ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT FRAUD PROTECTION: THE CASE OF ETHIOPIAN REVENUE AND CUSTOM AUTHORITY

DANIEL MAMO

January, 2013

SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT FRAUD PROTECTION: THE CASE OF ETHIOPIAN REVENUE AND CUSTOM AUTHORITY

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirement for the Degree of

Master of Science in Information Science

BY

DANIEL MAMO

January 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT FRAUD PROTECTION: THE CASE OF ETHIOPIAN REVENUE AND CUSTOM AUTHORITY

BY

DANIEL MAMO

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
Meftaha Hassen	Chairperson	----------	----------
Dereje Teferi (PhD)	Advisor	----------	----------
Rahel Bekele (PhD)	Examiner	----------	----------

DEDICATION

I would like to dedicate this thesis work to my late mother Asnakech Mulat, my late sisters Berhane Mamo, Muluneshe Mamo, Mehret Mamo and my late brothers Kindie Mamo and Abraham Mamo.

ACKNOWLEDGMENT

First and foremost my special thanks go to the almighty God for his forgiveness with the courage and endurance to successfully complete this research work.

Next to this I would like to express my sincerest gratitude and heartfelt thanks to my advisor, Dr. Dereje Teferi. I am really grateful for his constructive comments and critical readings of the study. I am very thankful to Dr. Million Meshesha for his support. I am also very thankful to my instructors and all staff members of the School of Information Science for their contribution in one way or another for success of my study. My thanks also go to staff of Institute of Ethiopian Studies.

I would like to extend my appreciation to my friends Belachew Regani, Guesh Dagnew, Girma Aweke and Belay Abebe for their all rounded help during our stay at Addis Ababa University. I am also very grateful to my father Mamo Abebe and my sister Fasika Mamo to their all rounded support.

I would like to thank ERCA, Information Technology Management Directorate members Ato Getenet Abebaw and Bezabehe Shumet for providing me the necessary data for the study and for their unreserved help throughout the study time.

My special thanks also go to Meseret Ayano for all rounded support.

My deepest gratitude also goes to Adamseged Birru for his unreserved professional supports and advice until the end of the thesis. Besides I appreciate his commitment and punctuality.

Finally, I would like to thank my wife Yirgalem Tilahun for her priceless support; your support helps me to finalize the thesis.

LIST OF ABBREVIATIONS

AI: Artificial Intelligence

ANN: Artificial Neural Network

ARFF: Attribute Relation File Format

ASYCUDA: Automation System Customs Data

CRISP-DM: Cross Industry Standard Process for Data Mining

CRM: Customer Relationship Management

CSV: Comma Separated Values

DM: Data Mining

ERCA: Ethiopian Revenues and Customs Authority

FIRS: Federal Inland Revenue Service

ITMD: Information Technology Management Directorate

KDD: Knowledge Discovery in Databases

SEMMA: Sample Explore Modify Model Assess

SIGTAS: Standard Integrated Government Tax Administration System

TIN: Tax Identification Number

WEKA: Waikato Environment for Knowledge Analysis

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Taxes are important sources of public revenue. The existence of collective consumption of goods and services necessitates putting some of our income into government hands. However, collection of tax is the main source of income for the government; it is facing difficulties with fraud. Fraud involves one or more persons who intentionally act secretly to deprive the government income and use for their own benefit. Fraud is as old as humanity itself and can take an unlimited variety of different forms. Fraudulent claims account for a significant portion of all claims received by auditors, and cost billions of dollars annually.

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

To collect the data the researcher used interview and observation for primary data and database analysis for secondary data. The experiments have been conducted following the six-step Cios et al. (2000) KDD process model. For the experiment, the collected tax payers' dataset is preprocessed to remove outliers, fill in ITMD values, select relevant attributes, integrate data and derive attributes. The preprocessing phase of this study really took the highest portion of the study time. In this study, different characteristics of the ERCA customers' data were collected from the customs ASYCUDA database. A total of 11080 tax payers' records are used for training the models, while a separate 2200 records are used for testing the performance of the model. The model developed using the J48 decision tree algorithm has showed highest classification accuracy of 99.98%. This model is then tested with the 2200 testing dataset and scored a prediction accuracy of 97.19%. The results of this study have showed that the data mining techniques are valuable for tax fraud detection. Hence future research directions are pointed out to come up with an applicable system in the area

CHAPTER ONE

BACKGROUND

1.1 INTRODUCTION

Taxes are important sources of public revenue. The existence of collective consumption of goods and services necessitates putting some of our income into government hands. Such public goods like roads, power, municipal services, and other public infrastructures have favorable results on many families, business enterprises, industries and the general public. Public goods are normally supplied by public agencies due to their natures of non-rivalry and non-excludability (Ethiopian Chamber of Commerce, 2005).

The main problem in tax collection activity is fraud. Fraud involves one or more persons who intentionally act secretly to deprive the government income and use for their own benefit. Fraud is as old as humanity itself and can take an unlimited variety of different forms (Palshikar, 2002). Traditional ways of data analysis have been in use since long time as a method of detecting fraud. They require complex and time-consuming investigations that deal with different domains of knowledge like financial, economics, business practices and law.

Fraud often consists of many instances or incidents involving repeated transgressions using the same method. Fraud instances can be similar in content and appearance but usually are not identical (Hand, et al, 2001). The Institute of Internal Auditors' International Professional Practices Framework (IPPF) defines fraud as: "…. Any illegal act characterized by deceit, concealment, or violation of trust". These acts are independent upon the threat of violence or physical force. Frauds are perpetrated by parties and organizations to secure personal or business advantage through unlawful act to obtain money, property, services or to avoid payment or loss of services.

Fraud management is a knowledge-intensive activity. The main Artificial Intelligence techniques used for fraud management includes data mining to classify, cluster, and segment the data and automatically

find associations and rules in the data that may signify interesting patterns, including those related to fraud.

Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules (Piatetsky et al, 1996). Organizations use this information to detect existing fraud and noncompliance, and to prevent future occurrences.

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

Data mining enables data exploration and analysis without any specific hypothesis in mind, as opposed to traditional statistical analysis, in which experiments are designed around a particular hypothesis. While this openness adds a strong exploratory aspect to data mining projects, it also requires that organizations use a systematic approach in order to achieve usable results.

The application of Data Mining techniques for financial classification is a fertile research area. Many law enforcement and special investigative units, whose mission it is to identify fraudulent activities, have also used Data Mining successfully. However, as opposed to other well-examined fields like bankruptcy prediction or financial distress, research on the application of DM techniques for the purpose of management fraud detection has been rather minimal (Kirkos & Manolopoulos, 2004).

The first industries to use data analysis techniques to prevent fraud were the telephony companies, the insurance companies and the banks (Fawcett, 1997). One early example of successful implementation of data analysis techniques in the banking industry is the Falcon fraud assessment system, which is based on a neural network shell (Phua et al., 2005).

ERCA traces its origin to July 7, 2008 as a result of the merger of the Ministry of Revenues, the Ethiopian Customs Authority and the Federal Inland Revenues into one giant organization. The Ethiopian Revenue and Custom Authority (ERCA) receive the taxpayers' declaration from taxpayers and based on their declaration it collects the tax. ERCA has the following objectives: -

1. To establish modern revenue assessment and collection system and provide customers with equitable, efficient and quality service,

2. To cause taxpayers voluntarily discharge their tax obligations,

3. To enforce tax and customs laws by preventing and controlling contraband as well as tax fraud and evasion,

4. To collect timely and effectively tax revenues generated by the economy,

5. To provide the necessary support to regions with a view to harmonizing federal and regional tax administration systems.

1.2 Statement of the Problem

Having stated some of the functions of government to the citizens using taxation as a tool, the objective of taxation can therefore be summed up as in (Nightingale, 2002; Lyme and Oats, 2010):

- Raising revenue to finance government expenditure.
- Redistribution of wealth and income to promote the welfare and equality of the citizens.
- Regulation of the economy thereby creating enabling environment for business to thrive.

Taxation is therefore, one among other means of revenue generation of any government to meet the need of the citizens some of which have been pointed out above. According to Public Finance General Directorate (2009) the purpose of taxation as enshrined in the French laws is "for the maintenance of public force and administrative expenses". Miller and Oats (2006:3) maintained, "Taxation is required to finance public expenditure". It is worthy of note however, that there are other sources of revenue generation by the government e.g. borrowing, grants etc. If taxation is for public expenditure, public goods ought to have been consumed equally.

But during tax collection activity the main problem is to get the exact income report from the taxpayers' to the tax collector offices. If the tax collector offices cannot collect the tax based on the taxpayers' income the problem will direct to government annual budget. Annual expenditure of the government depends on its income.

According to different documents described, currently Ethiopia is one of the countries which don't collect enough amount tax from the taxpayers properly.

The main responsibility of the Ethiopian Revenue and Custom Authority (ERCA) is to collect tax from all business sectors to change the government plan into practice. Currently the government is striving to cover its annual expenditure from local financial resources. ERCA is playing significant role to implement the development plan of this country. During Tax collection the main problem of the authority is fraudulent protection on taxpayers' declaration. To prevent this problem, the first task should be investigating the taxpayers' declaration data (ERCA BPR Document, 2005).

Rather than using highly curative method it is advantageous to make predictions about the future based up on the hidden knowledge that are constructed from the database or repositories of the records of taxpayers and this technique can help to take appropriate actions.

Tax investigation offices should predict the future based up on the extracted knowledge from data that would tells about the previous and current to predict the future situations by using data mining technology to prevent fraud activities that applies investigation techniques which is useful for prevention mechanisms rather than dealing after fraud has occurred. And the ERCA do not utilize its fraud activity data in a way that enables it to extract new knowledge that is important to forecasting the problem of fraud activities and take preventive actions.

This research is focused on the application of data mining techniques for revenue and custom authority in supporting its efforts in fraud detection and to enhance its responsiveness to the current and prospective tax payers needs and create strong tax collection mechanism.

Revenue and custom authority collects tax yearly in Billions of Birr (www.erca.gov.et). At this time it is trying to register all business enterprises and individuals who are engaged in the Business sector. On the one hand the organization has rich taxpayers' related data on which the application of data mining techniques, especially classification and clustering methods could result in valuable information for decision making. On the other hand, there is no integrated system or model being applied to investigate it the tax payers' declaration is free from fraud.

Generally, this study investigates the compliance or noncompliance behavior of the taxpayers from the available ERCA data using data mining techniques.

To this end, this study attempts to obtain answers for the following main research questions:

What is the pattern that characterizes whether a given claim is fraudulent or not?

Which data mining technique is more appropriate to identify determining factors for fraud detection?

To accomplish the tax collection task the authority needs to use the data mining techniques to protect fraud and improve loyalty.

Besides, the environment is rapidly changing because of information and communication technologies (ICTs). Hence, adoption of the recent trend is becoming mandatory.

1.3 Objectives of the Study

1.3.1 General Objective

The general objective of this research is to apply data mining for creating a predictive model that determine the compliance and noncompliance behavior of taxpayers for the purposes of developing an effective tax collection by Ethiopian Revenues and Customs Authority.

1.3.2 Specific Objectives

The specific Objectives of the Study include:

- To understand fraud problem based on the review of data mining literatures and analysis of ERCA data and documents.
- To make an interview with domain experts and analyze relevant documents. This would help to identify the relevant attributes.
- To construct ERCA target dataset used for data mining tools following major data preparation and pre-processing steps.
- To select proper model that helps in finding out and uncovers factors at the back of the compliance and noncompliance taxpayers.
- To investigate patterns that discloses the relationships of taxpayer status with other variables.
- To conduct experimentation to evaluate the accuracy of the system.
- To draw recommendation based on the findings of the study.

1.4 Scope and Limitation of the Study

The main aim of this research is exploring the applicability of DM for fraud claim prediction and detection in the ERCA. This research focuses only to the audit task process owner. Specifically the mined data for knowledge discovery is obtained from the ITMD in Addis Ababa. Auditing is the most effective working area of ERCA.

The authority categorized its taxpayers in to three. Category ‘A’ taxpayers annual income is more than 4,000,000 Birr. Category ‘B’ taxpayers annual income is between 500,000 and 4,000,000 Birr. Category ‘C’ taxpayers annual income is less than 500,000. The scope of this research is limited to analysis of the category of “A” and category ‘B’ taxpayers data of the year 2009-2011 G.C. Since Category ‘A’ and

category'B' taxpayers are large in number when compared to other catagories, and they are more diversed; it is believed that the researcher of this paper gets sufficient data od category A and B for the purpose of the research. Due to the absence of sufficient data , it is difficult to cover all taxpayers who are found in other categories . To achive the objective of this study, a two step DM technique is used; i.e first the study apply clustering technique to define the natural group of records and then classification to develop prediction model, which helps to identify tax fraud suspicion.

1.5 Research Methodology

In order to define the research problem properly, primary data was collected by interviewing concerned auditors as well as through observation (questions rose during the interview are presented in appendix part. Then based on the information obtained from these attempts, the overall fraud prevention of the ERCA was described.

Relevant literatures on data mining techniques and fraud were reviewed. The potential of data mining in general and particularly successful data mining applications in fraud prevention were assessed.

An important issue for the design of a framework is a knowledge discovery process. Many researchers explained a series of steps that comprise the Knowledgy Discovery process to be followed by practitioners when executing a DM project. In this research WEKA (open source software) is employed to implement most of the technical aspects of the CRISP-DM standard data mining methodology that has been adopted. Business understanding, data understanding, data preprocessing, and selection of modeling technique, model building and model evaluation is undertaken in this research. To implement the proposed study WEKA machine learning software have been used. The main reason is: a) it is extensively documented and b) it is equipped with multiple features to handle the activities performed in any data mining method.

1.5.1 Research Design

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

Cios et al. (2000) model consists of understanding the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge steps.

1.5.2 Understanding of the Revenue and Custom Domain

A closer look of the problem environment is the first step taken in Knowledge Discovery DM research. The DM problem is defined on the basis of the insights gained from the phase. To clearly identify, understand, and analyze the business problems, the primary (observation and interview), and secondary (database analysis) data collection methods are employed, Interview is employed to define features selection with the domain experts while observation is conducted to understand some complex business processes. The main goal of the Information Technology Management Directorate of ERCA is to reduce operational expense through increased efficiency and productivity throughout the process chain, improve service levels by implementing a faster, more visible, and consistent approach to increase the income from tax and proactive fraud management.

The main Data Mining goal for this research is identifying and detecting fraudulent tax payers. For that matter, a model is developed using the different data mining techniques, which helps to predict fraudulent tax payments.

1.5.3 Understanding the Data

After understanding the problem to be addressed clearly in this study, the next step is analyzing and understanding the data available. The outcome of data mining and knowledge discovery heavily depends on the quality and quantity of available data (Cios et al. 2007). The data that is used in this research was initially collected, regarding tax collected from tax payers.

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

1.5.4 Preparation of the Data

This step is the key upon which the success of the whole knowledge discovery process depends. At this time real-world databases are highly susceptible to noisy, missing values, and inconsistent data due to

their typically huge size (mostly several gigabytes or more) and their likely origin from multiple, heterogeneous sources (Han and Kamber, 2006). The researcher decides, at this step, together with domain experts, the data that is used as input for applying the data mining techniques.

The different data preprocessing techniques are applied for processing the data used by the k-means and decision trees algorithms chosen for the research. Data cleaning routines are applied to fill in missing values (with the mean value), smooth out noise (by removing the record), and detect outliers (by removing or substituting with mean values) in the data. The cleaned data is further processed by feature selection consulting the domain experts and the WEKA attribute selection preprocessing techniques (to reduce dimensionality) and by derivation of new attributes. The result of these processes generates datasets for training and testing the clustering and classification algorithms selected in this study.

1.5.5 Data Mining

Based on the identified goals and the assessment of the available data, appropriate mining algorithm is chosen and run on the prepared data. In many applications of machine learning to data mining, the explicit knowledge structures that are acquired, namely the structural descriptions, are at least as important, and often more important, than the ability to perform well on new examples. The use of data mining to gain knowledge discovery regularities in the data and not just predictions is common (Witten and Frank, 2005).

Gaining knowledge, discovering irregularities for detecting fraudulent claims within the tax payers' dataset in particular, is certainly the purpose of this study. Having this purpose in mind, the unsupervised clustering technique and the supervised classification technique are adopted. The clustering technique is selected because of the reason that the dataset, which is obtained from ERCA doesn't have a feature indicating whether a claim was fraud or not. Therefore, the clustered dataset is used as an input for the classification algorithm for classifying instances of the dataset into similar class labels. For this purpose, the J48 decision tree algorithm and the Naïve Bayes classification methods are used to create model for detecting fraudulent activities.

1.5.5.1 Data Mining Tool Selection

For conducting this study the WEKA (Waikato Environment for Knowledge Analysis) data mining software is chosen. WEKA is chosen because of its widespread application in different data mining researches and familiarity of the researcher with the software.

WEKA, a machine-learning algorithm written in Java, is adopted for undertaking the experiment. WEKA constitutes several machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from one's own Java code. WEKA is open source software. The WEKA data mining software included classification, clustering, association rule learner, numeric prediction and several other schemes. In addition to the learning schemes, WEKA also comprises several tools that can be used for datasets pre-processing (Witten and Frank, 2005).

1.5.6 Evaluation of the Discovered Knowledge

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

1.6 Significance of the Study

This study is significant to ERCA because it introduces with the application of the data mining technique that enables it to extract new, potentially useful and novel knowledge from the data repository and predict other associated situations that will be used to give acceptable decisions in the protection of fraud in the form of proactive prevention technique that should be expected to save the time and increase income to the government and can achieve its development plan.

Auditors in fraud prevention and investigation authority of the respective authority can make use of the results of this study in order to make optimal deployment of resource in fraud prevention. Moreover, the output of the study shall be used for designing appropriate training programs and fraud prevention and investigation strategies.

The outcome of this research shall also be used as a benchmark for auditors as well as a source of methodological approach for studies dealing on the application of data mining on fraud management as well as other similar areas.

1.7 Organization of the Thesis

This research paper is organized in six chapters. The first chapter deals with the background of the study which mainly introduces the problem area, statement of the problem, the general and specific objectives of the study, the research methodology, the scope of the study, and significance of the study. The second chapter reviews the KDD or data mining technology. In this chapter, different data mining techniques, such as clustering and classification with their respective algorithms, are reviewed. Chapter three discusses data mining methods for fraud detection. Chapter four presents business understanding of ERCA. Chapter five presents the experimentation phase of the study. The last chapter, chapter six presents the conclusion of the result of the study and provides recommendation based on the investigation of the research.

