ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

CONSTRUCTING PREDICTIVE MODEL USING DATA MINING TECHNIQUES IN SUPPORT OF MOTOR INSURANCE POLICY RISK ASSESSMENT THE CASE OF ETHIOPIAN INSURANCE CORPORATION (EIC)

YIHENEW FEKADU

JUNE 2015

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

CONSTRUCTING PREDICTIVE MODEL USING DATA MINING TECHNIQUES IN SUPPORT OF MOTOR INSURANCE POLICY RISK ASSESSMENT THE CASE OF ETHIOPIAN INSURANCE CORPORATION (EIC)

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

Yihenew Fekadu

June 2015

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

CONSTRUCTING PREDICTIVE MODEL USING DATA MINING TECHNIQUES IN SUPPORT OF MOTOR INSURANCE POLICY RISK ASSESSMENT: THE CASE OF ETHIOPIAN INSURANCE CORPORATION (EIC)

Yihenew Fekadu

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
Dereje Teferi (PhD)	Advisor	_____	_____
Million Meshesa (PhD)	Examiner	_____	_____
Ato Ermias Abebe	Examiner	_____	_____

Dedicated to my beloved Mother,

Wude Zeleke

ACKNOWLEGMENT

Foremost, my deepest gratitude is to the almighty God for giving me the ability to face challenges and complete this research work.

Next to this, I would like to express my sincerest gratitude to my advisor Dr. Dereje Teferi for his critical reading, guidance and valuable comments.

I would like to thank EIC Information technology department employees particularly w/o Mirchaye Mulugeta for provide me all the necessary documents from the INSIS database. I would also like to express my sincere gratitude to all the five main branches' mangers, claim division managers and archivist for their cooperative approach and critical evaluations of the discovered knowledge.

A greater many people have contributed on the manual data collection process of this thesis. I owe my gratitude to Tesfaye Ashenafi and Getahun Bulte who have collected the 5200 records as per the preset schedule.

My special thanks also goes to my friend and classmate Rahel Wonde, for her friendly and cooperative approach.

Finally, I would like to thank my family for supporting spiritually throughout my life.

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ARFF	Attribute Relation File Format
CRISP-DM	Cross Industry Standard Process For Data Mining
CSV	Common Separated Values
DM	Data Mining
EIC	Ethiopian Insurance Corporation
INSIS	Integrated System for Insurance Solutions
KDD	Knowledge Discovery in Database
MIRAS	Motor Insurance Risk assessment System
SQL	Structural query language
WEKA	Waikato environment for knowledge Analysis

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In recent years data mining has attracted a great deal of attention in information industry due to the wide availability of huge amounts of data and the need to change them into useful information and knowledge for broad applications including market analysis, business management, decision support, risk assessment and fraud detection. This research applies data mining techniques in support of motor insurance risk assessment at the time of underwriting.

The research is implemented using the six-step suggested by Cisos et. al.(2000) DM process model. The data collection process has been done in two phases. Records about vehicles are collected from INSIS database where as records about drivers' are collected manually. The collected dataset is preprocessed using weka DM tools and Microsoft Excel in order to select attributes, derive new attributes, handle missing values and remove outliers.

In this study an attempt was made to apply data mining clustering and classification algorithms. K-means clustering algorithm is implemented to come up with the natural group of the claim records as low risk, medium risk, high risk and very high risk. The researcher implemented two classification algorithms, J48 decision tree classification algorithms and multiperceptron ANN. Using j48 decision tree classification algorithms different experimentations are conducted. The first experimentation with default parameter values and 10 fold cross validation test options has registered 94.63% accuracy. ANN experimentations have also been conducted. Accordingly the experimentation with default parameter values with 10-fold cross-validation test option registered 99.58% accuracy. The study also

registered an accuracy of 93.74% with percentage split test option by splitting the dataset into 80% to training set and 20% to test set.

The result of this study indicates that applying data mining to classify insurance customers to predict the risk level is very promising. The above prediction accuracy also indicates that data mining is a powerful tool to measure the uncertainty of losses in EIC. Hence, the research identified future research direction in order to implement applicable system in risk assessment process.

Chapter One

Introduction

1.1 Background

With the increasing power of computer technology, companies and institutions can nowadays store large amounts of data at reduced cost. The amount of available data is increasing exponentially and cheap disk storage makes it easy to store data that previously was thrown away. There is a huge amount of information locked up in databases that is potentially important but has not yet been explored. The growing size and complexity of the databases makes it hard to analyze the data manually, so it is important to have automated systems to support the process. Hence there is a great need of computational tools able to treat these large amounts of data and extract valuable information (Pozzolo, 2011). In addition to the above data analysis challenges, insurance companies' face different kinds of risk which are transferred from the insurer to the insurance companies.

According to Casual Actuarial Society (2003), there are four types of risk. These are Hazard risk, operational risk, financial risk and strategic risk. Since the establishment of casual actuarial society in the year 1914 insurance companies try to assess the above type of risk using actuarial science. Actuarial science is the discipline that applies mathematical and statistical methods to assess risk in insurance, finance and other industries and professions. Actuaries are professionals who are qualified in this field through education and experience.

DM methodology can often improve existing actuarial models by finding additional important variables, identifying interactions and detecting nonlinear relationships. DM can help insurance firms make crucial business decisions and turn the new found knowledge into actionable results in business practice such as product development, marketing, claim distribution analysis, asset liability managements and solvency analysis. To be more specific, DM can perform the following tasks (Guo, 2003).

- Identify risk factors that predict profits claims and losses
- Customer level analysis
- Developing new product line
- Estimation outstanding claim provision
- Structure reinsurance

An important general difference in the focus between existing actuarial techniques and DM is that DM is more oriented towards application than towards describing the basic nature of the underlying phenomena. For example drivers" age and auto type are not the main goal of DM. Instead, the focus is on producing a solution that can improve that prediction for future premiums. DM is very effective in determining how the premium related to multidimensional risk factors such as drivers' age and type of automobile (Guo,2003)

In this study, DM techniques were employed for processing large amounts of data that are already present in database. DM is used to automatically extract important patterns and trends from databases seeking regularities or patterns that can reveal the structure of the data and answer business problems (Kaufmann, 2005).

The study is conducted on motor insurance data obtained from EIC. The corporation was established in 1976 by proclamation No.68/1975. The Corporation came into existence by taking over all the assets and liabilities of the thirteen nationalized private insurance companies, with Birr 11 million (USD 1.29 million) paid up capital having the following objectives: Engage in all classes of insurance business in ethiopia and ensure the insurance services reach the broad mass of the people (www.eic.com).

Currently the corporation has 176,000 customers which are distributed in around 60 different classes of businesses. The researcher selected motor insurance due to the fact that motor class is characterized by relatively high claim frequency and cost (www.eic.com). According to domain experts at underwriting department of East branch, there are no mechanisms that can classify a customer as very high risk, high risk, medium risk and low risk. But currently in United States of America some insurance companies use sophisticated software to help them set their rates. The industry uses the process, called "price optimization," which is simply a way to be more efficient. Consumer advocates believe it's being used to get around risk-based pricing. CFA's director of insurance says price optimization is a data mining tool that lets insurance companies figure out which groups of customers are more likely to accept a price increase and which are more likely to shop around for a new policy.

This study is conducted to identify features and classify new policy based on the level of risk at EIC motor class of business. The hidden knowledge discovered by data mining helps the corporation to adjust motor class of business rate chart based on the risk level.

1.2 Statements of the problem

Road traffic accident is the cause of significant loss of human and economic resources worldwide. About 1.2 million people die and 50 million are injured annually worldwide (Tibebe and Hill 2011). More than 85% of these casualties occur in low and middle in-come countries. Road traffic accident in Ethiopia is a cause of significant loss of human and economic resources (Road Traffic, 2009). In Ethiopian 2007/8 fiscal year, police reported 15,086 accidents which caused the loss of 2,161 lives and over ETB 82 million (equivalent to US$4.1 million) cost estimate of property damage (Road Traffic, 2009). It would be impossible to attach a value to each case of human sacrifice and suffering, add up the values and produce a figure that captures the national social cost of road crashes and injuries. However, the economic costs of road traffic accidents are, evidently, a heavy burden for the national economy.

The above statistical data indicate that insurance companies have incurred losses due to the number of claims reported by the insurers. According to EIC 2014 report around 12000 motor claims were reported to which around 726,700,000 Birr is paid (www.eic.com). There are facts that should be taken in to account at the time of underwriting. As underwriter domain experts noted, there are very high risk, high, low and medium risk vehicles. On the other hand according to Addis Ababa police commission about 90 percent of the accidents occurred due to human error. Of these accidents drivers are indicated as responsible causes in about 89 percent (Road Traffic, 2009).

Today the use of data mining technology in support of business decision-making is growing fast. There are attempts made to apply data mining technology in solving business problems in Ethiopian context, including Airlines, Banking and telecommunication. There are also studies

conducted to investigate the application of data mining in the insurance business. Tesfaye (2002) used the data mining technique for motor insurance risk assessment in Nyala insurance company. He developed a predictive model using neural network classification technique. Mesfin (2005), attempts to investigate the possible application of predictive data mining techniques in the renewal process of personal accident policies of EIC. The other research, which is undertaken by Tariku (2008) also, attempts to investigate motor insurance fraud at Africa insurance share company. However, no one investigated the application of data mining to identify features on both drivers and vehicles side and classify possible features to determine the level of risk on motor insurance class of business. It is, therefore, with this understanding that this study predicts features that contribute for the risk incurred in motor insurance policy of EIC. Hence, this study attempts to answer the following research questions.

- What are the tasks to apply for preparing quality dataset for experimentation?
- Which classification algorithms can be more suitable for the purpose of identifying/predicting motor insurance policy?
- What is the possible segment that motor claims can be clustered according to their natural groupings? And which DM algorithm is more appropriate for doing this task?
- What attributes should be taken into account at the time of underwriting?

1.3 Objectives of the study

1.3.1 General objective

The general objective of the study is to construct a predictive model that enable motor insurance policy risk assessment.

.

1.3.2 Specific Objectives

To achieve the general objective of the study, the following specific objectives are identified.

- To review literatures on data mining techniques especially on classification techniques and their application on insurance business.
- To undertake business understanding and collect the required data from Ethiopian Insurance Corporation.
- To prepare quality dataset by applying the various data preprocessing task.
- To Explore and identify best classification algorithm that are suitable for insurance data analysis.
- To construct a working model that identify whether a policy is very high risk, high risk, medium risk and low risk.
- To evaluate the performance of the predictive model on test dataset.

1.4 Scope and limitations of the study

The insurance industry deals with risks of various classes of business such as marine, motor, fire, and engineering. The scope of this research is to examine the potential features available only on motor insurance class of business. Motor insurance class of business is one of the most risky class of business. The data mining goal of this project is to classify the customers as very high risk, high risk, medium risk and low risk. In order to accomplish this task the researcher use a classification and clustering techniques.

EIC has around 67 branches throughout the country. The data used in this study is collected only from the five Addis Ababa branches of the corporation from June 2012 to December 2013. These five branches are selected because of their underwriting and claim records in the corporation.

1.5 Significance of the study

The output of this study provides information to EIC whether those features are very high risk, high risk, medium risk or low risk. Those attribute that belong to high risk take the lion share for the decrease in profitability of motor class of business. This gives the corporation a chance to filter high risk attributes on motor insurance which enables the corporation to minimize the loss incurred by motor insurance.

As improving customer service is one of the major priorities of EIC. The result of this study will help to know those drivers that have low risk, appreciating low risk drivers help the company to minimize the losses of EIC. The study gives an opportunity to adjust the underwriting rate chart by the classification made on this study. This study also used as an input for other related research works.

1.6 Thesis Organization

This research is organized into six chapters. The first chapter discusses background of the study, statement of the problem, objective of the study research methodology, scope and limitation, and significance of the study. The second chapter reviews data mining and knowledge discovery. The chapter briefly discusses DM process model and application of DM in insurance industry. The third chapter discusses about DM technology, methodology and algorithm applied for risk assessment. It explains k-means clustering algorithm, j48 decision tree classification technique and Multilayerperceptron neural network algorithm. The Fourth chapter explains Business and data understanding and prepared the data for experiment. Activities preformed in this chapter include business understanding, data understanding and preparation of the data. The fifth chapter discuss about the experimentation of this study and the result of k-means clustering, J48 decision tree algorithm and multilayerperceptron neural network. The last chapter provides conclusion and recommendation for future work.

CHAPTER TWO

DATA MINING AND KNOWLEDGE DISCOVERY

We are in an age often referred to as the information age. In this information age, because People believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, and other related storage media, they have been collecting tremendous amounts of information from different sources. Initially, with the advent of computers and means for mass digital storage, people started collecting and storing all sorts of data, counting on the power of computers to help sort through this bulk of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and database management systems. The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of knowledge form the available raw data. Therefore, in order to fulfill the above needs, applying data mining technologies in the available raw data could be crucial (Zaïane, 1999).

Different scholars define data mining in different ways. According to Fayyad et al (1996) data mining is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures, from large amounts of data stored in databases, data warehouses, or other

information repositories (Fayyad, Shapiro, Smyth 1996). Others define as a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions (Kolyshkina and Brookes 2002).

In recent years data mining has attracted a great deal of attention in information industry, due to the wide availability and the need to change these huge amounts of data into useful information and knowledge for broad applications including market analysis, business management, decision support, risk assessment and fraud detection (Han and Kamber 2006).

2.1 Data mining Tasks

The tasks of data mining can be modeled as either Predictive or Descriptive (Siraj and Abdoulha, 2011). Predictive model data mining tasks include classification, prediction, regression and time series analysis. The descriptive task encompasses methods such as Clustering, Summarizations, Association Rules, and Sequence analysis.

2.1.1 Descriptive Modeling

Descriptive models are unsupervised learning functions. These functions do not predict a target value, but focus more on the intrinsic structure, relations, interconnectedness, etc. of the data.

Descriptive modeling is a mathematical process that describes real-world events and the relationships between factors responsible for them. The process is used by consumer-driven organizations to help them target their marketing and advertising efforts (Siraj and Abdoulha, 2011).

In descriptive modeling, customer groups are clustered according to demographics, purchasing behavior, expressed interests and other descriptive factors. Statistics can identify where the customer groups share similarities and where they differ.

Statistical Modeling in the direct marketing industry has two primary goals: describing current customers and behaviors (Descriptive Modeling) and predicting future behaviors or events (Predictive Modeling).

The distinction between these differing statistical methods is often confused in the market place. Since the goals and statistical conclusions drawn for each method are different, it is essential to the limitations and value-add of each method.

An important distinction to be made is that a Predictive Model will always include the descriptive component of any modeling initiative, as everything within a Predictive Model is based on the descriptive foundation. However, a Descriptive Model will not include the predictive statistics and multivariate methods (algorithm) used to score customers or a list of prospects.

A helpful analogy for descriptive modeling is 'telling the story'. We can think of Descriptive Modeling as the method an author uses to describe the characters in a novel. A good author (analyst) will help you to connect with the characters by describing complex and rich relationships (Siraj & Abdoulha, 2011).

For instance, in marketing this is exactly how it works in Descriptive Modeling. Primary customer groups are aggregated, and the customer file is enhanced with demographics, purchasing behaviors, interests and other important indicators that will describe them (story elements).

Statistics are used to identify where the customer groups share similarities, and where they differ. Special attention is given to the best or active customer, as this is the group that offers the greatest return on investment –whether it is a cross-sell opportunity or acquisition of look-alikes (Siraj & Abdoulha, 2011).

2.1.1.1 Approaches of Clustering

In general, the major clustering methods can be classified into the following categories as can be seen the below figure (Han and Kamber, 2006).

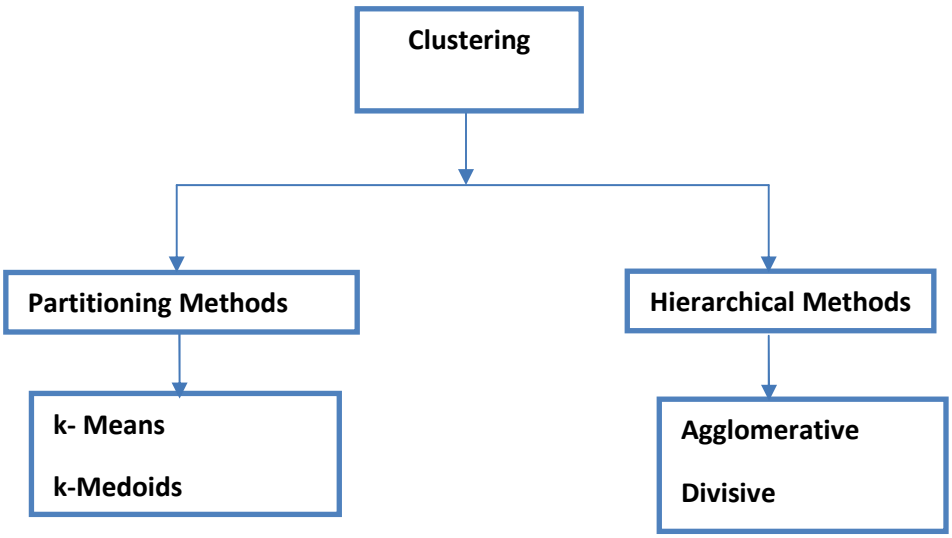


Figure 2.1 Major Approaches of Clustering

Partitioning methods: Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k<=n. That is, it classifies the data into k groups, which together satisfy the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group (Han and Kamber 2006).

Given k, the number of partitions to construct, a partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the same cluster are “close” or related to each other, whereas objects of different clusters are “far apart” or very different. There are various kinds of other criteria for judging the quality of partitions. To achieve global optimality in partitioning-based clustering would require the exhaustive enumeration of all of the

possible partitions. Instead, most applications adopt one of a few popular heuristic methods, such as (1) the k-means algorithm, where each cluster is represented by the mean value of the objects in the cluster, and (2) the k-medoids algorithm, where each cluster is represented by one of the objects located near the center of the cluster (Han and Kamber 2006). These heuristic clustering methods work well for finding spherical-shaped clusters in small to medium-sized databases. To find clusters with complex shapes and for clustering very large data sets, partitioning-based methods need to be extended.

Hierarchical methods: A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds. The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds (Han and Kamber 2006).

Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone. This rigidity is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices .However, such techniques cannot correct erroneous decisions. There are two approaches to improving the quality of hierarchical clustering (Han and Kamber 2006): (1) perform careful analysis of object "linkages" at each hierarchical partitioning, such as in Chameleon, or (2) integrate hierarchical agglomeration and other approaches by first using a hierarchical agglomerative algorithm to group objects into micro clusters, and

then performing macro clustering on the micro clusters using another clustering method such as iterative relocation (Han and Kamber 2006).

Applications of Clustering

Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing (Han and Kamber 2006). In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value, and geographic location, as well as the identification of groups of automobile insurance policy holders with a high average claim cost. It can also be used to help classify documents on the Web for information discovery.

Clustering can also be used for outlier detection, where outliers (values that are "far away" from any cluster) may be more interesting than common cases. Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce. For example, exceptional cases in credit card transactions, such as very expensive and frequent purchases, may be of interest as possible fraudulent activity. As a data mining function, cluster analysis can be used as a stand-alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Alternatively, it may serve as a preprocessing step for other algorithms, such as characterization, attribute subset selection, and classification, which would then operate on the detected clusters and the selected attributes or features (Han and Kamber 2006).

2.1.1.2 Association Rule Discovery

Frequent pattern discovery attempts to discover hidden linkage between data items. It produces dependency rules which will predict occurrence of an item based on occurrences of other items. An itemset is a set of one or more items such as X = $\{x_1, \ldots, x_k\}$ such as K-itemsets. Frequent patterns are patterns (such as itemsets, subsequences, or substructures) that appear in a data set frequently. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as sub graphs, sub trees, or sub lattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Finding such frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data classification, clustering, and other data mining tasks as well. Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining research (Han and Kamber 2006).

