ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

MINING INSURANCE DATA FOR FRAUD DETECTION: THE CASE OF AFRICA INSURANCE SHARE COMPANY

TARIKU ADANE

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

MINING INSURANCE DATA FOR FRAUD DETECTION: THE CASE OF AFRICA INSURANCE SHARE COMPANY

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfilment of the Requirements for the Degree of Master of Science in Information Science

By

TARIKU ADANE

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

MINING INSURANCE DATA FOR FRAUD DETECTION: THE CASE OF AFRICA INSURANCE SHARE COMPANY

By

TARIKU ADANE

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
Ato Aminu	Chairperson		
Million Meshesha (PhD)	Advisor		
Gashaw Kebede (PhD)	Examiner		

DEDICATION

I would like to dedicate this thesis work to my mother and sister, W/o Belaynesh Admassu and Dinkayehu Mersha who have always been the spice of my life.

ACKNOWLEDGMENT

First and foremost my special thanks goes to the almighty God for his kindness, blessings and forgiveness of my everyday sins with the courage and endurance to successfully complete this research work.

Next to this I would like to express my sincerest gratitude and heartfelt thanks to my advisor, Dr. Million Meshesha, for his keen insight, guidance, and unreserved advising. I am really grateful for his constructive comments and critical readings of the study.

I am also very thankful to my instructors and all staff members of the School of Information Science for their contribution in one way or another for the success of my study. My thanks also goes to Mekelle University (MU) for granting me study leave with the necessary benefits, without which I could not have been able to join my M.Sc. study here in Addis Ababa University (AAU).

I am truly grateful to Professor Conrad Lichtenstein for his kindness, moral and financial support. I would also like to extend my appreciation to my friends Belete Biazen and Kindie Alebachew for their all rounded help during our stay at Addis. I am also very grateful to my brother Mollalet Wasihune and Alemitu D. for always being there when I really need them in my life.

I would also like to thank AIC officials and employs particularly Ato Dejene Megerssa and Ato Asnake Amare for providing me the necessary data for the study and for their unreserved help throughout the study time. I am also very grateful to Ato Getacher Alemu for his support during the data-preprocessing phase of the study. I am also thankful to all my friends especially, Shumet T. & Tadele A. for their support in conducting this research.

Finally, I want to thank my parents for their love, understanding and support in my everyday life.

LIST OF ABBREVIATIONS

AIC: Africa Insurance share Company

ANN: Artificial Neural Network

ARFF: Attribute Relation File Format

CRISP-DM: Cross Industry Standard Process for Data Mining

CRM: Customer Relationship Management

CSV: Comma Separated Values

DM: Data Mining

KDD: Knowledge Discovery in Databases

PREMIA: Profitable Relationship through Effective Management of Insurance Activities

SEMMA: Sample Explore Modify Model Assess

WEKA: Waikato Environment for Knowledge Analysis

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The insurance industry has historically been a growing industry. It plays an important role in insuring the economic well being of one country. But ever since it's beginning as a commercial enterprise, the industry is facing difficulties with insurance fraud. Insurance fraud is very costly and has become a world concern in recent years. Fraudulent claims account for a significant portion of all claims received by insurers, and cost billions of dollars annually. Nowadays, great efforts have been made to develop models to identify potentially fraudulent claims for special investigations using the data mining technology.

This study is initiated with the aim of exploring the potential applicability of the data mining technology in developing models that can detect and predict fraud suspicious in insurance claims with a particular emphasis to Africa Insurance Company. The research has tried to apply first the clustering algorithm followed by classification techniques for developing the predictive model. K-Means clustering algorithm is employed to find the natural grouping of the different insurance claims as fraud and non-fraud. The resulting cluster is then used for developing the classification model. The classification task of this study is carried out using the J48 decision tree and Naïve Bayes algorithms in order to create the model that best classify fraud suspicious insurance claims.

The experiments have been conducted following the six-step Cios et al. (2000) process model. For the experiment, the collected insurance dataset is preprocessed to remove outliers, fill in missing values, select attributes, integrate data and derive attributes. The preprocessing phase of this study really took the highest portion of the study time.

A total of 17810 insurance claim records are used for training the models, while a separate 2210 records are used for testing their performance. The model developed using the J48 decision tree algorithm has showed highest classification accuracy of 99.96%. This model is then tested with the 2210 testing dataset and scored a prediction accuracy of 97.19%. The results of this study have showed that the data mining techniques are valuable for insurance fraud detection. Hence future research directions are pointed out to come up with an applicable system in the area.

CHAPTEER ONE

BACKGROUND

1.1 Introduction

The insurance industry has historically been a growing industry. It plays an important role in insuring the economic well being of one country. The insurance services provided to industry and individuals have far-reaching benefits both for those who insure and for the country as a whole. It provides a form of peace of mind, or security (Dickson and Stein 1999), which is a vital importance in the industry and commerce. Because of the rapid progress of information technology, the amount of information stored in insurance databases is rapidly increasing. These huge databases contain a wealth of data and constitute a potential goldmine of valuable business information. As new and evolving loss exposures emerge in the ever-changing insurance environment, the form and structure of insurance databases change. In addition, new applications such as dynamic financial analysis and catastrophe modeling require the storage, retrieval, and analysis of complex multimedia objects, which are often represented by high-dimensional feature vectors (Guo 2003). Because of the vast amounts of data that are available in those databases, finding the valuable information hidden there and identifying appropriate models is a difficult task.

The insurance industry entertained lots of claims from the customers' everyday. The claims presented to an insurance company for payment may include a variety of different components (DíArcy et al. 2006). One component is a valid expense that should be paid in full by the insurer, since both the amount is appropriate and the coverage is applicable. Another component could be an excessive charge on a claim that would otherwise be covered. A charge is considered excessive if it is judged by the insurer to be unreasonable; most insurance policies cover only reasonable charges with reasonability defined by context and ultimately determined by negotiation, arbitration or, if necessary, lawsuit. A third component could be a claim for a service that is not covered although other services would be covered. A final component could be for an incident that is not

covered by the insurance policy. Sorting out the different components of a claim efficiently is a constant process with a claims department of insurance company.

Nowadays, researchers attempt to apply data mining (DM) for claim investigation (Woodfield 2005; Koh and Gervais 2010; Rejesus et al. 2004). DM is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules (Berry and Linoff 2000). DM has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration (Han and Kamber 2006).

DM combines techniques from machine learning, pattern recognition, statistics, database theory, and visualization to extract concepts, concept interrelations, and interesting patterns automatically from large corporate databases (Guo 2003). Its primary goal is to extract knowledge from data to support the decision-making, planning and problem solving process. Two primary functions of DM are prediction and description (Han and Kamber 2006). Prediction involves finding unknown values/relationships/patterns from known values; and description provides interpretation of a large database. Classification is useful for prediction, whereas clustering, pattern discovery and deviation detection are for description of patterns in the data.

DM methodology often can improve upon traditional statistical approaches to solving business solutions (SAS Institute Inc. 1999). For example, linear regression may be used to solve a problem because insurance industry regulators require easily interpretable models and model parameters. DM often can improve existing models by finding additional, important variables, identifying interaction terms and detecting nonlinear relationships. Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs. According to SAS Institute Inc. (1999), DM can help, specifically, insurance firms in business practices such as:

- Optimizing products and pricing
- Acquiring new customers
- Retaining existing customers
- Performing sophisticated campaign management
- Estimating outstanding loss reserve
- Detecting fraudulent claims

Fraud encompasses a wide range of illicit practices and illegal acts involving intentional deception or misrepresentation. The Institute of Internal Auditors' International Professional Practices Framework (IPPF) defines fraud as: "… Any illegal act characterized by deceit, concealment, or violation of trust." These acts are independent upon the threat of violence or physical force. Frauds are perpetrated by parties and organizations to secure personal or business advantage through unlawful act to obtain money, property, services or to avoid payment or loss of services.

Fraud is a well-known phenomenon in insurance markets. It has probably existed ever since the inception of the insurance industry itself (ACL Services Ltd. 2010). Policyholders have the ability to use their informational advantage about the occurrence of an insured loss to report losses that never have happened. Economists usually concentrate on an incentive compatible truth-telling mechanism that induces the agent to reveal his private information honestly. But in many situations that mechanism is not feasible, especially in the insurance fraud context, when the principal cannot commit based on forecasts rather than actual results to his audit strategy (Schiller 2003).

SAS Institute Inc. (1999) stated that fraudulent claims are typically not the biggest claims, claims that will cost the insurer high, because perpetrators are well aware that the big claims are scrutinized more rigorously than average claims, claims that costs medium. Perpetrators of fraud use more subtle approaches. As a result, when searching for fraudulent claims, analysts must look for unusual associations, anomalies or outlying patterns in the data.

DM can be applied in detecting fraudulent claims in the insurance industry. Modern success stories for data-mining fraud detection include applications in credit-card fraud

and telecommunications fraud, which are much easier to model than insurance fraud (Woodfield 2005). Woodfield further noted that, credit card and telecommunications fraud is easily identified within a relatively short period of time because a customer contests a bill or reports a stolen credit card. Essentially, the absence of a payment is a key component of fraud in credit and telecommunications applications. With insurance fraud, if the fraudulent behavior is not discovered, the insurer never knows that the fraud has occurred.

1.2 Statement of the Problem

Companies in the Insurance industry collected enormous amounts of data about their clients. This is invaluable information about customers' behaviour, activities, and preferences. To extract information from the whole amount of raw data Insurance firms lost time and efforts. The data extracting process resulted in developing new products and services to meet customers' needs. It unearthed information on customers, markets, and competitors. But some essential dependencies and patterns just cannot be discovered by a human in terabytes of data, while increasing the competition made Insurance companies to become more effective and customer-centric.

Insurance underwriting has to do with the selection of subjects for insurance in such a manner that general company objectives are met. The profitability of underwriting depends on accurate fraud detection. Hence one of the main goals of the underwriting management is to protect the company against high claim cost due to fraudulent claims. An increasing claim cost can lead to very serious problem for the company and, more often than not, adequate supervision at an early stage could help avert these problems (Dockrill 2001).

While insurers have increased their efforts in detecting fraud much more can still be done with the help of information technology like DM. Most carriers use human generated rules to screen exceptional claims. These business rules are derived from the opinions of experts with substantial experience in detecting claims that need further investigation or referral to case managers. The process involves taking sample claims by employees who are trained in the use of these business rules to spot abnormal patterns. The success rate

of these workers varies from individual to individual. This approach is very laboured intensive and hence costly and error prone. Also many claims go unaudited due to the sampling process that is necessitated by the high labour cost of checking claims by existing workers.

Fraudulent claims are a serious financial burden on insurers and result in higher overall insurance costs (IBM Corporation 2010). According to IBM Corporation, Insurance companies lose millions of dollars each year through fraudulent claims, largely because they do not have a way to easily determine which claims are legitimate and which may be fraudulent. Insurance fraud affects not only the financial health of the insurers, but also of innocent people seeking effective insurance coverage. As indicated by IBM Corporation, fraud can range from exaggerated losses to deliberately causing an accident for the payout.

The business of insurance is to pay claims in a timely and efficient manner. Companies are well aware that claimants and providers may have opportunities and incentives to take advantage of accidents, even fabricate or cause them to happen, to obtain payments they might otherwise not deserve (Derrig 2002). As further indicated by Derrig, the claim adjusting process is in theory a narrowing of the information asymmetry (the claimant knows exactly what happened; the company knows some of what happened) to the point that an appropriate payment is made or the claim is denied. Adjusters routinely investigate claims and negotiate settlement.

As the researcher understood from the discussion conducted with experts of AIC, there are obstacles in processing claims. For example, for processing a single claim it takes much time, which delays the response time for the policyholders. The company is now trying to handle fraud suspicious claims through its experienced experts in claim processing. Currently, AIC has no any set procedures that are used for handling fraud suspicious insurance claims, which are arriving everyday for the company. Due to the sheer number of claims submitted each day, it would be far too expensive for the company to have employees check each claim for symptoms of fraud. That means, there is shortage of experts for surveying the different claims reported to the company. And also the decisions given by different experts are not always the same, which will lead a

fraud to happen. Generally, this costs the company a lot both financially and time wise. So the manual processing of claims is not effective and efficient in identifying fraudulent cases.

Today the use of DM technology as a support in business decision-making is growing fast. There are attempts made to apply the DM technology in solving business problems in the Ethiopian context, including Airlines (Gobena 2000; Henock 2002; Denekew 2003), Banking (Askale 2001; Samson 2009; Tilahun 2009), Telecommunications (Fekadu 2004; Melkamu 2009), shipping (Kumneger 2006), and healthcare (Shegaw 2002; Biru 2009).

There are also studies conducted to investigate the application of DM in the insurance business. Tesfaye (2002) used the DM technique for insurance risk assessment in Nyala insurance company. He develops a predictive model using neural network classification technique. The other research, which is undertaken by Mesfin (2005), also attempts to investigate the possible application of predictive DM techniques in the renewal process of personal accident policies as a case study on Ethiopian Insurance Corporation. However, as to the researcher knowledge no one investigates the application of DM for fraud detection in local context. It is, therefore, with this understanding that the present study is conducted to develop a predictive model using DM technology for predicting and detecting insurance claim fraud in Africa Insurance Company (AIC).

As can be seen from the current practice of the company, there is no as such an automatic means, which will help the company to show those claims which have a high degree of exposition to be fraudulent. This research attempts to answer:

- What is the pattern that characterizes whether a given claim is fraudulent or not?
- Which DM algorithm can be more suitable for the purpose of identifying/predicting fraudulent insurance claims?
- What is the possible segment that claims can be clustered according to their natural groupings? And which DM algorithm is suitable for doing this task?

1.3 Objectives of the Study

1.3.1 General objective

The general objective of this study is applying the DM techniques for predicting and detecting fraudulent insurance claims, in order to provide effective and efficient services for customers and keeping the company stable and competent enough in the market environment.

1.3.2 Specific objectives

For the realization of the general objective stated above, the following specific objectives are formulated.

- To review literature on DM technology and their application in the insurance industry for the purpose of getting information that will help in this research
- To identify sources and collect the required data from Africa Insurance Company
- To prepare the data for analysis and model building, by cleaning and transforming the data into a format suitable for the selected DM algorithms
- To identify the features of insurance fraud, and select the appropriate clustering and classification algorithms to be used based on the type of data and objectives of the study
- To train and develop a classification model that will help to predict and detect fraudulent insurance claims
- To test and compare the resulting performances of the clustering and classification models and recommend the overall best results of the clustering and classification models.
- To report findings of the result and forward recommendation for further research.

1.4 Scope and Limitation of the Study

The main aim of this research is exploring the applicability of DM for claim fraud prediction and detection in the insurance sector. This research focuses only to the claim department of AIC. Specifically the mined data for knowledge discovery is obtained from the Addis Ababa branches of AIC, which are networked to each other at the main office. Out of the various insurances offered by the company, the data related to motor insurance is considered. Motor insurance is one of the most effective working areas of AIC.

To achieve the objectives of this study, a two step DM technique is used; i.e. first the study apply clustering technique to define the natural group of records and then classification to develop prediction model, which helps to identify motor insurance fraud.

This research was aimed to include all-important information for solving the study problem. However, some attributes like garage report, and traffic police report are not included in this study because the data was not available. As described above this research only attempted to apply DM techniques in predicting and detecting fraudulent motor insurance claims. The motor insurance section was selected because of its wide coverage from the other services that the company provides. However, researches can also be conducted in other sections of the company other than the motor section. Although this research was aimed to include all branches of the company throughout the country, the data used in this study was only collected from the five Addis Ababa branches of the company. These five branches were chosen because of their activeness in the company currently. Due to time and financial matters this research didn't include the data from the regional branches of the company. So, further research can be conducted including the data from these branches.

1.5 Research Methodology

The design of a framework for a knowledge discovery process is an important issue. Several researchers described a series of steps that constitute the Knowledge Discovery process to be followed by practitioners when executing a DM project. As described by Cios and Kurgan (2005), the process models range from very simple models,

incorporating few steps to more sophisticated models (like the nine-step model proposed by Fayyad et al. 1996). The DM process model describes procedures that are performed in each of its steps (Jinhong et al. 2009). It is primarily used to plan, work through, and reduce the cost of any given project.

1.5.1 Research design

For the purpose of conducting this research the six-step process model of Cios et al. (2000) is selected. This model was developed, by adopting the CRISP-DM model to the needs of academic research community. Unlike the CRISP-DM process model, which is fully industrial, the Cios et al. process model is both academic and industrial. The main extensions of the Cios et al. process model include providing a more general, research-oriented description of the steps, an introduction of a DM step instead of the modeling step, and an integration of several explicit feedback mechanisms. The Cios et al. (2000) model consists of understanding the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge steps.

1.5.2 Understanding of the Insurance problem domain

A close look of the problem environment is the first step taken in this Knowledge discovery DM research. On the basis of the insights gained from this phase, the DM problem is defined. To clearly identify, understand, and analyze the business problems, the primary (observation and interview), and secondary (database analysis) data collection methods are employed. Interview is employed to define features selection with the domain experts while observation is conducted to understand some complex business processes. Further, databases such as claim, vehicle, and policy along with insured's information are consulted to gather the pertinent data for the present research.

The main goal of the claim processing department of AIC is to reduce operational expense through increased efficiency and productivity throughout the process chain, improve service levels by implementing a faster, more visible, and consistent approach to claims settlement and reduce indemnity cost through leakage control and proactive fraud management.

The main DM goal for this research is identifying and detecting fraudulent insurance claims in order to attain the goals of the claims processing and underwriting departments. For that matter, a model is developed using the different DM techniques, which helps to predict fraudulent insurance claims.

1.5.3 Understanding the data

After understanding the problem to be addressed clearly in this study, the next step is analyzing and understanding the data available. The outcome of DM and knowledge discovery heavily depends on the quality and quantity of available data (Cios et al. 2007). The data that is used in this research was initially collected, regarding motor insurance claim, from AIC, which is a privately owned insurance company.

At this stage, the data that is used in this research is described briefly. The description includes listing out attributes, their respective values, data types, and evaluation of their importance etc… as well as visualization of the data to see data distribution branch wise. Careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the particular DM tasks to be performed.

1.5.4 Preparation of the data

This is the key step upon which the success of the entire knowledge discovery process depends. It usually consumes about half of the entire research effort. Today's real-world databases are highly susceptible to noisy, missing values, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources (Han and Kamber 2006). In this step, the researcher decides, together with domain experts, the data that is used as input for applying the DM techniques.

The different data preprocessing techniques are applied for processing the data used by the algorithms chosen for the research. Data cleaning (or data cleansing) routines are applied to fill in missing values (with the mean value), smooth out noise (by removing the record), and detect outliers (by removing or substituting with mean values) in the

data. The cleaned data is further processed by feature selection consulting the domain experts and the Weka attribute selection preprocessing techniques (to reduce dimensionality) and by derivation of new attributes. The result of these processes generates datasets for training and testing the clustering and classification algorithms selected in this study.

1.5.5 Data mining

Based on the identified goals and the assessment of the available data, appropriate mining algorithm is chosen and run on the prepared data. In many applications of machine learning to DM, the explicit knowledge structures that are acquired, namely the structural descriptions, are at least as important, and often very much more important, than the ability to perform well on new examples. The use of DM to gain knowledge discovery regularities in the data and not just predictions is common (Witten and Frank, 2000).

As one can tell from the objectives of the research, gaining knowledge, discovering regularities for detecting fraudulent claims with in the insurance dataset in particular, is certainly the purpose of this study. Having this purpose in mind, the unsupervised clustering technique and the supervised classification technique are adopted. The clustering technique is selected because of the reason that the dataset, which is obtained from AIC doesn't have a feature indicating whether a claim was fraud or not. So, the clustering technique is used to group the insurance claims that are susceptible to fraud, into a similar cluster. For doing this the K-Means clustering algorithm is employed since we have two classes only. The clustered dataset is then used as an input for the classification algorithm for classifying instances of the dataset into similar class labels. For this purpose, the J48 decision tree algorithm and the Naïve Bayes classification methods are used to create model for detecting fraudulent insurance claims.

1.5.5.1 Data mining tool selection

For conducting this research the WEKA (Waikato Environment for Knowledge Analysis) version 3.7.0 (for Mac OS) DM software is chosen. Weka is chosen because of its widespread application in different DM researches and familiarity of the researcher with the software.

Weka, a machine-learning algorithm in Java, is adopted for undertaking the experiment. Weka constitutes several machine learning algorithms for solving real-world DM problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from one's own Java code. Weka is open source software issued under the GNU General Public License. The Weka DM software included classification, clustering, association rule learner, numeric prediction and several other schemes. In addition to the learning schemes, Weka also comprises several tools that can be used for datasets pre-processing (Witten and Frank 2000).

1.5.6 Evaluation of the discovered knowledge

In DM evaluation serves two purposes. First, it helps to envisage how well the final model will work in the future (or even whether it should be used at all). Second, as an integral part of many learning methods, it helps to explore the model that best represents the training data. The model is evaluated together with the domain experts regarding its interestingness and novelty. The different clustering models that are developed in this research are evaluated based on the within cluster sum of squared error values, number of iteration the algorithm takes to converge, the attributes average values that satisfy the threshold, and experts' judgement. Likewise, classification models that are developed in this research are evaluated using a test dataset based on their classification accuracy.

1.6 Significance of the Study

The results of this study can help Africa Insurance Share Company and other related insurance companies:

- To save time and money which is wasted for processing and surveying the different claims that are reported to the organization everyday
- To improve the decision making process for experts especially in the claim and underwriting departments
- To identify those fraudulent areas and take timely measure accordingly
- To improve rules and procedures that are related to claims handling
- To ratify procedures for investigating the fraudulent exposure of claims.

- To create a smooth relation with its clients

This study can also be an input for further research in this and other related areas in the context of our country. Finally, this study can give hands on experience for the researcher for understanding studies in the future.

1.7 Organization of the Thesis

This thesis is organized into six chapters. The first chapter briefly discusses background to the problem area and DM technology, and states the problem, objective of the study, research methodology, scope and limitation, and significance of the results of the research.

The second chapter deals about DM technology, methods/techniques and algorithms, the different methodologies, the DM process, the different tasks of DM, and its application in the insurance sector.

The third chapter provides discussions about the different DM methods and algorithms that are used in this research. This includes discussions about decision tree method and its J48 algorithm, the Naïve Bayes method and the K-Means clustering algorithm.

The fourth chapter provides introduction about AIC and discussions about the different DM steps that are undertaken by the methodology used in this research work. This includes the business understanding, data understanding, and data preprocessing phases.

