accuracy scores k-fold cross-validation 40
Table 13. Multiclass Confusion Matrix using the Naive Bayes operator 40
Table 14. Results of analysis 42

List of Figures

Figure 1. K&R Analysis steps 10
Figure 2. Problem cluster 11
Figure 3. Design Science Research Methodology (Peffers et al., 2007) 12
Figure 4. Example of budget (in Dutch) (K&R, 2019) 14
Figure 5. General concept of machine learning 19
Figure 6. Types of machine learning 19
Figure 7. Division of dataset 23
Figure 8. Confusion matrix (Narkhede, 2018) 23
Figure 9. Multiclass confusion matrix 24
Figure 10. ROC curve (Narkhede, 2018) 25
Figure 11. K-fold cross-validation (K-Fold Cross-Validation in Machine learning?, 2020) 26
Figure 12. Process of filtering out missing values 29
Figure 132. Process of giving a role to data 29
Figure 14. Histogram of number of appearances of codes in the data 30
Figure 15. Histogram of elements which appear less frequently 30
Figure 16. Scoring process in RapidMiner 31
Figure 17. Testing and training process in RapidMiner 32
Figure 18. Cross-validation process 34
Figure 19. Sub-processes operator Cross-Validation 34
Figure 20. Histogram of k-fold cross-validation analysis of Naive Bayes algorithm 38
Figure 21. Histogram of k-fold cross-validation analysis of k-Nearest Neighbour 38
Figure 22. Histogram of k-fold cross-validation analysis of ID3 39
Figure 23. Histogram of k-fold cross-validation analysis of Gradient Boosted Trees 39
Figure 24. K&R Analysis steps 43

1. Introduction

In this chapter, the reader is provided with the 1.1) Research Aim, 1.2) Context, 1.3) Methodological Framework and 1.6) Structure of the Thesis.

1.1 Research aim

Automated machine learning represents a fundamental shift in the way organizations of all sizes approach machine learning and data science. Automated machine learning is the process of automating the process of applying machine learning to real-world problems. Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed (What is Machine Learning? A definition, 2020). Applying traditional machine learning methods to real-world business problems is time-consuming, resource-intensive, and challenging. It requires experts in several disciplines. Automated machine learning changes that, making it easier to build and use machine learning models in the real world by running systematic processes on raw data and selecting models that pull the most relevant information form the data ("What Is Automated Machine Learning", n.d.). Building machine learning models is an experiment-driven science. It is not sure what will work well. Therefore, experiments need to be designed, they need to be run and the results need to be analysed (Calomme, 2019).

K&R Consultants is the company where the research described in this document is taking place. They are currently doing most of their analysis work manually. In this thesis, the automation of their manual analysis will be researched.

1.2 Context

1.2.1 K&R Consultants

K&R Consultants is a consultancy company in the construction sector located in Apeldoorn. They provide advice on costs as well as advice on installation and management. One of the activities of K&R Consultants is to analyse public tenders of contractors. The analysis is performed to find out of the prices on the budgets are fair. The budget includes all costs for a particular building project, such as building materials, quantities, and work hours. K&R Consultants are often hired by companies or other institutions, to help them analyse the public tenders from different contractors. The companies and institutions mentioned often do not have the expertise to make the right choices as to what contractor to choose. This is where K&R comes in to help make the right decisions.

K&R Consultants come to solutions through a collaborative approach. The approach focuses on both costs and the process.

1.2.2 Problem Identification

The problem K&R Consultants is currently facing lies in the analysis they are doing for their customers. The analysis consists of six steps and can be seen in Figure 1.

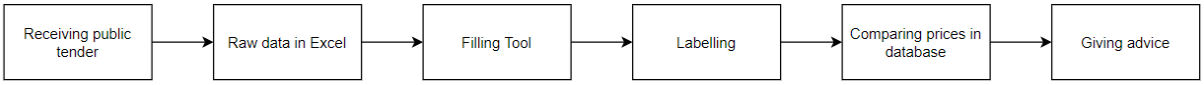

Figure 1. K&R Analysis steps

First of all, K&R Consultants receive a public tender from a contractor. These public tenders contain architectural elaboration, a technical description, and the budget for the construction. These are often in a PDF file. Since K&R uses Excel for its analysis tool they need to convert the PDF file to Excel. They use Optical Character Recognition (OCR) for this. OCR is a mechanical conversion of images of typed, handwritten, or printed text into machine-encoded text. After this, they will fill their tool in Excel with

the data of the budget. Next, the employees at K&R Consultants will start to manually label every element in the budget. This is done by giving every line in the budget a code that represents that element on that specific line. These codes are designed specifically for K&R Consultants. After this, they can compare the budget with their database and advise the contractor on the findings. The steps will also be explained in more detail in Chapter 2.

Since most contractors use different styles and formats for their budgets, it takes K&R Consultants a lot of time to make the budgets comparable to their database. Next to this, the process is mostly manual, which means that human mistakes can be made.

At this moment the labelling process takes the most time and because K&R Consultants have to manually do this step it is a very repetitive task. This leads to the fact that the analysis of the budgets is taking too long and is prone to human mistakes. To get more insight into the causality between the problems, a problem cluster was created from the relevant problems, which can be seen in Figure 2.

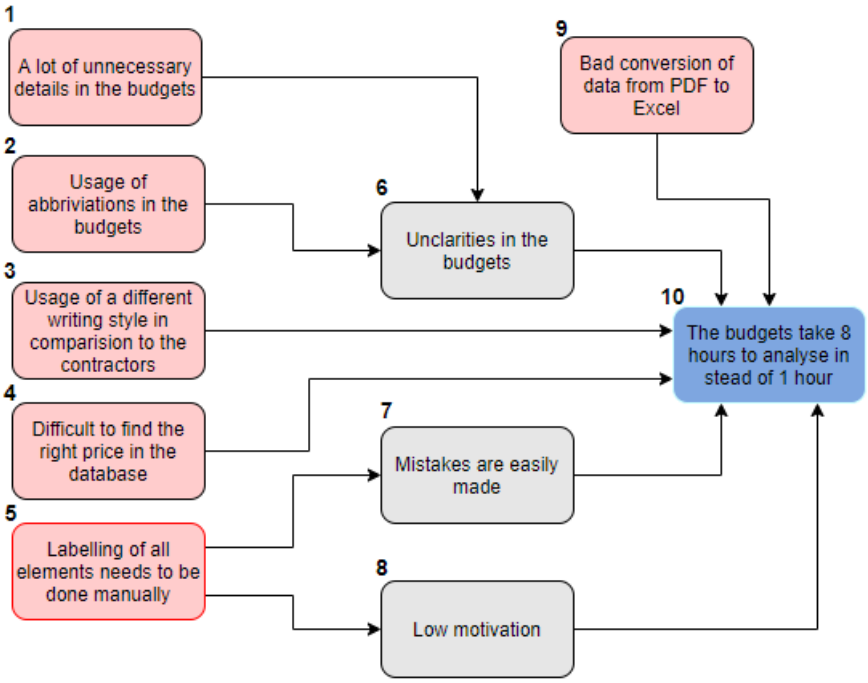

Figure 2. Problem cluster

Each block in the problem cluster represents a problem as stated by the personnel from K&R Consultants. The causality between problems is shown in arrow form. For instance, problem 2 is the cause of problem 6. On the right, problem 10, the action problem is displayed. The action problem captures the discrepancy between norm and reality (Heerkens, 2017). The research proposed in this report is aimed at solving this problem. However, the action problem can not be solved directly. Some problems have no causes and directly or indirectly influence the action problem. These problems are called candidate core problems and are visualized in red. From the six candidate core problems, the actual core problem must be identified. In the end, after solving the core problem the action problem will also be solved.

To find the core problem all the candidate core problems must be evaluated. Problems 1 and 2 can not be influenced easily. This is made clear by the employees of K&R Consultants. It is very unlikely for the contractors to start using budgets with different layout and clearer elements. Since the usage of an unclear budget can help them in making more profit. Problem 3 is also a problem that can not be solved easily, since the contractors do not want to change their way of making the budgets and for the analysis, the writing style of K&R Consultants is better. Problem 9 is something that can be solved by

looking into more suitable OCR programs. Therefore this problem is not suitable to solve for the scope of this report. Problem 4 is being caused since K&R Consultants does too little analysis per year to make a valid database with market conform prices. This leaves problem 5. This problem can be influenced and reduces the time to spend on doing the analysis. Therefore, problem 5 is chosen as the core problem.

To solve this problem an extra tool could be implemented to the analysis tool that K&R Consultants is currently using. This tool could make it easier to label the elements. This would in turn also increase the lead time of the process.

Machine learning algorithms have been suitable for analysing data as provided by K&R Consultants in the past. An algorithm is a set of instructions to solve a certain problem. Machine learning implies that a computer teaches itself to recognise certain patterns in the data. For the patterns, a computer can make predictions. Hence, machine learning could be useful in the labelling process of K&R Consultants. The machine learning algorithms can be analysed and validated by using the program RapidMiner, which is a data science software platform for machine learning.

Therefore, the research question can be formulated as follows:

> *How can the application of Machine Learning algorithms help K&R Consultants to automate the labelling process and in turn decrease the lead time of the process?*

1.3 Methodological Framework

The methodology framework applied for this research is the Design Science Research Methodology (DSRM) (Peffers et al., 2007). This methodology is focused on the creation of 'things' and is developed since the descriptive research was often not applicable to the solution of problems encountered in research and practice. Design science is of importance in a discipline-oriented to the creation of successful artifacts. The DSRM presented here incorporates principles, practices, and procedures required to carry out such research and meets three objectives: it is consistent with prior literature, it provides a process model for doing design science research, and it provides a mental model for presenting and evaluating design science research in information systems.

The DSRM consists of six phases and is used as a guideline throughout this report. These phases are visualised in Figure 3.

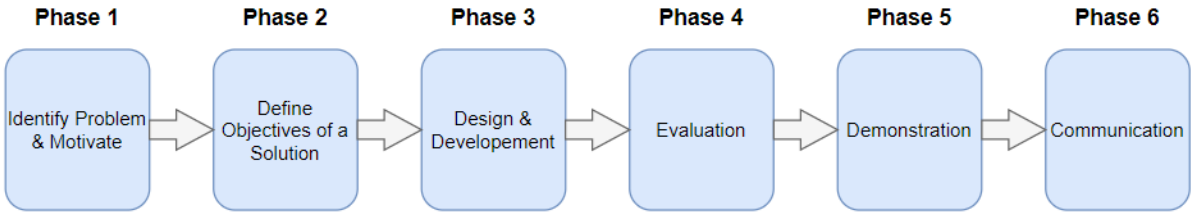

Figure 3. Design Science Research Methodology (Peffers et al., 2007)

In phase 1, the specific research problem is defined and the value of a solution is justified. By justifying the value of a solution the researcher and the audience will be motivated to pursue the solution and to accept the results. This phase is described in Chapter 1.

In phase 2, the objectives of a solution from the problem definition and knowledge of what is possible and feasible are inferred. The objectives can be quantitative, terms in which a desirable solution would be better than current ones. This phase can be found in Chapter 1.

In phase 3, the artifact is created. Such artifacts are potentially constructs, models, methods, or instantiations. This phase can be found in Chapter 4.

In phase 4, the use of the artifact demonstrates how to solve one or more instances of the problem. This could involve its use in experimentation, simulation, case study, or proof. This phase can be found in Chapter 5.

In phase 5, the artifact is observed and measured how well it supports the solution to the problem. This activity involves comparing the objectives of a solution to actual observed results from the use of the artifact in the demonstration. This phase can be found in Chapter 6.

In phase 6, the problem and its importance are communicated. This phase can be found in Chapter 6.

1.4 Structure of the Thesis

After identifying the core problem, the research concentrates on the testing and validating of the effect of machine learning algorithms on the data of K&R Consultants. To be able to answer the research question multiple knowledge questions are formulated. Chapter 2 describes the current situation of the manual analysis of K&R Consultants. In Chapter 3, the machine learning algorithms available are noted down as well as techniques for validating the machine learning algorithms. Chapter 4 shows how the analysis has been done. In Chapter 5, the results of the test will be shown. In Chapter 6, the recommendation and conclusion are described.

In Chapter 2 the following questions are answered:

1. How is the manual analysis of price lists at K&R Consultants currently done?
    a) What are the steps in the manual analysis of K&R Consultants?
    b) What are the points of improvement of the manual analysis of K&R Consultants?

In Chapter 3 the following questions are answered:

2. What are the different machine learning algorithms in supervised learning?
    a) What machine learning algorithms are available?
    b) How can machine learning algorithms be evaluated and compared?

In Chapter 4 the following questions are answered:

3. How can RapidMiner be used to analyse the machine learning algorithms?
    a) What steps need to be taken to analyse the machine learning algorithms?
    b) How can the validation and testing of the algorithms be done in RapidMiner?

