ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Mining e-filing Data for Predicting Fraud: The Case of Ethiopian Revenue and Custom Authority

Beris Geremew

2017

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

.....

Beris Geremew

June 2017

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

---

Dr. Wondwossen Mulugeta

June2017

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Mining e-filing Data for Predicting Fraud: The Case of Ethiopian Revenue and Custom Authority

A Thesis Submitted to the College of Natural Science of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

Beris Geremew

June 2017

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Mining e-filing Data for Predicting Fraud: The Case of Ethiopian Revenue and Custom Authority

By

Beris Geremew

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
Dr.Wondwossen Mulugeta	Advisor	_____	_____
Dr.Dereje Teferi	Examiner	_____	_____
Dr.Million Meshesha	Examiner	_____	_____

DEDICATION

I would like to dedicate this thesis work to my mother Asnaqech Tadesse, my father Geremew Lemma and my brother Tollossa Geremew.

ACKNOWLEDGEMENT

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. Wondwossen Mulugeta. I am really grateful for his constructive comments and critical readings of the study. I am also very thankful to my instructors and all staff members of the School of Information Science for their contribution in one way or another for success of my study.

I would like to thank ERCA, Information Technology Management Directorate members Ato Getenet Abebaw. Thanks also go to my staff in ERCA.

My special thanks also go to Birhan lakew for all rounded support.

LIST OF ACRONYMS

ARFF: Attribute Relation File Format

CRISP-DM: Cross Industry Standard Process for Data Mining

CSV: Comma Separated Values

DM: Data Mining

E-Filing: electronic Filing

E-Payment: Electronic Payment

ERCA: Ethiopian Revenue and Custom Authority

E-Tax: Electronic Tax

FIRS: Federal Inland Revenue Service

ITMD: Information Technology Management Directorate

KDD: Knowledge Discovery in Databases

SEMMA: Sample Modify Model Assess

SIGTAX: Standard Integrated Government Tax Administration System

TIN: Tax Identification Number

WEKA: Waikato Environment for Knowledge Analysis

TABLE CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

Nowadays the technological advancement is at improving stage. These technological advancements have their own side effect (loop hole) on the growing economy and the taxation system of a nation. Fraud is one of the risks in this digital environment of tax. Beside the technological advancement, the controlling and monitoring environment is necessary.

In this study, experiments were conducted by strictly following the six step Cios et al. (2000) process model. It start from business understanding in ERCA taxation system and fraud, specifically on E-filed data set. By taking the data from database of ERCA and understanding of the data with the help of domain expert and literature. In data preprocessing; inconsistencies, missing value, outliers and related issue handled properly. After that, construction of models and analysis of the result done to facilitate decision making in the business risk analysis.

For this study, used a total of 2954 records to training the classifier model. Experiment on deferent classification algorithms including J48, random forest and multilayered perception algorithms were done. We have compared the result of the various models to find the best model using 10-fold cross validation and percentage split (66/34%) evaluation methods.

The study, finds that J48 classification algorithm performs with best accuracy when cross checked with deferent testing mechanisms. J48 recorded an accuracy of 94.72% where 2798 instances are correctly classified out of 2954 test cases. Future research directions are also forwarded to come up with an applicable system in the area of the study.

Key word: Data Mining, e-filing, classification, j48, fraud.

Chapter One
Background

1.1 Introduction

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 [1].

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 [2]. 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 [6]. 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. Many researches shows that these techniques are not applied as often to problems in the developing world.

Taxes are important sources of public revenue. Such public revenue for goods like roads, power, municipal services, and other public infrastructures have favorable results on all 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 [5].

According to the income tax proclamation 979/2016 every beneficiary must pay monthly and yearly taxes according to the law. Tax in Ethiopian revenue and custom authority was based on estimation and manual bases for many years. This situation is changing from time to time [35].

ERCA introduced E-tax and E-filing since 2011 G.C. The purpose of E-filing system is sending commercial transaction of taxpayer to the server of Ethiopian Revenue and Custom Authority (ERCA). E-filing system is for the purpose of increasing compliance of taxpayers without pre audit to send the financial report online. Many researches show that frauds suspect on E-filing is very high[43].

According to new tax administration proclamation 983 article 111 No 2 will be "any taxpayer must start to use the E-filing system and declare taxes by the system otherwise penalized 50,000 birr".

The Institute of Internal Auditors [1]" 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 [1].

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 variety of different forms [4]. Fraudulent claims account for a significant portion of all claims received by auditors, and cost billions of dollars annually [1].

1.1.1 ERCA mission Vision value and Objective/goal

ERCA's vision is being a leading, fair and modern Tax and Customs Administration in Africa by 2025 that will finance Government expenditure through domestic tax revenue collection [35]. ERCA's mission is to contribute to economic development and social welfare by developing a modern Tax and Customs Administration that employs professional and highly skilled staff who promote voluntary compliance amongst individuals and businesses, and take swift action against those who do not comply.

ERCA understands its customers and their needs, treat them with trust and respect and help them meet their obligations. It acts with integrity, transparency, accountability and

professionalism to enforce customs and tax related laws. It works closely with stake holders and ensures the participation of women [35].

The ERCA has the following objectives [35]:

- Establish modern revenue assessment and collection system; and render fair, efficient and quality service;
- Assess, collect and account for all revenues in accordance with tax and customs laws set out in legislation;
- Equitably enforce the tax and customs laws by preventing and controlling contraband as well as tax fraud and evasion;
- Collect timely and effectively all the federal and Addis Ababa tax revenues generated by economy, and
- Provide the necessary support to the regional states with the objective of harmonizing federal and regional tax administration systems.

1.2 Statement of the Problem

Today, African governments and their public sector agencies everywhere have not been spared from the pressure to perform more efficiently and effectively. Previously, the traditional methods for addressing risk have served many authorities well, but there is now a need to use more advanced technologies to combat fraud, error and waste in the Tax Administration sector such as Business Intelligence, Data Warehouse and Data Mining [3].

The internal revenue service (IRS) institution responsible for administering taxes in the United States, has also used data mining techniques for various purposes, among which are measuring the risk of taxpayer compliance, the detection of tax evasion and criminal financial activities electronic fraud detection, detection of housing tax abuse, detection of fraud by taxpayers who receive income from tax credits and money laundering [10] [11].

The Ethiopian Revenue and Custom Authority use E-fining system for controlling commercial transaction of taxpayers. 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 taxpayer's income, the problem will cascade to government annual budget. Annual expenditure of the government depends on its income [1]. The annual reports of the authority and several documents show that ERCA can't effectively collect revenue generated by the economy.

Discovering non-files with potential tax liabilities, Identify potential under-reporting taxpayers, improving compliance of tax dedicators, Identify non-compliance in service sector and implicit linkages for effective investigation is important [9]. So, detecting fraud using normal audit procedures is a difficult task [7]. First, there is a shortage of knowledge concerning the characteristics of fraud. Second, most auditors lack the experience necessary to detect it. Finally, financial managers and accountants are deliberately trying to deceive the auditors [8].

The application of data mining tool on E-filing system data is important because of massive data collected from each taxpayer every day to the authority that could not be easily managed in a traditional manner. ERCA use E-fining system for tax monitor and facilitate commercial transaction of taxpayers. The main problem is to get the exact income report from the taxpayers to the tax collector offices. If a company uses E-filing system to declare whatever they want online including null and credit on monthly and yearly taxes, that information could be used efficiently. The authority aims to implement this E-filing system to all taxpayers and also implementing E-payment system soon. On this situation, ERCA implementing data mining tools and technique needs for resolving the problem.

Research Question

- Which data mining techniques perform well in developing a model that can predict fraud?
- What are the main determinant factors (attributes) of fraud from the E-filing data?

1.3 Objective

1.3.1 General objective

The main aim of this research is exploring the applicability of DM for fraud prediction based on E-filing data.

1.3.2 Specific objective

- Review of data mining literatures to understands fraud problem and analysis of E-filing data.
- Conduct interview with domain experts, taxpayers and analyze relevant documents to get insight of the problem domain.
- Build target dataset used for data mining tools following major data preparation and pre-processing steps.
- Explore patterns that relate the relationships of taxpayer status with other variables.
- Conduct recurring experimentation to evaluate the accuracy of the system.

1.4 Research Methodology

Secondary data is taken from database of E-filing system of ERCA for analysis. Then, based on the information acquired from this system, the fraud prediction of the Ethiopian revenue and custom authority is described. Relevant literatures on data mining tools (techniques), fraud are reviewed.

For the purpose of study, interview of experts and auditors of ERCA and review literatures is done. The data mining tool (technique) beginning from understand of the problem domain up to evaluating models and reporting the result is implemented on this E-filing system data.

Many researchers clarify that a series of steps that comprise the Knowledge Discovery process are to be followed by practitioners when executing a data mining project. In this research, used to implement six step Cios et al. (2000) process model standard data mining methodology that has been selected. Understanding problem domain, data understanding, data preprocessing, and selection of modeling technique, model building and model evaluation was undertaken in this research.

This model was developed, by adopting the CRISP-DM model to the needs of academic research community. The model consists of six step Cios et al. (2000) process model [31] as discussed below.

- **Understanding of the problem domain:** in this step one works closely with domain experts to define the problem and determine the research goal, identify key people and learn about current solutions to the problem.
- **Understanding of the data:** this step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important.
- **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 used as input for data mining tools of step is decided. It may involve sampling of data, data cleaning like checking completeness of data records, removing or correcting for noise, etc. the cleaned data can be, further processed by feature selection and extraction algorithms(to reduce dimensionality), and by derivation of new attributes(say by discretization). The result would be new data records, meeting specific input requirements for the planned to be used data mining tools.
- **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, pre processing techniques, machine learning etc. This step involves the use of several data mining 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 data mining tools; the generated data model is verified by using testing procedures.
- **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 data mining process may be revisited to identify which alternative actions could have been taken to improve the results.

- **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.

Six step Cios et al. (2000) process model is selected because of its recurring and ERCA is service giving organization. E-filing is new for ERCA to try like a project than a research. 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]. The nature of business in ERCA is more sophisticated and based on laws and rules for that matter returning way of study is needed.

1.4.1 Data mining tools selection

The basic aim of using data mining tool is to discover hidden knowledge from a large database. So, for the smooth course of action, selecting appropriate data mining tool is indispensable [31]. As a result, to get useful knowledge, convenient data mining tool was selected. The researcher used an open source data mining tool, WEKA, which is developed by the University of Waikato in New Zealand (Shigeki, 2006). For this research Weka 3.7.13 is selected for rules mining and Microsoft Excel is employed for preprocessing the dataset. WEKA was selected since the researcher is familiar with this tool.

1.5 Scope and Limitation of the Study

The scope of the study is classifying fraud based on the ERCA E-filing data for academic purpose. Basically, by using the knowledge discovered through data mining under six step Cios et al. (2000) process model approach, the authority may try to put in place better fraud controlling and management procedures.

The study forwards recommendations for Ethiopian revenue and custom authority for further implementation and knowledge base development.

The data is collected from ERCA data set it have 2998 records the number of data in this research is limited with it.

The approach follow in this research is rule based data mining for fraud prediction in ERCA E-filing data set.

1.6 Significance of the Research

This study helps the Authority to classify the taxpayers in their risk level. The study will help to save time, cost and resources of the Authority by automating the process and effectively identify risk zones. The purpose of the study is to predict fraud for better implementation of E-filing and E-payment at ERCA.

Beside, the massive data sent to the server also increases which need to be utilized and handled to its useful format. The study can support strategic decision made by the institution.

The finding of this study shall also be used as a point of indication for risk department in ERCA as well as a source of methodological approach for studies dealing on the application of data mining on fraud risk level management and other similar business activity.