CHAPTER TWO

DATA MINING AND KNOWLEDGE DISCOVERY

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases (Hand, et al., 2004). This has occurred in all areas of human endeavor. Such as supermarket transaction data, credit card usage records, telephone call details, and government offices statistics to the more exotic. Database technology since the mid-1980s has been characterized by the popular adoption of relational technology and an upsurge of research and development activities on new and powerful database systems (Han and Kamber, 2006). These employ advanced data models such as extended – relational, object oriented, object-relational, and deductive models. Database sizes have significantly increased into larger size of data which have a complete advantage if the hidden information is extracted. To uncover this hidden information and knowledge, data mining is being used both to increase revenues and to reduce costs.

To bridge the gap of analyzing large volume of data and extracting useful information and knowledge for decision making that the new generation of computerization methods known as Data Mining (DM) or Knowledge Discovery in Databases (KDD) has emerged in recent years. Data mining is 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 (Two Crows Corporation, 2005). Information is basically extracted from the huge size of data. Data mining is as a means for detecting fraud, assessing risk, and product retailing (Jeffrey, 2004). Data mining involves the use of data analysis, tools to discover previously unknown and valid patterns and relationships in large data sets. At this time data mining is used to identify different kinds of crimes such as illegal money transfer, and tax fraud. Using data mining in different organizations like bank, insurance, supermarket and research institutions is common.

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large datasets (Two Crows Corporation, 2005). These tools can include statistical models, mathematical algorithms, and machine learning methods. Accordingly, DM consists of more than collecting and managing data; it also includes analysis and prediction and use of algorithms that improve their performance automatically through experience, such as neural networks or decision trees.

DM is an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, data visualization, optimization, information retrieval, high and computing, and others (Guo, 2003). Data Mining (DM) is an iterative process within which progress is defined by discovery, either through automatic or manual methods. DM is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an ‘‘interesting’’ outcome (Kantardzic, 2002). The application of Data Mining techniques for financial classification is a fertile research area. Many law enforcement and special investigative units, whose mission it is to identify fraudulent activities, have also used Data Mining successfully.

2.1 The Data Mining Process

Because it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive *(*Han and Kamber, 2006*)*. For databases containing a huge amount of data, appropriate sampling techniques can first be applied to facilitate interactive data exploration. Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. Specifically, knowledge should be mined by drilling down, rolling up, and pivoting through the data space and knowledge space interactively. In this way, the user can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.

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

2.1.1 Data Acquisition

Data acquisition is the process of sampling signals that measure real world physical conditions and converting the resulting samples into digital numerical values that can be manipulated by a computer. In DM primarily, to select the type of data is used. However, a target dataset has been created for discovery in some applications, DM can be performed on a set of variables or data samples in a larger database called training set to create and model while holding back some of the datasets (test dataset) for latter validation of the model.

2.1.2 Data Preprocessing

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources (Han and Kamber, 2006). Low-quality data will lead to low-quality mining results. The researchers are using a number of data preprocessing techniques to remove the noise, outliers and irrelevant data if necessary and decide on strategies for dealing with missing data fields and accounting for time sequence information or known changes for instance, numeric ones into categorical.

2.1.3 Model Building

This stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome. After the data has been accessed and modified, analysts can use data modeling techniques to construct models that explain patterns in the data. Modeling techniques in data mining include neural networks, tree-based models, logistic models and other statistical models, such as time series analysis and survival analysis. Once a DM technique is chosen, the next step is to select a particular algorithm within the DM technique chosen. Choosing a DM algorithm includes a method to search for pattern in the data, such as deciding which models and parameters may be appropriate and matching a particular DM technique with the overall objective of DM. At the end with selected algorithm the data is mined to extract novel patterns hidden in databases.

2.1.4 Interpretation and Model Evaluation

At this forth step products can help the user understand the results by providing measures (of accuracy, significance, etc.) in useful formats such as confusion matrices and ROI charts, by allowing the user to perform sensitivity analysis on the result, and by presenting the result in alternative ways, such as graphically. The extracted knowledge is also evaluated in terms of its usefulness to a decision maker and to a business goal.

2.2 Data Mining Tasks

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive (Han and Kamber, 2006). Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

2.2.1 Predictive Modeling

A predictive model is constructed based on the analysis of the values of the other attributes or dimensions' describing the data objects (tuples). The set clause can be used to fix the values of these other attributes (Han and Kamber, 2006). Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is the process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to specific variable(s) you are trying to predict. It is a supervised learning because the classes are predefined before the examination of the target data.

Regression is a data mining function that predicts a number. Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques. In the time series analysis the value of an attribute is examined as it varies over time and the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

2.2.2 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. This classification is 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. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction. DM creates classification models by examining already classified data (cases) and inductively finding a predictive pattern (Two Crows Corporation, 2005). Based on this explanation, these existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. For example, a sample of a mailing list would be sent an offer, and the results of the mailing used to develop a classification model to be applied to the entire database. Sometimes an expert classifies a sample of the database, and this classification is

then used to create the model which will be applied to the entire database. Decision tree, neural network, genetic algorithm, Naïve Bayes are algorithms used for classification purpose.

Decision Tree

Decision trees are a way of representing a series of rules that lead to a class or value (Two Crows Corporation, 2005). A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions (Han and Kamber, 2006). Decision trees can easily be converted to classification rules. Depending on the algorithm, each node may have two or more branches. For example, CART generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed it is called a multiway tree. Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node. By navigating the decision tree can be assigned a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached. Each node uses the data from the case to choose the appropriate branch.

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

A decision tree is a collection of nodes, arranged as a binary tree. The leaves render decisions; in our case, the decision would be "likes" or "doesn't like." Each interior node is a condition on the objects being classified; in our case the condition would be a predicate involving one or more features of an item (Rajaraman and Ullman, 2011).

To classify an item, we start at the root, and apply the predicate at the root to the item. If the predicate is true, go to the left child, and if it is false, go to the right child. Then repeat the same process at the node visited, until a leaf is reached. That leaf is classified the item as liked or not.

The initial state of a decision tree is the root node that is assigned all the examples from the training set. If it is the case that all examples belong to the same class then no further decisions need to partition the examples and the solution is complete. If examples at this node belong to two or more classes then a test is made at the node that will result in a split. The process is recursively repeated for each of the new

intermediate nodes until a completely discriminating tree is obtained. A decision tree at this stage is potentially an over-fitted solution i.e. it may have components that are too specific to noise and outliers that may be present in the training data. As Apte and Weiss(1997) indicated, to relax this over-fitting most decision tree methods go through a second phase called pruning that tries to generalize the tree by eliminating sub trees that seem too specific. Error estimation techniques play a major role in tree pruning. Most modern decision tree modeling algorithms are a combination of a specific type of a splitting criterion for growing a full tree and a specific type of a pruning criterion for pruning tree.

Decision trees are a simple, but powerful form of multiple variable analyses. They provide unique capabilities to supplement, complement, and substitute for

• Traditional statistical forms of analysis (such as multiple linear regressions)

• A variety of data mining tools and techniques (such as neural networks)

• Recently developed multidimensional forms of reporting and analysis found in the field of business intelligence.

A series of improvements to ID3 culminated in a practical and influential system for decision tree induction called C4.5. These improvements include methods for dealing with numeric attributes, missing values, noisy data, and generating rules from trees.

Decision tree can be implemented with several algorithms. Some of them are J48, ID3, C4.5, CART, etc. The decision tree C4.5 algorithm is a practical method for inductive inference (Mitchell, 1997). Connect *J48* to the cross-validation fold maker in the usual way, but make the connection twice by first choosing trainingset and then choosing testset from the pop-up menu for the cross-validation fold maker. Amongst other enhancements (compared to the ID3 algorithm) the C4.5 algorithm includes different pruning techniques and can handle numerical and missing attribute values. C4.5 avoids over fitting the data by determining a decision tree, it handles continuous attributes, is able to choose an appropriate attribute selection measure, and handles training data with missing attribute values and improves computation efficiency. C4.5 builds the tree from a set of data items using the best attribute to test in order to divide the data item into subsets and then it uses the same procedure on each sub set recursively. The main problem in decision tree is deciding the attribute, which will best partition the data into various classes (Meera and Srivatsa, 2010). The ID3 algorithm is useful to solve this problem.

Neural Networks

Neural network (Quinlan, 1993) technology uses a multilayered approach that approximates complex mathematical functions to process data. Today neural networks can be trained to solve problems that are difficult for conventional computers or human beings. As shown a situation in Fig 1, neural networks are trained that a particular input leads to a specific target output. Based on a comparison of the output and the target, the network is trained until the network output matches the target. Typically many such input/target pairs are used to train a network.

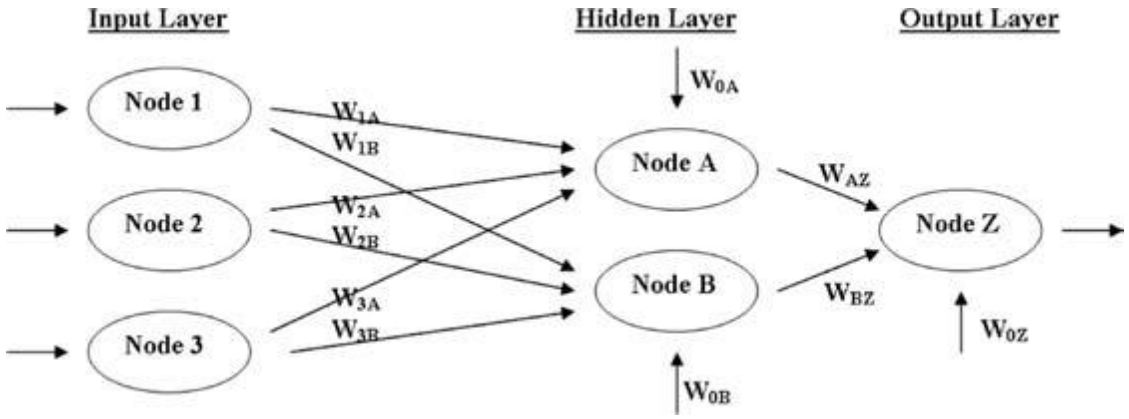

Fig.2. 1 Neural network algorithm

Neural network consists of many processing elements or nodes that work in parallel. Nodes are connected to each other in layers, and layers are interconnected. These nodes are simple mathematical functions; the connections between these nodes, which weight the data transformation from each node and send the information to the next node or output layer, are how neural networks "think" (Ponce and Karahoca, 2009). As the complexity of the task increases, the network size increases, and the number of nodes increases rapidly. To properly train a neural network, the developer feeds the model a variety of real-life examples, called training sets. The data sets normally contain input data and output data. The neural network creates connections and learns patterns based on these input and output data sets. Each pattern creates a unique configuration of network structure with a unique set of connection strengths or weights. A neural network adapts to changing inputs and learns trends from data. A set of examples of

the data or images is presented to the neural network, which then weights the connections between nodes based on each training example.

Neural networks are remarkable for their learning efficiency and tend to outperform other methods (like decision trees) when no highly relevant attributes exist, but many weakly relevant ones are present. Furthermore, ANN can easily be adjusted as new examples accumulate. However according to [Lu *et* al. (1996)], the drawbacks of applying neural networks to data mining include: difficulty in interpreting the model, difficulty in incorporating prior knowledge about the application domain in a neural network, and, also, long learning time, both in terms of CPU time, and of manually finding parameter settings that will enable successful learning.

The rule extraction algorithm, described in (Lu et al., 1996), makes an effective use of the neural network structure, though the weights of the links between the nodes remain meaningless, and the rules are extracted in a deterministic (Boolean) form. The network is pruned by removing redundant links and units, but removal of entire attributes (Feature selection) is not considered.

In a feed-forward neural network the connection between neurons only occurs with neurons in different layers (or in different sub-layers of the hidden layer). Each such connection has an associated weight (Carlos, 2006).

A neuron is only responsible for determining the activation level and firing an output. Its activation level is the sum of the activation levels of the neurons connected to it from the previous layer weighted by the connection strength. The output is a function of the activation level. Typically the logistic function is used: $f(x) = 1 / (1 + e^{x})$, where x is the activation level. Let us call the output of a neuron *i*, *Oi*.

Initially connection weights are set to random values and the aim of training a neural network is to determine the best weights for the neuron connections so that the sum of the squares of the errors is minimized. During the training process, as the weights are being adjusted, the prediction errors decrease and eventually the weights on the network no longer change significantly and so the error stabilizes.

The learning rate is a global neural network parameter that defines the rate at which the neural networks adapts to new weights. If this value is too small the network converges very slowly, if the value is too big the weights may jump between extremes and never converge.

The default stop criteria is the persistence which checks if the network has not improve prediction for K (200 by default in Clementine) cycles then the training phase is terminated. Other stop criteria area: time elapsed, percentage of error obtained (may never be reached!) and number of records used.
The main advantages of a neural network are its robustness to data noise, its capacity of running in parallel as well as its capacity of approximation any function.
The more serious disadvantage of a neural network is that the model (eg: the weights of the neuron connections) is incomprehensible for a human and no business knowledge can be extracted from it.

Neural network topologies can be divided into feed forward and recurrent classes according to their connectivity (Singh and Chauhan, 2005). The feed forward neural network was the first and arguably simplest type of artificial neural network devised.

2.2.1.1 Prediction

Prediction is the other example of predictive modeling. It can be accomplished by independently estimating the probabilities, and securely multiplying and comparing to obtain the predicted class (Shrikant, 2004). The predictions can be evaluated based on real observations on experimental data sets.

2.2.2 Descriptive Modeling

A model is a high-level description, summarizing a large collection of data and describing its important features. Often a model is global in the sense that it applies to all points in the measurement space.
The goal of a descriptive model is describe all of the data (or the process generating the data). Examples of such descriptions include models for the overall probability distribution of the data (density estimation), partitioning of the *p*-dimensional space into groups (cluster analysis and segmentation), and models describing the relationship between variables (dependency modeling) (Hand et al., 2001). Clustering is similar to classification except that the groups are not predefined, but are defined by the data alone (Han and Kamber, 2006). The association rule finds the association between the different attributes. Association rule mining is a two-step process. The first one is finding all frequent item sets, generating strong association rules from the frequent item sets. Secondly, sequence discovery is a process of finding the sequence patterns in data.

2.2.2.1 Clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering (Han and Kamber, 2006). A cluster is a collection of data objects that are similar to one

another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Although classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labeling of a large set of training tuples or patterns, which the classifier uses to model each group. It is often more desirable to proceed in the reverse direction: First partition the set of data into groups based on data similarity (e.g., using clustering), and then assign labels to the relatively small number of groups (Hajizadeh et al., 2010). Additional advantages of such a clustering-based process are that it is adaptable to changes and helps single out useful features that distinguish different groups.

Clustering tools assign groups of records to the same cluster if they have something in common, making it easier to discover meaningful patterns from the dataset. Clustering often serves as a starting point for some supervised DM techniques or modeling.

Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. 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 (Qiu et al., 2010). 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.

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

Cluster analysis tools based on *k*-means, *k*-medoids, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS (Han and Kamber, 2006). In machine learning, clustering is an example of unsupervised learning. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labeled

training examples. For this reason, clustering is a form of learning by observation, rather than learning by examples. In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large databases.

Hierarchical clustering approach is further subdivided into agglomerative and divisive,

- ✓ Agglomerative is started with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. It is a bottom up clustering technique.
- ✓ Divisive is started with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. It is a top down clustering technique. In this case, we need to decide, at each step, which cluster to split and how to perform the split.

K-Means clustering algorithm

The *k*-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intracluster similarity is high but the intercluster similarity is low (Han and Kamber, 2006). Cluster similarity is measured in regard to the *mean* value of the objects in a cluster, which can be viewed as the cluster's Centroid or center of gravity.

The *k*-means algorithm works as follows. First, it randomly selects *k* of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster.

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

2.2.2.2 Association Rule Discovery

The aim of association rule discovery is the derivation of *if-then-rules* based on the predicates *ax* defined in the previous subsection. An example of such a rule is "if a market basket contains orange juice then it also contains bread" (Hegland, 2003).

One possible way to interpret such a rule is to identify it with the predicate "either the market basket contains bread or it doesn't contain orange juice".

The analysis could then determine the support of this predicate. However, this is not a predicate in the class we have defined in the previous section. Of course it could be determined from the $s(ax)$. However, this measure also is not directly interpretable and thus not directly applicable in retail.

Association is suitable if the problem is to extract any structure that exists in the data at hand. Association rules can be developed in order to determine arrangements of items on store shelves in a given supermarket so that items often bought together will be found arranged closer to location (Berry and Linoff, 1997). Association rule is a powerful tool for discovering correlations among massive databases and the concept was introduced for analyzing market basket data to mine customer shopping pattern. The input used in the process of generating association rule is taken from a table where corresponding values in each data items have correlations to one another. For example, 65 percent of records that contain item A also contain item B. An association rule uses measures called "support" and "confidence" to represent the strength of association. The percentage of occurrences, 65 percent in this case is the confidence factor of the association.

There are a variety of algorithms to identify association rules. The most widely used association rule algorithms are Apriori and FP-growth tree. Apriori is an influential algorithm for finding frequent item sets using candidate generation (Wu et al., 2007). Frequent-pattern tree or FP-tree in short is an extended prefix-tree structure storing crucial, quantitative information about frequent patterns. FP-growth method is an efficient and scalable mining for both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm (Han et al., 2004).

2.3 Types of Data Mining Systems

Data mining systems can be classified into different categories. The classification is based on type of data source, data model, kind of knowledge discovered, and mining techniques used, also the degree of user interaction involved in the data mining process (Dunham and Sridhar, 2006). The classification of data mining systems according to the type of data source mined is based on the type of data handled for mining purpose such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

The other classification of DM systems is according to the data model. This classification is based on data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc. Further classification of DM systems is according to the kind of knowledge discovered. This classification of DM systems based on the kind of knowledge discovered or DM functionalities,

such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several DM functionalities together. Finally, DM systems can be classified according to mining techniques used. This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, and visualization.

2.4 The Data Mining Models

Model building is a key objective of data mining and data analysis applications. In the past, such applications required only a few models built by a single data analyst (Liu *and* Tuzhilin, 2008). As more and more data has been collected and real world problems have become more complex, it has become increasingly difficult for that data analyst to build all the required models and manage them manually. Building a system to help data analysts construct and manage large collections of models is a pressing issue.

The choice of a data model is a very crucial and complex task in data mining. Not only should the model represent the data precisely but it should also be appropriate for the mining technique used. For instance the data inputs of a neural network technique are different from the inputs of a support vector machine or a hidden Markov model (Ponce and Karahoca, 2009).

We usually divide the dimensions into two categories: primary and secondary. Primary dimensions are the main dimensions that characterize the data itself. The secondary dimensions are informative but they can play a huge role when associated with the inputs of a given mining technique. The difficulty here is that there is no general rule for how to select appropriate secondary dimensions (inputs) for a given mining process. The selection depends largely on the experience and the expertise of the user within that specific domain or application.

