Application of knowledge discovery in databases: Automating manual tasks

Biruk Yemane Habteselassie

University of Tampere
School of Information Sciences
Computer Sciences
M.Sc. thesis
Supervisor: Kati Iltanen
December 2016

University of Tampere
School of Information Sciences
Computer Sciences
Biruk Yemane Habteselassie: Application of knowledge discovery in databases: Automating manual tasks
M.Sc. thesis, 47 pages, and 1 index page
December 2016

Businesses have large data stored in databases and data warehouses that is beyond the scope of traditional analysis methods. Knowledge discovery in databases (KDD) has been applied to get insight from this large business data. In this study, I investigated the application of KDD to automate two manual tasks in a Finnish company that provides financial automation solutions. The objective of the study was to develop models from historical data and use the models to handle future transactions to minimize or omit the manual tasks.

Historical data about the manual tasks was extracted from the database. The data was prepared and three machine learning methods were used to develop classification models from the data. The three machine learning methods used are decision tree, Naïve Bayes, and k-nearest neighbor. The developed models were evaluated on test data.

The models were evaluated based on accuracy and prediction rate. Overall, decision tree had the highest accuracy while k-nearest neighbor has the highest prediction rate. However, there were significant differences in performance across datasets.

Overall, the results show that there are patterns in the data that can be used to automate the manual tasks. Due to time constraints data preparation was not done thoroughly. In future iterations, a better data preparation could result in a better result. Moreover, further study to determine the effect of type of transactions on modeling is required. It can be concluded that knowledge discovery methods and tools can be used to automate the manual tasks.

Key words and terms: Knowledge discovery in databases, data mining, business applications, automating manual tasks.

Contents

1. Introduction

Finding patterns and meaning from business data is an old practice done by business analysts. However, the practice of analyzing business data has changed with businesses adapting to information technology. Business transaction data are stored in large databases and data warehouses ready to be analyzed. Analyzing these data give competitive advantage for businesses. Moreover, advances in artificial intelligence have resulted in machine learning methods that automate the tedious process of discovering patterns in databases. These factors have changed how business data are analyzed. [Bose and Mahapatra, 2011]

Knowledge discovery in databases (KDD) is an interdisciplinary field that studies how to extract useful information (knowledge) from large data sets. KDD is an iterative process overseen by a human expert. The CRISP-DM (Cross Industry Standard Process for Data Mining) is one of the popular KDD process models and it consists of business understanding, data understanding, data preparation, modeling, evaluation, and deployment [North, 2012]. Data mining is one of the steps in the KDD process concerned with finding patterns from data. Data mining tasks include classification, clustering, association rule mining, regression, anomaly detection, and summarization [Fayyad *et al*., 1996]. Bose and Mahapatra [2011] have done a literature review of different data mining applications with application area, technique used, and problem type. KDD applications in business include financial data analysis, marketing, retail industry, fraud detection, telecommunication, manufacturing, and investment [Fayyad *et al*., 1996; Gheware *et al.*, 2014]. E-business mine Web data to improve their marketing and sales operation and provide personalized web service such as product recommendation [Ismail *et al.*, 2015; Nayak, 2002].

This thesis presents a study conducted in a Finnish company that provides financial automation solutions either outsourced or as a service. Moreover, a literature review of KDD, data mining and, application of KDD in business are presented. The study investigates possibilities to automate mostly repetitive manual tasks by applying KDD. Hundreds of clients outsource their financial business processes to the company and each client has its own instructions how to handle different transactions defined in a document called customer instruction. The customer instruction document is usually in MS Word or MS Excel format and it is updated by the customer whenever there is a change. However, handling a specific transaction is not a straightforward application of the customer instructions, rather it involves analyzing the transaction by taking into consideration the customer instructions, applicable Finnish law, domain knowledge and experience, and familiarity with the client business. Some cases of transactions are simple in nature and it is easy to handle them, on the other hand some transactions are complex and require

more analysis. The experts who handle the transactions do these manual tasks (of handling daily transactions by analyzing) repetitively. Taking into consideration the company have hundreds of clients, handling numerous transactions every day is very tedious and labor-intensive, time consuming and costly.

Earlier, a rule automation feature was developed in the current system that allows experts to add rules in If-Then format. The objective is to add to the system rules that are general enough and correct and help automate handling of transactions. To see how it works let's say an expert that handles the business process of a client notices that he is always handling a certain type of transaction in the same way and he can state it clearly as a rule in If-Then format. The rule he formulates is based on the application of customer instructions, applicable Finnish law, domain knowledge, and familiarity with client business. The expert can then add this rule to the rule automation feature. In the future, similar transactions will be handled automatically based on the rule without the interference of any human expert. However, the rule automation feature is not used extensively because the company personnel are hesitant to add rules to the rule automation system because they are not always sure if the current case they handled is general enough to be stated as a rule. Moreover, since the customer instructions change there is a need to change the derived rules added to the rules automating system and there is no one who can master all the customer instructions to make the necessary changes in the system. The current system is labor-intensive, taking much time and considerable effort, and resulting in higher cost and lower customer satisfaction. Automating the generation of rules and using the rules (by feeding them to the system) to handle future transactions will result in considerable reduction of cost, time, and effort.

The company wanted to investigate if there are artificial intelligence (AI)-based solutions that can automate the manual tasks of handling daily transactions. In the company case, AI refers to a self-learning system that learns (generates) rules from historical data and based on the learnt rules predicts how to handle future transactions. Moreover, the self-learning system should update the learnt rules as customer instructions change. It was found that there is neither generic AI solution that can be easily customized nor a specific AI solution for this business problem. Therefore, KDD was applied to discover patterns and models that represent the rules from historical data. The objective was to assess if there were patterns in the data that can be used to automate the manual tasks. The scope of the study does not include deployment i.e. using the discovered patterns in the real-world business scenario. Historical data were extracted, and preprocessed, three machine learning methods (decision tree, Naïve Bayes and K-nearest neighbor (K-NN)) were used for data mining, and finally the models or patterns discovered were evaluated.

This thesis proceeds as follows. Next, a literature review of KDD, data mining, and applications of KDD in business is presented. Chapters 2 and 3 give an overview of knowledge discovery in databases and its process. In chapter 4 data mining tasks, machine learning algorithms used in data mining, and data mining tools used in business are discussed. Applications of KDD in business are presented in chapter 5. The methodology which includes data preparation, modeling, and evaluation methods is discussed in chapter 6. The results are presented and evaluated in chapter 7. Finally, the results are summarized and recommended solutions are suggested in chapter 8.

2. Knowledge discovery in databases and data mining

The art of finding useful patterns has been given a variety of names such as knowledge discovery, data mining, knowledge extraction, information archaeology, information harvesting, data archaeology, and data pattern process. Each of these terms has one thing in common, finding patterns from data. The term "knowledge discovery in databases" was coined by Gregory Piatetsky-Shapiro at the first KDD workshop in 1989 to emphasize knowledge as the end result. The term "data mining" was introduced to the database community around 1990. The term KDD became popular in the AI and machine learning communities whereas the term data mining was more popular in the business and press communities. [Fayyad *et al*., 1996]

Data mining and KDD are often used interchangeably because data mining is a key part of KDD [Bose and Mahapatra, 2011; Priyadharsini and Thanamani, 2014]. However, it is important to understand the difference between KDD and data mining. KDD refers to the overall process of discovering useful patterns (knowledge) from data. The basic steps of the KDD process include understanding of domain knowledge and identifying goal of the discovery process, data selection, pre-processing, data transformation, data mining (pattern searching), interpretation/evaluation of patterns, and deployment of the discovered patterns. Data mining is used at the pattern discovery step of the KDD process. [Fayyad *et al.*,1996; Fu,1997]

One popular definition of KDD states that *"KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."* [Fayyad *et al.*,1996, pp. 40]. In this definition, "*data"* refers to recorded facts such as records in a database and "*pattern"* refers to a high-level description of a set of data which can be fitting a model or finding some structure. "*Process"* implies that KDD is an iterative process. Moreover, the patterns must be nontrivial, valid, useful and novel. By non-trivial, we mean that the pattern should not be a direct computation such as average or summary but rather it should involve some search and inference. When applied on new data, the patterns should be valid with reliable certainty. Moreover, the patterns should be useful, novel, and understandable by human. [Fayyad *et al*., 1996]

KDD is an interdisciplinary field that relies on other related fields such as statistics, pattern recognition, databases, AI, machine learning, and data visualization [Fayyad *et al*., 1996]. Data mining is sometimes mistakenly regarded as a subset of statistics but that is not realistic as data mining uses ideas, tools, and methods from other areas such as database technology and machine learning. Data mining extends traditional data analysis and statistical approaches by employing analytical techniques drawn from other fields. Classical statistical procedures such as logistic regression, discriminant analysis, and cluster analysis are used. Machine learning techniques used includes neural networks, decision

trees, genetic algorithms, inductive concept learning and, conceptual clustering. Database-oriented methods include attribute-oriented induction, iterative database scanning for frequent items, and attribute focusing. Basically, any method that helps to get more info about data can be used in data mining. [Jackson, 2002; Fu, 1997; Goebel and Gruenwald, 1999]

3. Knowledge discovery process

Although there are different variations of KDD process model the basic steps of the KDD process are described below using the Cross-Industry Standard Process for Data Mining (CRISP-DM), which was developed in 1999 to standardize the approach for data mining [North, 2012].

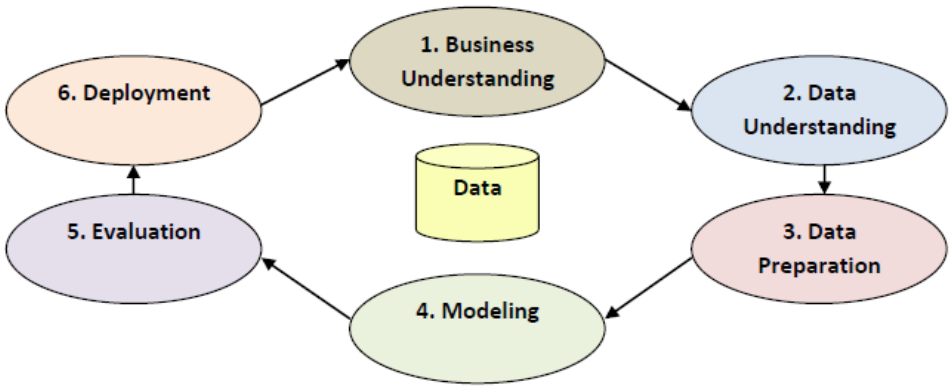

Figure 1. CRISP-DM conceptual model. [North, 2012]

Business understanding refers to clearly defining the problem we want to solve or the question we want to answer to [North, 2012]. Therefore, it is very important that the people who perform the KDD process have a good domain knowledge of the context in which the knowledge discovery process takes place [Tomar and Agarwal, 2014]. For example, it can be to understand our customers buying behavior which will be used to develop a new marketing campaign. Another example can be a bank applying KDD on the credit history data to find a pattern that will be used to predict if a new credit applicant will pay or default on a loan.

The data understanding step involves gathering, identifying, and understanding our data [North, 2012]. The data can be extracted from a data warehouse, transactional or operational database, or data marts [Jackson, 2002]. For example, this could mean importing the data from the data source in file formats such as CSV or facilitating direct access to the data source. This step defines our target data set on which we are going to do the analysis or knowledge discovery process.

Data preparation or preprocessing activities includes joining two or more datasets, reducing the dataset by selecting the attributes (columns) that are relevant, data cleaning (scrubbing) and, reformatting data for consistency [North, 2012]. Data cleaning refers to handling missing values and eliminating noisy data [Tomar and Agarwal, 2014]. Missing values are data that do not exist in the database commonly referred to as 'null ' in database

terminology. Depending on the nature of the data and the data mining objective records with missing values are either kept, filtered out (data reduction) or substituted with other value [North, 2012]. In addition, data transformation refers to transforming the data into a format that is suitable for the data mining techniques [Tomar and Agarwal, 2014].

Modeling is the most interesting step in which data mining is applied to find models. A model in data mining refers to a computerized representation of the real-world observation and it involves the application of algorithms to search, identify, and display patterns in the data [North, 2012]. The models and patterns give knowledge or insight to solve the problem stated at the first step of KDD i.e. business understanding. The nature of the business problem stated in the business understanding step should dictate the nature of the data mining task. Therefore, before doing the modeling we need to map the business problem with common data mining tasks. Common data mining tasks includes classification, clustering, regression, and association analysis.

Since the objective is to use the model we discovered it is necessary to estimate the accuracy of the model and in general to evaluate whether it is useful and interesting. A model or pattern discovered in the mining step needs to be validated whether it also applies to wider data sets. For example, a model that works well on the training examples might perform poor on the test data due to noise in the training data or small number of training examples (unrepresentative sample). In such as case, it is said the model *overfits* the training data [Mitchell, 1997]. The case of overfitting model shows that testing the accuracy of a model on training set data can give a highly-biased estimate of the models accuracy. To avoid this problem models are tested on a test dataset and their accuracy is estimated based on their performance on the test data. Statistical methods for estimating hypothesis accuracy are used to estimate the model's accuracy on additional examples from its observed accuracy on limited sample data [Mitchell, 1997]. One of the techniques used to present and compare classifiers is a ROC graph. It is a widely-used technique to visualize, organize and select classifiers based on their performance. It plots true positive rate (correct classification) on the Y-axis and false positive rate (misclassification) on the X-axis [Fawcett, 2004].