The fifth chapter provides a detailed discussion about the experimentation part of this study. This includes the clustering and classification experimentation phases. Evaluation of the discovered knowledge is also discussed at this section.

The last chapter is devoted to concluding remarks and recommendations forwarded based on the research findings of the present study.

CHAPTER TWO

DATA MINING AND KNOWLEDGE DISCOVERY

It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100% (Witten and Frank 2000). As the volume of data increases, the proportion of information in which people could understand decreases substantially. This reveals that the level of understanding of people about the data at hand could not keep pace with the rate of generation of data in various forms, which results in increasing information gap. Consequently, scholars begin to realize this bottleneck and to look into possible remedies. Current technological progress permits the storage and access of large amounts of data at virtually no cost. Although many times preached, the main problem in a current information-centric world remains to properly put the collected raw data to use (Kurgan and Musilek 2006). The true value is not in storing the data, but rather in our ability to extract useful reports and to find interesting trends and correlations, through the use of statistical analysis and inference, to support decisions and policies made by scientists and businesses (Fayyad et al. 1996). To bridge the gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerized methods known as Data Mining (DM) or Knowledge Discovery in Databases (KDD) has emerged in recent years.

Different scholars provided different definitions about DM. According to Berry and Linoff (2000); Han and Kamber (2006), DM is the process of extracting or “mining” knowledge from large amounts of data in order to discover meaningful patterns and rules. Witten and Frank (2000) have also noted that DM is valuable to discover implicit, potentially useful information from huge data stored in databases via building computer programs that sift through databases automatically or semi-automatically, seeking meaningful patterns.

DM involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large datasets (Two Crows Corporation 1999). These

tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, DM consists of more than collecting and managing data; it also includes analysis and prediction and use of algorithms that improve their performance automatically through experience, such as neural networks or decision trees.

According to Han and Kamber (2006), the major reason that DM has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. DM tools perform data analysis and may uncover important data patterns. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration (Han and Kamber 2006).

DM is an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, data visualization, optimization, information retrieval, high end computing, and others (Guo 2003; Han and Kamber 2006). DM methodology often can improve upon traditional statistical approaches for solving business solutions by finding additional, important variables, by identifying interaction among terms and detecting nonlinear relationships (SAS Institute Inc. 1999). Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs.

2.1 The Data Mining Process

DM requires massive collection of data to generate valuable information (Han and Kamber 2006; Deshpande and Thakare 2010). The data can range from simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. Deshpande and Thakare indicated that the data retrieval is simply not enough to take complete advantage of data. It requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

A typical DM process includes data acquisition, data integration, data exploration, model building, and model validation (Deshpande and Thakare 2010). Both expert opinion and DM techniques play an important role at each step of this knowledge discovery process.

2.1.1 Data acquisition

The first step in DM is to select the types of data to be used. Although a target dataset has been created for discovery in some applications, DM can be performed on a set of variables or data samples in a larger database called training set to create and model while holding back some of the datasets (test dataset) for latter validation of the model.

2.1.2 Data pre-processing

Once the target data is selected, the data is then pre-processed for cleaning, scrubbing, and transforming to improve the effectiveness of discovery. During this pre-processing step, researchers remove the noise or outliers if necessary and decide on strategies for dealing with missing data fields and accounting for time sequence information or known changes. Then data is transformed to reduce the number of variables by converting one type of data to another (e.g., numeric ones into categorical) or deriving new attributes.

2.1.3 Building model

The third step of DM refers to a series of activities such as deciding on the type of DM operations, selecting the DM algorithms, and mining the data. First, the type of DM operation (classification, regression, clustering, association rule discovery, segmentation, and deviation detection) must be chosen. Based on the operations chosen for the application, an appropriate DM technique is then selected based on the nature of the knowledge to be mined. Once a DM technique is chosen, the next step is to select a particular algorithm within the DM technique chosen. Choosing a DM algorithm includes a method to search for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular DM technique with the overall objective of DM. After an appropriate algorithm is selected, the data is finally mined using the algorithm to extract novel patterns hidden in databases.

2.1.4 Interpretation and model evaluation

The fourth step of DM process is the interpretation and evaluation of discovered patterns. This task includes filtering the information to be presented by removing redundant or irrelevant patterns, visualizing graphically or logically the useful ones, and translating them into understandable terms by users. In the interpretation of results, the researcher determines and resolves potential conflicts with previously known or decides redo any of the previous steps. The extracted knowledge is also evaluated in terms of its usefulness to a decision maker and to a business goal.

2.2 Data Mining Tasks

The DM tasks are of different types depending on the use of DM result (Hand et al. 2001). Predictive modeling, descriptive modeling, exploratory data analysis, patterns and rules discovery, and retrieval by content are some of the DM tasks.

2.2.1 Predictive modelling

Predictive modeling permits the value of one variable to be predicted from the known values of other variables. Classification, Regression, Time series analysis, Prediction etc. are some examples of predictive modeling. As Tan et al. (2009) indicated many of the DM applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes. It is a supervised learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that maps data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

2.3.1.1 Classification

Classification is the process of finding a model, which describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown (Han and Kamber 2006). The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). Classification problems aim to identify the characteristics that indicate the group to which each case belongs (Two Crows Corporation 1999). This pattern can be used both to understand the existing data and to predict how new instances will behave.

DM creates classification models by examining already classified data (cases) and inductively finding a predictive pattern (Two Crows Corporation 1999). According to the Two Crows Corporation, these existing cases may come from a historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. For example, a sample of a mailing list would be sent an offer, and the results of the mailing used to develop a classification model to be applied to the entire database. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model, which will be applied to the entire database. There are different algorithms that are used for classification purpose such as, decision tree, neural network, genetic algorithm, naïve bayes, etc.

Decision tree

A decision tree is a flow-chart-like tree structure where each internal node denotes a test on an attribute each branch represents an outcome of the test and leaf nodes represent classes or class distributions (Han and Kamber 2006). Decision trees are trees that classify instances by sorting them based on feature values (Two Crows Corporation 2005). They are a way of representing a series of rules that lead to a class or value. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values (Phyu 2009). In DM, a decision tree is

a predictive model, which can be used to represent both classifiers and regression models (Hajizadeh et al. 2010).

A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable (Hajizadeh et al. 2010). The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the categories, or to classify the record by assigning it to the most likely class. Decision tree can also be used to estimate the value of continuous variable.

The decision node, branches and leaves are the basic components of a decision tree (Apte and Weiss 1997). Depending on a decision tree algorithm, each node may have two or more branches. For example, CART (Classification and Regression Tree) generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed it is called a multi-way tree. Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node. By navigating the decision tree you can assign a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached. Each node uses the data from the case to choose the appropriate branch.

Decision trees are generated from training data in a top down general to specific direction (Apte and Weiss 1997). The initial state of a decision tree is the root node that is assigned all the examples from the training set. If it is the case that all examples belong to the same class then no further decisions need to be made to partition the examples and the solution is complete. If examples at this node belong to two or more classes then a test is made at the node that will result in a split. The process is recursively repeated for each of the new intermediate nodes until a completely discriminating tree is obtained. A decision tree at this stage is potentially an over-fitted solution i.e. it may have components that are too specific to noise and outliers that may be present in the training data. As Apte and Weiss (1997) indicated, to relax this over-fitting most decision tree methods go through a second phase called pruning that tries to generalize the tree by eliminating sub trees that seem too specific. Error estimation techniques play a major role in tree pruning. Most modern decision tree modeling algorithms are a combination of a specific type of a

splitting criterion for growing a full tree and a specific type of a pruning criterion for pruning tree.

The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules (Hajizadeh et al. 2010). Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved. Decision tree has its own properties. The following are some of them:

- Learns with positive and negative examples
- Noise tolerant
- General-to-specific search (reverse for pruning)
- Follows Divide-and-Conquer strategy, which has weaknesses of fracturing and diminishing training data.
- Learns discriminating rules

Decision tree can be implemented with several algorithms. Some of them are J48, ID3, C4.5, CART, etc. J48 is an implementation of C4.5 release 8(3) that produces decision trees (Meera and Srivatsa 2010). This is a standard algorithm that is used for machine learning. C4.5 is a decision tree-learning algorithm that builds upon the ID3 algorithm as indicated by Lavesson (2003). Amongst other enhancements (compared to the ID3 algorithm) the C4.5 algorithm includes different pruning techniques and can handle numerical and missing attribute values. C4.5 avoids over fitting the data by determining a decision tree, it handles continuous attributes, is able to choose an appropriate attribute selection measure, handles training data with missing attribute values and improves computation efficiency. C4.5 builds the tree from a set of data items using the best attribute to test in order to divide the data item into subsets and then it uses the same procedure on each sub set recursively. The main problem in decision tree is deciding the attribute, which will best partition the data into various classes (Meera and Srivatsa 2010). The ID3 algorithm is useful to solve this problem.

Neural networks

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, they imitate the way the human brain learns and use rules inferred from data patterns to construct hidden layers of logic for analysis (Singh and Chauhan 2005). Neural networks constitute the most widely used technique in DM. As Hajek (2005) stated, a neural network is a massively parallel-distributed processor that has a natural tendency for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network through a learning process
2. Interneuron connection strengths known as synaptic weights are used to store the knowledge.

A neural network is first and foremost a graph, with patterns represented in terms of numerical values attached to the nodes of the graph and transformations between patterns achieved via simple message-passing algorithms (Jordan and Bishop 1996). Generally, a neural network can be described as a directed graph in which each node performs a transfer function of the form

$$yi = (f i\sum_{j=1}^{n} Wij\,Xj - Qi) \qquad \text{…….. ….2.1}$$

Where y_i is the output of the node i, x_j is the j^{th} input to the node, and W_{ij} is the connection weight between nodes i and j. Q_i is the threshold (or bias) of the node. Certain of the nodes in the graph are generally distinguished as being input nodes or output nodes, and the graph as a whole can be viewed as a representation of a multivariate function linking inputs to outputs. Numerical values (weights) are attached to the links of the graph, which parameterize the input/output function and allowing it to be adjusted via a learning algorithm.

Neural network topologies can be divided into feed forward and recurrent classes

according to their connectivity (Yao 1999; Singh and Chauhan 2005). The feed forward neural network was the first and arguably simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. A neural network is feed forward if there exists a method, which numbers all the nodes in the network such that there is no connection from a node with a large number to a node with a smaller number. All the connections are from nodes with small numbers to nodes with larger numbers. A neural network is recurrent if such a numbering method does not exist. Contrary to feed forward networks, recurrent neural networks (RNs) are models with bi-directional data flow. While a feed forward network propagates data linearly from input to output, RNs also propagate data from later processing stages to earlier stages.

Learning in ANN's can roughly be divided into supervised, unsupervised, and reinforcement learning (Yao 1999; Singh and Chauhan 2005). Supervised learning or Associative learning is based on direct comparison between the actual output of an ANN and the desired correct output, also known as the target output. Reinforcement learning is a special case of supervised learning where the exact desired output is unknown. It is based only on the information of whether or not the actual output is correct. Unsupervised learning or Self-organization is solely based on the correlations among input data. No information on "correct output" is available for learning.

According to Larose (2006) there are two general categories of neural net algorithms: supervised and unsupervised. Supervised neural net algorithms such as Back propagation and Perceptron require predefined output values to develop a classification model. Among the many algorithms, Back propagation is the most popular supervised neural net algorithm (Han and Kamber 2006). Unsupervised neural net algorithms such as ART do not require predefined output values for input data in the training set and employ self organizing learning schemes to segment the target dataset.

For organizations with a great depth of statistical information, ANNs are ideal because they can identify and analyze changes in patterns, situations, or tactics far more quickly than any human mind, as indicated by Guo (2003). Although the neural net technique has strong representational power, interpreting the information encapsulated in the weighted

links can be very difficult. One important characteristic of neural networks is that they are opaque, which means there is not much explanation of how the results come about and what rules are used. Therefore, some doubt is cast on the results of the DM.

2.2.1.1 Prediction

Prediction is the other example of predictive modeling. It can be viewed as the construction and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of an attribute that a given object is likely to have (Han and Kamber 2006).

2.2.2 Descriptive modelling

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined (Deshpande and Thakare 2010). It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables. Clustering, Association rule discovery, Sequence discovery, Summarization, etc. are some of the examples. Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone (Han and Kamber 2006). Summarization is the technique of presenting the summarized information from the data. The association rule finds the association between the different attributes. Association rule mining is a two-step process: Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend.

2.2.2.1 Clustering

Clustering is a DM (machine learning) technique that finds similarities between data according to the characteristics found in the data and group's similar data objects into one cluster. The objective of clustering is to distribute cases (people, objects, events etc.) into groups, so that the degree of association can be strong between members of the same cluster and weak between members of different clusters (Hajizadeh et al. 2010). Clustering techniques are employed to segment a database into clusters, each of which

shares common and interesting properties (Two Crows Corporation 2005). The purpose of segmenting a database is often to summarize the contents of the target database by considering the common characteristics shared in a cluster. Clusters are also created to support the other types of DM operations, e.g. link analysis within a cluster (Guo 2003). Clustering tools assign groups of records to the same cluster if they have something in common, making it easier to discover meaningful patterns from the dataset (Qiu et al. 2004). Clustering often serves as a starting point for some supervised DM techniques or modeling.

Clustering is one of the most useful tasks in DM process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given dataset into groups such that the data points in a cluster are more similar to each other than points in different clusters (Guha et al. 1998). For example, segmenting existing insurance policyholders into groups and associating a distinct profile with each group can help future rate making strategies.

Clustering methods perform disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative variables and seeds that are generated and updated by the algorithm. You can specify the clustering criterion that is used to measure the distance between data observations and seeds. The observations are divided into clusters such that every observation belongs to at most one cluster.

Clustering studies are also referred to as unsupervised learning and/or segmentation. Unsupervised learning is a process of classification with an unknown target, that is, the class of each case is unknown. The aim is to segment the cases into disjoint classes that are homogenous with respect to the inputs. Clustering studies have no dependent variables. You are not profiling a specific trait as in classification studies.

Cluster analysis is related to other techniques that are used to divide data objects into groups. For instance, clustering can be regarded as a form of classification in that it creates a labeling of objects with class (cluster) labels.

Clustering techniques are heuristic in nature (Qiu et al. 2004). Almost all techniques

have a number of arbitrary parameters that can be "adjusted" to improve results. According to Rao (2003), clustering techniques can be divided broadly into two approaches:

- Partitioning clustering approach: - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors. K-Means clustering and expectation maximization (EM) clustering are the two methods of partitioning clustering.
- Hierarchical clustering approach: - Create a hierarchical decomposition of the set of data (or objects). It can be visualized as a dendrogram; a tree like diagram that records the sequences of merges or splits.

Hierarchical clustering approach is further subdivided into agglomerative and divisive.

a) Agglomerative: Start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. It is a Bottom Up clustering technique. This requires a definition of cluster similarity or distance.
b) Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. It is a Top Down clustering technique. In this case, we need to decide, at each step, which cluster to split and how to perform the split.

K-Means clustering algorithm

The k-Means algorithm is very widely used to produce clustering of data, due to its simplicity and speed (Graham 2006; Wu et al. 2007). It is a simple iterative method to partition a given dataset into a user-specified number of clusters, k (Wu et al. 2007). The idea is based around clustering items using centroids. These are points in the metric space that define the clusters. Each centroid defines a single cluster, and each point from the data is associated with the cluster defined by its closest centroid.

The algorithm operates on a set of d-dimensional vectors, $D = \{xi \mid i = 1 \ldots N\}$, where xi $\in \Re^d$ denotes the i^{th} data point. Picking k points in $\Re^d$ as the initial k cluster representatives or "centroids" initializes the algorithm. Techniques for selecting these

initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data *k* times. Then the algorithm iterates between two steps till convergence (Wu et al. 2007). The first step is Data Assignment. Here each data point is assigned to its closest centroid, with ties broken arbitrarily. This results in a partitioning of the data. The second step is Relocation of "means". Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions.

The K-Means algorithm is simple, easily understandable and reasonably scalable, and can be easily modified to deal with streaming data. However, one of its drawbacks is the requirement for the number of clusters, K, to be specified before the algorithm is applied (Pham et al. 2005).

2.2.2.2 Association rule discovery

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database (Kotsiantis and Kanellopoulos 2006). The problem is usually decomposed into two sub problems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence.

Association rule aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories (Kotsiantis and Kanellopoulos 2006). Given a collection of items and a set of records containing some of these items, association discovery techniques discover the rules to identify affinities among the collection of items as reflected in the examined records (Guo 2003). For example, 65 percent of records that contain item A also contain item B. An association rule uses measures called "support" and "confidence" to represent the strength of association. The percentage of occurrences, 65 percent in this case, is the confidence factor of the association.

According to Guo (2003), the efficiency with which association discovery algorithms can organize the events that make up an association or transaction is one of the differentiators among the association discovery algorithms. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. There are a variety of algorithms to identify association rules. The most widely used association rule algorithms are Apriori and FP-growth tree. Apriori is an influential algorithm for finding frequent itemsets using candidate generation (Wu et al. 2007). Frequent-pattern tree, or FP-tree in short is an extended prefix-tree structure storing crucial, quantitative information about frequent patterns. FP-growth method is an efficient and scalable mining for both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm (Han et al. 2004).

2.3 Types of Data Mining Systems

There are many DM systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited DM functionalities, other are more versatile and comprehensive. DM systems can be categorized according to various criteria (Han and Kamber 2006; Deshpande and Thakare 2010).

DM systems can be classified according to the type of data source mined. This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc. The other Classification of DM systems are according to the data model. This classification is based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc. Further Classification of DM systems are according to the kind of knowledge discovered. This classification of DM systems based on the kind of knowledge discovered or DM functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several DM functionalities together. Finally, DM systems can be classified according to mining techniques used. This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, and visualization.

The classification can also take into account the degree of user interaction involved in the DM process such as query-driven systems, interactive exploratory systems, or autonomous systems (Deshpande and Thakare 2010). According to Deshpande and Thakare, a comprehensive system would provide a wide variety of DM techniques to fit different situations and options, and offer different degrees of user interaction.

2.4 The Data Mining Models

There are different DM process model standards. The six step Cios et al. (2000) model, KDD process (Knowledge Discovery in Databases), CRISP-DM (Cross Industry Standard Process for Data Mining), and SEMMA (Sample Explore Modify Model Assess), are some of the models that are used in different DM projects.

2.4.1 The six step Cios model

As described in Section 1.4 of Chapter 1 this model was developed, by adopting the CRISP-DM model to the needs of academic research community. The model consists of six steps (Cios and Kurgan 2005).

1. **Understanding of the problem domain:** In this step one works closely with domain experts to define the problem and determine the research goals, identify key people, and learn about current solutions to the problem. A description of the problem including its restrictions is done. The research goals then need to be translated into the DM goals, and include initial selection of the DM tools.
2. **Understanding of the data:** This step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the DM goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.
3. **Preparation of the data:** This is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire research effort. In this step, which data will be used as input for DM tools of step 4, is decided. It may involve sampling of data, data cleaning like checking

completeness of data records, removing or correcting for noise, etc. The cleaned data can be, further processed by feature selection and extraction algorithms (to reduce dimensionality), and by derivation of new attributes (say by discretization). The result would be new data records, meeting specific input requirements for the planned to be used DM tools.

4. **Data mining:** This is another key step in the knowledge discovery process. Although it is the DM tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned DM tools and selection of the new ones. DM tools include many types of algorithms, such as neural networks, clustering, preprocessing techniques, Bayesian methods, machine learning, etc. This step involves the use of several DM tools on data prepared in step 3. First, the training and testing procedures are designed and the data model is constructed using one of the chosen DM tools; the generated data model is verified by using testing procedures.
5. **Evaluation of the discovered knowledge:** This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only the approved models are retained. The entire DM process may be revisited to identify which alternative actions could have been taken to improve the results.
6. **Using the discovered knowledge:** This step is entirely in the hands of the owner of the database. It consists of planning where & how the discovered knowledge will be used. The application area in the current domain should be extended to other domains.

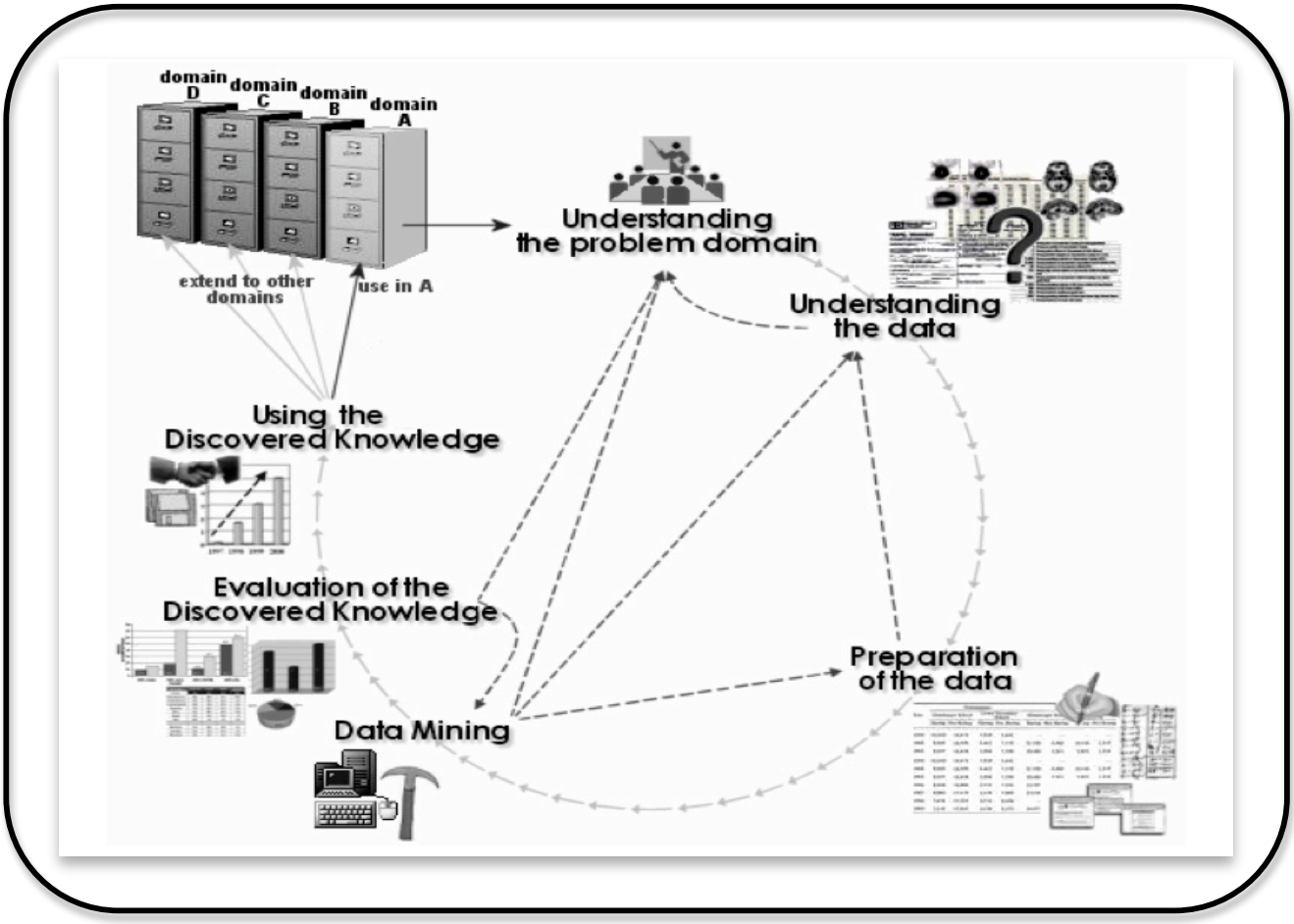

Figure 2. 1 The Six Step Cios et al. (2000) process model

2.4.2 The KDD process model

KDD process is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database as presented by Azevedo and Santos (2008). It is an interactive and iterative process, comprising a number of phases requiring the user to make several decisions. Generally, there are five steps in the KDD process (Two Crows Corporation 1999; Azevedo and Santos 2008):

1. **Data selection:** This stage consists on creating a target dataset, or focusing on a subset of variables or data samples, on which discovery is to be

performed. The data relevant to the analysis is decided on and retrieved from the data collection.