In Chapter 5 the focus lies on the testing of the machine learning algorithms. In this chapter, the following questions are answered:

4. What Machine Learning algorithms work for K&R Consultants?

In Chapter 6 the following questions are answered:

5. What steps should K&R Consultants take to implement the Machine Learning algorithm and improve the decision making process?

2. Current Situation

Currently, K&R Consultants go through a couple of steps to analyse the budgets they get from contractors. This overview can be seen in Figure 1. In this chapter, the reader is provided with the sections 2.1) Receiving budget, 2.2) Filling in the tool, 2.3) Labelling, 2.4) Comparing prices, 2.5) Conclusion.

2.1 Receiving budget

First of all, the analysis begins with the gathering of information. K&R Consultants receive the budget from contractors and often these are sent in a PDF file. An example of a budget can be seen in Figure 4.

Calculatie: 00104510/06020 Offertepost: 20.03.2019 Blad : 1

Offertepost : KRACHTINSTALLATIE Werk:
Valuta : EUR Aanvrager:
Bestek:

Reg	Mat-kode	Fabr.	Omschrijving	Aantal	Eh	Netto	Netto-tot	Norm	Norm-tot
1			** KRACHTINSTALLATIE **						
2									
3			** SPECIALE RUIMTEN VOORZIENINGEN **						
4	9999999900		SKILLS LAB; CONTACTDOZEN VOOR BEDDEN	1,00	POS	250,00	250,00	8,0000	8,00
5	9999999900		CREW RESOURCE MANAGEMENT; DIIVERSE WCD	1,00	POS	350,00	350,00	12,0000	12,00
6	9999999900		REGIERUIMTE; WCD PC'S	2,00	POS	100,00	200,00	4,0000	8,00
7	9999999900		PRESENTATIE SCHERMEN WCD'S	44,00	POS	50,00	2200,00	2,0000	88,00
8	9999999900		INFORMATIESCHERMEN	7,00	POS	50,00	350,00	2,0000	14,00
9	9999999900		AANSLUITIGN VOOR PRINTER	10,00	POS	25,00	250,00	1,0000	10,00
10	9999999900		AANSLUITINGEN KOFFIECORNER	8,00	POS	150,00	1200,00	4,0000	32,00
11	9999999900		AANSLUITING BOILER IN WERKKAST	10,00	POS	50,00	500,00	1,5000	15,00
12									
13			** ALGEMENE AANSLUITINGEN **						
14	S580800000		INB.DOOS U50	259,00	STK	1,86	481,74	0,3100	80,29
15	S481600020		INBOUWDOOS VOOR WANDGOOT (2 STUKS)	217,00	STK	7,86	1705,62	0,2100	45,57
16	S481700000		LASDOOS VOOR WANDGOOT	217,00	STK	3,93	852,81	0,1700	36,89
17	S610321720		WKD 1-V MET R.A. INBOUW	19,00	STK	3,43	65,17	0,1700	3,23
18	S610323720		WKD 2-V MET R.A. INBOUW	120,00	STK	6,86	823,20	0,3400	40,80
19	S610323722		WKD 2-V MET R.A. (WG) INBOUW	217,00	STK	6,86	1488,62	0,3400	73,78
20	S681600000		AFDEKPLAAT 1-VOUDIG	22,00	STK	1,48	32,56	0,0000	0,00
21	S681700000		AFDEKPLAAT 2-VOUDIG	361,00	STK	2,53	913,33	0,0000	0,00
22	S610321540		WKD 1-V MET R.A. OPBOUW	12,00	STK	4,71	56,52	0,2800	3,36
23	S630323540		WKD 2-V MET R.A. OPBOUW W.D.	7,00	STK	15,83	110,81	0,4200	2,94
24	9999999900		AANSLUITPUNT 230V	7,00	STK	3,50	24,50	0,5000	3,50
25	0000213281	MENNEKES	CEE-WANDCONTACTDOOS WCD TwinCONTACT 16A 4P 6H 400V IP44	2,00	STK	6,60	13,20	0,4200	0,84
26	0000213253	MENNEKES	CEE-WANDCONTACTDOOS WCD opb 16A5P 6H400V I	2,00	STK	6,60	13,20	0,4400	0,88

Figure 4. Example of budget (in Dutch)

K&R Consultants use Excel to perform the analysis. Therefore, the data in the PDF files need to be converted into raw data in Excel. Because the contractors all use different formats for the budgets, does this often lead to problems with the layout in Excel. The converter tool from Adobe regularly converts different lines in the PDF file into one line in Excel. This can lead to several problems when labelling the data, important data could get lost or some posts will be added together which do not belong together. An example of a converted PDF file into an Excel file can be seen in Table 1. In this case, the conversion was done correctly, since the descriptions in the budget were clear and had a good layout.

Table 1. Example of converted Excel file (in Dutch)

Reg Mat-kade	Fabr. Omschrijving	Aantal	Eh	Netto	Netto-tot
	* * ALGEMENE AANSLUITINGEN * *				
4 S580800000	!NB.DOOS USO	186,00	STK	1,70	316,20
5 S481600020	INBOUWDOOS VOOR WANDGOOT (2 STUKS)	18,00	STK	7,86	141,48
6 S481700000	LASDOOS VOOR WANDGOOT	6,00	STK	3,93	23,58
7 S610321720	WKD 1-V MET R.A. INBOUW	1,00	STK	3,43	3,43
8 S610323720	WKD 2-V MET R.A. INBOUW	105,00	STK	6,86	123,48
9 S610323722	WKD 2-V MET R.A. (WG) INBOUW	18,00	STK	6,86	123,48

The conversion from PDF to Excel is not the phase that takes the longest but is it frustrating work and also the highly educated employees of K&R Consultant do not want to waste their time on this. Nevertheless, is it important that no mistakes are made in this phase since it could make the analysis useless.

2.2 Filling in the tool

After the data is converted into Excel the data will be put into a tool in Excel which K&R Consultants has developed. This tool, in which also the database with the market conform prices are included, enables the employees of K&R Consultants to do a big part of their analysis.

In Figure 4 can be seen how a budget is put together. The column names used in the budget are also present in the tool in Excel, such as 'description', 'amount', or 'price'. In every column, there is supposed to only be one sort of information, so that in the end it is easy to calculate prices and the total amount. This step shows how important it is that in the first step the data is converted correctly.

2.3 Labelling

The next phase in the analysis of the budget is the labelling of all the elements. In the columns, A till J will be put all the information from the budget from the contractor. The labelling will be done in 'Hoofdcode', 'Kolom2', and 'Kolom3', respectively 'R', 'T', 'U', which can be seen in Table 2. All the elements in the database of K&R Consultants are divided into codes, where every code stands for a different element.

The first two numbers show to which main code ('Hoofdcode') the element belongs. Examples of these main codes are 'draining', 'air treatment', and 'power flow'. Consequently, within these main codes, all the elements will be present. Every main code can be divided into different variants. These names can be found in the so-named column 'Kolom2' in Table 2. When looking at the main code 'draining', examples of variants would be 'HWA outside appendages' and 'VWA piping'. At last, different variants can be divided into different elements. These elements can be found in the column 'Kolom3'. This means that every element has a unique code of six numbers, where the first two numbers represent the main code, the second two numbers on the variant, and the last two numbers on the unique element.

Currently, the process of labelling is that the employees of K&R look at each row and label each row individually, which is in line with the materials in the budget. The description is rarely the same, so the employees need to look good and find the information they need. Based on the descriptions and their knowledge of the database, the employees can give a label to all the elements. After the employees filled in the label, the tool will automatically fill in the corresponding code.

When labelling a wall socket all the elements such as the wiring and the cover plates get the same label. This means that there will be one price for the element wall socket instead of all separate prices for the separate components.

Table 2. Labelling process in Excel (in Dutch)

Hoofdcode	Kolom2	Kolom3	Code K&R
..55	Koelmachines	Koelmachines_LOD200	55.05.06
..55	Koelmachines	Koelmachines_LOD200	55.05.05
..55	Koelmachines	Koelmachines_LOD200	55.08.05
..55	Koelmachines	Koelmachines_LOD200	55.03.00
..55	Koelmachines	Koelmachines_LOD200	55.05.05
..55	Afsluiters_koudedistributie	Regelafsluiter_GKW_DN_150	55.01.20
..55	Leidingen_schacht_en_TR_koudeopwekking	Leidingen_GKW_DN_150_staal	55.12.03
..55	Pompen_koudedistributie	Circulatie_pomp_GKW_DN_150	55.12.03
..55	Inregelafsluiters_GKW_koudedistributie	Inregelafsluiter_GKW_DN_150	55.10.06
..55	Toebehoren_koudeopwekking	Appendages_gkw	55.10.06
..55	Toebehoren_koudeopwekking	Appendages_gkw	55.10.06
..55	Toebehoren_koudeopwekking	Appendages_gkw	55.10.02
..55	Pompen_koudedistributie	Circulatie_pomp_GKW_DN_50	55.10.02
..55	Afsluiters_koudedistributie	Regelafsluiter_GKW_DN_50	55.10.02
..55	Inregelafsluiters_GKW_koudedistributie	Inregelafsluiter_GKW_DN_50	55.10.02
..55	Decentrale_systemen_eigen_opwekking	Splitunit	55.10.03
..55	Leidingen_verdieping_koudeopwekking	Aansluitleidingen_gkw_staal	55.10.06
..55	Toebehoren_koudeopwekking	Appendages_gkw	55.04.07
..56	Afsluiters_CV_warmtedistributie	Afsluiter_CV_DN_200	56.05.06
..56	Leidingen_CV_schacht_en_TR	Leidingen_CV_DN_200_staal	56.05.05

After the labelling of all the rows, there are often more rows that belong to one label. This means that the employees of K&R Consultants need to adjust the numbers. When looking at for example a wall socket the contractors make a difference between the socket itself and the cover plate. Due to this, it shows that there are for example 110 sockets and 110 cover plates needed. For K&R Consultants this all falls under the same 110 wall sockets. Therefore the numbers in the rows need to be adjusted. The adjusting of the numbers takes a lot of time and can also result in the employees making mistakes. This step is necessary since the tool would count 220 wall sockets instead of the 110 that are needed.

2.4 Comparing prices

After the labelling of every row, the prices from the budget can be compared with the prices of the database. This can be seen in Table 4. The table is split into two since it would not fit on one page. After every label, the number, the total costs of all individual elements, and the price of the contractor are automatically filled in. In the columns next to that the price from K&Rs database is filled in. There can also be seen the absolute and percentage difference between the total costs from the contractor and the database. A percentage turns green when the total costs of K&R Consultants exceed the of the contractor and turn red when the total costs of K&R Consultants lie under the contractor costs.

The labels where the percentage price difference is more than approximately 30% positive and negative are marked with a colour and/or notes. These are the labels where K&R Consultants need to discuss with the contractors how the difference can be this big. In this phase with the expertise, experience, and knowledge of the employees of K&R and with the help of the tool a conclusion will be made over the prices in the budget.

Table 3. Comparing prices with database, part 1 (in Dutch)

code	Beschrijving	aantal	materiaal / derden	arbeid	Toeslagen	Totaal
51.03.02	expansievat_30l	1	63	128	22	213
51.05.01	Warmtepompen_met_bron_water	38	23.131	643	2.631	26.405
52.00.00	Afvoeren_LOD100_c	15	369	245	69	683
52.01.09	Straatkolk	2	156	268	48	472
52.01.12	polder_expansie_stukken	1	29	22	6	57
52.02.04	Afvoerleiding_PE_volvul	116	2.187	6.983	1.038	10.208
52.03.11	Afvoertrechter_vol_vul_systeem	6	566	402	108	1.076

Table 4. Comparing prices with database, part 2 (in Dutch)

Eenheidsprijs Aannemer		Eenheidsprijs K & R	Totaal aantal K & R	Verschil absoluut	Verschil percentueel
213,21		350,00	350	137	39%
109,43		115,00	782	38	5%
694,86		518,00	19.684	6.721	34%
45,57		45,57	683	-	0%
235,75		750,00	1.500	1.028	69%
56,98		375,00	375	318	85%
88,00		72,00	8.352	1.856	22%
179,29		189,00	1.134	58	5%

2.5 Conclusion

After the description of the current situation, there are a couple of points where improvements can be made. First of all, in the conversion from PDF to Excel (section 1.2.2.1) there are a lot of problems with the layout that come up. Due to this problem, a lot has to be done by hand to fix the problem and to make sure that the problems that occurred in this step do not cause any problems later on in the analysis.

Next, the labelling phase (1.2.2.3), since K&R does this all by hand this takes the most time. This is because they have to go over each row individually. Therefore, this is very repetitive and prone to mistakes.

Since the first problem is not suitable for the scope of this research, will this research only be focusing on the labelling problem and not on the conversion problem.