There are different methods used for evaluation. One approach is hold-out method which involves to randomly select examples from data set for test data and leave the rest for training set [Mitchell, 1997; Tan *et al.*, 2004]. Random sampling is done by repeating the hold-out method several times by selecting different test and training data sets to improve the estimation [Tan *et al.*, 2004]. The other method is cross-validation, where the data set is divided into k equal-sized (and disjoint) subsets and a model is built k times where each subset is only once used as a test data [Mitchell, 1997]. In bootstrap, unlike other

methods, records are selected for training with replacement, i.e. a record chosen for training is put back to the original data set so that it can be redrawn again, and records not included in the training dataset are used for test dataset [Tan *et al.*, 2004].

From the discussion of KDD process, we can see that the success of KDD application depends on doing each of the steps in the KDD process correctly. The deployment step is about using the knowledge and insight we get from the application of KDD. Activities done includes automating our model, integrating it with other existing systems, discussing its results with users of the model and improving the performance of the model based feedback from its use [North, 2012].

4. Overview of data mining

In the next sections data mining tasks and machine learning algorithms used in data mining and data mining tools used in business are discussed shortly.

4.1. Data mining tasks

There are mainly two types of data mining tasks, predictive and descriptive. In the case of predictive tasks, the discovered patterns are used to predict future unknown values. The discovered patterns can be used for predictive tasks, this enables organizations to make pro-active, knowledge-driven decisions and answer questions that were too time consuming previously [Ramamohan *et al*, 2012]. On the other hand, the aim of descriptive tasks is to find patterns that will be presented to a human user to get insight into data [Fayyad *et al*., 1996]. Common data mining tasks are categorized as follows [Fayyad *et al*., 1996; 11, 12]:

- *Classification* – It refers to a function that maps (classifies) a data item into predefined classes. A classic application is a bank using historical loan data to develop a model that classifies loans as good or bad. The bank can use the model when a new application for a loan is made to approve or reject the loan.
- *Prediction* - It uses a predictive model to predict unknown value of a quantitative attribute of a data item based on other given attributes. For example, using a predictive model of credit card transactions to predict the likelihood that a specific transaction is fraudulent.
- *Clustering* - It involves identifying finite set of categories or clusters to describe the data. The categories can be mutually exclusive or overlapping and hierarchical. A classic application of clustering is identifying subgroups of consumers in a marketing database.
- *Association rule mining* - It refers to discovering patterns that describe significant dependencies between variables. The discovered patterns show the togetherness or connection of objects. A common application is in market basket analysis where retailers use it to determine the buying behavior of customers. The rules describe what products are frequently bought together. The patterns (rules) can be used for cross-selling, which refers to selling additional product to existing customer [Radhakrishnan, 2013].
- *Regression* - It involves a function that maps a data item into a real-valued prediction variable. For example, let say we have a data of advertising expenditure and consumer demand for products, both numerical values. We can develop a linear regression model (mathematical function) which maps advertising expenditure into consumer demand and we can use this model (function) to predict the consumer demand for a new product given advertising expenditure.

- *Anomaly detection (deviation detection)* - It refers to methods used to detect significant difference from the previously recorded data or normative data. It has wide application in fraud detection such as credit card fraud and accounting fraud.
- *Summarization* – It is the abstraction or generalization of data to a smaller set which gives general overview of data. It involves finding methods that provide a compact description of a subset of data. A good example is determining the mean and standard deviation of a column.

4.2. Machine learning algorithms in data mining

As stated in the previous sections machine learning algorithms are extensively used in the data mining step of KDD. Machine learning is the study of computational methods to for example, automate the acquisition of knowledge from examples. The term examples in machine learning or data mining terminology refers to individual records in a data set. The data set used for training the learning algorithms is called training data. The discovered patterns will be then used, for example, for prediction on new unseen examples. Numerous machine learning algorithms exist to implement general data mining tasks such as the ones discussed in the previous section. The main categories of machine learning algorithms used in data mining are [Bose and Mahapatra, 2011, Jackson, 2002]:

- *Rule induction* involves creating a decision tree or rule set from a training data. The examples in the training data are labeled with known classes. In the first iteration of creating the decision tree the root node represents all examples in the training data. If the examples in the root node belong to two or more classes, then the attribute with most discriminating power is selected for further splitting the data set. The creation of decision tree is an iterative process of attribute selection and splitting until the examples in the leaf nodes (terminal nodes) consist of similar class.
- *Neural networks (NN)* are modeled after the human brain simulating the neurons. NN are a network of nodes which consists of input nodes that are connected to output nodes. In between the input nodes and the output nodes are hidden nodes. Each node receives input signal, transforms it and then transmits it to other nodes connected to it. Since the NN consists of layered network of nodes the classification logic is buried inside the network. The complex nature of NN makes them difficult to be understood by human users [Bose and Mahapatra, 2011].
- In *case-based reasoning (CBR)* representative examples (with known classes) are selected from the training data and stored in a case-base. A case stores a problem and its associated solution. On a new problem, the solution is provided by matching it with a stored case. Nearest neighbor matching algorithm is used. The advantage of CBR is it allows to use domain knowledge as human experts

can add and edit the case-base. However, its highly sensitive to noise and missing data. Moreover, lack of tool support makes it difficult to manage the case-base.

- *Genetic algorithms (GAs)* use search algorithms based on natural selection and evolution theory. The procedures are modeled upon the evolutionary biological processes of selection, reproduction, mutation, and survival of the fittest to search for very good solutions. The main operations in GA are selection, crossover, and mutation. Items (records) are selected for mating pool based on criteria of fitness, the crossover operation will change part of an item with corresponding part of another item to create a new item. Mutation is used rarely to add variety by changing part of an item. Advantage of GA are ability to handle noisy data and they require little domain knowledge making them ideal for integration with other systems.
- *Inductive logic programming (ILP)* uses first order predicate language to define concepts. The expressive power of predicate logic enables ILP to represent complex relationships and it also allows to easily represent domain knowledge. In addition, the models represented by predicate logic are easy to understand. However, the predictive accuracy of the system declines with new data and it is very sensitive to noisy data.

4.3. Data mining tools used in businesses

Data mining tools can be divided as traditional data mining tools, dashboards, and text mining tools. Traditional data mining tools enable to find patterns by applying different machine learning and data mining algorithms on data. Dashboards on the other hand monitor changes in the database and present them as a chart or table. Text mining tools can mine data from text sources such as MS Word, pdf, email, and simple text files [Ramamohan *et al*, 2012].

There are numerous open source and commercial data mining tools available. Organization can buy data mining tools or develop their own custom data mining tools [Ramamohan *et al*, 2012]. It is necessary to have some criteria to evaluate the tools. These are the important factors to take into consideration [Andronie and Crişan, 2010]:

- The size of the data to be analyzed. If the data is very large it requires a more powerful and expensive tool.
- The amount of preprocessing required to make the data ready for mining. Data stored in relational database is easier to mine where as a text data requires a tool that handles a text input.

- How the data is stored or what is the data source is an important factor. For example, for a data stored in a database a tool that works with databases is required. Otherwise, the data must be extracted from the database and this can be time-consuming, prone to errors, and data-security threat. On the other hand, if the data comes from a data stream a data mining tool that handles real time analysis is necessary.
- For a complex analysis, a specialized tool is required while for a simpler analysis an affordable tool can be used.
- Data mining tasks to be done such as association, clustering, classification, and regression determine what type of tool is needed.
- Future analysis needs must be taken into consideration and tools that support the future analysis needs should be selected.
- In the case of mining a data stored in a database, the coupling with the data base in use is of high importance to access internal functions of the database and results in efficiency.
- The availability of API interfaces. Data mining tools that provide API function libraries allow for the integration of data mining functions in the software that a company is already using in its day to day business operations. This is a great advantage since it eliminates the need to use different applications; one for daily business activities and another one (data mining tool) for data analysis.
- Scalability is necessary in the case of the database of the company extends and it becomes necessary to analyze large volumes of data.
- User friendliness is important because usually users of the data mining tools are not IT specialists. Visualization makes the results of the analysis to be understood by end users.

The most commonly used commercial and open source data mining tools are summarized in table 1 [Petre, 2013; Andronie and Crişan, 2010, 17].

Data mining tool	Features
WEKA	An open source tool that supports data pre-processing, classification, clustering, regression, visualization, and feature selection. Graphical user interface (GUI) makes it easy to setup and use.
RapidMiner	It provides data loading and transformation (ETL), data preprocessing and visualization, modelling, evaluation, and deployment procedures.
Orange	Free and open source component-based tool that supports data loading and transformation (ETL), data preprocessing and visualization, modelling, evaluation, and deployment.
SAS Enterprise Miner	Commercial product that supports decision trees, neural networks, regression and, association rule mining.
IBM SPSS Statistics	Commercial product that came from a statistical application. It provides decision tree and other algorithms.
IBM SPSS Modeller	Commercially available and it supports data mining and text analytics. It has an easy graphical user interface and it supports clustering, classification, association rules, and anomaly detection.
Microsoft SQL Server	It has OLAP, data mining and reporting tool. It provides classification, regression, clustering, and association
Oracle data mining	It embeds data mining techniques in the Oracle database. It provides classification, regression, anomaly detection, clustering, association models and feature extraction.
STATISTICA	It is a statistics and analytics software. It provides data mining, statistics and data visualization, data preprocessing and cleaning tools. It supports clustering, classification, regression, association, and sequence analysis.
KXEN	Provides algorithms such as regression, time series analysis, classification. Supports working with OLAP data cubes and can access data from spreadsheet such as MS Excel.

Table 1. Popular commercial and open source data mining tools. [Petre, 2013; Andronie and Daniel Crişan, 2010; Ramamohan *et al.*, 2012]

5. Knowledge discovery in business

Large amounts of data have been generated and accumulated in large databases and data warehouses. Extracting business knowledge from the data gives businesses a competitive advantage. However, analyzing this data is beyond the scope of the traditional methods of analysis used by business analysts. Advances in machine learning methods have enabled analyzing large databases, leading to different applications of KDD in business.

5.1. Applying KDD in business

This section briefly presents organizational and data acquisition issues, integrating data mining into business applications, role of domain knowledge, data and operational characteristics, and current trends.

Among the organizational issues faced is access to domain experts. Experts do not tend to share their expertise because they believe it will make them less critical to the business and lose their job. Rather they tend to keep critical information to keep their power in the organization. Moreover, since the most valued experts are in great demand, they have very limited time to participate in KDD projects. However, most projects are dropped in general because they do not get full support and commitment from the customer. The other main challenge is data acquisition, which is the most time-consuming part of the KDD process. Often the data required for mining is not available. Though it may be possible to capture the unavailable data by modifying the business process, that is not often practical. The other issue is combining data from multiple sources. Combing data from multiple sources requires to have a common key which may not be always available, as separate business units in big companies can have different databases that use different keys to identify records. [Weiss, 2009]

Applying KDD techniques alone is not sufficient to solve many business problems. KDD must be integrated into the applications used by business users. Data mining functions are not stand-alone functions used by power users, rather today data mining functions are embedded and integrated into applications [Weiss, 2009]. Moreover, KDD is used with other analytics methods such as business optimization and decision management systems to solve some business problems [Brown *et al.*, 2011]. The other factor that plays big role is the use of domain knowledge. It has been shown that the effectiveness of a data mining system was improved by including knowledge of experienced domain experts in insurance application [Weiss, 2009].

Data mining applications are changing to meet new challenges. Combining integrated analytics and optimization algorithms will create new generation of decisions support systems that enable automating decisions in business process. A need to analyze the exponentially growing big data in real-time is critical as IBM forecasts a data growth from

800,000 petabytes to 35 zettabytes only in the coming decade. Social media mining to gain insight into buyer's behavior is becoming critical as social medias such as blogs and social networks are affecting buyer behavior. [Apte, 2011]

KDD applications in real world business scenarios have certain common characteristics. These characteristics can be divided as data characteristic and operations characteristics [Bose and Mahapatra, 2011].

The data characteristics are because of the nature of business data and they include [Bose and Mahapatra, 2011]:

- *Noisy data*. Business databases contain noisy data because of inaccuracies and inconsistency at data entry. In addition, noise can be introduced the data at the time of extracting the data from the source for analysis.
- *Missing data*. This is another common issue and it refers to attributes with no value (null). This can occur at data entry or at the time of exporting the data from the data source. The other reason could be that the case doesn't have value for certain attribute, for example it may not be applicable for that case.
- *Unavailable attributes*. All the attributes required for analysis may not be available in the data set. This can be because of uncoordinated database design.
- *Large data sets*. The size of the data sets can be from terabytes to several gigabytes. These data sets may have large number of attributes. The ability of the algorithms to handle large data sets is critical in this situation.
- *Various data types*. Todays' business databases contain different types of data types such as numeric, textual, nominal, ordinal, interval and ratio.

The operational characteristic refers to developing a model and deploying it in a real-world business scenario. The operational characteristics are [Bose and Mahapatra, 2011]:

- *Declining predictive accuracy*. In machine learning-based data mining methods the system is first trained on training data. However, the predictive accuracy of the system with real data decreases. Prediction on actual data is critical for business applications.
- *Explaining results*. The business users and managers are more interested if the models and results we get can be explained in business terms.
- *Technical simplicity and less preprocessing*. The degree of expertise required to effectively use data mining tools varies. Moreover, the amount of pre-processing required to prepare the data for analysis differs between different data mining techniques. Ease of understanding and less data pre-processing makes a data mining method ideal for business applications.

5.2. KDD applications in business

In this section the most popular KDD applications in business are presented. The application areas discussed include fraud detection, marketing, e-business, and financial applications.

5.2.1. Fraud detection

Fraud involves misleading others to gain personal benefits [Cepêda de Sousa, 2014]. Although fraud can take different patterns, the common characteristics of most fraudulent behavior is that they are different from the norm in some way [Baragoin *et al.*, 2011]. The common application areas of fraud detection include credit card fraud detection, accounting fraud detection, internal fraud detection (companies) and telecommunications fraud detection. In the case of detecting unauthorized call from stolen phones, money laundering, and insider trading the need for near real-time or very timely detecting is critical [Baragoin *et al.*, 2011].