2. **Data pre-processing:** This stage consists on the target data cleaning and pre processing in order to obtain consistent data
3. **Data transformation:** It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure. This stage consists on the transformation of the data using dimensionality reduction or transformation methods
4. **Data mining:** It is the crucial step in which clever techniques are applied to extract potentially useful patterns. It consists on the searching for patterns of interest in a particular representational form, depending on the DM objective.
5. **Interpretation/Evaluation:** This stage consists on the interpretation and evaluation of the mined patterns.

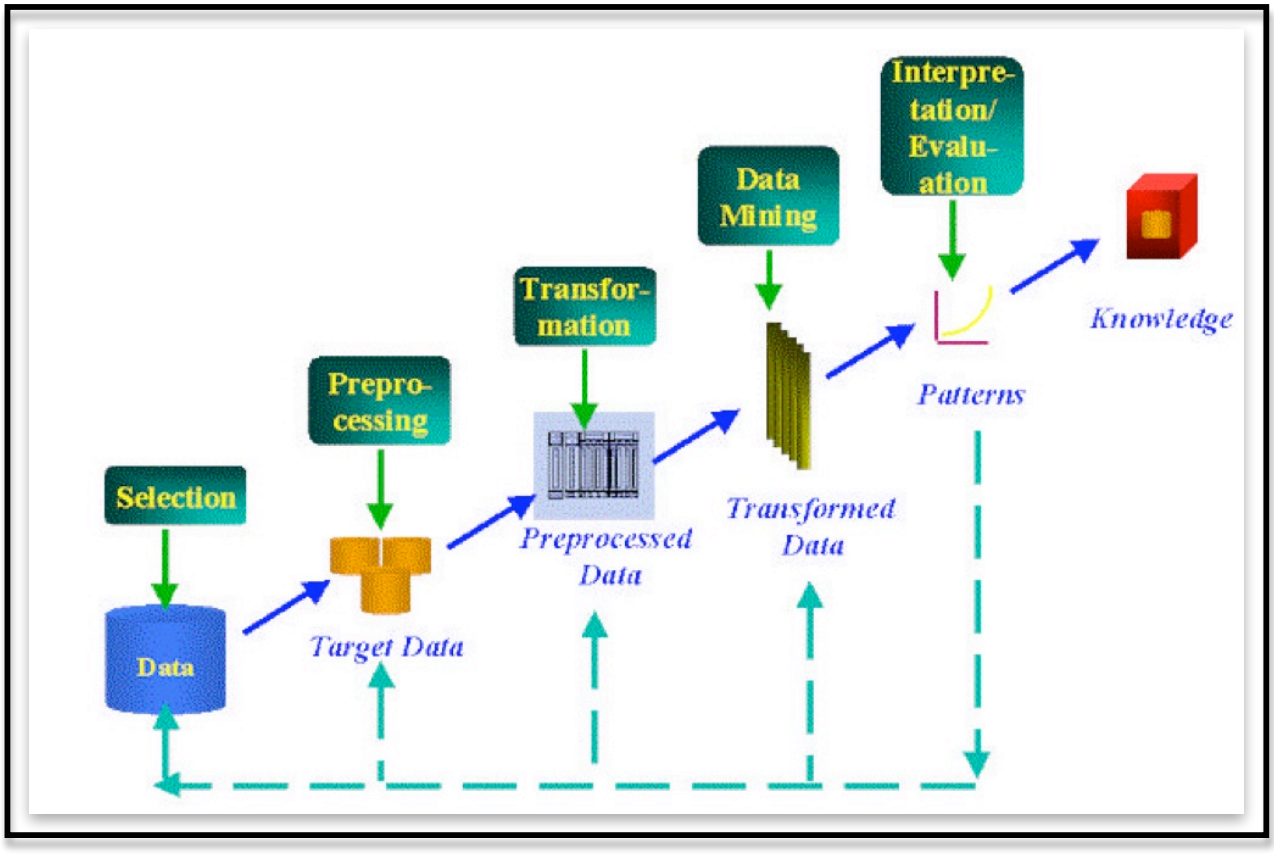

***Figure 2. 2** The KDD Process*

As indicated above, a KDD process involves preprocessing data, choosing a data-mining algorithm, and post processing the mining results. There are very many choices for each of these stages, and non-trivial interactions between them. Therefore both novices and DM specialists need assistance in KDD processes.

2.4.3 The CRISP-DM process

CRISP-DM (Cross Industry Standard Process for Data Mining), process model was first established by four companies in the late 1990s (Chapman et al. 2000; Kurgan and Musilek 2006; Azevedo and Santos 2008). These were Integral Solutions Ltd. (a provider of commercial DM solutions), NCR (a database provider), DaimlerChrysler (an automobile manufacturer), and OHRA (an insurance company).

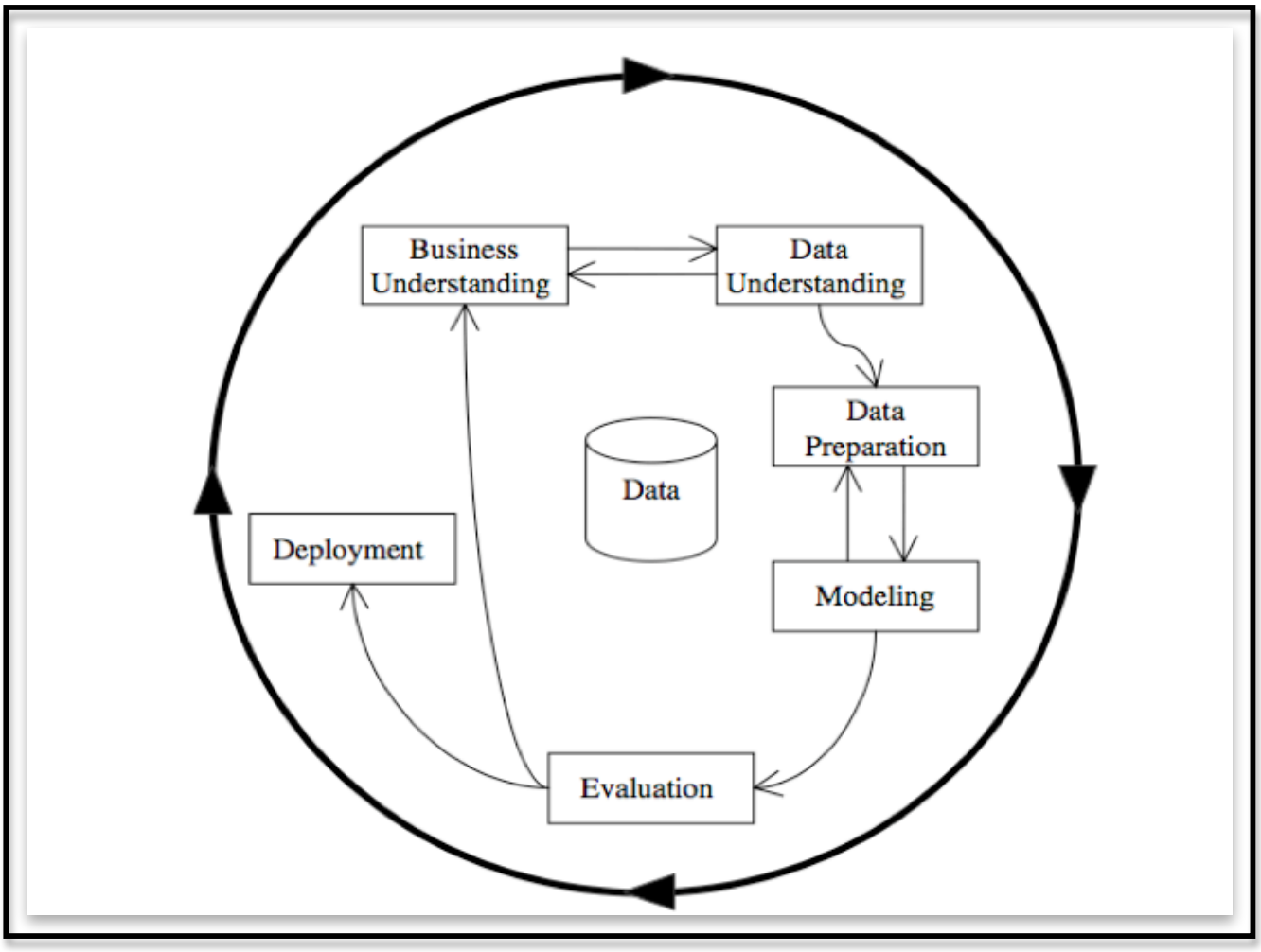

Figure 2. 3 *The CRISP-DM Process*

According to CRISP-DM the life cycle of a data-mining project consists of six phases (Chapman et al. 2000; Azevedo and Santos 2008). The sequence of the phases in the CRISP-DM process is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The CRISP-DM process has six stages.

1. **Business understanding:** This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
2. **Data understanding:** It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
3. **Data preparation:** It covers all activities to construct the final dataset from the initial raw data.
4. **Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
5. **Evaluation:** In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the DM results should be reached.
6. **Deployment:** The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

2.5 Application of Data Mining

DM is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use DM to reduce costs, enhance research, and increase sales (Seifert 2004). These days DM is widely used for fraud detection. For example, the insurance and banking industries use DM applications to detect fraud and assist in risk assessment (e.g., credit scoring). According to Seifert, using customer data collected over several years, companies can develop models that

predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely.

2.5.1 Data mining in the insurance industry

DM can help insurance firms make crucial business decisions and turn the new found knowledge into actionable results in business practices such as product development, marketing, claim distribution analysis, asset liability management and solvency analysis (Guo 2003). According to Guo, the Insurance industry can make better use of modern DM technologies to develop more accurate and better performing models that are generated in less time.

Insurance companies around the world lose more and more money through fraudulent claims each year (IBM Corporation 2010). They need to recoup this lost money so they can continue providing superior services for their customers. According to the IBM Corporation, agencies can now combine DM with existing fraud detection and prevention efforts to improve accuracy, decrease manpower and minimize loss.

Specifically, DM can help insurance firms in business practices (SAS Institute Inc. 1999) such as optimizing products and pricing, customer relationship management (such as Acquiring new customers, Retaining existing customers and customer segmentation), performing sophisticated campaign management, reinsurance, estimating outstanding loss reserve, and detecting fraudulent claims.

2.5.1.1 Optimizing products and pricing

As a result of changing demographics, economic factors and customer buying habits, it is critical that insurance companies identify and monitor the varying needs of their customers and adjust their product portfolio. Problems with profitability can occur if firms do not offer the right policy, rate or customer segment at the right time. For example, the most profitable customer segment might be higher-risk customers, which may command higher rates.

An important problem in actuarial science concerns rate setting or the pricing of each

policy. The goal is to set rates that reflect the risk level of the policyholder by establishing the "break-even" rate (or premium) for the policy. The lower the risk the lower the rate is. Although many risk factors that affect rates are obvious, subtle and non-intuitive relationships can exist among variables that are difficult if not impossible to identify without applying more sophisticated analysis. Modern DM models can more accurately predict risk, therefore insurance companies can set rates more accurately, which in turn results in lower costs and greater profits.

2.5.1.2 Customer relationship management (CRM)

CRM is an enterprise approach to understand and influence customer behaviour through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability (Kim et al. 2003). According to Parvatiyar and Sheth (2001) CRM can also be defined as a comprehensive strategy and process of acquiring, retaining, and partnering with selective customers for the purpose of creating superior value for both the company and the customer. To achieve a better efficiencies and effectiveness in delivering customer value, CRM involves the integration of marketing, sales, customer service, and the supply-chain functions of the organization. Generally, CRM involves acquisition of customers, customer retention and customer segmentation.

The acquisition of new customers is an important business problem related to ratemaking. Traditional approaches involve attempts to increase the customer base by simply expanding the efforts of the sales department. In contrast to traditional sales approach, DM strategies enable analysts to define the marketing focus. Analysts in the insurance industry can utilize advanced DM techniques that combine segmentations to group the high lifetime-value customers and produce predictive models to identify those in this group who are likely to respond to marketing campaign. Using a DM technique called association analysis, insurance firms can more accurately select which policies and services to offer to which customers.

As acquisition costs increase, insurance companies are beginning to place a greater emphasis on customer retention programs. Experience shows that a customer holding two

policies with the same company is much more likely to renew than is a customer holding a single policy. Similarly, a customer holding three policies is less likely to switch than a customer holding less than three. By offering quantity discounts and selling bundled packages to customers, such as home and auto policies, a firm adds value and thereby increases customer loyalty, reducing the likelihood the customer will switch to a rival firm.

Customer segmentation is the other endeavour that CRM involves. It is the practice of classifying your customer base into distinct groups (Farn and Huang 2009). Customer segmentation is also described as the process of dividing customers into homogeneous groups on the basis of shared or common attributes (Bounsaythip and Rinta-Runsala 2001). The goal of segmentation is to know your customer better and to apply that knowledge to increase profitability, reduce operational cost, and enhance customer service. Segmentation can provide a multidimensional view of the customer for better treatment targeting.

2.5.1.3 Performing sophisticated campaign management

Developing a customer relationship has a long-standing tradition in business. Small firms and many retailers are able to relate to their customers individually. However, as organizations grow larger, marketing departments often begin to think in terms of product development instead of customer relationship development and maintenance. It is not unusual for the sales and marketing units to focus on how fast the firm can bring a mass-appeal product to market rather than how they might better serve the needs of the individual customer.

Ultimately, the difficulty is that as markets become saturated, the effectiveness of mass marketing slows or halts completely. Fortunately, advanced DM technology enables insurance companies to return their marketing focus to the aspects of individual customer loyalty. Creative, data-driven, scientific marketing strategies are now paving the way back to the customer relationship management of simpler, efficient economies, while on a much grander, comprehensive scale.

2.5.1.4 Reinsurance

DM can be used to structure reinsurance more effectively than the using traditional methods. DM technology is commonly used for segmentation clarity. In the case of reinsurance, a group of paid claims would be used to model the expected claims experience of another group of policies. With more granular segmentation, analysts can expect higher levels of confidence in the model's outcome. The selection of policies for reinsurance can be based upon the model of experienced risk and not just the generalization that it is a long tailed book of business.

2.5.1.5 Estimating outstanding loss reserve

The settlement of claims is often subject to delay. For example, in liability insurance, the severity (magnitude) of a claim may not be known until years after the claim is reported. In cases of an employer's liability, there is even a delay involved in reporting the claim. Such delays may result in non-normal distribution of claims; specifically, skewed distributions and long-tailed distribution across time and across business classes.

Still the everyday running of the firm must continue, and an estimate of the claim severity is often used until the actual value of the settled claim is available. The estimate can depend on the severity of the claim, likely amount of time before settlement, and effects of financial variables, such as inflation and interest rates.

2.5.1.6 Detecting fraudulent claims

Another important application of DM in the insurance industry is detecting fraudulent insurance claims. Obviously fraudulent claims are an ever-present problem for insurance firms, and techniques for identifying and mitigating fraud are critical for the long-term success of insurance firms. In searching for fraudulent claims, analysts look for unusual associations, anomalies, or outlying patterns in the data. Specific analytical techniques adopt at finding such subtleties are market basket analysis, cluster analysis, and predictive modeling. Quite often, successful fraud detection analysis such as those from DM project can provide a very high return on investment.

2.5.2 Insurance fraud detection

Fraud encompasses a wide range of illicit practices and illegal acts involving intentional act or omission designed to deceive others, resulting in the victim suffering a loss and/or the perpetrator achieving a gain.

2.5.2.1 Types of fraud

According to SAS Institute Inc. (2008), there are two distinctly different types of fraud: opportunistic fraud and professional fraud. An individual who simply has a chance to inflate a claim or get an exaggerated estimate for losses or repairs from his or her insurance company usually perpetrates opportunistic fraud. This person might know an insider but generally isn't operating with an insider's knowledge of the insurer's fraud detection systems or thresholds. Opportunistic fraud is commonplace, but the amount per incident is relatively low. Professional fraud is often perpetrated by organized groups with multiple, false identities, targeting multiple organizations or brands. These criminals know how fraud detection systems work, and they routinely test thresholds to stay just under the radar. These crime rings often place or groom insiders to help them defraud the company through several channels at once. The incidence of organized fraud is lower than ordinary insurance fraud, but the amount per incident is far greater.

According to Trnka (2010), Fraud occurs in the following areas: Credit card fraud, Internet transaction fraud / E-Cash fraud, Insurance fraud and health care fraud, Money laundering, Intrusion into computers or computer networks, Telecommunications fraud, Voice Over IP (VOIP) fraud, and Subscription fraud / Identity theft.

The insurance fraud problem translates to two DM problems (Woodfield 2005).

1. Given a set of records (claims) that do not have known target values, use unsupervised learning techniques to divide the data into two or more clusters, and employ domain expertise to evaluate each cluster as likely to be either FRAUD=YES or FRAUD=NO. This can be an iterative process that involves modifying unsupervised learning options and criteria until domain experts are satisfied with the clusters that are produced.

2. Given a set of records (claims) that have known or estimated target values, construct a predictive model to score new cases with respect to a propensity for being fraudulent.

2.5.2.2 Fraud detection technologies

Fraud Detection is concerned with the detection of fraud cases from logged data of system and user behavior (Trnka 2010). Rapid advances in technology enable insurance companies to use more powerful techniques to not only detect fraudulent activity, but to prevent it (SAS Institute Inc. 2008). According to SAS Institute Inc., Insurers supported their anti-fraud procedures by implementing new technologies, such as:

- Rules and red flags: Identify specific patterns and highlight activities that look suspicious.
- Database searching: to pool data with other database subscribers to broaden claims investigations
- Exception reporting: to report events that exceed a threshold for a particular claims benchmark
- Query and analysis: examining large volumes of adjudicated claims to find discrepancies.
- Predictive modeling: compares claims to baselines or thresholds to create fraud propensity scores.
- Social networking analysis: shows links between entities to uncover abnormal claims patterns.

2.5.2.3 Fraud detection approaches

Generally, fraud detection can be done using the Clustering approach, Expectations approach, and Predictive modelling approach (Koh and Gervais 2010). While the first two approaches highlight suspicious cases for further fraud investigation, the last approach directly predicts the probability of fraud occurrence in a given transaction.

The clustering approach focuses on “normal” patterns/clusters and searches for deviations from the “norm”. These deviations flag suspicious cases that may be further investigated

for fraud. They indicate outliers only and not necessarily fraud cases. On the other hand, the expectations approach focuses on what should be the (expected) value and compares it with what is the (actual) value. Large deviations are suspicious. This approach requires a predictive model that generates the expectations.

Finally, the predictive modelling approach constructs a predictive model that predicts the probability of fraud. Such a model attempts to differentiate fraud from non-fraud cases and hence requires data from both categories. This data requirement may be difficult to satisfy in some types of fraud (e.g., motor insurance fraud). In particular, the fraud data may not be sufficient because there may not be many cases of confirmed fraud, relative to non-fraud cases.

2.6 Related Works

In our country, there are works done to assess the application of DM in the different sectors like Airlines, Banking, Insurance, HealthCare, and Customs. Henock (2002) and Denekew (2003) for example, conducted a research on the application of DM for customer relationship management in the airlines industry as a case study on Ethiopian Airlines. Both Henock and Denekew used clustering and classification techniques with k-Means and decision tree algorithms. In addition, Kumneger (2006) has also tried to study the application of DM techniques to support customer relationship management for the Ethiopian Shipping Lines. Kumneger has applied clustering and classification techniques with k-Means and decision tree algorithms.

Shegaw (2002) also conducted a research on the application of DM in predicting child mortality in Ethiopia as a case study in the Butajira Rural Health Project. Shegaw employed the classification technique, neural network and decision tree algorithms to develop the model for predicting child mortality. Additional case studies were also conducted regarding the application of DM in the different sectors. For example, Tilahun (2009) has tried to assess the possible application of DM techniques to target potential VISA card users in direct marketing at Dashen Bank. Melkamu (2009) also conducted a research to assess the applicability of DM techniques to CRM as a case study on Ethiopian Telecommunications Corporation (ETC).

In addition, Leul (2003) tried to apply the DM techniques for crime prevention as a case study on the Oromia Police Commission. Leul used the classification technique, decision tree and neural network algorithms to develop the model, which will help to classify crime records. Helen (2003) also tried to study the application of DM technology to identify significant patterns in census or survey data as a case of the 2001 child labor survey in Ethiopia. She has applied the association rule DM technique and the Apriori algorithm for identifying relationships between attributes within the 2001 child labor survey database that she used to clearly understand the nature of child labor problem in Ethiopia. Apart from the association rule technique the expectation maximization-clustering algorithm were used to categorize the final selected datasets.