1.7 Organization of the Research

This thesis has five chapters. The first chapter is overview of the research including background of the study, problem statement, objective, research methodology, scope and limitation, significance of the study. The second chapter is devoted to literature review of data mining technology, data mining tasks, data mining tools (technique), and application of data mining and local related work. The third include taxation in Ethiopia, e-filing in ERCA and ERCA risk criteria. The fourth chapter includes understanding of the problem domain; understanding the data, data preprocessing, modeling and evaluation. The fourth chapter deals with result and finding more on experimentation. Chapter five is about the conclusion and recommendation of the research.

Chapter Two
Literature Review

2.1 Data Mining Concepts

In this section, try to give the pictures of the research with related work and define and explore different concepts of data mining.

2.1.1 Introduction

Data Mining is a technology that uses various tools (techniques)to discover hidden knowledge from heterogeneous and distributed historical data stored in large databases, warehouses and other massive information repositories so as to find patterns in data that are: valid, novel, useful and understandable for the users[6].

It is clear that more than ever massive amount of data is produced in different fields of study. Studying the relationship, locating specific data group, and retrieving information from this bulk of data is challenging task. As a result of the aforementioned reasons and the wide interest on data mining, a way of grouping similar data into the same cluster and classifying into the same label based on their predetermined class is becoming important more than ever. To address this issue, many techniques have been developed over the years. But, there is long way ahead to get the most out of what clustering and classification can do in data mining [15].

Data mining is an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, data visualization, optimization, information retrieval, high and computing, and others [2]. It is used in many fields of studies such as biomedicine, gene functions, data analysis of DNA arrangement pattern, diagnosis of illnesses, retail data, telecommunication industry, selling, financial sectors analysis and astronomy [2].

Many researcher write about data mining is the study of finding the knowledge that enables us to find the pattern among the massive data collection which will be very useful for making decision about the future by using a computer program.

Data mining is about solving problems by analyzing data already present in databases. What about KDD? See the next subsection.

2.1.2 Data Mining and Knowledge Discovery

To investigating massive data and extracting useful information and knowledge for decision making that the new invention of computerization methods known as Data Mining or Knowledge Discovery in Databases 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 [26]. Information is basically extracted from the massive size of data. Data mining is as a means for detecting fraud, assessing risk, and product retailing [26] Data mining involves the use of data analysis, tools to discover hidden and valid patterns and relationships in huge data sets. At this time data mining is used to identify different kinds of crimes such as illegal money transfer(money laundry), and tax fraud. Using data mining in different organizations like bank, insurance, supermarket, hospitals and research institutions is common.

According to Fayyad et.al [17] the traditional method of turning data into knowledge relies on manual analysis and interpretation. For example, in the financial industry, it is common for specialists to analyze current trends and changes in financial data on a yearly or quarterly basis. The specialists then provide a report detailing the analysis to the organization; the report is then used as the basis for future decision making and planning for financial management. For these (and many other) applications, such manual probing of a dataset is slow, expensive, and highly subjective. When the scale of data manipulation, exploration, and inference grows beyond human capacities, people look to computer technology to automate the bookkeeping[16] . The problem of knowledge extraction from large databases involves many steps, ranging from data manipulation and retrieval to fundamental mathematical and statistical inference, search, and reasoning. Researchers and practitioners interested in these problems have been meeting since the first KDD Workshop in 1989 [17].

According to Jiawei et .al[33] data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and

prediction, and presenting the mining results using visualization tools. Some people don't distinguish data mining from knowledge discovery while others view data mining as an essential step in the process of knowledge discovery. Here is the list of steps involved in the knowledge discovery process: Some author define KDD as the whole process involving: data selection ➔ data pre-processing: cleaning ➔ data transformation ➔data mining ➔ result evaluation ➔ visualization. Data mining, on the other hand, refer to the modeling step using the various techniques to extract useful information/pattern from the data. KDD is the process model to find useful information and patterns in database.DM is the use of algorithms to extract hidden patterns & knowledge in data sets– thus our contention above that KDD is really about knowledge construction rather than discovery.

2.2 Data mining Models

2.2.1 Knowledge Discovery in Databases

According to Fayyad et.al[17] emphasize both the interactive and iterative nature of KDD, that humans make many decisions and that the various steps and methods within them are repeated frequently as knowledge is being refined.

As showed in figure 2.1 the KDD process consists of five steps [17].

- **Data selection** – having two subcomponents: (1) developing an understanding of the application domain and (2) creating a target dataset from the universe of available data;
- **Preprocessing** – including data cleaning (such as dealing with missing data or errors) and deciding on methods for modeling information, accounting for noise, or dealing with change over time;
- **Transformation** – using methods such as dimensionality reduction to reduce data complexity by reducing the effective number of variables under consideration;
- **Data mining** – having three subcomponents: (1) choosing the data mining task (e.g., classification, clustering, summarization), (2) choosing the algorithms to be used in searching for patterns, (3) and the actual search for patterns (applying the algorithms);

- **Interpretation/evaluation** – having two subcomponents: (1) interpretation of mined patterns (potentially leading to a repeat of earlier steps), and (2) consolidating discovered knowledge, which can include summarization and reporting as well as incorporating the knowledge in a performance system.

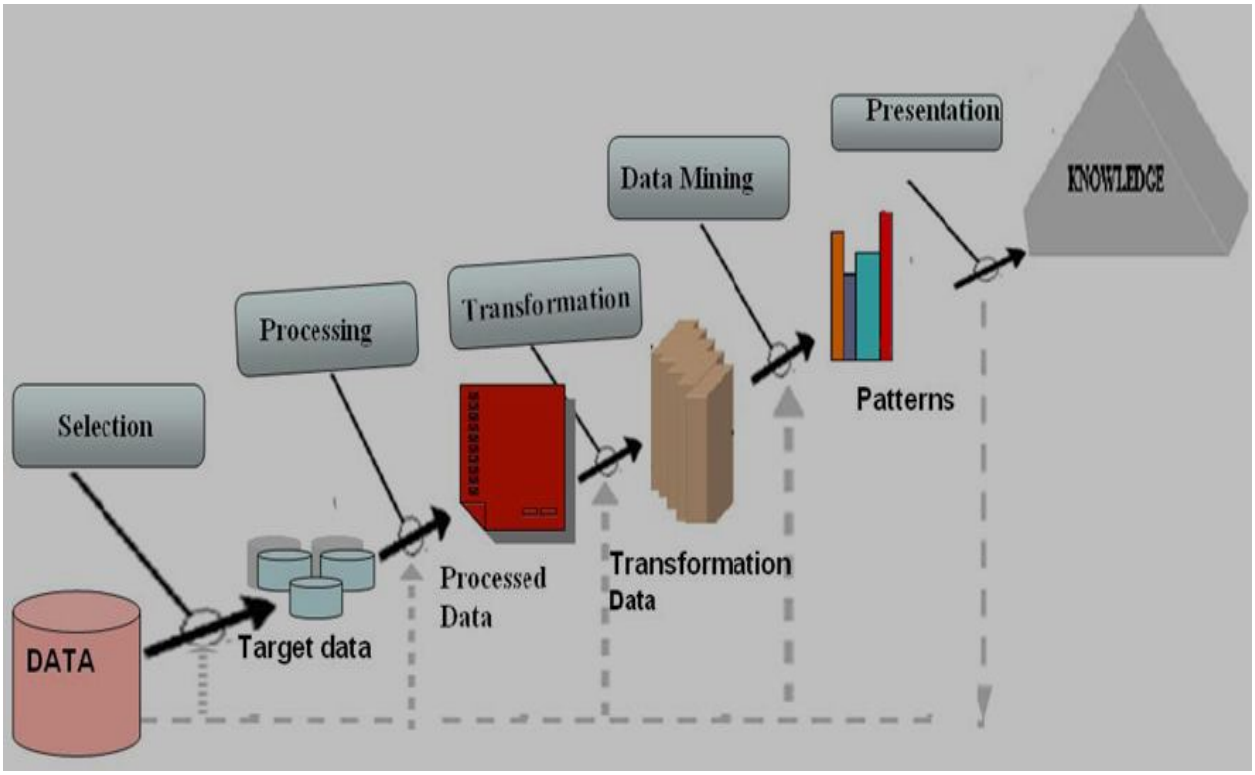

Figure 2.1 KDD processes[17].

2.2.1 The CRISP_DM Process Model

Cross Industry Standard Process for Data Mining (CRISP-DM) is the most used methodology for developing DM projects [18][19]. CRISP-DM is vendor independent so it can be used with any DM tool and it can be applied to solve any DM problem. CRISP-DM defines the phases to be carried out in a DM project. CRISP-DM also defines for each phase the tasks and the deliverables for each task. CRISP-DM is divided into six phases as shown in figure 2.2.

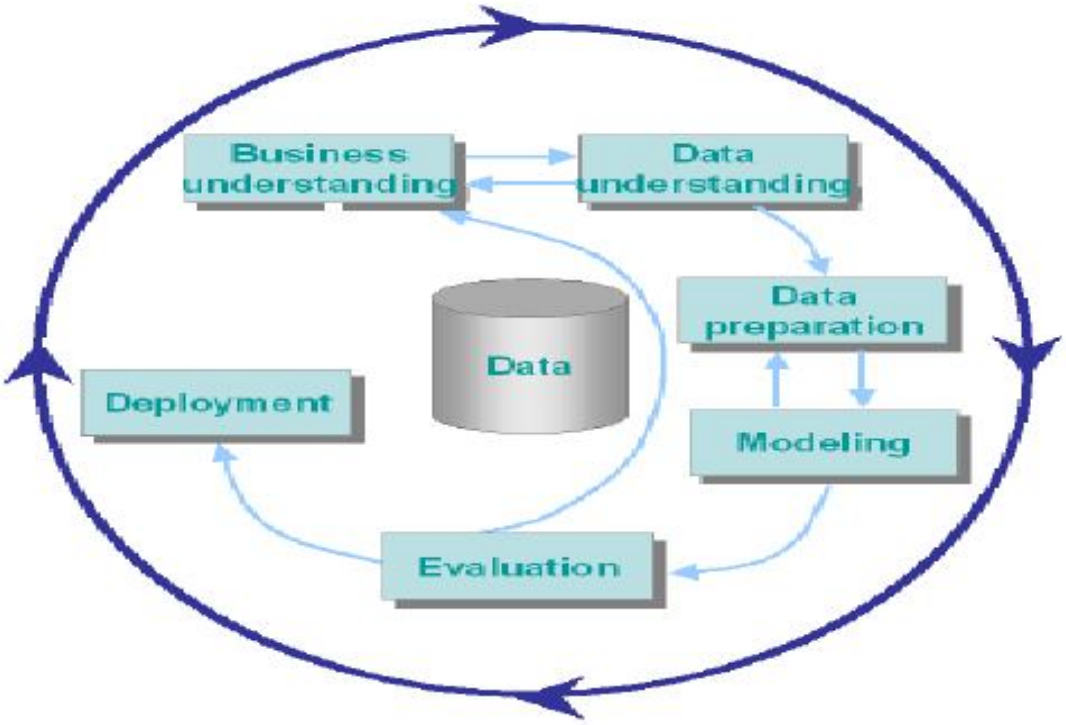

Figure 2.2 Cross Industry Standard Process for Data Mining CRISP-DM [12].

The sequence of the phases is not strict. 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 [19]. The life cycle of a data mining project consists of six phases.

- **Business understanding:** This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
- **Data understanding:** The data understanding 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:** The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.
- **Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, it is often necessary to step back to the data preparation phase.
- **Evaluation:** What are, from a data analysis perspective, seemingly high quality models will have been built by this stage? Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves t he business objectives. At the end of this phase, a decision should be reached on how to use of the DM results.
- **Deployment:** Model construction is generally not the end of the project. Even if 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.

2.2.3 Cios et.al model

This model was developed, by adopting the CRISP-DM model to the needs of academic research community. The model consists of six steps [31] as discussed below.