Telecommunications fraud is characterized by abusive usage of carrier services without the intention to pay, the victims can be the carrier or the client [Cepêda de Sousa, 2014]. The objective of most telecommunications fraud detection focus on detecting or preventing the methods of superimposed fraud and subscription fraud [Cepêda de Sousa, 2014]. In the case of credit card fraud detection, the objective is to identify those transactions that are fraudulent and to classify the transactions in the database as legitimate and fraudulent [Gayathri and Malathi, 2013]. Credit card frauds can be broadly divided as traditional card related, merchant related, and Internet related frauds [Bhatla *et al.*, 2003]. Forensic accounting is a field that studies fraudulent financial transactions, the analysis of funding mechanisms for terrorism is one area getting focus [Kovalerchuk and Vityaev, 2005]. Data mining is utilized to detect internal fraud in companies related to procurement fraud such as double payment of invoices and changing purchase order after release [Jans *et al.*, 2007]. In the insurance business data mining is used to detect if a claim is fraudulent [Petre, 2013].

The challenge in modelling fraud is you don't know what to model. The deviation detection technique overcomes this problem as it does not require a labelled data [Baragoin *et al.*, 2011]. Any transaction that deviates from the norm is detected by the deviation detection technique. Neural network, decision tree, naïve Bayes, K-NN and support vector machine are the most commonly used classification techniques in fraud detection [Cepêda de Sousa, 2014]. User profiling, neural networks and rule based systems are used to detect and prevent telecommunications fraud [Cepêda de Sousa, 2014]. Exemplary successful applications of fraud detection systems for telecommunications are published

widely. AT&T has developed a system for detecting international calling fraud [Piatetsky-Shapiro *et al.*, 1996]. The Clonedetector system that uses customer profile is developed by GTE to detect cellular cloning fraud [Piatetsky-Shapiro *et al.*, 1996].

5.2.2. Marketing

Strong competition, saturated market and maturity of products created a shift from quality to an information competition where detailed knowledge of behavior of customers and competitor is crucial [Piatetsky-Shapiro *et al.*, 1996]. Especially the retail market is dynamic one due to similarity of offered products by retailers and the Internet allowed new business concepts which intensified the competition [Garcke *et al.*, 2010]. Customer Relation Management (CRM) is the process of predicting customer behavior and using it to the benefit of the company, Data mining is useful in all the three phases of CRM: customer acquisition, increasing value of existing customer and customer retention [Chopra *et al.*, 2011].

In the case of customer acquisition the task is to find good prospects and target those prospects through marketing. Understanding our existing customers is essential for successful prospecting because once companies know what customer attributes and behaviors are currently driving their profitability they can use it to direct their prospecting efforts [Scridon, 2008]. Customer profiling and segmentation are marketing techniques used to understand existing customers. Profiling is describing or profiling customers based on data and segmentation is dividing the customer database into different groups [Scridon, 2008]. Data mining can be applied to the customer database to perform profiling and segmentation tasks.

The role of data mining is first defining what it means to be a good prospect and then finding rules that allow people with those characteristics to be targeted [Radhakrishnan, 2013]. The assumption is that similar customer data implies similar customer behavior, allowing to assess new customers based on former customers [Garcke *et al.*, 2010]. Clustering algorithms are used for segmentation while regression and classification algorithms are used for individualization (profiling) [Garcke *et al.*, 2010]. Applying data mining to customer database for marketing purposes is called database marketing. Database marketing analyzes database of customers using different techniques to identify customer groups or predict their behavior [Piatetsky-Shapiro *et al.*, 1996]. In Europe, leading market research companies such as A.C. Nielsen and information resources and in USA GfK and Infratest Burke apply KDD to the rapidly growing sales and marketing databases [Piatetsky-Shapiro *et al.*, 1996]. IDEA analyses effect of new promotions on market behavior in the telecommunications industry [Bose and Mahapatra, 2011]. Another study shows the application of genetic algorithm (GA) to identify groups of customers who will

likely respond to a marketing campaign. The application enabled to maximize return on advertising under a limited budget [Bose and Mahapatra, 2011].

In the case of existing customers, the focus is increasing profitability through cross-selling (offer additional products) and up-selling (offer higher valued products) [Radhakrishnan, 2013]. To offer other products to the customer market basket analysis is commonly used. Market basket analysis looks at associations between different products bought by the customers based on the association discovery algorithms [Piatetsky-Shapiro *et al.*, 1996]. Clustering approaches and content based methods (based on attribute of products such as color, description etc.) are also used to analyze products and categories [Garcke *et al.,* 2010]. Products bought together are placed near to each other in the store, in the case of e-commerce recommendation engines and avatars lead the customer to related products, and in shopping moles electronic devices such as the personal shopping assistant are becoming available enabling shoppers to get product information and related products [Garcke *et al.,* 2010]. An example of increasing value of existing customers is Charles Schwab, the investment company, which discovered that customers open accounts with few thousand dollars even if they had more stashed away in other accounts, and customers who transferred large balances into investment accounts did so in the first few months after they opened their first account. This knowledge enabled Charles Schwab to concentrate its effort during the first months than to send constant solicitations throughout the customer life cycle [Radhakrishnan, 2013].

Another application area of data mining is customer attrition which is concerned with lose of customers. Customer attrition is critical issue for all businesses, especially in mature industries where the initial period of growth is over [Radhakrishnan, 2013]. Customer attrition (also called customer churn) is concerned with identifying customer buying trends and adjust product portfolio, price, and promotion to avoid losing customers [Petre, 2013]. One of the challenges in modelling churn is deciding what it is and recognizing when it has occurred [Radhakrishnan, 2013]. There are two approaches to modelling churn. The first one treats churn as binary outcome and predicts which customers will leave and which will stay and the second one estimates customers' remaining lifetime [Radhakrishnan, 2013].

5.2.3. E-business

E-business refers to presence of a business on the Web in general, where as electronic commerce (e-commerce) which is a component of e-business implies goods and services can be purchased online [Baragoin *et al.*, 2011]. CRM is very critical for on-line businesses because face-to-face contact with customers is not possible and customer loyalty can be lost easily if customer is not satisfied [Chopra *et al.*, 2011]. Data mining can be

applied to customer data to get actionable information that helps a web-enabled e-business to improve its marketing, sales, and customer support operations [Nayak, 2002]. Applications of data mining in e-business includes customer profiling, personalization of service, basket analysis, merchandise planning, and market segmentation [Ismail *et al.*, 2015].

Today, e-businesses are generating huge amounts of data such as customer purchases, browsing patterns, usage times and preferences at an increasing rate [Nayak, 2002]. This huge volume of structured and unstructured data, which is called big data, provides opportunities for companies especially for those that use e-commerce [Ismail *et al.*, 2015]. Due to the heterogeneity and semi-structured or unstructured nature of Web data a pure application of traditional data mining techniques is not sufficient. This led to the development of Web mining [Liu, 2006]. The goal of Web mining is to find useful information and knowledge from the Web hyperlink structure, page content, and usage data [Liu, 2006]. It is important to understand that Web mining and data mining are not the same. Though Web mining uses traditional data mining techniques, many mining tasks and algorithms that are peculiar to Web mining were invented [Liu, 2006].

Mainly two types of data are collected in e-businesses, primary web data (actual web contents) and secondary web data (web server logs, proxy server logs, browser logs, user queries, cookies etc.). The aim of mining primary web data is to effectively interpret searched documents. This helps to organize retrieved information and increase precision of retrieval. The goal of mining secondary web data is to understand buying and traversing habits of customers. Its applications include targeting marketing for a certain group of customers based on web access logs, using link analysis to recommend products, and personalization of websites according to each individual's taste. [Nayak, 2002]

Personalization of websites using recommendation systems is one of the interesting applications in e-business. Web-based personalization aims to match the needs and preferences of the visitor to the online site, it is used by online auction sites such as eBay, camping equipment provider (REI), and Amazon [Weiss, 2009]. Amazon.com is at the forefront in the use of recommendation engines: Customers are shown related products and reviews based on their shopping basket and product search ("customers who bought this product also bought ...") [Garcke *et al.*, 2010].

5.2.4. Financial applications

The nature of uncertainty in the finance world makes predicting the future a fundamental problem in finance and banking [Bose and Mahapatra, 2011]. There are numerous applications of KDD in the financial industry. However, the details of such applications are not widely published by their developers to maintain competitive advantage [Piatetsky-

Shapiro *et al.*, 1996]. Applications in finance includes forecasting stock market, currency exchange rate, bank bankruptcies, understanding and managing financial risk, trading futures, credit rating, loan management, bank customer profiling, and money laundering [Kovalerchuk and Vityaev, 2005]. Another classic application in banking is credit scoring, where models are used to predict whether a new loan applicant will default on a loan and this information is used to grant or reject a loan for an applicant [Bose and Mahapatra, 2011; Petre, 2013]. Prediction tasks in finance are mainly prediction of market numeric characteristics such as stock return or exchange rate and predicting whether the market characteristics will increase or decrease. Another type of task is assessment of investing risk [Kovalerchuk and Vityaev, 2005].

Predictive modelling techniques such as statistical regression or neural networks are used in financial analysis applications for portfolio creation and optimization and trading model creation [Piatetsky-Shapiro *et al.*, 1996]. Many data mining methods used in financial modelling includes linear and non-linear models, multi-layer neural networks, k-means and hierarchical clustering; k-nearest neighbours, decision tree analysis, regression (logistic regression; general multiple regression), autoregressive integrated moving average (ARIMA), principal component analysis, and Bayesian learning [Kovalerchuk and Vityaev, 2005].

One of application areas is predicting the bankruptcy of a firm. It has been shown that neural networks (NNs) excel over discriminant analysis method in predicting bankruptcy of a firm [Bose and Mahapatra, 2011]. Rule induction (RI) is used to predict loan defaulters and assess reliability of credit card applicants [Bose and Mahapatra, 2011]. NNs and RI are used to forecast the price of S&P 500 Index [Bose and Mahapatra, 2011]. Automated Investor (AI), developed by Stanley and Co., identifies good trading opportunities [Piatetsky-Shapiro *et al.*, 1996]. Daiwa Securities developed a portfolio management tool that selects a portfolio based on the stock risk and expected rate of return [Piatetsky-Shapiro *et al.*, 1996]. In accounting GUHA, KEX and KnowledgeSeeker are used to identify periodically changing credit and debit balance patterns in a class of accounts from a financial transaction database [Bose and Mahapatra, 2011].

5.2.5. Other application areas

There are interesting applications of data mining in manufacturing. Scheduling is one of the complex problem in developing manufacturing systems. GA-based systems have been used to solve scheduling problems [Bose and Mahapatra, 2011]. CASSIOPEE troubleshooting systems, which received the European first prize for innovative applications, was developed in a joint venture between General Electric and SNECMA. Three major European airlines used it to diagnose and predict problems for the Boeing 737 [Fayyad *et al.*, 1996].

Management of telecommunication networks is another application area. A large amount of alarms is produced daily; these alarms contain a valuable information about the behaviour of the network. Analysing the alarms to find out the fault is a complex problem. Fault management systems can use the regularities in the alarms for filtering redundant alarms, locating problems in the network, and predicting severe faults [Piatetsky-Shapiro *et al.*, 1996]. The Telecommunication Alarm Sequence Analyser (TASA) was built at the University of Helsinki in the cooperation with telecommunication equipment manufacturer and three telephone networks [Piatetsky-Shapiro *et al.*, 1996].

6. Methodology

In this chapter the methodology used to do the study is presented. However, detailed information about the company and application area is not discussed due to confidentiality. In this chapter, the data mining tool used is briefly described and the steps of the CRISP-DM KDD process done are presented.

6.1. Data mining tools selection

A short list of popular data mining tools was prepared and a criterion for selection was defined. Polls, Internet search and literature were used to come up with a list of most popular data mining tools used in business. There were both open source and commercial data mining tools. The open source data mining tools considered are R, Weka, Python, and the commercial data mining tools considered are RapidMiner, MATLAB, SPSS, and SAS. The criteria for selection includes usability, cost, and availability of algorithms. RapidMiner was selected because it has more polished user interface and it is easier to use. Different licenses and prices of RapidMiner Studio at the time of doing the research are presented in figure 2. First, the free Starter edition of RapidMiner Studio 6 was used and later a free trial of the Professional edition was used with a cooperation from the vendor.