Many researchers applied different DM techniques and algorithms for detecting fraudulent claims in the insurance industry. Rejesus et al. (2004) conducted a research to study the application of DM technology in detecting crop insurance fraud for the US Department of Agriculture. Rejesus et al. applied the outlier analysis DM technique in helping to detect anomalous behaviour in an agricultural datasets. Guo (2003) also applied k-Means clustering, decision tree algorithm, and logistic regression for modelling insurance risk in the property/casualty insurance. In addition to this, Koh and Gervais (2010) applied DM techniques and algorithms for detecting motor insurance fraud. Koh and Gervais used the clustering and expectations approaches. Further, DM techniques such as outliers clustering and decision trees were used to generate the clustering results and expected/predicted values in this research. Woodfield (2005) also conducted a research for Predicting Workers' Compensation Insurance Fraud Using SAS Enterprise Miner and SAS Text Miner by applying the clustering technique.

To the knowledge of the researcher there are only two attempts in our country that have been done so far towards the application of DM in the insurance Industry. Mesfin (2005) had tried to study on the possible application of predictive DM techniques in the renewal process of personal accident policies as a case study on Ethiopian Insurance Corporation. The other research, which is undertaken by Tesfaye (2002), has tried to develop a predictive model using DM techniques in support of insurance risk assessment as a case study on Nyala insurance share company. He applied classification technique and neural

network algorithm for developing the model, which determines the risk exposure of motor insurance policies.

Both Mesfin and Tesfaye have tried to apply the different DM techniques in support of insurance risk assessment. But till now there is no any work that have been done so far regarding the application of DM in detecting insurance fraud in the insurance industry in Ethiopia. Hence this study has a great contribution in applying DM technology for the purpose of insurance fraud detection in the insurance industry.

CHAPTER THREE

DATA MINING METHODS FOR FRAUD DETECTION

The detection of insurance fraud can be done using computerized statistical analysis tools. Both supervised learning methods (where a dependent variable is available for training the model) and unsupervised learning methods (where no prior information of dependent variable is available for use) can be potentially employed to solve this problem. The DM techniques mostly used for fraud detection are clustering and classification (Koh and Gervais 2010).

In this research, unsupervised (K-Means clustering) and supervised (decision tree and naive bayes) DM tasks are experimented. Since the datasets contains unlabelled, the clustering algorithm is applied to segment insurance claims into different clusters depending on the fraudulence nature of the claims.

3.1 K-Means Clustering

Cluster analysis or clustering is the process of grouping the objects into subsets so that the objects in subset are similar in some sense (Nimmagadda et al. 2011). Clustering is a widely used technique in DM applications for discovering patterns in underlying data.

Han and Kamber (2006) define the clustering as the process of grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and dissimilar to the objects in another cluster. Clustering technique considers data tuples as objects. They partition the objects into groups or cluster, so that objects within a cluster are “similar” to one another and “dissimilar” to objects in other clusters. Similarity is commonly defined in terms of how “close” the objects are in space, based on distance function. The quality of the cluster may be represented by its diameter, the maximum distance between any two objects in the cluster. Centroid distance is an alternative measure of cluster quality and is

defined as the average distance of each cluster object from cluster centriod.

There are many different algorithms to do the clustering task of DM. These algorithms can be categorized as partitioning, hierarchal, Density-based, Grid-based, and model-based methods (Han and Kamber 2006; Nimmagadda et al. 2011). Despite the availability of these different methods, the most widely used one is K-Means algorithm, which is the partitioning method as indicated by Han and Kamber (2006). The present study uses the K-Means clustering algorithm to undertake the clustering task.

The reason why K-Means clustering algorithm is chosen for conducting the clustering process is described as follows:

- The K-Means algorithm is the best-known and easiest algorithm.
- The K-Means algorithm is relatively scalable and efficient in processing large datasets because the computational complexity of the algorithm is O(nkt), where n is the total number of objects, k is the number of clusters, and t is the number of iterations. Normally, k<<n and t<<n (Han and Kamber 2006).
- Han and Kamber (2006) further explain that the K-Means algorithm is applicable if it is possible to determine the mean of the cluster. Usually it works well if the dataset is numeric.
- Finally, K-Means algorithm is implemented in the WEKA data-mining tool, which has been applied in this research. Besides, the researcher found it easy to interpret the clustering results obtained from the K-Means algorithm.

3.1.1 K-Means algorithm

K-Means clustering algorithm is a prototype based clustering technique, which performs one level partition of the data objects (Nimmagadda et al. 2011). In this we first choose k initial centeroids, where k represents the number of clusters desired. Each point is then assigned to the closest centeroid, and each collection of points assigned to a centeroid is a cluster. The centeroids of each cluster is then updated based on the points assigned to the cluster. This assignment and update steps will continue until no point changes in cluster, equivalently, or centeroids remain the same.

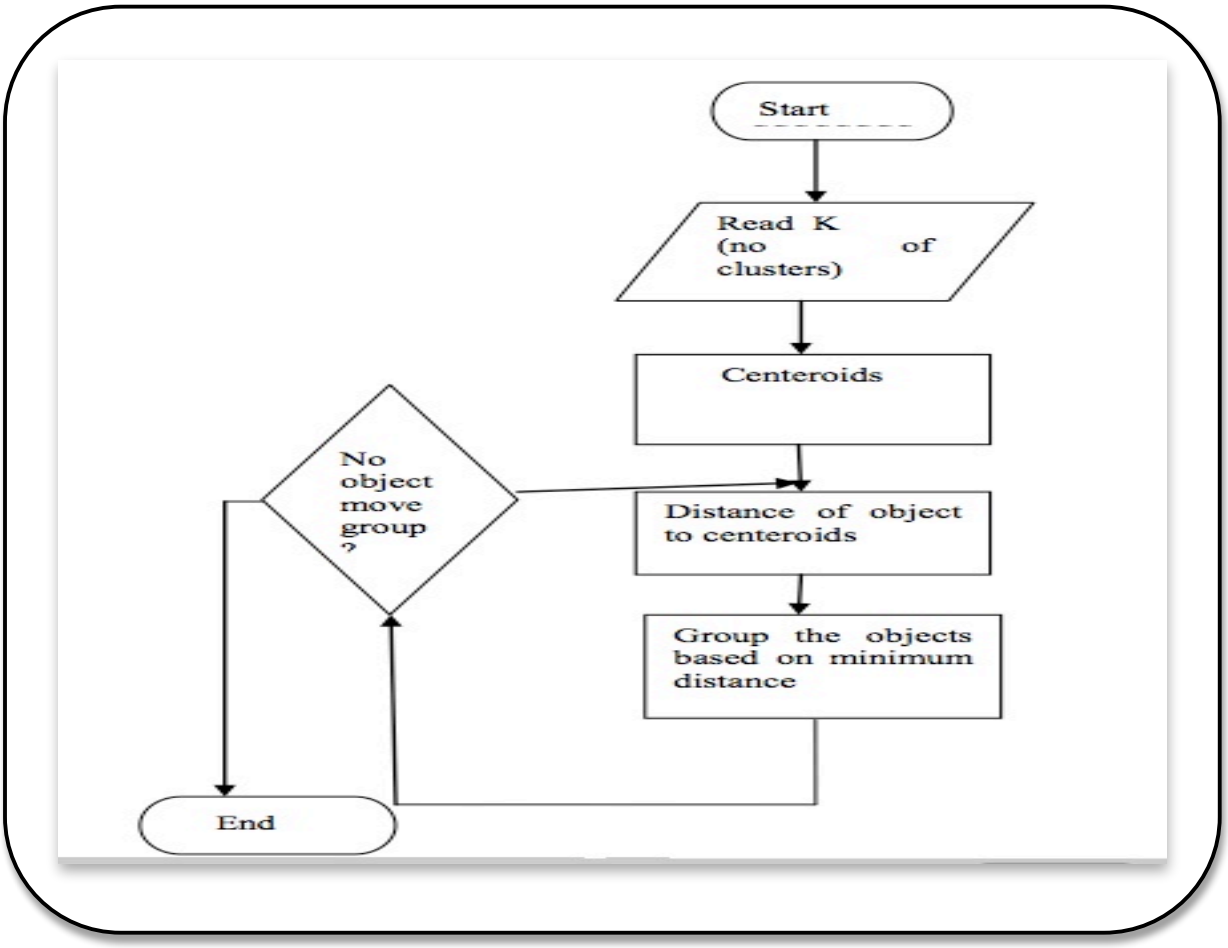

Figure 3. 1 *Flowchart Representing K-Means*

The K-Means algorithm is an iterative approach to finding K clusters based on distance (Berry and Linoff 2004). The algorithm divides a dataset into a predefined number of 'K' clusters. 'K' in the K-Means refers to the number of segments to partition the dataset, while 'means' refers to the average location of all of members (which are records from a database) of a particular cluster. The K-Means algorithm "self-organizes" to create clusters. The algorithm basically has three steps to do the clustering task (Berry and Linoff 2004). The first step, the algorithm randomly selects K data points to be the seeds. Each of the seeds is an embryonic cluster with only one element. The second step assigns each record to the closest seed. One way to do this is by finding the boundaries between two clusters. The boundaries between two clusters are the points that are equally close to each cluster. Finally, the third step is to calculate the centroid of the clusters; these now do a better job of characterizing the clusters than the initial seeds finding the centroid is

simply a matter of taking the average value of each dimension for all the records in the cluster. The centroids become the seeds for the next iteration of the algorithm (K-Means).

The 2nd Step (assigning records to the closest seed) is repeated, and each point is assigned to the cluster with the closest centroid. The process of assigning points to a cluster and then recalculating centroids continues until the cluster boundaries stop changing. In practice, K-Means algorithm usually finds a set of stable clusters after a few iterations.

According to Bounsaythip and Rinta-Runsala (2001), K-Means is based on a concept of distance, which requires a metric to determine distances. Euclidean distance can be used for continuous attributes, while for categorical variables; one has to find a suitable way to calculate the distance between attributes in the data.

The original choice of a value for K determines the number of clusters that is to be found. In addition, if this number does not match the natural structure of the data, the technique will not obtain good results. Unless the data-miner suspects the existence of a certain number of clusters, the experimenter will have to experiment with different values for K.

Berry and Linoff (2004) explain that, in general, the best set of clusters is the one that does the best job of keeping the distance between members of the same cluster small and the distance between members of adjacent clusters large. They further state the best set of clusters in the descriptive DM may be the one showing unexpected pattern in the data.

Some implementations of K-Means only allow numerical values for attributes (Nimmagadda et al. 2011). In that case, it may be necessary to convert the dataset into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales (e.g., "age" and "income"). While Weka provides filters to accomplish all of these preprocessing tasks, they are not necessary for clustering in Weka**.** This is because Weka Simple K-Means algorithm automatically handles a mixture of categorical and numerical attributes. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations. The Weka Simple K-Means algorithm uses Euclidean distance measure to compute distances between instances and clusters.

Once the clusters have been created using clustering algorithms, they need to be interpreted. Though there are several approaches to perform this, one of the approaches widely used for understanding clusters is building a decision tree with the cluster label as the target variable and using it to drive rules explaining how to assign new records to the correct cluster (Berry and Linoff 2004).

3.2 Decision Tree Classification Technique

According to Two Crows Corporation (1999), decision trees are a way of representing a series of rules that lead to a class or value. There are two types of decision trees, namely classification and regression trees. Classification trees label records and assign them to the appropriate class. Regression trees estimate the value of a target variable that takes on numeric values. Trees can grow in any forms. They could be binary trees of non-uniform depth, that is, each node has two children and the distance of a leaf to the root varies. In Figure 3.2**,** each node represents a 'yes' or 'no' question, the answer determines through which of the two paths a record proceeds to the next level of the tree.

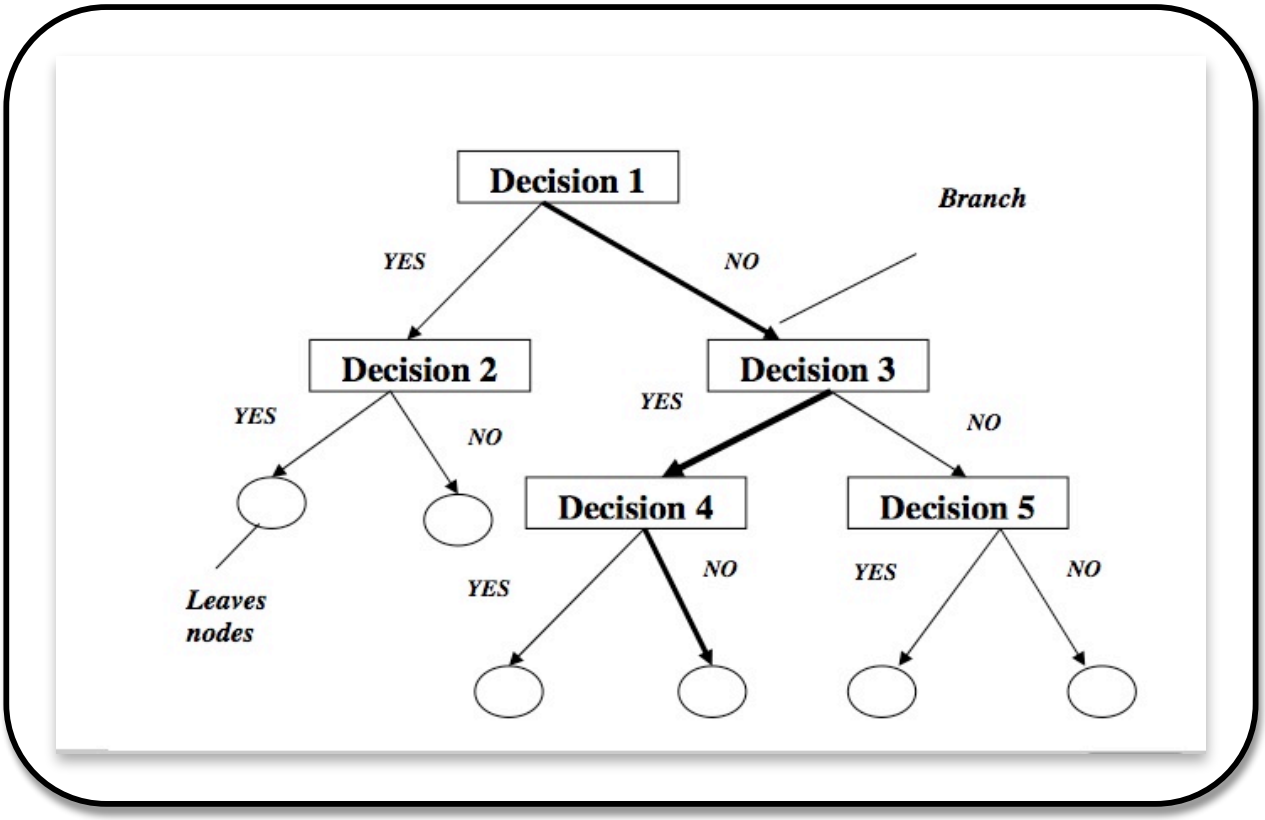

Figure 3. 2 *A Scenario that shows how decision tree is constructed*

Tree induction

The training process that creates the decision tree is called induction and requires a small number of passes through the training set. Most decision tree algorithms go through two phases: a tree-growing (splitting) phase followed by a pruning phase (Bounsaythip and Rinta-Runsala 2001).

Splitting: The tree-growing phase is an iterative process, which involves splitting the data into progressively smaller subsets. The first iteration considers the root node that contains all the data.

Subsequent iterations work on derivative nodes that will contain subset of the data. One important characteristics of splitting is that it is greedy, which means that the algorithm does not look forward in the tree to see if another decision would produce a better overall result.

Stopping criteria: Tree-building algorithms usually have several stopping rules. These rules are usually based on several factors including maximum tree depth, minimum number of elements in a node considered for splitting, or it's near equivalent, the minimum number of elements that must be in a new node. In most implementations the user can alter the parameters associated with these rules. Some algorithms, in fact, begin by building tree to their maximum depth. While such a tree can precisely predict all the instances in the training set (except conflicting records), the problem with such a tree is that, more than likely, it over fits the data.

Pruning: After a tree is grown, one can explore the model to find out nodes or sub trees that are undesirable because of over fitting or rules that are judged inappropriate. Pruning removes splits and the sub trees created by them. Pruning is a common technique used to make a tree more general. Algorithms that build trees to maximum depth will automatically invoke pruning. However, users have also the ability to prune the tree interactively.

Different decision tree algorithms

Decision tree algorithms that are commonly implemented include Chi-squared Automatic Detection (CHAID), Classification and Regression Trees (CART), ID3, C4.5 and C5.0. All are well suited for classification; some are also adaptable for regression. There are different distinguishing features between tree algorithms. These include (Bounsaythip and Rinta-Runsala 2001; Meera and Srivatsa 2010):

Target variables: Most tree algorithms require the target (dependent) variable be categorical. Such algorithms require that continuous variables be binned (grouped) for use with regression.

Splits: Many algorithms support only the binary splits, that is, each parent node can split into at most two child nodes. Others generate more than two splits and produce a branch for each value of categorical variables.

Split measures: help select which variables to use to split at a particular node. Common split measures include criteria based on gain, gain ratio, GINI, and chi-squared.

Rule generation: algorithms such as C4.5 and C5.0 include methods to generalize rules associated with a tree; this removes redundancies. Other algorithms simply accumulate all the tests between the root node and the leaf node to produce the rules.

The decision classification algorithm was selected due to the following reasons:

- Decision trees are easy to understand
- Decision trees are easily converted to a set of production rules
- Decision trees can classify both categorical and numerical data, but the output attribute must be categorical
- There are no a priori assumptions about the nature of the data.

As stated in Chapter 2, different decision tree algorithms were discussed and C4.5 is among one of the algorithms that includes methods to generalize rules associated with a tree.

3.2.1 The J48 decision tree algorithm

J48 algorithm is the Weka's implementation of the C4.5 (Witten and Frank 2000). In fact, J48 implements a later and slightly improved version of which is known as C4.5 Revision 8, which was the last public version of this family version of algorithms before C5.0 (new version of C4.5), a commercial implementation was released (Witten and Frank 2000). According to Bharti et al. (2010), J48 decision tree algorithm is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. It can be applied on discrete data, continuous or categorical data.

J48 is the decision tree algorithm that is used in this study to classify the segmented insurance claims as fraud or non-fraud suspicious claims. The J48 decision tree can serve as a model for classification as it generates simpler rules and remove irrelevant attributes at a stage prior to tree induction. In several cases, it was seen that j48 decision trees had a higher accuracy than other algorithms (Witten and Frank 2000). J48 offer also a fast and powerful way to express structures in data.

The J48 algorithm gives several options related to tree pruning. Pruning produces fewer, more easily interpreted results. More importantly, pruning can be used as a tool to correct for potential over fitting. J48 recursively classifies until each leaf is pure, meaning that the data has been categorized as close to perfectly as possible. When tested on new data, the rules may be less effective. Pruning always reduces the accuracy of a model on training data. This is because pruning employs various means to relax the specificity of the decision tree, hopefully improving its performance on test data.

J48 employs two pruning methods (Witten and Frank 2000). The first is known as sub-tree replacement. This means that nodes in a decision tree may be replaced with a leaf basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The second type of pruning used in J48 is termed sub-tree raising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Sub-tree raising often has a negligible effect on decision tree models. There is often no clear way

to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that sub-tree raising can be somewhat computationally complex.

Error rates are used to make actual decisions about which parts of the tree to replace or raise. There are multiple ways to do this. The simplest is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential over-fitting. Other error rate methods statistically analyze the training data and estimate the amount of error inherent in it.

There are several other options that determine the specificity of the model. The minimum number of instances per leaf is one powerful option. This allows you to dictate the lowest number of instances that can constitute a leaf. The higher the number of instances the more general the tree is. Lowering the number will produce more specific trees, as the leaves become more granular.

The binary split option is used with numerical data. If turned on, this option will take any numeric attribute and split it into two ranges using an inequality. This greatly limits the number of possible decision points. Rather than allowing for multiple splits based on numeric ranges, this option effectively treats the data as a nominal value. Turning this encourages more generalized trees. There is also an option available for using Laplace smoothing for predicted probabilities. Laplace smoothing is used to prevent probabilities from ever being calculated as zero. This is mainly to avoid possible complications that can arise from zero probabilities. Generally, the process of the J48 algorithm to build a decision tree can be expressed as follows:

1. Choose an attribute that best differentiates the output attribute values.
2. Create a separate tree branch for each value of the chosen attribute.
3. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.
4. For each subgroup, terminate the attribute selection process if:

a. All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.
b. The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances.

5. For each subgroup created in (3) that has not been labeled as terminal, repeat the above process.

The algorithm is applied to the training data. The created decision tree is tested on a test dataset, if available. If test data is not available, J48 performs a cross-validation using the training data itself.

3.3 Naive Bayes Classification Technique

Naïve Bayesian Classifier is a machine-learning algorithm that maps (classifies) a data example into one of several predefined classes. It attempts to maximize the posterior probability in determining the class. Observations show that Naïve Bayes performs consistently before and after reduction of number of attributes (Anbarasi et al. 2010). They are effective in handling Auto insurance fraud (Bhowmik 2011).

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes (Han and Kamber 2006). This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naïve." Naïve Bayes analyses the relationship between each input attribute and the dependent attribute to derive a conditional probability for each relationship (Ibrahim 1999).

The figure below shows the distribution of an input associated with each class, for example, given the variable **X** with a value at $\mathbf{x_i}$ the probability of it being in Class **A** is greater than the probability of it being in Class **B**. In mathematical terms, if one knows how $\mathbf{P(X \mid C)}$ and the densities $\mathbf{P(x_i)}$ and $\mathbf{P(c_j)}$ (prior probabilities) are known, then the class $\mathbf{c_j}$ assigned to datum $\mathbf{x_i}$ if **C** has the highest posterior probability given the data.