3. Theoretical Framework

In this chapter, the reader is provided with the 3.1) What is machine learning?, 3.2) Types of machine learning, 3.3) Machine learning algorithms, 3.4) Validation, 3.5) Machine learning algorithm evolution.

3.1 What is machine learning?

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. A general concept of machine learning is visualised in Figure 5 (Kulkarni, 2012). Machine learning is about designing algorithms that allow a computer to analyse data from various sources, select relevant data, and use those to predict the behaviour of the system in another similar and if possible different scenarios. An algorithm is a sequence of instructions that should be carried out to transform the input to output (Alpaydin,2010).

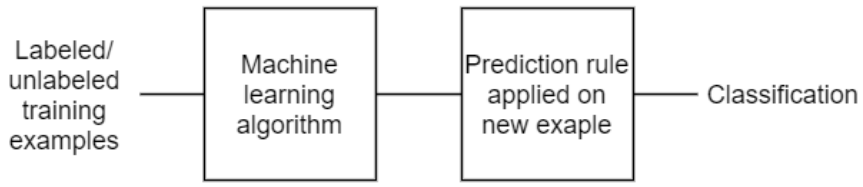

Figure 5. General concept of machine learning

The learning that is being done is always based on some sort of observation or data, such as examples, direct experience, or instruction. Learning does not necessarily involve consciousness but learning is a matter of finding statistical regularities or other patterns in the data. Furthermore, machine learning also classifies objects and behaviours to make decisions for the new input scenarios (Kulkarni, 2012). So, machine learning is about learning to do better in the future on what was experienced in the past.

3.2 Types of Machine Learning

In general, there are three types of machine learning: 1) reinforcement learning, 2) unsupervised learning, and 3) supervised learning (Figure 6) (Zhang, 2010). The type of learning used at K&R Consultants is supervised learning since they make use of labels.

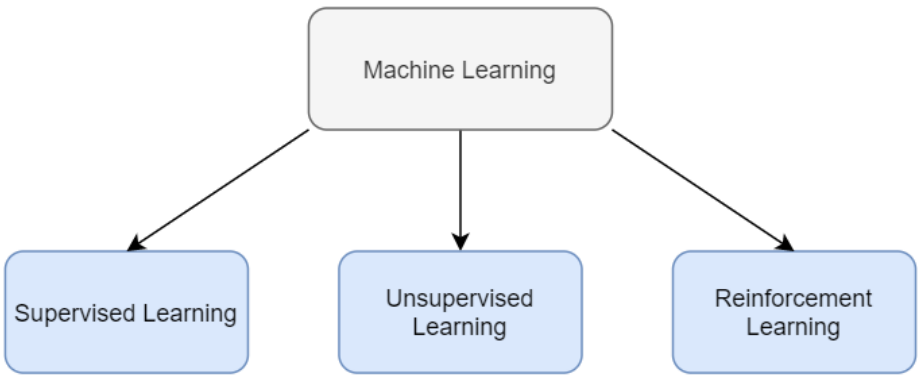

Figure 6. Types of machine learning

Supervised learning is where you have input variables and output variables and you use an algorithm to learn the mapping function from the input to the output. The goal is to approximate the mapping

function of the machine learning in such a way that when you have new input data that you can predict the output variables for the data. An example of a dataset used in supervised learning can be seen in Table 5. This is a labelled dataset. That implies that the used dataset contains both input and output data. The machine learning algorithm can use the input data (colour, body shape, and hair type) to make predictions. In this case, the algorithm will make predictions on the characteristic of the dog. For example, the machine learning algorithm gets the input black, big, and poodle from the input colour, body shape, and hair type respectively. Given this input data, the machine learning algorithm will give "Danger" as output data. The machine learning algorithm also checks whether or not the output data is correct. If this is not the case the machine learning algorithm will adjust the prediction strategy.

Table 5. Example dataset

Input data			**Output data**
Colour	**Body Shape**	**Hair Type**	**Characteristics**
Black	Big	Poodle	Danger
Brown	Big	Smooth	Danger
Black	Small	Smooth	Safe
Black	Medium	Poodle	Safe

Supervised learning can be divided into two categories, regression, and classification. They both share the same concept of utilizing known datasets to make predictions, but there are also some differences. The main difference is that the output variable in the regression is numerical (or continuous) while the output variable for classification is categorical (or discrete). In the case of K&R Consultants, the output variable is discrete, therefore classification should be used.

3.3 Machine learning algorithms

For supervised learning, there are some machine learning algorithms available. Kotsiantis (2007) in his paper, *Supervised Learning: A Review of Classification Techniques*, described the most effective supervised machine learning algorithms. Kotsiantis categorised these algorithms into six types: 1) Decision trees, 2) Rule-based learning, 3) Neural networks, 4) Naive Bayes, 5) K-Nearest Neighbour, 6) Support Vector Machines. The Decision tree type includes decision trees, random forest, gradient boosted trees, and ID3. These are all explained in the following sections.

3.3.1 Decision trees

Decision trees are trees that classify instances by sorting them based on feature values. A feature is a variable in the dataset where the data could be split on. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or outcomes. The decision nodes are where the data is split based on a certain feature from the data (Seif, 2018).

3.3.2 Random forest

In Random Forest, there are multiple trees as opposed to a single tree used in Decision Trees. The multiple trees operate together. Each individual tree in the random forest predicts a class prediction and the class with the most votes becomes the model's prediction. There are a large number of relatively uncorrelated models (trees) that are operating together. Uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The trees protect each other from their individual errors (Yiu, 2019).

3.3.3 Gradient boosted trees

The gradient boosting algorithm can be most easily explained by first introducing the AdaBoost Algorithm. The AdaBoost Algorithm begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, the weights of those observations that are difficult to classify are increased and the weights for those that are easy to classify are decreased. The second tree is grown on this weighted data. The idea is to improve upon the predictions of the first tree. Predictions of the final model are therefore the weighted sum of the predictions made by the previous tree models. Gradient Boosting trains many models in a gradual, additive, and sequential manner. The difference between AdaBoost and Gradient Boosting is how the two algorithms identify the shortcomings of weak learners. Gradient Boosting performs the same by using gradients in the loss function. The loss function describes how well the model will perform given the current set of parameters (weights and biases), and gradients are used to find the best set of parameters (Brownlee, 2016).

3.3.4 ID3

ID3 stands for Iterative Dichotomiser 3 and is named such because the algorithm repeatedly divides features into two or more groups at each step. ID3 uses a top-down greedy approach to build a decision tree. The greedy approach means that at each iteration the best feature is selected at the present moment to create a node (Sakkaf, 2020).

3.3.5 Rule-based learning

The defining characteristic of a rule-based machine learner is the identification and utilization of a set of relational rules that collectively represent the knowledge captured by the system. The goal is to construct the smallest rule-set that is consistent with the training data. A large number of learned rules is usually a sign that the learning algorithm is attempting to 'remember' the training set, instead of discovering the assumptions that are present in the data (Kotsiantis, 2007).

3.3.6 Neural networks

In a neural network, information is transmitted in a single direction through a network, where each layer is connected to its neighbours, from the input to the output layers. A neural network consists of at least three neurons joined together in a pattern of connections. A neuron takes a group of weighted inputs, applies a function, and returns an output. Neurons are usually divided into three classes: input neurons, which receive information; output neurons, where the results of the processed information are found; and neurons in between known as hidden neurons. Hidden units are mini-functions with unique parameters that must be learned (Ruder, 2016).

A neuron receives input and based on that input, fires off an output that is used by another neuron. The neural network simulates this behaviour in learning about the collected data and then predicting outcomes.

3.3.7 Naive Bayes

Naive Bayes is a classification algorithm for binary and multi-class classification problems. The algorithm is called naive because the calculation of the probabilities for each hypothesis is simplified to make the calculations tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value (Brownlee, 2016). The Bayes theorem can be seen in Equation 1.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad 1.$$

P(A|B) is the probability of hypothesis A given the data B. P(B|A) is the probability of the data B given that hypothesis A was true. P(A) is the probability of hypothesis A being true. P(B) is the probability of the data.

3.3.8 K-Nearest Neighbour (k-NN)

K-Nearest Neighbour uses the entire data set as the training set, rather than splitting the data into a training set and a test set. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours.

3.3.9 Support Vector Machines (SVM)

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. Hyperplanes are decision boundaries that help classify the data points. Data points that fall on either side of the hyperplane can be attributed to different classes. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, the maximum of the margin of the classifier can be decided. In the support vector machines algorithm, the goal is to maximize the margin between the data points and the hyperplane (Ghandi, 2018).

3.4 Validation

In the following sections, the ways to validate and test the machine learning algorithms will be explained. These include 1) Division of datasets, 2) Confusion matrix, 3) ROC curve, 4) K-fold cross-validation.

3.4.1 Division of datasets

To be able to validate and evaluate the machine learning algorithms the data provided by K&R Consultants must first be divided into different sets, a train, validation, and a test set. This is visualized in Figure 7. The training dataset is used to train the model. The model sees and learns from the data, it tries to find patterns between the input and the output of the data, in the case of K&R Consultants the output would be the labels. Next, the validation set is used to evaluate a given model and select the best model. After the selection of the best model, the accuracy of the best model should be measured. To measure the accuracy the test set is introduced. The test dataset is used to provide an evaluation of the final model.

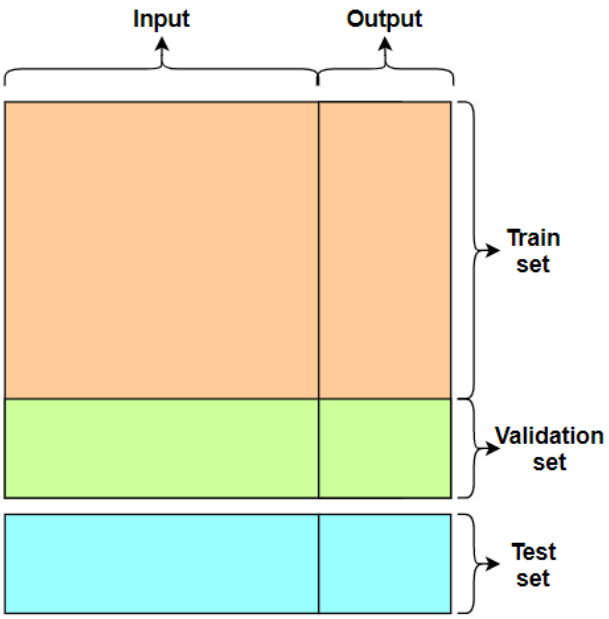

Figure 7. Division of dataset

3.4.2 Confusion matrix

A confusion matrix is a form of a contingency table showing the differences between true and predicted classes for a set of labelled examples (Bradley, 1997). It is a technique for summarizing the performance of a classification algorithm. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of error it is making.

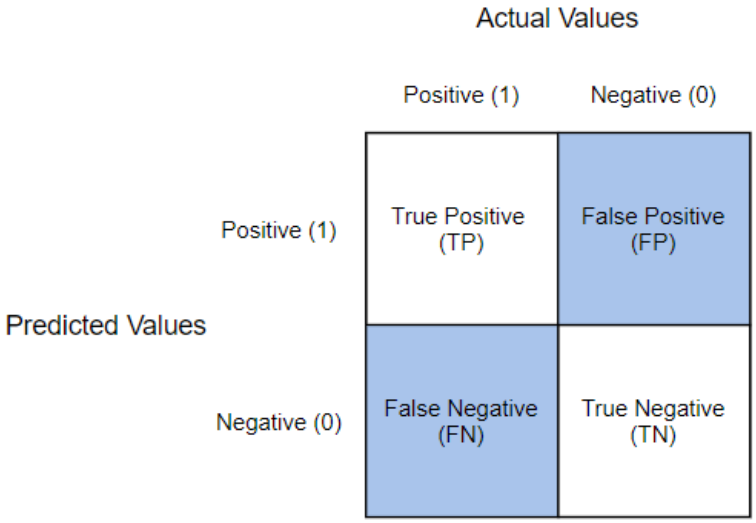

Figure 8. Confusion matrix (Narkhede, 2018)

In Figure 8, an example of a confusion matrix can be seen. In this example, there are only two possible states: positive or negative. If the machine learning algorithm predicts the future state to be positive and the state is indeed positive, the prediction falls into the category True Positive (TP). In the case that the machine learning algorithm predicts the future state to be negative and the state is indeed negative, the prediction falls into the category True Negative (TN). Then there are two categories left

False Negative (FN) and False Positive (FP). If the prediction of the algorithm is positive but the state is negative, the prediction will fall into the category False Positive, and vice versa the prediction will fall into False Negative.

In the case of K&R Consultants, they are dealing with a multiclass confusion matrix. Unlike binary classification, there are no positive or negative classes here. For multiclass classification, the TP, TN, FP, and FN have to be found for each individual class. For example, the TP of the class Apples is 7. The TN of the class Apples is (2+3+2+1) = 8. The FP of the class Apples is (8+9) = 17. The FN of the class Apples is (1+3) = 4. This is visualized in Figure 9.