Interesting rules are different from prevalent rules since the latter are already known by analysts. Hence, interesting rules are those that deviate from prior expectation and create a surprising phenomenon. Rules become surprising phenomena when they do not match prior expectation and cannot be trivially derived from simpler rules.

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. Support, s, is the fraction of transactions that contains X i.e., the probability that a transaction contains X. Support of X and Y greater than user defined threshold s; that is, support probability of s that a transaction contains $X \cup Y$. An itemset

X is *frequent* if X's support is no less than a minimum support threshold. Confidence is the probability of finding Y in a transaction with all X_1, X2,…,X_n. In other terms, confidence, *c,* conditional probability that a transaction having X also contains *Y; i.e.* conditional probability (confidence) of Y given X >the defined threshold c. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Additional analysis can be performed to uncover interesting statistical correlations between associated items (Han and Kamber 2006).

In general, association rule mining can be viewed as a two-step process (Han and Kamber 2006).

- The first step is finding frequent patterns from large item sets that satisfy the minimum support threshold value. By definition, each of these item sets will occur at least as frequently as a predetermined minimum support count.

- The second step is generating strong association rules from the frequent item sets. Association rules are defined as statements of the form $\{X_1,X_2,\ldots,X_n\}$ -> Y, which means that Y is likely to be present in the transaction if $X_1,X_2,\ldots,X_n$ are all in the transaction.

While the second step is straightforward, the first step needs more attention. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets (item combinations). The set of possible item sets is the power set over and has huge size (excluding the empty set which is not a valid item set). Although the size of the power set grows exponentially as the number of items in, efficient search is possible using the downward closure property of support. It guarantees that for a frequent item set, all its subsets are also frequent and those for non-frequent item set, all its supersets must also be non-frequent. This property is known as apriori property and applied in apriori algorithm of candidate item set generation (Han and Kamber 2006).

Applications of Association Rule Discovery

Frequent item set mining helps to discover associations and correlations among items in large transactional or relational data sets. It detects hidden linkages which seem to be totally unrelated data. Those linkages can provide rules which help actions to be taken. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes, such as catalog design, cross-marketing, and customer shopping behavior analysis and shelf management (such as of Supermarket, Pharmacy and Book shop). A typical example of frequent item set mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets". The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. Such kind of information can help to generate higher volumes of sales by helping retailers do selective marketing and plan their shelf space. For instance, market basket analysis may help to design different store layouts. In one strategy, items that are frequently purchased together can be placed in proximity in order to encourage the sale of such items together. If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items. In an alternative strategy, placing hardware and software at opposite ends of the store may tempt customers who purchase such items to pick up other items along the way. For instance, after deciding on an expensive computer, a customer may observe security systems for sale while heading toward the software display to purchase antivirus software and may decide to purchase a home security system as well. Market basket analysis can also help retailers plan which items to put on sale at reduced prices. If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers as well as computers. Association rule analysis

can also be used in other types of data such as spatial data, multimedia data, time series data, etc (Han and Kamber 2006).

2.1.2 Predictive Modeling

Predictive Analytics is the process of dealing with variety of data and apply various mathematical formulas to discover the best decision for a given situation. Predictive analytics gives company a competitive edge (Bellazzi, Zupan 2008). It is the decision science that removes guesswork out of the decision making process and applies proven scientific guidelines to find right solution in the shortest time possible. Predictive analytics is a solution used by many businesses today to gain more value out of large amounts of raw data by applying techniques that are used to predict future behaviors within an organization, its customer base, its products and services. Predictive analytics encompasses a variety of techniques from data mining, statistics and game theory that analyze current and historical facts to make predictions about future events (Bellazzi and Zupan 2008).

The term Predictive data mining is usually applied to identify data mining projects with the goal to identify a statistical or neural network model or set of models that can be used to predict some response of interest (Stockwell, 2008), For example, a credit card company may want to engage in predictive data mining, to derive a (trained) model or set of models (e.g., neural networks, meta-learner) that can quickly identify transactions which have a high probability of being fraudulent. Other types of data mining projects may be more exploratory in nature (e.g., to identify cluster or segments of customers), in which case drill-down descriptive and exploratory methods would be applied. Data reduction is another possible objective for data mining. Business metrics do a great job summarizing the past (Davies et al 1995). But if you want to predict how customers will respond in the future, there is one place to turn - predictive analytics. By learning from your abundant historical data, predictive

analytics provides the marketer something beyond standard business reports and sales forecasts: actionable predictions for each customer. These redactions encompass all channels, both online and off, foreseeing which customers will buy, click, respond, convert or cancel. The customer predictions generated by predictive analytics deliver more relevant content to each customer, improving response rates, click rates, buying behavior, retention and overall profit (Higgins 2005 and Stockwell, 2008).

2.1.2.1 Classification

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. For example, in a customer database the training set would have relevant customers' information recorded previously, where the prediction attribute is whether or not the given customer is risky.

Decision Trees

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an *internal* or test node. All other nodes are called leaves (also known as terminal

or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range (Rokach and Maimon 2005).

Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path (Rokach and Maimon 2005).

Neural Network

Neural networks represent a brain metaphor for information processing. These models are biologically inspired rather than an exact replica of how the brain actually functions. Neural networks have been shown to be very promising systems in many forecasting applications and business classification applications due to their ability to "learn" from the data, their nonparametric nature (i.e., no rigid assumptions), and their ability to generalize. Neural computing refers to a pattern recognition methodology for machine learning. The resulting model from neural computing is often called an artificial neural network (ANN) or a *neural network*. Neural networks have been used in many business applications for pattern recognition, forecasting, prediction, and classification. Neural network computing is a key component of any data mining tool kit (Priyanka Gaur 2012).

Neural network can be applied in classification, clustering, feature mining, prediction and pattern recognition. It imitates the neurons structure of animals, bases on the M-P model and Hebb learning rule, so in essence it is a

distributed matrix structure. Through training data mining, the neural network method gradually calculates (including repeated iteration or cumulative calculation) the weights the neural network connected. The neural network model can be broadly divided into the following three types (Priyanka Gaur 2012*):*

1. Feed-forward networks*:* It regards the perception back-propagation model and the function network as representatives, and mainly used in the areas such as prediction and pattern recognition;

2. Feedback network*:* It regards Hopfield discrete model and continuous model as representatives, and mainly used for associative memory and optimization calculation;

3. Self-organization networks*:* it regards adaptive resonance theory (ART) model and Kohonen model as representatives, and mainly used for cluster analysis.

2.2 Types of Data Mining Systems

There are many data mining systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited data mining functionalities, other are more versatile and comprehensive. Data mining systems can be categorized according to various criteria (Zaïane, 1999)

Classification according to the type of data source mined: this classification categorizes data mining systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

Classification according to the data model drawn on: this classification categorizes data mining systems based on the data model involved such as

relational database, object-oriented database, data warehouse, transactional, etc.

Classification according to the kind of knowledge discovered: this classification categorizes data mining systems based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

Classification according to mining techniques used: Data mining systems employ and provide different techniques. This classification categorizes data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems (Osmar R. Zaïane, 1999).

2.3 The Data Mining Models

There are many data mining and knowledge discovery methodologies and process models have been developed with various degree of success. In this study we describe the most used process model in industrial and academic projects. These includes six step cios et al process model also called hybrid model, KDD (knowledge Discovery in Database) and CRISP-DM (Cross Industry Standard Process for data Mining)

2.3.1 The KDD Process Model

The KDD Process stands for the Knowledge Discovery in Databases. According to Fayyad (1996) there are five steps: Selection, Pre-processing, Transformation, Data Mining and Interpretation. These five steps are passed

through iteratively. Every step can be seen as a work-through phase. Such a phase requires the supervision of a user and can lead to multiple results. The best of these results is used for the next iteration, the others should be documented.

In the ***Selection***-step the significant data gets selected or created. Hence forward the KDD process is maintained on the gathered target data. Only relevant information is selected, and also metadata or data that represents background knowledge. Sometimes the combination of data from ubiquitous sources can be useful, but possible matters of compatibility have to be observed.

1. A good result after applying data mining depends on an appropriate data preparation in the beginning. Important elements of the provided data have to be detected and filtered out. These kinds of things are settled in the ***Pre-processing*** phase. To detect knowledge the effective main task is to pre-process the data properly and not only to apply data mining tools. The less noise contained in data the higher is the efficiency of data mining. Elements of the pre-processing span the cleaning of wrong data, the treatment of missing values and the creation of new attributes.
2. That data also needs to be transferred into a data-mining-capable format. The ***Transformation*** phase of the data may result in a number of different data formats, since variable data mining tools may require variable formats. The data also is manually or automatically reduced. The reduction can be made via lossless aggregation or a loss full selection of only the most important elements. A representative selection can be used to draw conclusions to the entire data.
3. In the ***Data Mining*** phase, the data mining task is approached. Fayyad gives a classified overview over existing data mining techniques. He makes suggestions, which technique may be used for which objectives,

but most of the techniques are now improved. The output of this step is detected patterns.

4. The ***interpretation*** of the detected pattern reveals whether or not the pattern is interesting. That is, whether they contain knowledge at all. This is why this step is also called evaluation. The duty is to represent the result in an appropriate way so it can be examined thoroughly. If the located pattern is not interesting, the cause for it has to be found out. It will probably be necessary to fall back on a previous step for another attempt.

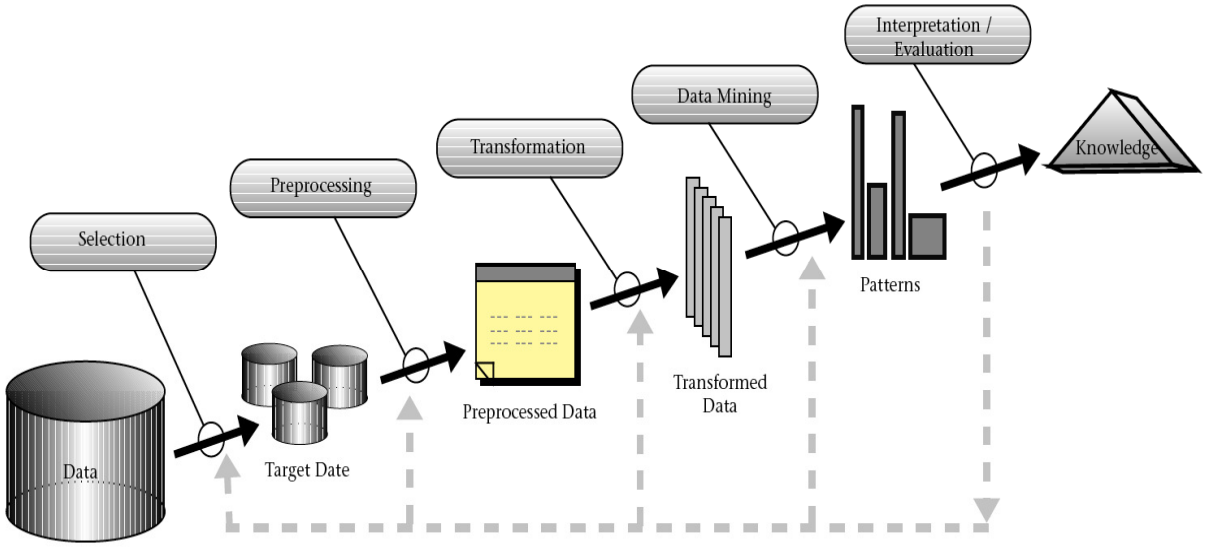


Fig 2.2 The KDD process Model (Malik, 2012)

2.3.2 Crisp DM Process

In response to common issues and needs in data mining project in the mid 90's, a group of organizations involved in data mining (Teradata, SPSS -ISL-, Daimler-Chrysler and OHRA) proposed a reference guide to develop data

mining projects, named CRISP-DM (Cross Industry Standard Process for Data Mining) (Chapman et al., 2000).

1. **Business Understanding:** Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
2. **Data Understanding:** Start by collecting data, then 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 about hidden information.
3. **Data Preparation:** Includes all activities required to construct the final data set (data that will be fed into the modeling tool) from the initial raw data. Tasks include table, case, and attribute selection as well as transformation and cleaning of data for modeling tools.
4. **Modeling:** Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.
5. **Evaluation:** Thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. Determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results is reached.
6. **Deployment:** Organize and present the results of data mining. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process.

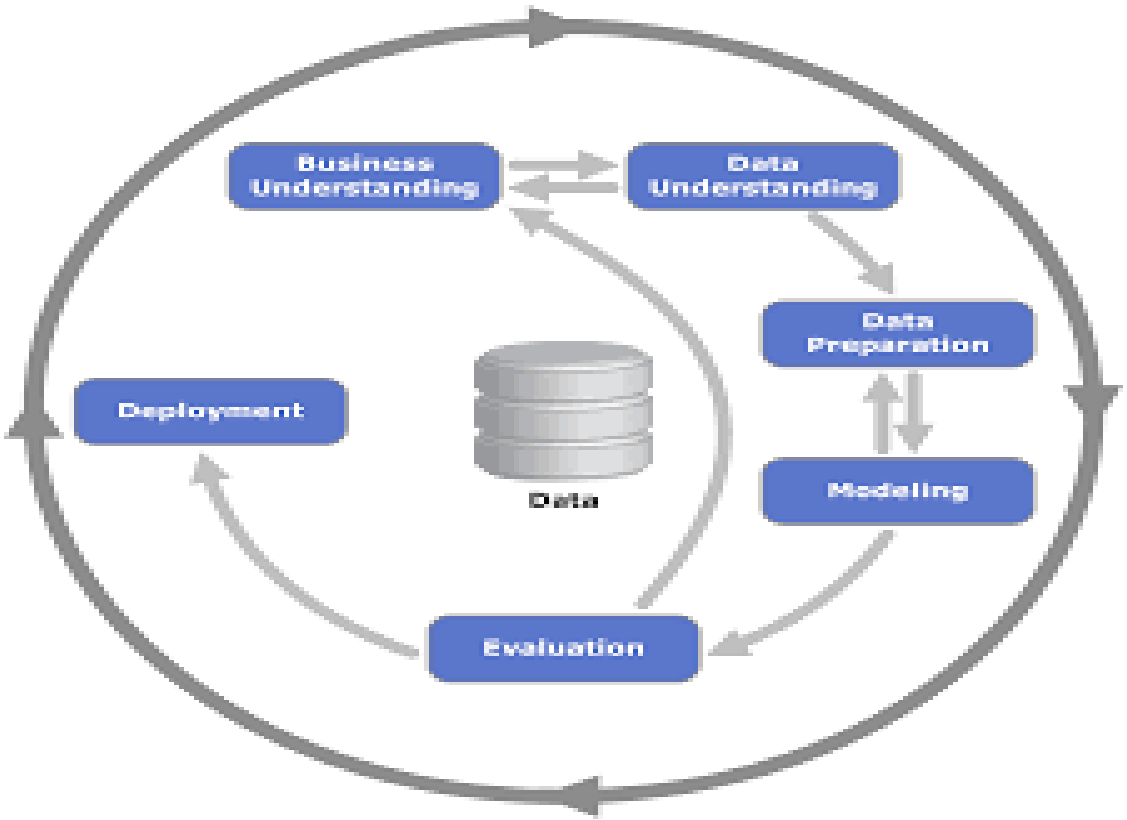


Fig 2.3 Crisp DM Process Model (Chapman et al., 2000)

2.4 Application of Data Mining

Data Mining was originally developed to act as expert system to discover problems and did not require assumption to be made about data. Nowadays companies and individuals apply data mining in order to achieve their goals.

2.4.1 Data Mining in the Insurance Industry

Data mining is becoming common in both the insurance sectors like private and public. Data of the customer are one of the most valuable assets of any firm. The traditional methods, which were used for handling huge amounts of data generated by insurance transactions, are too complex. For transferring huge amount of data for decision making, data mining makes the methodology. Insurance firms use the data mining methodologies to enhance research and

increase sales among the customers. The data mining used for various tasks in the insurance sector as follows **(**Umamaheswari and Janakiraman 2014).

Acquiring new customers:

Acquisition of new customer is most important scenario of any firm. Traditionally, the insurance companies used the services of brokers to acquire the customers, but today a lot of ways helps to acquire the new customers. Insurance firm focused of both acquiring new customer & retaining existing ones.

Cluster Analysis used in the private sector to identify target group of customers. It involves targeting the population who are most likely to become customers or most profitable to the company.

Customer level analysis

It is analysis of customer purchase patterns and behavior. Using associated discovery technique, most insurance firms accurately select which policies and services to offer which customers. According to Zhikun et al (2014), it used data mining technology for insurance settlement and analyzed the customer records and also developed function structure model for customer analysis using data mining method.

Customer Segmentation

Segment based products for targeting the customers. Data mining can be used for customer segmentation, for promoting the cross-selling of services, and in increasing customer retention. Customers are assigned to lifestyle segment based on their purchase history. Market segmentation is the key issue for the development of loyal relationships among the customers (Miguéis et al, 2012).

Policy designing and policy selection

The insurance firm made the investigation whether people tend to purchase policy for the reason and the policy designed. In that case, to compete successful in the market, insurance companies used data mining technologies.

Prediction

Data mining used for variety of applications such as predicting and classifying customer's and clustering customer characteristics for achievement of profitability. From the customer point of view, predictive analytics provides some benefits such as simplified claim handling process, reduced policy premium for low risk customers, faster and automated claim settlement. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. It performs inference on the current data (insurance dataset) order to make predictions (Balaji and Srinivasta, 2012)

Claims management

It is one of the most function is the insurance; data mining handled the claims management function such as claim analysis and fraud analysis.

Developing new product lines

To develop the new product/plan depends on the customer needs. Insurance firms utilized all of their available information to better develop new product and marketing campaigns **(**Umamaheswari and Janakiraman 2014)

Underwriting and Policy management

Data mining can be used in this application to optimize the function of the insurance value chain (Premium Analysis and Loss analysis)

Risk management

One of the main stages in the process of risk management is risk financing is, of course, insurance. Insurance industry is keen in identifying the risks pertaining to their business.

Reinsurance:

Reinsurance comes under in the fields of risk management. The reinsurer may be either a specialist reinsurance company, which only undertakes reinsurance business, or another insurance company. Data mining tools can develop predictive models to arrive at the reinsurance level for the book of business based on the historical claims data. These predictive models can be identified suitable policies for reinsurance based on the loss experience of similar policies in past.

Fraud detection

Detecting fraud claims is important in the insurance firm. Data mining isolates the factors that lead to fraud waste and abuse. To identify which transactions are most likely to be fraudulent. This is called as Fraud anomaly detection. In medical insurance, various medical insurance agencies suffered due to fraud claim in the health insurance; here he developed a model with three steps for the health insurance fraud detection. And he discussed the characteristics of fraud detection are high claims payment data is incorrect, suspicious data analysis, problem of hospital or physician (Kuo-Chung Lin and Ching-Long Yeh

2012). The types of fraud and how much of fraud activities in the insurance firm discussed and he developed a claim sorting algorithm for the claim processing systems (Derrig 2002).

Trend analysis

Trend analysis often refers to the science of studying changes in social patterns, including fashion, technology, and consumer behavior. Trend analysis also Applicable in insurance to reveal difference between the typical customers of this month and last month (Venkatesh 2013). Data mining used in the different service industry especially in insurance firms the most frequently used applications for customer segmentation, customer retention, risk assessment and fraud detection and Policy approval process.