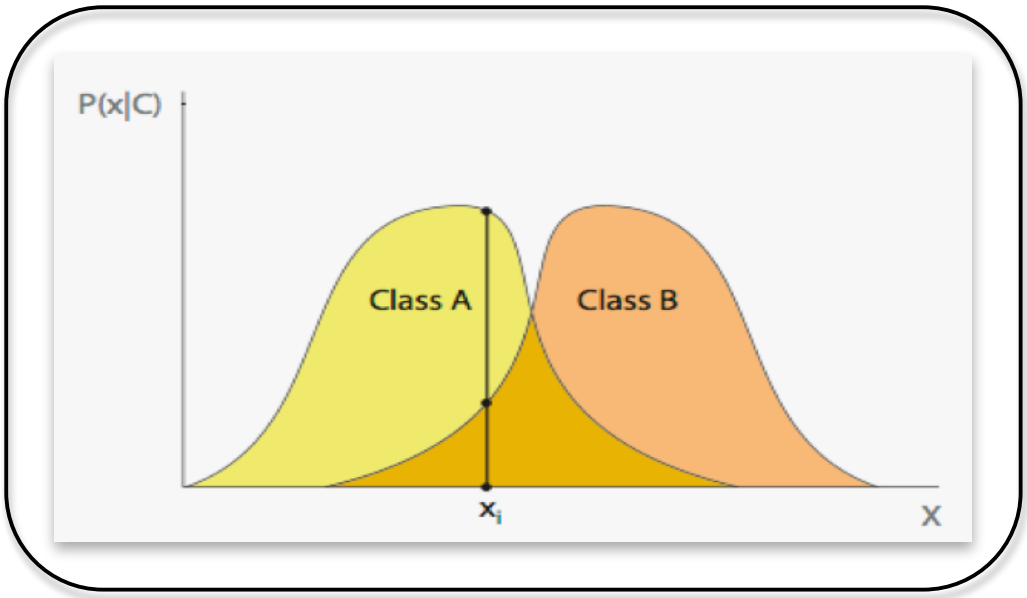

Figure 3. 3 *Naïve Bayes distribution of an input associated with each class*

Naïve Bayes works very well when tested on many real world datasets (Witten and Frank 2000). By theory, this classifier has minimum error rate but it may not be the case always (Anbarasi et al. 2010). However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. An advantage of Naïve Bayes algorithm over some other algorithms is that it requires only one pass through the training set to generate a classification model. In addition, Naïve Bayes can also obtain results that are much better than other sophisticated algorithms. However, if a particular attribute value does not occur in the training set in conjunction with every class value, then Naïve Bayes may not perform very well. It can also perform poorly on some datasets because attributes were treated as though they are independent, whereas in reality they are correlated.

3.3.1 Naive Bayes algorithm

The Naive Bayes algorithm can be used for both binary and multiclass classification problems. It builds a model to classify new examples based on observed probabilities and supporting evidence from the training data. The dataset for this type of problem is one with input attributes and a known outcome, or class.

Naïve Bayes makes predictions using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence. Bayes' Theorem states that the probability of event A occurring given that event B has occurred (P(A|B)) is proportional to the probability of event B occurring given that event A has occurred multiplied by the probability of event A occurring (P(B|A)P(A)).

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \text{----------3.1}$$

The naïve Bayesian classifier, or simple Bayesian classifier, works as follows (Han and Kamber 2006):

1. Let **D** be a training set of tuples and their associated class labels. Each tuple is represented by an n-dimensional attribute vector, $X = (x_1, x_2, \ldots x_n)$, depicting **n** measurements made on the tuple from **n** attributes, respectively, $A_1, A_2, \ldots A_n$.
2. Suppose that there are **m** classes, $C_1, C_2, \ldots C_m$. Given a tuple, **X**, the classifier will predict that **X** belongs to the class having the highest posterior probability, conditioned on **X**. That is, the naïve Bayesian classifier predicts that tuple **X** belongs to the class $\mathbf{C_i}$ if and only if

   $$P(C_i \mid X) > P(C_j \mid X) \text{ for } 1 \le j \le m, j \ne i.$$

   Thus we maximize $P(C_i \mid X)$. The class C_i for which $P(C_i \mid X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem (Equation 3.1),

   $$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}$$

3. As P(X) is constant for all classes, only $P(X \mid C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \cdots = P(C_m)$, and we would therefore maximize $P(X \mid C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = |C_i,D|/|D|$, where $|C_i,D|$ is the number of training tuples of class C_i in D.
4. Given datasets with many attributes, it would be extremely computationally expensive to compute $P(X \mid C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the naive assumption of class conditional independence is made. This

presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(X \mid Ci) = \prod_{k=1}^{n} P(Xk \mid Ci) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \cdots \times P(x_n \mid C_i)$$

We can easily estimate the probabilities $P(x_1 \mid C_i)$, $P(x_2 \mid C_i)$, . . . , $P(x_n \mid C_i)$ from the training tuples. Recall that here x_k refers to the value of attribute A_k for tuple X. For each attribute, we look at whether the attribute is categorical or continuous.

5. In order to predict the class label of X, $P(X \mid C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of tuple X is the class C_i if & only if

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \text{ for } 1 \leq j \leq m, j \neq i.$$

In other words, the predicted class label is the class C_i for which $P(X \mid C_i)P(C_i)$ is the maximum.

Generally, the techniques and the algorithms that are discussed before are used to conduct the experimentations of this study for developing the model used for predicting fraudulent insurance claims.

CHAPTER FOUR

BUSINESS AND DATA UNDERSTANDING

4.1 Introduction to AIC

AIC is a privately owned company, which was established in 1994 in accordance with the Licensing and Supervision of Insurance Business Proclamation No.86/1994, and the Commercial Code of Ethiopia. The company was established with Birr 30 million authorized and Birr 30 million paid up capital, which is raised to Birr 60 million in 2010 (Muluken 2010) with the purpose of honourably serve the needs of the society by producing, delivering and maintaining full range of superior quality insurance services at a competitive price. AIC has 15 branches all over the country, eight branches in Addis Ababa, and seven branches in the regions. Ever since its establishment starting from very low levels the company has been registering remarkable sustained increase and improvement in sales and production of life and non life institutional, financial, human resource, materials and technological capacities. The management's friendly and responsiveness to the researcher's enquiries to get access to and collect the required data for this study encouraged the researcher to conduct this research at AIC.

To achieve its vision, mission and objectives, AIC is structured into eight departments at the head office (AIC 2003). These are Marketing and Business Development, Finance, Underwriting, Claims, Reinsurance, Human Resource Development, Information system services, and Legal services departments. The management of the company constitutes Marketing, Finance, Underwriting, Claims, Reinsurance and Human Resource Development Managers that forms the management team of AIC, which is technically led by the Managing Director/CEO. From the eight departments of AIC the claims department is responsible for handling insured's claims.

Africa Insurance Company engages itself in all classes of general insurance. In August 1998 the Company expanded its services to include Life Assurance in its coverage. Today, AIC provides insurance covering Life, Property and Liability Risks. Currently, the company carries out transactions of all types of insurance policies.

Type of Insurance	Description
Motor	Insurance purchased for cars, trucks, and other vehicles
Fire	Insurance against loss due to fire
Marine	Insurance covers for the loss or damage of ships
Hull and Cargo	In Ocean Marine and Aviation insurance, insurance against physical damage to plane or ship & Insurance covers for the loss or damage of cargo, terminals, and any transport or property by which cargo is transferred, acquired, or held between the points of origin and final destination
Workmen's Compensation	Provides wage replacement and medical benefits for employees who are injured in the course of employment, in exchange for mandatory relinquishment of the employee's right to sue his /her employer for the tort of negligence.
Personal & Group Accident	Provides protection against death or dismemberment arising solely and directly from an accident
Money	Provides cover for money when carried by the insured or employees
Fidelity Guarantee	Covers crime and theft by a firms employees
Burglary and House Breaking	Cover against loss or damage by burglary and/or by housebreaking
Plate Glass	Cover for accidental loss or damage to the plate glass
Liability	Coverage against harm that an insured have caused to a third party
Domestic Package	Provides financial protection against misfortunes befalling your house
Engineering	Coverage against losses associated with an engineering firm
Computer All risks and others	Covers the losses from loss of computer equipments, system and data and the other related losses

Table 4. 1 *Types of Insurance Services AIC Provides*

4.2 AIC Motor Policy

There are different types of insurance policies that AIC is currently providing. For example, motor, life, workmen's compensation, marine, money, personal and group accident, etc are some of the policies that AIC is providing. Motor Policy is a contract concluded between the insurer and the insured to cover the insured's legal liability and/or property up to an agreed limit as stipulated on the contract on payment of a sum of money called a premium to the insurer for an agreed period of time (AIC Underwriting Department 2004).

4.2.1 Classification of motor policies

Motor contractors or policies are classified in line to the use and type of the vehicles into private and commercial vehicles. Motor vehicles are classified as "private" if they are used only for social, domestic, pleasure and professional purposes or business calls of the insured. The term "Private Purpose" excludes use for hiring, racing, pace making, speed testing and use for any purpose in connection with the motor trade. Vehicles which transport goods or which carry passengers for business purpose are classified as commercial vehicles. Commercial vehicles can be sub divided as: Commercial Vehicles Own Goods Use, Commercial Vehicles For General Cartage, Buses, Agricultural Vehicles, Motorcycle, and Motor Trade.

4.2.2 Classification of cover of motor policies

According to AIC Underwriting Department (2004), policies are classified according to the indemnity, which they provide and the indemnities provided by the company as follows:

Comprehensive cover: a comprehensive cover is the widest form of cover available, although it cannot protect against every conceivable risk. It provides protection against a wide range of contingencies including not only an indemnity in respect of the insured's legal liability for death of or bodily injury to or damage caused to the property of third parties arising out of the insured vehicle but also will indemnify the insured in respect of all damage/loss to the vehicle and its accessories caused by any accidental, external and

physical means as a result of Collision, Overturning, Fire, Self-Ignition, Lightning, Explosion, Burglary, and House Breaking, Theft and Malicious acts. The policy exclusion among other things are wear and tear and/or depreciation of any motor vehicle or any parts of such motors, mechanical fracture and/or mechanical or electrical break down or failure of any part of any motor vehicle. Generally, the main comprehensive covers can be summarized as:

- Insured's legal liability for death or bodily injury.
- Insured's legal liability for damage caused to the property of third party, arising out of the use of the insured's vehicles.
- Indemnity for damage to the insured's vehicle.

Third party cover: a third party cover indemnifies the insured against liability at (Ethiopian) law for damages in respect of death of or bodily injury to third parties and damage to the property of such persons in the event of an accident caused by or through or in connection with any motor car described in the schedule of the policy.

Third party fire and theft: third party can be extended to cover insured's vehicle against the risk of fire and theft in addition to the third party.

4.3 Business Understanding

AIC investigate the different insurance claims, through its experienced investigators. Although there is no any written criterion by the company, investigators use different variables to check whether the claims that are reported to the company are deceitful or not informally. Some of these are the length of accident report date, traffic police report, garage report, age of the vehicle, and the insured years of stay with the insurance company.

The company informally rates those insurance claims that are reported lately after the occurrence of the accident with an old age of vehicle, a short stay of the insured, no traffic police report, as fraud suspicious claims. In addition to this, those claims that are reported within a short time range with medium age of vehicle, and a medium stay of the

insured are also rated as fraud suspicious claims.

Insurance claims that are reported quickly with new vehicle, traffic police report, and a long stay of the insured are rated as suspicious of non-fraudulent claims. Apart from this, those claims with a quick accident report date, traffic police report, new vehicle, and with a short stay of the insured are also rated as suspicious of non-fraudulent claims.

In addition to the aforementioned criteria's, which are used by experts for judging whether an insurance claim is fraud suspicious or not, the type of accident occurred can also be employed for the purpose of investigation of claims whether they are suspicious of fraud or not. Those insurance claims with a collision accident especially with a third party are considered as suspicious of fraud. In addition to this commercial use of vehicles is also considered for showing fraud suspicious claims mostly, while claims with private use of vehicles are mostly believed to be free of freak.

4.3.1 Claims handling processes

The claim department of AIC undertake processes, such as, notification of claim, completion of claim form, recording, reporting, survey/investigation, and claims payment approval. Immediate notification of accident enables insurers to investigate circumstances while they are fresh and before witness memory about the incidence is faded. Notification of claims could be verbal or in writing.

In order to substantiate a claim, completion of a claim notification form is required. The form will elicit information pertaining the insured, the place of loss, the nature of loss, details of damage on the vehicle, details of damage on the third party's property, etc. Underwriting and accounts verification helps to get the information about Name of insured, Period of insurance, Type of coverage, vehicles Plate number, Chassis Number, Engine Number, and Year of make, Excess, Sum insured, Outstanding premium, Bank or other interest (if any), and Previous claims records if there exist any.

All the information gathered during claim processing is recorded in the register book. Accordingly, the department open claim file for each cases. All claims are then reported to the claim clerk supervisor who is in turn shall advise the department head Manager and

if reinsurance exists to reinsurance.

Claim related to motor insurance needs further survey and investigation. This helps the insurer to compare garage repair estimates with the estimates given by the surveyor and reach into a compromise or in cases of total loss claims, it helps the insurer to determine the market value of the object (vehicles) immediately before the loss and the salvage value of the object.

There is always a line of authority on the decision made on claims approval. The line of authority in AIC takes the pattern that payment up to Birr 100,000.00 is undertaken by the Claim's Department Manager while payment above Birr 100,000.00 is undertaken by the Managing Director (CEO).

4.3.2 Current practice of the company

Currently AIC is trying to entertain huge amounts of claims, which are arriving everyday to the claims department. The department undertakes the different processes that were discussed before, in order to solve the problems associated with the reported claims. AIC primarily trusts its customers. As explained before, when the customers come up with a claim the company accepts it and based on this information the company conducts a survey.

Claim related to motor insurance needs further survey and investigation. Insurance investigations are usually conducted to investigate matters pertaining to insurance claims that are suspicious or otherwise in doubt for some reason. Some insurance companies have their own in-house investigation teams while other companies sub-contract the work to private investigators or private investigation firms.

The survey and investigation of claims helps AIC to compare garage repair estimates with the estimates given by the surveyor(s) and reach into a compromise. If the surveyor's report is similar with the garage repair value and a traffic police report is submitted, the company then offers a payment for the claimant. But if the two reports contradicted with each other and no traffic police report, the company may refuse to pay for the claimant. This is the only way that the company is currently following to detect suspicious fraudulent claims.

4.4 Understanding the Data

A precondition to any DM is the data itself. A good source of data for DM purpose is identified to be the corporate data warehouse (Berry and Linoff 2000). The reason for this is that the data is stored in common format with consistent definitions for fields and their values. To fulfill this requirement, raw data was initially collected regarding motor insurance claims and related issues from the PREMIA database of AIC and careful analysis of the data and its structure is done together with the business (domain) experts by evaluating the relationship of the data with the problem at hand and the particular DM tasks to be performed. The following sections describe the nature and structure of the collected data.

4.4.1 Initial data collection

For the purpose of data collection, five service units were chosen from the Addis Ababa branches of AIC. The service units were chosen by the suggestion of experts, because of their activeness within the company. All these service units are connected with each other at the head office of the company, which is found at Bole sub city Alem Building. So, all the necessary data about each service unit is collected from the PREMIA database of the main office.

The PREMIA database was thoroughly studied and, as a result of the study, six basic relations (tables) were found to be important. All the six basic relations, PT_CLAIM, PT_VEHICE, PT_DRIVER,PT_CLAIM_DTLS,PT_POLICY, PT_CLAIM_ESTIMATE and PT_CLAIM_PAID, were extracted in Excel format directly from the target database.

Next, the attributes within each of the claims, vehicle, policy, and the other tables that are listed above are combined into a single database. Before doing this, those important attributes from each table were chosen first.

The number of records collected from the five service units is summarized in Table 4.2.

Service Unit	Number of records collected
Arada	2413
Filwuha	7846
Head Office	2592
Kirkos	2585
Tekele Hayimanote	2374
Total	17810

Table 4. 2 *Distribution of collected data with respect to sample service units*

As can be seen from Table 4.2 the distribution of the collected data from the five branches of the company is almost balanced except the Filwuha branch, which contains higher number of claims relative to the other branches.

4.4.2 Description of the data collected

As indicated before, the pertinent data to carry out the research is collected from six different tables of the PREMIA database. These are claim, policy, detail claim, vehicle, claim estimate and claim paid tables. Though each of the tables has had lots of attributes in the original dataset, the tables bellow show initially selected attributes of each table.

No.	Attribute Name	Data Type	Description
1.	CLM_POL_NO	Varchar2	The policy number of the claim
2.	CLM_LOSS_DT	Date	The date that the accident has occurred
3.	CLM_INTM_DT	Date	The date the insured reported to the company about the accident
4.	CLM_LOSS_TYPE	Varchar2	The type of loss that has occurred
5.	CLM_LOSS_DED_YN	Varchar2	Existence of deductibles
6.	CLM_DIVN_CODE	Number	Branch service unit.
7.	CLM_SC_CODE	Varchar2	Subclass code
8.	CLM_CLOSE_DT	Date	The date that the claim has been closed
9.	CLM_CLAIM_DESC	Varchar2	Description of the claim about the accident
10.	CLM_CLOSE_REASON	Varchar2	The reason that the claim has been closed

Table 4. 3 *Description of the PT_CLAIM Table*

- Originally, the PT_CLAIM table has more than sixty attributes. But most of them are unrelated with the problem at hand and some are empty. Because of this the above initial attributes are selected together with the domain expert.

No.	Attribute Name	Data Type	Description
1.	CE_LC_AMT	Number	The amount of money that is estimated to be paid for the claim in local currency
2.	CE_SRC_BUS_CODE	Varchar2	Source of business
3.	CE_EST_NO	Varchar2	Claim estimate number
4.	CE_LOSS_TYPE	Varchar2	The type of loss

Table 4. 4 *Description of the PT_CLAIM_ESTIMATE Table*

- Originally, the PT_CLAIM_ESTIMATE table has more than 40 attributes. But most of the attributes are similar with the PT_CLAIM table.

No.	Attribute Name	Data Type	Description
1.	VEH_YEAR	Number	Vehicle year of manufacture
2.	VEH_TYPE	Varchar2	The type of vehicle
3.	VEH_AGE	Number	The original age of the vehicle

Table 4. 5 *Description of the PT_VEHICE Table*

- The PT_VEHICE table originally has 79 attributes. But for conducting this research only three of them are selected discussing with the domain expert.

No.	Attribute Name	Data Type	Description
1.	POL_UW_YEAR	Number	Policy underwriting year
2.	POL_ISSUE_DT	Date	The date that the policy is issued
3.	POL_END_EFF_FM_DT	Date	Policy effective end date
4.	POL_STATUS	Varchar2	Status of the policy

Table 4. 6 *Description of the PT_POLICY Table*

- This table originally has 171 attributes. But discussing with the domain expert only three of them, which are important for the problem domain, are selected for conducting this research.

No.	Attribute Name	Data Type	Description
1.	CD_RISK_TYPE	Varchar2	The type of vehicle use. It is to mean that whether the vehicle is used for private or commercial purpose
2.	CD_LOCATION_CODE	Number	The location of the accident

Table 4. 7 *Description of the PT_CLAIM_DTLS Table*

- This table originally has 47 attributes. But discussing with the domain expert only two of them, which are important for the problem domain, are selected for conducting this research.

The CP_LC_PAID_AMT attribute, whose data type is Number, is only selected from the PT_CLAIM_PAID table. The CP_LC_PAID_AMT is the amount of money paid for the claimant in local currency.

4.4.3 Data quality assurance

The collected data contains missing, incomplete and irrelevant data. Some important information regarding the driver's age, marital status, health condition, date of driving license, the place and time of the accident occurred, traffic police report and some others are missing. Generally, most of the attributes of the tables in the PREMIA database have no that much use and it was too difficult to understand the database because of its too complexity. Only some of the attributes are applicable for the problem at hand. Apart from that some of the attributes are duplicated in some of the tables.

4.5 Preparation of the Data

While DM is a key stage in the knowledge discovery process, the data preprocessing process often require considerable effort. The purpose of the preprocessing stage is to cleanse the data as much as possible and to put it into a form that is suitable for use in later stages. Starting from the data extracted from the source database maintained by AIC, a number of transformations are performed before a working dataset was built.

4.5.1 Data selection

The whole target dataset may not be taken for the DM task. Irrelevant or unnecessary data are eliminated from the DM database before starting the actual DM function. Originally there were around 25000 records. From this around 2500 of the records are randomly selected for testing purpose. The rest of the dataset is used for training purpose. Since this dataset contains irrelevant and unnecessary data, all are not used for training. So, after eliminating irrelevant and unnecessary data only a total of 17810 datasets are used for the purpose of conducting this study.

The above-described tables of the PREMIA database consist of more than 30 attributes each. The first task was to remove from the database those fields or attributes, which were irrelevant to the task at hand. As shown in Appendix 1, the following are the initial sets of attributes, which are further preprocessed to select the final attributes used in the study. CLM_INTM_DT, CLM_LOSS_DT, CLM_POL_NO, CLM_LOSS_TYPE, CLM_LOSS_DED_YN, CLM_DIVN_CODE, CLM_SC_CODE, CLM_CLOSE_DT, CLM_CLAIM_DESC, CLM_CLOSE_REASON,CE_LC_AMT, CE_SRC_BUS_CODE, CE_EST_NO, CE_LOSS_TYPE, VEH_YEAR, VEH_TYPE, VEH_AGE, POL_UW_YEAR, POL_ISSUE_DT, POL_END_EFF_FM_DT, POL_STATUS, CD_RISK_TYPE, CD_LOCATION_CODE and CP_LC_PAID_AMT were the initial set of attributes that are selected apart from the derived attributes.

4.5.2 Data cleaning

The data was cleaned, by removing the records that had incomplete (invalid) data and/or missing values under each column. Removing of such records was done as the records with this nature are few and their removal does not affect the entire dataset. The researcher makes use of the MS-EXCEL application for cleaning the data.

Accordingly, the CLM_LOSS_DED_YN, CLM_LOSS_TYPE, CE_EST_NO and CE_LOSS_TYPE attributes have only one value. The researcher then deleted the whole column since it is meaningless to use those attributes, which are having similar values throughout the records. Although it was a very important variable, the CLM_CLOSE_REASON attribute is also deleted because almost all of its values are

missed. Some of the VEH_AGE values were missing and the researcher used the mean value to fill in the missing. Apart from this information regarding the driver like driver's age, marital status, health condition, date of driving license and some other attributes are missing. Generally, the value for all these attributes is empty. Though very important, the researcher deleted all these attributes.

4.5.3 Data construction

The other important step in preprocessing is deriving other fields from the existing ones. Adding fields that represent the relationships in the data are likely to be important in increasing the chance of the knowledge discovery process yield useful result (Berry and Linoff 2000). In consultation with the domain experts at AIC, the following fields that are considered essential in determining the fraudulence of claims were derived from the existing fields. CLM_REP_DT_G (this refers the length of accident report date since its occurrence) is derived from the CLM_LOSS_DT and CLM_INTM_DT columns of the PT_CLAIM table. INSU_YR_S (refers for how many years the insured stay together with the insurance company) is derived from the POL_END_EFF_YEAR and POL_ISSUE_DT columns of the PT_POLICY table. CLM_REC_AMT (refers to the amount of money that the company recovered from previous claims of the claimant) is also derived from the CE_LC_AMT and CP_LC_PAID_AMT columns of the PT_CLAIM_ESTIMATE and PT_CLAIM_PAID tables.