1. **Understanding of the problem domain:** in this step one works closely with domain experts to define the problem and determine the research goal, identify key people and learn about current solutions to the problem. A description of the problem including its restrictions is done. The research goals then need to be translated into the data mining goals, and include initial selection of data mining tools.

2. **Understanding of the data:** this step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, verification of the usefulness of the

data is needed in respect to the data mining goal. Data needs to be checked for completeness, redundancy, missing value, 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 is decided. It may involve sampling of data, data cleaning like checking completeness of data records, removing or correcting for noise, etc. the cleaned data can be, further processed by feature selection and extraction algorithms(to reduce dimensionality), and by derivation of new attributes(say by discretization). The result would be new data records, meeting specific input requirements for the planned to be used data mining 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, pre processing techniques, machine learning etc. This step involves the use of several data mining 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 data mining 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 data mining 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.

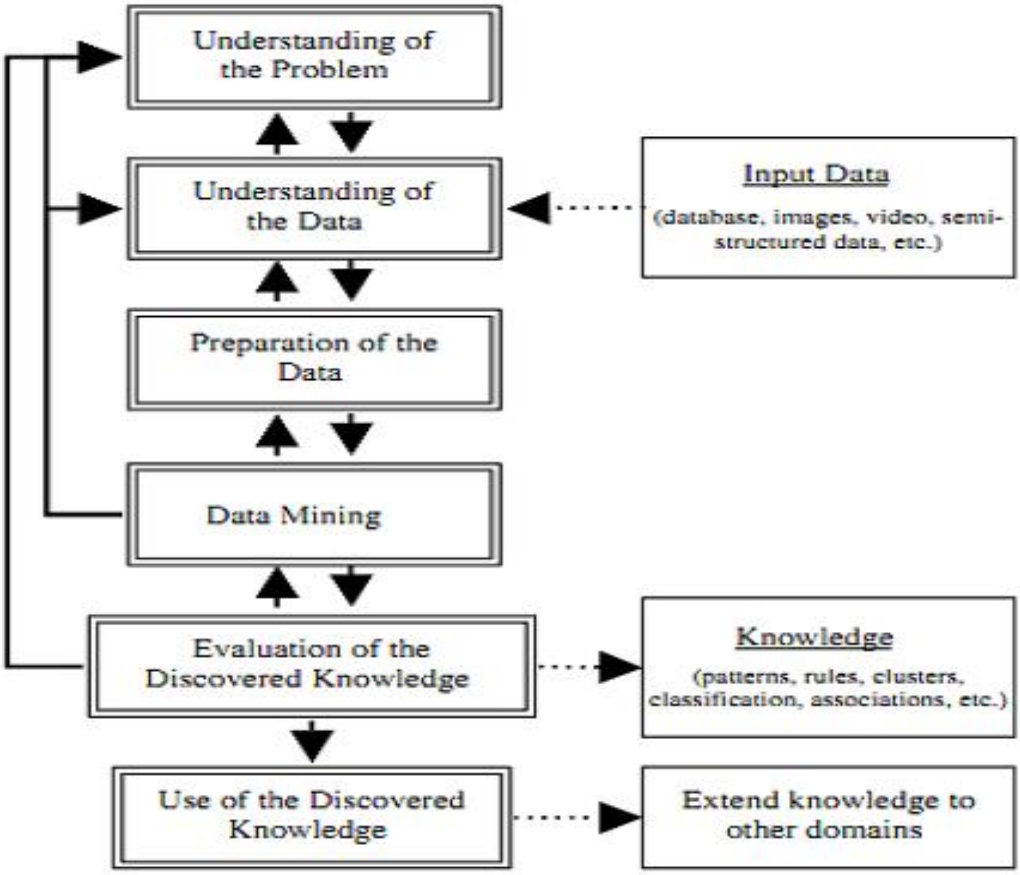

Figure 2.3 Cios et al. model (hybrid way of Knowledge Discovery Process)[31]

2.3 Comparison of Different Data Mining Models

In the late 1980s, when the Knowledge discovery process term was first coined [20], there was a rush to develop data mining algorithms that were capable of solving all problems of searching for knowledge in data. The Knowledge discovery process [21] has a process model component because it establishes all the steps to be taken to develop a data mining project, but it is not a methodology because its definition does not set out how to do each of the proposed tasks. It is also a life cycle. The 5 A's [23] is a process model that proposes the tasks that should be performed to develop a data mining project and was one of CRISP data mining forerunners. Therefore, they share the same philosophy, 5 A's proposes the tasks but does not propose how they should be performed. Its life cycle is similar to the one proposed in CRISP data mining.

A people-focused data mining proposal is presented in [23]: Human Centered Approach to data mining. This proposal describes the processes to be enacted to carry out a data mining project, considering people's involvement in each process and taking into account that the target user is the data engineer. Sample, Explore, Modify, Model, and Assess (SEMMA) (SAS, 2012) is the methodology that SAS proposed for developing data mining products. Although it is a methodology, it is based on the technical part of the project only. Like the above approaches, SEMMA also sets out a waterfall life cycle, as the project is developed right through to the end. The two models by [24, 25] are based on knowledge discovery with few changes and have similar features. Like the knowledge discovery process, [26] is a process model and waterfall life cycle. At no point does it set out how to do the established data mining project development tasks. CRISP data mining states which tasks have to be carried out to successfully complete a data mining project [19]. It is therefore a process model. It is also a waterfall life cycle. CRISP data mining also has a methodological component, as it gives recommendations on how to do some tasks. Even so these recommendations are confined to proposing other tasks and give no guidance about how to do them.

2.4 Data Mining Main Task

The data mining tasks are of different types depending on the use of data mining product [1]. Predictive modeling, descriptive modeling, exploratory data analysis, patterns and rules discovery, and retrieval by content are some of the data mining tasks. Among the major tasks are predictive and descriptive models discussed below in figure 2.4.

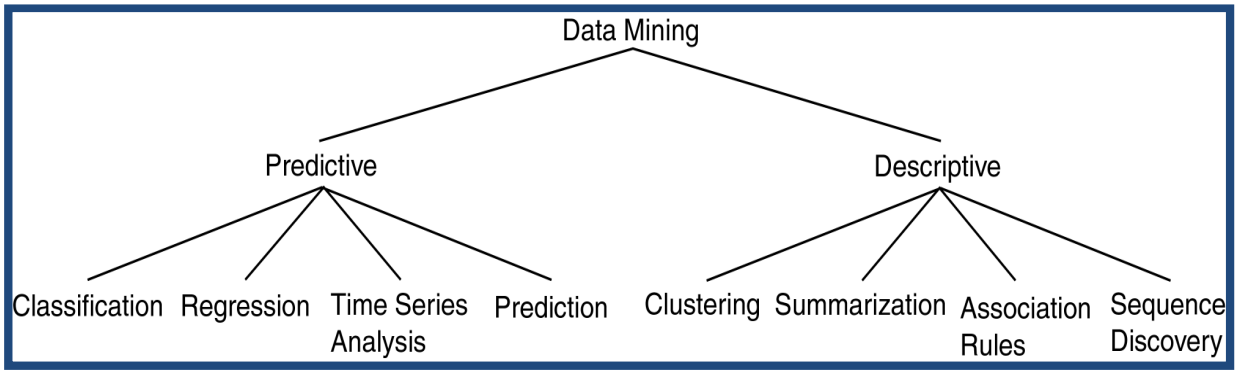

Figure 2.4 Data mining main tasks [33].

2.4.1 Predictive Modeling

A predictive model makes a prediction (forecast) use existing variables to predict unknown or future values of other variables. It's about values of data using known results found from different historical data [29]. Classification is a technique of mapping the target data to the predefine groups or classes. It is a supervised learning because the classes are predefined before the test of the target data. The time series analysis is determine the pattern of Knowledge's over time. It is 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. The regression involves the learning of function that maps data item to real valued prediction variable. It is a data mining function that predicts a number, Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques.

2.4.2 Descriptive Modeling

Descriptive model identifies patterns or association in data, unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties [37]. It finds human-interpretable patterns that describe and find natural groupings of the data. It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables. Examples clustering, association rule discovery, sequence discovery, summarization etc.

Descriptive modeling techniques, such as Summarization, Association Rule, Sequence Analysis and clustering which produces classes (or categories), are not known in advance. Summarization maps data into subsets with associated simple descriptions. Basic statistics such as Mean, Standard Deviation, Variance, Mode and Median can be used as Summarization approach [37]. Association Rules a popular technique for market basket analysis because all possible combinations of potentially interesting product groupings can be explored. The investigation of relationships between items over a period of time is also often referred to as Sequence Analysis. Sequence Analysis is used to determine sequential patterns in data .The patterns in the dataset are based on time sequence of actions, and they

are similar to association data, however the relationship is based on time. In Market Basket analysis, the items are to be purchased at the same time, on the other hand, for Sequence Analysis the items are purchased over time in some order [37].

2.5 Data Mining Techniques

2.5.1 Classification

Classification divides data samples into target classes. In supervised learning, classification refers to the mapping of data items into one of the pre defined classes. In the development of data mining tools that use statistical approaches, one of the critical tasks is to create a classification model, known as a classifier, which will predict the class of some entities or patterns based on the values of the input attributes. Choosing the right classifier is a critical step in the pattern recognition process. A variety of techniques have been used to obtain good classifiers. Some of the more widely used and well known techniques that are used in data mining include decision trees, logistic regression, neural networks, NaïveBayesand nearest neighbor approach[30].

2.5.1.1 Decision Trees

Decision tree performs classification by constructing a tree based on training instances with leaves having class labels. Decision tree models are constructed in a top-down recursive divide-and conquer manner. The most popular algorithm is J48 decision tree. Same of tree algorithmLMT, M5P, Hoeffdinh tree, decisions Stump are some of known algorithm.

Some Property of tree

- Can handle huge datasets
- Can handle mixed predictors—quantitative and qualitative
- Easily ignore redundant variables
- Handle missing data elegantly
- Small trees are easy to interpret

2.5.1.2 Neural Network

It is represented as a layered set of interconnected processors. This processor node has a relationship with the neurons of the brain. Each node has a weighted connection to several

other nodes in adjacent layers. Individual nodes take the input received from connected nodes and use the weights together to compute output values. A network with the input and output layer only is called single-layered neural network. Whereas multilayer neural network is a generalized one with one or more hidden layer.

They require a number of parameters that are typically best determined empirically such as the network topology or "structure." Neural networks have been criticized for their poor interpretability. For example, it is difficult for humans to interpret the symbolic meaning behind the learned weights and of "hidden units" in the network. These features initially made neural networks less desirable for data mining [33].

According to [33] some feature of neural network as follows:

- Their high tolerance of noisy data.
- Their ability to classify patterns on which they have not been trained.
- They can be used when you may have little knowledge of the relationships between attributes and classes.
- They are well suited for continuous-valued inputs and outputs, unlike most decision tree algorithms.
- They have been successful on a wide array of real-world data, including handwritten character recognition, pathology and laboratory medicine, and training a computer to pronounce English text.

2.5.1.3 Random Forest

It resolve over fitting and is used for unbalanced data. Prediction is aggregate majority vote for classification and average for regression. Many research says that the classifier exhibit a substantial performance improvement over single tree such as CATR and C4.5

The specific algorithm process is as follows [22]:

Step 1: Sample "m" (the number of samples chosen) (m< M, where M is the number for the entire sample) samples randomly from all samples with a bootstrap method Liaw and Wiener (2002).

Step 2: Construct a decision tree with the extracted sample, which is no pruning.

Step 3: Repeat steps 1, 2 and build a large number of decision trees and develop decision tree classification sequence $\{h_1(X), h_2(X), \ldots\ldots h_{ntree}(X)\}$.

Step 4: Final classification is determined by each record vote from the results of the decision tree classification.