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

2.4.1 The Six Step Cios Model

As described in section 1.5.1 of Chapter one this model was developed by adopting the CRISP-DM model to the needs of academic research community. The model of six steps (Cios and Kurgan, 2005).

1. Understanding the problem domain.

In this step one works closely with domain experts to define the problem and determine the research goals, identifies key people, and learns about current solutions to the problem. It involves learning domain-specific terminology. A description of the problem including its restrictions is done. The research goals then need to be translated into the DM goals, and include initial selection of potential DM tools.

2. Understanding the data.

This step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the DM goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

3. Preparation of the data.

This is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire research effort. In this step, which data will be used as input for data mining tools of step 4, is decided. It may involve sampling of data, running, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be further processed by feature selection and extraction algorithms (to reduce dimensionality), and by derivation of new attributes (say by discretization), and by summarization of data (data granularization). The result would be new data records, meeting specific input requirements for the planned to be used DM tools.

4. Data mining.

This is another key step in the knowledge discovery process. Although it is the data mining tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. Data mining tools include many types of algorithms, such as neural networks, clustering, preprocessing techniques, etc. This step involves the use of several DM tools on data prepared in step 3. First, the training and testing procedures are designed and the data model is constructed using one of the chosen DM tools; the generated data model is verified by using testing procedures.

5. Evaluation of the discovered knowledge.

This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered

knowledge. Only the approved models are retained. The entire DM process may be revisited to identify which alternative actions could have been taken to improve the results.

6. Using the discovered knowledge.

This step is entirely in the hands of the owner of the database. It consists of planning where and how the discovered knowledge will be used. The application area in the current domain should be extended to other domains.

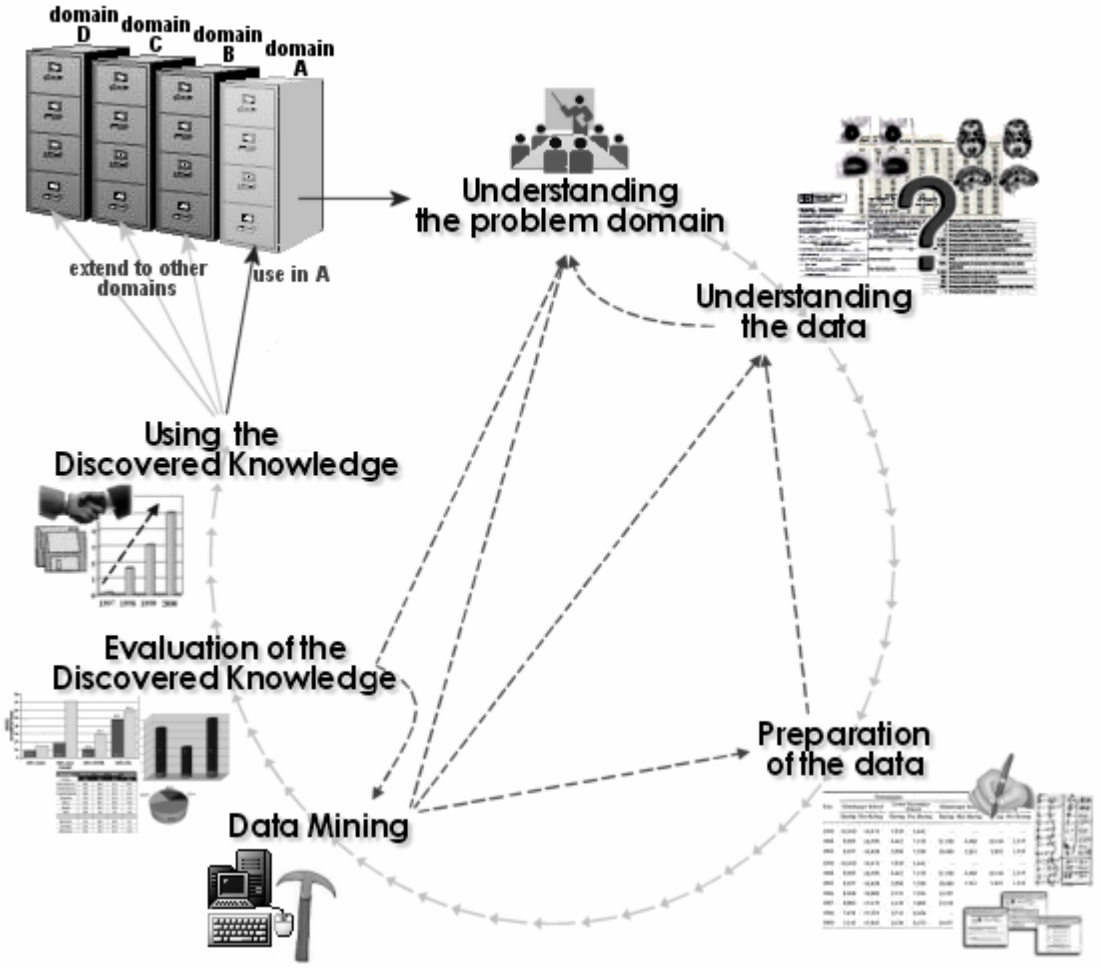

Fig 2. 2 The Six Step Cios et al. process model

2.4.2 The KDD Process Model

As stated by Cao. (2007) Knowledge-Discovery and Data Mining (KDD), is the process of automatically searching large volumes of data for hidden, interesting, unknown and potentially useful patterns. It is an interactive and iterative process, comprising a number of phases requiring the user to make several decisions. There are five steps in the KDD process.

1. **Data Selection:** This step focused on creating a target dataset, or a subset of variables or data samples, on which discovery is to be performed. The data relevant to the analysis is decided on and retrieved from the data collection.
2. **Data Pre-Processing:** This stage consists on the target data cleaning and pre-processing in order to obtain consistent data.
3. **Data Transformation:** is data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure. This stage is consists on the transformation of the data using dimensionality reduction or transformation methods.
4. **Data Mining:** This is the most important step in which clever techniques are applied to extract potentially useful patterns. It consists on the searching for patterns of interest in a particular representational form, depending on the data mining objective.
5. **Interpretation/ Evaluation:** This stage focused on the interpretation and evaluation of the data mined patterns.

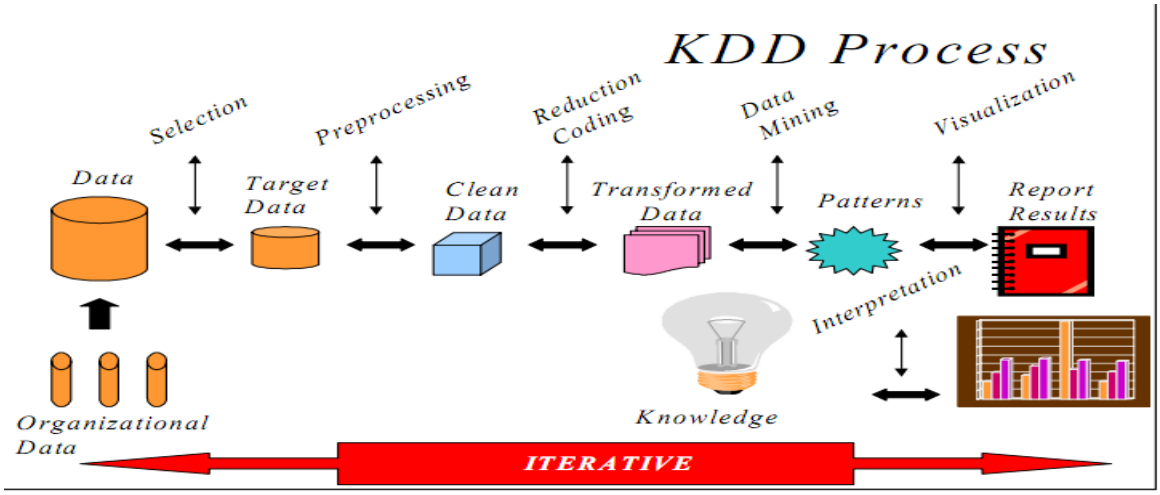

Fig 2.3 The KDD process

2.4.3 The CRISP-DM Process

Chapman, et al (2000) stated that CRISP-DM was conceived in late 1996 by three "veterans" of the young and immature data mining market. DaimlerChrysler (then Daimler-Benz) was already experienced, ahead of most industrial and commercial organizations, in applying data mining in its business operations.

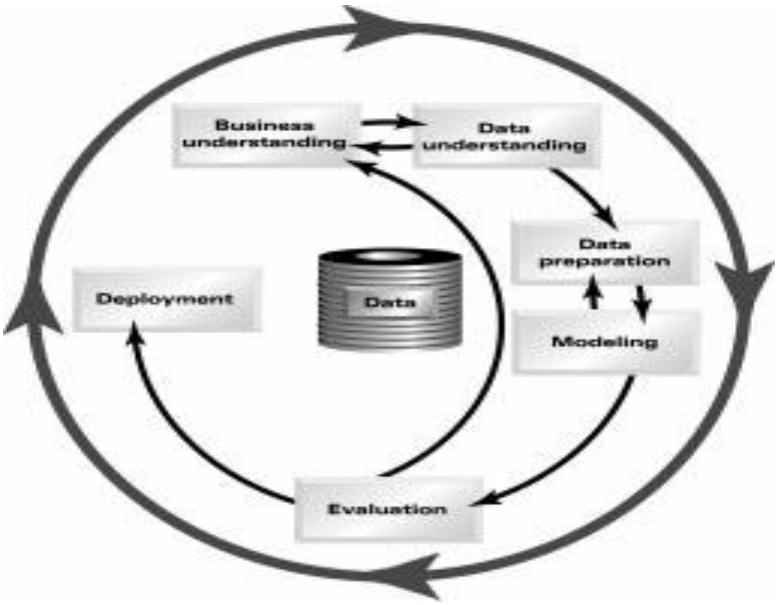

Fig 2.4 the CRISP-DM Process

The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase which phase or which particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases (Chapman et al. 2000). The life cycle of a data mining project consists of six phases.

Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data understanding

This phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data preparation

This phase covers all activities to construct the final dataset from the initial raw data. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Modeling

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

Evaluation

At this stage, before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

2.5 Application of Data Mining

Currently, different firms are realizing the numerous advantages that come with DM. And organizations are using data mining to manage all phases of the customer life cycle (acquiring new customers, increasing revenue from existing customers, and retaining good customers) (Two Crows Corporation, 1999). It provides clear and competitive advantage across a broad variety of industries by identifying potentially useful information from the huge amounts of data collected and stored. Today it is primarily used by companies with a strong consumer focus on retail, financial, tax revenue, communication, and marketing organizations.

Data mining enables these companies to determine relationships among internal factors such as price, product positioning, or staff skills, and external factors such as economic indicators, competition, and customer demographics. Furthermore, it enables to determine the impact on sales, customer satisfaction, corporate profits by drilling down into summery of information (Berry, and Linoff, 2004).

2.5.1 Data Mining in the Tax Administration

Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules (Fayyad, 1996). Organizations use this information to detect existing fraud and noncompliance, and to prevent future occurrences.

Audit selection is one of many possible data mining applications in tax administration. In the area of tax compliance alone, tax collectors face diverse challenges, from underreporting to non-reporting to enforcement. Data mining offers many valuable techniques for increasing the efficiency and success rate of tax collection. Data mining also leverages the variety and quantity of internal and external data sources available in modern data warehousing systems.

Data mining is the process of exploration and analysis, by automatic means, of large quantities of data in order to discover meaningful patterns and rules. Typical techniques employed include such processes as neural networks and regression trees. Essentially, for a given set of data (including the results of past case audit activity), the data mining software is asked to distinguish the characteristics of taxpayers that have been non-compliant (usually on the basis of past audit results) from those that are compliant (Organization for economic co-operation and development, 2004). The software can analyze thousands of characteristics simultaneously, and find patterns in the data that can be used to provide new criteria for identifying non-compliance. It is an example of how technology can be used to supplement human auditing experience.

Tax agencies in the state of Texas in the United States, United Kingdom and Australia rely on data mining to help find delinquent taxpayers and in making effective resource allocation decisions.

2.5.2 Revenue and Custom Fraud Detection

The delineation of fraud to "occupational fraud and abuse" is one way to categorize fraud. There are numerous other ways of classifying fraud. A classification that resembles however this first delineation is the distinction Bologna and Lindquist (1995) make between internal versus external fraud. This classification, applied in the field of corporate fraud (fraud in an organizational setting), is based on whether the perpetrator is internal or external to the victim company. Frauds committed by vendors, suppliers or contractors are examples of external fraud, while an employee stealing from the company or a manager cooking the books are examples of internal fraud.

Fraud management is a knowledge-intensive activity. The main AI techniques used for fraud management include: Data mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud.

As advised in the Fraud Control Guidelines, fraud against the Commonwealth is defined as 'dishonestly obtaining a benefit, or causing a loss, by deception or other means'. A result of the Australian Institute of Criminology's is in 2007–08 Annual Reporting Questionnaire indicated that of the external fraud incidents, the focus of the highest number of activities was on entitlements. This category includes obtaining a Commonwealth payment, for example, a social, health or welfare payment by deceit. It also

includes revenue fraud, which is, deliberately avoiding obligations for payment to government, including income, customs or excise taxes (ANAO Audit Report No.30, 2008)

Most entities that collect revenue or administer government payments conduct reviews across the various revenue and payment types. Based on previous experience, knowledge of their customers, and evidence from within their systems or from outside information, entities may undertake reviews that examine a recipient's circumstances where there is a perceived risk of fraud. The aim of such reviews is to detect a deliberate error, omission, misrepresentation or fraud on the part of a customer (Attorney-General's Department, 2011).

The opportunities to enhance tax administration and compliance functions through the use of data and advanced data analytics are significant. Revenue agencies face a growing list of challenges including the continued pressures of shrinking operating budgets, the loss of experienced workers, and the growing occurrence of evasion and fraud schemes that understate tax liabilities and/or exploit vulnerabilities in traditional returns processing, especially refund returns (SAS Enterprise Miner, 2003). As evasion and fraud schemes become more complex and pervasive, the need to leverage data and data analytics to optimize processes and detect and predict return anomalies and errors, whether intentional or unintentional, is a critical core competency of tax administration.

Data Mining (DM) is an iterative process within which progress is defined by discovery, either through automatic or manual methods. DM is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an ‘‘interesting’’ outcome (Kantardzic, 2002). The application of Data Mining techniques for financial classification is a fertile research area. Many law enforcement and special investigative units, whose mission is to identify fraudulent activities, have also used Data mining successfully, however, as opposed to other well-examined fields like bankruptcy prediction or financial distress, research on the application of DM techniques for the purpose of management fraud detection has been rather minimal (Kirkos & Manolopoulos, 2004).

Fraud that involves cell phones, insurance claims, tax return claims, credit card transactions etc represent significant problems for governments and businesses, but yet detecting and preventing fraud is not a simple task. Fraud is an adaptive crime, so it needs special methods of intelligent data analysis to detect and prevent it. These methods exist in the areas of Knowledge Discovery in Databases (KDD),

Data Mining, Machine Learning and Statistics. They offer applicable and successful solutions in different areas of fraud crimes.

Techniques used for fraud detection fall into two primary classes: statistical techniques and artificial intelligence (Palshikar 2002). Examples of statistical data analysis techniques are: Data preprocessing techniques for detection, validation, error correction, and filling up of missing or incorrect data. Calculation of various statistical parameters such as averages, quantiles, performance metrics, probability distributions, and so on.

Data preprocessing techniques are used for detection, validation, error correction, and filling up of missing or incorrect data.

Review activity should be targeted to areas of higher risk, and an entity should pursue the most productive method for undertaking reviews. Data mining / matching is a cost-effective method of supporting reviews, including cross-organizational approaches.

During tax collection the tax agencies collect vast data of the tax payers. The Revenue and Custom authority should handle and investigate its data to identify compliance from non-compliance tax payers, so that data mining techniques are of particular importance.

Due to this fact the research is done in this area of tax fraud detection. Among them some are discussed below.

According to Oluba (2011), as it does in many other spheres of our national and business life, data mining has many existing and potential applications in tax administration. For instance, predictive modeling will obviously assist the Nigerian FIRS to be in the best position to identify noncompliant taxpayers as well as ensure that tax auditing resources are channeled more appropriately on the accounts that will most likely yield the most desirable tax adjustments. There are many inherent advantages in this. First, the tax authority will be able to harness and manage its human resources more appropriately. Secondly, it will minimize the wastage of resources on compliant taxpayers. Thirdly, it will also enable the modification of the deployment of the hitherto traditional audit selection strategies for one that surely produces more efficacious outcomes.

Data or knowledge discovery (or data mining) encompasses the process of discovering correlations or patterns among lots of fields in large relational databases. The process involves the analysis of data and organizing them into useful information which in many cases (particularly in organizations) are for the purposes of revenue enhancement, cost cutting or both. With the existence of vast historical data on tax payers, it is easy for tax authorities to predict potential degrees of non-compliance. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Generally, data mining algorithms scour databases for hidden patterns, in search of predictive information that experts may miss because they lie outside their expectations (Oluba, 2011).

Phua, et al. (2005) states that, it is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms. Evolved from numerous research communities, especially those from developed countries, the analytical engine within these solutions and software are driven by artificial immune systems, artificial intelligence, auditing, database, distributed and parallel computing, econometrics, expert systems, fuzzy logic, genetic algorithms, machine learning, neural networks, pattern recognition, statistics, visualization and others. There are plenty of specialized fraud detection solutions and software which protect businesses such as credit card, e-commerce, insurance, retail, and telecommunications industries.

The Inland Revenue office of Australia (2009) stated that internationally, additional resources are being applied to address compliance issues similar to those highlighted above. For example the Australian Government will invest in excess of $600m over the next four years in the Australian Tax Office (ATO). This will be used to address key compliance areas such as the cash economy, abuse of tax havens, managing compliance risks to Australia's economic recovery and other public awareness campaigns.
Additional investment would also be made in Inland Revenue's intelligence tools and processes, such as increased automated data-matching. This would increase revenue by ensuring that Inland Revenue can identify cases of non-compliance more quickly and accurately. This would allow us to intervene more quickly and appropriately. Inland Revenue would also invest additional funding in proactive compliance, which will increase revenue over time by encouraging taxpayer compliance through education, awareness and influencing social norms.