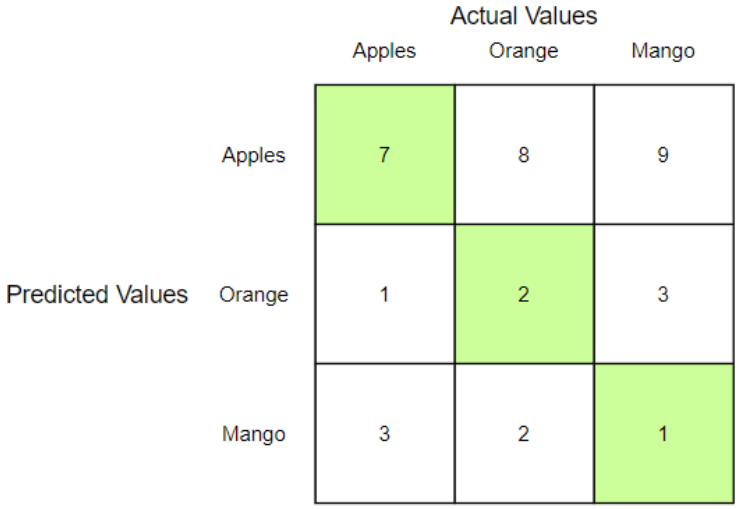

Figure 9. Multiclass confusion matrix

After the algorithm has been trained and validated it is applied to the test set. The predictions made by the algorithm could be either correct or incorrect. The accuracy of the algorithm can be calculated by dividing the number of correct predictions by the total number of predictions (Kotsiantis, 2007).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (2.)$$

However, it is not sufficient to only base the performance of the algorithm on the accuracy, since a high accuracy does not necessarily imply a good algorithm. The model will only have high predictive accuracy if the model is balanced and it will give low predictive accuracy if there is a class imbalance. A class imbalance is the problem of classification when there is an unequal distribution of classes in the training dataset, so there are a few labels that appear more frequently than others. To support the accuracy another measure of the test's accuracy is the F_β-score.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \quad (3.)$$

When choosing beta in your F-score you should decide how important recall is over precision. Precision is the proportion of correctly identified positives and recall is the proportions of actual positives that were identified correctly. For example, F1-score precision and recall are even important, with the F2-score recall is twice as important as precision.

$$Recall = \frac{TP}{TP + FN} \quad (4.)$$

$$Precision = \frac{TP}{TP + FP} \quad (5.)$$

The F1-score is the balance score of a machine learning algorithm. A high F1-score implies that the predictions are well balanced, while a low F1-score implies the opposite. The F1-score is best if there is some sort of balance between precision and recall in the system and is best if you have an uneven class distribution.

3.4.3 Receiver operating characteristic curve

A receiver operating characteristic (ROC) curve is a way to visualize the performance of a machine learning algorithm. The curve plots two parameters, True Positive Rate (TPR) and the False Positive Rate (FPR). The ROC Curve method plots a ROC curve for each code metric and then retrieves the optimal threshold value by maximizing the sum of the True Positive Rate and the False Positive Rate.

$$TPR = \frac{TP}{TP + FN} \quad 6.$$

$$FPR = \frac{FP}{FP + TN} \quad 7.$$

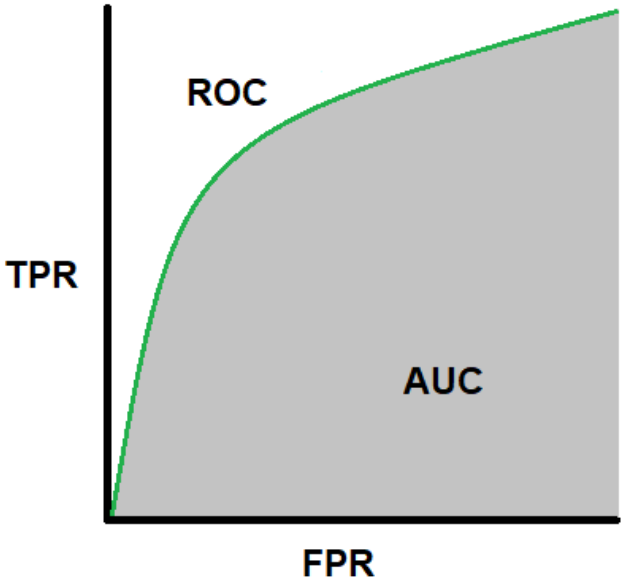

Figure 10. ROC curve (Narkhede, 2018)

An example of the ROC curve can be seen in Figure 10. The X-axis of the plot is mapping the FPR and the Y-axis is mapping the TPR, at various threshold settings. Each TPR and FPR are obtained from the confusion matrix using Equation 6 and 7. The threshold value retained is the one that maximizes both the TPR and the FPR. A test with perfect discrimination has a ROC curve that passes through the upper left corner. Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.

The Area Under the Curve (AUC) is a way to summarize the performance in a single number. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is classifying correctly (Narkhede, 2018).

3.4.4 K-fold cross-validation

Cross-validation is a statistical method used to estimate the performance of machine learning models. The method has a single parameter called k that refers to the number of groups that a given data sample is split into. With k-fold cross-validation, you divide the complete dataset you have into k

disjoint parts of the same size. After that, you train k different models on k-1 parts each while you test those models with the remaining part of the data. This can be seen in Figure 11.

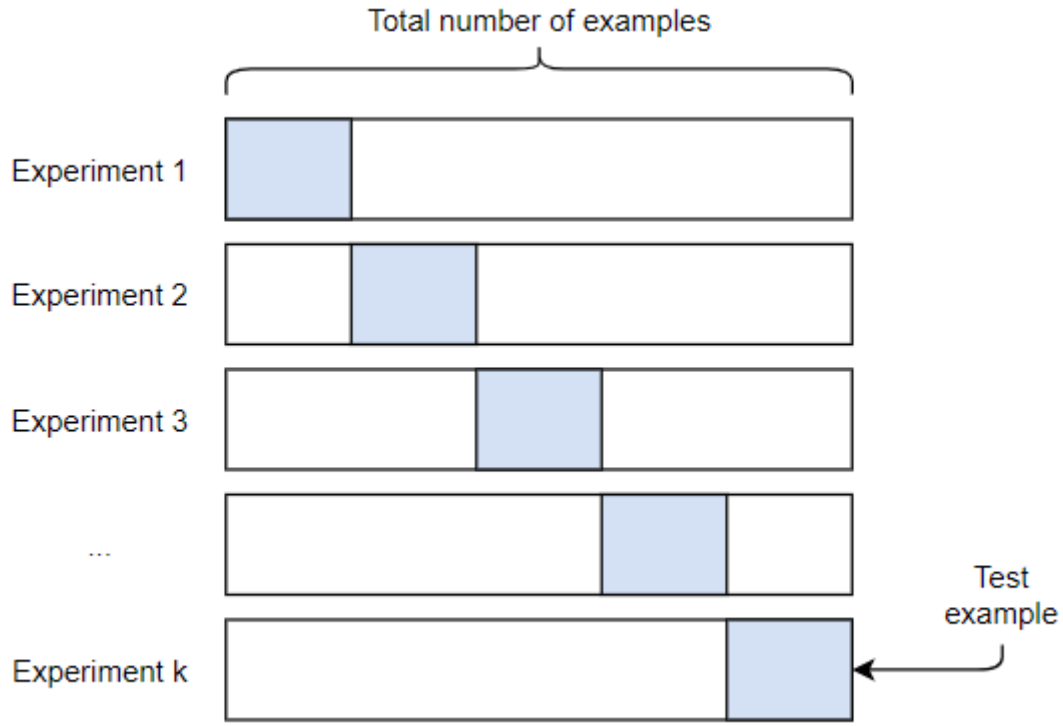

Figure 11. K-fold cross-validation (K-Fold Cross-Validation in Machine learning?, 2020)

This method is used to validate the future accuracy of a predictive model. From this test you will get multiple test errors, which allow you to calculate the average and a standard deviation, giving you a better idea about the range the actual model accuracy will likely be.

Stratified k-fold is a variation of k-fold which returns stratified folds which means that each set contains approximately the same percentage of samples of each target class as the complete set. This version of k-fold cross-validation preserves the imbalanced class distribution in each fold and will enforce the class distribution in each split of the data to match the distribution in the complete training dataset. In the case of class imbalances, it is useful to use stratified k-fold cross-validation, which ensures that the proportion of positive to negative examples found in the original distribution is respected in all the folds.

3.5 Machine learning algorithms evolution

Scientists design experiments and make observations and collect data from those observations. After that, they try to extract knowledge by finding simple models that explain the data they observed. However, we have now reached a point that such an analysis of data can take a long time when done by people. This is due to the fact of the datasets being huge, people who can do such an analysis are rare and manual analysis is very costly. This explains the reason why there is a growing interest in computer models that can analyse data and can extract the information automatically.

Data Mining (DM) has been identified as one of the most critical to transforming data into added value and new knowledge. It describes the search for hidden patterns or associations in data to aid understanding of systems and/or their processes. One of the main challenges of data mining is the lack of guidance in choosing the right analytics method for a given problem. It has been observed how several analyses of the same dataset can provide contradictory conclusions when analysed by two

independent data scientists without a common set of guidelines for conducting properly. The paper *Environmental Modelling & Software* provides an innovative methodology to help non-expert data scientist to identify DM methods suitable to properly analyse a certain kind of data when addressing a specific type of question (Gibert et al., (2018).

The final choice of technique depends on two parameters: the main goal of the target problem; and the structure of the available data set. The proposed methodology in *Environmental Modelling & Software* first introduces the DM methods conceptual map (DMMCM) as a reference decision map to browse through big groups of methods answering similar questions, and, at a second level, the DM methods templates (DMMTs), with more specific information about those groups of methods, helping to narrow down to a box of the DMMCM containing the suitable type of technique.

In September 2019, a patent was published which presents systems and methods for efficiently training a machine learning model (U.S. Patent No. 2019/0286943, 2019). The training data can be organized such that the initial state of the training data is relatively easy for a machine learning model to differentiate. Once trained on the initial training data, the training data is updated such that the differentiating becomes more difficult. By iteratively updating the difficulty of the training data and training the machine learning model on the updated training data, the speed that the machine learning model reaches the desired level and the accuracy is significantly improved.

A big part of machine learning is testing. The usage of the word testing concerning machine learning models is primarily used for testing the model performance in terms of accuracy or precision of the model. Moreover, it can be noted that the word testing means different for conventional software development and machine learning models development. Blackbox testing can be used to test the machine learning models from the quality assurance perspective. Blackbox testing of machine learning models is budding as a quality assurance approach that evaluates the model's functioning without internal knowledge. Machine learning development requires extensive use of data and algorithms that demand in-depth monitoring of functions not always known to the tester themselves. Therefore, techniques such as Blackbox and white box testing have been applied and quality control checks are performed on machine learning models. Blackbox testing is testing the functionality of an application without knowing the details of its implementation including internal program structure and data structures.

3.6 Conclusion

From the problem seen at K&R Consultants, it becomes clear that a way should be found to automate the manual analysis. From the literature, it becomes clear that this can be done by using machine learning. Because K&R Consultants make use of labels, supervised learning should be able to help K&R in the automation of their analysis. Next to this, there should be looked into classification algorithms, since the labels of K&R Consultants are discrete.

Furthermore, from this chapter, it becomes clear that there are machine learning algorithms that could be a match for K&R Consultants, but this should be analysed in further chapters to make sure that they are a fit.

There are a few ways to validate if the machine learning algorithm fit to the dataset of K&R Consultants. In the case of the confusion matrix, this should be done in a multiclass way to be able to fit this validation method to the data of K&R. Furthermore, also the ROC curve, k-fold cross-validation, and the f-score should be able to validate the machine learning algorithms.

4. Design and development

In this chapter, the reader is provided with 4.1) RapidMiner, 4.2) Data preparation, 4.3) Balance of the dataset, 4.4) Building predictive models, 4.5) Testing and training data, 4.6) K-fold cross-validation, 4.7) Selection of Algorithms, 4.8) Comparing Algorithms, 4.9) ROC curve, 4.10) Conclusion.

In this chapter, a way to analyse which machine learning algorithm fits the data of K&R Consultants best will be designed. This will be done by using the platform RapidMiner. Before the data can be analysed the data must be prepared, by removing missing values and checking whether there is a balance in the dataset. After this, the predictive models using the machine learning algorithms can be built in RapidMiner. Following this, the models should be tested and validated. This can be done by using confusion matrixes, k-fold cross-validation, the ROC curve, and f-scores.