2.5.2 Clustering

Clustering is an unsupervised learning method that is different from classification. Clustering is unlike to classification since it has no predefined classes. In clustering large database are separated into the form of small different subgroups or clusters. Clustering partitioned the data points based on the similarity measure [28]. Clustering approach is used to identify similarities between data points. Each data points within the same cluster are having greater similarity as compare to the data points belongs to other cluster. Clustering method the datasets having 'n' data points partitioned into 'k' groups or clusters. Each cluster has at least one data point and each data point must belong to only one cluster. In this clustering approach there is a need to define the number of cluster before partitioning the datasets into groups[27].

Based on the clustering technique, clusters can be expressed in different ways:

- Identified clusters may be exclusive; any instance belongs to only one cluster.
- They may be overlapping: an instance may belong to several clusters.
- They may be probabilistic: whereby an instance belongs to each cluster with a certain probability.
- Clusters might have hierarchical structure: having crude division of instances at highest level of hierarchy, which is then refined to sub-clusters at lower levels.

Clustering is a technique useful for exploring data. It is particularly useful where there are many cases and no obvious natural groupings. Here, clustering data mining algorithms can be used to find whatever natural groupings may exist. Clustering analysis identifies clusters embedded in the data. A cluster is a collection of data objects that are similar in some sense to one another. A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high; in other words, members of a cluster are more like each other than they are like members of a different cluster. Clustering

can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build predictive models [39].

Clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The ultimate aim of the clustering is to provide a grouping of similar records. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre-defined classes, whereas in clustering the classes are formed. The term “class” is in fact frequently used as synonym to the term “cluster” [37].

2.6 Application of Data Mining

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 [32].

Currently, organizations are use data mining to manage all phases of the customer life cycle (acquiring new customers, increasing revenue from existing customers, and retaining good customers) [26]. 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 [1].

2.6.1 Revenue and Tax Fraud Detection

Around the world today, [44] tax authorities are experiencing growing pressure to collect extra tax revenues, to discover underreporting taxpayers, and predict the irregular behavior of non-paying taxpayers.

According to [43] the numbers surrounding tax fraud have always been significant. For instance, in the US, tax evasion resulted in revenue loss of over $300 billion during 2010. In the UK; the tax gap amounted to £35 billion in between 2012 and 2013.The EU is estimated to have lost over €193 billion in Value-Added Tax (VAT) revenues due to non-compliance or

non-collection. Such large numbers have traditionally kept Government tax and welfare authorities on their toes in the course of time. Authorities have been engaged in a constant battle with fraudsters.

The advent of digital technologies is helping tax authorities have a slow, but steady, impact on traditional fraud. Data analytics techniques help classify patterns and identify outliers that could indicate fraud (Majesty's Revenue and Customs in the UK).Analytics techniques are used to analyze extensive data such as income, tax paid and asset ownership to spot fraudsters.

Today, African governments and their public sector agencies everywhere have not been spared from the pressure to perform more efficiently and effectively. Previously, the traditional methods for addressing risk have served many authorities well, but there is now a need to use more advanced technologies to combat fraud, error and waste in the Tax Administration sector such as Business Intelligence, Data Warehouse and Data Mining. However, not much research has been done on these technologies in Africa as indicated by the Google, Google scholar and Science Direct Search where every research returns very little about the Data Mining and the applied algorithms in the Tax Administration sector. Botswana, a rapidly developing country with many new organizations establishing presence every year also acknowledges the challenges to analyze data effectively and efficiently in order to gain important strategic and competitive advantage [45].

In Tanzania, Arusha region, [46] a research study indicates and suggests that Data mining is very important for healthcare as it can improve the industry as well as the well-being of the residents.

To arm themselves for this battle, [47] more and more tax authorities have turned to data mining and analytics to improve their business processes, resulting in better compliance.

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[1].

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[1].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 [48]. 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 [49].

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),

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.

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.

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.

Based on [50] 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.

2.7 Local Related Works

Local researches are conducted in ERCA in deferent data set and agenda to assess the application of DM. Belete.B [13] attempted knowledge discovery for effective customer segmentation for ERCA by using customs ASYCUDA database. He used K-means clustering algorithm for clustering and J48 decision tree and multi layer preceptor ANN algorithms for classification. Using J48 decision tree algorithm with default 10-fold cross-validation shows better performance which is99.95% of overall accuracy rate. Danil.M [1] attempted 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 and developed using the J48 decision tree algorithm has showed highest classification accuracy of 99.98%.

Other related application of DM in the different sectors like Airlines, Banking, Insurance, telecommunication, 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. He 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. He 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 (2010) 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.

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 and 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
Taxation in Ethiopia

3.1 Overview

The Ethiopian Revenues and Customs Authority (ERCA) was established by the proclamation No .587/2008 on 14 July 2008, by the merger of the Ministry of Revenue, Ethiopian Customs Authority and the Federal Inland Revenue Authority for the purpose of enhancing the mobilization of government revenues, while providing effective tax and Customs administration and sustainability in revenue collection. The main objective of the establishment of ERCA was to streamline the public revenue generation function by bringing the relevant agencies under the umbrella of the central revenue collector body. This structuring aimed at improving service delivering, facilitating trade, enforcing the tax and customs laws and thereby enhancing mobilization of Government revenue in sustainable manner. Presently, the Authority is exercising the powers and duties that were granted to the Ministry of Revenue, the Federal Inland Revenue Authority and the Customs Authority by existing laws [35].

According to the income tax proclamation 979 ERCA authorized to collect around 17 varieties of taxes. Basically ERCA has divided the taxes as, direct and indirect tax. Direct tax means collected from the gain of income (net income). Indirect tax means collected from sales or production or value add in product not collect from income (net income).

Direct taxes are:-

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

Indirect taxes are: -

- VAT
- Excise
- Turnover…

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

No	**Direct tax**	**Tax description**
1	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.
2	Personal Income tax	A tax is imposed on employee on monthly income or basic salary.
3	Rental Income tax	A tax that is imposed on the income from rental of buildings.
4	Dividend	Tax imposed on shareholders when dividing the profit of a company.
5	Royalty	Tax imposed on gain from innovation, patent, copyright, official document …etc
6	Technical service	Tax imposed on gain from technical services when service provider company is not legally registered in Ethiopia.
7	Game of chance	Every person deriving income from winning at games of chance/for example, lotteries, tom bolas, and other similar activities.
No	**Indirect tax**	**Tax description**
1	Value added tax	Tax on added value only (or compensate the paid amount that will be deducted before the payment).
2	Excise tax	Tax on production cost, consumption and luxury goods.
3	Turnover tax	Tax on total sales the other name of turnover tax is sales tax.

Table 3.1 Tax description

3.2 E-filing in ERCA

According to ERCA mission promoting voluntary compliance for declaring their tax. This gives opportunity for tax payers for self assessment and pay as they want. E-tax and E-filing is one instrument for fulfilling the above mission.

According to the income tax proclamation 979/2016 every beneficiary must pay monthly and yearly taxes according to the law.

Tax in Ethiopian revenue and custom authority based on estimation and manual bases for many years. This situation is changing from time to time. In 2003 G.C ERCA introduces SIGTAS (standard integrated government tax administration system) it is networked system to administer tax collection and centralized tax data administration system by identifying each taxpayers by TIN (tax identification number). ERCA introduce E-tax and E-filing in 2011 G.C. The purpose of E-filing system is sending commercial transaction of taxpayer to the server of Ethiopian Revenue and Custom Authority (ERCA). E-filing system is for the purpose of increasing compliance of taxpayers without pre audit them to send the financial report online. Many researches show that frauds suspect on data of E-filing is very high.

E-Tax is linked to the SIGTAS database, so all saved activities occurring in E-Tax will be registered in SIGTAS. And also E-Tax can bring all Notices and Reminders produced in SIGTAS to the customer message account.

The first step to using E-filing system is registration and agreement with the authority after that get username and password and begin submission of the file. The authority gives two day training for the user of the system.

Tax payers have the following roles after registered on E-tax:-

- To declare all taxes online from anywhere any time without any limitation.
- They can check there account balance(history in SIGTAS database)
- They can request clearance and refund remotely.

Basically E-tax begins in 3 branches of ERCA to mention that Large Taxpayers Branch Office, Addis Ababa Eastern Branch and Addis Ababa Western Branch.

3.3ERCA Risk Selection Criteria

Based on the risk department document more than 29 criteria's are there see the below table.

ERCA Risk Criteria				
No.	**Name**	**Logic**	**Interval**	**Score**
General Criteria				
1	C1 (Date of Commencement of the Business)	Corresponds to the field Registration Date in SIGTAS	Date of Commencement >3year	0
			Date of Commencement 3year	1
			Date of Commencement 2 years	2
			Date of Commencement <=1 year	3
2	C2 (Size of Business)	Corresponds to the field 'Annual Turnover' in SIGTAS. This field is updated automatically in SIGTAS by summing up Total Gross Income or Gross Sales of Schedule B, C-Normal and C-Mining, but can be empty if there is no declaration. Note that if several Tax Accounts exist, and if the Taxpayer has several Tax Accounts and files only 1 of them, the Annual Turnover will be update only by the amount of this Tax Account (hence it might show a much lower figure than what it should be). Update is done annually. If no value is available in SIGTAS, then default the score value to 3.	<35,000,000	0
			>35,000,000< = 70,000,000	1
			>70,000,000 < = 220,000,000	2
			>220,000,000	3
3	C3 (Size of Employment)	Corresponds to the field 'Total No. of Employees' in SIGTAS. This field is updated automatically in SIGTAS from Line 10 of all Schedule A- PAYE tax accounts (if multiple establishments), but can be empty if there is no declaration. Update is done monthly. Value of '0' in case there is no employee (e.g. only an owner in the enterprise being paid through dividends).	Employee = 0	0
			Employee >1 <=25workers	1
			Employee >25 <=265workers	2
			Employees > 265workers	3
4	C4 (Branches/ Subsidiaries)	A branch is defined in SIGTAS as an Establishment. A value of '0' indicates that the main location is the only location. No Additional Establishment. Sometimes the information is not updated in SIGTAS (e.g. Commercial Bank of Ethiopia has 1134 branches but only HQ is registered in SIGTAS).	0 branches	0
			>1 <=2	1
			>2 <=4	2
			>4	3
5	Existence of	Considered as 3rd party criterion, only	No Existence of Branch	0