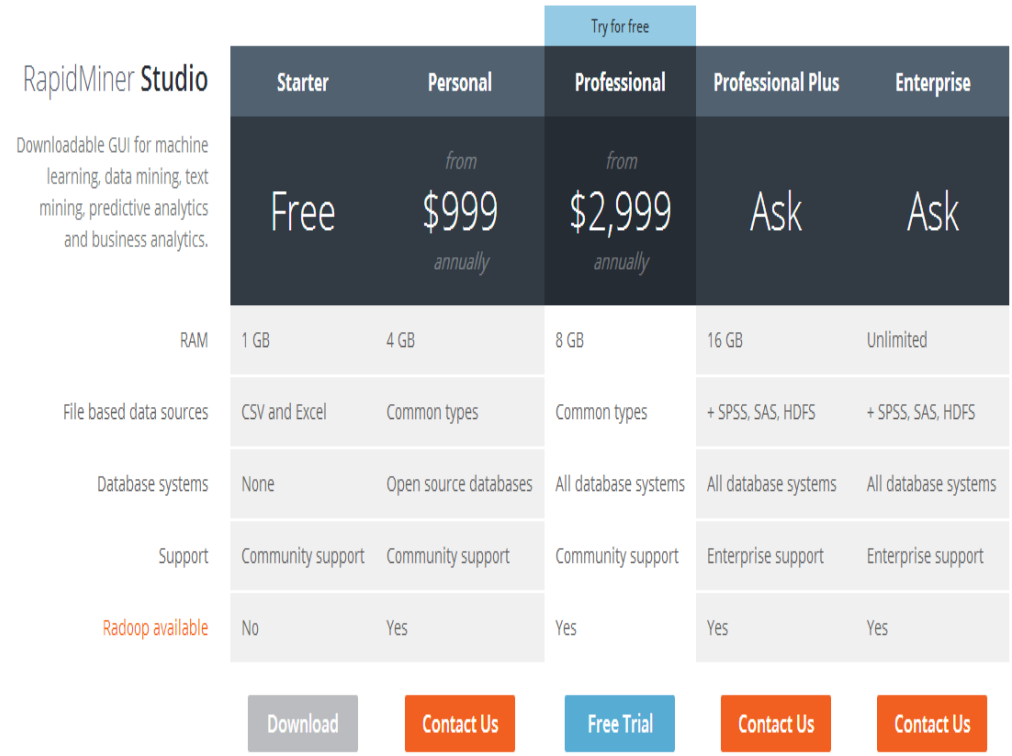

Figure 2. Rapid miner Studio 6 licenses, prices, features at the time of doing the research. [RapidMiner pricing, 2014]

6.2. Business understanding

As already mentioned the first step in the CRISP-DM model is 'business understanding'. The current system and its problem is studied to understand what business problem is being solved. The current system under study is one of the automated financial solutions provided by the company that enables to automate financial business process. This solution is a web-based workflow system.

The solution is provided in different business models which includes:

- *SaaS (software as a service)* – The clients just use the solution provided as SaaS but the work is done by their own personnel.
- *BPO (business process outsourcing)* – The clients fully outsource their business process to the company. The company personnel do the work using the company's solution.
- *Implementation* – The client buys a license and the solution is implemented for the client. In this case the solution is sold as an application.

Our study focuses on the BPO business model i.e. when clients fully outsource their business process to the company. The business process handled by the work flow systems includes:

- *Receive data* - Receive transaction data in different format. The transaction data are received both in paper and electronic format. Object character recognition (OCR) software is used to extract data from paper-based input data.
- *Data entry* - Enter the data into the workflow system
- *Review and processing* - The company personnel check the validity of the data and if it is not complete send it back to the company who sent it.
- Do *Manual task 1* and *Manual task 2* (not mentioned for confidentiality).

The focus of the study is on the two manual tasks in the workflow system. The current system is highly labour-intensive. The manual tasks considered mainly involve daily operational decision making. The current approach is the personnel do it manually using rules, experience, knowledge, and judgement. In addition, there is a rule automation feature that allows to create rules using If-Then format. The objective is the personnel should create rules for repetitive tasks by observing the pattern. However, the rule automation feature is not used extensively. The main reasons for the failure of the rule automation is, that the personnel are not sure if the rule they created can work for all cases and no one can master all the rules making it almost impossible to track and adjust the rules when they change.

To summarize, the business problem is to investigate if KDD can be applied to automate the manual tasks.

6.3. Data preparation

The first step was identifying the necessary data and accessing it. Different data are required to do analysis about the manual tasks. First, there is the input data which the company receives about the transaction of its clients. The other data needed is the data created as a result of processing (handling) the transactions based on the input data received i.e. because of the two manual tasks.

Regarding the input data, the company accepts data about financial transaction of its clients either in electronic or paper documents. The electronic documents have more data and it is very easy to extract the necessary data and enter it to database. However, in case of paper-based documents, data is extracted from the paper documents using OCR and it is not economical to capture all data. Though both electronic and paper based documents are entered to the same set of tables, in the paper-based transaction some information is missing resulting in *null* value in the respective attributes. Due to this reason, it was necessary at the time of extraction to separate the data as electronic-based and paper-based. However, it should be clear that both electronic and paper based data set are almost the same except there are few more attributes that are present in the case of the electronic-based.

Each company was treated separately and the respective data were extracted in separate data sets. This is logical because each company is different and we want to develop a model for each company separately. In extracting the data sets the mechanism used is to use SQL to create a *view*. For each company, the tables that contain the necessary data were joined to form a view, this enabled to create a dataset for each company. Even though the data set of each company is different, the number and type of attributes is the same for all the data sets. The number of attributes extracted was more than 50. The dataset (view) for each company was exported to a comma separated value (CSV) file. A CSV file is a text file. Finally, the CSV file was imported to RapidMiner Studio. The

rows of the datasets ranged from 2,000 up to 900,000, this is expected as some companies have larger transaction data and others smaller.

Company code	Source of data	Period
4110	electronic	Jan – May 2014 (5 month)
4111	electronic	Jan – May 2014 (5 month)
4113	electronic	Jan – May 2014 (5 month)
4114	electronic	Jan – May 2014 (5 month)
4112	electronic	Jan – Sep 2014 (9 months)
9170	electronic	Jan – Sep 2014 (9 months)
4112	paper	Jan – Sep 2014 (9 months)
9170	paper	Jan – Sep 2014 (9 months)

Table 2. The extracted data for each selected company.

The next step was preparing (preprocessing) the data which includes:

- attribute reduction.
- handling missing values.
- handling noisy data.

Attribute reduction

As already mentioned, each of the data sets extracted have more than 50 attributes. It was necessary to reduce the attributes to those attributes that are relevant to predicting the manual tasks. This was not easy because there were many attributes (>50) and it was not easy to determine which ones are relevant. The recommendation of domain experts was

used in selecting the relevant attributes. Moreover, it was found logical to start with fewer attributes at the first stage to reduce complexity.

In RapidMiner, a data set is only once imported to RapidMiner Studio and attribute reduction is done by selecting relevant attributes at the time of building a model. In our case, the data sets of all companies were reduced to six attributes (by selecting) while building models. The only exception was the paper-based data sets; they were reduced to three attributes. The paper-based data sets have fewer attributes because the attributes that were in the electronic-based datasets were not available.

Handling missing values

The data sets have a lot of missing values. This is typical of real-world business databases. The first option was to filter the missing values. This option didn't give a better result rather a loss of considerable data. The second option was to replace missing values, using average values or most frequently occurring values. However, this option didn't work for our scenario. The values of the attributes are not numerical rather categorical making it difficult to apply simple replacement methods such as mean and mode. Moreover, since each attribute has hundreds of different values replacing missing values using these kinds of techniques might not result in a realistic data. Therefore, the records with missing value were kept.

Handling noisy data

There was a considerable amount of inconsistency and noise in the data. Particularly, inconsistency in the way data had been entered was a big issue. The same data had been entered bit differently such as using different names, terms, and descriptions. There were also significant number of spelling errors and data entry errors. Data cleaning was done by removing outliers. However, this is not done thoroughly due to time constraint.

6.4. Modelling

In this chapter the modelling process is presented. The data mining task and the machine learning methods used are discussed. The use of confidence levels to improve the models performance is also discussed.

6.4.1. Classification tasks

As stated in the business understanding step the objective is to automate the two manual tasks by finding pattern of the rules used by the company personnel to do the manual tasks. The basic business problem is how to handle a transaction given some known variables about the transaction. This problem was mapped to data mining classification task. Classification is the task of predicting to which group (class) a new instance belongs. Predicting means determining unknown value of a variable based on other given variables. The discovered classification model or pattern can be used on new instances of

transactions to predict unknown values. In that way, the model can be deployed to eliminate or reduce the manual tasks. The two manual tasks considered were reduced to two classification tasks. Let us call them classification task 1 (for manual task 1) and classification task 2 (for manual task 2).

6.4.2. Machine learning algorithms used

Models for each type of classification task were developed using three machine learning algorithms. The three machine learning algorithms are decision tree, naïve Bayes and K-nearest neighbor (K-NN). Supervised learning was used to train the machine learning algorithms. In supervised learning, all the instances in the data set are given with known labels i.e. the correct output [Kotsiantis *et al.*, 2006]. A brief description of the machine learning algorithms used is presented below. Decision trees belong to the family of Top-Down Induction of Decision Trees (TDIDT) [Quinlan, 1986] whereas Naïve Bayes and K-NN belong to family of classification techniques that are based on statistical approach [Kotsiantis *et al.*, 2006].

Decision tree

Decision trees (DT) are inverted trees that classify instances by sorting them based on attribute values [Robles-Granda and Belik, 2010]. Each inner DT node represents an attribute of the instance to be classified and each branch represents a value that the node can have [Kotsiantis *et al.*, 2006]. The classification starts at the root node; we compare attribute value of the new instance with the branches of the root node. We follow the branch that matches the attribute value of the instance. This process is repeated until we reach a leaf node (node without branches). A leaf node tells us to which class the instance belongs to. DT is widely used because it is simple to interpret, good performance on large dataset, and high-level of robustness [Robles-Granda and Belik, 2010]. A simple decision tree based on a dataset that have attributes *outlook*, *temperature*, *humidity,* and *windy* and a class variable *Class* is shown in figure 3.

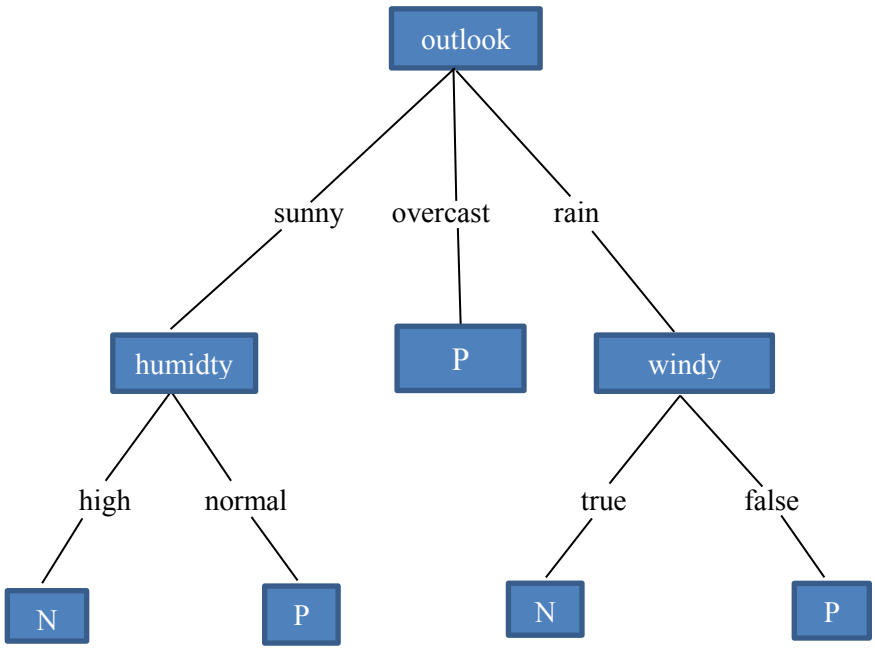

Figure 3. A simple decision tree. [Quinlan, 1986]

The aim of constructing decision trees is to construct a DT that will not only classify instances in the training data but also other unseen instances. The training dataset is a representative sample of the population and the aim is to generalize (induce) from the training dataset. One problem with DTs is overfitting. A DT h is said to overfit training data if another DT h' has a larger error on the training data, but smaller error on the entire dataset than h [Kotsiantis *et al.*, 2006]. An overfitted DT have a better performance on training data than on unseen data, where as a DT that generalized from the training data have a better performance on unseen data than the training data. Pre-pruning which involves not allowing the DT to grow to its full size is the most straightforward way of solving overfitting [Robles-Granda and Belik, 2010].

Naïve Bayes

Naïve Bayes classifiers are family of simple probabilistic classifiers based on Bayes' theorem that assume strong (naive) independence between the features [Robles-Granda and Belik, 2010; Han *et al.*, 2011]. Though, naïve Bayes has very simple mathematical assumptions, it is effective in solving complicated problems [Robles-Granda and Belik, 2010]. Naïve Bayesian classifiers are a simple class of Bayesian networks. A Bayesian network is a graphical model where the structure of the network is a directed acyclic graph (DAG) that represent probability relationships between set of features (variables) [Kotsiantis *et al.*, 2006]. The nodes represent features (attributes) and the directed edges represent dependence between variables. In a DAG, all the edges are directed in one direction and there are no cycles i.e. if we start from one node and traverse along the di-

rected edges we cannot arrive back at the starting node [Stephenson, 2002]. Naïve Bayesian networks are very simple Bayesian networks with only one parent node (representing the unknown variable) and several child nodes (representing known variables) [Kotsiantis *et al.*, 2006]. Following the same dataset example used for making a decision tree, a naïve Bayes network is shown in figure 4.

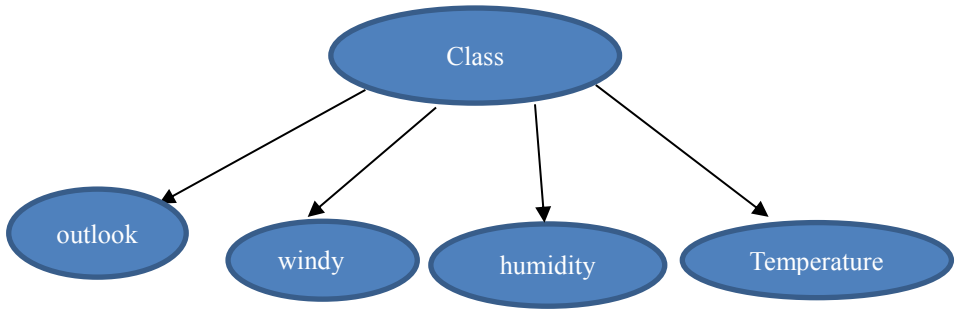

Figure 4. A naïve Bayes network.