4.1 RapidMiner

RapidMiner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It provides an advanced analytical solution through template-based frameworks that speed delivery and reduce errors by nearly eliminating the need to write code. It allows for data cleansing and transformation, algorithm selection and validation, and model deployment and optimization.

4.2 Data preparation

Before the analysis of the data can be done the data should first be prepared. The preparation starts with importing the data into RapidMiner. This data consists of the excel files K&R Consultants uses for the analysis of the budgets. In these Excel files, two columns are necessary for the analysis. These are the descriptions and the code of K&R, an example can be seen in Table 6.

Table 6. Example data used for analysis of budget

Description	Code K&R
Afsluiters DN 150	55.12.08
Leiding compensatoren DN 150	55.06.06
Circulatiepompen 26,3 l/s/ 175 kPA	55.11.09
Inregelafsluiters STA-F 150	55.13.08
Vul en aftapkranen	55.08.06
Thermometers	55.08.06
Ontluchtingspotten	55.08.06
Circulatiepomp 1,95 l/s - 125 kPA	55.11.04
Afsluiters DN 50	55.12.03
Inregleafsluiters sta-d 50	55.13.03
Multispilt spysteem NOC / Spreekkamer	55.01.01
CRAC units, 15 kW met ondensing unit	55.20.02
Koeltechnisch leidingwerk, geisoleerd koper l = 48 meter	55.05.06
flensenset DN 150	55.08.06
Afsluiters DN 150	55.12.08
Leiding compensatoren DN 150	55.06.06

In the dataset, there are some blank rows or some rows which do have a description but no code or vice versa. These rows are not of importance in the analysis, so the first step is to remove the missing values from the table.

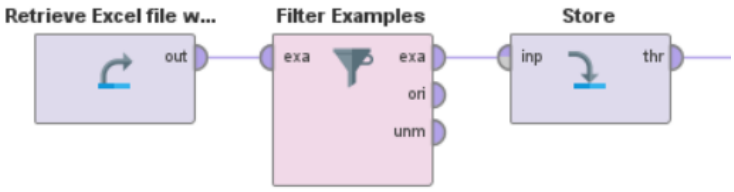

Figure 12. Process of filtering out missing values

The missing values will be filtered out with the Filter Examples operator, which can be seen in Figure 12. In the operator, the condition class can be defined, so that there will be no more missing attributes. By removing the missing values and the blank rows the analysis will take less time.

After the missing values are removed, the dataset is now referred to as an unlabelled dataset. However, training data with known labels are needed as input for supervised machine learning. Therefore, a labelled dataset needs to be created.

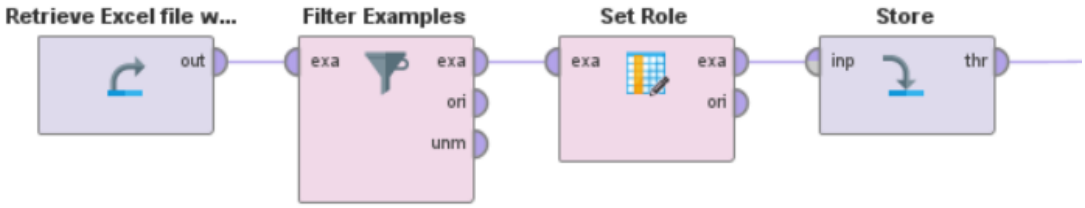

Figure 132. Process of giving a role to data

The process of creating a labelled dataset is almost identical to the process of making an unlabelled dataset. First, the missing attributes are removed using the Filter Examples operator. After that, an attribute can be given a role by the Set Role operator, which can be seen in Figure 13. The attribute Code K&R will be given the label role and after that, the file will be stored so that it can be used for the analysis.

4.3 Balance of the dataset

After doing preparing the dataset, the dataset should also be analysed on the balance. There is a total of around 2500 different codes in the database of K&R Consultants. Some of these codes occur more than others. This leads to an imbalance in the dataset. In Figure 14 the first twenty elements of the sample of 442 elements are shown. Here, one can already see that there is an imbalance in the dataset. The element that appears most often is 56.10.06 with an occurrence of 411 times.

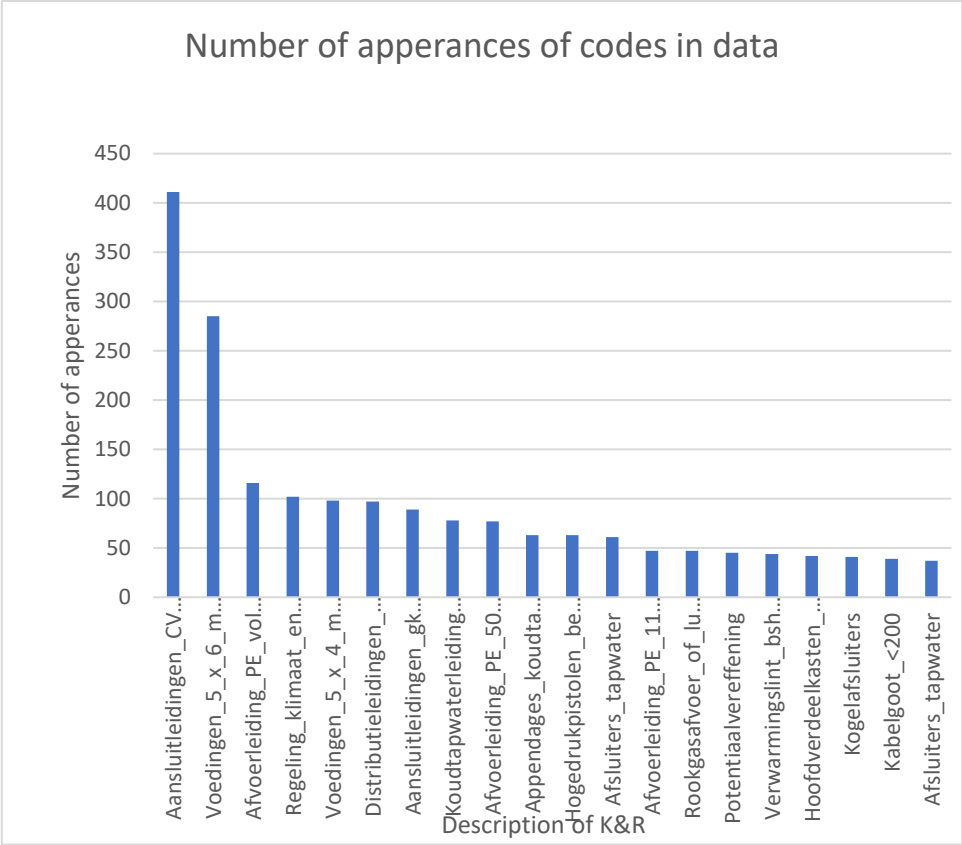

Figure 14. Histogram of number of appearances of codes in the data

In Figure 15 the sum of elements is shown which appears the least amount of times. As one can see 74 elements only appear once in the dataset. The elements who appear 20 times or less make up for around 87% of the dataset.

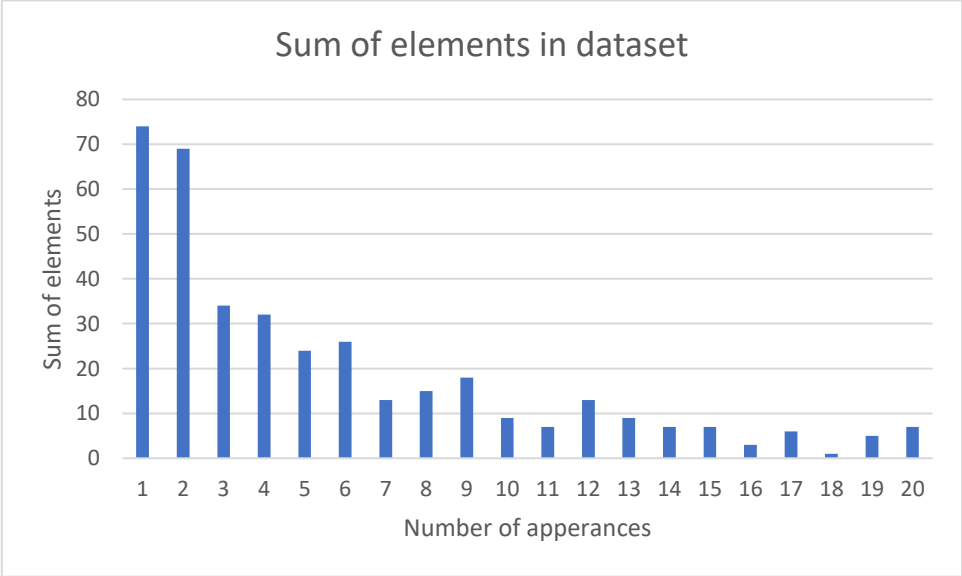

Figure 15. Histogram of elements which appear less frequently

To improve the balance in the dataset upsampling can be used. This way the difference between the minority class and the majority class will become less and the prediction of the machine learning algorithms will become more accurate. This is an efficient way to make the outcome of the algorithms

more accurate. However, the more upsampling is used the more time the analysis will take. A balance should be found in this process.

4.4 Building predictive models

Predictive Modelling is a set of machine learning techniques that search for patterns in big data sets and use those patterns to create predictions for new situations. Those predictions can be categorical (classification learning) or numerical (regression learning). These types of models are also used if you want to understand more about the underlying processes leading to certain outcomes.

RapidMiner can be used to find patterns and reveal insights in big data sets and consequently, the insights can be used to predict future outcomes. Using a model to generate predictions for new data points and find out what the most likely answer is, is called scoring. This process can be seen in Figure 16.

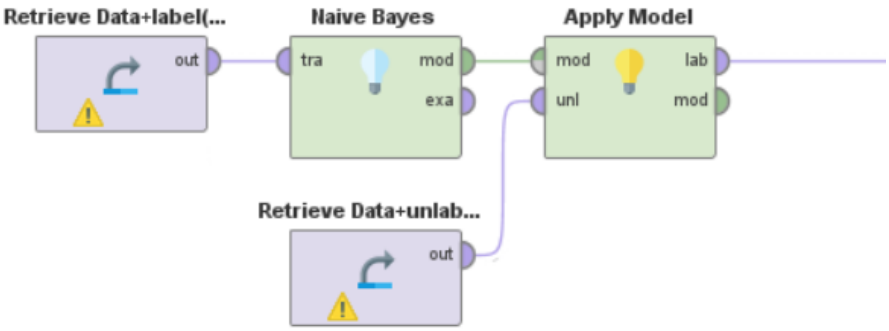

Figure 16. Scoring process in RapidMiner

The process shown in Figure 16 builds a Naive Bayes model. First of all, the labelled dataset is retrieved and is used as input for the Naive Bayes operator. Next, the Apply Model operator is used to create predictions for a new, unlabelled dataset. After, adding the dataset with unlabelled data the Apply Model operator takes the unlabelled data as an input, applies the Naive Bayes operator, and outputs a dataset with a label, which are the predictions made by the model. The result is the original unlabelled data with a column for the predicted outcome and additional columns for the confidences of the different codes of K&R. After every code received a confidence prediction the code with the highest confidence will be chosen as the prediction. A small part of this table can be seen in Table 7.

Table 7. Output of using Naive Bayes operator in scoring process in RapidMiner

Prediction(Code)	**Confidence(52.02.08)**	**Confidence(52.10.05)**	**Confidence(52.04.14)**
52.02.08	0.495	0.000	0.485
52.02.08	0.495	0.000	0.485
52.02.08	0.486	0.000	0.476
52.02.08	0.486	0.000	0.476
58.06.09	0.000	0.240	0.000
52.02.08	0.495	0.000	0.485
52.02.08	0.495	0.000	0.485
52.02.08	0.486	0.000	0.476
52.02.08	0.486	0.000	0.476
53.01.00	0.000	0.000	0.000
53.06.04	0.000	0.000	0.000
53.01.26	0.000	0.000	0.000
53.03.03	0.000	0.000	0.000

4.5 Testing and training data

After building predictive models a question that arises is "How well is this model going to perform"? The way to test this is to split the labelled data into test and training data. This way you can compare the predictions with the actual outcomes and calculate how often the model was right on these cases. The process of splitting the data into test and training sets can be seen in Figure 17.