	International Branches	applicable to LTO. This information is not in SIGTAS. Therefore a mean will be provided in Paradox to capture the value. To facilitate data-entry, 'Existence of Branch' will be the default value so only the balance of taxpayers' data has to be adjusted.	Existence of Branch	3
6	Sister Companies	This information does not exist in SIGTAS. Therefore a mean will be provided in Paradox to capture the value (number of sister companies).	2 branches	0
			>2 <=3	1
			>3 <=4	2
			>4	3
7	C5 (Taxpayer's Compliance)	Check for all penalty types for all Tax Types in the last 2 years from the selected Year. The idea is if the TP fails to maintain his books of accounts, it's also TP compliance and he's more at risk.	0 & 1Penalty	0
			only 2	1
			UP to 3	2
			From 3>	3
8	C6 (Comparison to Date of Previous Audit)	Any Assessment linked to an audit case. Look at the most recent one. If the case is active (not closed), the score is 0. If it's closed, the score will depend on its year of creation. If there is no audit case for this Taxpayer, the maximum score (3) is attributed.	Date of the Previous Audit <=1 Year	0
			Date of the Previous Audit >1 Year <=2 years	1
			Date of the Previous Audit >2 Year <=3 years	2
			Date of the Previous Audit >3 Years	3
9	C7 (Audit Case Assessment's Results)	The most recent 'Closed' case is looked for. A score of 0 implies that: - There are no additional Reassessments in a closed case, or - A reassessment of less than 5% of increase exists, or - No audit case with a closed status exists for this Taxpayer (even if some Reassessments exist in this opened case), or - The result of the Reassessment diminished the Tax Liability in comparison with the original Assessment. The total of the Assessments are compared to the total of the Reassessments. If there are 3 Assessments and only 2 have a Reassessment, only the amounts of the 2 Assessments and Reassessments will be used in the calculation.	No reassement or Additional Assessment up to 5%	0
			Additional Assessment >5% <=10%	1
			Additional Assessment >10% <=25%	2
			Additional Assessment > =25%	3
10	C8 (Claim for Refund)	Existence of Claim for refund under one of Schedules B, C- Normal, C- Mining or VAT, in the last 2 years. The criterion looks at the difference between the amount requested and the amount	No refund or difference between amounts requested and approved is less or equal to 1%	0
			>1% and <=5%	1
			>5% and <=7%	2

		approved.	> 7%	3
11	C9 (Loss Incurring Business)	Existence of loss in either Schedule B, C-Normal or C-Mining in the last 3 completed fiscal years of the taxpayer. Net Income/ Loss is negative on either Schedule C- Normal or C- Mining form (line 110) for the year selected. Taxable Rental Income/ Loss is negative on Schedule B form (line 200) for the year selected. Nothing for VAT. No score of '1' is possible.	No loss	0
			only 1 Loss in the last 3 years	2
			More than 1 losses in the last 3 years	3
12	C10 (Schedule C- Normal Average % of Gross Profit per SIC)	(Sum of Gross Profit (all Taxpayers of the specified SIC)/ Total Gross Income (all Taxpayers of the specified SIC)) – (Gross Profit (1 Taxpayer having the specified SIC)/ Total Gross Income (1 Taxpayer having the specified SIC)) *100, where: Gross Profit = Line 25; Total Gross Income = Line 15. SIC (International Standard Industry Code) represents the business activity as registered in SIGTAS. Note: This ratio doesn't exist for Schedule B since Gross Profit and Total Gross Income are the same. Default value is 3 is the taxpayer has not filed his last fiscal year.	Deviation >10%	0
			Deviation >10% and <=20%	1
			Deviation >20% and <=30%	2
			Deviation >30%	3
13	C15 (Schedule B- Change in Sale/Turnover from previous year)	(|Previous Year Gross Rental Income - Current Year Gross Rental Income |/ Previous Year Gross Rental Income) *100, where Gross Rental Income = Line 10. Note: Annual Turnover equals Gross Rental Income and not the total Annual Turnover of all 3 Tax Types (B, C- Normal and C- Mining) because the ratio is calculated based on the Tax Type. Default value is 3 is the taxpayer has not filed his last fiscal year.	Deviation <=5%	0
			Deviation >5% and <=10%	1
			Deviation >10% and <=20%	2
			Deviation >20%	3
14	C16 (Schedule C- Normal Change in Sale/Turnover from previous year)	(|Previous Year Business Income or Sales_Turnover - Current Year Business Income or Sales_Turnover |/ Previous Year Business Income or Sales_Turnover) *100, where Business Income or Sales_Turnover = Line 5. Note: Annual Turnover equals Business	Deviation <=5%	0
			Deviation >5% and <=10%	1
			Deviation >10% and <=20%	2
			Deviation >20%	3

		Income or Sales_Turnover and not the total Annual Turnover of all 3 Tax Types (B, C-Normal and C- Mining) because the ratio is calculated based on the Tax Type. Default value is 3 is the taxpayer has not filed his last fiscal year.		
15	C25 (C-Schedule C-Normal Average % of each individual expense compared to total sales/turnover per SIC)	(Sum of Group of Expenses (all Taxpayers of the specified SIC)/ Business Income or Sales_Turnover (all Taxpayers of the specified SIC)) – (Group of Expenses (1 Taxpayer having the specified SIC)/ Business Income or Sales_Turnover (1 Taxpayer having the specified SIC)) *100, where: Group of Expenses = Line 45; Business Income or Sales_Turnover = Line 5. Default value is 3 is the taxpayer has not filed his last fiscal year.	Deviation <=5%	0
			Deviation >5% and <=10%	1
			Deviation >10% and <=20%	2
			Deviation >20%	3
16	C36 (Schedule C- Normal Change in Tax payable from previous year)	(|Previous Year Tax to Pay - Current Year Tax to Pay|/ Previous Year Tax to Pay) *100, where Tax to Pay = Line 130 (Profit Income Tax Payable). Default value is 3 is the taxpayer has not filed his last fiscal year.	Deviation >10%	0
			Deviation >10% and <=20%	1
			Deviation >20% and <=30%	2
			Deviation >30%	3
17	C38 (Schedule B Change in Total Assets from previous year)	(|Previous Year Total Assets - Current Year Total Assets|/ Previous Year Total Assets), where Total Assets = Line 230 of the Financial Statement. Default value is 3 is the taxpayer has not filed his last fiscal year.	Deviation <=5%	0
			Deviation >5% and <=10%	1
			Deviation >10% and <=20%	2
			Deviation >20%	3
18	C39 (Schedule C- Normal Change in Total Assets from previous year)	(|Previous Year Total Assets - Current Year Total Assets|/ Previous Year Total Assets), where Total Assets = Line 230 of the Financial Statement. Default value is 3 is the taxpayer has not filed his last fiscal year.	Deviation <=5%	0
			Deviation >5% and <=10%	1
			Deviation >10% and <=20%	2
			Deviation >20%	3
19	C42 (Schedule C- Normal Change in Total Liabilities from previous year)	((Previous Year Total Liabilities - Previous Year Tax Payable) - (Current Year Total Liabilities - Current Year Total Tax Payable)/ (Previous Year Total Liabilities - Previous Year Total Tax Payable), where - Total Liabilities = Line 320 of the Balance Sheet; - Tax Payable= Line 270 of the Balance Sheet. Default value is 3 is the taxpayer has not filed his last fiscal year.	Deviation <=5%	0
			Deviation >5% and <=10%	1
			Deviation >10% and <=20%	2
			Deviation >20%	3

20	C43 (Schedule C- Mining Change in Total Liabilities from previous year)	((Previous Year Total Liabilities - Previous Year Tax Payable) - (Current Year Total Liabilities - Current Year Total Tax Payable)/ (Previous Year Total Liabilities - Previous Year Total Tax Payable), where - Total Liabilities = Line 320 of the Balance Sheet; - Tax Payable= Line 270 of the Balance Sheet. Default value is 3 is the taxpayer has not filed his last fiscal year.	Deviation <=5%	0
			Deviation >5% and <=10%	1
			Deviation >10% and <=20%	2
			Deviation >20%	3
21	Taxpayer Financial Report / Financial Record	Considered as 3rd party criterion, only applicable to LTO. This information is not in SIGTAS. Therefore a mean will be provided in Paradox to capture the value. To facilitate data-entry, 'Acceptable' will be the default value so only the balance of taxpayers' data has to be adjusted.	Acceptable	0
			Not acceptable	3
22	Audit Opinion from External Auditors	Considered as 3rd party criterion, only applicable to LTO. This information is not in SIGTAS. Therefore a mean will be provided in Paradox to capture the value. To facilitate data-entry, 'Disclaimer of Opinion' will be the default value so only the balance of taxpayers' data has to be adjusted.	Unqualified opinion	0
			Qualified opinion	1
			Adverse opinion	2
			Disclaimer of opinion	3
23	Existence of Tax Holiday / Grace Period	A tax holiday implies an increasing potential risk as time passes. Further analysis will be conducted to identify if the information is available in SIGTAS. If not, a mean will be provided in Paradox to capture the value with the default value set to 'Less than 6 months'.	with in grace period	0
			after the grace period pass/Bypass grace period	3
Intelligence Criteria				
24	Multi Business Relationship	This information does not exist in SIGTAS. Therefore a mean will be provided in Paradox to capture the value.	Independent Manager	0
			Share Company	1
			Family Based Company	2
25	Computerisation of Financial Records	Considered as 3rd party criterion, only applicable to LTO. This information is not in SIGTAS. Therefore a mean will be provided in Paradox to capture the value. To facilitate data-entry, 'Mix with manual...' will be the default value so only the balance of taxpayers' data has to be adjusted.	Computerised Finance System and Check Payment	0
			Mix with Manual Record Keeping with Computer and Mix Payment System with Check and Cash	1
			Finance System is Fully Based on Manual and Cash Payment	2

26	Industry Surveillance Report	Considered as 3rd party criterion, only applicable to LTO. This information is not in SIGTAS. Therefore a mean will be provided in Paradox to capture the value. No default value is required.	Tax Payment Amount is Proportional to the Growth of the Business	0
			Declaring 5% Profit Margin Despite the Growth of the Business for the Last 5 Years	1
27	Receipt and Cash Register	Considered as 3rd party criterion, only applicable to LTO. This information is not in SIGTAS. Therefore a mean will be provided in Paradox to capture the value. To facilitate data-entry, 'Receipts and Cash Register are not Working for More than 2 Times' will be the default value so only the balance of taxpayers' data has to be adjusted.	All Receipts and Cash Registers are Working Properly	0
			Receipts and Cash Register are not Working Properly for 2 Times	1
			Receipts and Cash Register are not Working for More than 2 Times	2
28	Information Collected from Collaborative Agencies	Considered as 3rd party criterion, only applicable to LTO. This information is not in SIGTAS. Therefore a mean will be provided in Paradox to capture the value. No default value is required.	Information Collected Previously was not Effective	0
			Information Collected previously was OK but Needs mor Facts	1
			Information Collected Previously was Effective and Based on Facts	2
Risk Profile for Imports & Exports				
29	Total Risk Analysed by Customs	This is the total result of the risk analysis conducted at Customs for the Taxpayer. Not in SIGTAS and therefore a mean will be provided in Paradox to capture the value. No default value is required.	Green	0
			Yellow	1
			Red	2

Table 3.2 ERCA risk criteria

Chapter Four
Data Preparation

4.1 Overview

This chapter describes, sources of data and techniques used in preprocessing and model building phases. It also deals with the description of the data mining process undertaken based on the six step Cios et al. (2000) process model approach. All the data mining process of this research has been done in line with the six step Cios et al. (2000) process model. Its method is described in terms of a hierarchical process model, consisting of sets of tasks described at six levels: understanding of the problem domain, data understanding, preparation of the data, modeling, evaluation discovered knowledge and using the discovered knowledge.

4.2 Understanding of the Problem Domain

The first thing for DM is understand the business area and risk of fraud on the sector. So, to gain the true picture of the business area, the researcher undertook interview of domain expert up to the managers and taxpayers for the purpose of problem understanding. Besides, the researcher reviewed necessary literature in the field of fraud and E-tax to understand the business.

To understand business and coin data mining problems, the researcher reviewed deferent document of the institutions. Interviews and discussion was also made with senior managers of the organization. Domain expert consultation had been to have brief understanding on the problem area of E-tax and E-filing. Besides, the dataset is also thoroughly examined with domain experts.

From business understanding perspective, currently, there is a very poor traditional means of risk assessment(fraud detection). Through interviews and discussion made with senior managers of the organization, the institution is attempting to add more risk criteria(see in chapter three section 3.3 risk criteria). However, the institution is not using

appropriate tools and technique that helps them to identify fraud in the sector. So, the institution could not pass proper decisions to classify their taxpayer. As the auditors say, the current system, do not helps them to identify as well as to predict fraud.

Therefore, the researcher proposed a data mining technique that helps to predicts fraud, so that the institution can pass proper decisions and identifying fraudster area. This in turn has a significant impact in improving performance of tax collection in the institution.

4.3 Understanding the Data

Data understanding and data preprocessing tasks should be carried out carefully to come up with good output in data mining. The reason is that the models that will be built mainly depend on these tasks. Hence, these sections discuss about the activities performed to understand data. The data elements (attributes) found in the e-filing system are the following with detailed description and use heuristic attribute selection methods (see section 4.4.4 attribute selection).