In general, developing a Bayesian network model consists of first learning the DAG structure of the network and then constructing a table for each variable (node) that shows probability parameters. The probability distribution table tell us the probability of a node given value of other variables. The probability distribution table will be referred when trying to classify a new instance with given features. Given a new instance to be classified, represented by $X = (x_1, \ldots, x_n)$, the probabilistic model assigns to it probabilities for each of the possible classes C_k which can be formulated as [Han *et al*, 2011]:

$$p\ (C_k|x_{1,\ \ldots,}\ x_n).$$

This can be read as the probability of class C_k given the instance $X = (x_1, \ldots, x_n)$. The probability is calculated using naïve Bayes theorem, which simplifies the problem by assuming strong independence between variables. The only dependence each variable have is on the class (parent node) (see figure 4).

K-Nearest Neighbor (K-NN)

K-NN is one of the most straightforward instance-based learning (IBL) algorithms [Kotsiantis *et al.*, 2006]. IBL algorithms use known cases to solve new problems. K-NN is based on learning by analogy, a given a new instance is compared to similar instances in the training data set [Han *et al*, 2011]. K-NN assumes that instances in a dataset exist in close proximity to other instances with similar properties [Kotsiantis *et al.*, 2006]. The instances that are close to each other are called neighbors. K is a positive, usually small odd number. Given a new unclassified instance its class is predicted by finding its k nearest neighbors and assigning it the class that is most frequent among its neighbors [Kotsiantis *et al.*, 2006].

K is a positive number, typically a small number. The selection of K affects the classification success and among many ways to select K the simplest one is to run the algorithm with different K values and choose the one with best performance [Guo *et al.*, 2003]. Given a new instance K-NN searches for k training samples that are close to the new instance based on similarity measure such as Euclidian distance or Cosine similarity [Guo *et al.*, 2003].

6.4.3. Confidence level

Obviously, predicting using a model have uncertainty and it is necessary to measure a models performance. The common practice is to evaluate the model performance on the whole test dataset using measures such as classification rate, sensitivity, specificity, and receiver operating characteristics (ROC) but it does not make a difference between instances of data points predicted [Alasalmi *et al.*, 2016]. If a model has 90% prediction accuracy, it tells us 90 out of 100 times the model predicts correctly. However, this does not tell us the models performance on predicting a single data point $X = (x_1, \ldots, x_n)$, i.e. how confident the model is when predicting an instance. The confidence level helps us measure certainty of our prediction at a single data point.

Since our application area was automating manual tasks which are sensitive, using confidence levels helps to improve the models usage in real world. Using different confidence levels values either to predict, recommend or drop the prediction was proposed. For example, when the confidence interval is high (say 90%) predict. Such high confidence level predictions can be used to automate the manual tasks without human intervention. On the other hand, when confidence level is relatively high (between 70% - 90%) propose or recommend the prediction and a human expert makes decision. Finally, when confidence level is low (<70%) drop the prediction. RapidMiner have a functionality that allows to drop predictions that are below a certain confidence level. This feature was used by setting the confidence level at different values to see how it affects the overall performance of the models.

6.4.4. Overview of the modeling process

Models were developed for each of the two types of classification tasks described in section 6.4.1. The datasets used to develop models, machine learning methods used and issues encountered are summarized in table 3.

Model	Datasets used	Learning methods trained
Classification task 1	Six feature small datasets: • 4110, 4111, 4113, 4114	• Decision tree, naïve Bayes, and K-NN
	Six feature large datasets: • 4112, 9170	• Only naïve Bayes model for 4112 and naïve Bayes and k-NN for 9170 were trained. • There was a memory problem while trying to train the other learning methods
	Two feature large datasets: • 4112, 9170	• Decision tree, naïve Bayes and, K-NN
Classification task 2	Six feature small datasets: • 4110, 4111, 4113	• Decision tree, naïve Bayes and, K-NN
	Six feature large datasets: • 4112, 9170	• Only naïve Bayes model was trained • There was a memory problem while trying to train the other learning methods
	Two feature large datasets: • 4112, 9170	• Decision tree, naïve Bayes and, K-NN

Table 3. Summary of the modeling process.

6.5. Evaluation methods

The models were tested on the two classification tasks across the set of datasets. Cross validation technique was used to evaluate the performance of the models. Cross validation technique was discussed in chapter 3 under the evaluation step of the KDD process.

6.6. Evaluation measurements

RapidMiner Studio provides functionalities to determine performance criteria values, for polynomial classification tasks (more than one class values) accuracy and kappa statistic

are shown [RapidMiner documentation, 2016]. The result of the evaluation is presented as a confusion matrix and shows the average accuracy. A confusion matrix is a table that allows to analyze how well the classifier predicts the classes [Han *et al*, 2011]. For example, table 4 presents a classification problem with two classes *Yes* and *No*. Each instance in the test is labeled as *Yes* or *No*. In the confusion matrix, the column presents the predicted class of an instance and the row presents the actual class of an instance. True positive and true negative indicate the prediction was correct and false positive and false negative indicate the prediction was wrong.

	Predicted class		
Actual class		Yes	No
	Yes	TP (true positive)	FN (false positive)
	No	FN (false negative)	TN (true negative)

Table 4. A simple confusion matrix [Han *et al*, 2011].

The metrics used to evaluate a model's performance are accuracy and prediction rate. Accuracy is the percentage of correct predictions out of all predictions. It can be calculated in the following way [Han *et al*, 2011]:

$$ACC = \frac{TP + TN}{P + N}$$

where TP is number of true positive cases, TN is number of true negative cases and, P+N is the total number of cases.

Note that in our scenario the total number of cases (P+N) can be different from the number of instances in the test dataset. This is because some predictions might be dropped because they are less than the confidence level. So, in that case the total number of cases is the total predictions made without including the dropped predictions.

Prediction rate is how many percentage we have predicted out of all instances in the test dataset. This metric was necessary because confidence level was used to drop predictions that are below a certain confidence level. Reasonable prediction rate is necessary to use the models in real world. A model that have high accuracy but dropped 80% of predictions is of no use obviously. The prediction rate is calculated as:

$$Prediction\ rate = predicted\ instances / total\ instances\ in\ the\ test\ dataset$$

7. Evaluation results

In this chapter, the evaluation results are presented and discussed. First, the evaluation results on the two classification tasks are presented using graphs. In section 7.1 the evaluation results for classification task 1 are presented and in section 7.2 the evaluation results for classification task 2 are presented. Then, the results for both classification tasks are analyzed by taking into consideration different factors that affect the performance of the models. Section 7.3 presents the analysis of the results.

7.1. Results on classification task 1

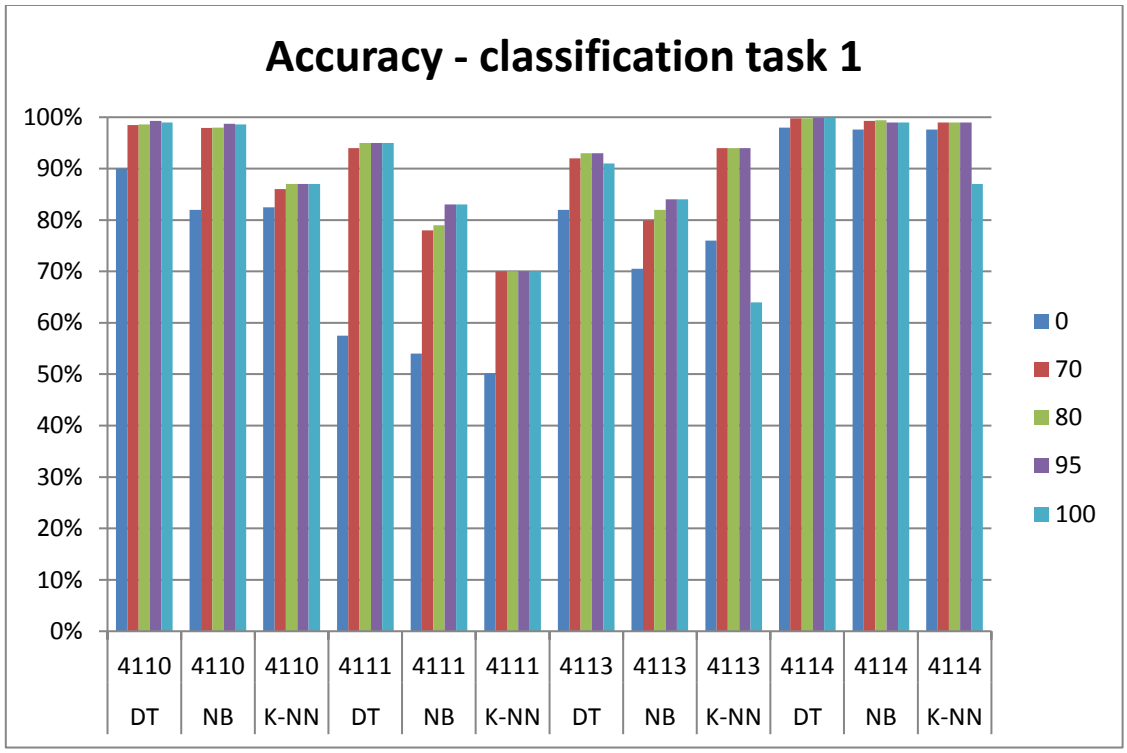

Figure 5. Accuracy on classification task 1 on the set of datasets with six features. In the figure 0, 70, 80, 95, and 100 are confidence levels.

In figure 5, the y-axis shows accuracy of a model and the x-axis shows machine learning algorithm and dataset used. Moreover, different colors of bars are used to show confidence levels values used. For example, the first bar tells us accuracy of a model developed using decision tree (DT) on dataset 4110 at different confidence levels. The accuracy is 90% at confidence level 0 and its near 100% at other confidence levels. Other graphs presenting accuracy of models could be read in similar fashion.

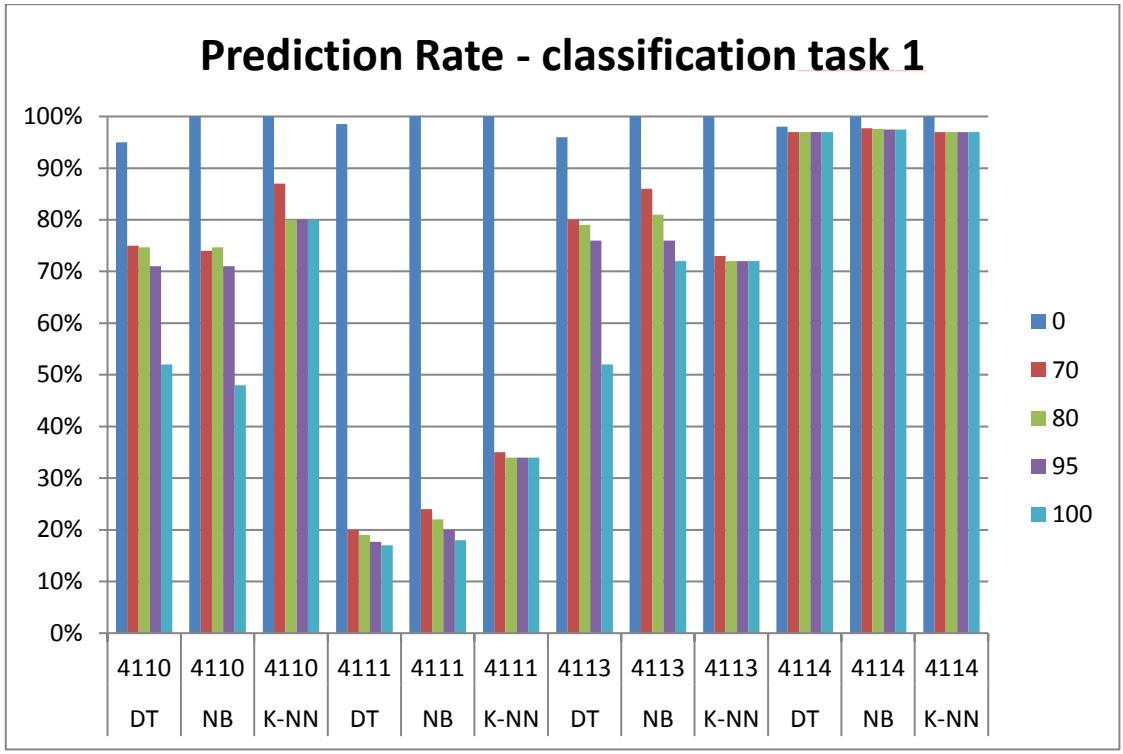

Figure 6. Prediction rate on classification task 1 on the set of datasets with six features. In the figure 0, 70, 80, 95, and 100 are confidence levels.

In Figure 6, prediction rate of a model is shown on the y-axis and on the x-axis machine learning algorithm and dataset used are presented. Moreover, different colors of bars are used to show confidence levels values used. Other graphs presenting prediction rate of models could be read in similar fashion.