Gupta and Nagadevara, (2008) stated that selecting returns for Audits are like looking for a needle in a haystack. Every year, a large number of taxpayers fail to declare their tax liabilities correctly, and the Tax Administration is forced to tackle a tough task – to detect them (and enforce compliance from them) without increasing the compliance costs of the tax compliant taxpayers. It is not possible to identify the likely tax-evaders by simple query and reporting tools. Tax departments have access to enormous amounts of taxpayer data. However, it is impossible to ascertain the legitimacy of and intention behind a claim or a declaration in the tax return by simply looking at the return or a profile of a taxpayer. Given this reality, the best cost effective option is to tease-out possible indications of fraudulent claims/declarations from the available data using data mining algorithms.

Based on Luciano, (2009) explanation, once again, neural networks have been widely used. The main fraud detection software of the fraud solution unit of Nortel Network uses a combination of profiling and neural networks. Moreover, on the continuous basis that made a research on user profiling by using neural network and probabilistic models and by other research which is call-based fraud detection by using a hierarchical regime switching model.

From the above very influential researches we can generalize that the tax administration task will become even more complicated over time with more opportunity for fraud. At present the extent of fraud is based on the financial statement and evasion of income.

There are several approaches to deal with fraud detection. We highlight the use of neural networks, Bayesian networks, expert system, rule based systems and the detection of statistical outliers. These approaches can be subdivided in two groups: supervised and unsupervised. In the supervised approaches there is a training set of operations that are labeled either as fraudulent or normal. These operations are used as input to some systems, such as neural network systems, that need labeled inputs to construct the model that will be used to detect frauds. Alternative strategies have been employed in without benefits, but a novel approach, described in, achieved significant improvements in some performance measures. The unsupervised approaches do not need labeled inputs, as they use a set of rules to classify an operation as a fraud or compare each one with the previous operations to identify those that might be considered suspicious (outliers).Rule based systems are unsupervised approaches that use a set of rules to classify the operations as fraudulent or normal, or to assign a value to each operation corresponding to the chance an operation has to be a fraud. The rules are typically constructed following advises of experts. These systems have the advantage of being unsupervised and taking account of the experts' knowledge to construct the rules that evaluate each operation. One of the disadvantages of these systems

is the fact that the rules frequently need to be updated to deal with new fraudulent behaviors. Otherwise, the rules will eventually become obsolete.

2.5.3 Local Related Works

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

Shegaw (2002) also conducted a research on the application of DM in predicting child mortality in Ethiopia as a case study in the Butajira Rural Health Project. Shegaw employed the classification technique, neural network and decision tree algorithms to develop the model for predicting child mortality. Additional case studies were also conducted regarding the application of DM in the different sectors. For example, Tilahun (2009) has tried to assess the possible application of DM techniques to target potential VISA card users in direct marketing at Dashen Bank. Melkamu (2009) also conducted a research to assess the applicability of DM techniques to CRM as a case study on Ethiopian Telecommunications Corporation (ETC).Additionally, Leul (2003) tried to apply the DM techniques for crime prevention as a case study on the Oromia Police Commission. Leul used the classification technique, decision tree and neural network algorithms to develop the model, which will help to classify crime records. Helen (2003) also tried to study the application of DM technology to identify significant patterns in census or survey data as a case of the 2001 child labor survey in Ethiopia. She has applied the association rule DM technique and the Apriori algorithm for identifying relationships between attributes within the 2001 child labor survey database that she used to clearly understand the nature of child labor problem in Ethiopia. Apart from the association rule technique the expectation maximization-clustering algorithm were used to categorize the final selected datasets. Tariku (2011) tried to develop models that can detect and predict fraud in insurance claims with a particular emphasis to Africa Insurance Company. He tried to apply clustering algorithm followed by classification techniques for developing

the predictive model. K-means clustering algorithm is employed to find the natural grouping of the different insurance claims as fraud and non-fraud.

Local researches are conducted to assess the application of DM in the different sectors like Airlines, Banking, HealthCare, and Customs. The main intension of all researchers is to investigate the applicability of Data Mining in the above mentioned sectors. Most researchers used clustering and classification techniques with k-Means and decision tree algorithms. In addition most of the researches are implemented for specific domain area.

Similarly, the proposed DM techniques is in this study are conducted to explore the applicability of DM in ERCA. The main objective of this research study is to apply data mining for creating a predictive model that determine the compliance and noncompliance behavior of taxpayers for the purposes of developing an effective tax collection by Ethiopian Revenues and Customs Authority. Therefore, to accomplish the tax collection task the authority needs to use the data mining techniques to protect fraud and improve loyalty.

CHAPTER THREE

DATA MINING METHODS FOR FRAUD DETECTION

As a consequence of the pervasion of electronic computing systems into private and business events of everyday life a lot of targets for misuse have evolved. A very high potential of financial damage can be found in the areas of electronic banking, telecommunications and tax collecting agencies.

Traditional ways of data analysis have been in use since long time as a method of detecting fraud. They require complex and time-consuming investigations that deal with different domains of knowledge like financial, economics, business practices and law. Fraud often consists of many instances or incidents involving repeated transgressions using the same method. Fraud instances can be similar in content and appearance but usually are not identical (Palshikar, 2002).

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

3.1 K-Means Clustering

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

To do the clustering task of DM, there are different algorithms. These algorithms can be categorized as partitioning, hierarchal, Density-based, Grid-based, and model-based methods (Han and Kamber, 2006). As Han and Kamber (2006) indicated that despite the availability of these different methods, the most widely used one is K-means algorithm, which is the partitioning method.

Out of different reasons why the K-means clustering algorithm is chosen for conducting the clustering process is this research described as follows:

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

3.1.1 K-Means Algorithm

Bramer (2007) defines that *k*-means clustering is an exclusive clustering algorithm. Each object is assigned to precisely one of a set of clusters. (There are other methods that allow objects to be in more than one cluster.)

For this method of clustering we start by deciding how many clusters we would like to form from our data. We call this value *k*. The value of *k* is generally a small integer, such as 2, 3, 4 or 5, but may be larger.

According to Han and Kamber (2006) the *k*-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. Cluster similarity is measured in regard to the *mean* value of the objects in a cluster, which can be viewed as the cluster's Centroid or center of gravity.

The K-Means algorithm is an iterative approach to finding K clusters based on distance (Berry and Linoff, 2004). The algorithm divides a dataset into a predefined number of 'K' clusters. 'K' in the K-Means refers to the number of segments to partition the dataset, while 'means' refers to the average

location of all of members (which are records from a database) of a particular cluster. The K-Means algorithm "self-organizes" to create clusters. The algorithm basically has three steps to do the clustering task (Berry and Linoff, 2004). The first step, the algorithm randomly selects K data points to be the seeds. Each of the seeds is an embryonic cluster with only one element. The second step assigns each record to the closest seed. One way to do this is by finding the boundaries between two clusters. The boundaries between two clusters are the points that are equally close to each cluster. Lastly, the third step is to calculate the Centroid of the clusters; these now do a better job of characterizing the clusters than the initial seeds finding the Centroid is simply a matter of taking the average value of each dimension for all the records in the cluster. The Centroid becomes the seeds for the next iteration of the K-Means algorithm.

The second step is repeated, and each point is assigned to the cluster with the closest Centroid. The process of assigning points to a cluster and then recalculating Centroid continues until the cluster boundaries stop changing. In practice, K-Means algorithm usually finds a set of stable clusters after a few iterations.

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

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

Some implementation of K-Means only allows numerical values for attributes (Nimmagadda et al. 2011). In that case, it may be necessary to convert the dataset into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales (e.g. “protein” and “fat”). While WEKA provides filters to accomplish all of these preprocessing tasks, they are not necessary for clustering in WEKA. This is because WEKA Simple K-Means algorithm automatically handles a mixture of categorical and numerical attributes. Furthermore, the algorithm automatically normalizes numerical attributes when

doing distance computations. The WEKA Simple K-Means algorithm uses Euclidean distance measure to compute distances between instances and clusters.

Let us apply the k-Means clustering algorithm by example: -

Food item #	**Protein content, P**	**Fat content, F**
Food item #1	1.1	60
Food item #2	8.2	20
Food item #3	4.2	35
Food item #4	1.5	21
Food item #5	7.6	15
Food item #6	2.0	55
Food item #7	3.9	39

The four clusters chosen are:

Cluster number	**Protein content, P**	**Fat content, F**
C1	1.1	60
C2	8.2	20
C3	4.2	35
C4	1.5	21

Also, we observe that point 1 is close to point 6. So, both can be taken as one cluster. The resulting cluster is called C16 cluster. The value of **P** for C16 Centroid is (1.1 + 2.0)/2 = 1.55 and **F** for C16 Centroid is (60 + 55)/2 = 57.50.

Upon closer observation, the point 2 can be merged with the C5 cluster. The resulting cluster is called C25 cluster. The values of **P** for C25 Centroid is (8.2 + 7.6)/2 = 7.9 and **F** for C25 Centroid is (20 + 15)/2 = 17.50

The point 3 is close to point 7. They can be merged into C37 cluster. The values of **P** for C37 Centroid is (4.2 + 3.9)/2 = 4.05 and **F** for C37 Centroid is (35 + 39)/2 = 37.

The point 4 is not close to any point. So, it is assigned to cluster number 4 i.e., C4 with the value of **P** for C4 Centroid as 1.5 and **F** for C4 Centroid is 21.

Finally, four clusters with three Centroids have been obtained.

Cluster number	Protein content, P	Fat content, F
C16	1.55	57.50
C25	7.9	17.5
C37	4.05	37
C4	1.5	21

In the above example it was quite easy to estimate the distance between the points. In cases in which it is more difficult to estimate the distance, one has to use *Euclidean metric* to measure the distance between two points to assign a point to a cluster.

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

3.2 Decision Tree Classification Technique

A "divide- and – conquer" approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree. Nodes in a decision tree involve testing a particular attribute (Witten and Frank, 2005). Usually, the test at a node compares an attribute value with a constant. However, some trees compare two attributes with each other, or use some function of one or more attributes. Leaf nodes give a classification that applies to all instances that reach the leaf or a set of classifications, or a probability distribution over all possible classifications. To the leaf classify an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached the instance is classified according to the class assigned to the leaf.

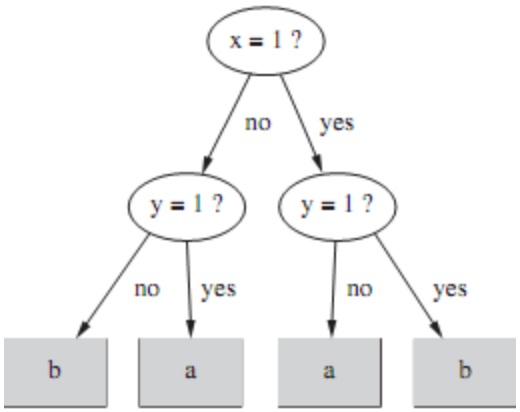

Fig 3.1 A Scenario that shows how decision tree is constructed

Trees can grow in any forms. They could be binary trees of non-uniform depth that is, each node has two children and the distance of a leaf to the root varies. In Fig 3.1 each node represents a 'yes' or 'no' question, the answer determines through which of the two paths a record proceeds to the next level of the tree.

Tree induction

Tree induction is the learning of decision trees class-labeled training tuples. A decision tree is a flow chart- like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node (Han and Kamber, 2006).

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

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

Stopping criteria: Tree-building algorithms usually have several stopping rules. These rules are usually based on several factors including maximum tree depth, minimum number of elements in a node considered for splitting, or it is near equivalent, the minimum number of elements that must be in a new node. In most implementations the user can alter the parameters associated with these rules. Some algorithms, in fact, begin by building tree to their maximum depth. While such a tree can precisely

predicate all the instances in the training set (except conflicting records), the problem with such a tree is that, more than likely, it over fits the data.

Pruning: When a decision tree is built many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over fitting the data. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data.

Different decision tree algorithms

Decision tree algorithms, such as ID3, C4.5, and CART, were originally intended for classification. Some decision tree algorithms produce only binary trees (where each internal node branches to exactly two other nodes), where as others can produce non binary trees (Han and Kamber, 2006). Differences in decision tree algorithms include how the attributes are selected in creating the tree and the mechanisms used for pruning.

Target variables: The target or dependent variable is being categorical for most tree algorithms. Such algorithms require that continuous variables be binned (grouped) for use with regression.

Splits: A lot of algorithms support only the binary splits. Each parent node can split into at most two child nodes. Other algorithms generate more than two splits and produce a branch for each value of categorical variables.

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

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

Because of the following reasons the decision classification was selected:

- ✓ Decision trees are easy to understand
- ✓ Decision trees are easily converted to a set of prediction rules
- ✓ Decision trees can classify both categorical and numerical data, but the output attribute must be categorical

- ✓ There are no a priori assumptions about the nature of the data.

Different decision tree algorithms were discussed in chapter two and C4.5 is among one of the algorithms that includes methods to generalize rules associated with a tree.

3.2.1 The J48 Decision Tree Algorithm

J48 adopt an approach in which decision tree models are constructed in a top-down recursive divide-and –conquer manner. Witten and Frank (2005) emphasized the importance of understanding the variety of options during implementation of J48 algorithm. J48 decision tree algorithm is a predictive machine learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. It can be applied on discrete data, continuous or categorical data (Bharti et al., 2010).

J48 is the decision tree algorithm that is used in this study to classify the tax claims as fraud or non-fraud suspicious claims. The J48 decision tree can serve as a model for classification as it generates simpler rules and remove irrelevant attributes at a stage prior to tree induction. In several cases, it was seen that J48 decision trees had a higher accuracy than other algorithms (Witten and Frank, 2005). J48 offer also a fast and powerful way to express structures in data.

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

J48 employs two pruning methods (Witten and Frank, 2000). Sub-tree replacement is the first one. This means that nodes in a decision tree may be replaced with a leaf basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Sub-tree rising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that sub-tree rising can be somewhat computationally complex.

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

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

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

a) Choose an attribute that best differentiates the output attribute values.
b) Create a separate tree branch for each value of the chosen attribute.
c) Divide the instances into subgroups so as to reflect the attribute values of the chosen node.
d) For each subgroup, terminate the attribute selection process if:
    1) All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with specified value.
    2) The subgroup contains a single node or no further distinguishing attributes can be determined. As in (1), label the branch with the output value seen by the majority of remaining instances.
e) For each subgroup created in (c) that has not been labeled as terminal, repeated the above process. The algorithm is applied to the training data. The created decision tree is tested on a test

dataset, if available. If test data is not available, J48 performs a cross-validation using the training data itself.

3.3 Naïve Bayes Classification Technique

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class (Han and Kamber, 2000).

Bayesian classification is based on Bayes theorem. Studies comparing classifications algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

Observations show that Naïve Bayes performs consistently before and after reduction of number of attributes (Anbarasi et al. 2010). Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. Naïve Bayes analyses the relationship between each input attribute and the dependent attribute to derive a conditional probability for each relationship (Ibrahim, 1999).

As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2)= ----=P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)$ $P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = (C_{i,} D|/|D|$, where $|C_{i,} D|$ is the number of training tuples of class C_i in D.

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

3.3.1 Naïve Bayes Algorithm

Naïve Bayes implements the probabilistic Naïve Bayes classifier. Naïve Bayes Simple uses the normal distribution to model numeric attributes. Naïve Bayes can use kernel density estimators, which improves performance if the normality assumption is grossly incorrect; it can also handle numeric attributes using supervised discretization. Naïve Bayes Updateable is an incremental version that processes one instance at a time; it can use a kernel estimator but not discretization. Naïve Bayes Multinomial implements the multinomial Bayes classifier.

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

1. Let D be a training set of tuples and their associated class labels. Each tuple is represented by an n-dimensional attribute vector, $X = (x_1 , x_2,\ldots x_n)$, depicting **n** measurements made on the tuple from **n** attributes, respectively, $A_1, A_{2,\ldots} A_n$.
2. Suppose that there are **m** classes, $C_1, C_2,\ldots\ldots C_m$. Given a tuple, **X**, the classifier will predict that **X** belongs to the class having the highest posterior probability, conditioned on **X**. That is, the Naïve Bayesian classifier predicts that tuple **X** belongs to the class $\mathbf{C_i}$ if and only if
   $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, j \neq i$.
   Thus we maximize $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$
3. As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \ldots = P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = |C_i,D|/|D|$, where $|C_i,D|$ is the number of training tuples of class C_i in D.
4. Given datasets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the Naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that are no dependence relationships among the attributes). Thus,

$$p(X \mid C_j) \propto \prod_{k=1}^{d} p(x_k \mid C_j)$$

$= P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots\ldots\ldots \times P(x_n \mid C_i)$

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

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

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

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

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

CHAPTER FOUR

BUSINESS AND DATA UNDERSTANDING

4.1 Introduction to ERCA

ERCA is the body responsible for collecting revenue from customs duties and domestic taxes in Ethiopia. In addition to raising revenue, ERCA is responsible to protect the society from adverse effects of smuggling. It seizes and takes legal action on the people and vehicles involved in the act of smuggling while it facilitates the legitimate movement of goods and people across the border. ERCA is established on July 7, 2008 as a result of the merger of the Ministry of Revenues, the Ethiopian Customs Authority and the Federal Inland Revenues into one giant organization (www.erca.gov.et).

According to article 3 of the proclamation No. 587/2008, the Authority is looked upon as "an autonomous federal agency having its own legal personality". The Authority came into existence on 14 July 2008, by the merger of the Ministry of Revenue, Ethiopian Customs Authority and the Federal Inland Revenue Authority who formerly were responsible to raise revenue for the Federal government and to prevent controband. Reasons for the merge of the foregoing administrations into a single autonomous Authority are varied and complex.

Some of those reasons include:

1. To provide the basis for modern tax and customs administrations
2. To cut through the red tape or avoid unnecessary and redundant procedure that results delay and considered cost-inefficient etc.
3. To be much more effective and efficient in keeping and utilizing information, promoting law and order, resource utilization and service delivery.
4. To transform the efficiency of the revenue sector to a high level.

ERCA has its headquarters in Addis Ababa. It is led by a Director General who reports to the Prime Minister and is assisted by five Deputy Director Generals, namely D/Director General for Program Designing of Operation and Development Businesses; D/Director General for Branch offices' Coordination and Support; D/Director General of Enforcement Division; D/Director General, Corporate

Functions Division; Change Management and Support Sector; and Enforcement Sector. Each deputy director general oversees at least four directorates. Both the Director General and the Deputies are appointed by the Prime Minister (ERCA booklet, 2008).

ERCA has divided the type of taxes as, indirect and direct tax.

Indirect taxes are: -

- VAT
- Excise
- Turnover

Direct taxes are: -

- Personal income tax
- Rental tax
- Business profit tax
- Withholding tax
- Other tax

The ERCA's direct and indirect tax types have been described in the following table.