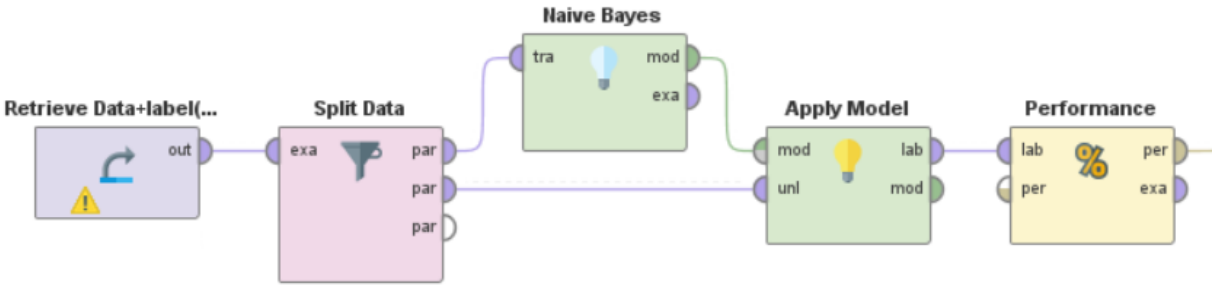

Figure 17. Testing and training process in RapidMiner

First of all, the labelled data is used as input for the Split Data operator. The Split Data operator takes a dataset and divides it into partitions. In this case, the data is divided into a partition of 70% and a partition of 30%. The 70% partitions will become the training set and the set where the model is built on. The remaining 30% will become the test set to which the models' predictions can be compared. The Performance operator can give different performance measurements, examples are the accuracy of the model, the weighted mean recall, and the weighted mean precision. Another output of the process is the confusion matrix. The confusion shows different kinds of errors. For example, the number of cases where have been predicted 'no' when they were 'yes'. The numbers on the diagonal are the right predictions. The predictions that are not on the diagonal are either false positives or false negatives. The confusion matrix can be seen in Table 8.

Table 8. Confusion matrix using Naive Bayes operator

	true 52.02.08	true 52.10.05	true 52.04.14	true 53.01.00	true 53.06.04	true 53.01.26
pred. 52.02.08	0	0	1	0	0	0
pred. 52.10.05	0	1	0	0	0	0
pred. 52.04.14	2	0	0	0	0	0
pred. 53.01.00	0	0	0	0	0	0
pred. 53.06.04	0	0	0	0	0	0
pred. 53.01.26	0	0	0	0	0	0

The Apply Model operator gives the following output which includes the prediction that the model makes and the confidences of all the codes. The code with the highest confidence will be chosen by the model. A small part of this table can be seen in Table 9. This shows the code of K&R and the prediction of the model applied in the process and the confidences of all predictions. In this table, you can check how the algorithm performs.

Table 9. Output of using Naive Bayes operator in testing and training process in RapidMiner

Code of K&R	**Predicted (Code)**	**Confidence(52.02.08)**	**Confidence(52.10.05)**	**Confidence(52.04.14)**
99.01.27	55.08.06	0.00123	0.00934	0.0129
57.18.04	57.18.04	0.00	0.00245	0.00
99.01.00	99.01.00	0.00	0.00312	0.00
99.01.00	55.05.06	0.00104	0.00790	0.00172
99.01.00	99.01.00	0.00	0.00142	0.00
99.01.00	99.01.00	0.00	0.00297	0.00
99.01.00	99.01.00	0.00	0.00322	0.00
99.01.00	99.01.00	0.00	0.00294	0.00
99.01.00	99.01.00	0.00	0.00240	0.00
55.05.06	55.05.06	0.00	0.00124	0.00
55.05.06	55.05.06	0.00129	0.00436	0.00
55.05.06	55.05.06	0.00	0.00175	0.00
55.05.06	55.05.06	0.00	0.00648	0.00
61.02.04	61.02.03	0.00	0.00185	0.00
61.02.03	52.03.11	0.00113	0.00855	0.00114
61.02.03	61.02.03	0.00	0.00342	0.00
61.08.20	61.08.20	0.00	0.00703	0.00

Table 9 can be converted into a table that has more value for K&R Consultants. This table can be seen in Table 10. Here the code which has to be predicted, the predicted code, and the three highest confidences are shown. Sometimes it can happen that the code which gets the highest confidence is not the right prediction, but for example, the code which gets the second-highest confidence is the right prediction. Therefore, it is very valuable to have the three highest confidences shown next to the prediction. For example, when looking at the fourth row of Table 10, you can see that the third-highest confidence is the right prediction. When implementing a machine learning algorithm into the analysis of K&R Consultants, it is therefore very valuable to create a dashboard that shows the top three highest confidences.

Table 10. Prediction of model with top three highest confidences

Code	**Predicted Code**	**Highest confidence**	**Second highest confidence**	**Third highest confidence**
52.06.07	56.11.01	56.11.01	53.01.17	52.06.01
52.06.07	56.11.01	56.11.01	53.01.17	52.06.01
52.06.07	56.11.01	56.11.01	53.01.17	52.06.01
52.06.08	52.10.05	52.10.05	52.06.01	52.06.08
52.06.08	52.10.05	52.10.05	52.06.01	52.06.08
52.06.09	52.06.09	52.06.09	52.06.02	52.10.05
52.02.08	52.04.14	52.04.14	52.02.08	52.10.05
52.02.08	52.04.14	52.04.14	52.02.08	52.10.05
52.02.08	52.04.14	52.04.14	52.02.08	52.10.05
52.10.05	58.06.09	58.06.09	52.10.05	52.01.15
52.04.14	52.04.14	52.04.14	52.02.08	52.10.05
52.04.14	52.04.14	52.04.14	52.02.08	52.10.05

4.6 K-fold cross-validation

K-fold cross-validation divides the dataset into equal parts and rotates through all parts, using one for testing and all others for the training of the model. In the end, the average of all testing accuracies is delivered as result. The Cross-Validation operator can split the data into for example ten different parts, so this can be called 10-fold cross-validation. This process can be seen in Figure 18.

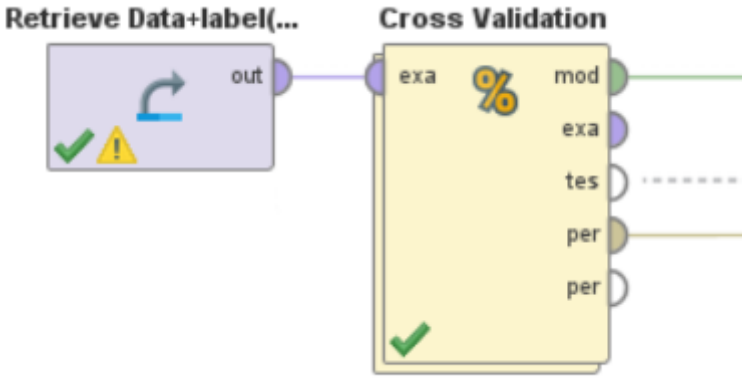

Figure 18. Cross-validation process

The Cross-Validation operator has two sub-processes, one for training the model and one for testing it. This can be seen in Figure 19.

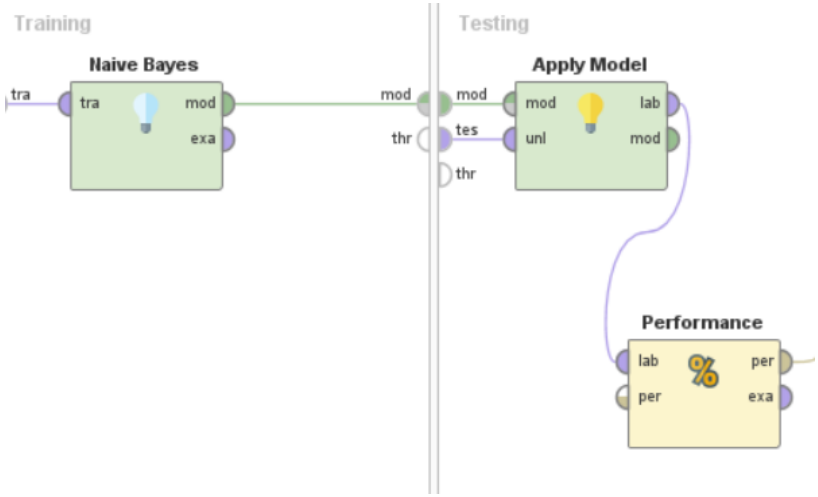

Figure 19. Sub-processes operator Cross-Validation

The Performance operator will output the accuracy. However, the accuracy now has an additional number indicating the standard deviation of the performances from the cross-validation. The standard deviation gives an idea of how robust the model is, the smaller the standard deviation, the less dependent the model performance is on the test dataset. For example accuracy: 44.52% +/- 1.43%.

Before the k-fold cross-validation can be conducted the value for k should be decided first. A poorly chosen value for k may result in a misrepresentative idea of the skill of the model, such as a score with high variance, or a high bias. As k gets larger, the difference in size between the training set and the resampling subsets gets smaller. As this difference decreases, the bias of the technique becomes smaller.

4.7 Selection of Algorithms

After constructing the processes in RapidMiner the data is analysed on different machine learning algorithms. However, both the Neural Networks algorithm and the Support Vector Machine algorithm did not support polynomial labels in RapidMiner. This is the reason why these two algorithms are not included in the analysis. The rule-based learning algorithm was first used in the analysis however the processing time of the algorithm was already longer compared to the other algorithms with less data than the data used for the final analysis.

4.8 Comparing Algorithms

A Key Performance Indicator is a measurable variable that makes it possible to measure performance. For this research, the KPI's accuracy score and F-score are used to compare the machine learning algorithms.

4.8.1 Accuracy scores

There are a lot of codes in the database of K&R Consultants, some of which are used very frequently and some which are used very rarely. Since accuracy is most valuable for K&R consultants, it was decided to concentrate on the accuracy score. The balance score is of less importance but is also taken into account. This way there will still be looked into if the codes that are used less frequently are predicted correctly.

4.8.2 F-score

The F-score can be calculated with the recall and precision. The recall and precision are calculated for each class included in the analysis. However, since there are a lot of classes included in the analysis the weighted mean of the recall and the precision are calculated. These weighted means can be used to calculate the f-score. In Table 10 one can be seen how the precision and recall are calculated for each separate class and how after that the weighted means are computed.

4.9 ROC curve

The ROC Curve method plots a ROC curve for each code metric and then retrieves the optimal threshold value by maximizing the sum of the True Positive Rate and the False Positive Rate. These rates can be calculated by looking at a confusion matrix. However, with a multiclass machine learning model, there is a slight difference as can be seen in Figure 9. The TP, FN, FP, TN can be calculated for every class and after that, the True Positive Rate and the False Positive Rate can be calculated for every class. After the FPR and TPR of all classes are calculated the average can be taken to create the ROC curve.

4.10 Speed of process

To measure the impact the machine learning algorithm will have on the labelling process, two situations must be compared. In the first situation, the employees of K&R Consultants should label the codes manually and in the second situation, the analysis with the machine learning algorithm included should be executed to label the codes. The time it takes to label all the codes in one analysis should be measured and compared in both situations.

4.11 Conclusion

In short, before the analysis can be done the data should be prepared. The data is useful for the analysis if there are no more blanc rows and missing values. Moreover, the data should be assessed on the balance of how many times every element appears. An imbalance in the data could lead to a lower accuracy score. After this, predictive models can be built. These predictive models can reveal insights about the data and also shows if the algorithm is suitable for the data. Next to this, is the testing and validating of the algorithms. The k-fold cross-validation can be used to see the bias in the algorithm.

The f-score can be used to calculate the accuracy of the algorithm. The ROC curve shows the performance of the model. For multiclass classification, the TPR and FPR have to be found differently. These first have to be calculated for each separate class and afterward, the average can be taken. The last step of the analysis is to compare the scores of all the algorithms with each other. To in the end be able to measure the reduction in speed of the analysis a comparison should be made.

5. Implementation & Demonstration

5.1 Results

After analysing the different machine learning algorithms the following results were realized, which can be seen in Table 11.

Table 11. Results of machine learning algorithms

Algorithm	Ac (%)	Fs (%)
Decision trees	2.99	0.02
Random Forest	2.99	0.02
Gradient Boosted Trees	77.96	83.63
ID3	78.42	83.71
Naive Bayes	78.56	83.74
k-Nearest Neighbour	68.55	73.14

From Table 11, it becomes clear that the Naive Bayes algorithm performs best in terms of accuracy and terms of F-score. This can be seen since the Naive Bayes algorithm has the highest percentages. The higher the percentage the better the algorithm is performing. Since the accuracy score is the most important score for the data of K&R Consultants the Naive Bayes algorithm would be considered the best algorithm. However, the differences are minimal in the top three algorithms. This shows that also the Gradient Boosted Trees or the ID3 algorithm could be a good match to use with the data of K&R Consultants.

The reasons for the Decision tree and Random Forest algorithm not working is because you need very clear decisions points in the data. These decision points are not clear in the database of K&R Consultants.

5.2 K-fold cross-validation

In this section, the four best-performing algorithms will be compared based on the k-fold cross-validation. The algorithms will be tested on different k's ranging from two up to fifteen. When increasing the k-fold, the smaller test set you get. When using k is equal to fifteen, the total dataset that is being used for testing is 1/15. The higher the k value the more data is available for the training of the dataset. This will in return often lead to higher accuracy.