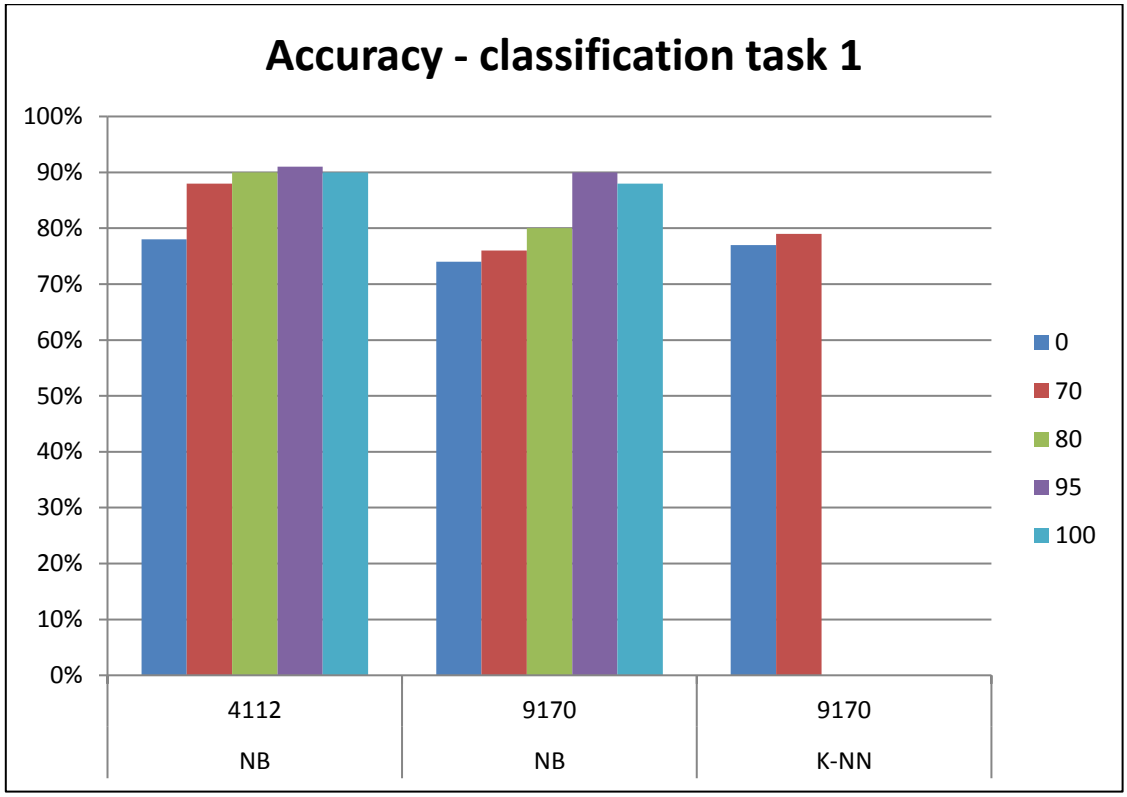

Figure 7. Accuracy on classification task 1 on the set of large datasets with six features.

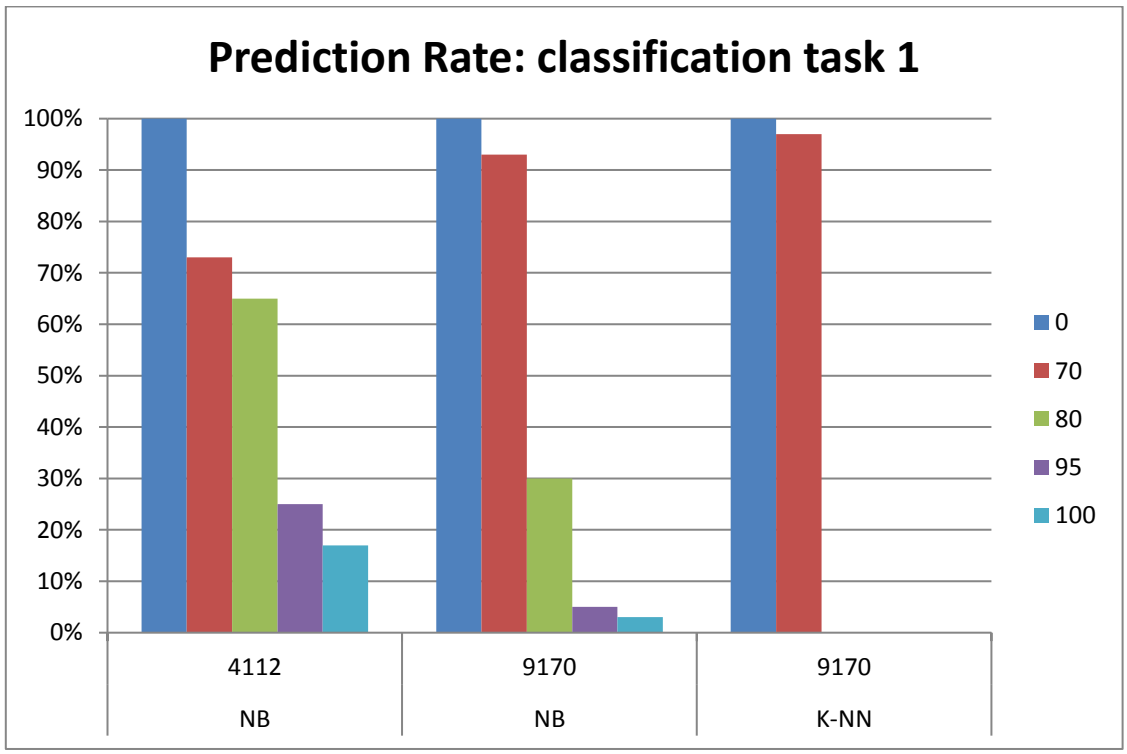

Figure 8. Prediction rate on classification task 1 on the set of large datasets with six features.

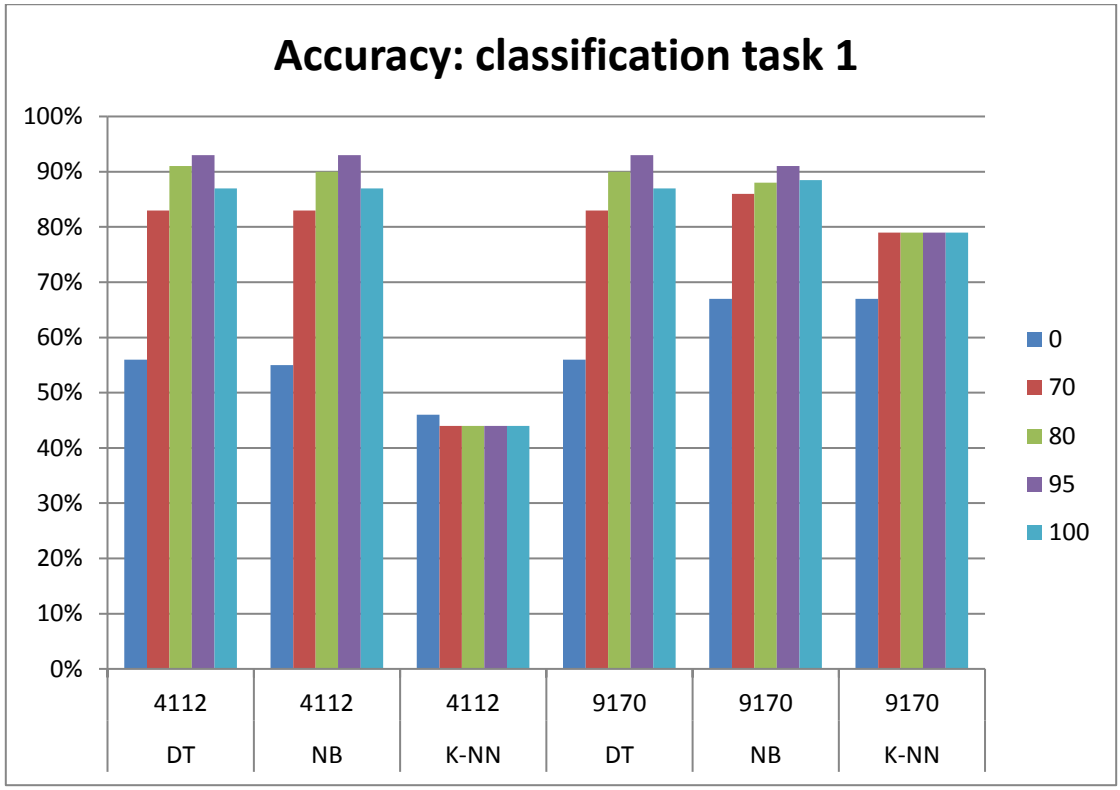

Figure 9. Accuracy on classification task 1 on the set of large datasets with two features.

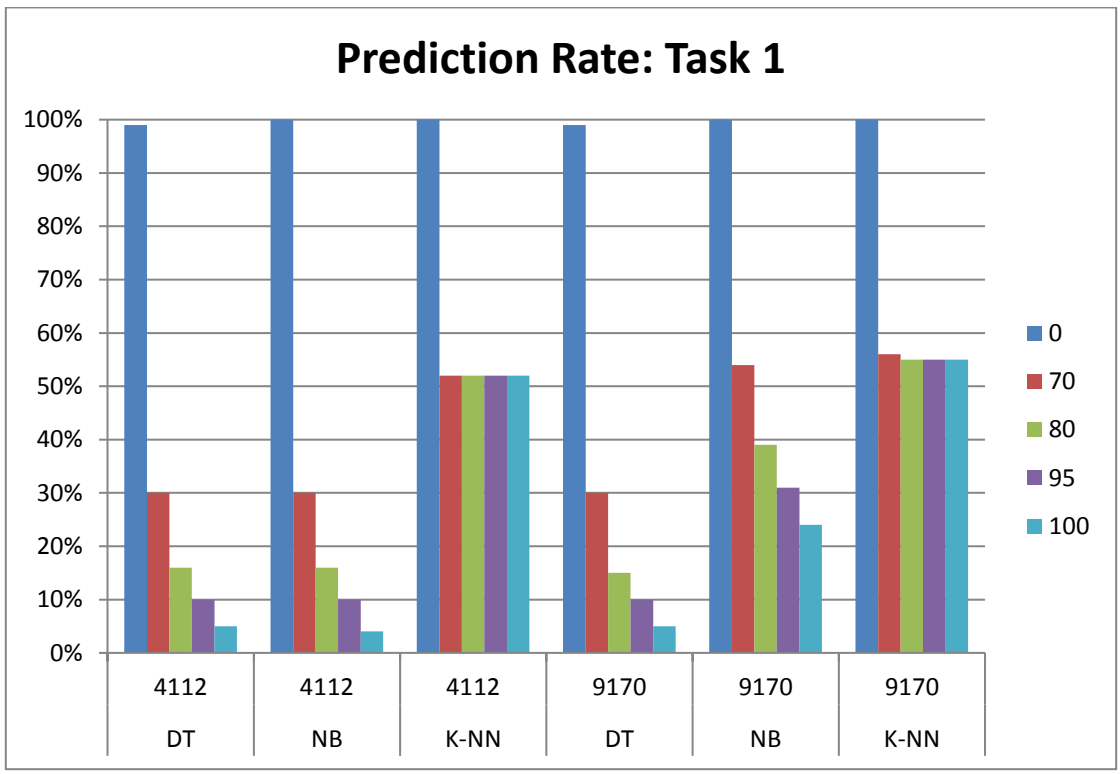

Figure 10. Prediction rate on classification task 1 on the set of large datasets with two features.

7.2. Results on classification task 2

The results on classification task 2 are presented in this section.

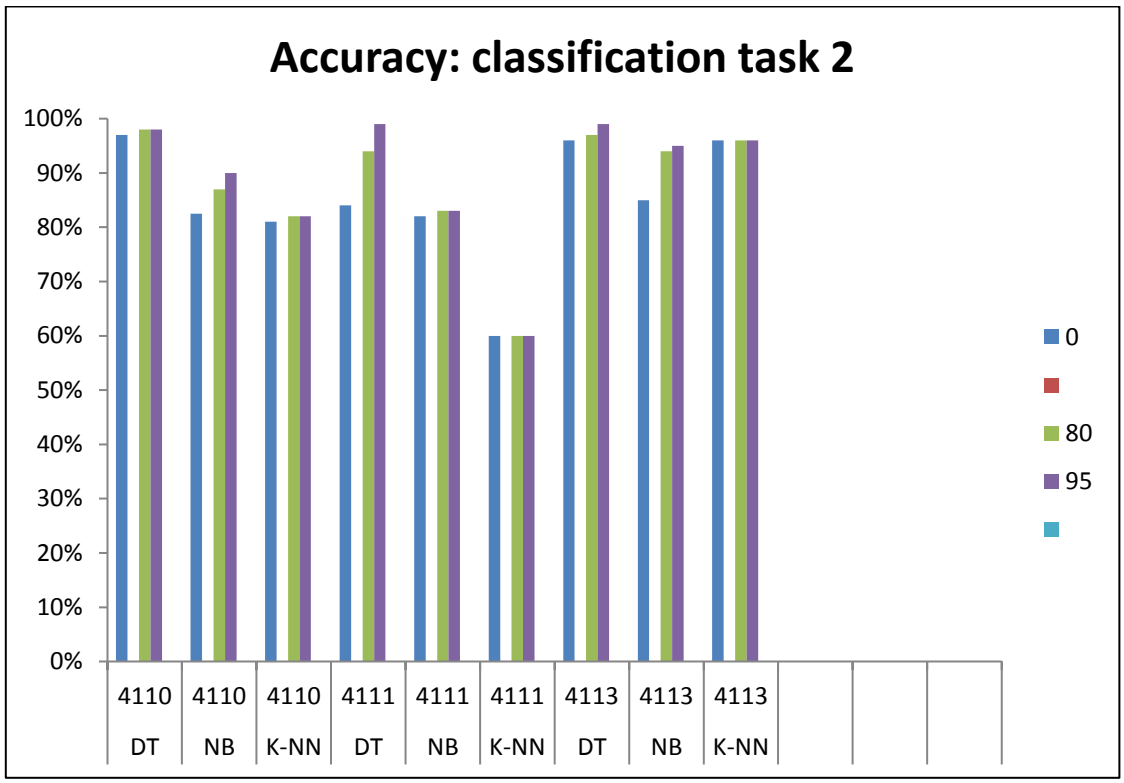

Figure 11. Accuracy on classification task 2 on the set of datasets with six features.