No	Attribute Names	Description of Attribute	Data type
1	TIN	Taxpayers identification number	Number
2	Company name	Company name	Text
3	Sector Activity	Sector of the company work	Text
4	Business Activity	Based on the sector detail business activity	Text
5	Fiscal year	The company working fiscal year	Number
6	Seles income	Main business income	Number
7	Other income	Not direct income for example wastage and other related	Number
8	Gross income	Total yearly income	Number
9	Cost of good soled	Direct cost or expenses	Number

10	Gross profit	Gross profit	Number
11	Total expense	Administration and other expenses	Number
12	Net income or loss	Net profit or loss	Number
13	Loss carry forwarded	Loss came from previous years	Number
14	Loss carry balance adjusted	Adjusted loss balance	Number
15	Taxable income	Taxable income	Number
16	Tax payable	Tax to be payable	Number
17	Forwarded Withholding total	Prepayment of every sales	Number
18	Tax paid		Number

Table 4.1list of attribute in the data.

4.3.1 Collection of Initial Data

The initial data was collected from ERCA data base SIGTAS. The data was selected on the requirement of the study data of E-filing from ERCA IT directorate data base.

4.3.2 Data Quality Verification

ERCA's information technology management directorate collects data from the company's financial report with E-filing system. The collected data contain missing value and incomplete data. All companies didn't register in their proper sector. The data for all companies has focused mainly on their total expenses.

4.4 Preparation of the Data

According to [27], preprocessing helps to fill some missing values; to detect some outliers that may jeopardize the result of data mining; and to detect and remove/correct some noisy data. In relation to this, data normalization, discretization, etc need to be

performed. Moreover, to conduct the experimentation, the dataset must be prepared in the appropriate format.

To accomplish data preprocessing the first thing is gathering the data, refine it and preparing for data mining process. The dataset which has been collected from SIGTAS database cannot directly be used in data mining process. It needs further preprocessing tasks. Here, data cleansing, data reduction and data transformation are performed.

In this research, from total of 2,998 records 2,954 clean data are selected for modeling. The reason, 44 records are incomplete data.

4.4.1 Data Construction

The other important step in preparation 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 [32]. In consultation with the domain experts and risk assessment team at ERCA, the following fields that are considered essential in determining the fraudulence of claims were derived from the existing fields.

Gross profit margin is computed, total yearly income of the company mines direct cost of the company (cost of good soled) in that specific year and divided by total income (sales).

Gross Profit Margin = Gross profit/Total Seles…… (3.1)

Expense margin is computed, total yearly expense of the company is divided by total income (sales) in that specific year.

Expense Margin = Total Expense/Total Seles…….. (3.2)

Net profit margin is computed, net income of the company divided by total income (sales) in that specific year.

Net Profit Margin = Net Income/Total Seles…….. (3.3)

4.4.2 Data Cleaning

Data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. In particular, data collected from different sources has to

be managed into a form that will allow the data mining tools to be used to best advantage. This process of data cleaning and pre-processing is highly dependent on the technique to be employed. In this research since the collected dataset had some inconsistent values, such data were removed. Likely, some data contained missing value and they were also removed from the original dataset. Hence data cleaning had been required and performed on the selected dataset.

The following activities were done in data cleaning phase of the study:

- The data cleaning is done on Microsoft excel 2007.WEKA could also have been used for further data cleaning data but the researcher select excel because of familiarity and easy for work.
- 44 incomplete records are removed from the dataset because the values areloss sales income, gross profit and expenses records. For that matter unable to fill manually or Expected Maximization (EM) method.
- Then, after going through the data cleaning, the data is saved as CSV file format in which the values are saved in comma delimited form in order to create an ARFF format file.

4.4.3 Data Transformation and Concept Hierarchy

Data transformation that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.

According to [33] discretization techniques can be categorized based on how the discretization is performed, such as whether it uses class information or which direction it proceeds (i.e., top-down vs. bottom-up). If the discretization process uses class information, then we say it is supervised discretization. Otherwise, it is unsupervised. If the process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range, and then repeats this recursively on the resulting intervals, it is called top-down discretization or splitting. This contrasts with bottom-up discretization or merging, which starts by considering all of the continuous values as potential split-points, removes some by merging neighborhood values to form intervals, and then recursively applies this

process to the resulting intervals. Discretization can be performed recursively on an attribute to provide a hierarchical or multiresolution partitioning of the attribute values, known as a concept hierarchy. Concept hierarchies are useful for mining at multiple levels of abstraction.

A concept hierarchy for a given numerical attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute sales) with higher-level concepts (such as very high, high, medium, or low). Although detail is lost by such data generalization, the generalized data may be more meaningful and easier to interpret. This contributes to a consistent representation of data mining results among multiple mining tasks, which is a common requirement. In addition, mining on a reduced data set requires fewer input/output operations and is more efficient than mining on a larger, ungeneralized data set. Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining as a preprocessing step, rather than during mining.

Feature selection is extremely useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy (inclusion of irrelevant features can introduce noise into the data, thus obscuring relevant features). It is worth noting that even though some machine learning algorithms perform some degree of feature selection themselves (such as classification trees); feature space reduction can be useful even for these algorithms. Reducing the dimensionality of the data reduces the size of the hypothesis space and thus results in faster execution time.

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data.

In this research the discretization of the data with consultation of the domain expert and literature review on selected attribute.

The transform of the data as follows:-

Sector attribute have 32 deferent sector activity so the researcher took narrowed sector (big picture) divided in to 9 main sector see in the below table.

Selected main sector	Detail sector on the data
Agriculture	Agriculture, hunting, forestry, fishing…
Banking and Insurance	Banks, insurance, micro finance…..
Construction	Construction, engineering and consultation, Real estate ….
Services	Educational service, health service, professional, technical, rental and hire, transportation, storage services, other services…
Export/Import	Import, export
Hotel and related tour	Accommodation and food, entertainment and recreation, restaurant and bar, hotel, tour …
Mining	Mining, quarrying …
Manufacturing	Hand craft and cottage industries, metal and non metal industries, plastic and fiber industries …
Wholes sales and related	Durable and non durable whole sales trade, retail trade, sundry items….

Table 4.2 selected attribute with description

In Seles income attribute the department of risk discrete the sales as follows

Annual sales of the company	Risk level	Risk label
<35,000,000	0	Low
>35,000,000< = 70,000,000	1	Medium
>70,000,000 < = 220,000,000	2	High
>220,000,000	3	Very high

Table 4.3 sales attribute discretization

For gross profit margin and expense margin discretization the table below shows risk department working criteria in the authority.

Where '0' means low risk, '1' means medium risk, '2' means high risk and '3' means very high risk. Deviation means the calculated result in percent for that specific sector. Deviation of Expense margin=total expense*100/total sales and deviation of Gross profit margin=gross profit*100/total sales.

Selection Criteria Business sector Base			
No	**Business sector**	**Expense margin**	**Gross profit margin**
1	**Agriculture total**	Deviation <=10% 0 Deviation >10% and <=20% 1 Deviation >20% and <=30% 2 Deviation >30% 3	Deviation >15% 0 Deviation >10% and <=15% 1 Deviation >5% and <=10% 2 Deviation <=5% 3
2	**Banking and insurance**	Deviation <=5% 0 Deviation >5% and <=15% 1 Deviation >15% and <=25% 2 Deviation >25% 3	Deviation >50% 0 Deviation >40% and <=50% 1 Deviation >30% and <=40% 2 Deviation <=30% 3
3	**CONSTRUCTION**	Deviation <10=% 0 Deviation >10% and <=20% 1 Deviation >20% and <=30% 2 Deviation >30% 3	Deviation >30% 0 Deviation >20% and <=30% 1 Deviation >15% and <=20% 2 Deviation <=15% 3
4	**Different service sectors**	Deviation <3=% 0 Deviation >3% and <=10% 1 Deviation >10% and <=15% 2	Deviation >40% 0 Deviation >25% and <=40% 1 Deviation >10% and <=25% 2

		Deviation >15% 3	Deviation <=10% 3
5	**Hotel and related tour**	Deviation <10=% 0 Deviation >10% and <=20% 1 Deviation >20% and <=30% 2 Deviation >30% 3	Deviation >40% 0 Deviation >30% and <=40% 1 Deviation >10% and <=30% 2 Deviation <=10% 3
6	**IMPORT/EXPORT**	Deviation <5=% 0 Deviation >5% and <=10% 1 Deviation >10% and <=20% 2 Deviation >20% 3	Deviation >40% 0 Deviation >30% and <=40% 1 Deviation >10% and <=30% 2 Deviation <=10% 3
7	**MINING**	Deviation <2=% 0 Deviation >2% and <=5% 1 Deviation >5% and <=10% 2 Deviation >10% 3	Deviation >15% 0 Deviation >10% and <=15% 1 Deviation >5% and <=10% 2 Deviation <=5% 3
8	**MANUFACTURING**	Deviation <5=% 0 Deviation >5% and <=20% 1 Deviation >20% and <=30% 2 Deviation >30% 3	Deviation >25% 0 Deviation >20% and <=25% 1 Deviation >10% and <=20% 2 Deviation <=10% 3
9	**Wholesale durable goods /Whole sale**	Deviation <3=% 0 Deviation >3% and <=15% 1 Deviation >15% and <=25% 2	Deviation >30% 0 Deviation >25% and <=30% 1 Deviation >20% and <=25% 2

	non-durable	▯ Deviation >25%	3	▯ Deviation <=20%	3

Table 4.4 discretization of gross margin and expense margin

In loss carry forward attribute if a company declare loss in previous year and forwarded the loss to this year to offset the loss with current year profit. So, if have loss forward in the current year declaration discrete as 'YES' and if there is no loss forward in the current year declaration discrete as 'NO'.

In withholding credit forward attribute it is a prepayment from the daily sales 2% payment collected by agents. So based on the annual sales of the company discrete as low, medium, high and very high.

In Net profit margin attribute is based on the consultation of domain expert and common understanding every taxpayers need be profitable more than 5% because banks pay interest without any tax. So discrete if it is net profit margin more than 5% not suspected and less than 5% fraud suspected.

4.4.4 Attribute Selection

The attribute selection activities involves selection and identification of best variables or attributes for predicting fraud from business perspective in which the institutions involved and develop classification model in E-filing data.

Based on the objective of the study and the domain expert consultation, best attributes are selected. The data corpus has 18 attribute, out of these, addition three derived attributes are added to enhance the knowledge discovery and efficiency of the model. The most commonly used methods for attribute selection is the heuristic approach method.

Best step-wise attribute selection:

- Start with empty set of attributes.
- The best single-attribute is picked first.
- Then combine best attribute with the remaining to select the best combined two attributes, then three attributes.

- The process continues until the performance of the combined attributes starts to decline.

By consultation with domain expert and the above approach the following eight (8) attributers are selected:

- Sector Activity: represents the company business area.
- Fiscal Year : show the company on which period making which profit/loss
- Seles income/total sales: represent the company annually transaction of the year.
- Gross profit margin; it is derived attribute see data construction section.
- Expense margin: it is also derived attribute see data construction section.
- Loss carry forward: previously recorded loss it may be compensate with current year profit.
- Credit with holding total: it is prepaid tax and compensate with current tax payment
- Net profit margin: it is based on net income or remaining balance after deduction of cost, expense and tax. It is also derived attribute see data construction section.

4.5 Selecting of Models/Technique

Model selection for data mining is the process of providing the processed data to candidate classification algorithm and identifying the model that shows better performance. There are a number of tasks involved in the phase. Some of the tasks include selection of modeling technique; evaluate the model building model and selection of the best model.

4.5.1 Test Tesign

An experimental plan should first be set to guide the training, testing and evaluation process of the model. Mostly researchers split the dataset into training and test sets. Normally, training should be done on large proportion of the total data available, whereas testing is done on small percentage of the data that has been excluded during training of the model. In the case of decision tree classification models J48 algorithm, random forest algorithm and multi layer perception, different experiments have been done by splitting the dataset into training and testing set, used default parameter values and 10 fold cross validation. Finally, the classification model that shows better accuracy performance has been selected.