2.4.2 Insurance Risk Assessment

Insurance companies lose lots of money each year because of not profitable and risky customers which are attracted blindly. Risky customers are one of the most important treats to insurance companies; therefore some of these companies adopt a credit scoring and risk assessment approach for identifying profitable and risky customers. One of the most preferable methods for risk assessment is data mining. Data mining is a powerful new technology with great potential to help insurance firms focus on the most important information in the data they have collected about the behavior of their customers and potential customers

A proper assessment of the size of an insurance company's risk or, a good prediction of future expected claims, is of vital importance to the company for several reasons. First, correct assessment of future expected claim size is very important in calculating appropriate premiums, thus affecting profitability. Second, by charging risk-adequate premiums, the insurance company can

avoid adverse selection, i.e., the loss of good (low claim potential) insurance customers because its premiums are too high priced (Growitsch et al., 2006). A good risk assessment might even allow undercutting the premium level in certain lines of business, leading to a gain of market share in those segments. Third, proper risk assessment is becoming of increasing importance to rating agencies (S & P, 2005). A good rating is essential to lower refinancing costs of the insurance company and it also signals the company's reliability, which can enhance customer loyalty. Finally, the ability to select certain risks based on advanced risk assessment such as data mining can be helpful in identifying and classifying overall portfolio risk.

2.4.2.1 Types of Risk

Insurance risk is classified as pure risk, personal risk and Fundamental risk.

Pure Risk

When the risk is either all or none, it is called a pure or static risk. Pure risks are straight bets, and most insurance companies deal in these kinds of bets. This is because there are only two possible outcomes for the risk of insuring the person or property: either the risk will pay off, or it won't. This design is obviously at work in policies, such as life or flood insurance. These policies only pay off in the event of total loss of the insured item. The benefit of pure risk policies to the policyholder is a potentially large payoff in the event of a catastrophe; the benefit to the insurance company is the likelihood that the policy will remain active, and premiums will continue to be paid.

Personal Risk

When an individual is personally affected by the risk involved, this is known as personal risk. Personal risk is the basis behind a wide variety of insurance types, including unemployment, health, homeowner's and renter's insurance.

This is also where policyholders find the most ambiguity in their policies. Losses in a personal risk policy do not have to be total; and because the chances of at least a partial payout of the policy are good, many insurance companies look to specify the circumstances under which a loss is covered by the policy. For example, a health insurance policy may cover cancer treatment but only if that treatment falls within certain guidelines.

Fundamental Risk

Fundamental risk is one that involves the entire community. These types of risk include high inflation, stock market crashes, high instances of unemployment and widespread natural disasters. Insurance companies occasionally find themselves wrapped up in these types of fundamental risks (e.g., the homeowner's insurance companies were entangled in debts to homeowners from hurricane Katrina for years), but most fundamental risks must be insured by government agencies. Stock market crashes and bank runs are a good example of fundamental risks handled by government agencies, such as the Federal Reserve Bank (www.consumerfed.org).

2.5 Related Works

Today the use of data mining technology in support of business decision-making is growing fast. Accordingly attempts have been made to apply data mining technology in solving business problems in Ethiopia context. In his research decision tree and rule induction predictive data mining techniques were applied in driver and road factors for car accidents to identify hidden patterns in the accident dataset. As a methodology the researcher has used CRISP-DM 1.0 and the WEKA data mining tool while implementing ID3, J48 and PART algorithms. On the other hand, Belachew Microfinace (2013) research work applied clustering and then classification models to categorize and predict customers. In this study the cluster instances has been grouped

for similarity using Simple K-means algorithms and then J48 decision tree algorithm has been employed to classify the result obtained from clustering. The other study is on banking by Luel (2011), which assessed the application of data mining techniques, classification and clustering to support expansion of Electronic Fund Transfer (EFT) of POS service at Dashen Bank S.C. Again, k-means is applied for clustering along with j48 and neural network for further classification. The research conducted using 11000 records and J48 has registered an accuracy result of 99.54% whereas multilayerperceptron registered 99.90%. The researcher concluded that the output is very encouraging as it identifies high, medium and low value customer properly. The researcher also recommended that the company need to design knowledgebase system, which can provide advice for the domain experts. In addition similar researches of several sectors have also been made by different researchers including Askale (2001), Samson (2009) and Tilahun (2009) in banking, Tesfaye (2010) Airlines, Tariku (2013) and Girma (2012) in Medicine.

There are also studies conducted to investigate the application of data mining in the insurance business. Tesfaye (2002) used the data mining technique for insurance risk assessment in Nyala insurance company. The researcher developed a predictive model using decision tree and neural network classification technique in order to classify a given policy as low risk, medium and high risk. As a research methodology, the research used literature review, data collection, data preparation, training and building models, performance evaluation and prototype development. The researcher conducted using 1332 records and the result of decision tree and neural network reported. Accordingly, neural network registered accuracy level of 92.24 percent whereas j48 decision tree classification algorithm registered 95.69 percent. The research concluded that the problem in insurance risk assessment in particular assessment made during policy renewal could be leveraged using data mining techniques. The researcher also recommend on using alternate neural network models such as adaptive resonance neural networks and

probabilistic neural network. Mesfin (2005) attempts to investigate the possible application of predictive data mining techniques in the renewal process of personal accident policies as a case study on EIC. The other research, which is undertaken by Tariku (2008) also attempted to investigate fraud detection the case of Africa insurance share company. The research conducted using clustering and then classification. k-means algorithms has been implemented to categorize the data into fraud and non fraud and then cluster records were submitted for the classification module for model building using the J48 decision tree algorithm and naïve Bayes. The research conducted using 17810 records and the result indicated that j48 decision tree algorithm registered an accuracy level of 99.96% and naïve bayes 91.10%. The researcher also concludes that it is possible to identify those fraud suspicious insurance claims and suggest concrete solutions for detecting them, using the DM techniques.

However, no one investigates the application of data mining to identify features on both drivers and vehicles side to classify policy on motor insurance class of business as very high risk, high risk, medium risk and low risk. It is, therefore, with this understanding this study is conducted using data mining technology to predict features and report hidden knowledge discovered by data mining techniques in Ethiopian insurance corporation (EIC).

Chapter Three

Data Mining Methods for Risk Clustering and Classification

The ability to predict or classify customers risk level is very important in the insurance industry. A very promising ground to attend this objective is the use of data mining. Researchers apply different data mining methods in order to classify customers. In recent years both supervised learning and unsupervised learning methods are employed.

The data mining methods that are experimented in this research are unsupervised learning method (k-means clustering) and supervised learning method (decision tree and neural network). First k-means clustering algorithm is applied due to the unlabelled nature of the dataset and then in order to create the model classification algorithms are applied. Based on the result of the classification algorithm a system developed. The architecture of the system shown below:

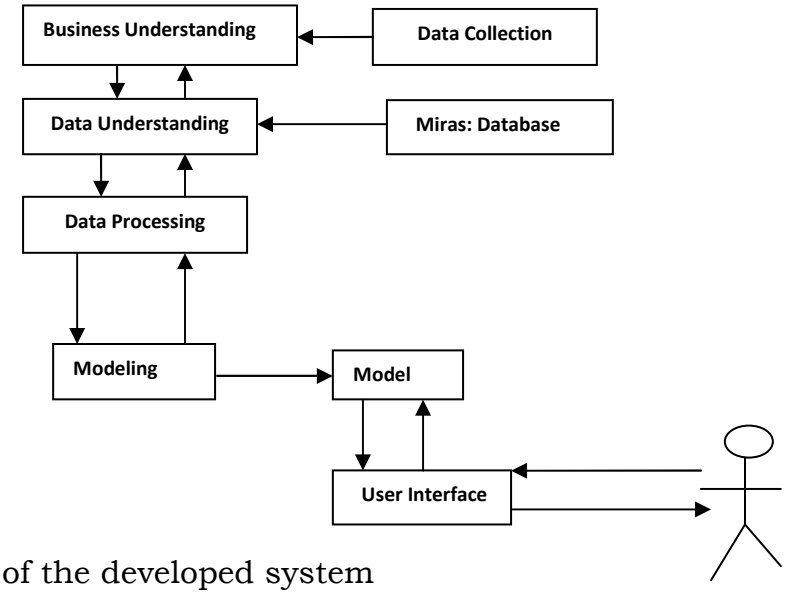


Fig 3.1 Architecture of the developed system

3.1 Hybrid DM Process Model

In order to achieve the general objective of this study we followed hybrid data mining process model. Hybrid data mining process model is selected due to the advantage of providing more general, research-oriented description of the steps and introducing a data mining step instead of the modeling step. The process model of Cios et al. was first proposed in 2000 by adapting the CRISP-DM model to the needs of academic research community (Cios et al., 2000; Cios & Kurgan, 2005). The model, as shown in Figure 2.2 consists of six steps: The following discuss the steps briefly

1. Understanding of the problem domain
2. Understanding of the data
3. Preparation of the data
4. Data mining
5. Evaluation of the discovered knowledge
6. Use of the discovered knowledge

3.1.1 Understanding of the problem domain

To understand, identify and analyze the problem domain, observation and discussion conducted with key selected people. Based on the explanation and evaluation made by the experts, there are different kinds of customers in the corporation. Some customers are high risk customers; these types of customers are those customers with high probability of having claim. On the other hand, there are also medium risk customers with less probability of having claim. The third types of customers are low risk customers with near zero probability of having claim. Thus, the main aim of this research is to differentiate customers whether they are very high, high, medium and low risk customers based on the claim reported in a given year.

3.1.2 Understanding of the data

The potential source of data used to undertake this research is mainly the policy and claims database used by selected branch of EIC. The manual formats used to collect information about the vehicle and the owner (driver) both at the time of underwriting and claims request are also used in support of preprocessing the collected data. From the available 67 branches this research addresses the branches located in North Western Main Branch, North Eastern Main Branch, Western Main Branch, Southern Main Branch, and Eastern Main Branch. The sampling technique used in the above selected branches is purposive sampling and the total sample data selected is 5176. The reason for purposive sampling for the branches located in Addis Ababa is mainly due to the problem of data access from the branches outside Addis Ababa and the short time span for research undertaking. Before applying the data mining techniques, the data need to be understood. Therefore, the researcher together with domain experts list out attributes, data type and short descriptions and then the data are checked for completeness, redundancy, missing values, plausibility of attribute values and finally the step includes verification of the usefulness of the data with respect to the DM goals.

3.1.3 Preparation of the data

The collected data are preprocessed and cleaned in to a form suitable for the particular data mining software that is used in the study. Different data mining preprocessing techniques are applied. For some of the data the following cleaning mechanisms are applied: these include handling noisy data and unknown values, as well as accounting for missing data fields, deriving new fields from the existing ones, and summarization of data. Then initial relevant

features based on the goal of the study are identified in consultation with domain experts.

3.1.4 Data mining

For conducting this research the WEKA (Waikato Environment for Knowledge Analysis) version 3.7.4 Data mining software is chosen. Weka is chosen because of its spread application in different data mining researches and familiarity of the researcher with the software. The research implemented both clustering and classification methods. For clustering, k-means clustering algorithm is employed due to the advantage of time complexity, space complexity and order independent (Rokach and Maimon, 2006). For classification neural network and j48 decision tree classification algorithms are employed. Decision tree classification algorithm selected due to the following reasons (kumar 2004).

- They are easily understandable. They build a model easy to understand for the user.
- They are one of the most used data mining techniques.
- Inexpensive to construct
- Extremely fast at classifying unknown records
- Easy to interpret for small-sized trees
- Accuracy is comparable to other classification techniques for many simple data sets.

Neural network selected due to the following advantage (Kumar 2004):

- Mapping Capabilities, map input patterns to their associated output patterns.
- Learn by example. NN architectures can trained with known examples of a problem before they are tested for their 'inference' capability on

unknown instances of the problem. They can, therefore , identify new objects previously in trained,

- Possess the capability to generalize. Predict new outcomes from past trends.
- Robust systems and are fault tolerant. i.e recall full patterns from incomplete, partial or noisy patterns. The algorithm alternated between these passes several times as it scans the training data.

3.1.5 Evaluation of the discovered knowledge

The Evaluation checked whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. The developed clustering models are evaluated based on sum of square errors values, the number of iteration and experts' judgments. The developed classification model also evaluated using separate test dataset.

3.1.6 Use of the discovered knowledge

In order to use the discovered knowledge an attempt is made to develop an early warning system prototype by selecting those features. SQL server 2008 is used to design the database. The selected features are designed and developed by using an object oriented programming language called C#.

3.2 K-Means Clustering

Cluster analysis or data clustering is a data exploration technique that allows objects with similar characteristics to be grouped together in order to facilitated their further processing (Pham, Dimov, and Nguyen 2004). Cluster Analysis of a data is an important task in Knowledge Discovery and Data Mining.

Clustering is the process to group the data on the basis of similarities and dissimilarities among the data elements. Clustering is the process of finding the group of objects such that object in one group will be similar to one another and different from the objects in the other group. A good clustering method will produce high quality clusters with high intra cluster distance similarity and low inter cluster distance similarity. According to Bhatia and Khurana (2013), there are different similarity measures available such as standard Euclidean distance, Manhattan distance, Murkowski distance. In Weka we implemented Euclidean distance and Manhattan distance functions are implemented in weka to cluster the give dataset using the k-means algorithms.

According to Rokach and Maimon (2005) clustering methods can be classified as hierarchical and partition. This study implemented K-means algorithm which is an example of partition method. The reason behind the popularity and implementation of the algorithm in this study are:

1. Its time complexity is $O(mkl)$, where m is the number of instances; k is the number of clusters; and l is the number of iterations taken by the algorithm to converge. Typically, k and l are fixed in advance and so the Algorithm has linear time complexity in the size of the data set.

2. Its space complexity is $O(k+m)$. It requires additional space to store the data matrix. It is possible to store the data matrix in a secondary memory and access each pattern based on need. However, this scheme requires a huge access time because of the iterative nature of the algorithm. As a Consequence, processing time increases enormously.

3. It is order-independent. For a given initial seed set of cluster centers, it generates the same partition of the data irrespective of the order in which the patterns are presented to the algorithm.

4. In addition to the above reasons Rokach and Maimon (2005) stated that k-means is simple and can be used for a wide variety of data types.

3.2.1 K-means Algorithm

K-means cluster algorithm was proposed by J. B. MacQueen in 1967, which is used to deal with the problem of data clustering, the algorithm is relatively simple, so generate a widely influence in the scientific field research and industrial applications. It is based on decomposition, using K as a parameter, divide n object into K relatively low similarity between clusters and minimize the total distance between the values in each cluster to the cluster center. The cluster center of each cluster is the mean value of the cluster. The calculation of similarity is done by mean value of the cluster objects (Bhatia and Khurana 2013).

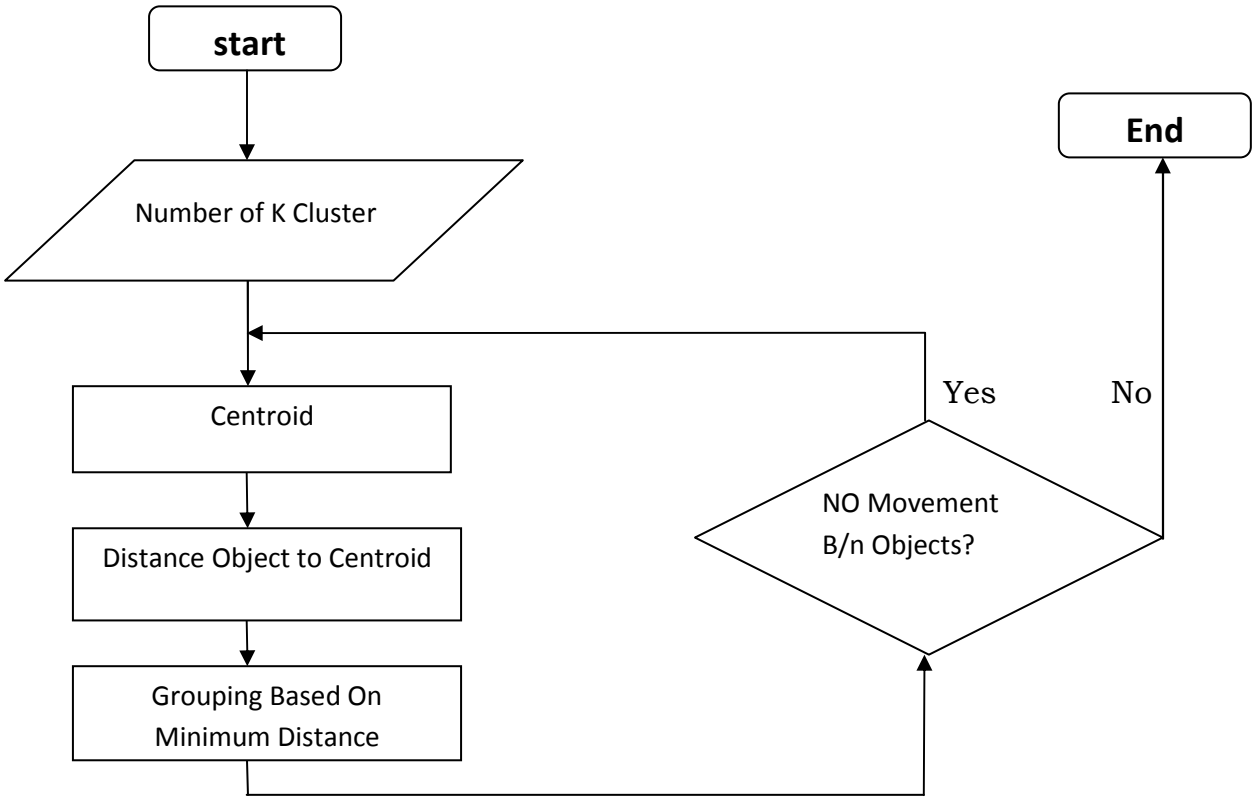


Fig:3.2 K-Means Algorithms Flow Chart

Step 1 Distribute all objects to K number of different cluster at random;

Step 2 Calculate the mean value of each cluster, and use this mean value to represent the cluster.

Step 3 Re-distribute the objects to the closest cluster according to its distance to the cluster center.

Step 4. Update the mean value of the cluster. That is to say, calculate the mean value of the objects in each cluster.

Step 5 Calculate the criterion function, until the criterion function converges. Usually, the K-means algorithm criterion function adopts square error criterion which is one example of partition method clustering.

3.3 Decision Tree Classification Techniques

In machine learning environment, methods that help in order to predict are commonly referred to as supervised learning. Supervised methods are methods that attempt to discover the relationship between independent variable and dependent variable. The relationship that is discovered is represented in a structure referred to as a model.

Supervised models can be categorized as classification (classifiers) and regression models. Regression maps the input model to the real-valued domain. For example using regression model one can predict the demand for a certain product given its characteristics. On the other hand classifiers map the input space into predefined classes. For example classifiers can be used to classify mortgage consumers as good (good pay back history) and Bad (delayed pay back). Depending on the type of problems, Data mining apply different classifiers; there are for example, support vector machines, decision tree, Bayesians network, neural network etc (www.worldscibooks.com).

In this study we implemented decision tree and neural network as classifiers. When decision tree is used for classification tasks, it is more appropriately referred to as classification tree. The paper used classification to classify policy to a predefined set of classes (Very High risk, High Risk, Medium Risk and Low risk) based on their attributes value.

3.3.1 Decision Tree Algorithms

There are a variety of algorithms for building decision trees that share the desired quality of interpretability. According to Rokach and Maimon (2005) CHAID (chi square Automatic Interaction Detection) algorithm is based on a statistical approach. It is different from other decision tree algorithms in the sense of attribute selection measure for tree formation. It uses chi square test to choose best split instead of information gain (reduction in entropy) as in C3.5 during tree generation. The second classification algorithm is QUEST which stands for Quick, Unbiased, and Efficient Statistical Tree. It is a binary classification algorithm for constructing decision trees. A major advantage for the development of the algorithm was to reduce the processing time required for large C&RT (classification & regression tree) analyses with either many variables or many cases, secondly decrease the trend found in classification tree algorithms to give priority to attributes that permit more splits. The third algorithm C 4.5 uses information gain measure to select the test attribute at each note in the tree such a parameter is termed as an attribute selections measure or a measure of goodness of split. The ID3 algorithm was originally developed by J.Ross Quinlan at the University of Sydney, first presented it in the 1975 book “Machine Learning”. It identities attributes that differentiate one class from another. All attributes must be known in advance, and must also be either continuous or selected from a set of known values. The other known decision tree classifier is CART algorithm which stands for Classification and Regression Trees algorithm. It was developed by Leo Breiman, Jerome Friendman and later joined by Richard Olshen and Charles Stone.