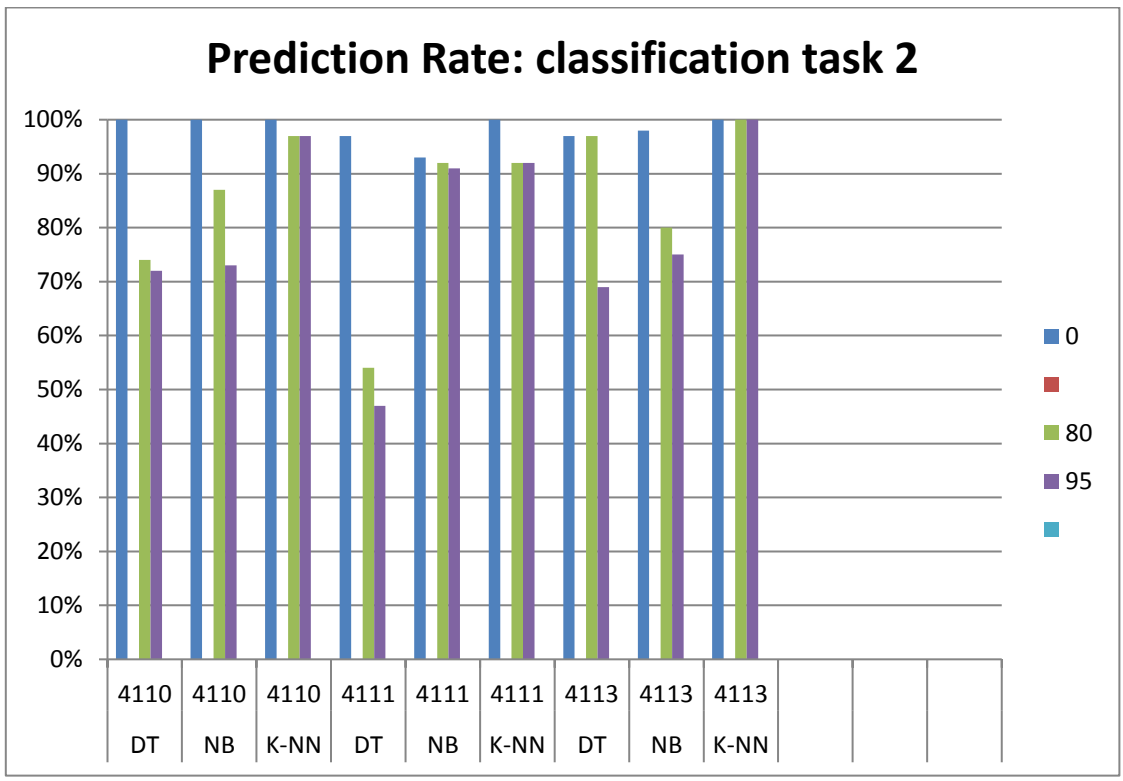

Figure 12. Prediction rate on classification task 2 on the set of datasets with six features.

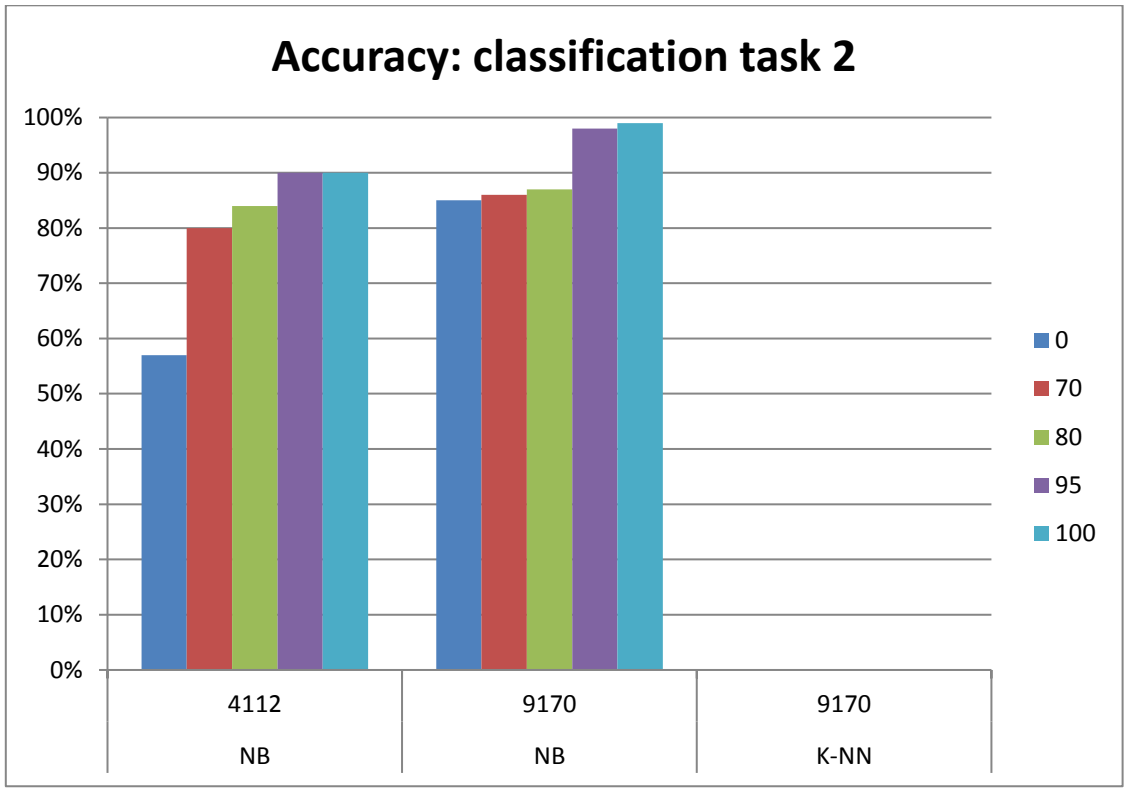

Figure 13. Accuracy on classification task 2 on the set of large datasets with six features.

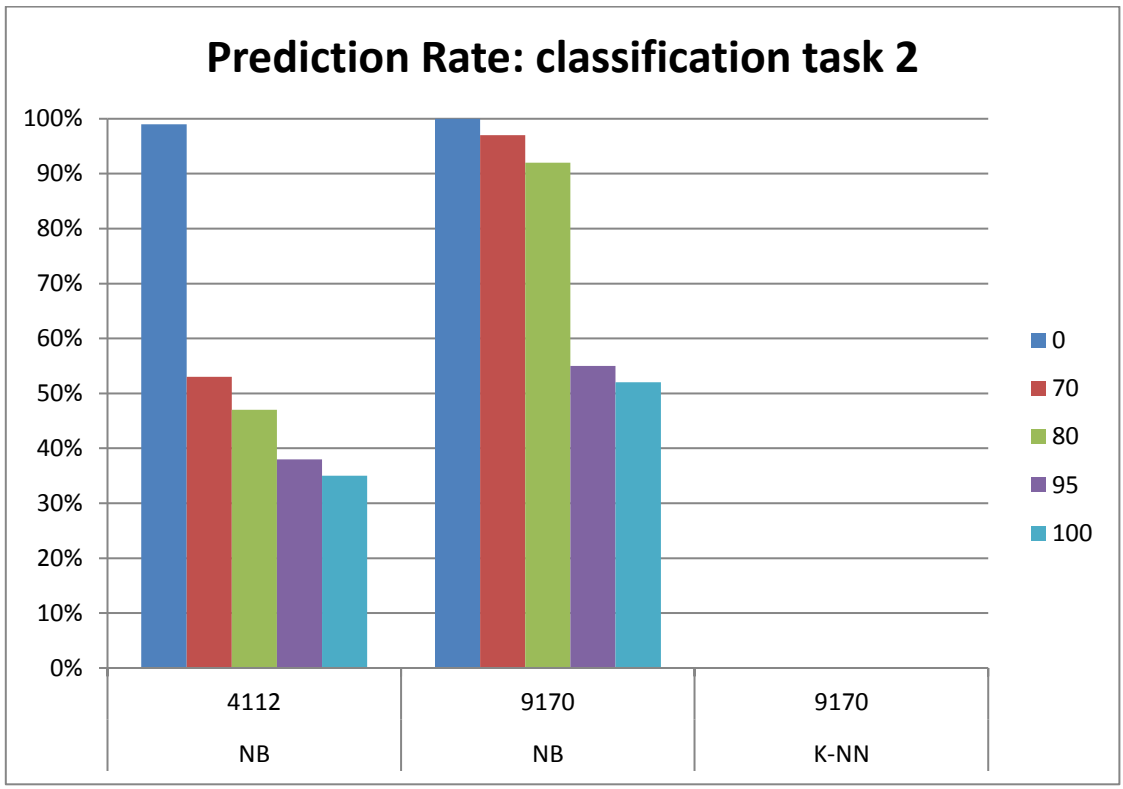

Figure 14. Prediction rate on classification task 2 on the set of large datasets with six features.

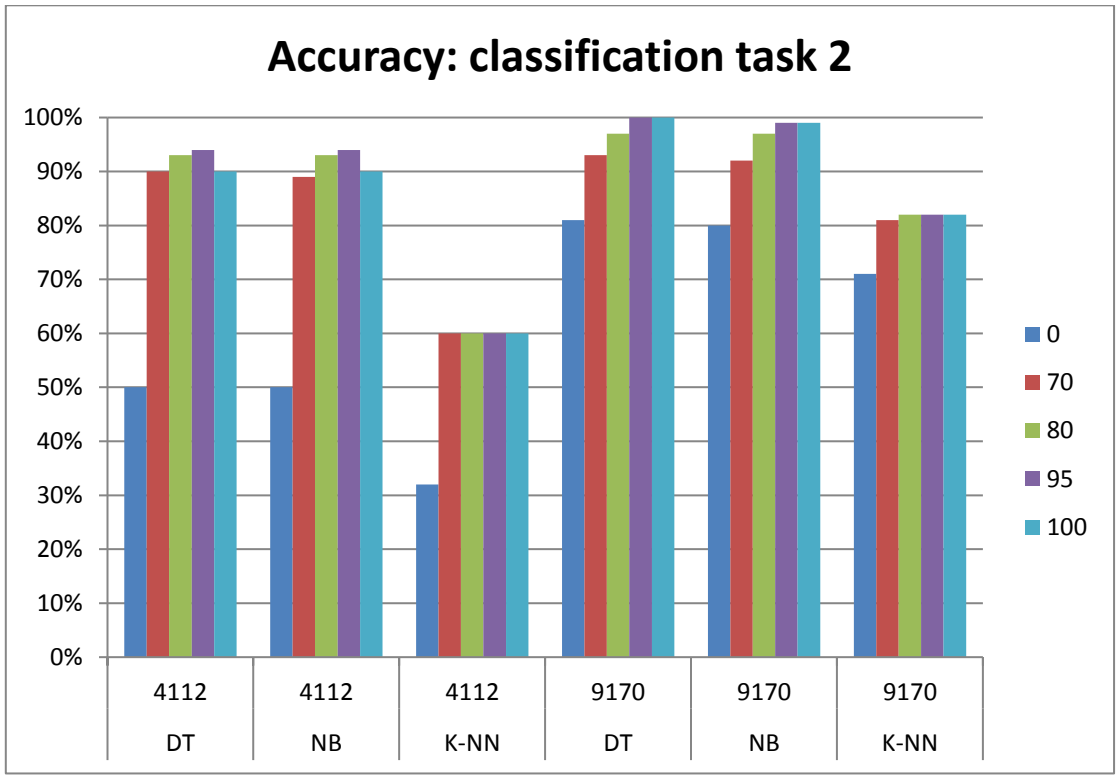

Figure 15. Accuracy on classification task 2 on the set of large datasets with two features.

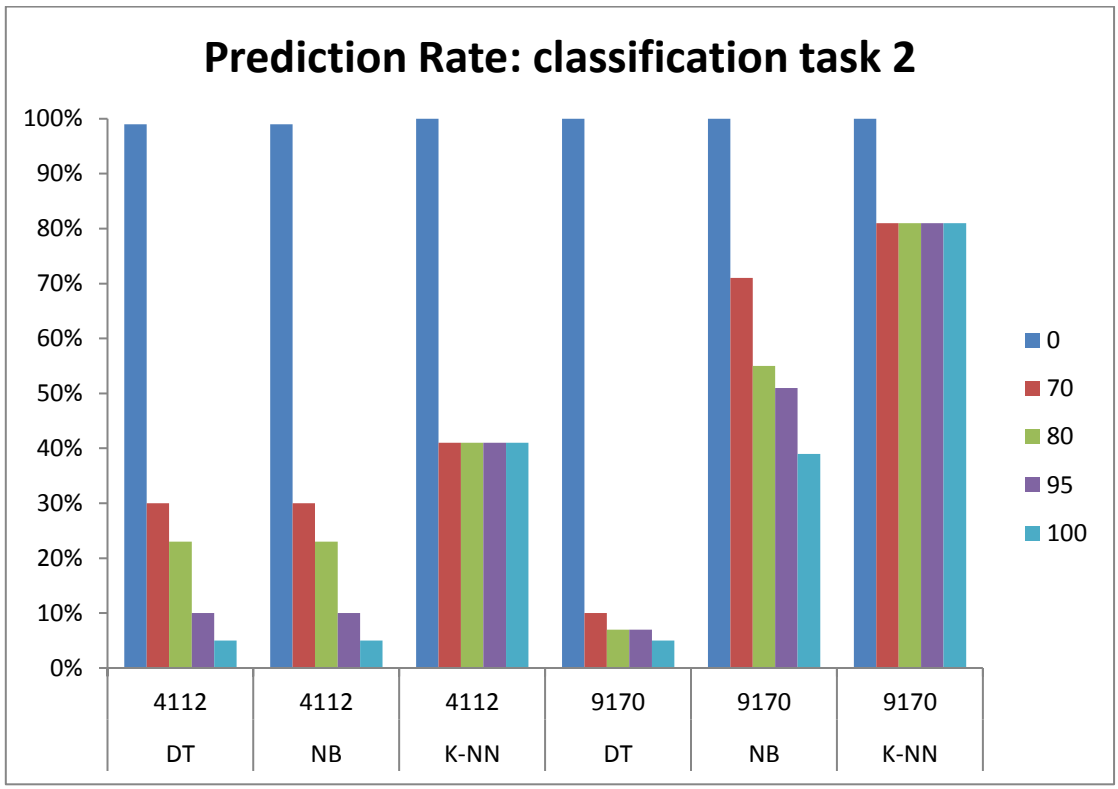

Figure 16. Prediction rate on classification task 2 on the set of large datasets with two features.