Indirect Taxes	**Description**
VALUE ADD TAX (VAT)	To a product by a business is the sale price charged to its customer, minus the cost of materials and other taxable inputs.
TURNOVER TAX	It is an indirect tax, typically on an ad valorem basis, applicable to a production process or stage.
EXCISE TAX	Is commonly referred to as an inland tax on the sale, or production for sale, of specific goods; or, more narrowly, as a tax on a good produced for sale, or sold, within a country or licenses for specific activities.
Direct Taxes	**Description**
WITHHOLDING TAX	Is a government requirement for the payer of an item of income to withhold or deduct tax from the payment, and pay that tax to the government?
PERSONAL INCOME TAX	Every person deriving income from employment or other private organization or non-government organization.
RENTAL TAX	A tax that is imposed on the income from rental of buildings.
BUSINESS PROFIT TAX	A tax is imposed on commercial, professional or vocational activity or any other activity recognized as trade by commercial code of Ethiopia.
COST SHARING	A portion of total project or program costs related to a sponsored agreement that is contributed by someone other than the sponsor.
SCHED D-GAMES OF CHANCE	Every person deriving income from winning at games of chance/for example, lotteries, tom bolas, and other similar activities.

Table 4.1 *Types of taxes used for national budget (Information Technology Management Directorate)*

4.1.1BusinessDriver

The key business drivers have been identified as follows:

- Revenue collection.
- Trade facilitation.
- Enforcement and security.
- Reliable data and statistics.
- Process oriented management.
- Good governance.

The 30 branch offices in Ethiopia comprise 22 Customs Control stations, 50 Checkpoints and 153 Tax Centers. Tax Center means a tax collection station administered under a branch office and located in the vicinity of taxpayers. Although the key businesses of ERCA are the above-mentioned six tasks, this research focused on the revenue collection. The computer system that enables ERCA to administer the taxes mentioned above is called Standard Integrated Government Tax Administration System (SIGTAS). The system allows ERCA to administer all aspects of most domestic taxes including: registration, assessment, cashing, and auditing in one easy-to-use integrated system. The system was introduced in ERCA in December1997 and is presently operational both at the head office and branch offices level. One of the main activities of the authority is auditing the customers' financial statement. The audit process and program development directorate is working with Information Technology Management Directorate closely.

4.1.2 Tax Audit Process and Program Development Directorate

Tax audit process and program development directorate is organized under operation program design and development activities sector. This directorate has the following activities:-

- To change the tax and audit policy and strategy of the authority into practice. Follow and improve the performance.
- To create functional systems improved tax and audit activities.
- To perform special investigation audit and transfer the result to criminal investigation directorate.
- To perform other activities those are related with tax and audit.

4.2 ERCA Audit Risk Criteria

ERCA needs to devise several screens and modify some others in order to generate the report that gives the result for the Audit Risk Criteria. The agency needs to modify TR121AS- Maintain non-Individual Enterprises and TR121BS- Maintain Individual Enterprises for criteria involved in the date of commencement of the organization and their Business Activity. In addition, a screen needs to be devised to create the different benchmarks used by the report. Another screen will be devised to capture the details of the calculation of each benchmark while another will be needed to execute the calculation of the scores of the benchmarks to be used in the report Audit Risk Criteria Report.

The Report's objective is to assess the overall risk of a specific set of taxpayers by ordering them according to risk factors. The report's intention does not necessarily (although possible) target a specific tax type. The main purpose of the report is then to target who should be audited. A proper audit case would follow. It can also be used during an audit case to guide or support the auditor. Since schedules B and C (normal and mining) are the main taxes and their sum is used to define legally the annual turnover. The majority of the information captured for audit risk criteria comes from them. In addition, information needed for financial ratios come from the financial statements and the tax declarations that are mandatory to be filled for those 2 (two) types. It's important to note that although there is a tax type selection criterion in the report, it is limited to schedules B, C-normal and C-mining.

4.3 Business Understanding

ERCA auditors are auditing the tax payers companies which are in different level. ERCA audit task process owner developed the risk management criteria. Based on the taxpayers' financial statement, the auditors are auditing the companies. The taxpayers financial statement included 43 (forty three) criteria. Some of these are the business activity, Total Gross Income, Gross profit/ Loss, Total Expenses and Repair and Maintenance Expenses.

The researcher took two categories of tax payers. The taxpayers categorized based on their yearly income. Category 'A' taxpayers' income is more than 4,000,000 Birr. Category 'B' taxpayers' income is between 500,000 and 4,000,000. Each department has 5660 (Five thousand six hundred and sixty) customers or taxpayers. From each category taxpayers 20% that means around 1132 (One thousand and thirty two) should be audited yearly. All taxpayers should be audited within five years. Most taxpayers are registering their financial report on the financial statement form and submitted to the audit task

process owner. The financial statement forms standard is designed by the authority. For the day to day activities some companies are using peach tree or excel. For information exchange and reporting purpose networking both parts were better; however, there is no such kind of facility.

The authority's audit task process owner auditors are checking the taxpayers' income and expenditure based on the tax payers' financial report. Currently, the main problem in the tax collection activity is income evasive. The authority rates informally the high gross profit/loss, interest expenses, net income/loss, net tax due/refundable amount, non operating income, low profit income tax, repair and maintenance expenses, selling and distribution expenses, depreciation expenses, high total expenses, and low total gross income as fraud suspicious claims.

The auditors rates the low gross profit/loss, interest expenses, net income/loss, net tax due/refundable amount, non operating income, low profit income tax, repair and maintenance expenses, selling and distribution expenses, depreciation expenses, low total expenses, and low total gross income as non-fraud suspicious claims.

In addition to the aforementioned criteria's, which are used by experts for judging whether a tax claim is fraud suspicious or not, the type of tax can also be employed for the purpose of investigation of claims whether they are suspicious of fraud or not. The private companies are also considered for showing fraud suspicious claims mostly, while claims with government companies are mostly believed to be free of freak.

All the information gathered during claim processing and submitted to head of the audit task process owner. The head assigns auditor/s to investigate the case. After the investigation the auditor/s reports the result to the head. Based on the investigation result the head make decision. The authority can take the case to the court if necessary.

4.3.1 Tax Collection Handling Processes

During the establishment of the companies the owners of the companies start the process of register their organization to ERCA. While the companies register in ERCA, they have their own Tax Identification Number (TIN). This number is unique. Every tax payer should have this number. This number is used to uniquely identify a tax payer for tax collection purpose.

4.4 Understanding the Data

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

4.4.1 Initial Data Collection

For the purpose of data collection, the central database of ERCA was chosen which is found in Addis Ababa. All higher taxpayers' data is found in ERCA ITMD task process owner database. The sub cities financial bureaus collect tax from the small scale enterprises. The sub cities financial bureaus are not networked with ERCA ITMD task process owner database which is found around Mexico square.

The database was thoroughly studied and, as a result of the study, 15 (fifteen) attributes were found to be important. All the fifteen attributes were extracted in Excel format directly from the target database.

The number of records collected from the ITMD task processes and owner is summarized in Table 4.2.

Department	**Taxpayers category**	**Number of records collected**
Information Technology Management Directorate (ITMD)	A	5541
Information Technology Management Directorate (ITMD)	B	5541

Table 4.2 *Distribution of collected data with respect to taxpayers' level.*

As can be seen from Table 4.2 the distribution of the collected data from the taxpayers' level is balanced.

4.4.2 Description of Data Collected

As indicated before, the pertinent data to carry out the research is collected from ERCA ITMD database. These are Business income or sales/turnover, Gross profit/loss, Interest expenses, Net income/loss, Net tax due/refundable amount, Non operating and other income, Profit income tax payable, Repair and maintenance expenses, Selling and distribution expenses, Depreciation expenses, Total expenses, Total gross income, and Category. Though, the table has lots of attributes in the original dataset, the table below show initially selected attributes.

No.	Attribute Name	Data Type	Description
1	Business Income or Sales/Turnover	Number	It is an indirect tax, typically on an ad valorem basis, applicable to a production process or stage.
2	Gross profit/loss	Number	It is a company's residual profit or loss after selling a product or service and deducting the cost associated with its production and sale.
3	Interest expenses	Number	It is calculated as a percentage of the amount of debt for each period of time. Points paid for a mortgage are a form of prepaid interest.
4	Net income/loss	Number	Net income is calculated by starting with a company's total revenue. From this, the cost of sales, along with any other expenses that the company incurred during the period, is removed to reach earnings before tax.
5	Net tax due/Refundable amount	Number	A tax credit is not limited by the amount of an individual's tax liability. Typically a tax credit only reduces an individual's tax liability to zero.
6	Non operating income	Number	Income is received by a business that is not derived from operations, such as manufacturing. Non-operating income usually does not occur on an ongoing basis, and is examined separately from operating income.
7	Profit income tax	Number	It is a tax levied on the income of individuals or

			businesses (corporations or other legal entities).
8	Repair and Maintenance expenses	Number	The costs incurred to bring an asset back to an earlier condition or to keep the asset operating at its present condition (as opposed to improving the asset).
9	Selling and Distribution expenses	Number	Selling and distribution expenses include sales commissions, advertising, promotional materials distributed, rent of the sales showroom, rent of the sales offices, salaries and fringe benefits of sales personnel, utilities and telephone usage in the sales department, etc.
10	Depreciation expenses	Number	An annual deduction that allows taxpayers to recover the cost of property used in a trade or business or held for the production of income
11	Total expenses	Number	These costs consist primarily of management fees and additional expenses such as trading fees, legal fees, auditor fees and other operational expenses
12	Total gross income	Number	The total gross income is being taxable income from all sources, reduced by any adjustments can be taken. Here are the most common income and adjustment items.

Table 4.3 *Description of the taxpayers*

4.4.3 Data Quality of Taxpayers

The collected Net income, Non-operating and other income, and Selling and distribution expenses data contain missing and incomplete data. All companies didn't register total number of their employee. All companies have been focused on their total expenses. ERCA's ITM directorate is collecting data from the companies' financial report. During data entry mistake the financial reports are not complete. The auditors are asking the taxpayers repeatedly to get full information.

4.5 Preparation of the Data

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

4.5.1 Data Selection

The whole target dataset may not be taken for the DM task. Irrelevant or unnecessary data are eliminated from the DM database before starting the actual DM function. Originally there were around 11080 records. From this, around 1000 of the records are randomly selected for testing purpose. The rest of the dataset is used for training purpose. So, after eliminating irrelevant and unnecessary data only a total of 11080 datasets are used for the purpose of conducting this study.

The above elaborated table of the ITMD database consists of more than 30 attributes. The first task was to remove from the database those fields or attributes, which were irrelevant to the task at hand. As shown in Appendix 1, the following are the initial sets of attributes, which are further preprocessed to select the final attributes used in the study. CODE, ENTERPRISE START DATE, BUSINESS ACTIVITY, SUBCITY, BUSINESS INCOME OR SALES/TURNOVER, GROSS PROFIT/LOSS, INTEREST EXPENSES, NET INCOME/LOSS, NET TAX DUE/ REFUNDABLE AMOUNT, NON OPERATING AND OTHER INCOME, PROFIT INCOME TAXPAYABLE, REPAIR AND MAINTENANCE EXPENSES, SELLING AND DISTRIBUTION EXPENSES, DEPRECIATION EXPENSES, TOTAL EXPENSES, TOTAL GROSS INCOME CATALOGE, AND YEAR were the initial set of attributes that are selected.

4.5.2 Data Cleaning

Data Cleaning is a process which fills in missing values, removes noise and corrects data inconsistency. Usually, real world database contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2006). Consequently, data cleaning has become a must in order to improve the quality of data so as to improve the performance of the accuracy and efficiency of the data mining techniques.

This technique involves removing the records that had incomplete, noise (invalid) data and filling missing values under each column. Removing of such records was done as the records with this nature are few and their removal does not affect the entire dataset.

As a result, the researcher use MS-Excel 2007 built-in functions like search and replace, filtering, and auto fill mechanisms, and WEKA to identify and fill missing value.

4.5.2.1 Missing Value Handling

Missing values refer to the values for one or more attributes in a data that do not exist. In real world application data are rarely complete. It can also be a particularly pernicious problem. Especially when the dataset is small or the number of missing fields is large, not all records with a missing field can be deleted from the sample.

Moreover, the fact that a value is missing may be significant itself. A widely applied approach is used to calculate a substitute value for missing fields, for example, the median or mean of a variable (Famili and Turney, 1997). Accordingly, the researcher has been analyzed the taxpayers' dataset and identified missing values and take measure to solve the problem as follows. The total records are 11080 and the missing values are 3180. From the total amount the missing data contains 28.7%. Net income/loss, profit income tax, Total gross income and depreciation expenses attribute which of them accounts of 10%, 10%, 15% and 8% respectively. To handle the problem of missing values of numerical data type attributes were recommended to be replaced by the mean of that value a whole (Two Crows Corporation, 2005).

Therefore, based on the above principles the researcher handles the missing value and WEKA preprocessing replace missing value techniques also used. WEKA fills using the most frequent (model) value methods which is same as the above principle. Additionally, manually tracing and fixing the missed value is other technique used by researcher.

As shown below in table 4.4 out of the selected 15 attributes 4 of them have registered with missing values. Accordingly, the researcher reacts to take appropriate action to clean the data.

Number	Attribute name and their data type	% of missing value	Data types	Reason/ Technique applied
1	Net income	10.12	Numeric	The mean of this attribute
2	Profit income tax	10.25	Numeric	The mean of this attribute
3	Total gross income	15	Numeric	The mean of this attribute
4	Depreciation expenses	8	Numeric	The mean of this attribute

Table 4.4 *Handling missing value*

The missing values were occurred by two reasons; the first one during data entry the clerk of the ITMD made a mistake and the other reason was the financial statements which are not filled by the taxpayers'.

4.5.2.2 Handling Outlier Value

The data stored in a database may reflect outlier – noise, exceptional case, or incomplete data object and random error in a measure of variable. These incorrect attribute values may be due to data entry problems, faulty data collection, inconsistency in naming convention or technology limitation (Han and Kamber, 2006). The authors have also explained four basic methods for handling of noise data. These are binning method, clustering, regression and combined computer and human inspection.

In this research, the researcher has identified and detected noise or outlier value from the taxpayers' data. With the help of domain experts the identified outlier was corrected manually. Accordingly, a combined effort of the researcher and domain expert were taken to identify and correct the problem of the incompleteness, noise or outlier.

4.5.3 Data Construction

The other important step in preprocessing is deriving other fields from the existing ones. Adding fields that represent the relationships in the data are likely to be important in increasing the chance of the knowledge discovery process yield useful result (Berry and Linoff 2000). In consultation with the domain experts at ERCA, the following fields that are considered essential in determining the fraudulence of claims were derived from the existing fields. Net Worth (this refers net resources of the company) is derived from the Total Asset and Total Liability columns of the database. Liquid Cash (refers for currently available cash of the company) is derived from the Current Asset and Current Liability. Net book value (refers the current price of the material) is derived from the Fixed Asset and Accumulated Depreciation.

Net Worth= **Total Asset-Total Liability**

Liquid Cash= **Current Liability-Current Asset**

Net Book Value= **Fixed Asset-Accumulated Depreciation**

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

No.	Attribute Name	Data Type	Description
1	Business Income or Sales/Turnover	Number	It is an indirect tax, typically on an ad valorem basis, applicable to a production process or stage.
2	Gross profit/loss	Number	It is a company's residual profit or loss after selling a product or service and deducting the cost associated with its production and sale.
3	Interest expenses	Number	It is calculated as a percentage of the amount of debt for each period of time. Points paid for a mortgage are a form of prepaid interest.
4	Net income/loss	Number	Net income is calculated by starting with a company's total revenue. From this, the cost of sales, along with any other expenses that the company incurred during the period, is removed to

			reach earnings before tax.
5	Net tax due/Refundable amount	Number	A tax credit is not limited by the amount of an individual's tax liability. Typically a tax credit only reduces an individual's tax liability to zero.
6	Non operating income	Number	Income is received by a business that is not derived from operations, such as manufacturing. Non-operating income usually does not occur on an ongoing basis, and is examined separately from operating income.
7	Profit income tax	Number	It is a tax levied on the income of individuals or businesses (corporations or other legal entities).
8	Repair and Maintenance expenses	Number	The costs incurred to bring an asset back to an earlier condition or to keep the asset operating at its present condition (as opposed to improving the asset).
9	Selling and Distribution expenses	Number	Selling and distribution expenses include sales commissions, advertising, promotional materials distributed, rent of the sales showroom, rent of the sales offices, salaries and fringe benefits of sales personnel, utilities and telephone usage in the sales department, etc.
10	Depreciation expenses	Number	An annual deduction that allows taxpayers to recover the cost of property used in a trade or business or held for the production of income
11	Total expenses	Number	These costs consist primarily of management fees and additional expenses such as trading fees, legal fees, auditor fees and other operational expenses
12	Total gross income	Number	The total gross income is being taxable income from all sources, reduced by any adjustments can be taken. Here are the most common income and adjustment items.

13	Net Worth	Number	It indicates that net resource of the company.
14	Liquid cash	Number	It shows that currently available cash.
15	Net book value	Number	It describes the current price of the material.

Table 4.5 *The Final List of Attributes used in the Study*

4.5.4 Data Integration

The data integration process was done before deriving the attributes. As described before the dataset for the database which was discussed above was available in different excel files. Data integration method for retrieving important fields from different files was done in the effort to prepare the data ready for the DM techniques to be undertaken in this research. The Oracle and SIGTAS databases were used to carry out the data integration process. These data integration process took a lot of time of the research. This was because of the reason that when the different excel files were combined together the size of the dataset increased by ten times. That means the records were duplicating tremendously. The integrated files are on table 4.5 showed fifteen attributes including Net Worth, Liquid Cash, and Net Book Value. Finally the data is integrated and put together into a single excel file.

4.5.5 Data Formatting

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

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

4.5.6 Attribute Selection

Most machine learning algorithms are designed to learn the most appropriate attributes to use for making their decisions (Witten and Frank, 2005). However, algorithms have their own limitations, such as considering insignificance attribute as crucial one and ignoring the most important attribute as irrelevant. Due to this limitation, the attributes for fraudulent protection should be selected after they

have been tested in the authority from business perspective. So, different models were built having different values of 'K', different attributes number and different iteration to identify attributes which have high information content to cluster tax payers and that would enable to develop a better clustering model according to the authority business objective. In order to select the best attributes from the initial collected dataset, the researcher evaluates the information content of the attributes using the select attribute techniques of WEKA with GainRationAttributeEval attribute evaluator and Ranker search method. So, there were variables to be discarded. However, it does not mean that these attributes have no importance rather these variables were believed to provide very little useful information for the problem at hand and since clustering is unsupervised learning they may reduce the accuracy of the clustering algorithm.

All attributes listed in the dataset were used for attribute selection experiments. The number of attributes in different experiments was similar. And the number of iteration (*i*) refers to the maximum number of times the algorithm reads the dataset to create clusters was also similar.