CLM_REP_DT_G = **CLM_INTM_DT - CLM_LOSS_DT**
INSU_YR_S = **POL_ISSUE_DT - POL_END_EFF_YEAR**
CLM_REC_AMT = **CE_LC_AMT - CP_LC_PAID_AMT**

Those attributes, which were used for deriving the above attributes, are not included in the final list of attributes used in the study.

4.5.4 Data integration

The data integration process was done before deriving the attributes. As described before the dataset for the tables which was discussed above were available in different excel files. Data integration method for retrieving important fields from different files and tables was done in the effort to prepare the data ready for the DM techniques to be

undertaken in this research. The Oracle and ASYCUDA databases were used to carry out the data integration process. These data integration process took a lot of time of the research. This was because of the reason that when the different excel files were combined together the size of the dataset increased by more than 10 times. That means the records were duplicating tremendously. Till understanding and solving the problem lots of time were lost. Finally the data is integrated and put together into a single excel file.

4.5.5 Data formatting

Like any other software, Weka needs data to be prepared in some formats and file types. The datasets provided to this software were prepared in a format that is acceptable for Weka software. Weka accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute-Relation File Format) file format (a file name with an extension of ARFF i.e. FileName.arff).

At first the integrated dataset was in an excel file format. To feed the final dataset into the Weka DM software the file is changed into other file format. The excel file was first changed into a comma delimited (CSV) file format. After changing the dataset into a CSV format the next step was opening the file with the Weka DM software. Then this file was saved with ARFF (Attribute Relation File Format) file extension. Now the dataset, which is in ARFF file format, is ready to be used.

4.5.6 Attribute selection

The importance of reducing the number of attributes from hundreds to within a few dozen, not only speed up the learning process, but also prevents most of the learning algorithms from getting fooled into generating an inferior model by the presence of many irrelevant or redundant attributes (Berry and Linoff 2000). This is mainly because most practical learning algorithms are necessarily heuristic in nature and they often are misled by the presence of many non-essential attributes. Very limited numbers of attributes that are most important for the study at hand are selected because of this reason.

The dataset containing the attributes indicated in Appendix 1 is fed into the Weka DM

software. In order to select the best attributes from this initial collected dataset, the researcher evaluates the information content of the attributes using the select attribute technique of Weka with GainRationAttributeEval attribute evaluator and Ranker search method. So the tool arranges the attributes in order of their gain ratio. Hence, the researcher together with the domain expert removes those attributes, which have less gain ratio. The table below shows the final list of attributes that have been used in this study.

No.	Attribute Name	Data Type	Description	Remark
1.	CLM_REP_DT_G	Number	The length of accident report date since the occurrence of accident	Derived
2.	BRANCH	Varchar2	Branch service unit. The data type this attribute was Number initially.	Initially it was named CLM_DIVN_CODE
3.	CLM_ACCID_TYPE	Varchar2	The type of accident that has occurred	Initially it was named CLM_CLAIM_DESC
4.	VEH_TYPE	Varchar2	Refers the type of vehicle	
5.	VEH_AGE	Number	The original age of the vehicle	
6.	VEH_USE	Varchar2	The type of vehicle use. It is to mean that whether the vehicle is used for private or commercial purpose	Initially it was named CD_RISK_TYPE
7.	INSU_YR_S	Number	For how many years the insured stay in the company since policy issue date	Derived
8.	POL_STATUS	Varchar2	Refers to the status of the policy	
9.	SRC_BUS	Varchar2	Refers about source of the business like governmental	Initially it was named CE_SRC_BUS_CODE
10.	CLM_REC_AMT	Number	The amount of recovered money from the previous claim by the insured	Derived

Table 4. 8 *The Final List of Attributes used in the Study*

CHAPTER FIVE

EXPERIMENTATION

This chapter presents steps and procedures followed during experimentations. The main objective of this research is, discovering regularities for predicting and detecting fraudulent claims within the insurance dataset. Having this purpose in mind, the model-building phase in the DM process of this investigation is carried out following a two-step DM approach. Hence, the unsupervised clustering technique and the supervised classification techniques are adopted. First, the given dataset is segmented into different clusters based on their similarity. Then the output of this clustering process is used for the classification task as an input. These techniques are implemented using Weka DM tool.

The description and evaluation of the performances of the clustering and classification models developed are presented. It applies the methods, techniques and algorithms of DM that are briefly discussed in Section 3.2 of Chapter Three to accomplish the objectives of the research.

For clustering purpose the K-Means clustering algorithm is selected, as it is a very good general-purpose clustering algorithm since K is easily determined (known) in our present study. This is followed by creating predictive model with the help of classification techniques such as, J48 Decision Tree and Naïve Bayesian classifiers, which are widely applicable in solving the current problem.

5.1 Experiment Design

A procedure or mechanism of how to test the model's quality and validity is needed to be set before the model is actually built. In order to perform the model building process of this study, 17810 training dataset is used to train the clustering and classification models. Once the clustering model is developed, the resulting clustered dataset is then used as an input for training the J48 decision tree and naïve bayes models. For validating the clustering result of this study the intra cluster similarity measure (within cluster sum of squared error) value, the number of iteration the algorithm has undergone to converge

and the domain experts' judgment is used. A threshold value is set to determine what patterns are discovered for each subsequent cluster models, which helps to identify and label the cluster dataset based on the fraudulence nature of insurance claims. The 10-fold cross validation and percentage split test options are used for training and testing the classification model. In addition to these a different 2210 testing dataset is used for testing the prediction performance of the classification model developed. These testing dataset is prepared by simple random sampling technique from the original dataset.

5.2 Cluster Modelling

Once the dataset is ready to be used, the next step is building the clustering model using the selected DM tool. As it was discussed before, the Weka version 3.7.0 (for the Mac OS) software is used for conducting this study. The Weka 3.7.0 explorer includes different parameters for K-Means clustering. Some of the basic parameters are discussed as follows:

- Distance Function: this option is used to choose the distance function that is used to perform the distance calculation.
- The number of clusters: this option is used to set the K value i.e. the number of clusters that need to be created.
- Seed size: this option is also used to set the random number of seed to be used. This defines the number of data tuples the cluster must start with.

The clustering task of segmenting insurance claims is done using the Weka simple K-Means algorithm. The number of clusters chosen should be driven by how many clusters the business can manage. Accordingly, the business experts have been consulted in setting the optimal value. They have suggested that the K value to be 2 (representing FRAUD and NON-FRAUD insurance claims). This cluster model is experimented and evaluated against its performance in creating dissimilar clusters/segments when the default parameters are changed. According to the works of Halkidi and Vazirgiannis (2001), the notion of "good" clustering is strictly related to the application domain and its specific requirements. Nevertheless the generally accepted criterions for validating the clustering results in different domains are the measures of separation among the clusters

and cohesion within clusters (i.e. inter and intra cluster similarities respectively). So, for validating the clustering result of this study the intra cluster similarity measure (within cluster sum of squared error) value is used. In addition to this the number of iteration the algorithm has undergone to converge and the domain experts judgment is used. Once the necessary parameters are set, the experiment is undertaken in a stepwise manner.

Before conducting the experiment, the threshold value is set for each numeric attributes used to build the clustering model. The threshold value for each attribute has been determined together with the domain experts in the claim department and with the aid of the Weka's minimum, maximum, and mean values displayed for each attribute. The need for determining the threshold values is solely to determine what patterns are discovered for each subsequent cluster models with K=2, and changing the other default parameters. This helps a lot to identify fraud suspicious segments easily. Table 5.1 depicts the threshold values set for each of the variables suggested by the domain experts.

CLM_REP_DT_G (CRDG)	**VEH_AGE (VA)**	**INSU_YR_S (IYS)**	**CLM_RECV_AMT (RA)**
CRDG<=3 Fast	VA>10 Old	IYS>5 Old	CRA>=50000 Very High
4<=CRDG<=10 Moderate	6<=VA<=10 Young	5>=IYS>1 Young	50000>CRA>10000 High
CRDG>10 Slow	VA<=5 New	IYS<=1 New	1500<CRA<=10000 Medium
			CRA<=1500 Low

Table 5. 1 *List of range of conditions (thresholds) used to assess the cluster result*

Table 5.2 depicts the abbreviated terms and attributes, which are used in the following clustering models analysis and comparison discussion.

List of Abbreviated Terms		List of Abbreviated Attributes	
Abbreviated Terms	**Description**	**Abbreviated Attributes**	**Description**
FT	Fast	CRDG	Claim Report Date Gap
MDT	Moderate	VA	Vehicle Age
SW	Slow	IYS	Insured years of stay
O	Old	CRA	Claim Recovered Amount
Y	Young	B	Branch
N	New	CA	Claim Accident Type Description
H	High	VT	Vehicle Type
VH	Very High	VU	Vehicle Use
M	Medium	PS	Policy Status
L	Low	CT	Claim Type
COL-TP	Collision with 3rd party	CSB	Claim Source of Business
WBR	Windshield breakage		
VHP	Vehicle hit pedestrian		
COL	Collision		
OT	Overturning		

Table 5. 2 *List of Abbreviated Terms and Attributes along with their Description*

For the purpose of developing the cluster model of this study, the following three experiments are conducted. The experiments are conducted by changing the default parameters of the simple K-Means algorithm. These experiments are presented and discussed in this section. From the three different experiments conducted in this study, one of the best is chosen for developing the final cluster model of this study.

5.2.1 Experimentation I

The first experimentation is done for K=2, with the default seed value and default distance function. All of the final selected attributes and 17810 records are used as an input for the experiment. In order to cluster the records based on their values, the model is trained by using the default values of the program. The use training set cluster mode is employed to make use of the dataset for training.

Table 5.3 exhibits the result of the first experiment and the resulting segments. The algorithm is instructed to segment the dataset into two clusters. As can be seen from this table, the distribution of the dataset for each cluster is presented.

Clustering Result of the 1st Experiment				
K	**Distance Function**	**Seed Value**	**Cluster Distribution**	
			C1	**C2**
2	EuclideanDistance	10	10471(59%)	7339(41%)

Table 5. 3 *Training of the first experiment by the default parameter values*

As can be seen from Table 5.3, the first experiment is conducted with default values of the simple K-Means algorithm (K = 2, Seed = 10, and Euclidean distance function). The table below shows results of the first experiment.

Clustering Result of the First Experiment											
Cluster #	**Dist. of instances (in %)**	**CRDG**	**VA**	**IYS**	**CRA**	**VU**	**CA**	**B**	**VT**	**PS**	**CSB**
1	10471(59%)	9.25	10.89	1.43	10597.71	COM	COL-TP	FILW	VT110	R	CRB6
2	7339(41%)	6.85	9.88	1.44	2394.30	PR	WBR	FILW	VT110	R	CRB4
Total	17810(100%)										
Cluster #	**Dist. of instances (in %)**										
1	10471(59%)	MDT	O	Y	H	COM	COL-TP				
2	7339(41%)	MDT	Y	Y	M	PR	WBR				
Total	17810(100%)										

Table 5. 4 *Cluster result of the first experiment for K=2, Seed =10, Euclidean distance function*

As can be seen from Table 5.4, the upper part exhibits the attributes' average values for each segment (1-2) while the bottom part shows the corresponding mapping of these values into the discrete values. The mapping is performed to facilitate the comparison among the segments generated from the clustering algorithm.

After the average value of each attribute in each cluster has been replaced with the corresponding discrete values, a description for each segment of the cluster has been done as shown in Table 5.5. The ranking is determined based on the fraudulence nature of the insurance claims.

Cluster #	Description	Rank
1.	Moderate report date, old vehicle age, young customer, high recovered amount, Filwuha branch, collision with third party accident, VT110 vehicle type, commercial use of vehicle, renewed policy status, SRB6 business category	1
2.	Moderate report date, young vehicle age, young customers, medium recovered amount, Filwuha branch, windshield breakage accident, VT110 vehicle type, private use of vehicle, renewed policy status, SRB4 business category	2

Table 5. 5 *Cluster summary of the first experiment for K=2, seed=10, Euclidean distance and rank of clusters*

As stated in the business understanding section of Chapter Four, fraudulent suspicious claims are those with late accident report date, an old age of vehicle, a short stay of the insured, and a very high-recovered amount of money.

So, the above ranking of the clusters has been assigned depending on the aforementioned facts of the business. As described in Table 5.5 the first cluster is ranked first. It is because of the reason that the claims that are assigned in this cluster are reported within a moderate time range (takes slightly long time to report after the accident has occurred), with an old age of vehicle, and a high-recovered amount of money from the previous claims. In addition the type of the accident that has been occurred is collision with third party and the vehicle is used for commercial purpose.

The second cluster is ranked as second because of the reason that the claims that are grouped in this cluster have medium accident report date, medium age of vehicle, medium recovered amount of money. In addition the vehicle was used for private purpose. Generally, cluster 1 (ranked 1$^{st)}$ is considered as containing those suspicious

fraudulent insurance claims, while Cluster 2 (ranked 2nd) is considered as containing non-fraudulent insurance claims.

Since the first experimentation has created greater number of fraudulent claims, which is assumed to be smaller, compared with the non-fraudulent insurance claims, conducting another experiment is necessary. In addition to this the result of this experiment showed that within cluster sum of squared error is higher, which means that the segmented instances within a cluster don't have that much similarity. Because of this reason the second experimentation is performed with changed seed value=100 and Euclidean distance.

5.2.2 Experimentation II

The second cluster experiment is done for K=2, changed default seed value (10) to 100 and with the default distance function (Euclidean distance). Similar to the first experiment, all of the final selected attributes and 17810 records are used as an input for conducting the experiment. Table 5.6 exhibits the result of the second experiment and the resulting segments.

Clustering Result of the 2nd Experiment				
K	**Distance Function**	**Seed Value**	**Cluster Distribution**	
			C1	**C2**
2	EuclideanDistance	100	10002(56%)	7808(44%)

Table 5. 6 *Training of the second experiment by changed Seed value =100 and other default parameter Values*

The result of this cluster experiment with K=2, seed =100 and Euclidean distance function is depicted in Table 5.7. The description of the resulting cluster model is then interpreted and briefly explained. The table below shows the result of this experiment.

Clustering Result of the 2nd Experiment											
Cluster #	**Distribution of instances (in %)**	**CRDG**	**VA**	**IYS**	**CRA**	**VU**	**CA**	**B**	**VT**	**PS**	**CSB**
1	10002(56%)	6.14	9.91	1.57	1115.98	COM	OT	FILW	VT110	R	CRB4
2	7808(44%)	10.99	11.19	1.27	15033.09	COM	COL-TP	FILW	VT110	E	CRB6
Total	17810(100%)										
Cluster #	**Distribution of instances (in %)**										
1	10002(56%)	MDT	Y	Y	L	COM	OT				
2	7808(44%)	SW	O	Y	H	COM	COL-TP				
Total	17810(100%)										

Table 5. 7 *Cluster result of the 2nd experiment for K=2, Seed=100, Euclidean distance function*

Similar to the first cluster experiment, two clusters are formed in the second experiment. This cluster has resulted in similar segment formation with the first experiment result. Table 5.8 below shows the description of the second experimentation. Interpretation of the result is also done.

Cluster #	Description	Rank
1.	Moderate report date, young vehicle age, young customer, low recovered amount, Filwuha branch, overturning accident, VT110 vehicle type, commercial use of vehicle, renewed policy status, SRB4 business category	2
2.	Late (slow) report date, old vehicle age, young customers, high recovered amount, Filwuha branch, collision with third party accident, VT110 vehicle type, commercial use of vehicle, endorsed policy status, SRB6 business category	1

Table 5. 8 *Cluster summary of the 2nd experimentation for K=2, seed=100, Euclidean distance function and rank of clusters*

As can be seen from Table 5.8 the second cluster is ranked 1st. It is because of the reason that the claims that are assigned in this cluster have late accident report date, an old age of vehicle, and high-recovered amount of money. In addition to this the type of accident that has been occurred is collision with third party.

The first cluster is ranked second because it has an old age of vehicle, moderate accident report date, and a low recovered amount of money. Cluster 2 is considered as suspicious of fraudulent claims while cluster 1 is considered as suspicious of non-fraudulent claims. Compared to the first run experimentation this one seems to creating dissimilar clusters with respect to the recovered amount of money. In addition to this the segmented clusters represent lower fraudulent insurance cases.

Compared with the first experimentation, the value of within clustered sum of squared error is lowered in this experiment. Apart from this the number of iteration that the algorithm used to converge is also minimized. The number of non-fraudulent insurance claims is also higher than the fraud suspicious claims in this experimentation. This showed that the result of this experiment is better than the first one in creating dissimilar clusters.

Though the second experimentation with a change in seed value seems a good segmentation for the problem at hand, conducting another experiment with a changed distance function and seed value is important in searching for a good clustering model.

5.2.3 Experimentation III

The final cluster experiment is done for K=2, changed default seed value (10) and distance function (Euclidean distance). Similar to the first two runs, all of the final selected attributes and 17810 records are used as an input for conducting the experiment. This experiment is conducted with changed default distance function and seed value. Table 5.9 exhibits the result of the third cluster experiment, with K=2, seed =1000, and Manhattan distance function.

Clustering Result of the 3rd Experiment				
K	**Distance Function**	**Seed Value**	**Cluster Distribution**	
			C1	**C2**
2	ManhattanDistance	1000	10438(59%)	7372(41%)

Table 5. 9 *Training of the third cluster experiment with K=2, Seed=1000 and Manhattan distance function*

The result of this cluster experiment with K=2, seed =1000, and Manhattan distance function is presented in the following table (Table 5.10). The description of the resulting cluster model is then interpreted and briefly explained.

Clustering Result of the 3rd Experiment											
Cluster #	**Dist. of instances (in %)**	**CRDG**	**VA**	**IYS**	**CRA**	**VU**	**CA**	**B**	**VT**	**PS**	**CSB**
1	10438(59%)	10	10	1	3250	COM	COL-TP	FILW	VT110	R	CRB6
2	7372(41%)	3	9	1	0	COM	OT	FILW	VT250	R	CRB4
Total	17810(100%)										
Cluster #	**Dist. of instances (in %)**										
1	10438(59%)	MDT	Y	N	M	COM	COL-TP				
2	7372(41%)	FT	Y	N	L	COM	OT				
Total	17810(100%)										

Table 5. 10 *Cluster result of the third experiment for K=2, Seed=1000, Manhattan distance function*

Similar to the first two cluster experimentations, two clusters are formed in this experiment. This cluster experiment has resulted in similar segment formation with the first and second experiment results. Table 5.11 shows the description of the third cluster experimentation.

Cluster #	Description	Rank
1	Moderate report date, young vehicle age, new customer, medium recovered amount, Filwuha branch, collision with third party accident, VT110 vehicle type, commercial use of vehicle, renewed policy status, SRB6 business category	2
2	Quick report date, young vehicle age, new customer, low recovered amount, Filwuha branch, overturning accident, VT110 vehicle type, commercial use of vehicle, renewed policy status, SRB4 business category	1

Table 5. 11 *Cluster Summary of the third Experiment for K=2, seed=1000, Manhattan Distance Function and Rank of Clusters*

As can be seen from Table 5.11, the second cluster is ranked 1^{st} because of the reason that the claims that are segmented in this class have a medium amount of recovered money, young vehicle age, a short stay of the insured with the company, and a collision accident with a third party. The first cluster is ranked as 2^{nd} because it has a short accident report date, a low recovered amount of money. The second cluster seems to show suspicious fraudulent insurance claims, while the first cluster shows suspicious non-fraudulent claims. But when we see this two clusters almost they are very similar. Compared with the 1^{st} and the 2^{nd} cluster experiments, this experiment doesn't create dissimilar segments.

Though this experiment was conducted with changed distance function (Manhattan distance function) and seed value, the resulting cluster is not better than the result of the first two experimentations. In this experiment the number of iteration that the algorithm takes to converge and the value of within cluster sum of squared error are increased compared with the preceding experimentations. Generally, this experiment is failed to create dissimilar clusters of insurance claims.

5.2.4 Choosing the best clustering model

Different experiments of the K-means algorithm were conducted with K=2 and with changed default seed values and distance function. The entire dataset output from the different cluster experiments was available together with their segment distribution and the resulting cluster output. This enabled the domain experts to compare resulting insurance claim segments from the different cluster experiments. Though experimentation was carried out for different seed values and distance function, only those that resulted in a good segmentation are presented.

As described in Section 5.2 of this chapter, the cluster validity is a very difficult issue and subject of endless arguments since the view of good clustering is strictly related to the application domain and its specific requirements. But usually in most domains the intra and inter cluster similarity values are used for the validation of good clusters. The value of within cluster sum of squared error is used to evaluate the goodness of clustering in the Weka DM tool. The lower value indicates that the records segmented within the same cluster are more related with each other. In addition to this the discovery of patterns requires that there is close interaction with domain experts, which allows them to interact with the output. So, the evaluation of the final clustering results also incorporated the suggestion of domain experts. Generally, in this study the best cluster model from the three cluster experimentations has been chosen based on the following criterions.

I. Within cluster sum of squared errors. This is a measure of the "goodness" of the clustering and tells how tight the clustering is overall. Lower values are better.
II. The number of iteration the algorithm has undergone to converge. This shows the algorithm has relocated all misplaced data items in their correct classes within a few looping. The minimum value exhibits K-Means algorithm has converged very soon.
III. The domain experts' judgment. Suggestions of experts about the best model with respect to the nature of the business.

The summary of the result of these criterions is depicted in Table 5.12.

Experimentation	Number of Iteration	Within Cluster Sum of Squared Error
I	4	3186
II	3	2772
III	4	7935

Table 5. 12 *Within cluster sum of squared error values of the three cluster experimentations*

As can be seen from Table 5.12, the second cluster experiment shows the least number of iterations and within cluster sum of squared errors compared with the first and the third cluster experimentations. This shows that the second experiment is good in creating dissimilar clusters. In addition to this the domain experts are consulted to give their suggestion whether the clustering result matches with the business. Accordingly, the experts suggested that the model of the second cluster experiment is good in segmenting the different fraud suspicious insurance claims compared to the other models developed by the first and third cluster experimentations. So the model developed in the second cluster experiment is selected as the final clustering model of this study.

5.3 Classification Modelling

Once the clustering model is developed, the next step of this study is developing the predictive model using the classification techniques. As can be seen in the foregoing discussion, the resulted clustering model identified segments of insurance claims that share high intra-class similarity and low inter-class similarity. Since the developed clustering model does not classify new instances of insurance claims into a certain segment, the classification process is carried out.

For starting the classification modeling experiments, the decision tree (in particular the J48 algorithm) and the naïve bayes methods are selected. In order to classify the records based on their values for the given cluster index, the model is trained by changing the default parameter values of the algorithms.