The above decision tree classification algorithms have their own advantage and disadvantage. For example one of the limitations of ID3 is that it is very sensitive to attributes with a large number of values, therefore to overcome this problems C4.5 algorithm extends the ID3 algorithm through the use of information gain to reduce the problem of artificially low entropy values for attributes.

3.3.1.1 The J48 Decision Tree Algorithm

J48 is an open source java implementation of the C4.5 algorithm in the weka data mining tool (Gholap 2012). C4.5 is a program that creates a decision tree based on a set of labeled input data. This algorithm was developed by Ross Quinlan. There are two methods supported by j48, the first method is known as subtree replacement, it works by replacing nodes in decision tree with leaf. Basically by reducing the number of test with certain path. It works with the process of starting from leaves that formed the overall tree and moves backward toward the root. The second method implemented in j48 is subtree raising by moving nodes upwards toward the root of tree and also replacing other nodes on the same way.

J48 algorithm is a popular machine learning algorithm. In Weka, the implementation of a particular learning algorithm is encapsulated in a class, and it may depend on other classes for some of its functionality. J48 class builds a C4.5 decision tree. Each time the Java virtual machine executes J48, it creates an instance of this class by allocating memory for building and storing a decision tree classifier. The algorithm, the classifier it builds, and a procedure for outputting the classifier is all part of that instantiation of the J48 class.

C4.5 builds decision trees from a set of training data in the same way as ID3. Both use the concept of information entropy. The training data is a set S= s1,s2,s3 …….Sn of already classified sample. Each sample in the training data set defined as Si=x1,x2,……Xn is a vector where x1,x2, ….Xn represent

attributes or features of the sample. The training data augmented with a vector C=c1,c2,.....Ck where c1,c2,..... Ck represent the class to which each sample belongs.

C4.5 selects one attribute based on the criteria of normalized information gain that most effectively splits its data of samples into subsets enriched in one class or the other. The attribute with the highest normalized information gain is chosen to make the decision. The algorithm has a few base cases (Sehgal et al, 2012) :

1) All the samples in the list belong to the same class. In this case the algorithm simply creates a leaf node for the decision tree saying to choose that class.
2) None of the features provide any information gain. In this case the algorithm creates a decision node higher up the tree using the expected value of the class.
3) Instance of previously-unseen class encountered. The algorithm handle such case by created a decision node hither up the tree using the expect value.

In this study j48 algorithm is selected as classifier due to the following advantage over the others decision tree classification algorithms. The first important point is due to the improvements made on the ID3. Some of these are:

1) The ability to handle both continues and discrete attributes- In order to handle continuous attributes C4.5 creates a threshold and then splits the list into those whose attributes value is above the threshold and those that are less than or equal to it.

2) The ability to handle training data with missing attributes value- C4.5 allows attributes values to be marked as? for Missing, missing attribute values are simply not used in gain and entropy calculations.
3) Handling attributes with differing costs.
4) Pruning tree after creation- C4.5 Remove branched that do not help by replacing them with lead nodes (Sehgal et al, 2012).

Moreover, there are papers that made comparison between decision tree classification algorithms. In Sharmal et al (2011) paper the data mining tools weka was used to compare ID3, CART, ADTree and J48. The experiment focused on categorizes the given dataset as spam or non spam. Their result indicated that J48 algorithm is more accurate than the other three classification algorithms. Zhao and Zhang (2007) also conducted a research on comparison of decison tree method. Their paper tries to compare decision tree algorithms (i.e REPTree, Random Tree, Decision Stump, Random Forest, J48, NBTree and AdTree). According to the experiment ADTree shows the best performance only in terms of accuracy, Decision stump is the best only in terms of speed, J48 is the optimal choice in terms of both accuracy and speed.

3.4 Neural Network

Artificial Neural Networks (ANNs), are computational models that consist of a number of simple processing units that communicate by sending signals to one another over a large number of weighted connections. ANNs are originally developed from the inspiration of human brains. Like human brains, neural networks also consist of processing units (artificial neurons) and connections (weights) between them. ANNs have powerful pattern classification and pattern recognition capabilities through learning and generalize from experience. ANNs are non-linear data driven self adaptive approach as opposed to the traditional model based methods. It is a powerful and popular tool especially where there is a relationship between dataset. ANNs imitate the learning process of the

human brain and can process problems involving non-linear and complicated data even if the data are imprecise and noisy (Singh and Chauhan 2009).

Neural networks are built from layers of neurons connected so that one layer receives input from the preceding layer of neurons and passes the output on to subsequent layer. A neuron is a real function of the input vector (y1,……., yk). The output is obtained as (Ganatra et al 2011):

$$f(x_J) = f\ (a_J + \sum_{i=1}^{k} w_{ij}\ y_i)$$

Where f is a function, typically the sigmoid (logistic or tangent hyperbolic)

3.4.1 Neural Network Algorithms

There are different kinds of architecture of ANNs however; the two most widely used ANNs are Feed forward neural network and recurrent networks. In our data mining implementation tool i.e weka we have different applicable neural network algorithms but for our purpose we have selected feed forward neural network.

3.4.1.1 Feed forward (multi-layer) Networks

It is one of the popular approaches in artificial neural network. In feed forward-back propagation (or simply back propagation) information flows in one direction along connection pathways from the input layer to the hidden layers to the final output layer (see Fig 3.2). There is no feedback (loops) i.e the output of any layer does not affect the same or preceding layer. The algorithm cycles through two distinct passes, the first one is forward pass followed by a backward pass through the layers of the network (Ganatra et al 2011).

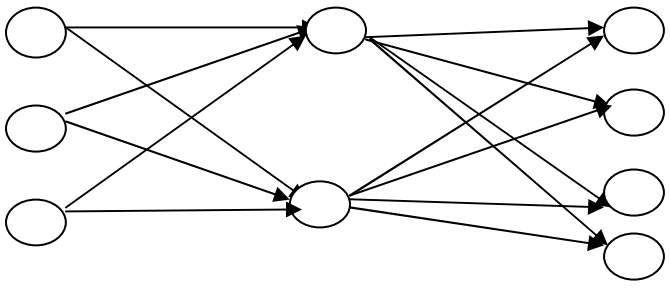
Forward Pass: Computation of outputs of all the neurons in the network

- The algorithm starts with the first hidden layer using as input values the independent variables of a case from the training data set.

- The neuron outputs are compute for all neurons in the first hidden layer by performing the relevant sum and activation function calculations are preformed to compute the outputs of second layer neurons.
- These outputs are the inputs for neurons in the second hidden layer. Again the relevant sum and activation function calculates are performed to compute the output of the second layer neurons.

Backward Pass: Propagation of error and adjustment of weights

- This phase begins with the computations of error at each neuron in the output layer. A popular error function is the squared difference between Ok the output of node OK and Yk the target value for the node
- The target value is just 1 for the output node corresponding to the class of the example and zero for other output nodes.
- The new value of the weigh wjk of the connection from node j to node k is given by: Wnewjk= woldjk+nojqk. Here n is an unimportant tuning parameter that is chosen by trial and error by repeated runs on the training data. Typical values for n are in the range 0.1 to 0.9.
- The backward propagation of weight adjustments along these lines continues until we reach the input layer.
- At this time we have a new set of weights on which we can make a new forward pass when presented with a training data observation.



Input Layer Hidden layer Output Layer

Fig: 3.3 A Multi-Layer Feed forward neural network

Why Back Propagation Network?

A study comprising feed forward network, recurrent neural network and time-delay neural Network show that highest correct classification rate is achieved by the fully connected feed forward neural network (Ganatra, Kosta Panchal and Gajjar 2011). According to Kumar (2004) back propagation neural network is a powerful system, often capable of modeling complex relationships between variables.

Chapter Four

Data Preparation

Insurance is an agreement whether, for a stipulated payment called the premium, one party (the insurer) agrees to pay to the other (the policyholder or his designated beneficiary) a defined amount (the claim payment or benefit) upon the occurrence of a specific loss (Anderson and Brown, 2006). Generally the operation of an insurance company can be divided in to claim department and underwriting department. This paper is more focused on the management of underwriting department due to the fact that underwriting is about risk assessment of the subject matter insured.

Underwriting is the function of evaluating the subject of insurance. Whether a person, property, profession, business, or other entity, and determining whether to insure it. Underwriter is a professional that has the ability to understand the risks to which the underwritten object is exposed to (Macedo, 2009*).*

There are different class of business underwrite all over the world. For a given insurance company, it is difficult to underwrite all the available class of businesses. EIC provides different class of businesses. The following table 4.1 describes some of the type of policies issued by the underwriting department.

Type of Policy	Description
Motor	Cover for loss or damage on road worthy Vehicle.
Marine	Cover for loss or damage on property transported by ship
Fire and Allied Perils	Cover for loss or damage due to fire
Burglary and House Breaking	Cover for loss or damage on property due burglary and house braking.
Plate Glasses	Cover for accidental loss or damage on plate Glass
Engineering	Cover for loss or damage on construction and machinery
Workmen's Compensation	Cover for medical expense or death on an employee at work place.
Good in Transit	Cover for loss or damage for good in transit.
Money	Cover for loss of money
Public liability	Cover for damage or loss on public property
Professional Indemnity	It is a type of insurance that gives cover for professionals

Table 4.1 Description of some of the policies issued by EIC

4.1 EIC Motor Insurance Policy

One of the available class of business underwrite by EIC is motor insurance policy. As compared to the other class of businesses, motor insurance class of business take the lion share in premium collection. According to a report by national bank of Ethiopia from the collected annual premium from all insurance companies in Ethiopia motor insurance contribute around 60 percent.

4.1.1 Classification of Motor Insurance policy.

EIC classifies motor insurance class of business base on the indemnity; accordingly the corporation has Comprehensive cover, Third party cover and

Third party fire and theft. These are described as follows (www.suncorp.com.au):

Comprehensive Cover: Comprehensive cover is the highest level of cover which gives cover for both own damage and third party. It covers accidental loss or damage to the vehicle caused by Hail, storm, flood, fire, Theft, attempted theft, malicious damage, collision and impact. EIC excluded accidental loss or damage due to act of God. According to EIC comprehensive cover policy, the indemnity amount limited to 10000 for third party property damage, 30000 third party deaths, and 3000 third party injury.

Third party Fire and Theft: Third party fire and theft cover indemnifies the insurer's vehicles from accidental loss or damage aroused by fire, theft or attempt of theft plus third party cover.

Third Party Cover: Third party cover is the minimum amount of insurance cover. It only covers insurer's vehicle for damage to someone else's vehicle or property, or injury to someone else in an accident. This includes accidents caused by insurer's passenger.

4.2 Business Understanding

According to a report by international actuarial association (2004) insurer risk can be categorized under four major headings underwriting, Credit, Market and Operational risk.

EIC measures the risk level of the subject matter insured at the time of underwriting. In Motor insurance class of business, risk level are measured based on the attribute values of the subject matter insured.

EIC considered commercial vehicles as more risky than private vehicles and among commercial old vehicles more risky than new ones. The above discussion clearly indicated by motor underwriting rate chart of the corporation.

4.2.1 Risk Assessment Process

There is no any written procedure or steps that the underwriting department of EIC has been followed. The first and the most important point is determine whether the vehicle is road worthy or not. These can be done by identifying some attributes or characteristic of the vehicle and physical survey. EIC has assigned surveyor or inspector in each main branch for the physical assessment of the vehicle. After the physical survey and all the attributives of the vehicle identified the next stage is measuring the risk level and categorization. There are decisions that can be made by the underwriter or the underwriting department head. According to the domain experts, these decisions are rejected, issued on a substandard basis, issued on a standard basis, or issued on a preferred basis.

Rejecting Applicants: This option is implemented when the underwriter finds the applicant represents a risk that falls outside to the underwriting standards established by the corporation.

Issuing policies on a Substandard Basis: This option is applied when a risk is not deemed to be outside underwriting standards, but considered to be of high risk with those standards. The insurer generally has three basic options:

1) Issue the policy with a higher premium than would be required for a standard policy
2) Issue the policy with limited benefits
3) Issue the policy with certain exclusions.

Issue policy on a standard Basis: Such policies are issued within the normal boundaries of underwriting standards.

Issuing Polices on a preferred Basis: Policy that represent the lowest rates offered by an insurer for its coverage.

4.2.2 Current practice of the Corporation

Currently the underwriting departments of EIC underwrite large number of applicants every day. All the available class of businesses have application forms. In Motor insurance class of business the available form is filled by the applicant which has an insurable interest on the subject matter insured. Some of the attributes collected at the time of motor underwriting are age of vehicle, make, purpose, chassis, Engine, sum insured, Horse Power (cc), claim experience, Body Type, Body color, and plate number.

The underwriters collect the necessary document such as application form from the applicant and survey report form from surveyor. These forms help the underwriter in order to measure the risk level. In calculating the premium the underwriters take different attributes in to consideration. According to the domain experts at southern main branch the main attributes of motor insurance that should be taken in to consideration are sum insured, hors Power, purpose and Year of Make. Based on the above necessary document the underwriters calculate the premium using motor underwriting rate chart.

EIC implemented INSIS Insurance software since 2011. The software helps underwriters to capture the necessary document and calculates the premium based on the input attributes. However, INSIS doesn't allow the underwriter at EIC to capture detail about drivers. Therefore, EIC doesn't take drivers attributes for premium calculation.

4.3 Understanding the data

The most important thing in any data mining project is the data itself and the source of the data. Data warehouse is a good source in order to understand the fields, values, data type and definitions of fields. All the necessary raw data about the motor vehicle collected from INSIS software and details about the drivers is collected manually from the available claim files. Analysis of the data and identifying the relationship of the data with the stated problem has been

done together with the domain expert. The following sections describe the structure of the collected data.

4.3.1 Initial data collection

The data collection process has been done in five different main branches of EIC. These five main branches are North Western Main Branch, North Eastern Main Branch, Western Main Branch, Southern Main Branch, and Eastern Main Branch. The total number of records collected from each main branches are summarized in the following table.

s.n	Main Branch	Records collected
1	North Western	1215
2	North Eastern	1235
3	Western	875
4	Southern	918
5	Eastern	980
Total		5223

Table 4.2 Distribution of collected records.

4.3.2 Description of the data collected

INSIS database was used to collect data related with the motor vehicles. The database contains more than 1000 tables. For our propose we have selected motor insurance claim and underwriting tables. These tables are CAR_TYPE, CAR_BODY, OCP1, MAN_COMP, CAR_USAGE, O_CAR,CLAIM_OBJECT. These attributes with their description, and data type is presented in table 4.3 -4.5.

s.n	Attribute Name	Data type	Description
1	CAR_TYPE	Nominal	The type of car
2	MODEL	String	Model of the vehicle
3	MAKE	string	Make of the vehicle
4	CAR_USAGE	Nominal	Use of the car Nominal [2-14]
5	SEATS_NUM	Numeric	Number of seats
6	PROD_YEAR	Numeric	Production year
7	CAR_AGE	Nominal	The age of the car
8	BODY_TYPE	Nominal	Body type of the car
9	ENGINE_POWER	Numeric	Engine power of the vehicle
10	OCP1	Nominal	Legal entity Nominal [1,2]
11	OCP2	Nominal	Terterial extension of the vehicle
12	CCM_TON	Numeric	Horse power of the vehicle

Table 4.3 Description of O_CAR table attributes and data type

The O_CAR table 4.3 originally contains 36 attributes but after discussion with the domain expert the above 12 attributes were selected initially.

s.n	Attribute Name	Data type	Description
1	CLAIM_ID	String	Claim ID of the vehicle
2	INSURANCE_TYPE	Numeric	Type of insurance
3	COVER_TYPE	String	The type of cover for the insured vehicle
4	RISK_TYPE	String	The risk type of the vehicle
5	LOSS_TYPE	String	The type of loss
6	INITIAL_RESERV_AMNT	Numeric	The initial amount reserved for the claim
7	LAST_RESERV_AMNT	Numeric	The last reserved amount for the claim

Table 4.4 Description of CLAIM OBJECTS table

Table 4.4 originally contained around 46 attributes, after discussion with the domain expert the numbers of attributes is reduced to 7.

s.n	Value	Description
1	2	MOTOR_CYCLE
2	3	TRUCK
3	4	TRACKTOR
4	5	TRACTOR (AUTHORIZED TO USE PUBLIC ROADS)
5	6	MOTORBIKE WITH ATTACH
6	7	BUS
7	8	TRAILERS AND SEMITRAILERS
8	10	TRADE PLATED
9	11	SPECIAL CONSUTRUCTION
10	12	TRANKER
11	13	STATION WAGONES
12	14	PICK-UP

Table 4.5 Descriptions of CAR_USAGE attribute

Table 4.5 is derived from attribute of CAR_USAGE which was under table O_CAR. It contains description of the values of the car type attributes in O_CAR table.

s.n	Attribute Name	Description
1	1	OTHERS
2	2	LEARNER
3	3	TAXI
4	4	CAR HIRE
5	5	SPECIAL CONSTRUCTION
6	6	SEAZONAL USAGE

8	8	FARE PAYING PASSENGERS
9	9	OWN GOODS
10	10	GENERAL CARTAGE
11	11	AGRICULTURAL ANY FARM
12	12	AMBULANCE
13	14	OWN SERVICE
14	15	COMMERCIAL VEHICEL
15	16	FIRE FIGHTING
16	20	OTHER OLD

Table 4.6 Description of CAR_TYPE attribute

Table 4.6 is derived from attribute of CAR_TYPE which was under table O_CAR. It contains description of the value of the CAR_TYPE attributes in O_CAR table.

s.n	**VALUE**	**Description**
1	1	COMMECA MEMBER
2	2	NON COMMRCIAL MEMBER

Table 4.7 Description of OCP1 attributes

Table 4.7 is derived from attribute of OCP1 which was under table O_CAR. It contains description of the value of the territorial extension attributes in O_CAR table.

s.n	**Attribute Name**	**Description**
1	1	PHYSICAL ENTITY
2	2	LEGAL ENTITYT
3	0	N/A

Table 4.8 Description of OCP2 attribute

Table 4.8 is derived from attribute of OCP2 which was under table O_CAR. It contains description of the value of the ownership attributes in O_CAR table.

4.3.3 Data Quality Assurance

The collected data contain incomplete, missing values, outliers, and irrelevant data. On the other hand some important attributes are not available in the database such as driver's age, marital status, health condition, gender, driving experience. Therefore in order to solve the above problems especially on the drivers' details it was mandatory to access the claim files. Accordingly around 5200 claim files were referred in order to collect driver's details.

4.4 Preparation of the data

Data preparation is a fundamental stage of data analysis. There are a number of different processes applied in the available data in order to come up with the working dataset. Some of the processes applied in the data are data cleaning, data construction, data integration, data formatting, and attribute selection.

4.4.1 Data Selection

At the beginning the collected data from the INSIS database are around13, 000 and 5223 records collected manually form claim files. But of the collected dataset irrelevant and unnecessary data are available; therefore all the data were not used for training purpose. After the elimination of irrelevant and unnecessary attributes we had 5176 dataset ready for this study.

Previously described tables of the INSIS database contain around 22 attributes. From the available attribute there are attributes which were irrelevant and need to be removed form the attribute list. The following attributes are selected initially and prepared for further preprocessing, CLAIM_NO,CLAIM_ID, INSR_TYPE, COVER_TYPE, RISK_TYPE, PAID, CAR_TYPE, CAR_TYPE, CAR_USAGE, CARBODY_TYPE, CCM_TON, MAKE, MODEL, SEATS_NUM, OCP1,OCP2,MAN_COMP,INITIAL_RESERVED_AMOUNT,LAST_RESERVED_AMOUNT, CCM_TON.