In Figure 20 the analysis on the k-fold cross-validation of the Naive Bayes algorithm is shown. The value of k in this analysis is ranging from two up to fifteen. As one can see the accuracy does not follow an increasingly straight line. However, when looking at the trendline the accuracy does increase when having a higher k value. However, there is not an extensive change in the accuracy of the algorithm when you increase the k value. Because the accuracy does not fluctuate too much from three to fifteen, this shows that there is very little bias in the algorithm and that it is a clear algorithm. Therefore, for the Naive Bayes algorithm, it would be good to go with a k-fold of 7, because you would save on time and computations and therefore can optimise those.

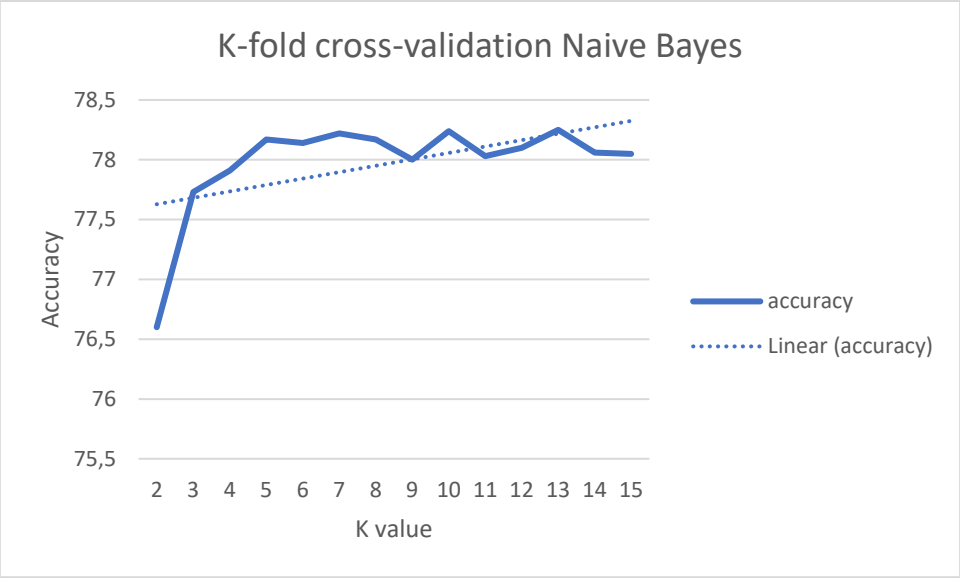

Figure 20. Histogram of k-fold cross-validation analysis of Naive Bayes algorithm

In Figure 21 the k-fold cross-validation of the k-Nearest Neighbour can be seen. Here you can see that the accuracy of the algorithm steadily increases till k has a value of eleven. From here on up the accuracy does not increase anymore and stays on about the same level. Therefore, the k value to use with a k-Nearest Neighbour on this data would be a value of around eleven, since you would save on time and computations and the accuracy is almost the same as a k value of fifteen.

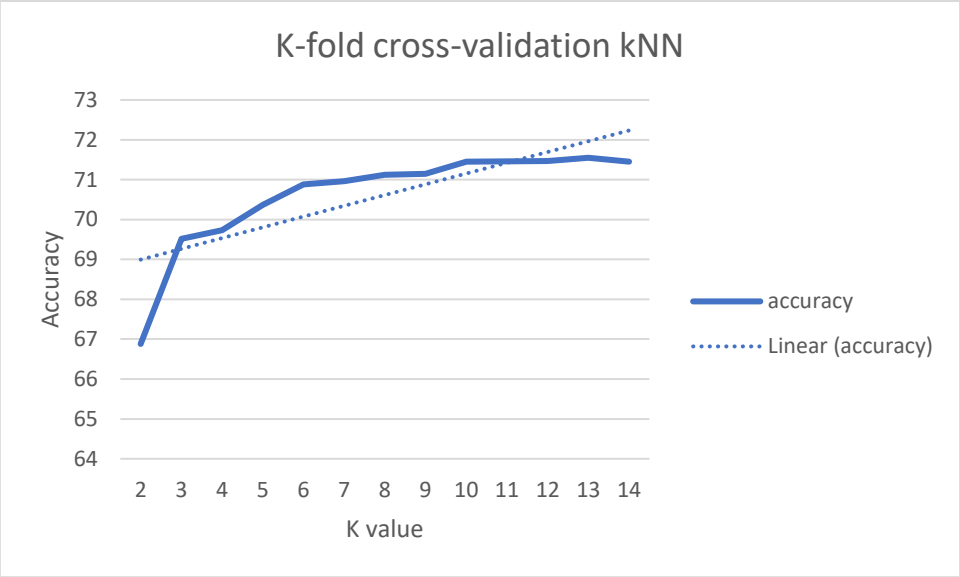

Figure 21. Histogram of k-fold cross-validation analysis of k-Nearest Neighbour

In Figure 22 the k-fold cross-validation of the ID3 algorithm can be seen. Here you can see that the accuracy of the algorithm steadily increases till k has a value of ten. From here on up the accuracy does not get higher than the accuracy at a k value of ten. Therefore, the k value to use with the ID3 algorithm on this data would be a value of around ten. This would save you on time and computations compared to a k value of fifteen.

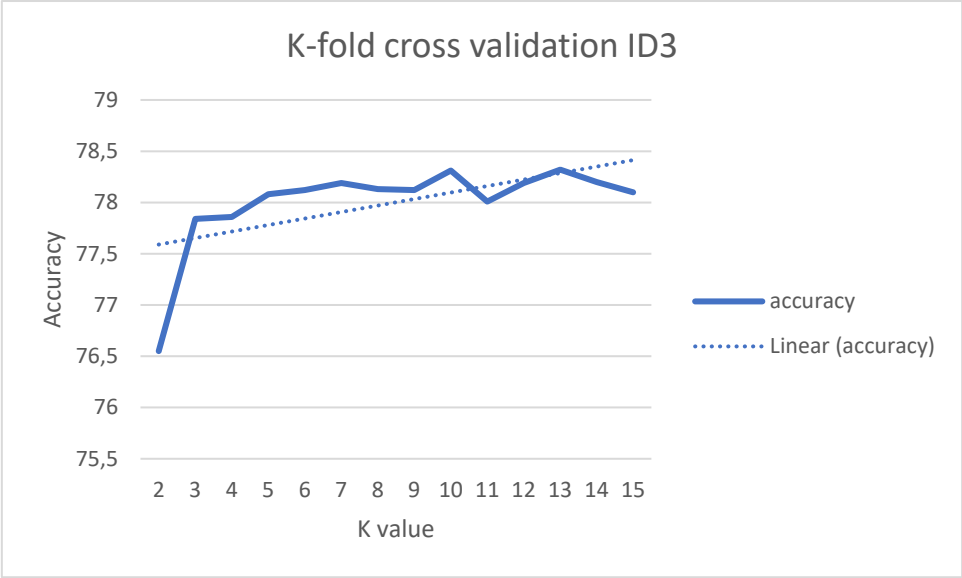

Figure 22. Histogram of k-fold cross-validation analysis of ID3

In Figure 23 the k-fold cross-validation of the Gradient Boosted Trees algorithm can be seen. The accuracy does not follow an increasingly straight line. However, there is not an extensive change in the accuracy of the algorithm when you increase the k value. Therefore, it means that there is very little bias in the algorithm and that it is a clear algorithm. For the Gradient Boosted Trees algorithm, it is good to go with a k-fold of 10, because then you would save on time and computations.

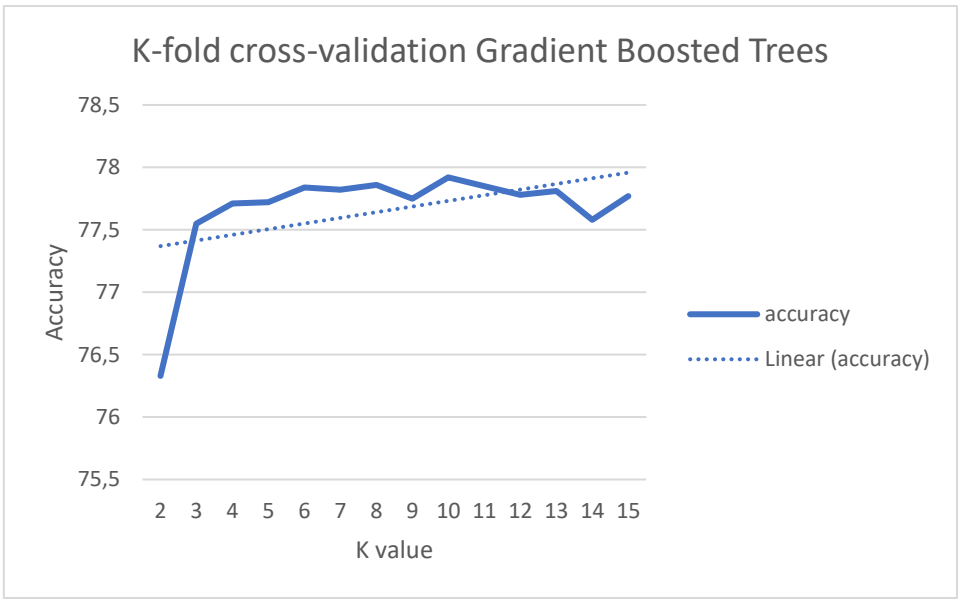

Figure 23. Histogram of k-fold cross-validation analysis of Gradient Boosted Trees

After analysing the four highest-scoring machine learning algorithms on the data of K&R Consultants a conclusion can be made. Table 11 shows the highest accuracy scores of the k-fold cross-validation and by which k-fold this accuracy score belongs. Where before Naive Bayes had the highest accuracy score, now the ID3 algorithm scores highest with the accuracy score.

Table 12. Comparison highest accuracy scores k-fold cross-validation

Algorithm	**K-fold**	**Accuracy**
Naive Bayes	13	78.25
k-NN	14	71.55
ID3	13	78.32
Gradient Boosted Trees	10	77.92

5.3 ROC curve

The ROC curve is not easy to create with the dataset of K&R Consultants. This is because first of all the dataset is a multiclass dataset. This means that there are a lot of possibilities to choose from for the algorithm to make a decision. Second of all, because there is uncertainty. Certain classes are more interdependent than others. This can be seen in Table 11, where most of the false positives come from a single class and not from the others. This shows that the error is not uniformly distributed in the matrix. In these cases, the ROC curve will not give you specific information that you want to look at in the case of validation of the algorithm.

Table 13. Multiclass Confusion Matrix using the Naive Bayes operator

	true 52.02.08	true 52.10.05	true 52.04.14	true 53.01.00	true 53.06.04	true 53.01.26
pred. 52.02.08	59	0	15	0	0	0
pred. 52.10.05	0	5	0	0	0	0
pred. 52.04.14	0	0	19	0	0	0
pred. 53.01.00	0	0	0	5	0	0
pred. 53.06.04	0	0	0	0	4	0
pred. 53.01.26	0	0	0	0	0	3

The best way to deal with these kinds of errors is to plot two pairs at the same time to create a binomial dataset and plot the ROC curve for these pairs. However, using a single classifier to predict every class would not be a valid representation of the data. ROC curves were considered in this theses, however, due to the reasons mentioned above, they will not be used to validate the algorithms.

5.4 Advantages and disadvantages algorithms

After the four best performing machine learning algorithms were identified, the advantages and disadvantages of all four algorithms will be discussed.

5.4.1 Naive Bayes

The advantage of using the Naive Bayes algorithm is that the algorithm may need a relatively small dataset for maximum prediction accuracy. Next to this, the algorithm requires little storage space

during both the training and classification stages. Furthermore, the algorithm is very transparent, as it is easily grasped by users.

The disadvantage of using the Naive Bayes algorithm is that the algorithm has high bias since it assumes that the dataset under consideration can be summarized by a single probability distribution and that this model is sufficient to discriminate between classes (Kotsiantis, 2007).

5.4.2 k-Nearest Neighbour

The advantage of using the k-Nearest Neighbour algorithm is that a basic k-NN usually only has a single parameter (k) which is relatively easy to tune.

The disadvantage of using the k-Nearest Neighbour algorithm is that the algorithm is very sensitive to irrelevant features. The algorithm is also considered as intolerant of noise. The similarity measures can be easily distorted by errors in attribute values, leading it to misclassify a new instance on the basis of the wrong nearest neighbour (Kotsiantis, 2007).

5.4.3 Gradient boosted trees

The advantage of using the Gradient Boosted Trees algorithm is that the algorithm has lots of flexibility. The algorithm can optimize different loss functions and provides several hyperparameter tuning options that make the function fit very flexible. Next to this, the pre-processing of data is not necessarily needed, since the algorithm works well with categorical and numerical values and the algorithm can also handle missing values.

The disadvantage of using the gradient boosted trees algorithm is that using the algorithm is computationally expensive. The gradient boosted trees algorithm often requires many trees which can be time and memory exhaustive. Next to this, gradient boosted trees will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting (Boehmke, 2020).