The training of the decision tree classification models of the experimentation is done by employing the 10-fold cross validation and the percentage split classification modes. The

classification is analyzed to measure the accuracy of the classifiers in categorizing the insurance claims into specified classes. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications (Baeza-Yates and Ribeiro-Neto 1999). The classification accuracy of each of these models is reported and their performance is compared in classifying new instances of records. A separate test dataset is used for testing the performance of the classification models.

5.3.1 J48 decision tree model building

In this phase of the study, the resulting clustering model is used as an input for building the decision tree model. Taken as a whole, a decision tree is a classifier. Any previously unseen record can be fed into the tree. At each node it will be sent either left or right according to some test. Eventually, it will reach a leaf node and be given the label associated with that leaf. Generally, this research is more interested in generating rules that best predict the fraud exposure of insurance claims and to come to an understanding of the most important factors (variables) affecting the insurance claims to be fraudulent.

As described before, the J48 algorithm is used for building the decision tree model. All of the attributes, which are selected for the cluster model building, are fed as independent variables and the cluster labels, which are assigned by the clustering algorithm, as dependent variable for the algorithm.

J48 algorithm contains some parameters that can be changed to further improve classification accuracy. Initially the classification model is built with the default parameter values of the J48 algorithm. The following table summarizes the default parameters with their values for the J48 decision tree algorithm.

Parameter	Description	Default Value
confidenceFactor	The confidence factor used for pruning (smaller values incur more pruning)	0.25
minNumObj	The minimum number of instances per leaf	2
Unpruned	Whether pruning is performed	False

Table 5. 13 *Some of the J48 algorithm parameters and their default values*

By changing the different default parameter values of the J48 algorithm, the experimentations of the decision tree model-building phase is carried out.

5.3.1.1 Experiment I

The first experimentation is performed with the default parameters. The default 10-fold cross validation test option is employed for training the classification model. Using these default parameters the classification model is developed with a J48 decision tree having 84 number of leaves and 122 tree size. The decision tree used six of the total ten variables for generating the tree. These are CLM_REP_DT_G, CLM_ACCID_TYPE, VEH_TYPE, POL_STATUS, SRC_BUS_CODE, and CLM_REC-AMT. The decision tree has also shown that the CLM_REC_AMT variable is the most determining one. Table 5.14 depicts the resulting confusion matrix of this model. The snapshot confusion matrix taken from the tool is depicted in Appendix 3-A.

Actual	**Predicted**		**Total**	**Correctly Classified**
	Cluster 1	Cluster 2		
Cluster 1	10002	0	10002	100%
Cluster 2	6	7802	7808	99.92%
	10008	7802	17810	99.96%

Table 5. 14 *Confusion matrix output of the J48 algorithm with default values*

As we can see from the resulting confusion matrix, the J48 learning algorithm scored an accuracy of 99.96%. This result shows that out of the total training datasets 17802 (99.96%) records are correctly classified, while only 6 (0.036%) of the records are incorrectly classified. The accuracy of the model shows us that the classification is good.

Furthermore, the resulting confusion matrix of this experiment has shown that 100% of the records are correctly classified in the first cluster (cluster 1). This shows that the algorithm classified all of the non-fraud suspicious insurance claims in their respective class. In addition to this out of the 7808 fraud suspicious insurance claims, who are described in cluster 2 of Table 5.7, 7802 (99.92 %) of them are classified correctly in their designated cluster, i.e. cluster 2, while only 6 (0.077 %) of them are misclassified in

cluster 1. Compared with the records that are classified in cluster 2, those records that are classified under cluster 1 are fully correctly classified.

The performance of the developed classification model with the default parameter values is then tested with the separately prepared 2210 testing dataset. The prediction performance of this model is tested using a java code. First the model, using the training dataset, is created. Once the model is created it is then called for predicting instances of the testing dataset into one of the predefined class labels (fraud or non-fraud). The commands used for training and testing J48 decision tree classification model are the following.

- Java weka.classifiers.trees.J48 -C 0.25 -M 2 -t [path]/TrainingdataName.arff -d [destination path]/ModelName.model
- Java weka.classifiers.trees.J48 –p *position of the class label to be predicted* [path]/ModelName.model -T [destination path]/TestingdataName.arff

The performance of the model with this testing dataset was 97.19%. This result shows that out of the 2210 testing datasets, the developed decision tree classification model predicted around 2147 (97.19%) records correctly.

As described before, the size of the tree and the number of leaves produced from this training was 122 and 84 respectively. This seems that it is difficult to traverse through all the nodes of the tree in order to come out with valid rule sets. Therefore, to make ease the process of generating rule sets or to make it more understandable, the researcher attempted to modify the default values of the parameters so as to minimize the size of the tree and number of leaves. With this objective, the minNumObj (minimum number of instances in a leaf) parameter was tried with 25, 20, 15, 10 and 5. But the minNumObj set to 20 gives a better tree size and accuracy compared with the other trials. With this value of the minNumObj the process of classifying records proceeds until the number of records at each leaf reached 20. Table 5.15 depicts the result of this experiment. The snapshot confusion matrix taken from the tool is depicted in Appendix 3-B.

Actual	Predicted		Total	Correctly Classified
	Cluster 1	Cluster 2		
Cluster 1	9979	23	10002	99.77%
Cluster 2	47	7761	7808	99.39%
	10026	7784	17810	99.61%

Table 5. 15 *Confusion matrix output of the J48 algorithm with changed minNumObj parameter set to 20*

This experiment has shown an improvement in the number of leaves and tree size. The size of the tree is lowered from 122 to 93 and the number of the leaves decreased to 66 from 84. As we can see from Table 5.22, the resulting confusion matrix shows that the J48 decision tree algorithm scored 99.61% accuracy. This result shows that out of the total 17810 records 17740 (99.61%) are correctly classified. Only 70 (0.39%) records of the total dataset are misclassified.

Furthermore, the confusion matrix of this experiment has shown that 9979 (99.77%) of the 10002 total non-fraudulent insurance claims, are correctly classified in cluster 1, while 23 (0.13%) of them are misclassified in cluster 2 of the fraudulent insurance claims class. In addition to this the confusion matrix also shows that out of 7808 total fraudulent insurance claims records of cluster two, 7761 (99.39%) of the records are correctly classified while 47 (0.71%) of the records are misclassified in the non-fraudulent insurance claims of cluster 1.

The resulting confusion matrix of this experiment has also shown that the first cluster of the non-fraudulent insurance claims class correctly classified the records than the second fraud suspicious insurance claims cluster. In general, though the tree size and the number of the leaves decreased from 122 and 84 to 93 and 66 respectively, the accuracy of the J48 decision tree algorithm in this experiment is poorer than the first experiment with the default parameter value of the minNumObj.

Although the size of the tree and the number of leaves are lowered in this experiment, their value has no that much difference compared with the first one. That means the complexity of the decision tree to generate rules is the same in both the experiments. So,

since there is no a tangible difference in the tree size and number of leaves in the two experiments and the accuracy of the model is decreased from 99.96% to 99.61% in this experiment, the first experiment with the default minNumObj parameter value is taken as the J48 decision tree model.

5.3.1.2 Experiment II

This experiment is performed, by changing the default testing option (the 10-fold cross validation). In this learning scheme a percentage split is used to partition the dataset into training and testing data. The purpose of using this parameter was to assess the performance of the learning scheme by increasing the proportion of testing dataset if it could achieved a better classification accuracy than the first experimentation. First this experiment has run with the default value of the percentage split (66%). But the one with the better classification accuracy is presented here. So the percentage split parameter set to 70, which is to mean 70% for training and 30% for testing, resulted with a better accuracy. The result of this learning scheme is summarized and presented in Table 5.16. The snapshot confusion matrix taken from the tool is depicted in Appendix 3-C.

Actual	**Predicted**		**Total**	**Correctly Classified**
	Cluster 1	Cluster 2		
Cluster 1	2969	4	2973	99.85%
Cluster 2	0	2370	2370	100%
	2969	2374	5343	99.92%

Table 5. 16 *Confusion matrix output of the J48 algorithm with the percentage-split set to 70%*

In this experiment out of the 17810 total records 12467 (70%) of the records are used for training purpose while 5339 (30%) of the records are used for testing purpose. As we can see from the confusion matrix of the model developed with this proportion, out of the 5339 testing records 99.92% of them are correctly classified. Only 4 (0.074) records are incorrectly classified.

In addition to this the resulting confusion matrix has shown that out of 2973 non-fraud suspicious records 2969 (99.85%) of them are correctly classified while only 4 (0.15) of

the records are misclassified in cluster 2 as a fraud suspicious instances. Furthermore, the confusion matrix of this experiment shown, that 100% of the records are correctly classified in the second cluster (fraud suspicious class). This shows that the model correctly classified those fraud suspicious insurance claims in their respective class.

By and large, in this experiment when the testing data is increased the performance of the algorithm for predicting the newly coming instances is also diminished as well. Though this experiment is conducted by varying the value of the training and the testing datasets, the accuracy of the algorithm for predicting new instances in their respective class couldn't be improved. This shows that the previous experiment conducted with the default 10-fold cross validation, is better than this experiment.

Generally, from the three experiments conducted before, the model developed with the default parameter values of the J48 decision tree algorithm and the default 10-fold cross validation test option gives a better classification accuracy of predicting newly arriving insurance claims in their respective class category. Therefore, among the different decision tree models built in the foregoing experimentations, the first model, with the default parameters' values and 10-fold cross validation, has been chosen due to its better overall and individual cluster classification accuracy. A tree generated from this model is depicted in Appendix 4**.**

5.3.2 Naïve Bayes model building

The second DM technique employed for the classification sub phase is the naïve bayes. To build the naïve bayes model, Weka software package is used and it employs the Naïve Bayes Simple algorithm in developing the model. In order to build the model, the resulting clustering model of the clustering phase is used as an input. The 10-fold cross validation, which is set by default, and the percentage split with 75-25 for training and testing the model test options are employed.

Naïve Bayes makes predictions using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other

attributes (Han and Kamber 2006). The first experiment of the naïve bayes model building is performed using the Naïve Bayes Simple algorithm with the default10-fold cross validation test option. Table 5.17 shows the resulting confusion matrix of the model developed using the Naïve Bayes Simple algorithm with the default 10-fold cross validation test option. The snapshot confusion matrix taken from the tool is depicted in Appendix 3-D.

Actual	**Predicted**		**Total**	**Correctly Classified**
	Cluster 1	Cluster 2		
Cluster 1	9620	382	10002	96.18%
Cluster 2	1099	6709	7808	85.92%
	10719	7091	17810	91.10%
10-fold Cross Validation				

Table 5. 17 *Confusion matrix output of the Naïve Bayes Simple algorithm*

As can be seen from the resulting confusion of this experiment, the Naïve Bayes Simple Algorithm scored an accuracy of 91.10%. This shows that out of the total 17810 records 16329 (91.10%) of the records are correctly classified, while 1481 (8.9%) of the records are misclassified.

Furthermore, the resulting confusion matrix also shows that out of the total 10002 non-fraud suspicious insurance claim records 9620 (96.18%) of them are correctly classified in their respective class, while 382 (3.82%) of the records are incorrectly classified in the fraud suspicious claims cluster segment (cluster 2). In addition to this out of the total 7808 fraud suspicious insurance claim records 6709 (85.92%) of them are correctly classified, while 1099 (14.08%) of the records are misclassified in the non-fraud suspicious cluster segment (cluster 1). Generally from the two clusters, cluster 1 (non-fraud suspicious insurance claims) correctly classified the records than the second cluster (fraud suspicious insurance claims).

The result from this experiment shows that the model developed with Naïve Bayes Simple Algorithm is poor in the accuracy of classifying new insurance claims to the respected class, compared with the decision tree model that is developed before.

The second experiment of the naïve bayes model building is performed using the Naïve Bayes Simple algorithm with the 75-25 training and testing percentage split test option. Though different experiments are conducted by changing the size of the training and testing datasets, the one with 75-25 training and testing dataset scored better classification accuracy and it is presented here. The result of this experiment is shown in Table 5.18.

Actual	**Predicted**		**Total**	**Correctly Classified**
	Cluster 1	Cluster 2		
Cluster 1	2337	131	2468	94.69%
Cluster 2	297	1687	1984	85.03%
	2634	1818	4452	90.30%
Percentage Split (75-25 Training and Testing) Test Option				

Table 5. 18 *Confusion matrix output of the Naïve Bayes Simple algorithm*

As can be seen from the confusion matrix that resulted from the model developed by the Naïve Bayes Simple Algorithm with the 75-25 percentage split, the model scored an accuracy of 90.30%. This shows that from the total 4454 test data, 4024 (90.30%) of the records are correctly classified, while 428 (9.7%) of them are misclassified. In addition to this the confusion matrix also shows that from the total 2468 non-fraudulent insurance claim records 2337 (94.69%) of them are correctly classified, while 131 (5.31%) of the records are misclassified in the fraud suspicious insurance claims cluster segment (cluster 2). Furthermore, the confusion matrix also shows that out of the total 1984 fraud suspicious insurance claims 1687 (85.03%) of the records are correctly classified in the fraud suspicious claims cluster segment, while 297 (14.97%) of them are incorrectly classified in the non-fraud suspicious claims cluster segment (cluster 1). Compared with the second cluster the first cluster is better in classifying insurance claims correctly.

Generally, the first experiment that is conducted using the Naïve Bayes Simple algorithm with the default 10-fold cross validation test option generates a better classification model with a better classification accuracy than the second one conducted with 75-25 training and testing percentage split test option.

5.3.3 Comparison of J48 decision tree and Naïve Bayes models

Selecting a better classification technique for building a model, which performs best in handling the prediction and detection of fraudulent insurance claims, is one of the aims of this study. For that reason, the decision tree (particularly the J48 algorithm) and the bayes (the Naïve Bayes Simple algorithm in particular) classification methods were applied for conducting experiments to build the best model. Summary of experimental result for the two classification algorithms is presented in table 5.19.

Classification Model	**Overall Accuracy (17810 records)**	
	Correctly Classified	Misclassified
Decision Tree	17802 (99.96%)	8 (0.04%)
Naïve Bayes	16329 (91.10%)	1481 (8.9%)
10-fold Cross Validation		

Table 5. 19 *Accuracy of the J48 decision tree and Naïve Bayes models*

The result showed that J48 decision tree outperforms naïve bayes by 8.86% in identifying suspicious insurance fraud claims. The reason for the J48 decision tree to perform better than naïve bayes is because of the linearity nature of the dataset. That means there is a clear demarcation point that can be defined by the algorithm to predict the class for a particular insurance claim. Regarding the Naïve Bayes, scoring a lower accuracy than the J48 decision tree is due to its assumption that each attribute is independent of other attributes, which is not true in reality especially in the insurance dataset. Moreover, in terms of ease and simplicity to the user the J48 decision tree is more self-explanatory. It generates rules that can be presented in simple human language.

Therefore, it is plausible to conclude that the J48 algorithm is more appropriate to this particular case than the Naïve bayes method. So, the model that is developed with the J48 decision tree classification technique is taken as the final working classification model.

5.4 Evaluation of the Discovered Knowledge

The data required to undergo any DM task is the core of every process. However, unfortunately the data required for effective knowledge mining is not readily available in a format that the DM algorithm required it. To make things worse some of the fields may contain outliers, missing values, inconsistent data types within a single field and many other possible anomalies. But this must be cleansed, integrated and transformed in a format suitable for the DM task to be undertaken. For that reason the researcher has taken considerable time for the data-preprocessing task. Data cleaning (handling missing values, and outlier detection and removal), and data integration tasks are carried out in a format suitable for the clustering and classification techniques.

As discussed before the cluster model developed with K=2, seed value=100 and Euclidean distance function was chosen as the final clustering model. This model was selected because of the reason that it has a lower within cluster sum of squared error relative to the other models and the algorithm takes minimum number of iteration to converge. In addition to these the domain expert suggests that this model is good in segmenting fraudulent insurance claims. Similarly, the classification model developed using the J48 decision tree algorithm is chosen as the final model for this study.

From the decision tree developed in the aforementioned experiment, it is possible to find out a set of rules simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node (Berry and Linoff 2004). This produces rules that are unambiguous in that it doesn't matter in what order they are executed. The following are some of the rules extracted from the decision tree.

Rule 1. If CLM_REC_AMT<=10000 and CLM_REP_DT_G<=4 and
POL_STATUS=R and SRC_BUS_CODE=SRB6 and
CLM_REC_AMT<=3000 THEN cluster 1 (non-fraud suspicious).

Rule 2. If CLM_REC_AMT<=10000 and CLM_REP_DT_G>4 and
POL_STATUS=A and SRC_BUS_CODE=SRB4 and
CLM_ACCID_TYPE=COLLISION-TP THEN cluster 2 (Fraud suspicious)

Rule 3. If CLM_REC_AMT<=10000 and CLM_REP_DT_G<=4 and POL_STATUS=R and SRC_BUS_CODE=SRB6 and CLM_REC_AMT>4550 and VEH_USE=PRIVATE THEN cluster 2 (Fraud suspicious)

Rule 4. If CLM_REC_AMT<=10000 and CLM_REP_DT_G<=4 and POL_STATUS=C and CLM_ACCID_TYPE= OVERTURNING THEN cluster 1 (Non-Fraud suspicious)

Rule 5. If CLM_REC_AMT<=10000 and CLM_REP_DT_G<=4 and POL_STATUS=C and CLM_ACCID_TYPE= COLLISION-TP THEN cluster 2 (fraud suspicious)

Rule 6. If CLM_REC_AMT<=10000 and CLM_REP_DT_G<=4 and POL_STATUS=R and SRC_BUS_CODE=SRB6 and CLM_REC_AMT>4550 and VEH_USE=COMMERCIALTHEN cluster 1 (Non-Fraud suspicious)

Rule 7. If CLM_REC_AMT<=10000 and CLM_REP_DT_G>4 and POL_STATUS=R and SRC_BUS_CODE=SRB6 and CLM_REC_AMT<=515 and CLM_ACCID_TYPE= OVERTURNING THEN cluster 1 (non-fraud)

Rule 8. If CLM_REC_AMT>10000 THEN cluster 2 (fraud suspicious)

The rules that are presented above indicate the possible conditions in which an insurance claim record could be classified in each of the fraud and non-fraud suspicious classes. Six of the total ten variables are used for constructing the decision tree model. These attributes are claim report date, recovered amount of money, policy status, source of business, vehicle use, and accident type which are the basis for building the decision tree. From these, the generated decision tree has shown that the recovered amount of money is the most determinant variable, which is the top splitting variable of the model. In addition to this the model has also shown that the length of claim reporting date is another determining variable for making decisions. A claim is more likely to be fraudulent if there is a long delay in contacting the company. Rule 2 shows this fact that long delay of reporting can cause the claim to be fraudulent. This finding is consistent with the general idea that time is required to building up an accident with fraud. As discussed in the

business understanding section of Chapter Four, the length of claim report date can be an important factor for suspicious fraudulent insurance claims. The rule generated at number 4 also showed that if the claim is reported quickly, the type of the accident is overturning, and the policy status is cancelled this claim is that much to be dishonest.

Moreover, this decision tree has shown that the policy status variable can also be used in the process of decision-making. The result of the study has showed that most of renewed policies are exposed to be fraudulent. The above rule (Rule 6) has also showed that the commercial use of vehicles has a lower probability of being involved in dishonest claims than private use of vehicles. But this rule is contradicted with the company's idea that commercial uses of vehicles are more likely to be involved in dishonest claims. The rule generated at number 8 shows that if the recovered amount of money is greater than 10000 then the claim is highly suspicious to be fraud. These generated rules also showed that collision accidents with a third party and a SRB6 business sources (private companies) are influencing variables for the insurance claim to be fraudulent. Generally, the generated rule has showed that renewed policies, private company business sources, and the recovered amount of money variables are more likely to be involved in dishonest claim activities.

However, age of vehicle, year of stay of the claimant, branch, and the vehicle type variables are not used in developing the decision tree. Though drivers with older cars may be tempted to obtain the cash value of the car from the insurance company before buying a new car, the vehicle age attribute does not have an influence in building the decision tree of this study.

The researcher has faced different challenges in conducting this study. The first challenge was the dataset obtained from the company, which does not have the target class of the study. Because of this the study employed a two-step data mining technique for solving this problem. The preprocessing task of this study was also very challenging. Especially, selecting those important attributes for the study, integrating the different tables to build a single dataset appropriate for the data-mining tool and many others.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

The application of DM technology has increasingly become very popular and proved to be relevant for many sectors such as insurance, airline, telecommunications, banking, and healthcare industries. Particularly in the insurance industry, DM technology has been applied for fraud detection. As a matter of fact insurance fraud is the most challenging problem in today's motor insurance business.

In this research, an attempt has been made to apply the DM technology in support of detecting and predicting fraudulent insurance claims in the insurance industry. The six-step Cios et al. (2000) process model has strictly been followed while undertaking the experimentation. This process model embraces understanding the problem domain, understanding the data, preparation of the data, DM, evaluation of the discovered knowledge, and using the discovered knowledge phases.

The data used in this research has been gathered from the PREMIA database of Africa Insurance Share Company. Once the data has been collected, it has been preprocessed and prepared in a format suitable for the DM tasks. This phase took considerable time of the study. The study was then conducted in two sub phases, first the clustering followed by classification phase.

The initial data collected from AIC didn't incorporate the target class for this study. The clustering sub phase has been then conducted using the K-Means clustering algorithm for segmenting the data into the target classes of FRAUD and NON-FRAUD. By changing the default parameters of the algorithm three different clustering experiments have been conducted for generating a plausible model that can create dissimilar clusters of insurance claims. The models from these three experimentations are interpreted and evaluated. Among the three models, the one with K=2, Seed value=100 and Euclidean distance

function has showed better segmentation of the insurance claims. This model created, dissimilar clusters of FRAUD and NON-FRAUD insurance claims. The model segmented those lately reported claims as fraud suspicious claims. This result of the model complies with the company's assumption that lately reported claims are more of fraud suspicious claims.

Cluster records are then submitted for the classification module for model building using the J48 decision tree algorithm. By changing the training test options and the default parameter values of the algorithm, different decision tree models have been created. The model developed with the 10-fold cross validation with the default parameter values has shown a better classification accuracy of 99.96% on the training dataset, with the CLM_REC_AMT as a splitting variable. This model is then valuated with a separate test dataset and scored an accuracy of 97.19% of classifying new insurance datasets as fraud and non-fraud suspicious claims.

In general, the results from this study are very promising. The study has shown that it is possible to identify those fraud suspicious insurance claims and suggest concrete solutions for detecting them, using the DM techniques.