4.6 Evaluation

To evaluate the various classification models, confusion matrix which contains values of true positive (correct classification) and false positives (incorrect classifications) have been used as follows.

Confusion metrics (standard metrics)		Predicted connection label	
		Not suspected	Fraud suspected
	Not suspected	True Negative (TN)	False positive (FP)
	Fraud suspected	False Negative (FN)	True Positive(TP)

Table 4.5Confusion metrics standard metrics for evaluations

The representation of True Positive (TP), True Negative (TN), False Positive (FP), and False

Negative (FN) are defined as follows:

- True Positive (TP): The number of malicious records that are correctly identified.
- True Negative (TN): The number of legitimate records that are correctly classified.
- False Positive (FP): The number of records that are incorrectly identified as suspected however in fact they are legitimate activities.
- False Negative (FN): The number of records that are incorrectly classified as legitimate activities however in fact they are malicious.

4.6.1 Performance Measure

General performance of fraud prediction systems is measured in terms of numbers of selected features and the classification accuracies of the DM algorithms giving the best classification results. As discussed by [41] there are different techniques used for performance measuring of the fraud suspicion systems. Good fraud suspicion systems require high detection rate, low false alarm rate and lower average misclassification cost [41]. Thus during developing fraud suspicion systems; overall classification rate (OCA), detection rate (DR), false Positive rate (FPR), average misclassification cost (AMC), Error rate, and training and testing time are considered.

4.6.1.1 Error Rate

The error rate, which is only an estimate of the true error rate and is expressed to be a good estimate, if the number of test data is large and representative of the population, is defined as [42].

Error Rate= [(Total Test samples- Total Correctly Classified Samples)*100%] /Total Test Samples …….. (3.4)

4.6.1.2 Accuracy

Overall Classification accuracy (OCA) is the most essential measure of the performance of a classifier. It determines the proportion of correctly classified examples in relation to the total number of examples of the test set i.e. the ratio of true positives and true negatives to the total number of examples. From the confusion matrix, we can say that accuracy is the percentage of correctly classified instances over the total number of instances in total test dataset, namely the situation TP and TN, thus accuracy can be defined as follows [41]:

Accuracy = [(TP+TN)*100%]/TP+TN+FP+FN ……. (3.5)

4.6.1.3 Detection Accuracy

Detection accuracy (rate) refers to the proportion of suspecting detected among all suspected data, namely, the situation of TP, thus detection rate is defined as follows [41]:

Detection Accuracy = (TP*100%)/ (TP+FN) …… (3.6)

4.6.1.4 False Positive Rate

False positive rate, also known as False Alarm Rate (FAR) measures the number of misclassified positive instances in relative to the total number of misclassified instances. It can be expressed also as the proportion that normal data is falsely detected as attack behavior, namely, the situation of FP. Thus false alarm rate is defined as follows [41]:

False Positive Rate= (FP*100%)/ (FP+TN) ….. (3.7)

4.6.1.5Precision and Recall

Recall and precision are two widely used metrics employed in applications where successful of one of the classes is considered more significant than detection of the other classes [42]

4.6.1.6 Precision

Precision is the number of class members classified correctly over the total number of instances classified as class members [42].

Precision= (TP*100%)/ (TP+FP)……. (3.8)

4.6.1.7 Recall

Also called True Positive Rate (TPR), Recall measures the number of correctly classified examples relative to the total number of positive examples. In other words the number of class members classified correctly over the total number of class members [42].

Recall= (TP*100%)/ (TP+FN) …….. (3.9)

Chapter Five
Experimentation and Result Discussion

This chapter presents the experimental activities and discussion of results. In this research labeled records are used. The dataset is in a spreadsheet (Excel) from ERCA data base. It is preprocessed and ready for experimentation, which are described in the previous chapter.

5.1 Experimentation

Experiments are performed in a computer with the configurations Intel(R) Core(TM) 5 CPU 2.5GHz, 4 GB RAM, and the operating system platform is Microsoft Windows 8. WEKA(3.7.13) and Microsoft Excel (2007) is used to filter record. The following are the steps used in this study experimentation:

- In the beginning, in order to build the experiment, the researcher selected the data mining software.
- The selected records are changed from Microsoft excel format to ARFF suitable for WEKA software for data mining.
- The preprocessing tasks are done (see chapter four section preprocessing) by using Microsoft Excel for resolving missing values(incomplete data).
- At this step open the ARRF file using WEKA tool. Those training models which scores better classification accuracy has selected for this study.
- Lastly, the selected model writes the rule for the selected best model for deployment and farther work.

5.1.1Decision Tree Model Building

A decision tree classifier is one of the most widely used supervised learning methods used for data exploration, approximating a function by piecewise constant regions, and does not necessitate previous information of the data distribution. Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. The true purpose of the decision tree is to classify the data into distinct groups or branches that generate the strongest separation in the values of the

dependent variable, being superior to predict segments with a desired individual behavior such as response or activation, thus providing an easily interpretable solution [34].

5.1.1.1 Experiment J48 Decision Tree Modeling

TheJ48 algorithm is used for building the decision tree model. The training of the decision tree classification models of the experimentation is done by employing the 10-fold cross validation and the percentage split classification models.

Test No	Records	No attribute	No leave	Size of the tree	Test modes	Time to build model	Accuracy (%)
1	2954	8	79	100	10-fold cross-validation	0.31 second	94.719%
2	2954	8	79	100	Percentage split 66/34%	0.2 second	94.621%

Table 5.1 summarized output of decision tree with different test modes for J48 algorithm

In table 5.1 the result of j48 shows that the accuracy of 10-fold cross validation and percentage split (66/34%), 94.719% and 94.621% respectively. The confusion matrix of all training set:-

Confusion Matrix

ab <-- classified as

1252 42 | a = not_suspected

84 1576 | b = Fraud_suspected

As shown in the resulting confusion matrix, the J48 learning algorithm tested using 10-fold cross validation scored an accuracy of 94.719%. This result shows that in the above table different result with the same parameter and with deferent testing mode.

Detailed Accuracy by their performance measurement for percentage split (66/34%) is shown in the below table.

class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Not suspected	0.965	0.068	0.914	0.965	0.939	0.892	0.971	0.947
Fraud suspected	0.932	0.035	0.973	0.932	0.952	0.892	0.971	0.977
Weighted Avg.	0.946	0.049	0.948	0.946	0.946	0.892	0.971	0.964

Table 5.2 Output of J48 algorithm with different measurement for percentage split (66/34%)

Detailed Accuracy by their performance measurement for 10-foled cross validation as show in below table.

class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Not suspected	0.958	0.061	0.924	0.958	0.941	0.894	0.964	0.936
Fraud suspected	0.939	0.042	0.967	0.939	0.952	0.894	0.964	0.959
Weighted Avg.	0.949	0.052	0.946	0.949	0.947	0.894	0.964	0.948

Table 5.3 Output of J48 algorithm with different measurement for 10-foled cross validation

5.1.2 Experiment Random Forest

The random forest algorithm is used for building the decision tree model. The training of the decision tree classification models of the experimentation is done by employing the 10-fold cross validation and the percentage split classification models.

Test No	Records	No attribute	Test modes	Time to build the model(second)	Accuracy (%)
3	2954	8	10-fold cross-validation	1.44 second	94.0081%
4	2954	8	Percentage split 66/34%	1.4 second	94.0239 %

Table 5.4summarized output of decision tree with different test modes for random forest algorithm

In table 5.4 the result of random forest shows that the accuracy of 10-fold cross validation and percentage split (66/34%), 94.008% and 94.239% respectively. The confusion matrix of all training set:-

Confusion Matrix

a b <-- classified as

1265 29 | a = not_suspected

42 1618 | b = Fraud_suspected

As shown in the resulting confusion matrix, the random forest learning algorithm. This result shows that in the above table different result with the same parameter and with deferent testing mode.

Detailed Accuracy by their performance measurement for percentage split (66/34%) as show in below table.

class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Not suspected	0.939	0.059	0.922	0.939	0.931	0.878	0.969	0.932
Fraud suspected	0.941	0.061	0.954	0.941	0.947	0.878	0.969	0.980
Weighted Avg.	0.940	0.060	0.940	0.940	0.940	0.878	0.969	0.960

Table 5.5 Output of random forest with different measurement for percentage split (66/34%)

Detailed Accuracy by their performance measurement for 10-foled cross validation as show in below table.

class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Not suspected	0.945	0.064	0.920	0.945	0.933	0.879	0.970	0.948
Fraud suspected	0.936	0.055	0.956	0.936	0.946	0.879	0.970	0.976
Weighted Avg.	0.940	0.059	0.941	0.940	0.940	0.879	0.970	0.964

Table 5.6 Output of random forest with different measurement for 10-foled cross validation

5.1.3 Neural Network

It is represented as a layered set of interconnected processors. These processor nodes have a relationship with the neurons of the brain. Each node has a weighted connection to several other nodes in adjacent layers. Individual nodes take the input received from connected nodes and use the weights together to compute output values.

4.1.3.1 Experiment Neural Network (Multi layer perception)

A network with the input and output layer only is called single-layered neural network. Whereas, a multilayer neural network is a generalized one with one or more hidden layer. For this experiment multilayer neural network algorithm is selected.

Test No	Records	No attribute	Test modes	Time to build the model(second)	Accuracy (%)
5	2954	8	10-fold cross-validation	125.11 second	93.0941 %
6	2954	8	Percentage split 66/34%	120.13 second	92.6295 %

Table 5.7summarized output of neural network with different test modes for multilayer perception algorithm

In table 5.7 the result of multilayer perception shows that the accuracy of 10-fold cross validation and percentage split (66/34%), 93.094% and 92.629% respectively. The confusion matrix of all training set:-

Confusion Matrix

ab <-- classified as

1260 34 | a = not_suspected

47 1613 | b = Fraud_suspected

As shown in the resulting confusion matrix, the multilayer perception learning algorithm. This result shows that in the above table different result with the same parameter and with deferent testing mode.

Detailed Accuracy by their performance measurement for percentage split (66/34%) as show in below table.

class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Not suspected	0.921	0.070	0.908	0.921	0.914	0.850	0.960	0.907
Fraud suspected	0.930	0.079	0.940	0.930	0.935	0.850	0.960	0.977
Weighted Avg.	0.926	0.075	0.926	0.926	0.926	0.850	0.960	0.947

Table 5.8 Output of multi layer perception with different measurement for percentage split (66/34%)

Detailed Accuracy by their performance measurement for 10-foled cross validation as show in below table.

class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Not suspected	0.935	0.072	0.910	0.935	0.922	0.860	0.967	0.933
Fraud suspected	0.928	0.065	0.948	0.928	0.938	0.860	0.967	0.977
Weighted Avg.	0.931	0.068	0.931	0.931	0.931	0.860	0.967	0.958

Table 5.9 Output of multi layer perception with different measurement for 10-foled cross validation

5.2 Comparison of the above Experiments

Comparing different classification techniques and selecting the best model for predicting the fraud prediction is one of the aims of this study. Accordingly the decision trees particularly the J48 algorithm, random forest algorithm and neural network particularly classification approaches were used for conducting experiments.

In table 5.10 the result of 10-fold cross validation break up data into groups of the same size hold one group for test and the rest for training repeat until all folds tested.

Experiment No	Classifier algorithm	Testing mode	Accuracy	Ranking
1	J48	10-fold cross-validation	94.72%	1
2	Random forest	10-fold cross-validation	94.01%	2
3	Multilayer perception	10-fold cross-validation	93.09%	3

Table 5.10 summarized 10-fold cross-validation experiment result.

In table 5.11 the result of percentage split up data into 66/33%, 66%of the data for training and 34% for test.

Experiment No	Classifier algorithm	Testing mode	Accuracy	Ranking
1	J48	Percentage split 66/34%	94.62%	1
2	Random forest	Percentage split 66/34%	94.02%	2
3	Multilayer perception	Percentage split 66/34%	92.63%	3

Table 5.11 summarized percentage split up data into 66/34% experiment result.