4.4.2 Data Cleaning

For The purpose of data cleaning this research uses MS_Excel. As we discussed previously all the necessary data for this research were collected in two ways i.e details about the vehicle form INSIS database and details about drivers manually from claim files. Therefore, removing incomplete and missing attributes/values were applied on vehicle attributes and values.

The values of LAST_RESERVED_AMOUNT AND INITIAL_RESERVED_AMOUNT column had the same value, therefore INITIAL_RESERVED_AMOUNT is deleted. Some of the collected data attributes are unrelated, different record to the problem at hand. Some of these attributes are MODEL, SEATS_NUM, ENGINE_POWER from O_CAR attribute CLAIM_ID from CALIM OBJECT attribute. OCP1 and OCP2 attributes from O_CAR table, contain missing values and difficult to estimate. Therefore, after discussion with domain experts the attributes are deleted from the dataset. Drivers' age attribute contains outlier values. Therefore, drivers' age greater than 75 years old considered as outliers and deleted from the dataset.

4.4.3 Data Construction

Data construction is another important step in data preparation. In this stage new fields were derived from the existing ones. This is done due to the fact that derived fields were found important for risk assessment process. After discussion with domain experts at eastern main branch of EIC, the following fields were derived. AGE of the vehicle is derived from PROD_YEAR and similarly AGE of the driver attribute derived from DATE_BIRTH and DRIVE_YEAR_ EXP attribute derived from data of issuance of the driving license attribute.

4.4.4 Data Integration

The data collection process has been done in two different sources. As we discussed in the previous section details about the vehicle were collected from the INSIS database and details about derivers were collected manually from the claim files. Therefore manually collected data must be integrated in to the respective vehicle details. Claim number attribute was available in both manually collected data and data from database files. From manually collected 5223 records around 47 records were not found in the collected records from the INSIS database. Therefore, these 47 records are deleted from the dataset. The remaining 5176 manually collected records are integrated into the dataset from INSIS database.

4.4.5 Data Transformation

According to Malik et al. (2012) in data transformation the data are transformed or consolidated into forms appropriate for mining. Discretization is the process of transforming continuous space valued series x=[x1, x1,x….xn} into a discrete valued series Y={y1,y2,…..yn}. Major discretization methods which are used can be majorly categorized as unsupervised and supervised discretization. This paper implemented unsupervised discritization method called binning. Binning method is implemented due to the fact that it is the simplest methods to discretize a continuous-valued attributed by creating a specified number of bins.

There are attributes which has number value and need to be grouped. These attributes are PROD_YEAR, DATE_BIRTH, DRIVER_EXP, CCM_TON (HOURS POWER), REPAIR_COST. The derived attribute “AGE” contains numerical values, after discussion with domain expert the AGE attribute is categorized into three group i.e AGE_GROUP_FIRST, AGE_GROUP_SECOND, AGE_GROUP_THIRD. This is done using equal-width (Distance) partitioning method which is an example of simple discretization binning.

Accoding to Chaudhari et al (2014) Equal width Discretization is a simples discretization methods that divides the range of observed values for a feature into k equal sized bin. The process involves finding values as the minimum (Min) and maximum (Max). The interval is computed by dividing the range of observed values for the variable into k number of equally sized bins using the formula Int= (max-min/k). From the dataset age attributes contains maximum value of 72 and min value of 20. Prviously we decided that our value of k is 3 (AGE_GROUP_FIRST, AGE_GROUP_SECOND, AGE_GROUP_THIRD)

- Bin 1 AGE_GROUP_FIRST [20,37] The value in the dataset with this range replaced by AGE_GROUP_FIRST
- Bin 2 AGE_GROUP_SECOND [38,53] The value in the dataset with this range replaced by AGE_GROUP_SECOND
- Bin 3 AGE_GROUP_THIRD [54,+) The value in the dataset with this range replaced by AGE_GROUP_THIRD

For the other nominal attributes the threshold value set by consulting the domain experts. Threshold value applied for attribute PROD_YEAR to group the attribute as VEH_AGE_GROUP_ONE, VEH_AGE_GROUP_TWO, VEH_AGE_GROUP_THREE, for attribute DRIVER_EXP the groups are DRV_EXP_GROUP_ONE, DRV_EXP_GROUP_TWO, and DRV_EXP_GROUP_THREE. CC (HOURS POWER) attribute the groups are CC_GROUP_ONE, CC_GROUP_TWO, CC_GROUP_THREE. For REPAIR_COST attribute the groups are REP_COST_GROUP_ONE, REP_COST_GROUP_TWO, REP_COST_GROUP_THREE.

4.4.6 Data Formatting

This study implemented weka as a data mining tool. The available dataset should be prepared in a format and data type which is suitable for weka. Weka data mining tool accept ARFF (Attribute-Relation File Format). This format contained attribute values whose values are separated by comma. The file extension for the file format ARFF is arff.

Initially the data collected from the original database exported to an excel file format. Then the excel file format converted to common delimiter (csv) file format. Next the CSV file format opened with weka and then saved with arff file extension.

4.4.7 Attribute Selection

There are many reasons that make the number of attributes to have a significant decrease from the original collection. Some of the reasons are, it uses to speed up the learning process, it makes simple to understand the generated rules. Due to the above reasons selecting attribute for measuring the risk level is mandatory. The above selected attribute run on weka using GainRatioAtttibuteEval. Based on the gain ration value and discussion with the domain expert at western main branch the following final lists of attributes were selected.

s.n	Attributes Name	Data Type	Description	Remark
1	INSR_TYPE	Number	The type of insurance policy issued. i.e. whether commercial or private	
2	COVER_TYPE	Varchar	The types of cover underwrite.	
3	RISK_TYPE	Varchar	The type of risk the policy covered	
4	REPR_COST	Nominal	The actual repair cost	Drived. The original attribute was final reserved amount
5	CAR_USAGE	Number	The type of vehicle use.	
6	CCM_TON	Number	Cubic capacity of the vehicle	
7	MAKE	Varchar 2	Make of the vehicle	
8	DRVR_SEX	Nominal	Drivers' sex	
9	DRVR_AGE	Nominal	Drivers' age	Derived, Originally it was date of birth
10	DRVR_EXPERNC	Nominal	Drivers' driving experiance	Derived
11	MAN_COMP	Number	Territorial extension	
12	VHCL_PROD_YEAR	Number	Year of vehicle manufactured	Derived

Table 4.9 Lists of attributes with description and data type

CHAPTER FIVE

EXPERIMENTATION

This chapter describes the hybrid model data mining process. First the given dataset is segmented into four different clusters. Then the best cluster is selected as a model and used as an input for the selected classification algorithms.

This chapter also describes different evaluation and performance measurement techniques in order to select the best model by applying the selected clustering and classification algorithms.

For clustering purpose the simple k-means clustering algorithm is selected, in our discussion in chapter three the value of k is known. The output model of k-means clustering is used as an input of j48 classification algorithm and Multilayerperceptron neural network classification algorithms. Finally the output of the classification algorithm create predictive model.

5.1 Experiment Design

Before any model is built, there has to be a mechanism in order to test the validity. To measure the goodness of the cluster there are different procedure we should follow. These include intra cluster similarity measure (sum of square error value), number of iteration and the domain experts'. For training and testing the classification model, the 10 fold cross validation and split test options are used. In order to test the prediction performance of the classification model we selected a separate 1035 records from the original dataset by using random sampling technique.

5.2 Cluster Modeling

In this research most of the original collected data contain numerical values. In order to build the cluster model these numerical attribute values are changed to categorical by using threshold values. The threshold values for each numerical data are set after a discussion with domain experts. Table 5.1 Show the threshold values for attributes which are suggested by the domain experts at eastern main branch.

DRVR_AGE (DA)	**VHCL_AGE (VA)**	**DRVR_EXP (DE)**	**CC_TON (CT)**	**CLM_PAID (CP)**
20<=DA<=37 Young	DA<=5 New	DE<=5 INEXPERIANCE	CT<=1500 Small	CP<=3000 Small
38 <=DA<=53 Old	6<=DA<=10 AVERAGE	6<=DE<=10 AVERAGE	1500<=CT<=4000 Pick_up	3001<=CP<= 10000 Medium
54<=DA<=73 Very old	11<=DA<=20 Old	11<=DE<=20 EXPERIANCED	4001<=CT<=6000 Station Wagon	10001<=CP<= 50000 High
	DA>20 Very Old	DE>20 Very EXPERIANCED	6001<=CT<=9000 Bus	CP>=50001 Very High
			CT>=9001 Track	

Table 5.1 List of attributes with threshold values

In order to develop the cluster model, three different experiments are conducted. These three experiments are created by changing the default parameter of the simple k-means algorithm. The following section describe the three experiment conducted and the selected final cluster model of the study.

5.2.1. Experimentation 1

The first experiment run on simple k means clustering algorithm with k=4, with default seed value and default distance function (Euclidean distance). In this experiment the research implemented the previous selected attributes with 5176 records. The training set cluster mode is used in the dataset.

Table 5.2 shows the result of the first experiment by dividing the dataset in to four clusters.

Number of Cluster	**Distance**	**Seed Value**	**Cluster Distribution**			
			C1	**C2**	**C3**	**C4**
4	Euclidean Distance	10	1846(36%)	854(16%)	1245(24%)	1231(24%)

Fig 5.2 Cluster experiment with k=4 with default seed value and default Distance function

In addition to the above result k means clustering algorithm identify attributes, value of attribute, sum of square errors, number of iterations and cluster value. Table 5.3 presented details description about the value in each cluster segments.

Cluster	Centroid	Rank	Level of risk
1	Commercial vehicle, comprehensive, own damage, Toyota, M, [20, 37), [1,5],[6,10], Pick-up, (1500,4000], (10000,50000], Own Goods.	2	High risk
2	Private vehicle, comprehensive, own damage, Toyota, M, [37,53), [6,10],[6,10],Station Wagon, [4000,6000],(10000,50000], Private Use	4	Low risk
3	Private vehicle, comprehensive, own damage, Toyota, M, [20,37), [1,5],[20,47),Small Vehicle, [0,1500],[10000,50000], private use	3	Medium risk
4	Commercial vehicle, comprehensive, own damage, Isuzu ,M, [20,37), [1,5],[6,10],Track, [9000,14000],[50000,2000000], General Cartage	1	Very high risk

Table 5.3 Details attribute value of each cluster with k=4, default seed value and default distance function.

The first cluster indicate that the type of insurance is Commercial, the cover type is comprehensive, the risk type is own damage, make of the vehicle Toyota, driver's sex male, driver's age between 20 and 37, driver's experience less than five years, vehicle type Pick-up, Horse power of the vehicle between 1500 and 4000, repair cost between 10000 and 50000 and purpose of the vehicles own damage. After discussion with domain experts cluster 1 ranked 2 with risk level of high.

The second cluster attributes are characterized by the following unique feature i.e. type of insurance is Private, the cover type is comprehensive, the risk type is own damage, make of the vehicle Toyota, driver's sex male, driver's age between 37 and 53, driver's experience between 6 and 10 years, vehicle type Station wagon, Horse power of the vehicle between 4000 and 6000, repair cost

between 10000 and 50000 and purpose of the vehicles own damage. Based on the above features of cluster 2 and discussion with domain experts, this cluster ranked 4 with risk level of low.

The third cluster contains attribute value with Type of insurance is Private, the cover type is Comprehensive, the risk type is own damage, make of the vehicle Toyota, driver's sex male, driver's age between 20 and 37, driver's experience between 1 and 5 years, vehicle type small (Automobile), Horse power of the vehicle is less than 1500, repair cost between 10000 and 50000 and purpose of the vehicles private. After consulting the domain experts, cluster 3 ranked 3 with risk level of medium.

The attribute value of cluster four indicate that the Type of insurance is Commercial, the cover type is comprehensive, the risk type is own damage, make of the vehicle Isuzu, driver's sex male, driver's age between 20 and 37, driver's experience less than 5 years, vehicle type heavy Track, Horse power of the vehicle is greater than 9000, repair cost is greater than 50,000.00 and purpose of the vehicles own damage. According to the domain experts, cluster 4 ranked 1 with risk level of very high

Generally, in the first experiment our parameter values are set to default value except the value of k. Therefore in order to understand the dataset it is better to come up with Parameter values other than the default values.

5.2.2 Experimentation 2

The second experiment is run on simple k means clustering algorithm with k=4, with seed value=100 and default distance function (Euclidean distance). In this experiment the research implemented the previously selected attributes with 5176 records.

Table 5.4 shows the result of the second experiment by dividing the dataset in to four clusters.

Number of cluster	Distance	Seed Value	Cluster Distribution			
			C1	C2	C3	C4
4	Euclidean Distance	100	1126(22%)	990(19%)	1138(22%)	1922(37%)

Fig 5.4 Cluster experiment with k=4, seed=100 and default distance function

The second experiment has the same number of clusters with the first experiment. This experiment also has similar segment formation with the first experiment result. Table 5.5 shows detail description of the value of the second experiment.

Cluster	Centroid	Rank	Risk level
1	Private vehicle, comprehensive, own damage, Toyota, M, [20, 37), [1,5],[6,10], Small(Automobile), (1500,4000], (10000,50000], Private Use.	3	Medium
2	Private vehicle, comprehensive, own damage, Toyota , M, [37,53), [6,10],[20,47],Small (Automobile) ,[0,1500],(10000,50000], Private Use	4	Low
3	Commercial vehicle, comprehensive, own damage, ISUZU ,M, [20,37), [1,5],[6,10],Track,[4000,6000],[50000, 250000000], General Cartage	1	Very high
4	Commercial vehicle, comprehensive, own damage, Toyota ,M, [37,53), [1,5],[6,10],Pick-up, (1500,4000],[10000,50000], Own Goods	2	High

Table 5.5 Details attribute value of each cluster with k=4, seed value=100 and default distance function (Euclidean distance function).

In this experiment the first cluster is characterized by the following value of attributes, type of insurance is Private, the cover type is comprehensive, the risk type is own damage, make of the vehicle Toyota, driver's sex male, driver's age between 20 and 37, driver's experience less than five years, age of the vehicle between 6 and 10 years, vehicle type Small (Automobile), Horse power of the vehicle is between 1500 and 4000, repair cost between 10000 and 50000 and purpose of the vehicles Private Use. After a discussion with domain experts, cluster 1 ranked 3rd with Medium level of risk.

The second cluster contains attributes value with the type of insurance is Private, the cover type is comprehensive, the risk type is own damage, make of the vehicle Toyota, driver's sex male, driver's age between 38 and 53, driver experience between 20 and 47 years,, vehicle type Small (Automobile), Horse power of the vehicle less than 1500, repair cost between 10000 and 50000 and purpose of the vehicles Private Use. The domain expert suggested that cluster 2 ranked 4th with Low level of risk.

The third cluster uniquely identified by attribute value of type of insurance is Commercial, the cover type is comprehensive, the risk type is own damage, make of the vehicle Isuzu, driver's sex male, driver's age between 20 and 37, driver's experience less than 5 years, age of the vehicle between 6 and 10, vehicle type heavy Track, Horse power of the vehicle is between 4000 and 6000, repair cost is between 10000 and 50,000.00 and purpose of the vehicles General Cartage. Based on the above features, cluster 3 ranked 1st with very high level of risk.

The fourth cluster attributes are characterized by the following unique feature i.e. type of insurance is Commercial, the cover type is comprehensive, the risk type is own damage, make of the vehicle Toyota, driver's sex male, driver's age between 37 and 53, driver experience less than 5 years, Age of the vehicle between 6 and 10, vehicle type pick-up, Horse power of the vehicle between 1500 and 4000, repair cost between 10000 and 50000 and purpose of the

vehicles own damage. The domain expert suggested this cluster to be ranked 2nd with high level of risk.

Generally, the result of the second experiment shows that there are 5 iteration with 19028 sum of square errors. As compared to the first experimentation the number of iteration decrease by one but sum of squarer errors value increase form 18246 to 19028.

5.2.3 Experimentation 3

The third and the final clustering experiment is run with k= 4, seed value= 1000 and Manhattan distance measurement function. Similar with the above experiment, this experiment also apply the final selected attribute with 5176 records. Table 5.6 shows the result of the third experiment.

Similar to the above experiment the third experiment also identifies attribute values that belong to the available cluster. Table 5.6 show details about attribute value of each cluster.

Number of cluster	**Distance Function**	**Seed Value**	**Cluster Distribution**			
			C1	**C2**	**C3**	**C4**
4	Manhattan Distance	1000	1410(27%)	1246(24%)	1527(30%)	993(19%)

Fig 5.6 Cluster experiment with k=4, seed=1000 and Manhattan Distance

We can see from the previous experiment as the default seed value changed the clusters percentage, Number of iteration, sum of square error also changed. The third experiment also supports this idea. Table 5.7 show the value of attributes distribution in each cluster.

Cluster	Centroid	Rank	Risk level
1	Commercial vehicle, comprehensive, own damage, ISUZU, M, [20, 37), [1, 5], [6, 10], Track, (6000, 9000], (50000, 2500000], General Cartage.	1	Very high
2	Commercial, comprehensive, own damage, Toyota, M, [38,53), [6,10],[6,10],pick-up, (1500,4000],(3000,10000], Own Goods	4	Low risk
3	Private, comprehensive, own damage, ISUZU ,M, [20,37), [1,5],[6,10],Small (Automobile), [0,1500],[10000, 50000], Private Use	2	High risk
4	Commercial vehicle, comprehensive, own damage, Toyota ,M, [20,37), [1,5],[6,10],Pick-up, (1500,4000],[10000,50000], Own Goods	3	Medium risk

Table 5.7 Details attributes value of each cluster with k=4, seed value=1000 and distance function (Manhattan distance function).

In this experiment the first cluster is characterized by the following value of attributes, type of insurance is Commercial, the cover type is comprehensive, the risk type is own damage, make of the vehicle isuzu, driver's sex male, driver's age between 20 and 37, driver's experience less than five years, age of the vehicle between 6 and 10 years, vehicle type Truck, Horse power of the vehicle is between 6000and 9000, repair cost between 50000 and 2500000 and purpose of the vehicles General Cartage. Based on a discussion with domain experts, cluster 1 ranked 1st with very high risk level.

The second cluster contains attributes value with the type of insurance is Commercial, the cover type is comprehensive, the risk type is own damage, make of the vehicle Toyota, driver's sex male, driver's age between 38 and 53, driver's experience between 6 to 10 years, vehicle age between 6 to 10, vehicle

type pick-up, Horse power of the vehicle between 1500 to 4000, repair cost between 3000 and 10000 and purpose of the vehicles Own Goods. The domain experts suggested that cluster 2 ranked 4th with low level of risk.

The third cluster uniquely identified by attribute value of type of insurance is Private, the cover type is comprehensive, the risk type is own damage, make of the vehicle Isuzu, driver's sex male, driver's age between 20 and 37, driver's experience less than 5 years, age of the vehicle between 6 and 10, vehicle type small (Automobile), Horse power of the vehicle is less than 1500, repair cost is between 10000 and 50,000.00 and purpose of the vehicles Private use. Based on the above features the domain experts suggested that cluster 3 ranked 2 with high level of risk.

The fourth cluster attributes are characterized by the following unique feature i.e. type of insurance is Commercial, the cover type is comprehensive, the risk type is own damage, make of the vehicle Toyota, driver's sex male, driver's age between 20 and 37, driver's experience less than 5 years, Age of the vehicle between 6 and 10, vehicle type pick-up, Horse power of the vehicle between 1500 and 4000, repair cost between 10000 and 50000 and purpose of the vehicles own Goods. The domain experts suggested that the 4th cluster rank to be 3 with medium level of risk.