Finally, the selected attributes are the followings: - Gross profit/loss, Net worth, Interest expenses, Net income/loss, Net tax due/Refundable amount, Liquid Cash, Non operating income, Profit income tax, Repair and Maintenance expenses, Selling and Distribution expenses, Depreciation expenses, Net book value, Total expenses, and Total gross income.

CHAPTER FIVE

EXPERIMENTATION

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

The description and evaluation of the performances of the clustering and classification models developed are presented. It applies the methods, techniques and algorithms of DM that are briefly discussed in section 3.1 of chapter Three to accomplish the objective of the research.

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

5.1 Experiment Design

A procedure or mechanism of how to test the model's quality and validity is needed to be set before the model is actually built. In order to perform the model building process of this research, 11080 training dataset is used to train the clustering and classification models. Once the clustering model is developed, the resulting clustered dataset is then used as an input for training the J48 decision tree and naïve Bayes models. For validating the clustering result of this research the intra cluster similarity measure (within cluster sum of squared error) value, the number of iteration the algorithm has undergone to coverage and the domain experts' judgment is used. A threshold value is set to determine what patterns are discovered for each subsequent cluster models, which helps to identify and label the cluster dataset based on the fraudulence nature of tax claims. The 10-fold cross validation and percentage split test options are used for training and testing the classification model. These testing dataset is prepared by simple random sampling techniques from the original dataset.

5.2 Cluster Modeling

Once the dataset is ready to be used, the next step is building the clustering model using the selected DM tool. As it was discussed before, the WEKA version 3.7.5 (for the Mac OS) software is used for conducting this research. The WEKA 3.7.5 explorer includes different parameters for K-Means clustering. Some of the basic parameters are discussed as follows:

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

The clustering task of segmenting tax claims is done using the WEKA simple K-Means algorithm. The number of clusters chosen should be driven by how many clusters the business can manage. Accordingly, the business experts have been consulted in setting the optimal value. They have suggested that the K value to be 2 (two) (representing FRAUD suspicious and NON-FRAUD suspicious tax claims). This cluster model is experimented and evaluated against its performance in creating dissimilar clusters/segments when the default parameters are changed. Halkidi and Vazirgiannis (2001) stated that the notion of "good" clustering is strictly related to the application domain and its specific requirements. Nevertheless the generally accepted criterions for validating the clustering results in different domains are the measures of separation among the clusters and cohesion within clusters (i.e. inter and intra cluster similarities respectively). So, for validating the clustering result of this research the intra cluster similarity and the measure (within cluster sum of squared error) value is used. In addition to this the number of iteration the algorithm has undergone to converge and the domain experts' judgment is used. Once the necessary parameters are set, the experiment is undertaken in a stepwise manner.

Before conducting the experiment, the threshold value is set for each numeric attributes used to build the clustering model. The threshold value for each attributes has been determined together with the domain experts in the audit department and with the aid of the WEKA's minimum, maximum, and mean values displayed for each attribute. The need for determining the threshold values is solely to determine what patterns are discovered for each subsequent cluster models with K=2, and changing the other default

parameters. This helps a lot to indentify fraud suspicious segments easily. Table 5.1 depicts the threshold values set for each of the variables.

Gross Profit/Loss	Profit Income tax	Liquid Cash	Net Worth	Net Book Value
GP/L<=99,999 Low	PIT<=99,999 Low	LC<=99,999 Low	NW<=99,999 Low	NBV<=99,999 Low
GP/L<=999,999 Medium	PIT<=999,999 Medium	LC<=999,999 Medium	NW<=999,999 Medium	NBV<=999,999 Medium
GP/L>=1,000,000 High	PIT>=1,000,000 High	LC>=1,000,000 High	NW>=1,000,000 High	NBV>=1,000,000 High

Table 5.1 *List of range of conditions (thresholds) used to assess the cluster result*

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

5.2.1 Experimentation I

Before starting this experimentation part, the researcher believes that it is important to mention the fact that there was a discussion with the experts at the ERCA. This discussion focused on assessing the influential factors for being a tax payer. Generally the experts were discussing some of the most important variables that are used to select variables. Thus the researcher would like to point out the important points raised by the experts in the following paragraphs.

There is no actual quantitative definition of a good segmentation output, assessing the clusters based on certain decisive attribute is sensible (Pritscer and Hans, 2008). Thus the attributes “Gross Profit/Loss”, “Liquid cash”, “Net book value”, “Profit income tax”, “Net worth”, “Interest expense”, “Net income/Loss”, “Net tax due/Refundable amount”, “Non operating income”, “Repair & maintenance expenses”, “Selling & distribution expenses”, “Depreciation expenses”, “Total expenses”, and “Total gross income” were given a very high weight by the experts. Consequently, in the experimentation part the analysis and interpretation of each and every cluster was highly dependent on these attributes. But

this doesn’t mean that the rest of the attributes have no importance, rather it is to note the weight given to these variables in the real world by the experts. As the experts explained, if a tax payer financial statement report has the following characteristics as having a higher probability to be fraud suspicious.

Gross Profit/Loss, low, Interest Expenses, low, Net income/Loss, low, Net tax due/ Refundable amount, low, Non operating income, low, Profit income tax, low, Repair & maintenance expenses, medium, Selling & distribution expenses, low, Depreciation expenses, low, Total expenses, low, Total gross income, high, Liquid cash, low, Net worth, high, Net book value, high depicted in Appendix 2-A.

On the other hand, if a tax payer financial statement report has the following characteristics as having a higher probability to be Non- fraud suspicious.

Gross Profit/Loss, low, Interest Expenses, low, Net income/Loss, low, Net tax due/ Refundable amount, low, Non operating income, low, Profit income tax, low, Repair & maintenance expenses, low, Selling & distribution expenses, low, Depreciation expenses, low, Total expenses, high, Total gross income, high, Liquid cash, low, Net worth, medium, Net book value, medium.

The first experiment is done for K=2 with the default seed value and default distance function. All of the final selected attributes and 11080 records are used as an input for the experiment. In order to cluster the records based on their values, the model is trained by using the default values of the program. The use training set cluster model is employed to make use of the dataset for training. Table 5.2 exhibits the result of the first experiment and the resulting segments. The algorithm is instructed to segment the dataset into two clusters. As can be seen from this table, the distribution of the dataset for each cluster is presented.

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

Clustering Result of the 1st Experiment				
K	**Distance Function**	**Seed Value**	**Cluster Distribution**	
			C1	**C2**
2	EuclideanDistance	10	6323 (57%)	4757 (43%)

Table 5.2 *Training of the first experiment by the default parameter values*

As can be seen from Table 5.2, the first experiment is conducted with default values of the simple K-Means algorithm (K = 2, Seed = 10, and Euclidean distance function). Table 5.3 shows results of the first experiment.

Clustering Result of the First Experiment															
Cluster #	Dist. of instances (In %)	GP/L	IE	NI/L	NTD/RA	NOI	PIT	RME	SDE	DE	TE	TGI	LC	NW	NBV
1	6323(57%)	L	L	L	L	L	L	L	L	L	L	M	L	M	M
2	4757(43%)	L	L	L	L	L	L	M	L	L	L	H	L	H	H

Table5.3 *Cluster result of the first experiment for K=2, Seed = 10, Euclidean distance function*

As can be seen from Table 5.3, the discrete value is generated from the clustering algorithm. After the average value of each attribute in each cluster has been replaced with the corresponding discrete values, a description for each segment of the cluster has been done as shown in Table 5.4. The ranking is determined based on the fraudulence nature of the audit claims.

Cluster #	**Description**	**Rank**
1.	Gross Profit/Loss, low, Interest Expenses, low, Net income/Loss, low, Net tax due/ Refundable amount, low, Non operating income, low, Profit income tax, low, Repair & maintenance expenses, low, Selling & distribution expenses, low, Depreciation expenses, low, Total expenses, high, Total gross income, high, Liquid cash, low, Net worth, medium, Net book value,	2

	medium.	
2.	Gross Profit/Loss, low, Interest Expenses, low, Net income/Loss, low, Net tax due/ Refundable amount, low, Non operating income, low, Profit income tax, low, Repair & maintenance expenses, medium, Selling & distribution expenses, low, Depreciation expenses, low, Total expenses, low, Total gross income, high, Liquid cash, low, Net worth, high, Net book value, high.	1

Table 5.4 *Cluster summary of the first experiment for K = 2, seed = 10, Euclidean distance and rank of clusters*

As stated in the business understanding section of Chapter Four, suspicious claims are those with high gross profit/loss, low profit income tax, high total expenses, and total gross income.

So, table 5.4 ranking of the clusters has been assigned depending on the aforementioned facts of the business. As described in Table 5.4 the first cluster is ranked second. It is because of the reason that the claims that are assigned in this cluster are reported within a low gross profit/loss, with low profit income tax, high total expenses, and high total gross income.

The second cluster is ranked as first because of the reason that the claims that are grouped in this cluster have low gross profit/loss, with low profit income tax, low total expenses, and high total gross income. Generally, cluster 2 (ranked 1st) is considered as containing those suspicious fraudulent audit claims, while cluster 1 (ranked 2nd) is considered as containing non- fraudulent audit claims.

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

5.2.2 Experimentation II

The second cluster experiment is done for K = 2, changed default seed value (10) to 100 and with the default distance function (Euclidean distance). Similar to the first experiment, all of the final selected

attributes and 11080 records are used as an input for conducting the experiment. Table 5.5 exhibits the result of the second experiment and the resulting segments.

Clustering Result of the 2nd Experiment				
K	**Distance Function**	**Seed Value**	**Cluster Distribution**	
			C1	**C2**
2	Euclidean Distance	100	7884 (71%)	3196 (29%)

Table 5.5 *Training of the second experiment by changed Seed value = 100 and other default parameter Values*

The result of this cluster experiment with K = 2, seed = 100 and Euclidean distance function is depicted in Table 5.6. The description of the resulting cluster model is then interpreted and briefly explained. The table below shows the result of this experiment.

Clustering Result of the Second Experiment															
Cluster #	Dist. of instances (in %)	GP/L	IE	NI/L	NTD/RA	NOI	PIT	RME	SDE	DE	TE	TGI	LC	NW	NBV
1	7884(71%)	L	L	L	L	L	L	M	L	L	L	M	L	M	M
2	3196(29%)	L	L	L	L	L	L	M	L	L	L	H	L	H	H

Table 5.6 *Cluster result of the 2nd experiment for K = 2, Seed = 100, Euclidean distance function*

Similar to the first cluster experiment, two clusters are formed in the second experiment. This cluster has resulted in similar segment formation with the first experiment result. Table 5.7 below shows the description of the second experimentation. The result of the attribute and the value are on the left side and on the right side respectively.

Cluster #	**Description**	**Rank**
1.	Gross Profit/Loss, low, Interest Expenses, low, Net income/Loss, low, Net tax due/ Refundable amount, low, Non operating income, low, Profit income tax, low, Repair & maintenance expenses, low, Selling & distribution expenses, low, Depreciation expenses, low, Total expenses, high, Total gross income, high, Liquid cash, low, Net worth, medium, Net	2

	book value, medium.	
2.	Gross Profit/Loss, low, Interest Expenses, low, Net income/Loss, low, Net tax due/ Refundable amount, low, Non operating income, low, Profit income tax, low, Repair & maintenance expenses, medium, Selling & distribution expenses, low, Depreciation expenses, low, Total expenses, low, Total gross income, high, Liquid cash, low, Net worth, high, Net book value, high.	1

Table 5.7 *Cluster summary of the 2nd experimentation for K = 2, Seed = 100, Euclidean distance function and rank of clusters*

As can be seen from Table 5.7 the second cluster is ranked 1st. It is because of the reason that the claims that are assigned in this cluster have gross profit/loss, low, profit income tax, low, total expenses, low, total gross income, high, liquid cash, low, net worth, high, and net book value, high.

The first cluster is ranked second because it has gross profit/Loss, low, Gross Profit/Loss, low, Interest Expenses, low, Net income/Loss, low, Net tax due/ Refundable amount, low, Non operating income, low, Profit income tax, low, Repair & maintenance expenses, medium, Selling & distribution expenses, low, Depreciation expenses, low, Total expenses, low, Total gross income, low, Liquid cash, low, Net worth, medium, Net book value, medium. Cluster two is considered as suspicious of fraudulent claims while cluster one is considered as suspicious of non-fraudulent claims. These segmented clusters represent lower fraudulent tax cases.

Compared with the first experimentation, the value of within clustered sum of squared error is lowered in this experiment. Apart from this the number of iteration that the algorithm used to converge is also minimized from 43% to 29%. The number of non-fraudulent tax claims is also higher than the fraud suspicious claims in this experimentation. This showed that the result of this experiment is better than the first one in creating dissimilar clusters.

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

5.2.3 Experimentation III

The final cluster experiment is done for K = 2, changed default seed value (1000) and distance function (Euclidean distance). Similar to the first two runs, all of the final selected attributes and 11080 records are used as an input for conducting the experiment. This experiment is conducted with changed default distance function and seed value. Table 5.8 exhibits the result of the third cluster experiment, with K = 2, seed = 1000, and Manhattan distance function.

Clustering Result of the 3rd Experiment				
K	**Distance Function**	**Seed Value**	**Cluster Distribution**	
			C1	**C2**
2	Manhattan Distance	1000	5605 (51%)	5475 (49%)

Table 5.8 *Training of the third cluster experiment with K = 2, Seed = 1000 and Manhattan distance function*

The result of this cluster experiment with K = 2, seed = 1000, and Manhattan distance function is presented in table 5.9. The description of the resulting cluster model is then interpreted and briefly explained.

Clustering Result of the 3rd Experiment															
Cluster #	Dist. of instances (in %)	GP/L	IE	NI/L	NTD/RA	NOI	PIT	RME	SDE	DE	TE	TGI	LC	NW	NBV
1	5605(51%)	L	L	L	L	L	L	L	L	L	L	L	L	L	L
2	5475(49%)	L	L	L	L	L	L	M	L	L	L	H	L	H	H

Table 5.9 *Cluster result of the third experiment for K = 2, Seed = 1000 Manhattan distance function*

Similar to the first two cluster experimentations, two clusters are formed in this experiment. This cluster experiment has resulted in similar segment formation with the first and second experiment results. Table 5.10 shows the description of the third cluster experimentation.

Cluster #	**Description**	**Rank**
1	Gross Profit/Loss, low, Interest Expenses, low, Net income/Loss, low, Net tax due/ Refundable amount, low, Non operating income, low, Profit	2

	income tax, low, Repair & maintenance expenses, medium, Selling & distribution expenses, low, Depreciation expenses, low, Total expenses, low, Total gross income, medium, Liquid cash, low, Net worth, medium, Net book value, medium.	
2	Gross Profit/Loss, low, Interest Expenses, low, Net income/Loss, low, Net tax due/ Refundable amount, low, Non operating income, low, Profit income tax, low, Repair & maintenance expenses, medium, Selling & distribution expenses, low, Depreciation expenses, low, Total expenses, low, Total gross income, high, Liquid cash, low, Net worth, high, Net book value, high.	1

Table 5.10 *Cluster Summary of the third Experiment for K = 2, seed = 1000, Manhattan distance function and rank of clusters*

As can be seen from Table 5.10, the second cluster is ranked 1st because of the reason that the claims that are segmented in this class have a low gross profit/loss, low total expenses, high total gross income, low liquid cash, high net worth, and high net book value. The first cluster is ranked as 2nd because it has a low gross profit/loss, low total expenses, medium total gross income, low liquid cash, medium net worth, and medium net book value. The second cluster seems to show suspicious fraudulent tax claims, while the first cluster shows suspicious non-fraudulent claims.

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

5.2.4 Choosing the Best Clustering Model

Different experiments of the K-means algorithm were conducted with K = 2 and with changed default seed values and distance function. The entire dataset output from the different cluster experiments was available together with their segment distribution and the resulting cluster output. This enabled the domain experts to compare resulting tax claim segments from the different cluster experiments. Though

experimentation was carried out for different seed values and distance function, only those that resulted in a good segmentation are presented.

As described in Section 5.2 of this chapter, the cluster validity is a very difficult issue and subject of endless arguments since the view of good clustering is strictly related to the application domain and its specific requirements. But usually in most domains the intra and inter cluster similarity values are used for the validation of good clusters. The value of within cluster sum of squared error is used to evaluate the goodness of clustering in the WEKA DM tool. The lower value indicates that the records segmented within the same cluster are more related with each other. In addition to this the discovery of patterns requires that there is close interaction with domain experts, which allows them to interact with the output. So, the evaluation of the final clustering results also incorporated the suggestion of domain experts. Generally, in this study the best cluster model from the three cluster experimentations has been chosen based on the following criteria (Halkidi and Vazirgiannis, 2001).

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

The summery of the result of these criterions is depicted in Table 5.11

Experimentation	**Number of Iteration**	**Within cluster sum of squared Error**
I	3	3429
II	3	2724
III	3	6440

Table 5.11 *Within cluster sum of squared error values of the three cluster experimentations*

As can be seen from Table 5.11, the second cluster experiment shows the least number of iterations and within cluster sum of squared errors compared with the first and the third cluster experimentations. This shows that the second experiment is good in creating dissimilar clusters. In

addition to this the domain experts are consulted to give their suggestion whether the clustering result matches with the business. Accordingly, the experts suggested that the model of the second cluster experiment is good in segmenting the different fraud suspicious tax claims compared to the other models developed by the first and third cluster experimentations. So the model developed in the second cluster experiment is selected as the final clustering model of this study.

5.3 Classification Modeling

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

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

The training of the decisions tree classification models of the experimentation is done by employing the 10-fold cross validation and the percentage split classification models. The classification is analyzed to measure the accuracy of the classifiers in categorizing the tax claims into specified classes. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications (Baeza and Ribeiro 1999). The classification accuracy of each of these models is reported and their performance is compared in classifying new instances of records. A separate test dataset is used for testing the performance of the classification models.

5.3.1 J48 Decision Tree Model Building

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

tax claims and to come to an understanding of the most important factors (variables) affecting the tax claims to be fraudulent.

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

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

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

Table 5.12 *Some of the J48 algorithm parameters and their default values*

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

5.3.1.1 Experiment I

The first experimentation is performed with the default parameters. The default 10-fold cross validation test option is employed for training the classification model. Using these default parameters the classification model is developed with a J48 decision tree having 17 numbers of leaves and 25 tree size. The decision tree used five of the total fourteen variables for generating the tree, due to their influential level. These are TOTAL GROSS INCOME, NET WORTH, INTEREST EXPENSES, LIQUID CASH, and DEPRECIATION EXPENSES. The decision tree has also shown that the TOTAL GROSS INCOME variable is the most determining one. Table 5.13 depicts the resulting confusion matrix of this model. The snapshot confusion matrix taken from the tool is depicted in Appendix 3-A.