6.2 Recommendations

This research is mainly conducted for an academic purpose. However, that the results of this study are found promising to be applied to address practical problems of insurance fraud. This research work can contribute a lot towards a comprehensive study in this area in the future, in the context of our country. The results of this study have also shown that the DM technology particularly the K-Means clustering and the J48 decision tree classification technique are well applicable in the efforts of insurance fraud detection.

Hence, based on the findings of this study, the following recommendations are forwarded.

- The predictive model, which is developed in this research, generated various patterns and rules. For the company to use it effectively there is a need to design a knowledge base system, which can provide advice for the domain experts.
- The model building process in this investigation was carried out in two sub-phases. For clustering the researcher uses the simple K-Means algorithm whereas for classification J48 decision tree algorithm. Though, the results were encouraging, accuracy decreases on sample test datasets. So further investigation needs to be done using other classification techniques such as Neural Networks and Support Vector Machine.
- For this work, only a limited number of all possible attributes are available with their values in the database of the company. There are inconsistency and missing values in the database. There is no record related to age, sex, health status, driving experience, and marital status of driver and the place where and when the accident has occurred, and traffic police report, which are important fraud factors. Since data is the most important component in DM research, the company has to design a data warehouse where operational and non-operational data can be kept.
- In this research we didn't consider records related to age, sex, health status, driving experience, and marital status of driver and the place where and when the accident has occurred, and traffic police report. Future research can be undertaken by including these attributes.

- This research attempted to assess the potential application of DM techniques in detecting fraudulent motor insurance claims. However, researches can also be conducted in other insurance claims, other than motor insurance claim.
- Fraud is not only occurred in the insurance claims of the insured, it can also occur within the company by experts, governors and other staffs. These can also be taken as another area for further research.

REFERENCES

ACL Services Ltd. 2010, *Fraud Detection Using Data Analytics in the Insurance Industry*, Global Fraud Study: Report to the Nations on Occupational Fraud and Abuse, ACL Services Ltd.

AIC 2003, *The Study of Organizational Structure and Preparation of the Manual*, Africa Insurance Corporation S.C., Addis Ababa: Ethiopia.

AIC Underwriting Department 2004, *Underwriting and Procedures Manual*, Africa Insurance Corporation S.C. Underwriting Department, Addis Ababa: Ethiopia.

Anbarasi, M Anupriya, E and Iyengar, N 2010, 'Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm', *International Journal of Engineering Science and Technology*, Vol. 2(10), pp. 5370-5376.

Apte, C and Weiss, M 1997, *Data Mining with Decision Trees and Decision Rules*, Future Generation Computer Systems, New York.
www.research.ibm.com/dar/papers/pdf/fgcsapteweiss_with_cover.pdf
Access Date: December 20, 2010.

Askale, W 2001, 'Data Mining Application in Support of Loan Disbursement Activity at Dashen Bank SC.', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Azevedo, A and Santos, F 2008, 'KDD, SEMMA AND CRISP-DM: A Parallel Overview', *IADIS European Conference Data Mining*, Portugal, pp. 182-185.

Baeza-Yates, R and Ribeiro-Neto, B 1999, *Modern Information Retrieval*, ACM Press, Addison Wesley.

Berry, M & Linoff, G 2004, *Data Mining Techniques for Marketing, Sales and Customer Relationship Management*, 2nd edn, Wiley Publishing, Inc., Indianapolis, Indiana.

Berry, M and Linoff, G 2000, 'Mastering Data Mining: The Art of Science of Customer Relationship Management', *John Willy and Sons Inc*, New York.

Bharti, K Jain, S and Shukla, S 2010, 'Fuzzy K-mean Clustering Via J48 for Intrusiion

Detection System', *International Journal of Computer Science and Information Technologies*, Vol. 1 (4) , pp. 315-318.

Bhowmik, R 2011, 'Detecting Auto Insurance Fraud by Data Mining Techniques', *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 2(4)

Biru, A 2009, 'Applicability of Data Mining Techniques to Support Voluntary Counseling and Testing (VCT) for HIV: The Case of Center for Disease Control and Prevention (CDC)*'*, Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Bounsaythip, C and Rinta-Runsala, E 2001, 'Overview of Data Mining for Customer Behaviour Modelling', *VTT Information Technology*, Vol. 18, pp. 1-53.

Chapman, P Clinton, J Kerber, R Khabaza, T Reinartz, T Shearer, C and Wirth, R 2000, 'CRISP-DM 1.0: Step-by-step Data Mining Guide'*,* SPSS Inc., USA.

Cios, K and Kurgan, L 2005, 'Trends in Data Mining and Knowledge Discovery', In Pal, N.R., and Jain L.C. (Eds.), *Advanced Techniques in Knowledge Discovery and Data Mining*, Springer Verlag, London, pp. 1–26.

Cios, K Witold, P Roman, S and Kurgan A. 2007, *Data Mining: A Knowledge Discovery Approach,* Springer, New York: USA.

Denekew, A 2003, 'The Application of Data Mining to Support Customer Relationship Management at Ethiopian Airlines**'**, Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Derrig, A 2002, 'Insurance Fraud', *Journal of Risk and Insurance*, Vol. 69(3), pp. 271-287.

Deshpande, SP and Thakare, VM 2010, 'Data Mining System and Applications: A Review', *International Journal of Distributed and Parallel systems (IJDPS),* Vol.1 (1), pp. 32-44.

DíArcy*,* P Derrig, A and Weisberg, I 2006, *'*The Economics of Insurance Fraud Investigation: Evidence of a Nash Equilibrium'*, Journal of Risk and Insurance*, Vol. 69(3).

Dickson, G and Stein, WM 1999, *Risk and Insurance,* CII Publishing Division, London.

Dockrill, M 2001, *Underwriting Management*, CII publishing Division, London.

Farn, C and Huang, T 2009, 'A Study on Industrial Customers Loyalty to Application Service Providers: The case of logistics information services', *International Journal of Computers*, Vol. 3, pp. 151-160.

Fayyad, U Piatetsky-Shapiro, G and Smyth, P 1996, 'From Data Mining to Knowledge Discovery in Databases*', AI Magazine,* July 27, pp. 37-54.

Fayyad, U Piatetsky-Shapiro, G and Smyth, P 1996, 'Knowledge Discovery and Data Mining: Towards a Unifying Framework', *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD96)*, Portland, OR. AAAI Press.

Fekadu, M 2004, 'Application of Data Mining Techniques to Support Customer Relationship Management at Ethiopian Telecommunications Corporation', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Gobena, M 2000, 'Flight Revenue Information Support System for Ethiopian Airlines', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Guha, S Rastogi, R and Shim, K 1998, 'CURE: An Efficient Clustering Algorithm for Large Databases', *Proceedings of the ACM SIGMOD Conference*, Seattle, USA.

Guo, L 2003, 'Applying Data Mining Techniques in Property & Casualty Insurance,' USA. http://www.casact.org/pubs/forum/03wforum/03wf001.pdf
Access date: December 4, 2010.

Hajek, M 2005, 'Neural Networks',
http://www.cs.ukzn.ac.za/notes/NeuralNetworks2005.pdf
Access Date: December 25, 2010.

Hajizadeh, E Ardakani, D and Shahrabi, J 2010, 'Application of Data Mining Techniques in Stock Markets: A Survey', *Journal of Economics and International Finance*, Vol. 2(7), pp. 109-118.

Halkidi, M and Vazirgiannis, M 2001, 'Evaluating the Validity of Clustering Results Based on Density Criteria and Multi-representatives,' Greece.

Han, J and Kamber, M 2006, *Data Mining: Concepts and Techniques*, 2nd edn, Morgan kufman Publishers, San Francisco.

Han, J Pei, J Yin, Y and Mao, R 2004, 'Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach', *Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, Netherland, pp. 53–87.

Hand, D Mannila, H and Smyth, P 2001, *Principles of Data Mining*, A Bradford Book, The MIT Press Cambridge, Massachusetts London, England.

Helen, T 2003, 'Application of Data Mining Technology to Identify Significant Patterns in Census or Survey Data: The Case of 2001 Child Labor Survey in Ethiopia', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Henock, W 2002, '*Application of Data Mining Techniques to Support Customer Relationship Management at Ethiopian Airlines* ", Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

IBM Corporation 2010, *Using Data Mining to Detect Insurance Fraud*, IBM Corporation, Chicago.

Ibrahim, S 1999, 'Data Mining of Machine Learning Performance Data', Master of Applied Science (Information Technology) Thesis, RMIT University: Melbourne, Australia.

Jinhong, L Bingru, Y and Wei, S 2009, 'A New Data Mining Process Model for Aluminum Electrolysis', *Proceedings of the International Symposium on Intelligent Information Systems and Applications (IISA'09) Qingdao, P. R. China Oct. 28-30*, Academy Publisher, pp. 193-195.

Jordan, I and Bishop, M 1996, 'Neural Networks', *ACM Computing Surveys*, CRC Press, Vol. 28(1), pp. 73-75.

Kim, J Suh, E and Hwang, H 2003, 'A Model for Evaluating the Effectiveness of CRM Using the Balanced Scorecard', *Journal of Interactive Marketing,* Vol. 17, pp. 5-19.

Koh, C and Gervais, G 2010, 'Fraud Detection Using Data Mining Techniques: Applications in the Motor Insurance Industry', Singapore.

Kotsiantis, S and Kanellopoulos, D 2006, 'Association Rules Mining: A Recent Overview', *GESTS International Transactions on Computer Science and Engineering*, Vol.32 (1), pp. 71-82.

Kumneger, F 2006, 'Application of Data Mining Techniques to Support Customer Relationship Management for Ethiopian Shipping Lines (ESL)', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Kurgan, A and Musilek, P 2006, 'A Survey of Knowledge Discovery and Data Mining Process Models', *The Knowledge Engineering Review*, Vol. 21(1), Cambridge University Press, pp. 1-24.

Larose, T 2006, *Data Mining Methods and Models*, John Wiley & Sons Inc. Publisher, Hoboken: New Jersey.

Lavesson, N 2003, 'Evaluation of Classifier Performance and the Impact of Learning Algorithm Parameters', Master Thesis in Software Engineering, Department of Software Engineering and Computer Science, Blekinge Institute of Technology, Sweden.

Leul, W 2003, 'The Application of Data Mining in Crime Prevention: The Case of Oromia Police Commission', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Meera, G and Srivatsa, SK 2010, 'Adaptive Machine Learning Algorithm (AMLA) Using J48 Classifier for an NIDS Environment', *Advances in Computational Sciences and Technology*, Research India Publications, Vol. 3(3) pp. 291–304.

Melkamu, G 2009, 'Applicability of Data Mining Techniques to Customer Relationship Management: The Case of Ethiopian Telecommunications Corporation (ETC) Code Division Multiple Access (CDMA Telephone Service)', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Mesfin, F 2005, 'Predictive Data Mining Technique in Insurance: The Case of Ethiopian

Insurance Corporation', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Muluken, Y 2010, 'Africa Insurance Collects 19 million Birr Profit', *Capital,* Sunday December 19, p. 6.

Nimmagadda, R Kanakamedala, P And Yaramala, B 2011, 'Implementation of Clustering Through Machine Learning Tool', *IJCSI International Journal of Computer Science Issues*, Vol. 8(1), pp. 395-401.

Parvatiyar, A and Sheth, N 2001, 'Customer Relationship Management: Emerging Practice, Process, and Discipline', *Journal of Economic and Social Research,* Vol. 3, pp. 1-34.

Pham, DT Dimov, SS and Nguyen, CD 2005, 'Selection of K in K-means Clustering', *Journal of Mechanical Engineering Science,* Vol. 219, pp. 103-119.

Phyu, N 2009, 'Survey of Classification Techniques in Data Mining', *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong.

Qiu, M Davis, S and Ikem, F 2004, 'Evaluation of Clustering Techniques in Data Mining Tools', *Issues in Information Systems*, Vol. 5 (1), pp. 254-260.

Rao, R 2003, 'Data Mining and Clustering Techniques'*, DRTC Workshop on Semantic Web*, DRTC: Bangalore.

Rejesus, M Little, B and Lovell, C 2004, 'Using Data Mining to Detect Crop Insurance Fraud: Is There a Role for Social Scientists?' *Journal of Financial Crime*, Vol. 12 (1), pp. 24-32.

Samson, T 2009, 'Possible Application of Data Mining Technology in Supporting Term Loan Risk Assessment: The Case of United Bank S.C.', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

SAS Institute Inc. 1999, *Data Mining in the Insurance Industry: Solving Business Problems Using SAS® Enterprise Miner™ Software*, SAS Institute Inc.

SAS Institute Inc. 2001, *Finding the Solution to Data Mining: Exploring the Features and Components of Enterprise Miner ™, Release 4.1*, SAS Institute Inc.

SAS Institute Inc. 2008, *Combating Insurance Claims Fraud: How to recognize and reduce opportunistic and organized claims fraud*, SAS Institute Inc. http://www.sas.com/resources/whitepaper/wp_4422.pdf
Access date: December 15, 2010.

Schiller, J 2003, 'The Impact of Insurance Fraud Detection Systems', *Journal of Risk and Insurance*, Vol. 73 (3), pp. 421-438.

Seifert, W 2004, *Data Mining: An Overview*, CRS Report for Congress, Congressional Research Service: The Library of Congress.

Shegaw, A 2002, 'Application of Data Mining Technology to Predict Child Mortality Patterns: The Case of Butajira Rural Health Project (BRHP)', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Singh, Y and Chauhan, S 2005, 'Neural Networks in Data Mining', *Journal of Theoretical and Applied Information Technology*, pp. 37-42.

Tan, P Steinbach, M and Kumar, V 2009, *Introduction to Data Mining*, 3rd edn, Pearson Education, New Delhi.

Tesfaye, H 2002, 'Predictive Modeling Using Data Mining Techniques in Support of Insurance Risk Assessment: The Case of Nyala Insurance Corporation', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Tilahun, M 2009, 'Possible Application of Data Mining Techniques to Target Potential VISA Card Users in Direct Marketing: The Case of Dashen Bank S.C.', Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Trnka, A 2010, 'Six Sigma Methodology with Fraud Detection', *Advances in Data Networks, Communications, Computers*, Trnava, Slovakia, pp. 162-165.

Two Crows Corporation 1999, *Introduction to Data Mining and Knowledge Discovery*, 3rd edn, Two Crows Corporation, Potomac: U.S.A.

Two Crows Corporation 2005, *Introduction to Data Mining and Knowledge Discovery*,

3rd edn, Two Crows Corporation, Potomac: U.S.A.

Witten, I and Frank, E 2000, *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 2nd edn, Morgan Kaufmann publishers, San Francisco.

Woodfield, J 2005, 'Predicting Workers' Compensation Insurance Fraud Using SAS® Enterprise Miner™ 5.1 and SAS® Text Miner', *Data Mining and Predictive Modeling*, SAS Institute Inc, USA.

Wu, X Kumar, V Quinlan, R Ghosh, J Yang, Q Motoda, H McLachlan, J Ng, A Liu, B Yu, S Zhou, Z Steinbach, M Hand J and Steinberg, D 2007, '*Top 10 Algorithms in Data Mining*', Springer-Verlag, London.

Yao, X 1999, 'Evolving Artificial Neural Networks', *Proceedings of the IEEE*, Vol. 87(9), pp. 1423-1447.

APPENDICES

Appendix 1: Initial list of original attributes with their description

No.	Attribute Name	Data Type	Description
1.	CLM_POL_NO	Varchar2	The policy number of the claim
2.	CLM_LOSS_DT	Date	The date that the accident has occurred
3.	CLM_INTM_DT	Date	The date the insured reported to the company about the accident
4.	CLM_LOSS_TYPE	Varchar2	The type of loss that has occurred
5.	CLM_LOSS_DED_YN	Varchar2	Existence of deductibles
6.	CLM_DIVN_CODE	Number	Branch service unit. The data type this attribute was Varchar2 originally.
7.	CLM_UW_YEAR	Number	Underwriting year
8.	CLM_SC_CODE	Varchar2	Subclass code of the motor vehicle
9.	CLM_CLOSE_DT	Date	The data that the claim has been closed
10.	CLM_CLAIM_DESC	Varchar2	Description of the claim about the accident
11.	CLM_CLOSE_REASON	Varchar2	The reason that the claim has been closed
12.	CE_LC_AMT	Number	The amount of money that is estimated to be paid for the claim in local currency
13.	CE_CLM_TYPE	Number	The type of claim
14.	CE_SRC_BUS_CODE	Varchar2	Source of business (based on the purpose of use of vehicle)
15.	CE_EST_NO	Varchar2	Claim estimate number
16.	CE_LOSS_TYPE	Varchar2	The type of loss
17.	VEH_YEAR	Number	Vehicle year of manufacture
18.	VEH_TYPE	Varchar2	The type of vehicle
19.	VEH_AGE	Number	The original age of the vehicle

20.	POL_UW_YEAR	Number	Policy underwriting year
21.	POL_ISSUE_DT	Date	The date that the policy is issued
22.	POL_END_EFF_FM_DT	Date	Policy effective end date
23.	POL_STATUS	Varchar2	Status of the policy
24.	CD_RISK_TYPE	Varchar2	The type of vehicle use. It is to mean that whether the vehicle is used for private or commercial purpose
25.	CD_LOCATION_CODE	Number	The location of the accident
26.	CP_LC_PAID_AMT	Number	The amount of money paid for the claimant in local currency

Appendix 2: Sample values of the final selected attribute

CLM_REP_DT_G	BRANCH	CLM_ACCID_TYPE	VEH_TYPE	VEH_AGE	VEH_USE	INSU_YR_S	POL_STATUS	SRC_BUS_CODE	CLM_REC_AMT
3	ARADA	COLLISSION-TP	VT160	8	PRIVATE	5	E	SRB4	5920
7	HEADO	COLLISSION-TP	VT220	4	COMMERCIAL	0	A	SRB6	0
3	ARADA	COLLISSION-TP	VT160	5	PRIVATE	1	R	SRB4	-1900
14	FILW	'WINDSHIELD BREAKAGE'	VT110	9	PRIVATE	1	E	SRB5	0
3	ARADA	COLLISSION-TP	VT160	3	PRIVATE	3	R	SRB4	-1900
16	FILW	COLLISSION-TP	VT110	10	COMMERCIAL	3	R	SRB6	0
10	KIRKOS	COLLISSION-TP	VT110	7	COMMERCIAL	0	E	SRB4	3250
16	FILW	COLLISSION-TP	VT110	8	COMMERCIAL	0	A	SRB6	0
10	KIRKOS	COLLISSION-TP	VT110	8	COMMERCIAL	2	R	SRB4	0
	FIL	OVERTU	VT25		COMMERC				

					IAL				
4	FIL W	OVERTU RNING	VT25 0	9	COM MERC IAL	2	R	SRB4	4907
3	ARA DA	COLLISS ION-TP	VT16 0	8	PRIV ATE	3	R	SRB4	0

Appendix 3: Confusion matrix results of the classification techniques

A) Confusion matrix result of J48 algorithm with default values

```
=== Confusion Matrix ===

    a     b   <-- classified as
10002     0 |     a = cluster0
    6  7802 |     b = cluster1
```

B) Confusion matrix result of J48 algorithm with minNumObj=20

```
=== Confusion Matrix ===

    a     b   <-- classified as
 9979    23 |     a = cluster0
   47  7761 |     b = cluster1
```

C) Confusion matrix result of J48 algorithm with 70% percentage split

```
=== Confusion Matrix ===

    a     b   <-- classified as
 2969     4 |     a = cluster0
    0  2370 |     b = cluster1
```

D) Confusion matrix result of Naïve Bayes algorithm with 10-fold cross validation

```
=== Confusion Matrix ===

    a     b   <-- classified as
 9620   382 |     a = cluster0
 1099  6709 |     b = cluster1
```

Appendix 4: A sample decision tree generated from the J48 decision tree learner with the 10-fold cross validation and the default parameters.

```
J48 pruned tree
------------------

CLM_REC_AMT <= 10000
|   CLM_REP_DT_G <= 4
|   |   POL_STATUS = R
|   |   |   SRC_BUS_CODE = SRB6
|   |   |   |   CLM_REC_AMT <= 3000: cluster0 (855.0)
|   |   |   |   CLM_REC_AMT > 3000
|   |   |   |   |   CLM_REC_AMT <= 4550: cluster1 (13.0)
|   |   |   |   |   CLM_REC_AMT > 4550
|   |   |   |   |   |   VEH_USE = COMMERCIAL: cluster0 (59.0)
|   |   |   |   |   |   VEH_USE = PRIVATE: cluster1 (4.0)
|   |   |   SRC_BUS_CODE = SRB4: cluster0 (2210.0)
|   |   |   SRC_BUS_CODE = SRB5: cluster0 (0.0)
|   |   POL_STATUS = C
|   |   |   CLM_ACCID_TYPE = OVERTURNING: cluster0 (202.0)
|   |   |   CLM_ACCID_TYPE = COLLISSION-TP
|   |   |   |   SRC_BUS_CODE = SRB6: cluster1 (2.0)
|   |   |   |   SRC_BUS_CODE = SRB4: cluster0 (76.0)
|   |   |   |   SRC_BUS_CODE = SRB5: cluster0 (0.0)
|   |   |   CLM_ACCID_TYPE = VHP: cluster0 (9.0)
|   |   |   CLM_ACCID_TYPE = COLLISSION: cluster0 (35.0)
|   |   |   CLM_ACCID_TYPE = WINDSHIELD BREAKAGE
|   |   |   |   SRC_BUS_CODE = SRB6: cluster1 (9.0)
|   |   |   |   SRC_BUS_CODE = SRB4: cluster0 (12.0)
|   |   |   |   SRC_BUS_CODE = SRB5: cluster0 (0.0)
|   |   POL_STATUS = E
|   |   |   SRC_BUS_CODE = SRB6: cluster1 (339.0)
|   |   |   SRC_BUS_CODE = SRB4
|   |   |   |   CLM_ACCID_TYPE = OVERTURNING: cluster0 (497.0)
|   |   |   |   CLM_ACCID_TYPE = COLLISSION-TP:cluster1(123.0)
|   |   |   |   CLM_ACCID_TYPE = VHP: cluster0 (65.0)
|   |   |   |   CLM_ACCID_TYPE = COLLISSION: cluster0 (178.0)
|   |   |   |   CLM_ACCID_TYPE=WINDSHIELD BREAKAGE: cluster0
(41.0)
|   |   |   SRC_BUS_CODE = SRB5: cluster0 (0.0)
```

DECLARATION

I declare that the thesis is my original work and has not been presented for a degree in any other university.

_____

Date

This thesis has been submitted for examination with my approval as university advisor.

_____