The third experiment indicates that the number of iteration is 6 and sum of square errors become 18807. As compare to the first and second experiment, this experiment has equal number of iteration with the first experiment and it shows an increment from the second experiment. The sum of square errors value also lay between the first and the second experiment.

5.2.4 Selecting the Best Clustering Model

There are different selections criteria employed in order to select the best clustering model from the above developed models. According to Rokach and Maimon (2005) a good clustering model can be measured in two categories as

internal and external. Internal quality metrics usually measure the compactness of the clusters using some similarity measure. It usually measures the intra-cluster homogeneity, the inter-cluster separability or a combination of these two. Examples of internal quality criterions are sum of squared error (SSE), minimum variance Criteria, Scatter criteria and Edge Cut Metrics. External quality criteria are useful for examining whether the structure of the clusters match to some predefined classification of the instances. Example of external quality criterions are Mutual Information Based Measure, precision Recall measure and Rand Index. In General, this research selected the best clustering model base on the following criterions.

I) Cluster sum of square errors (SSE). It measure the goodness of each cluster by identify how tight the cluster members.

II) The Number of Iteration. The number of iteration that weka used in order to converge the dataset in to clusters.

III) The Domain Experts' Judgment. The domain experts play a vital role in identifying the similarity with in a cluster and dissimilarity between clusters.

Experimentation	**Number of Iteration**	**Cluster Sum of square Error(SSE)**
First	6	18246
Second	5	19028
Third	6	18807

Table 5.8 The result of number of iteration and sum of square errors

The first experiment indicates that the number of iteration is equal with the third experiment and less by one with the second experiment. On the other hand the first experimentation Sum of square errors shows the least as compare to the other two. This indicates that the second experiment is good in creating dissimilar clusters. The other important point in selecting the best

cluster is the domain expert suggestion. According to the domain experts the model created by the first cluster is good in segmenting the claim record as compare to the other two clustering. Due to the above reasons the first model created by the first experiment selected as the final clustering model.

5.3 Classification Modeling

There are different data mining classification algorithms available depending on the type of data mining tools used. In our pervious discussion on chapter one this research selected weka data mining version 3.7.4 as a data mining tools. As we have seen in the previous section of this chapter the clustering model has been created using simple k-means clustering algorithm. The result of clustering model taken as an input for the selected classification algorithms i.e. J48 decision tree classification algorithm and Multilayerperceptron neural network. Appling classification algorithm on the given dataset helps to classify new instance of insurance claim records into specified classes.

There are different parameters weka classification algorithms to be set for the development of the classification model. For developing a decision tree classification model we have 10 fold cross validation and percentage split classification modes. For neural network classification the experiment has been done by employing 10 fold cross validation. The classification and performance of each of these models are compared in order to come up with the best classification model.

5.3.1 J48 decision tree model building

The clustering model created by k-means clustering algorithm is used as an input to develop j48 decision tree model. From the available attributes used in clustering algorithm, 11 of them are used as independent variables and class label attributes is used as dependent attribute.

There are default parameters which need to be changed in order to get alternate j48 models. The next section discusses three alternate experiments which were produced by changing the default parameter value of the j48 classification algorithm. The following table describes the default parameters used by j48 algorithm in weka.

Parameter	Description	Default Value
ConfidenceFactor	The value that show the confidence factor used for pruning	0.25
minNumObj	The minimum number of instance per leaf	2
Unpruned	True/false true value pruning is not performed and false, pruning performed	False

Table 5.9 some of default values of j48 decision tree

5.3.1.1 Experimentation 1

The first experiment employed the default parameter of j48 algorithm. Some of the default parameters used in this experiment is confidenceFactor, minNumObj, Unpruned and 10 fold cross validation. Using these default parameters the classification model is developed with j48 decision tree which contains 210 numbers of leaves and 278 size of the tree. Table 5.10 Show the result confusion matrix of the model.

Actual	Predicted				Total	Correctly Classified
	Cluster 1	Cluster 2	Cluster 3	Cluster 4		
Cluster 1	1812	8	5	21	1846	98.12%
Cluster 2	8	770	27	49	854	89.09%
Cluster 3	7	22	1208	8	1245	96.94%
Cluster 4	33	24	8	1166	1231	94.43%
	1830	824	1248	1244	5176	94.63%

Table 5.10 the result of confusion matrix using j48 algorithm with default values

The above confusion matrix indicates that j48 algorithm scored an accuracy result of 94.63%. This result shows that out of the total training dataset of 5176, 94.63% of the records are correctly classified, whereas only 5.37% classified in wrong clusters. The accuracy result indicates that the algorithm is good in classifying the cluster model.

In addition to the above result, the confusion matrix also contains the recall result of each cluster. Cluster 1 shows an accuracy result of 98.12%, this indicate that out of the total 1846 record 1812 records are correctly classified, whereas only 33 records are wrongly classified. In cluster 2 the least accuracy percentage registered i.e out of the total 854 records 770 records are correctly classified, whereas 84 records are incorrectly classified. Cluster 3 contains a total of 1245 records, out of these records 1208 records are correctly classified, whereas 37 records are incorrectly classified. Out of the total record of 1201 in cluster 4, 1166 records are correctly classified, whereas 65 records are incorrectly classified.

In our previous discussion of this experiment, the total number of leave and the size of leave are 210 and 278 respectively. From these numbers we can

understand that it is difficult to traverse through the whole tree structure and come up with defined rules. Therefore in order to understand the tree structure and to generate interesting rule, the value of minNumObj parameter has to be changed. Accordingly in this experiment the value of minNumObj changed to different values i.e. 5, 10, 15, 20, 25, 30. Table 5.11discrib the result obtained with minNumObj =15.

Actual	**Predicted**				**Total**	**Correctly Classified**
	Cluster 1	Cluster 2	Cluster 3	Cluster 4		
Cluster 1	1771	42	1	32	1846	95.77%
Cluster 2	35	688	62	69	854	75.87%
Cluster 3	13	42	1185	5	1245	94.94%
Cluster 4	53	41	6	1131	1231	91.16%
	1830	824	1248	1244	5146	89.44%

Table 5.11 the result confusion matrix using j48 algorithm with minNumObj =15.

The result of this experimentation indicates that the number of leaves and size of tree shows significant decrease. The size of the tree decrease form 210 to 95 and the number of leaves decrease from 278 to 122. The result of the confusion matrix shows an accuracy result of 89.44%.. This result indicate that from the total 5146 records, 89.44% or 4603 records are correctly classified, whereas 10.56% or 573 records are incorrectly classified.

The above confusion matrix also shows that cluster 1 incurred an accuracy result of 95.77. From the total number of 1846 records, 42 records classified as cluster2, 1 records classified as cluster 3 and 32 records classified as cluster 4. Cluster 2 registered 75.87% accuracy results. From the total number of 854 records, 35 classified as cluster 1, 62 records classified as cluster and 69 records classified as cluster 4. Cluster 2 shows an accuracy result of 94.94%. In this cluster out of the total 1245 records, 13 records classified as cluster 1,

42 records classified as cluster 2 and 5 records classified as cluster 4. Cluster 4 show accuracy results of 91.16%. This result show that out of the total 1231 records, 53 records classified as cluster 1, 41 records classified as cluster 2 and 6 records classified as cluster 3.

Even though the size of the tree and the number of leaves shows significant decrease, the accuracy percentage decrease as compare to the pervious experiment. Therefore, the first experiment with default minNumObj parameter selected as j48 classification model.

5.3.1.2 Experimentation 2

This experiment implemented by changing the default value of the 10 fold cross validation. The other available test option in j48 decision tree is the percentage split. This experiment run on default value of percentage split (66%), percentage split (70%), percentage split (75%) and percentage split (80%). After the above value of percentage split experimented the one with percentage split 80% scored better accuracy. This percentage score indicate that 80% of the total dataset allocated to training data and 20% of the total dataset allocated for testing. Table 5.12 Show the confusion matrix of the percentage split test option.

Actual	**Predicted**				**Total**	**Correctly Classified**
	Cluster 1	Cluster 2	Cluster 3	Cluster 4		
Cluster 1	376	1	0	4	381	98.67%
Cluster 2	3	147	3	14	167	86.39%
Cluster 3	0	4	244	4	252	96.72%
Cluster 4	7	5	3	220	235	93.18%
	386	157	250	242	1035	93.74%

Table 5.12 the result confusion matrix using j48 algorithm with percentage split 80%

The above confusion matrix show that out of 5176 records 4140(80%) records are used for training whereas 1035 (20%) of the records are used for test purpose. The accuracy result of this experiment indicate that out of the total 1035 test records 970(93.74%) of the records classified correctly, whereas 65(6.26%) of the records are incorrectly classified.

The above confusion matrix not only shows the general accuracy level, but it also identified the accuracy level of each clusters. The accuracy of cluster 1 is 98.67, this result indicate that from the total 381 record 376 records are classified correctly whereas 5 records are incorrectly classified. The accuracy of cluster 2 is 86.39 %. This result indicates that from the total 167 records 147 records are correctly classified, where as 20 records are incorrectly classified. The accuracy level of cluster 3 is 96.72%. This result indicates that from the total 252 records 244 records are correctly classified, whereas 8 records are incorrectly classified. The accuracy level of the final cluster is 93.18%. This percentage indicates that out of the total 235 test records 220 records are correctly classified, whereas 15 records are incorrectly classified.

The other trials of this experiment contain an accuracy value of less than the percentage split 80%. This indicates that the accuracy level registered in experiment 1 is better than this experiment.

The above three experiment registered different accuracy result. From these accuracy results we can understand that as the default value of the parameter changed the distribution of records between classes changed. From the three classification model the better classification model is the one which is good in predicting new underwriting records in their correct class category. Accordingly experiment 1 with default parameter value and 10 fold cross validation test option registered better result. See appendix 4 tree generated from the model. Based on clustering result labeled dataset is prepared manually which is used for classification task (using artificial neural network and decision tree)

5.3.2 Artificial Neural Network (ANN) Classification Model Building

In order to apply artificial neural network, the collected data need to be normalized to the range [-1,1]. The normalized data is useful to crate the ANN model very fast. WEKA's pre-processing is employed in order to normalize the values of the data attributes to be in the range [-1, 1].

More than half of the attributes in this research contains categorical or nominal value. Therefore, the value has to be changed into numeric values for normalization. Table 5.12 Show how the categorical attributes are changed into numerical values.

Attribute	Value assigned
SEX	MALE=1 FEMALE=2
AGE_GROUP	AGE_GROUP_ONE=1,AGE_GROUP_TWO=2 AGE_GROUP_THREE= 3 AGE_GROUP_FOUR=4
EXP_GROUP	EXP_GROUP_ONE=1,EXP_GROUP_TWO=2, EXP_GROUP_THREE=3
VHCL_GROUP	VHCL_GROUP_ONE=1,VHCL_GROUP_TWO=2, VHCL_GROUP_THREE=3, VHCL_GROUP_FOUR=4
CAR_TYPE	SMALL_GROUP_ONE=1,STA_GROUP_FIVE=5 TRACKTOR_GROUP_THREE=3,TANKER_GROUP_SIX=6, PICK_GROUP_SEVEN=7,BUS_GROUP_FOUR=4, TRK_GROUP_TWO=2
CC_TON	CC_GROUP_ZERO=1 CC_GROUP_ONE=2 CC_GROUP_TWO=3 CC_GROUP_THREE=4,CC_GROUP_FOUR=5, CC_GROUP_FIVE=6
REP_COST	REP_GROUP_ONE=1,REP_GROUP_TWO=2, REP_GROUP_THREE=3, REP_GROUP_FOUR=4

CAR_USAGE	OWN_GOODS=1,OWN_SERVICES=2,FAIR_PAYING_PASSANGER=3 GENERAL_CARTAGE=4 PRIVATE_USE=5
COVER_TYPE	M_COMPREH=1, M_COMPTPL=2,m_LOSDAM_V=3,M_TPONLY=4, M_TPR_PROC=5,
INSR_TYPE	1201=1, 1202=2
RISK_TYPE	M_LOADAM_IV=1, TPL_PROP=2

Table 5.13 Representation of categorical attributes to numerical values

After the above mapping process has been completed, weka's pre-processing facility has been employed to normalize to the value suitable for multilayerperceptron neural network algorithm. All the attributes and value of attributes employed in decision tree are used in developing multilayerperceptron model. In order to produce different model, the experiments are implemented by changing the default value of the following multilayerperceptron parameters.

- HiddenLayer: One of multilayerperceptron parameter which accepts numerical values, but the default value of this parameter is 'a' which represent (number of classes + attributes)/2, in our case the value of 'a' is (12+4)/2= 8.
- LearningRate: The other important parameter of multiperceptron neural network is learningRate with default value of 0.3.

5.3.1 Expermentation 1

The first experiment is conducted with default parameter value. The 10-fold cross validation test option is used. Table 5.13 shows the distribution of all the records with the respective Cluster.

Records in each cluster	**Clusters**				**Total**
	Cluster1	**Cluster2**	**Cluster 3**	**Cluster4**	
Total Number of Records	630	1331	1098	2115	5174
Correctly classified records	626	1324	1098	2104	5152
Incorrectly classified records	4	7	0	11	22
Correctly Classified in % accuracy	99.37	99.47	100	99.48	99.58
Incorrectly classified %	0.63	0.53	0	0.52	0.42

Table 5.14 *summary result of experimentation 1 with default parameter values.*

The neural network result in the above table is built with 0.3 of learningRate and 8 hidden Layer which is the default value of multilayerperceptron parameter. In addition to the above default parameter, 10-fold cross validation option is selected as a test option. In General, from the total record 99.58 % of the records are correctly classified, whereas 0.42% of the total records are incorrectly classified. Out of the total 630 cluster 1 records, 626 records are classified correctly, whereas only 4 records are incorrectly classified. From the total 1331 cluster 2 records, 1324 records are correctly classified, whereas only 7 records are incorrectly classified. The highest level of accuracy registered in cluster three. In this cluster from the total record of 1098, all the records are

correctly classified. From the available 2115 records in cluster 4, 2104 records are correctly classified, whereas 11 records are incorrectly classified.

5.3.2 Experimentation 2

The second experiment is generated with learningRate of 0.4 and hiddenlayer of 8. The resulting experiment of the multilayerperceptron algorithm depicted in table 5.14 below.

Records in each cluster	**Clusters**				**Total/Average**
	Cluster1	**Cluster2**	**Cluster 3**	**Cluster4**	
Total Number of Records	630	1331	1098	2115	5174
Correctly classified records	625	1301	1098	2106	5130
Incorrectly classified records	5	30	0	9	44
Correctly Classified in % accuracy	99.21	97.75	100	99.48	99.13
Incorrectly classified %	0.79	2.25	0	0.43	0.87

Table 5.15 *summary result of experimentation 2 with learningRate=0.4 and default hidden layer*

The above table is derived from the confusion matrix of multilayerperceptron algorithm. Generally the accuracy level of this experiment is 99.13%. The experiment also shows the accuracy level of individual clusters. In cluster 1

from the total 630 records, 625 records are correctly classified, whereas 5 records are incorrectly classified. From the total 1331 records in cluster 2, 1301 records are correctly classified, whereas 30 records are incorrectly classified. From the total 1098 records in cluster 3, all the records are correctly classified. Out of the total 2115 records in cluster 4, 2106 records are correctly classified, whereas only 9 records are incorrectly classed.

5.3.3 Experimentation 3

MultilayerPerceptron algorithm of the neural network also experimented with default value of 0.3 learning rate and 10 hidden layers. Table 5.15 depicted details result of this experiment.

Records in each cluster	**Clusters**				**Total/Average**
	Cluster1	**Cluster2**	**Cluster 3**	**Cluster4**	
Total Number of Records	630	1331	1098	2115	5174
Correctly classified records	627	1311	1098	2103	5130
Incorrectly classified records	3	20	0	12	35
Correctly Classified in % accuracy	99.52	98.48	100	99.43	99.36
Incorrectly classified %	0.48	0.52	0	0.57	0.87

Table5.16 *summary result of experimentation 3 with default learning rate and 10 hidden layer*

The above table is derived from the confusion matrix of multilayerperceptron algorithm. The accuracy level of this experiment is 99.36%. The experiment also shows the accuracy level of individual clusters. In cluster 1 from the total 630 records, 627 records are correctly classified, whereas 5 records are incorrectly classified. From the total 1331 records in cluster 2, 1311 records are correctly classified, whereas 20 records are incorrectly classified. From the total 1098 records in cluster 3, all the records are correctly classified. Out of the total 2115 records in cluster 4, 2103 records are correctly classified, whereas only 12 records are incorrectly classed.

Generally, from the above experiment we can understand that neural network models built with multilayerperceptron registered better accuracy performance in both average and individual clusters. Among the three neural network models, the model created by the default parameter values and 10-fold cross-validation test option registered better accuracy performance.

5.3.3 Decision Tree and Neural Network models comparison

The research presented two classification models developed by j48 decision tree classification algorithm and multilayerperceprton Neural Network classification algorithm. Both models created by decision tree and neural network are built with 10-fold cross-validation test option and default parameter values.

The researcher together with the domain experts set criterions in order to select the best classification model. These criterions are the overall accuracy registered by the models, the accuracy level for High risk customers and the accuracy level of Very high risk customers. Accordingly, the following section describes in detail the two models in respect of the above criteria.

- The model created by decision tree classification algorithm registered an overall accuracy of 4898 (94.63%) which are correctly classified and 278 (5.37%) records were incorrectly classified. From the total 1244 very high risk customers, 1166 (94.43%) records classified correctly and 78 records

are incorrectly classified. The other selected criteria is High risk customer, according to the confusion matrix from the total 1830 high risk customer, 1812 (98.12%) records are classified correctly and 18(1.88%) records are incorrectly classified.

- The model created by Neural Network classification algorithm has an overall accuracy of 5130 (99.58%) which are correctly classified and 35 (5.37%) records are incorrectly classified. Out of the total 630 very high risk customer, 627 (99.52%) records classified correctly and only 3 (0.48%) records are incorrectly classified. The other selected criteria is High risk customer, according to the neural network confusion matrix from the total 1098 high risk customer, 1098 (100%) records are classified correctly.

The above analysis indicates that from the overall accuracy level decision tree registered 94.63% while neural network registered 99.58%. On the other hand from the high risk customer (cluster 1), decision tree registered an accuracy percentage of 94.43% while neural network registered 99.52%.

Generally, the data correctly classified by neural network is higher than the number registered by decision tree classification algorithm and the numbers of incorrectly classified by neural network are less than the number registered by decision tree classification algorithm. Therefore, the model created by neural network multiperepetron algorithm is better than decision tree classification algorithm.

5.4 Evaluation of the discovered knowledge

Usually data collected for the purpose of data mining contains different format. The values of the data might contain outliers, missing values,

inconsistent with in a single attributes, therefore such values need to be transformed, cleaned and integrated to the format suitable for data mining.

The model building process of this research includes a clustering model building using the k-means algorithm. Using k-means clustering different experiments are conducted to segment customers. From the experimented clusters, the cluster with k=4 and with default parameters values registered better sum of square errors as compared to the other three experiments.

Based on the above clustering model, classification models are built using weka 3.7.4. The research implemented the classification model j48 decision tree and neural network multilayerperceptron algorithm. The experiment is conducted with 10-fold cross validation and by splitting the data set into 80% training and 20% for testing. In addition to the above criteria different parameter values are implemented for both decision tree and neural network models. From the created models using j48 decision tree classification algorithms, the one with default parameter values registered better accuracy in the overall and individual cluster classification. Multilayerperceptron Neural network classification algorithms with default parameter values registered better accuracy performance in both individual and overall clusters. Finally, the two models generated by j48 decision tree and multilayerperceptron are compared. The accuracy level registered by neural network is better than decision tree classification in both individual and overall accuracy.

The above results indicate that applying data mining technology to solve problems in the insurance industry is very promising. Generally, the developed model clearly classified insurance customers as low risk, medium risk, high risk and very high risk.