5.4.4 ID3

The advantage of using the ID3 algorithm is that this algorithm is considered a very simple decision tree. The algorithm builds very fast and short trees. It only needs to test enough attributes until all data is classified. Next to this, understandable prediction rules are created form the training data.

The disadvantages of using the ID3 algorithm is that data may be over-fitted or over-classified, if only a small sample is tested. Next to this, the algorithm does not handle numeric attributes, missing values, and is susceptible to outliers (Singh & Gupta, 2014).

6. Recommendations & Conclusion

Based on the previous chapters recommendation and conclusions can be made.

6.1 Conclusions

6.1.1 Problem

The problem K&R Consultants is currently facing is that the process of making the budgets comparable to the database K&R Consultants is using takes a lot of time and is mostly manual. After it was established that this was the problem that K&R Consultants is facing, the manual analysis was analysed. The manual analysis of K&R consist of a couple of steps and these steps are in detail explained in Chapter 2. After this, a problem cluster was created to get more insight into the causality between the relevant problems. Concluding from the problem cluster the core problem could be found. The core problem is that the labelling of the elements has to be done manually.

6.1.2 Research question

After the establishment of the core problem, a research question could be formulated. The research questions were formulated as follows: *How can the application of machine learning algorithms help K&R Consultants to automate the labelling process and in turn decrease the lead time of the process?*

This research question was divided into smaller sub-questions. The sub-questions can be found in Chapter 6.1.4.

6.1.3 Answering the Research Question

This study aimed to solve the core problem. As described in Chapter 1, the core problem was that K&R Consultants do the labelling process manually.

To answer the research question, the sub-questions have been answered in the thesis. From this six machine learning algorithms were selected and tested. The result show, given the data from K&R Consultants, that four algorithms would be good to build predictive models. The accuracy and F-scores in Table 14 showed that there is still a slight imbalance in the data, but it gives reason to believe that the labelling process of K&R Consultants can be automated using machine learning algorithms.

Table 14. Results of analysis

Algorithm	Ac (%)	Fs (%)
Gradient Boosted Trees	77.96	83.63
ID3	78.42	83.71
Naive Bayes	78.56	83.74
k-Nearest Neighbour	68.55	73.14

Therefore, the results show that the application of machine learning algorithms can help the automation of the labelling process. After the experiment described in Section 4.10 is conducted the conclusion can be made if the inclusion of the machine learning algorithm decreased the lead time of the process. However, in all likelihood, this will be the case.

6.1.4 Answering the Sub Questions

In Chapter 1, the main research questions were divided into 4 sub-questions to help answer the main question.

The first sub-question was:

***1:** How is the manual analysis of price lists at K&R Consultants currently done?*

To answer this sub-question, an analysis at K&R Consultants was done. This leads to a process with a couple of steps which can be seen in Figure 23.

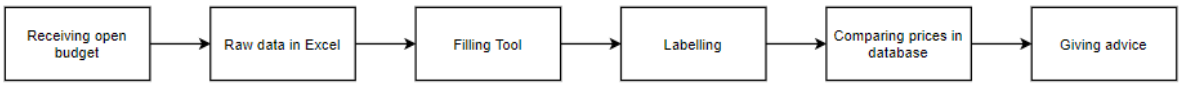

Figure 24. K&R Analysis steps

After this, the point of improvement in the analysis was mentioned. The first point of improvement is the conversion from PDF to Excel, there are a lot of problems with the layout that come up. Due to this problem, a lot has to be done by hand to fix the problem and to make sure that the problems that occurred in this step do not cause any problems later on in the analysis. The second point of improvement is the labelling phase, since K&R does this all by hand this takes the most time because they have to go over each row is this very repetitive and prone to mistakes.

The second sub-question was:

***2:** What are the different machine learning algorithms?*

To answer this sub-question, a literature study was done. The following supervised machine learning algorithms are available: decision trees, random forest, gradient boosted trees, ID3, Rule-based learning, neural networks, Naive Bayes, k-Nearest Neighbour, and Support Vector Machines.

There are a few ways to evaluate machine learning algorithms. One of them is the confusion matrix. From the confusion matrix, both the accuracy score and the f-score can be determined. Next to this, the ROC curve can be used to validate the machine learning algorithms. However, the ROC curve is not used to validate the algorithm with the data of K&R Consultants, since there are too many uncertainties. At last, k-fold cross-validation was also used to validate the machine learning algorithms.

The third sub-question was:

***3:** How can RapidMiner be used to analyse the machine learning algorithms?*

To be able to analyse the machine learning algorithms on the data of K&R Consultants the data should first be prepared. The preparation includes the removal of missing values and blanc rows. Furthermore, the data should be assessed on the balance of the occurrence of different elements. An imbalance in the data could lead to a lower accuracy score. Secondly, predictive models can be built. After this, the predictive models should be tested and validated. This can be done by using k-fold cross-validation, the f-score, and the ROC score. These validation measures can be used in RapidMiner by dragging different operators to the predictive models and then altering the settings to use them to validate the machine learning algorithms. Concluding the scores of all validation measures should be compared to each other.

The fourth sub-question was:

***4:** What machine learning algorithms work for K&R Consultants?*

To answer this sub-question, an analysis in RapidMiner was done. The detailed analysis can be found in Chapter 4 Design & Development. In the end, the following machine learning algorithms were tested on the data of K&R Consultants: decision trees, random forest, gradient boosted trees, ID3, Naive Bayes, and k-Nearest Neighbour. In the end, all the machine learning algorithms were compared to the validation measures. The numbers showed in Table 14 show that the Gradient Boosted Trees, ID3,

Naive Bayes, and as well as k-Nearest Neighbour algorithm could work for K&R Consultants. Next to this, all the algorithms have advantages and disadvantages. These could influence the decision for one of the algorithms.

Lastly, the fifth sub-question was:

***5:** What steps should K&R Consultants take to implement the machine learning algorithm and improve the decision making process?*

This sub-question is answered in Chapter 6.2 Recommendations and 6.3 Discussion.

6.1.5 Findings

After conducting the analysis it is clear that there are a couple of machine learning algorithms that could be suitable for the data of K&R Consultants. These are Naive Bayes, ID3, Gradient Boosted Trees, and the k-Nearest Neighbour. The Naive Bayes algorithm performed best on the analysis in RapidMiner with an accuracy of 78.56 percent. However, the Gradient Boosted Trees and the ID3 closely follow the Naive Bayes algorithm. Moreover, the ID3 algorithm performed best on the k-fold cross-validation analysis. So this means that one of these four algorithms can help K&R Consultants to automate the labelling process by implementing it into the analysis of K&R.

6.2 Recommendations

6.2.1 Application

After the analysis, it is recommended to use one of the top four algorithms for the application of the labelling process. The accuracy and F-score of the four highest-scoring algorithms show that there is potential in using machine learning to automate the labelling process of K&R Consultants. When choosing to create a dashboard or implement the algorithm in the analysis it is recommended to show the three highest predictions for every prediction. It occurs sometimes that the second prediction is the right prediction when showing the three highest predictions the right prediction could be picked more efficiently by an expert.

6.2.2 Continuous improvement

After the application has been programmed, a few pitfalls should be kept in mind.

Firstly, when new data is created not all the data should be added to the dataset. Adding the new data could result in an imbalance in the dataset or the data could contain errors. Therefore, an expert on the data should decide whether to add new data or not. If the data is updated, it is also recommended to execute the machine learning algorithm comparison again, since the updated data could lead to the finding of a more suitable algorithm to the data.

Next to this, also the priority of the new and old data should be assessed. Assume that one code occurred very frequently in the past. However, since things can change in the future, now another code should be selected. If old data is evenly important as the new data there is a possibility that the old code is predicted instead of the new code. To prevent this, the new data should be given a higher priority to the old data. This can be done by for example adding weights to the data. At last, the irrelevant data should be removed from the dataset.

6.2.3 Noise

There is a lot of noise in the description of the unique elements where the predictions of the labels are based on. This can lead to mispredicted labels. Therefore, it is recommended to let an expert look at the predictions and correct the predictions when not correct. After this, the algorithm should be updated, so the algorithm will predict the right label the next time. Next to that, to get the accuracy and f-score of the analysis higher there should be looked into how to remove the noise from the description in an effective way.

6.2.4 Speed of process

To be able to measure the impact the machine learning algorithm will have on the speed of the analysis of K&R Consultants the experiment described in Section 4.10 should be executed. This can be done after the machine learning algorithm is included in the analysis. After this, the two situations, with and without a machine learning algorithm, can be compared.

6.3 Discussion

6.3.1 Choice of algorithms

The machine learning algorithms applied in the comparison were determined through a literature study. It is, however, imaginable that a more suitable algorithm will be developed. For future research, it might be interesting to add this, possibly more accurate, machine learning algorithm to the comparison program.

6.3.2 Other algorithms

There are a lot of different codes in the database of K&R Consultants. This leads to some codes that appear very frequently and to codes that very sporadically. This causes an imbalance in the dataset. The upsampling of the codes that appear less frequently does help with the imbalance, however, there will always be a slight imbalance. Therefore, there is a possibility that when new data is added another algorithm is more favourable for that dataset. This can already be seen in the results that the Naive Bayes algorithm, the ID3 algorithm, and the Gradient Boosted Trees perform almost the same. Therefore, there could be more (new) algorithms that could be a good match for the dataset.

References

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). Massachusetts Institute of Technology.

Boehmke, B. (2020). *Gradient Boosting Machines · UC Business Analytics R Programming Guide*. Uc-r.github.io. Retrieved from http://uc-r.github.io/gbm_regression#proscons.

Bradley, A. (1997). *The use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms.* Elsevier.

Brownlee, J., (2016). *A Gentle Introduction To The Gradient Boosting Algorithm For Machine Learning*. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Brownlee, J., (2016). *Naive Bayes For Machine Learning*. Machine Learning Mastery. Retrieved from: https://machinelearningmastery.com/naive-bayes-for-machine-learning/

Calomme, V., (2019). *Introduction To Automated Machine Learning*. Mediaan. Retrieved from https://www.mediaan.com/mediaan-blog/automated-machine-learning

Expert System. (2020). *What Is Machine Learning? A Definition.* Retrieved from https://expertsystem.com/machine-learning-definition/#:~:text=Machine%20learning%20is%20an%20application,use%20it%20learn%20for%20themselves.

Gandhi, R., (2018). *Support Vector Machine — Introduction To Machine Learning Algorithms*. Medium. Retrieved from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Gibert, K., Izquierdo, J., Sànches-Marrè, M., Hamilton, S., Rodrìguez-Roda, I., Holmes, G. (2018). *Environmental Modelling & Software*. Elsevier.

Heerkens, H., van Winden, A. (2017). *Solving managerial problems systematically*. Noordhoff Uitgevers BV.

Kotsiantis, S. (2007). *Supervised machine learning: A review of classification techniques (31st ed., pp. 249-268). Informatica.*

Kulkarni, P. (2012). *Reinforcement and Systematic Machine Learning for Decision Making*. Institute of Electrical and Electronics Engineers.

Leskovec, J., Eksombatchai, C., Chen, K., He, R., Ying, R. (2019). U.S. Patent No. 2019/0286943. Washington, DC: U.S. Patent and Trademark Office.

MLTut. (2020). *K Fold Cross-Validation In Machine Learning? How Does K Fold Work?*. Retrieved from https://www.mltut.com/k-fold-cross-validation-in-machine-learning-how-does-k-fold-work/

Narkhede, S., (2018). *Understanding AUC - ROC Curve*. Medium. Retrieved from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Narkhede, S., (2018). *Understanding Confusion Matrix*. Medium. Retrieved from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Peffers, K., Tuunanen, T., Rothenberger, M., Chatterjee, S. (2007). *A Design Research Methodology for Information Systems Research*. Journal of Management Information Systems.

Ruder, S., (2016). *An Overview Of Gradient Descent Optimization Algorithms*. Sebastian Ruder. Retrieved from https://ruder.io/optimizing-gradient-descent/

Sakkaf, Y., (2020). *Decision Trees For Classification: ID3 Algorithm Explained*. Medium. Retrieved from https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1

Seif, G., (2018). *A Guide To Decision Trees For Machine Learning And Data Science*. Medium. Retrieved from https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956.

Singh, S., Gupta, P. (2014). *Comparitive study ID3, CART and C4.5 decision tree algorithm* (27th ed., pp. 97-103). IJAIST. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.685.4929&rep=rep1&type=pdf.

Sundaram, S., Ramasamy, S., Narasimhan, S. (2013). *Supervised Learning with Complex-valued Neural Networks*. Springer.

Yiu, T., (2019). *Understanding Random Forest*. Medium. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Zhang, Y. (2010). *New Advances in Machine Learning*. InTech.