Actual	**Predicted**		**Total**	Correctly Classified
	Cluster 1	Cluster 2		
Cluster 1	6288	3	6291	99.97%
Cluster 2	0	4789	4789	100%
	6288	4792	11080	99.98%

Table 5.13 *Confusion matrix output of the J48 algorithm with default values*

As we can see from the resulting confusion matrix, the J48 learning algorithm scored an accuracy of 99.98%. This result shows that out of the total training datasets 11080 (99.98%) records are correctly classified, while only 3 (0.027%) of the records are incorrectly classified. The accuracy of the model shows us that the classification is good.

Furthermore, the resulting confusion matrix of this experiment has shown that 100% of the records are correctly classified in the second cluster (cluster 2). This shows that the algorithm classified all of the fraud suspicious tax claims in their respective class. In addition to this out of the 6291 fraud suspicious tax claims, who are described in cluster 1 of Table 5.13, 6288 (99.97%) of them are classified correctly in their designated cluster, i.e. cluster 1, while only 3(0.027%) of them are misclassified in cluster 2. Compared with the records that are classified in cluster 1, those records are classified under cluster 2 are fully correctly classified.

As described before, the size of the tree and the number of leaves produced from this training was 122 and 84 respectively. This seems that it is difficult to traverse through all the nodes of the tree in order to come out with valid rule sets. Therefore, to ease the process of generating rule sets or to make it more understandable, the researcher attempted to modify the default values of the parameters so as to minimize the size of the tree and number of leaves. With this objective, the minNumObj (minimum number of instances in a leaf) parameter was tried with 25, 20, 15, 10 and 5. But the minNumObj set to 20 gives a better tree size and accuracy compared with the other trials. With this value of the minNumObj the process of classifying records proceeds until the number of records at each leaf reached 20. Table 5.14 depicts the result of this experiment. The snapshot confusion matrix taken from the tool is depicted in Appendix 3-B.

Actual	Predicted		Total	Correctly Classified
	Cluster 1	Cluster 2		
Cluster 1	6285	6	6291	99.90%
Cluster 2	2	4787	4789	99.95%
	6287	4793	11080	99.93%

Table 5.14 *Confusion matrix output of the J48 algorithm with changed minNumObj parameter set to 20*

This experiment has shown an improvement in the number of leaves and tree size. The size of the tree is lowered from 25 to 22 and the number of the leaves decreased to 15 from 17. As we can see from Table 5.14, the resulting confusion matrix shows that the J48 decision tree algorithm scored 99.93% accuracy. This result shows that out of the total 11080 records 11072 (99.93%) are correctly classified. Only 8 (0.07%) records of the total dataset are misclassified.

Furthermore, the confusion matrix of this experiment has shown that 6285 (99.90%) of the 6291 total non-fraudulent tax claims, are correctly classified in cluster 1, while 6 (0.10%) of them are misclassified in cluster 2 of the fraudulent tax claims class. In addition to this the confusion matrix also shows that out of 4789 total fraudulent tax claims records of cluster two, 4787 (99.95%) of the records are correctly classified while 2 (0.042%) of the records are misclassified in the non-fraudulent tax claims of cluster one.

The resulting confusion matrix of this experiment has also shown that the second cluster of the fraudulent tax claims class correctly classified the records than the first non- fraud suspicious tax claims cluster. In general, though the tree size and the number of the leaves decreased from 25 and 17 to 22 and 15 respectively, the accuracy of the J48 decision tree algorithm in this experiment is poorer than the first experiment with the default parameter value of the minNumObj.

Although the size of the tree and the number of leaves are lowered in this experiment, their value has no that much difference compared with the first one. That means the complexity of the decision tree to generate rules is the same in both the experiment. So, since there is no a tangible difference in the tree size and number of leaves in the two experiments and the accuracy of the model is decreased from 99.98% to 99.93% in this experiment, the first experiment with the default minNumObj parameter value is taken as the J48 decision tree model.

5.3.1.2 Experiment II

The experiment is performed by changing the default testing option (the 10-fold cross validation). In this learning scheme a percentage split is used to partition the dataset into training and testing data. The purpose of using this parameter was to assess the performance of the learning scheme by increasing the proportion of testing dataset if it could achieve better classification accuracy than the first experimentation. First this experiment has run with the default value of the percentage split (66%). But the one with the better classification accuracy is presented here. So the percentage split parameter set to 70, which is to mean 70% for training and 30% for testing, resulted with a better accuracy. The result of this learning scheme is summarized and presented in Table 5.15. The snapshot confusion matrix taken from the tool is depicted in Appendix 3-C.

Actual	**Predicted**		**Total**	**Correctly Classified**
	Cluster 1	Cluster 2		
Cluster 1	2258	31	2289	98.65%
Cluster 2	3	1031	1034	99.70%
	2261	1062	3323	99.18%

Table 5.15 *Confusion matrix output of the J48 algorithm with the percentage-split set to 30%*

In this experiment, out of the 11080 total records 7756 (70%) of the records are used for training purpose while 3323 (30%) of the records are used for testing purpose. As we can see from the confusion matrix of the model developed with this proportion, out of the 3323 testing records 99.18% of them are correctly classified. Only 34 (1.02%) records are incorrectly classified.

In addition to this the resulting confusion matrix has shown that out of 2289 non-fraud suspicious records 2258 (98.65%) of them are correctly classified while only 31 (1.4%) of the records are misclassified in cluster 2 as a fraud suspicious instances. Furthermore, the confusion matrix of this experiment shown, that 99.70% of the records are correctly classified in the second cluster (fraud suspicious class). This shows that the model correctly classified those fraud suspicious tax claims in their respective class.

By and large, in this experiment when the testing data is increased the performance of the algorithm for predicting the newly coming instances is also diminished as well. Though this experiment is conducted by varying the value of the training and the testing datasets, the accuracy of the algorithm for predicting

new instances in their respective class couldn't be improved. This shows that the previous experiment conducted with the default 10-fold cross validation, is better than this experiment.

Generally, from the three experiments conducted, the model developed with the default parameter values of the J48 decision tree algorithm and the default 10-fold cross validation test option gives a better classification accuracy of predicting newly arriving tax claims in their respective class category. Therefore, among the different decision tree models built in the foregoing experimentations, the first model with the default parameters' values and 10-fold cross validation has been chosen due to its better overall and individual cluster classification accuracy. A tree generated from this model is depicted in Appendix 3-A.

5.3.2 Naïve Bayes Model Building

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

Naïve Bayes makes predictions using Bayes' Theorem which derives the probability of a prediction from the underlying evidence. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes (Han and Kamber, 2006). The first experiment of the Naïve Bayes model building is performed using the Naïve Bayes Simple algorithm with the default 10-fold cross validation test option. Table5. 16 show the resulting confusion matrix of the model developed using the Naïve Bayes Simple algorithm with the default 10-fold cross validation test option. The snapshot confusion matrix taken from the tool is depicted in Appendix 3-D.

Actual	**Predicted**		**Total**	**Correctly Classified**
	Cluster 1	Cluster 2		
Cluster 1	5809	39	5848	99.33%
Cluster 2	482	4750	5232	90.79%
	6291	4789	11080	95.10%
10-fold Cross Validation				

Table5.16 *Confusion matrix output of the Naïve Bayes Simple algorithm*

As can be seen from the resulting confusion matrix of this experiment, the Naïve Bayes Simple algorithm scored an accuracy of 95.10%. This shows that out of the total 11080 records 10537 (95.10%) of the records are correctly classified, while 521 (4.7%) of the records are misclassified.

Furthermore, the resulting confusion matrix also shows that out of the total 5848 non-fraud suspicious tax claim records 5809 (99.33%) of them are correctly classified in their respective class, while 39 (0.67%) of the records are incorrectly classified in the fraud suspicious claims cluster segment (cluster 2). In addition to this out of the total 5232 fraud suspicious tax claim records 4750 (90.79%) of them are correctly classified, while 482 (9.21%) of the records are misclassified in the non-fraud suspicious cluster segment (cluster 1). Generally from the two clusters, cluster 1 (non-fraud suspicious tax claims) correctly classified the records than the second cluster (fraud suspicious tax claims).

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

The second experiment of the Naïve Bayes model building is performed using the Naïve Bayes Simple algorithm with the 75-25 training and testing percentage split test option. Though different experiments are conducted by changing the size of the training and testing datasets, the one with 75-25 training and testing dataset scored better classification accuracy and it is presented here. The result of this experiment is shown in Table5.17.

Actual	**Predicted**		**Total**	**Correctly Classified**
	Cluster 1	Cluster 2		
Cluster 1	1484	14	1498	99.10%
Cluster 2	139	1133	1272	89.10%
	1623	1147	2770	94.10%
Percentage Split (75-25 Training and Testing) Test Option				

Table5.17 *Confusion matrix output of the Naïve Bayes Simple algorithm*

As can be seen from the confusion matrix that resulted from the model developed by the Naïve Bayes Simple Algorithm with the 75-25 percentage split, the model scored an accuracy of 94.10%. This shows that from the total 2770 test data, 2607 (94.10%) of the records are correctly classified, while 153 (5.5%) of them are misclassified. In addition to this the confusion matrix also shows that from the total

1498 non-fraudulent tax claim records 1484 (99.10%) of them are correctly classified, while 14 (0.9%) of the records are misclassified in the fraud suspicious tax claims cluster segment (cluster 2). Furthermore, the confusion matrix also shows that out of the total 1272 fraud suspicious tax claims 1133 (89.10%) of the records are correctly classified in the fraud suspicious claims cluster segment, while 139 (10.9%) of them are incorrectly classified in the non-fraud suspicious claims cluster segment (cluster 1). Compared with the second cluster the first cluster is better in classifying tax claims correctly.

Generally, the first experiment that is conducted using the Naïve Bayes Simple algorithm with the default 10-fold cross validation test option generates a better classification model with a better classification accuracy than the second one conducted with75-25 training and testing percentage split test option.

5.3.3 Comparison of J48 Decision Tree and Naïve Bayes Models

Selecting a better classification technique for building a model, which performs best in handling the prediction and detection of fraudulent tax claims, is one of the aims of this study. For that reason, the decision tree (particularly the J48 algorithm) and the Bayes (the Naïve Bayes Simple algorithm in particular) classification methods were applied for conducting experiments to build the best model. Summary of experimental result for the two classification algorithms is presented in table 5.18.

Classification Model	**Overall Accuracy (11080 records)**	
	Correctly Classified	misclassified
Decision Tree	11072 (99.93%)	8(0.07%)
Naïve Bayes	10537(95.10%)	543(4.9%)
10-fold Cross Validation		

Table5.18 *Accuracy of the J48 decision tree and Naïve Bayes models*

The result showed that J48 decision tree outperforms Naïve Bayes by 4.9% in identifying fraud suspicious tax claims. The reason for the J48 decision tree to perform better than Naïve Bayes is because of the linearity nature of the dataset. That means there is a clear demarcation point that can be defined by the algorithm to predict the class for a particular tax claim. Regarding the Naïve Bayes, scoring a lower accuracy than the J48 decision tree is due to its assumption that each attribute is independent of other attributes, which is not true in reality especially in the tax dataset. Moreover, in terms of ease and

simplicity to the user the J48 decision tree is more self-explanatory. It generates rules that can be presented in simple human language.

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

5.4 Evaluation of the Discovered Knowledge

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

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

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

Rule1. If Total gross income = low and Gross profit = low and Liquid cash = low and Net worth = medium and Net book value = medium THEN cluster 1 (Non-fraud suspicious).

Rule2. If Total gross income = high and Gross profit = low and Liquid cash = low and Net worth = low and Net book value = low THEN cluster 2 (Fraud suspicious).

Rule3. If Total gross income = low and Gross profit = high and Liquid cash = high and Net worth = medium and Net book value = high THEN cluster 2 (Fraud suspicious).

Rule4. If Total gross income = medium and Gross profit = medium and Liquid cash = medium and Net worth = medium and Net book value = medium THEN cluster 1 (Non-fraud suspicious).

Rule5. If Total gross income = low and Gross profit = low and Liquid cash = high and Net worth = high and Net book value = high THEN cluster 2 (Fraud suspicious).

Rule6. If Total gross income = low and Gross profit = low and Liquid cash = low and Net worth = low and Net book value = low THEN cluster 1 (Non-fraud suspicious).

Rule7. If Total gross income = high and Gross profit = high and Liquid cash = high and Net worth = high and Net book value = high THEN cluster 1 (Non-fraud suspicious).

Rule8. If Total gross income = high and Gross profit = high and Liquid cash = low and Net worth = low and Net book value = low THEN cluster 2 (Fraud suspicious).

The rules that are presented above indicate the possible conditions in which a tax claim record could be classified in each of the fraud and non-fraud suspicious classes. Five of the total fourteen variables are used for constructing the decision tree model. These attributes are claim total gross income, gross profit, liquid cash, net worth, and net book value which are basis for building the decision tree. From these, the generated decision tree has shown that the total gross income is the most determinant variable, which is the top splitting variable of the model. In addition to this the model has also shown that net worth is another determining variable for making decisions. A claim is more likely to be fraudulent if there is high total gross income in contacting the tax payer. Rule 2 shows this fact that total gross income can cause the claim to be fraudulent. This finding is consistent with the general idea that amount of money is required to building up tax with fraud. As discussed in the business understanding section of Chapter Four, the rate of the total gross income can be an important factor suspicious fraudulent tax claims. The rule generated at number 4 also showed that if the total gross income, the gross profit is medium, the liquid cash is medium and net book value is medium this claim non-fraud suspicious.

Moreover, this decision tree has shown that the net worth variable can also be used in the process of decision- making. The result of the study has shown that most of renewed total gross incomes are exposed to be fraudulent. The above rule (Rule 6) has also shown that the total gross income and others are low, this indicated that involved in honest claims. But this rule is contradicted with the auditors' idea that the incomes rate didn't be similar. The rule generated at number 8 shows that if the total gross income and gross profit are high and others are low then the claim is highly suspicious to be fraud. Generally, the generated rule has shown that liquid cash, net worth and net book value variables are more likely to be involved in claim activities.

However, enterprise start, number of employees, business activity, sub city, and year variables are not used in developing the decision tree.

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

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

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

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

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

The initial data collected from ERCA didn't incorporate the target class for this study. The clustering sub phase has then been conducted using the K-Means clustering algorithm for segmenting the data into the target classes of FRAUD suspicious and NON-FRAUD suspicious. By changing the default parameters of the algorithm three different clustering experiments have been conducted for generating a plausible model that can create dissimilar clusters of tax claims. The models from these three experimentations are interpreted and evaluated. Among the three models, the one with K = 2, Seed value = 100 and Euclidean distance function has shown better segmentation of the tax claims. This model created dissimilar clusters of FRAUD suspicious and NON-FRAUD suspicious tax claims. The model segmented those different rated claims as fraud suspicious claims. This result of the model complies with the authority's assumption that different rated claims are more of fraud suspicious claims.

Cluster records are then submitted for the classification module for model building using the J48 decision tree algorithm. By changing the training test options and the default parameter values of the algorithm, different decision tree models have been created. The model developed with the 10-fold cross validation with the default parameter values has shown a better classification accuracy of 99.98% on the training dataset, with the Total gross income as a splitting variable. This model is then valuated with a separate test dataset and scored an accuracy of 98.65% of classifying new tax datasets as fraud and non-fraud suspicious claims.

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

6.2 Recommendations

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

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

- The predictive model, which is developed in this research, generated various patterns and rules. For the authority to use it effectively there is a need to design a knowledge base system, which can provide advice for the domain experts.
- The model building process in this investigation was carried out in two sub phases. For clustering the researcher uses the simple K-Means algorithm whereas for classification J48 decision tree algorithm. Though, the results were encouraging, accuracy decreases on sample test datasets. So, further investigation needs to be done using other classification techniques such as Neural Networks and Support Vector Machine.
- For this work only a limited number of all possible attributes are available with their values in the database of the authority. There are inconsistency and missing values in the database. There is no data related to number of employees in the firms, the total material and cash assets. Since data is the most important component in DM research, the authority has to design data warehouse where operational and non-operational data can be kept.
- In this research we didn't consider records related to enterprise starting date, number of employees, business activity, sub city, and year. Future research can be conducted including these attributes.
- Fraud does not only occur in tax collection, it can also occur within the authority by experts, auditors and other staffs. These can also be taken as another area for further research.

REFERENCES

ANAO audit report No. 30 2007-08. *The Australian taxation office's use of data matching and analytics in tax administration*, Canberra, 2008.

Anbarasi, M Anupriya, E. and Iyengar, N. 2010. *Enhance prediction of heart disease with feature subset selection using genetic algorithm*, International Journal of Engineering Science and Technology, Vol. 2(10), pp. 5370-5376.

Apte, C. and Weiss, M. 1997. *Data mining with decision trees and decision rules, future Generation computer systems*, New York. www.research.ibm.com/dar/papers/pdf/fgcsapteweiss_with_cover.pdf access date: January 18, 2012.

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

Berry, M. and Linoff, G. 1997. *Data mining techniques for marketing, sales and customer relationship management, 2nd*., Wiley Publishing, Inc. Indianapolis, Indiana.

Bharti, K Jain, S. and Shukla, S. 2010. *Fuzzy K-means clustering via J48 for intrusion detection system*, International Journal of Computer Science and Information Technologies, Vol. 1(4), pp. 315-318.

Bologna and Lindquist , 1995. *Fraud auditing and forensic accounting*. John Wiley and Sons, New Jersey.

Cahill, M. Chen, F. Lambert, D. Pinheiro, J. and Sun, D. 2002. *Detecting fraud in the real world*. Handbook of massive datasets, San Francisco: Jossey-Bass.

Cao. L, 2007. *Fraud detection using Data mining*. Wiley IEEE Press, London.

Chapman, P Clinton, J Kerber, R Khabaza, T Reinartz, T Shearer, C and Wirth, R. 2000. *CRISP-DM 1.0: Step-by-Step data mining guide*, SPSS Inc., USA.

Cho, S. 2002. *Incorporating soft computing techniques into a probabilistic intrusion detection system. IEEE transactions on systems*, Man and Cybernetics 32 (2): pp.154-160.

Cios. 2000. *Data Mining: A knowledge Discovery Approach*. Springer, New York: USA.

Cios. 2007. *Data Mining: A knowledge Discovery Approach*. Springer, New York: USA.

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

Deshpande, S. P. and Thakare, V.M., 2010. *Data mining system and applications: A review.* International Journal of distributed and parallel systems (IJDPS), V.I, pp. 32-44.

Dunham, M.H. and Sridhar,S. 2006. *Data mining: Introductory and advanced topics*. New Delhi, Pearson Education, Inc.