5.5 Use of the discovered Knowledge

In this research an attempt was made to develop applicable prototype. The prototype called Motor Insurance Risk Assessment System (MIRAS). The system uses the model developed by j48 decision tree due to the fact that it is easily converted to if then statements. First the researcher together with the domain experts derived interesting rules from the above developed model and then the rules are coded using C#. Fig 5.1 Show the snapshoot of MIRAS.

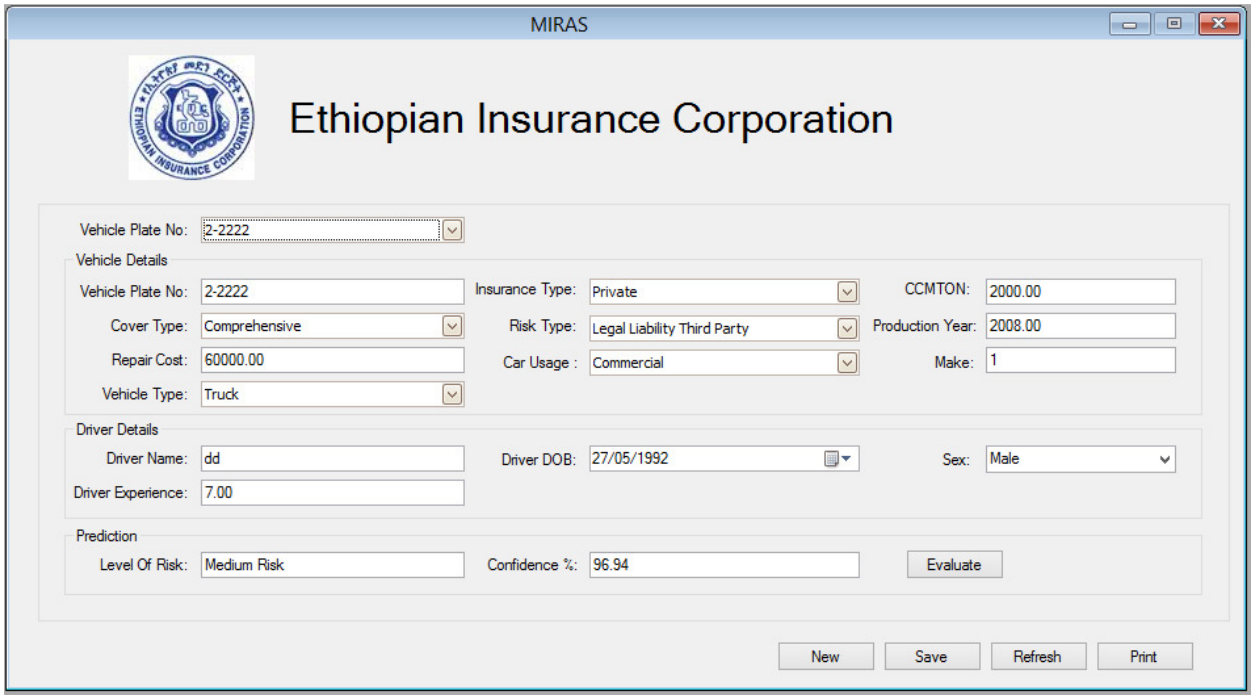
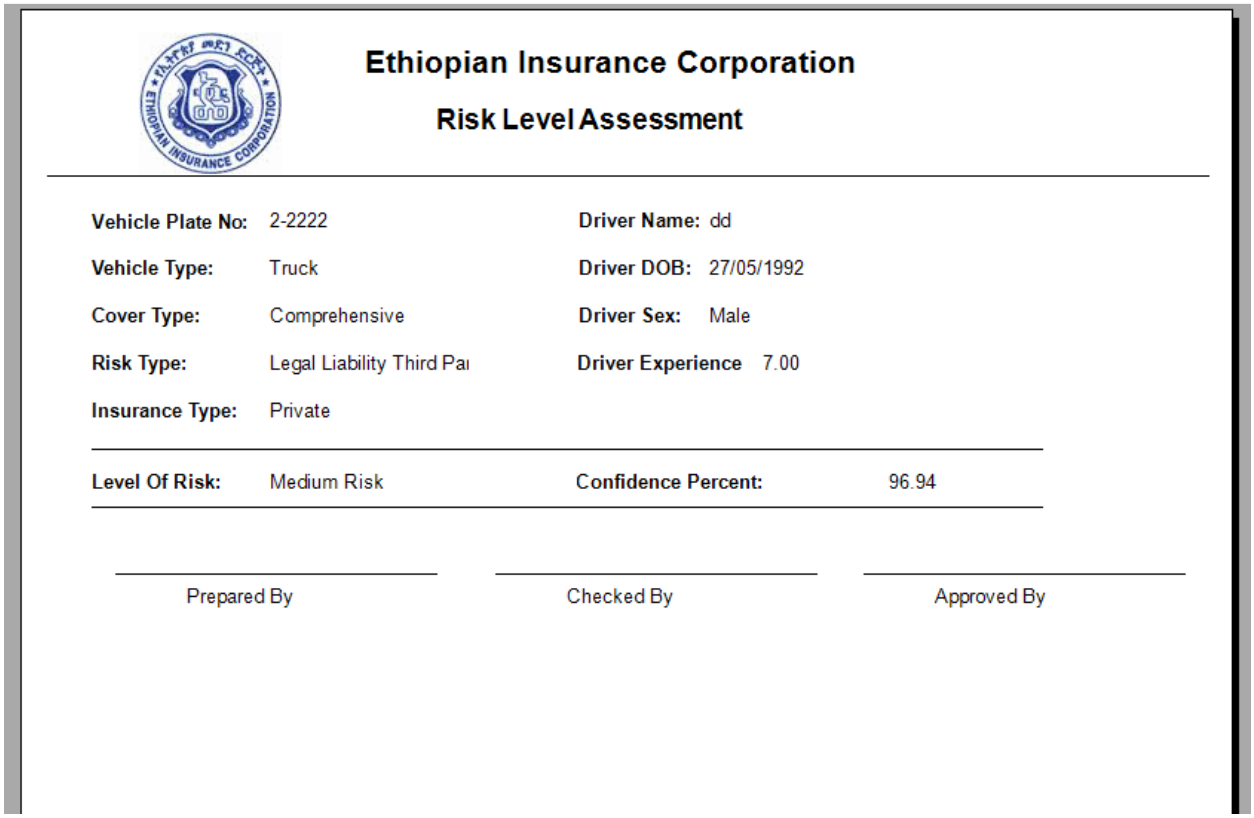


Fig 5.1 Home page of MIRAS

The system allows the underwriter to generate reports. The generated report has all the necessary attribute values which help the underwriter to calculate premium. Fig 5.2 shows snapshoot of a report.



Ethiopian Insurance Corporation

Risk Level Assessment

Vehicle Plate No:	2-2222	**Driver Name:**	dd
Vehicle Type:	Truck	**Driver DOB:**	27/05/1992
Cover Type:	Comprehensive	**Driver Sex:**	Male
Risk Type:	Legal Liability Third Pai	**Driver Experience**	7.00
Insurance Type:	Private		
Level Of Risk:	Medium Risk	**Confidence Percent:**	96.94

Prepared By | Checked By | Approved By

Fig 5.2 A Report that shows medium risk customers attribute values.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

Nowadays the application of data mining technology has become increasing in different sectors such as insurance, bank, airlines, telecommunication and other related industries. Insurance industry applies data mining technology for fraud detection, customer segmentation and risk assessment. The most challenging job in insurance underwriting is measuring the risk level of the subject matter insured.

In this research an attempt has been made to apply data mining in support of underwriting risk level measurement. The research implemented the six step Cios et al. (2000) process model. This process model contains six phases i.e. understanding the problem domain, understanding the data, preparation of the data, DM, evaluation of the discovered knowledge and using of the discovered knowledge.

The data collection process has been done in two phases. In the first phase the data is collected from INSIS database of EIC. The second phase of the data collection process has been conducted in five main branches of EIC. These main branches are North Western Main Branch, North Eastern Main Branch, Western Main Branch, Southern Main Branch, and Eastern Main Branch. This phase was the most challenging phase of the data mining. The researcher assigned two individuals in order to collect manually details about the drivers from the claim files of the above selected five main branches. On the collected data the researcher applies different data preprocessing techniques in order to assure the quality of the collected dataset. Those techniques that are applied in the collected dataset are, data cleaning, data construction, data integration, data transformation, data formatting and attribute selection.

The data mining tool is implemented in order to cluster the dataset in to four clusters and then based on the clustered model, classification algorithms are applied to classify the clustered model.

The raw data collected from both sources didn't identify the target classes for this research. Therefore, to identify the target classes, k-means clustering algorithm was implemented to segment the dataset as low risk, Medium risk, High risk and Very High risk customers. In order to come up with good clustering model, three different cluster experimentations have been implemented. The model with k=4, seed value =10 and Euclidean distance registered better segmentation. According to the segmentation, commercial vehicles were registered high level of risk than private vehicles and also young and inexperienced peoples registered higher level of risk than old and experienced.

The selected cluster model is used as an input for creating the classification models. In this regard two classification algorithms have been implemented i.e j48 decision tree and Multilayerperceptron neural network. Using j48 three different models were created by changing the default parameters of the algorithms. Based on the result, the model created by using 10-fold cross-validation test option with default parameter value registered a classification accuracy of 94.63%. On the other hand three experiments were conducted using Multilayerperceptron neural network. Accordingly, the one with 10-fold cross-validation test option to gather with default parameter values registered an accuracy level of 99.58%. Based on the above result neural network has registered better accuracy level than decision tree.

The findings result of the study indicate that in order to construct the predicative model, first the researcher to gather with domain experts have to identify attributes that best meet the required objectives. Accordingly the study identified attributes on both the drivers and vehicle side and the dataset prepared based on the selected attributes.

Finally, the working model that has been created using DM tool is coded in order to use the discovered knowledge. The developed prototype based on the discovered knowledge gives advice to the domain experts at the time of underwriting.

Generally, the result of this research implies that data mining technology is an important tool to classify insurance customer as low risk, medium risk, high risk and very high risk based on features on both drivers and vehicles side.

6.2 Recommendations

This research has created a model using data mining techniques to classify insurance customers' based on the risk level. The study contributes a lot for future researches in the area of insurance risk level measurement. Even though this research is conducted for an academic purpose, the findings indicated that there are interesting rules generated both on the drivers and vehicles side. The researchers believe that the interesting rules derived specially on the drivers' side will lead the insurance underwriters at the EIC to capture derivers' details and apply the rules to measure the risk level.

Based on the findings of this research, the following recommendations are given by the researcher.

- This research couldn't find records of the following attributes and did not consider attributes related with driver's marital status, Driver's level of education, place of accident, traffic police report and nature of the accident Future research can include the above attributer in order to discover new knowledge.
- The database of EIC contains many attributes related with vehicles' details, but no records are found in the database that is related with driver's age, driver's experience, and driver's sex. Due to this limitation the researcher forced to collect only 5176 records. As the above attributes are important attributes, the corporation has to add the

attributes as an addition field in the INSIS software so that underwriter at EIC used as risk level measurement for named based driver policy and researchers can apply more records than the records used in this research. EIC also issued non named base policy. Therefore, this study recommends underwriter at EIC to issues such policy by considering attributes from vehicle side.

- The model building process carried out in two phases. In the first phase cluster algorithm k-means implemented, and then for classification j48 decision tree and neural network multilayerperceptron algorithm implemented. Using these algorithms the classification results registered better accuracy. From the results we can understand that neural network has registered better result than decision tree classification algorithm. This indicate that further investigation needs to be conducted using different clustering and classification algorithms such as time serious, summarization and support vector machine.
- This research has implemented DM techniques to classified motor insurance customers based on the risk level. Different class of businesses are underwrite by insurance companies. Therefore researchers can also apply DM in other class of business other than motor insurance class of business.
- The quality of the data matters on the performance of the data mining algorithms. In this study much effort has made on data cleaning, integration and data transformation due to the fact that the corporation doesn't have data warehouse that contains all the important attributes. Therefore, EIC needs to have data warehouse to generate novel and interesting knowledge with minimum efforts.

References

Aman Kumar Sharma 2011 A comparative study of classification algorithms for spam email data analysis, International Journal on Computer Science and Engineering (IJCSE)

Amrender kumar 2004 Artificial neural networks for data mining I,A.S.R.I., Library Avenue,pusa, new delhi -110012. http://www.iasri.res.in/sscnars/data_mining/4-Artificial%20Neural%20Networks_Amrender.pdf. Access date Jan 20 2015.

Amit Ganatra , Y P Kosta 2, Gaurang Panchal , Chintan Gajjar 2011 Initial Classification Through Back Propagation In a Neural Network Following Optimization Through GA to Evaluate the Fitness of an Algorithm, International Journal of Computer Science and Information Technology(IJCSIT), Vol 3, No 1.

Andrea Dal Pozzolo 2010/2011 Comparison of data mining techniques for insurance claim prediction, Anno Accademico Sessione II.

Anteneh Fentahun 2011 Mining Road Traffic Accident Data For Predicting Accidental severity to improve public health-Role of Driver and Road Factors in the case of Addis Ababa. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Belete Biazen June 2011 Knowledge discovery for effective customer segmentations: The case of Ethiopian Revenue and customs Authority. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Belachew Reganie 2013 Application of data mining techniques for customers segmentation and prediction: Case of Buusaa Gonofa Microfinance Institution. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Cios, K.J.; Pedrycz, W.; Swiniarski, R.W.; Kurgan, L. (2007) Data Mining A Knowledge Discovery Approach.

Data mining with decision trees- Theory and Applications world Scientific Publishing Co. Pte Ltd. http://www.worldscibooks.com/compsci/6604.html Access Date Feb 26 2015

Derrig R. A., 2002 "Insurance fraud," Journal of Risk and Insurance, vol. 69.3, pp. 271-287.

D T Pham, S S Dimov, and C D Nguyen 2004 Manufacturing Engineering Centre, Cardiff University, Cardiff, UK, Proc.IMechE Vol.219 Part C:J Mechanical Engineering Science.

Fadzilah Siraj and Mansour Ali Abdoulha 2011 Mining Enrolment Data Using Predictive and Descriptive Approaches, College of Arts & Sciences, University Utara Malaysia, http://www.intechopen.com/books/knowledge-oriented-applications-in-datamining/mining-enrollment-data-using-descriptive-and-predictive-approaches, Feb 20 2015

Girma Aweke 2012 Predicting HIC infection risk factor using volantoru counseling and testing data. A case of Africa aids initiative initiative international (AAII). Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

G.ROSAN I.H. Witten and E. Frank. Data Mining: 2005 Practical machine learning tools and techniques. Morgan Kaufmann.

Han J. and Kamber. M. (2006) Data Mining Concepts and techniques. Morgan kuffmann publishers, San Francisco.

Helen Tefera Kidane 2003 Application of Data Mining Technology to Identify Significant patterns in Census or Survey Data: The case of 2001Child Labour Survey in Ethiopia. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Higgins J, 2005, the Radical Statistician, Prentice Hall Publishing.

Inna Kolyshkina, Richard Brookes 2002 Data mining approaches to modelling insurance risk. http://docs.salford-systems.com/insurance4211.pdf. Access date Feb 20 2015

Jay Gholap 2012 Performance tunining of j48 algorithm for prediction of soil fertility, Dept. of Computer Engineering college of Engineering, Pune, Maharashtra, India. Asian Journal of computer science and Informaion technology vol.(2):L8 P251-252.

Jasdeep Singh Malik, Prachi Goyal, Mr. Akhilesh K Sharman 2012 A Comprehensive Approach towards Data Preprocessing Techniques and association Rules. International Journal of Emerging Technology and advance engineering Technology Volum 4 issue 10.

John R. Davies, Stephen V. Coggeshall, Roger D. Jones, and Daniel Schutzer, "Intelligent Security Systems," 1987 in Freedman, Roy S., Flein, Robert. http://samples.sainsburysebooks.co.uk/9781743045299_sample_137778.pdf. Access Date Feb 20, 2015

Judy Feldman Anderson, FSA and Robert L. Brown, FSA (2005) Risk and Insurance, Society of Actuaries.

K. Umamaheswari, Dr. S. Janakiraman Internation 2014 Role of Data mining in Insurance Industry International Journal of advance computer technology (compusoft)

Kuo-Chung Lin , Ching-Long Yeh," 2012 Use of Data Mining Techniques to Detect Medical Fraud in Health Insurance", International Journal of Engineering and Technology Innovation, vol. 2, no. 2, pp. 126-137

Laveena Sehgal, Neeraj Mohan, and Dr. Parvinder S. Sandhu (ICCEMT'2012) September 8-9, 2012 Bangkok (Thailand). Quality Prediction of Function Based Software Using Decision Tree Approach International Conference on Computer Engineering and Multimedia Technologies.

Lionel Macedo Sep 2009 The Role of the underwriter in Insurance The International Bank for Reconstruction and Development/ The world bank1818 H Street, NW www.wpr;dbank.org/nbfi Access date Feb 15 2015.

Lior Rokach and Oded Maimon 2005 Data Mining and knowledge discovery Chapter 15 Clustering methods hand book Department of Industrial Engineering .

Lior Rokach and Oded Maimon 2005 Data Mining and Knowledge discovery Hand book Page 166

Lijia Guo. PH.D., A.S.A university of Florida Causality actuarial society Enterprise risk management committee (2003).
Applying data mining techniques in casual/property insurance. https://www.casact.org/pubs/forum/03wforum/03wf001.pdf. Access Date Feb 20, 2015

Luel Berhe 2011 The Role of data mining technology in electronic transaction expansion at Dashn Bank S.C. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Miguéis V.L., Camanho A.S., João Falcão e Cunha (2012), "Customer data mining for lifestyle segmentation", Expert Systems with Applications,Vol.39, pp. 9359-9366.

Mihreteab Negash 2012 Neural Network Based Data-Driven Clinked Quality prediction: Case study on Muger Cement Factory. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

M. Venkatesh June 2013 "A Study Of Trend Analysis In Insurance Sector In India ",International Journal Of Engineering And Science (IJES) ,Volume 2 ,Issue 6,pp 01-05,.

Minale Tefera 2012 Application of data mining techniques to predict urinary fistula surgical repair outcome. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

M.P.S Bhatia1 and Deepika Khurana 2013 Experimental study of Data clustering using k-Means and modified algorithms, International Journal of Data Mining and Knowledge Managemnet Process (IJDKP) Vol.3, No. 3

Osmar R. Zaïane, 1999 Principles of Knowledge Discovery in Databases University of Alberta Department of Computing Science.

Priyanka Gaur 2012 Neural Networks in Data Mining International Journal of Electronics and Computer Science Engineering.

P. Chaudhari, D. P. Rana, R. G. Mehta, N. J. Mistry, M. M. Raghuwanshi 2014 Discretization of Temporal Data A survey. Internation Journal of Computer science and information security Volum 12 No. 2.

Riccardo Bellazzi , Blaz Zupan 2008 "Predictive data mining in clinical medicine: Current issues and guidelines", International Journal of Medical Informatics 77 81-97.

Stockwell I, 2008, Introduction to Correlation and Regression analysis, SAS Global Forum, Paper-364, pp. 1-8.

S. Balaji and Dr. K. Srinivasta, August 2012 "Naïve Bayes Classification Approach for Mining Life Insurance Databases for Effective Prediction of Customer Preferences over Life Insurance Products" , International journal of Computer Applications, vol.51, No.3, pp.22-26.

Tariku Adane june 2011. Mining insurance data for fraud detection: the case of Africa insurance share company. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Tan, Steinbach, Kumar 2004 Introduction to Data Mining and knowledge discovery approach. https://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap1_intro.pdf. Access Date Feb 26, 2015

Tariku Debela 2013 developing a predictive model for fertility preference of women of reproduction age using data mining techniques. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Tesfaye Hintsay Atsmo June 2002 Predictive modeling using data mining techniques in support of insurance risk assessment. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Tibebe Beshah, Shawndra 2010 Hill Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia.