7.3. Analysis of results

The evaluation results show performance of the developed models on the two classification tasks based on evaluation metrics accuracy and prediction rate. Obviously, the models have different performance. There are different factors that affect performance of a model. In our case, learning method used, confidence level value and, size and number of features of dataset affect the performance of a model. The overall observed effects of this factors are summarized in Table 5.

Let us see the effect of machine learning method used on performance of a model. Figures 5, 9, 11 and 15 present accuracy of models and almost in all cases decision tree has the highest performance followed by naïve Bayes. Though, figure 7 and figure 13 present accuracy of models they are not used to compare models accuracy by learning algorithms because they do not present results for all the three learning methods. In a similar way, figures 6,10, 12 and 16 present prediction rate of models and K-NN has the highest performance followed by naïve Bayes. Again, figures 8 and 14 are not used because they don't present results for all the three learning methods. There is an interesting relationship between the effect of a learning method on accuracy and prediction rate. Decision tree which has the highest accuracy has the lowest prediction rate and K-NN which has the highest prediction rate has the lowest accuracy. This tradeoff between accuracy and prediction rate could be explained by confidence level, which was discussed in section 6.4.3. In general, the optimal accuracy and prediction rate depends on the application area. If the cost of error is high, then high accuracy is better than high prediction rate. In automating manual tasks considered in this study, high accuracy is essential. Since decision tree has overall high accuracy it can be concluded that it has a higher performance. However, since no single learning method has high performance always, the use of ensemble methods might give a better performance. Ensemble methods use a weighted vote of multiple classifiers predictions to classify a new instance [Dietterich, 2000].

Confidence level is one of the factors that affect performance of models. Analyzing the graphs that present accuracy (figures 5, 7, 9, 11, 13 and 15) shows that in general accuracy increases as confidence value increases and vice versa. In similar way, analyzing graphs that present prediction rate (figures 6,8,10,12,14, and 16) shows in general prediction rate decreases as confidence value increases and vice versa. However, in some cases (for example figure 9) a confidence level of 100% results in less accuracy than 80% confidence level. This is against the expected tradeoff between accuracy and prediction rate. Obviously, the problem is with the confidence estimation. Assuming a correct confidence estimation, a better result is expected at higher confidence values. However, generating reliable confidence estimates that work always is challenging and it is difficult to avoid the situation where predictions labeled confident become incorrect [Delany *et al.*, 2005].

Moreover, there is no measure to decide which confidence level value gives optimal accuracy and prediction rate. Therefore, it is necessary to try different confidence levels to find the one that gives the best results.

Let's see the effect of dataset size and the number of attributes used to get more insight about the effect of dataset on performance. The prediction rate for large datasets (figure 8 and 10) and datasets with two features (figure 10) are low compared with the prediction rate for small size datasets with six features (figure 6). However, this is especially true for datasets with two features as the prediction rate falls dramatically when higher confidence values are used with only little gain in accuracy. Though, there could be many reasons why performances drop, better data preparation could give better results.

Metrics	**Factors**		
	Learning method	**Confidence level**	**Dataset**
Accuracy	Overall, decision tree has a high accuracy	In general, accuracy increases when a high confidence level is used.	• Accuracy is very high or low on some datasets. • The high accuracy on large datasets and datasets with two features was with a very low prediction rate
Prediction rate	Overall, K-NN have a better prediction rate	In general, prediction rate falls with high confidence level.	• Prediction rate on large datasets and datasets with two features is low, • Prediction rate is very high or low on some datasets.

Table 5. Summary of the results presented in section 7.1 and 7.2.

Though, performance difference on different datasets is expected because each dataset represents different company, there were very high or low results on some datasets that

need further investigation. For example, in figure 6 the prediction rate for dataset 4111 falls from 100% up to 20%. Obviously, there is significant difference between the type of business each of these companies do. Some are very small companies doing smaller transactions while others are very large corporations doing large number of different types of business. The nature of the transactions recorded in the datasets is from simple transactions to very complex transactions. Studying closely the nature of the transactions results in a better domain knowledge. Therefore, it is necessary to involve domain experts and closely study the effects of these transaction differences on performances.

The overall results show that the manual tasks can be automated by applying KDD. The models showed good performance considering this is the first iteration of applying the KDD process. Moreover, the models developed were very simple as they were trained on few attributes. Developing more powerful models with the help of domain experts would result in better performance. Moreover, the data mining software used to develop the models was user-friendly, easy to learn and, it can be integrated with business solutions.

8. Conclusion

8.1. Summary of results

The aim of the study was to assess if KDD can be applied to automate manual tasks. To achieve this aim, the following research objectives were set:

- Compare different learning methods performance.
- Use set of datasets to validate the results.
- Assess the tools used to see their practicality in business scenario.

The overall evaluation results show that applying KDD have the potential to automate the manual tasks. Accuracy and prediction rate were used to evaluate performance of the models. The models' performance on the two manual tasks were evaluated across set of datasets. The results were analyzed by comparing the learning methods used, confidence level values used and the nature of the datasets used.

The comparison of learning methods shows that there is a tradeoff between accuracy and prediction rate. Models that have high accuracy have low prediction rate and vice versa. Decision trees have the highest accuracy outperforming other learning methods and K-NN have the highest prediction rate. Since no single method have a high performance, the use of ensemble methods might improve the results.

It was found that the results vary across different datasets. The models have low performance on large datasets and datasets with two features. Moreover, the models perform very high or low on some datasets. Though, a further study to find out the reasons for variation of performance on different data sets could lead to a better understanding of the problem.

The use of confidence levels resulted in boosting the performance of model by increasing accuracy. However, there was a tradeoff between accuracy and prediction rate. A gain in accuracy results a loss in prediction rate. It is shown that there is no single confidence level that gives optimal accuracy and prediction rate. Therefore, trying different confidence levels is required to get the optimal performance.

8.2. Recommendations and future work

Data mining is an iterative process and results get better on each iteration and as more data are collected. A lot of improvements are required to get better results. To get better results better data preparation, trying different modeling methods, and more involvement of domain experts are required.

Though, data preparation was done in this study it was not done thoroughly due to time limitation. In future iterations, data preparation should be done with strong emphasis. As discussed in section 5.1 business databases are by nature characterized by missing values and noise. Improving the quality of data through better data preparation would probably give much better results.

The results show that performances vary across datasets. Each dataset represents different company's data. Since the companies considered are engaged in different businesses, the transactions data stored in the datasets have different nature. The effect of the nature of transactions on the models performance needs further study.

In conclusion, the practical approaches used and the results found in this study lay a foundation for future studies.

References

Tuomo Alasalmi, Heli Koskimäki, Jaakko Suutala, and Juha Röning. 2016. Instance level classification confidence estimation. In: Distributed Computing and Artificial Intelligence, 13th International Conference. Springer International Publishing, 275-282.

Mihai Andronie and Daniel Crişan. 2010. Commercially Available Data Mining Tools used in the Economic Environment. Database Systems Journal 1(2), 45-54.

Chid Apte. 2011. The Role of Data Mining in Business Optimization. IBM research, 2011.

Corinne Baragoin, Ronnie Chan, Helena Gottschalk, Gregor Meyer, Paulo Pereira, and Jaap Verhees. 2002. Enhance Your Business Applications. Simple Integration of Advanced Data Mining Functions. IBM Press, 2002.

Tej P. Bhatla, Vikram Prabhu, and Amit Dua. 2003. Understanding Credit Card Frauds. Tata Consultancy Services, 2003.

Indranil Bose and Radha K. Mahapatra. 2011. Business data mining - A machine learning perspective. Information & Management 39(3), 211-225.

D. E. Brown, Fazel Famili, Gerhard Paass, and Sebastián Maldonado. 2011. Future trends in business analytics and optimization. Intelligent Data Analysis. 15(6), 1001-1017.

João Cepêda de Sousa. 2014. Telecommunication Fraud Detection Using Data Mining techniques. M. Sc. Thesis, Faculty of Engineering, University of Porto

Babita Chopra, Vivek Bhambri, and Balram Krishan. 2011. Implementation of data mining techniques for strategic CRM issues. International Journal of Computer Technology and Applications, 2(4), 879-883.

Sarah J. Delany, Padraig Cunningham, Donal Coyle, and Anton Zamolotskikh. 2005. Generating estimates of classification confidence for a case-based spam filter. In: Case-Based Reasoning Research and Development. Berlin Heidelberg, 177-190.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In: Multiple Classifier Systems. Berlin Heidelberg, 1-15.

Tom Fawcett. 2004. ROC Graphs: Notes and Practical Considerations for Researchers. Machine learning 31(1), 1-38.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3), 37-54.

Yongjian Fu. 1997. Data mining: Tasks, techniques, and Applications. IEEE Potentials, 16(4), 18-20.

Jochen Garcke, Michael Griebel and Michael Thess. 2010. Data Mining for the category management in the retail market. In: Martin Grötschel, Klaus Lucas, and Volker Mehrmann (eds.), Production Factor Mathematics. Springer, 81-92.

R. Gayathri and A. Malathi. 2013. Investigation of data mining techniques in fraud detection: credit card. International Journal of Computer Applications, 82(9), 12-15.

Shital Gheware, Anand Kejkar, and Santosh Tondare. 2014. Data Mining: Task, Tools, Techniques, and Applications. International Journal of Advanced Research in Computer and Communication Engineering 3(10), 8095-8098.

M. Goebel and L. Gruenwald. 1999. A Survey of Data Mining and Knowledge Discovery Software Tools. SIGKDD Explorations 1(1), 20-33.

Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. KNN model-based approach in classification. In: On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Springer Berlin Heidelberg, 986-996.

Jiawei Han, Micheline Kamber, and Jian Pei. 2011. Data mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc.

M. Ismail, M. Ibrahim, Z. Sanusi, and M Nat. 2015. Data Mining in Electronic Commerce: Benefits and Challenges. International Journal of Communications, Network, and System Sciences, 8, 501-509.

Joyce Jackson. 2002. DATA MINING: A CONCEPTUAL OVERVIEW. Communications of the Association for Information Systems 8, 267-296.

Mieke J. Jans, Nadine Lybaert, and Koen Vanhoof. 2007. Data mining for fraud detection: Toward an improvement on internal control systems?. In: Proc. of the 30th Annual Congress European Accounting Association (EAA2007), 2007.

S.B. Kotsiantis, I.D. Zaharakis, and P.E. Pintelas. 2006. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3),159-190.

Boris Kovalerchuk and E. E. Vityaev. 2005. Data Mining for Financial Applications. In: Oded Maimon and Lior Rokach (eds.), Data Mining and Knowledge Discovery Handbook. Springer-Verlag New York, Inc., 1203-1224.

Bing Liu. 2006. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer-Verlag New York, Inc., 2006.

Thomas M. Mitchell. 1997. Machine Learning. McGraw-Hill, Inc., 1997.

Richi Nayak. 2002. Data Mining for Web-Enabled Electronic Business Applications. In: Architectural Issues of Web-Enabled Electronic Business. IGI Publishing, 129-138.

Martin A. North. 2012. *Data Mining for The Masses*. Available: https://rapidminer.com/resource/data-mining-masses

Ruxandra-Stefania Petre. 2013. Data Mining Solutions for the Business Environment. Database Systems Journal 4 (4), 21-29.

Gregory Piatetsky-Shapiro, Ron Brachman, Tom Khabaza, Willi Kloesgen, and Evangelos Simoudis. 1996. An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications. In: Proc. KDD-96, 89-95.

Priyadharsini.C and Dr. Antony Selvadoss Thanamani. 2014. An Overview of Knowledge Discovery Database and Data Mining Techniques. International Journal of Innovative Research in Computer and Communication Engineering 2(1).

J.R. Quinlan. 1986. Induction of decision tree. Machine Learning, 1(1), 81-106.

Radhakrishnan B, Shineraj G, and Anver Muhammed K.M. 2013. Application of Data Mining in Marketing. International Journal of Computer Science and Network, 2(5), 41-46.

Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, and A.S.K. Ratnam. 2012. A Study of Data Mining Tools in Knowledge Discovery Process. International Journal of Soft Computing and Engineering 2(3), 191-1994.

RapidMiner documentation. 2016. Available: http://docs.rapidminer.com

RapidMiner pricing. Available: Rapidminer.com/pricing. (checked on Jun, 2014)

Pablo D. Robles-Granda and Ivan V. Belik. 2010. A Comparison of Machine Learning Classifiers Applied to Financial Datasets. Lecture Notes in Engineering and Computer Science, 2186(1), 454-459.

Mircea A. Scridon. 2008. Understanding customers - profiling and segmentation. Management & Marketing, 6(1), 175-184.

T.A. Stephenson. 2002. An Introduction to Bayesian Network Theory and Usage, IDIAP Research Report 00-03.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2004. Introduction to Data Mining. Addison-Wesley Longman Publishing Co., Inc., 2004.

Divya Tomar and Sonali Agarwal. 2014. A Survey on Pre-processing and Post-Processing Techniques in Data Mining. International Journal of Database Theory and Application 7(4), 99-128.

Gary M. Weiss. 2009. Data mining in the real world: experiences, challenges, and recommendations. In: Proc. The 2009 International Conference on Data Mining, DMIN 2009, 124-130.