Ethiopian Revenue and Custom Authority (ERCA), 2008. *BPR document*. www.erca.gov.et , Addis Ababa, Ethiopia. access date: March 20, 2012

Ethiopian Revenue and Custom Authority (ERCA), 2005. *BPR document*, Addis Ababa, Ethiopia.

Famili, A and Turney, P., 1997. *Data preprocessing and Intelligent Data analysis.* Institute of Information Technology, National research council, Canada

Fawcett, T., 1997. *Data Mining and Knowledge Discovery*. Science and Technology, New York, USA.

Fayyad, U Piatetsky-Shapiro, G Smyth,P Uthurusamy, R. ed. 1996. *Advances in Knowledge Discovery and Data mining*. AAAI/MIT Press.

Guo,L., 2003. *Applying Data Mining Techniques in property and casualty Insurance Science*. and Technology, New York, USA.

http:// www.casact.org/pubs/forum/03wforum/03wf001.pdf access date: March 4, 2012.

Hajizadeh, E. Ardakani, D. and Shahrabi, J. 2010. *Application of data mining techniques in stock markets*: A survey, Journal of Economics and International Finance, V. 2(7), pp. 109-118.

Halkidi, M and Vazirgiannis, M, 2001. *Evaluating the validity of clustering results based on density criteria and mulit-representatives*, Greece.

Han, J. and Kamber, M. 2006. *Data Mining: Concepts and Techniques, 2nd ed.* Morgan Kaufman publishers, San Francisco.

Hand, D. 2006. *Principles of data mining*. Prentice-Hall of India, New Delhi, India.

Hegland, M. 2003. *Data mining- challenges, models, methods, and algorithm*. Springer-Verlag,

Helen, T. 2003. *Application of Data Mining Technology to Identify Significant Patterns in Census or Survey Data: the case of 2001 Child Labor Survey in Ethiopia*, Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

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

Ibrahim, S. 1999. *Data mining of machine learning performance data*, Master of applied science (Information Technology) Thesis, RMIT University: Melbourne, Australia.

Jeffrey, A., 2004. *Computer Security*. 19th IEEE Computer Security Foundations Workshop. IEEE Press.

Kantardzic, M., 2002. *Data mining: Concepts, models, methods, and algorithm*. Wiley IEEE Press.

Kirkos and Manolopoulos, 2007. *Applying Data Mining Methodologies for Auditor Selection.* Expert Systems with Applications 32 (2007).pp. 995–1003, Greece

Koh, C and Gervais, G. 2010. *Fraud detection using data mining techniques: Applications in the motor insurance industry's*. Singapore.

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

Leul, W. 2003. *The Application of Data Mining in Crime Prevention: the case of Oromia Police ComITMDsion*, Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Liu,B and Tuzhilin, A. 2008. *Communications of the ACM*, February/2008. Journal of Mechanical Engineering Science. Vol. 51, No. 2. London.

Luciano, A. 2009. *Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection System.* Institute of Computing. Brazil

Marcer, L. (1990). *Fraud detection via regression analysis*. Computers and Security 9: pp.331-338.

Meera, G. and Srivatsa, SK. 2010. *Adaptive machine learning algorithm (AMLA) using J48 classifier for an NIDS environment.* Advanced in Computational Sciences and Technology, Research India Publications, V. 3(3) pp. 291-304.

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

Mitchell, M. T., 1997. *Machine Learning*. McGraw-Hill, London.

Nimmagadda, R Kanakamedala, P and Yaramala, B. 2011. *Implementation of clustering through machine learning tool*, International Journal of Computer Science Issues. Vol. 8 (1), pp. 395- 401.

Oluba, M. 2011. *Nigeria's FIRS can continuously triple tax revenue through data mining*. Value Fronteira Press.Abuja. Nigeria.

Organization for Economic Co-operation and Development, 2004. *Compliance Risk Management: Managing and Improving Tax compliance*. OECD Press. New York, USA.

Palshikar, G. 2002. *Data Analysis Techniques for Fraud Detection*,

Pathak, J., Vidyarthi, N. and Summers, S. (2003). *A fuzzy-based algorithm for auditors to detect element of fraud in settled insurance claims*, Odette School of Business Administration.

Pham, DT. Dimov, SS. And Nguyen, CD. 2005. *Selection of K in K-means clustering*, Journal of

Mechanical Engineering Science, Vol. 219, pp. 103-119.

Phua, C., Alahakoon, D. and Lee, V. (2005). *Minority report in fraud detection: Classification of skewed data,* SIGKDD Explorations 6(1): pp.50-59.

Piatetsky .1996. *From Data Mining to Knowledge Discovery in Databases.* American Association for Artificial Intelligence, USA

Ponce, J. and Karahoca, A. 2009. *Data mining and Knowledge Discovery in real life applications,* I-Tech education and publishing KG, Vienna, Austria.

Pritscher, L. and Hans, F. 2008. *Data mining and strategic marketing in the airlines industry*. Zurich-Airport, Switzerland.

http://www.luc.ac.be/iteo/article/pritcher/pdf. access date: on March 28, 2011.

Qui, M. Davis, S. and Ikem, F. 2004. *Evaluation of clustering techniques in data mining tools*. Issues in Information Systems, Vol. 5(1), pp. 254-160.

Quinlan, J. R. 1993. *C4.5: Programs for machine learning*. Morgan Kaufman Publisher, USA.

SAS Institute Inc. 2003. *Finding the solution to data mining: Exploring the features and components of enterprise miner*, Release 4.1, SAS Institute Inc.

Shao, H., Zhao, H. and Chang, G. (2002). *Applying data mining to detect fraud behavior in customs declaration.* Proceeding of 1st International Conference on Machine Learning and Cybernetics, pp.1241-1244.

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

Shrikant, J. 2004. *Privacy preserving data mining over vertically partitioned data*, Master Thesis, Purdue University.

Singh, Y. and Chauhan, A. 2009. *Neural networks in data mining*. India: Journal of theoretical and applied Information Technology. pp. 37-42.

Summers, S. and Sweeney, J. (1998). *Fraudulently misstated financial statements and insider trading: An empirical analysis*. The accounting review January: pp.131-146.

Tariku, A. (2011). *Mining insurance data for fraud detection: the case of African Insurance Share Company*, Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

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

Two Crows Corporation, 2005. *.Introduction to data mining and knowledge discovery, 3rd*. Potomac, MD 20854, USA.

Weatherford, M. 2002. *Mining for fraud*. IEEE intelligent systems July/ August: pp. 4-6.

Witten, H. Ian and Frank, E., 2005. Data *mining practical machine learning tools and techniques, 2nd ed.*, Morgan Kaufman, USA.

Wu, X Kumar, V Quinlan, R Ggosh, J Yang, Q Motoda, H McLachlan, JNg, A Liu, B Yu, S Zhou, Z Steinbach, M Hand, J and Steinberg, D. 2007. *Top 10 algorithm in data mining*, Springer-Verlag, London.

APPENDICES

Appendix 1: Initial list of original attributes with their description

1	Gross profit/loss	Number	It is a company's residual profit or loss after selling a product or service and deducting the cost associated with its production and sale.
2	Net Worth	Number	Net resource of the company. (Derived)
3	Interest expenses	Number	It is calculated as a percentage of the amount of debt for each period of time. Points paid for a mortgage are a form of prepaid interest.
4	Net income/loss	Number	Net income is calculated by starting with a company's total revenue. From this, the cost of sales, along with any other expenses that the company incurred during the period, is removed to reach earnings before tax.
5	Net tax due/Refundable amount	Number	A tax credit is not limited by the amount of an individual's tax liability. Typically a tax credit only reduces an individual's tax liability to zero.
6	Liquid Cash	Number	Currently available cash (Derived).
7	Non operating income	Number	Income received by a business that is not derived from operations, such as manufacturing. Non-operating income usually does not occur on an ongoing basis, and is examined separately from operating income.
8	Profit income tax	Number	It is a tax levied on the income of individuals or businesses (corporations or other legal entities).
9	Repair and Maintenance expenses	Number	The costs incurred to bring an asset back to an earlier condition or to keep the asset operating at its present condition (as opposed to improving the asset).

10	Selling and Distribution expenses	Number	Selling and distribution expenses include sales commissions, advertising, promotional materials distributed, rent of the sales showroom, rent of the sales offices, salaries and fringe benefits of sales personnel, utilities and telephone usage in the sales department, etc.
11	Depreciation expenses	Number	An annual deduction that allows taxpayers to recover the cost of property used in a trade or business or held for the production of income
12	Net book value	Number	Shows the current price of the material (Derived).
13	Total expenses	Number	These costs consist primarily of management fees and additional expenses such as trading fees, legal fees, auditor fees and other operational expenses
14	Total gross income	Number	The total gross income is being taxable income from all sources, reduced by any adjustments can be taken. Here are the most common income and adjustment items.

Appendix 2: Confusion matrix results of the classification techniques

A) Confusion matrix result of J48 algorithm with default values

=== Confusion Matrix ===

a	b	<-- classified as
4789	0 |	a = cluster2
3	6288 |	b = cluster1

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

=== Confusion Matrix ===

a	b	<-- classified as
4787	2 |	a = cluster2
6	6285 |	b = cluster1

C) Confusion matrix result of J48 algorithm with 30% percentage split

=== Confusion Matrix ===

a	b	<-- classified as
2258	31 |	a = cluster2
3	1031 |	b = cluster1

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

=== Confusion Matrix ===

a	b	<-- classified as
4750	39 |	a = cluster2
482	5809 |	b = cluster1

Appendix 2-A: A sample dataset of cluster 1 and cluster 2.

Gross Profit/Loss	Interest Expenses	Net income/Loss	Net tax due/Refundable amount	Non operating income	Profit income tax	Repair and maintenance expenses
Medium	Low	Medium	Low	Low	Medium	Low
Medium	Low	Medium	Medium	Low	Low	Low
Medium	Low	Medium	Low	Low	Medium	Low
High	Low	High	Low	Low	Medium	Low
High	High	Medium	Low	Low	Low	Low
High	High	Medium	Low	Low	Low	Low
High	High	Medium	Low	Medium	Low	Low
High	High	Low	Low	Low	Low	Low
High	High	Low	Low	Low	Low	Low
High	High	Medium	Medium	Low	Medium	Low
High	High	Low	Low	Low	Low	Low
Low	High	Low	Low	Low	Medium	Low
Low	Low	Medium	Medium	Low	Medium	Low
Low	Low	Medium	Medium	Low	Medium	Low
Low	Low	Medium	Medium	Low	Medium	Low
High	Low	Low	Medium	Low	Medium	Low
Low	Medium	High	Medium	Low	Low	Low
Low	Low	Medium	Low	Low	Low	Low
Medium	Low	Medium	Low	Low	Low	Low
Medium	Low	High	Low	Low	Low	Low
High	Low	High	Low	Low	Medium	Low
High	Medium	Low	Low	Medium	Medium	Low
Medium	Low	Low	Medium	Low	Low	Low
Medium	Low	Low	Low	Low	Medium	Low
Low	Low	Low	Medium	Medium	Medium	Medium
High	Low	Medium	Medium	Low	Medium	Low
Medium	Low	Low	Medium	Low	Low	Low
Medium	Low	Medium	Low	Low	Low	Low
Medium	Low	Low	Low	Low	Low	Low

Selling and distribution expenses	Depreciation expenses	Total expenses	Total gross income	Liquid Cash	Net worth	Net book value	Cluster
Low	Low	Medium	High	Low	High	High	cluster2
Low	Low	Medium	High	Low	High	High	cluster2
Low	Low	Medium	High	Low	High	High	cluster2
Low	Low	Medium	High	Low	High	High	cluster2
Medium	Medium	High	High	Medium	High	High	cluster2
Medium	Medium	High	High	Medium	High	High	cluster2
Medium	High	High	High	High	Medium	High	cluster1
Medium	Low	High	High	Low	High	High	cluster2
Medium	Low	High	High	Low	High	High	cluster2
Medium	Medium	High	High	Medium	High	High	cluster2
Low	Medium	High	High	Medium	High	High	cluster2
Low	High	High	High	High	High	High	cluster2
Medium	Medium	High	High	Medium	High	High	cluster2
Low	Low	Low	Medium	Low	Medium	Medium	cluster1
Medium	Low	Low	Low	Low	Low	Low	cluster1
Medium	Low	Low	Medium	Low	Medium	Medium	cluster1
Medium	Low	Low	Low	Low	Low	Low	cluster1
Low	Low	Low	Low	Low	Low	Low	cluster1
High	Low	Low	High	Low	High	High	cluster2
Medium	Medium	Low	High	Medium	Medium	High	cluster1
Low	Low	High	Low	Low	Low	Low	cluster1
Low	Low	High	High	Low	High	High	cluster2
Low	Low	Low	Medium	Low	Medium	Medium	cluster1
Low	High	Low	High	High	Medium	High	cluster1
Low	Low	Medium	High	Low	High	High	cluster2
Low	Low	High	Medium	Low	Medium	Medium	cluster1
Medium	High	Medium	Medium	Medium	High	High	cluster1
Low	Low	Medium	Medium	Low	Medium	Medium	cluster1
Low	Low	High	Medium	Low	Medium	Medium	cluster1

Appendix 3-A: A sample decision tree generated from the J48 decision tree learner with the 10-fold cross validation and the default parameters.

J48 pruned tree

Total gross income = High

| Net worth = High: cluster2 (3038.0)

| Net worth = Medium

| | Repair and maintenance expenses = Low: cluster1 (39.0)

| | Repair and maintenance expenses = Medium: cluster2 (43.0)

| | Repair and maintenance expenses = High: cluster1 (2.0)

| Net worth = Low

| | Repair and maintenance expenses = Low: cluster1 (3.0)

| | Repair and maintenance expenses = Medium: cluster2 (3.0)

| | Repair and maintenance expenses = High: cluster2 (0.0)

Total gross income = Medium

| Net worth = High

| | Repair and maintenance expenses = Low: cluster1 (28.0)

| | Repair and maintenance expenses = Medium: cluster2 (34.0)

| | Repair and maintenance expenses = High: cluster2 (0.0)

| Net worth = Medium: cluster1 (4172.0)

| Net worth = Low: cluster1 (388.0)

Total gross income = Low

| Repair and maintenance expenses = Low: cluster1 (1443.0)

| Repair and maintenance expenses = Medium

| | Net worth = High: cluster2 (35.0)

| | Net worth = Medium: cluster1 (154.0)

| | Net worth = Low: cluster2 (1636.0)

| Repair and maintenance expenses = High: cluster1 (62.0)

Appendix 3-B: A sample decision tree generated from the J48 decision tree learner with parameter set to 20.

J48 pruned tree

Total gross income = High

| Net worth = High: cluster2 (927.0)

| Net worth = Medium

| | Repair and maintenance expenses = Low: cluster1 (23.0)

| | Repair and maintenance expenses = Medium: cluster2 (4.0)

| | Repair and maintenance expenses = High: cluster1 (1.0)

| Net worth = Low: cluster1 (2.0)

Total gross income = Medium: cluster1 (1367.0)

Total gross income = Low

| Repair and maintenance expenses = Low: cluster1 (880.0)

| Repair and maintenance expenses = Medium

| | Net worth = High: cluster2 (3.0)

| | Net worth = Medium: cluster1 (7.0)

| | Net worth = Low: cluster2 (100.0)

| Repair and maintenance expenses = High: cluster1 (9.0)

Appendix 3-C: A sample decision tree generated from the J48 decision tree learner with percentage set to 30.

J48 pruned tree

Total gross income = High

| Net worth = High: cluster2 (927.0)

| Net worth = Medium

| | Repair and maintenance expenses = Low: cluster1 (23.0)

| | Repair and maintenance expenses = Medium: cluster2 (4.0)

| | Repair and maintenance expenses = High: cluster1 (1.0)

| Net worth = Low: cluster1 (2.0)

Total gross income = Medium: cluster1 (1367.0)

Total gross income = Low

| Repair and maintenance expenses = Low: cluster1 (880.0)

| Repair and maintenance expenses = Medium

| | Net worth = High: cluster2 (3.0)

| | Net worth = Medium: cluster1 (7.0)

| | Net worth = Low: cluster2 (100.0)

| Repair and maintenance expenses = High: cluster1 (9.0)

Appendix 3-D: A sample confusion matrix output of the Naïve Bayes

	Class	
Attribute	cluster2	cluster1
	(0.38)	(0.62)
Gross Profit/Loss		
Medium	1065.0	1654.0
High	389.0	471.0
Low	1491.0	2691.0
[total]	2945.0	4816.0
Interest Expenses		
Low	2165.0	3499.0
High	239.0	385.0
Medium	541.0	932.0
[total]	2945.0	4816.0
Net income/Loss		
Medium	658.0	1066.0
High	156.0	251.0
Low	2131.0	3499.0
[total]	2945.0	4816.0
Net tax due/Refundable amount		
Low	2318.0	3840.0
Medium	509.0	786.0

High	118.0	190.0
[total]	2945.0	4816.0

Non operating income

Low	2308.0	3764.0
Medium	538.0	852.0
High	99.0	200.0
[total]	2945.0	4816.0

Profit income tax

Medium	487.0	803.0
Low	2387.0	3888.0
High	71.0	125.0
[total]	2945.0	4816.0

Repair and maintenance expenses

Low	1359.0	3478.0
Medium	1563.0	1253.0
High	23.0	85.0
[total]	2945.0	4816.0

Selling and distribution expenses

Low	2221.0	3722.0
Medium	597.0	915.0
High	127.0	179.0
[total]	2945.0	4816.0

Depreciation expenses

Low	2614.0	4085.0
Medium	287.0	653.0
High	44.0	78.0
[total]	2945.0	4816.0

Total expenses

Medium	1098.0	1702.0
High	214.0	352.0
Low	1633.0	2762.0
[total]	2945.0	4816.0

Total gross income

High	2135.0	45.0
Medium	18.0	3220.0
Low	792.0	1551.0
[total]	2945.0	4816.0

Liquid Cash

Low	2606.0	4093.0
Medium	294.0	651.0
High	45.0	72.0
[total]	2945.0	4816.0

Net worth

High	2143.0	54.0
Medium	22.0	3147.0
Low	780.0	1615.0
[total]	2945.0	4816.0

Net book value

High	2136.0	49.0
Medium	19.0	3232.0
Low	790.0	1535.0
[total]	2945.0	4816.0

Appendix4. INTERVIEW

1. How often the authority's auditors audit the taxpayers yearly?
2. Which taxpayers get priority from the auditors?
3. How is the process of the auditing task?
4. What is the main tool used by auditors during audit activities?
5. What is the current activity to protect suspicious of fraud?

DECLARATION

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

---------- ----------

Signature Date

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

Dereje Teferi (PhD)