Types of Insurance Risk
http://www.ehow.com/list_6612287_types-insurance-risks.html
Access Date Jan 2015

U.M. Fayyad, G. Piatessky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). 1996 Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

United Nation Economic Commission for Africa Road Safety in Ethiopia Case study 2009 ECA/NRID/019

www.eic.com.et EIC official website Access Date Dec 2014

www.Consumer Federation of America (CFA) and the Center for Economic Justice Access Date Dec 2014.

Xu Zhiku, Wang Yanwen and Liu Zhaohui, 2014 ”Optional Insurance Compensation Rate Selection and Evaluation in Financial Institutions “, International Journal of u- and e- Service, Science and Technology,Vol.7, No.1 , pp.233-242 ,

Yashpal Singh, Alok Singh Chauhan 2009 Neural network in data mining Journal of Theoretical and Applied Information Technology.

Appendices

Appendix 1 Initial collected attributes from INSIS database

s. n	Attributes Name	Data Type	Description
1	INSR_TYPE	Number	The type of insurance policy issued. i.e. whether commercial or private
2	COVER_TYPE	Varchar2	The types of cover underwrite.
3	RISK_TYPE	Varchar2	The type of risk the policy covered
4	REPR_COST	Number	The actual repair cost
5	CAR_USAGE	Number	The type of vehicle use.
6	CCM_TON	Number	Cubic capacity of the vehicle
7	MAKE	Varchar2	Make of the vehicle
8	DRVR_SEX	Varchar2	Drivers' sex
9	DRVR_AGE	Number	Drivers' age
10	DRVR_EXPERNC	Number	Drivers' driving experience
11	MAN_COMP	Number	Territorial extension
12	VHCL_PROD_YEAR	Number	Year of vehicle manufactured
13	INITIAL_RESERVED_AMOUNT	Number	Amount reserved for payment
14	CAR_TYPE	Varchar2	The type of car
15	CARBODY_TYPE	Number	Body type of the car
16	OPC1	Varchar2	Legal entity
17	OPC2	Varchar2	Teritorial extension
18	LAST_RESERVED_AMOUNT	Number	The amount of money remain form the reserved amount
19	CLAIM_NO	Number	Claim number
20	CLAIM_ID	Varchar2	Claim id

21	MODEL		Model of the car
22	SEATS_NUM	Number	Number of seats

Appendix 2. Sample Value of final selected attributes

INSR_TYPE	COVER_TYPE	RISK_TYPE	MAKE	SEX	AGE_GROUP	EXP_GROUP	VHCL_AGE	CC_TON	REP_COST	CAR_TYPE	CAR_USAGE
1201	M_COMPREH	M_LODAM_IV	TOYOTA	M	AGE_GROUP_ONE	EXP_GROUP_ONE	VHCL_GROUP_TWO	CC_GROUP_THREE	REP_GROUP_ONE	STA_GROUP_FIVE	PRIVAT_USE
1201	M_COMPRTPL	TPL_PROP	TOYOTA	M	AGE_GROUP_ONE	EXP_GROUP_ONE	VHCL_GROUP_TWO	CC_GROUP_THREE	REP_GROUP_ONE	STA_GROUP_FIVE	PRIVAT_USE
1201	M_COMPREH	M_LODAM_IV	TOYOTA	M	AGE_GROUP_ONE	EXP_GROUP_ONE	VHCL_GROUP_FOUR	CC_GROUP_ONE	REP_GROUP_ONE	SMALL_GROUP_ONE	PRIVAT_USE
1202	M_COMPREH	M_LODAM_IV	MITSUBISHI	M	AGE_GROUP_ONE	EXP_GROUP_ONE	VHCL_GROUP_THREE	CC_GROUP_TWO	REP_GROUP_ONE	PICK_GROUP_SEVEN	OWN_GOODS
1202	M_COMPREH	M_LODAM_IV	MITSUBISHI	M	AGE_GROUP_ONE	EXP_GROUP_TWO	VHCL_GROUP_THREE	CC_GROUP_TWO	REP_GROUP_ONE	PICK_GROUP_SEVEN	OWN_GOODS
1202	M_COMPREH	M_LODAM_IV	TOYOTA	M	AGE_GROUP_ONE	EXP_GROUP_ONE	VHCL_GROUP_TWO	CC_GROUP_THREE	REP_GROUP_TWO	BUS_GROUP_FOUR	OWN_SERVICE
1202	M_COMPREH	M_LODAM_IV	MERCEEDES	M	AGE_GROUP_ONE	EXP_GROUP_TWO	VHCL_GROUP_FOUR	CC_GROUP_ZERO	REP_GROUP_ONE	BUS_GROUP_FOUR	OWN_SERVICE
1202	M_COMPREH	M_LODAM_IV	TOYOTA	M	AGE_GROUP_ONE	EXP_GROUP_ONE	VHCL_GROUP_TWO	CC_GROUP_TWO	REP_GROUP_ONE	PICK_GROUP_SEVEN	OWN_GOODS
1202	M_COMPREH	M_LODAM_IV	TOYOTA	M	AGE_GROUP_ONE	EXP_GROUP_TWO	VHCL_GROUP_TWO	CC_GROUP_TWO	REP_GROUP_TWO	PICK_GROUP_SEVEN	OWN_GOODS
1202	M_COMPREH	M_LODAM_IV	MITSUBISHI	M	AGE_GROUP_ONE	EXP_GROUP_TWO	VHCL_GROUP_FOUR	CC_GROUP_TWO	REP_GROUP_TWO	PICK_GROUP_SEVEN	OWN_GOODS

Appendix 3 Confusion marix result of classification algorithms

1) Confusion matrix result of j48 with minNumObj=20

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1765   45    2   34 |    a = cluster0
   39  688   58   69 |    b = cluster1
   13   57 1168    7 |    c = cluster2
   41   54    6 1130 |    d = cluster3
```

2) Confusion matrix result of j48 with minNumObj=15

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1771   42    1   32 |    a = cluster0
   35  688   62   69 |    b = cluster1
   13   42 1185    5 |    c = cluster2
   53   41    6 1131 |    d = cluster3
```

3) Confusion matrix result of j48 with minNumObj=10

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1796   13    1   36 |    a = cluster0
   16  727   46   65 |    b = cluster1
   10   32 1195    8 |    c = cluster2
   42   29    6 1154 |    d = cluster3
```

4) Confusion matrix result of j48 with minNumObj=5

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1807    6    3   30 |    a = cluster0
   16  749   32   57 |    b = cluster1
    6   32 1199    8 |    c = cluster2
   30   20    9 1172 |    d = cluster3
```

5) Confusion matrix result of j48 with default value

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1812    8    5   21 |    a = cluster0
    8  770   27   49 |    b = cluster1
    7   22 1208    8 |    c = cluster2
   33   24    8 1166 |    d = cluster3
```

6) Confusion matrix result of j48 with 70% percentage split.

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 376   1   0   4 |   a = cluster0
   3 147   3  14 |   b = cluster1
   0   4 244   4 |   c = cluster2
   7   5   3 220 |   d = cluster3
```

7) Confusion matrix result of multiperceptron neural network with default values.

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1812    8    5   21 |    a = cluster0
    8  770   27   49 |    b = cluster1
    7   22 1208    8 |    c = cluster2
   33   24    8 1166 |    d = cluster3
```

Appendix 4 Partial view of the result of j48 decision tree with default values and 10-fold cross validation test option

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

INSR TYPE <= 1201
|   CAR_TYPE = STA_GROUP_FIVE
|   |   AGE_GROUP = AGE_GROUP_ONE
|   |   |   EXP_GROUP_ = EXP_GROUP_ONE
|   |   |   |   VHCL_AGE = VHCL_GROUP_TWO
|   |   |   |   |   CC_TON = CC_GROUP_THREE: cluster1 (61.03/0.03)
|   |   |   |   |   CC_TON = CC_GROUP_ONE: cluster1 (0.0)
|   |   |   |   |   CC_TON = CC_GROUP_TWO: cluster0 (6.0)
|   |   |   |   |   CC_TON = CC_GROUP_ZERO: cluster2 (11.01/0.01)
|   |   |   |   |   CC_TON = CC_GROUP_FIVE: cluster1 (0.0)
|   |   |   |   |   CC_TON = CC_GROUP_FOUR: cluster1 (0.0)
|   |   |   |   VHCL_AGE = VHCL_GROUP_FOUR: cluster2 (67.04/0.04)
|   |   |   |   VHCL_AGE = VHCL_GROUP_THREE: cluster2 (67.03/0.03)
|   |   |   |   VHCL_AGE = VHCL_GROUP_ONE: cluster2 (57.03/0.03)
|   |   |   EXP_GROUP_ = EXP_GROUP_TWO: cluster1 (48.03/1.03)
|   |   |   EXP_GROUP_ = EXP_GROUP_THREE: cluster1 (1.0/0.0)
|   |   AGE_GROUP = AGE_GROUP_THREE: cluster1 (61.03/4.03)
|   |   AGE_GROUP = AGE_GROUP_TWO: cluster1 (236.12/4.12)
|   CAR_TYPE = SMALL_GROUP_ONE
|   |   EXP_GROUP_ = EXP_GROUP_ONE: cluster2 (671.0/6.0)
```

Appendix 5: Source code of the developed MIRAS

```
using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.Linq;
using System.Text;
using System.Windows.Forms;
namespace MIRAS
{
    public partial class frmMiras : Form
    {
        # region Fields
        private string _PlateNo = string.Empty;
        private bool _isNew = false;
        private DataTable tblMira;
        private DataTable tblMiraList;
        private DataRow _Mira;

        private DataTable _tblCoverType;
        private DataTable _tblInsuranceType;
        private DataTable _tblRiskType;
        private DataTable _tblCarUsage;
        private DataTable _tblVehicleType;
```

```
#endregion
private void frmMiras_KeyDown(object sender, KeyEventArgs e)
{
    if (e.KeyCode == Keys.Enter)
    {
        SendKeys.Send("{TAB}");
    }
}

public frmMiras()
{
    InitializeComponent();
}
private void frmMiras_Load(object sender, EventArgs e)
{
    GetDataSources();
    FillData();
}
private void GetDataSources()
{
    DbConn Conn = new DbConn(StartUpClass.ConnString);
    try
    {
        Conn.OpenConnection();
   this.tblMiraList = Conn.GetDataTable("tblMiras");
   this.vehiclePlateNoLookUpEdit.Properties.DataSource = this.tblMiraList;
```

```
this._tblCoverType = Conn.GetDataTable("tblCoverType");
this.coverTypeIdLookUpEdit.Properties.DataSource = this._tblCoverType;

    this._tblInsuranceType = Conn.GetDataTable("tblInsuranceType"
    this.insuranceTypeIdLookUpEdit.Properties.DataSource = this._tblInsuranceType;

    this._tblRiskType = Conn.GetDataTable("tblRiskType");
    this.riskTypeIdLookUpEdit.Properties.DataSource = this._tblRiskType;

    this._tblVehicleType = Conn.GetDataTable("tblVehicleType");
    this.vechicleTypeIdLookUpEdit.Properties.DataSource = this._tblVehicleType;

    this._tblCarUsage = Conn.GetDataTable("tblCarUsage");
    this.carUsageIdLookUpEdit.Properties.DataSource = this._tblCarUsage;

    this.tblMira = Conn.GetDataTable("tblMiras");
    List<Control> cntrsList = new List<Control>();

    cntrsList.AddRange(new Control[]{

        this.vehiclePlateNoTextBox,
        this.insuranceTypeIdLookUpEdit,
        this.coverTypeIdLookUpEdit ,
        this.cCMTONTextBox ,
        this.riskTypeIdLookUpEdit ,
        this.vProductionYearTextBox ,
        this.repairCostTextBox ,
```

```
            this.carUsageIdLookUpEdit ,
            this.makeTextBox,
            this.driverNameTextBox ,
            this.driverDOBDateTimePicker ,
            this.driverSexComboBox ,
            this.driverExperienceTextBox ,
            this.levelOfRiskTextBox ,
            this.confidencePercentTextBox,
            this.vechicleTypeIdLookUpEdit
        });

        StartUpClass.BindControls(cntrsList, this.tblMira);
        cntrsList.Clear();

    }
    catch (Exception Ex)
    {

        MessageBox.Show(Ex.Message, "Error", MessageBoxButtons.OK,
        MessageBoxIcon.Exclamation);
    }
    finally
    {
        Conn.CloseConnection();
    }
}

private void FillData()
```

```
        {
            DbConn Conn = new DbConn(StartUpClass.ConnString);
            try
            {
                Conn.OpenConnection();
                Conn.FillData(this.tblMiraList, "select * from tblMiras");
                Conn.FillData(this._tblCoverType, "select * from tblCoverType");
                Conn.FillData(this._tblInsuranceType, "select * from tblInsuranceType");
                Conn.FillData(this._tblRiskType, "select * from tblRiskType");
                Conn.FillData(this._tblCarUsage, "select * from tblCarUsage");
                Conn.FillData(this._tblVehicleType, "select * from tblVehicleType");

            }
            catch (Exception Ex)
            {
                MessageBox.Show(Ex.Message, "Error", MessageBoxButtons.OK,
MessageBoxIcon.Exclamation);
            }
            finally
            {
                Conn.CloseConnection();
                this.Refreshbutton_Click(this, new EventArgs());
            }
        }

        private void GetMiraDetail(string plateNo)
        {
            DbConn Conn = new DbConn(StartUpClass.ConnString);
```

```
        object result;
        string drSex = string.Empty;
        if (!this.MiraDetailExists(plateNo))
        {
            this.GetNewEntry();
            return;
        }
        try
        {
            Conn.OpenConnection();
            this._isNew = false;
            Conn.FillData(this.tblMira, "Select * from tblMiras where VehiclePlateNo='" + plateNo
+ "'");

            result = Conn.GetDataScalar("Select DriverSex from tblMiras where
VehiclePlateNo='" + plateNo + "'");
            if (result != null && !Convert.IsDBNull(result))
            {
                this.driverSexComboBox.SelectedItem = Convert.ToString(result);
            }
            else
                this.driverSexComboBox.SelectedItem = null;

        }
        catch (Exception Ex)
        {
            MessageBox.Show(Ex.Message, "Error", MessageBoxButtons.OK,
MessageBoxIcon.Exclamation);
        }
```

```
        finally
        {
            Conn.CloseConnection();
        }
    }

    private bool MiraDetailExists(string plateNo)
    {
        object result;
        string cmdText;
        DbConn Conn = new DbConn(StartUpClass.ConnString);

        if (plateNo == string .Empty)
            return false;

        try
        {
            Conn.OpenConnection();
            cmdText = "Select VehiclePlateNo from tblMiras where VehiclePlateNo='" + plateNo +
"'";

            result = Conn.GetDataScalar(cmdText);
            if (result != null && !Convert.IsDBNull(result))
                return true;
            else
                return false;
        }
        catch (Exception Ex)
```

```
        {
            MessageBox.Show(Ex.Message, "Error", MessageBoxButtons.OK,
MessageBoxIcon.Exclamation);
            return false;
        }
        finally
        {
            Conn.CloseConnection();
        }
    }
    private void GetNewEntry()
    {
        this._isNew = true;
        this._Mira = this.tblMira.NewRow();
        if (this._PlateNo != string.Empty) this._Mira["VehiclePlateNo"] = this._PlateNo;
        this.tblMira.Rows.Clear();
        this.tblMira.Rows.Add(this._Mira);
    }
```

Appendix 6 sample rules generated to classified new records in to existing cluster.

Rule 1

If Insurance_type = Commercial and

Experiance_group(driving Experience) <= 5 years and

Vhcl_age(vehicle production year) <10 and >6 and

Car_type = Track and

Cc_ton >= 9000 and

Rep_cost(Repaire cost) >50,000 and

Car_usage = General Cartage and

Age_group (driver's age Group) < 37 and >20 and

Sex =Male

Then Cluster 1 (Very High Risk)

Rule 2

If Insurance_type = Private and

Experiance_group(driving experience) <10 and >6 and

Vhcl_age (age of the vehicle) <10 and >6 and

Car_type = station wagon and

Cc_ton <6000and >40000 and

Rep_cos (Repair cost) >10000 and <50000 and

Car_usage = private and

Age_group <53 and >37 and

sex=M

then Cluster 2 (Low Risk).

Rule 3

If Insurance_type = Private and

Exp_group(Driving Experience) < 5 and

Vhcl_age (age of vehicle) > 20 and

Car_type = small Vehicel and

Cc_ton < 1500 and

Rep_cost (Repair cost)<50000 and >10 000 and

Car_usage = Private and

Age_group(Driver's age) <37 and >20 and

sex = M

then Cluster 3 (Medium Risk)

Rule 4

If Insurnce_type = Commercial and

Exp_group (Driving Experience) < 5 and

Vhcl_age (Vhicle age) >6 and <10 and

Car_type = pick_up and

Cc_ton <4000 and >1500 and

Rep_cost (Repair Cost) >10000 and <50000 and

Car_usage = Own Goods and

Age_group (Driver's Age)) <37 and >20 and

sex =M

then Cluster 4 (High Risk)

Rule 5

If Insurance_type = Private and

Exp_group (Driver's Experiance) < 5 and

Vhcl_age (Age of vehicle)<10 and >6 and

Car_type =Station Wagon and

Cc_ton <6000 and >4000 and

Age_group (driver's age) <37 and >20 and

Then Cluster 2 (Low Risk)

Rule 6

If Insurance_type = Commercial and

Exp_group (driving Experiance) < 5 and

Car_type = Bus and

Cc_ton <6000 and >4000 and

Car_usage= Fair paying passenger and

Age_group <37 and > 20

then Cluster 3(Medium Risk)

Rule 7

If Insurance_type = Commercial and

Exp_group (driving experience) < 5 and

Car_type = Track and

Cc_ton >9000 and

Car_usage = Own Goods and

Then Cluster 1 (Very High Risk)

Rule 8

If Insurance_type = Commercial and

Exp_group (Driving Experiance)<10 and >6 and

Car_type =Track and

Cc_ton <6000 and >40000 and

Car_usage = Own Goods

Then Cluster 1 (Very High Risk).

Rule 9

If Insurance_type = Private and

Exp_group (driving experience) <20 and >10 and

Car_type = Small

sex=M

then Cluster 3 (Medium Risk).

Rule 10

If Insurance_type = Commercial and

Exp_group (Driving Experience) < 5 and

Car_type = Pick_up and

Cc_ton < 6000 and >4000 and

Car_usage= Commerical and

Then Cluster 4 (High Risk)

Rule 11

If Insurance_type = Private and

Exp_group (Driving Experience) < 5 and

Vhcl_age (Age of the vehicle) >20 and

Car_type = Station Wagon and

Age_group <37 and >20

Then Cluster 3 (Medium Risk)

Rule 12

If Insurnce_type = Private and

Exp_group (Driving Experiance) < 5 and

Vhcl_age <20 and >10 and

Car_type =Station Wagon and

Age_group (Drivers age) <37 and >20

then Cluster 3 (Medium Risk)

Rule 13

If Insurance_type = Private and

Exp_group (Driver's Experiance) < 5 and and

Car_type = Station Wagon and

Age_group (Drivers Age) <37 and >20

then Cluster 3 (Medium Risk)

Rule 14

If Insurance_type = Private and

Car_type = Station Wagon

Age_Group = <10 and >6

Then Cluster 2(Low Risk).

Rule 16

If Insurance_type = Commercial and

Exp_group (Driver's experiacne)<10 and >6 and

Car_type= Truck and

Cc_ton <4000 and >1500 and

Rep_cost (Repaire cost) > 50000

then Cluster 1 (Very High Risk)

DECLARATION

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

_____

Yihenew Fekadu

June 2015

The thesis has been submitted for examination with my approval as university advisors.

_____

Dereje Teferi (PhD)