The experiment and evaluation is done side by side in experimentation time and the result of each experiment based on their performance. Based on the result show in the above tables J48 is scores accuracy of 94.72% in 10-fold cross validation and scores accuracy of 94.62% in percentage split and best classifier than others algorithm.

5.3 Rules Generated with J48

There are several rules generated from the selected decision tree (J48) model for fraud prediction. But for simplicity and manageability, only those rules with large number of instances and correctly classified are taken as best rules and these are the following:-

Rule #1

If gross profit margin = low then the taxpayers classified as Fraud suspected (1053.0/4.0)

Rule #2

If gross profit margin = medium, Exp-margin = medium and Sector Activity = MANUFACTURING then the taxpayers classified as Fraud suspected (124.0/3.0)

Rule #3

If gross profit margin = medium and Exp-margin = very high then the taxpayers classified as Fraud suspected (89.0/2.0)

Rule #4

If gross profit margin = medium, Exp-margin = high and FCredi-WithHldg-Total = low then the taxpayers classified as Fraud suspected (86.0/1.0)

Rule #5

Rule #6

If gross profit margin = very high, Sector Activity = AGRICULTURE and Exp-margin = very high then the taxpayers classified as Fraud suspected (37.0/2.0)

Rule #7

If gross profit margin = very high, Sector Activity = Wholesale Trade Other Related, Exp-margin = very high FCredi-WithHldg-Total = low Sales Income = low then the taxpayers classified as Fraud suspected (31.0/2.0)

Rule #9

If gross profit margin = high and Sector Activity = AGRICULTURE then the taxpayers classified as Fraud suspected (18.0)

Based on domain expert and risk department comment the following is one of interesting rules, if companies declare low gross profit margin means based on Rule#1 and the

consultation of domain expert it is more fraud suspected than others. Beside the rule, low gross profit margin is computed only on deference of total sales and direct cost (cost of sales) so the company shows that loss before other expenses. In Rule#3 if a company declares medium gross profit and very high expense margin these shows the company try to hide the profit (in other word Fraud suspected). In Rule #9 if a company high gross profit margin and sector activity is agriculture the taxpayer classified as fraud suspected these taxpayers transfer cost into expense. Among those interesting rules and evaluation of the model from the domain expert and risk dep't perspective was evaluated and Rule #9, Rule #7, Rule #3 and Rule #1 are interesting rules.

Chapter six
Conclusion and Recommendation

6.1 Conclusion

Fraud is risk for the economy and development of developing countries. Data mining tools and techniques now came up with good result in developing countries in many field of study including finance, telecommunications, health, customer service management etc. Despite these enhancements, data mining practice is at infancy level in most developing and underdeveloped countries.

The researcher conduct interview and collect secondary data ERCA data set tried to meet the objective of the study by strictly following the six step Cios et.al (2000) data mining model on collected data set.

The result assured that with J48 10-fold cross validation 94.72% accuracy is registered. In this study the result show the business area is still at risk and need for data mining tools and techniques especially fraudster sector with sufficient level of confidence.

The study also tested the suitability of various classifier algorithms with data set arrangement and compared the results. For instance with percentage split (66/33)% resulted with J48 94.62% accuracy, random forest 94.02% accuracy and Multilayer perception 92.63% accuracy is registered. In 10-fold cross validation with J48 94.72% accuracy, random forest 94.01% accuracy and Multilayer perception 93.09% accuracy is registered.

In general, the results from this study are very promising. It is possible to identify those fraud suspicious tax filers and suggest concrete solutions for detecting them, using data mining techniques. The research is limited to current data set of e-filing data in the future researcher can continue on more data set and other approach to the better performance and improvement of the model.

6.2 Recommendation

In the belief of the researcher, findings of the study will support the institution, to work on the application of data mining techniques for successful achievement of the organizational goals.

Based on the findings, the recommendations are forwarded:

- The result of the research is promising. But this research was conducted for academic purpose. To deploy in the institution with little modification and to come up with more comprehensive models, it is recommended that experimental tests be conducted by the institution with inclusion of many dataset by using large training and testing datasets.
- ERCA should continue in this type of study in the further to get better improvement fraud prediction and for other issues. Nowadays, fraud is complex and assorted. Thus, the researcher believes that, application of other data mining techniques (rather than clustering and classification) with different algorithms in new tool is potential research area to improve performance of customer relationship management in the institution.
- Fraud does not only occur in E-filing, 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

[1] Denial mamo, 2013. “Application of data mining technology to support fraud protection, case study on ERCA”.Addis Ababa University: Addis Ababa, Ethiopia.

[2] Guo,L., 2003. “Applying Data Mining Techniques in property and casualty Insurance Science.and Technology”, New York, USA.

[3]Memorie Mwanzaand Jackson Phiri,2016.“Fraud Detection on Bulk Tax Data Using Business Intelligence Data Mining Tool: A Case of Zambia Revenue Authority”IJARCCE Vol. 5, Issue 3,

[4]Palshikar, G. 2002. “Data Analysis Techniques for Fraud Detection”,

[5] Ethiopian Chamber of Commerce, 2005. “Report” Addis Ababa, Ethiopia.

[6] Hand, D.2006. “Principles of data mining”. New Delhi, India.

[7]F. Coglitore, and G. Berryman,1988. "Analytical procedures: A defensive necessity," Auditing: A Journal of Practice & Theory, Vol. 7. No. 2, pp.150-163.

[8]Fanning et.al, 1998."Neural network detection of management fraud using published financial data," International Journal of Intelligent Systems in Accounting, Finance & Management, Vol. 7, No. 1, pp. 21 -24.

[9]Tewari R., 2014."Data Mining and other Application in financial and Tax Crime Investigations: Experience of India," Inter-American Center of Tax Administrations - CIAT, Rio de Janeiro.

[10]González, P., 2013."Characterization and detection of taxpayers with false invoices using data mining techniques," Expert Systems with Applications, International Journal, vol. 40, p. 1427–1436.

[11] Martikainen J.,2012. "Data Mining in Tax Administration - Using Analytics to enhance Tax Compliance,".

[12] Bruno et.al, 2010. “IDENTIFYING BANK FRAUDS USING CRISP-DM AND DECISION TREES” (IJCSIT) Vol.2, No.5.

[13] Belete.B,2011."Knowledge discovery for effective customer segmentation case study ERCA".Addis Ababa University: Addis Ababa, Ethiopia.

[14]Tariku .A,2011. "Mining Insurance data for fraud detection: case study of Africa insurance", Addis Ababa University: Addis Ababa, Ethiopia.

[15]Member et.al, 2005."Survey of Clustering Algorithms.IEEE Transactions on neural networks", vol.16.

[16]Nanda A., 2010. "Data Mining and Knowledge Discovery in Database: An AI perspective",Proceedings of national Seminar on Future Trends in Data Mining,

[17] Fayyad et.al, 1996."The KDD process for Extracting Useful Knowledgefrom Volumes of Data", Communications of the ACM, Vol. 39, PP. 27 -34.

[18]KdNuggets, "Data Mining Methodology", http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm,[Accessed date February 15, 2017]

[19]Chapman et.al, 2003."CRISPDM 1.0 step-by-step data mining guide", Technical report, CRISP-DM

[20]Piatetsky et.al, 1991. "Knowledge Discovery in Databases", AAAI/ MIT Press, MA

[21]Smyth et.al,1996."The Knowledge Discovery in Databases process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, Vol.18, PP.37-44.

[22]Chengwei et.al,2015. "Financial Fraud Detection Model: Based on Random Forest",International Journal of Economics and Finance; Vol. 7, No. 7.

[23]Brachman et.al,1996."The process of knowledge discovery in databases", PP. 37–57.

[24]Cabena et.al,1998."Discovering Data Mining: From Concepts to Implementation", Prentice Hall.

[25]Anand et.al,1998."Data Mining Methodology for Cross-sales", Knowledge Based Systems Journal., Vol.10, PP.449–461.

[26] Two Crows Corp,1999. "Introduction to Data Mining and Knowledge Discovery" , 3rd edition, USA.

[27]Kamber et.al,2006."Data mining concepts and techniques", 2nd edition.

[28]Silver et.al, 2001."Case study: how to apply data mining techniques in a healthcare data warehouse", Health Care Information Management, vol. 15, no. 2, pp. 155-164.

[29]Hian et.al, 2005."Data Mining Applications in Health Care", Journal of Health Care Information Management, Vol. 19, No. 2, pp. 64-72

[30]Rashmi et.al, 2010."Clustering in Data Mining", New daily, India.

[31]Cios et.al, 2005."Trends in Data Mining and Knowledge Discovery", Springer-Verlag, London, PP. 1–26.

[32]Berry et.al. 1997. "Data mining techniques for marketing, sales and customer relationship management", 2nd edition, Wiley Publishing, Inc. Indianapolis, Indiana.

[33]Jiawei et.al, 2006. "Data MiningConcepts and Techniques",2nd edition. New York,USA

[34]Vasileet.al, 2007."Analysis and Predictions on Students Behavior Using Decision Trees in Weka Environment". Proceedings of the ITI 29th Int. Conf. on Information Technology Interfaces, Cavtat, Croatia.

[35]www.erca.gov.et, [Accessed date February 1, 2017],

[36]Jeffrey A., 2004. "Computer Security".19thIEEE Computer Security Foundations Workshop.IEEE Press.

[37]Dunham et.al, 2003."Data mining introductory and advanced topics". Upper Saddle River, NJ: Pearson Education, Inc.

[38]Collier et.al, 1999."A methodology for Evaluating and selecting Data Mining Software".Proceedings of the 2ndHawaii International Conference on System Science.

[39]Pavel B., 2012. "A Survey of Clustering Data Mining Techniques". From http://msdn.microsoft.com/en-us/library/ms175595.aspx [Accessed on 25/08/2017],

[40]Charles E., 2001."The foundations of cost-sensitive learning".In Proceedings of the Seventeenth International Joint Conference of Artificial Intelligence, Seattle,Washington, PP. 973–978.

[41]Farhan et.al, 2010."Distributed and Cooperative Hierarchical Intrusion Detection on MANETs", International Journal of Computer Applications, Vol. 12, No.5, PP. 33-40.

[42] Ferri et.al, 2002. "Learning Decision Trees Using Area under the ROC Curve", Proceedings of the 19th International Conference on Machine Learning, PP. 139-146.

[43] Financial Times, 2012. "Ten ways HMRC checks if you are cheating", Closing in on Tax Evasion-HMRC's Approach,

[44]Wua et.al,2012. "Using data mining technique to enhance tax evasion detection performance", Expert Systems with Applications, An International Journal., no. 39, p.8769–8777

[45]Anderson G., 2013. "Understanding the potential of Data Mining in Botswana," Africa Journal of Computing and ICT, vol. 6, no. 1.

[46]Diwani S., 2013. "Overview Applications of Data Mining in Health Care: The case study of Arusha Region.," International Journal of Computational Engineering Research, vol. 3, no. 8.

[47] Cleary D., "irish-tax-and-customers,"http://www.sas.com/da_dk/customers/irish-tax-andcustomers.html,SAS Institute,[Accessed on 9 March 2017].

[48]KantardzicM., 2002. "Data mining: Concepts, models, methods, and algorithm", Wiley IEEEPress.

[49] Kirkos et.al, 2007. "Applying Data Mining Methodologies for Auditor Selection",Expert Systems with Applications 32 (2007).pp. 995–1003, Greece

[50] Luciano A., 2009. "Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection System".Institute of Computing.

Appendix

Appendix I: Decision Trees Generated With All Training Set Technique

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: FRAUDE DETECTION

Instances: 2954

Attributes: 8

SectorActivity

Year

Sales_Income

gross_profit_margin

Exp_margin

LossCFPrev

FCredi_WithHldg_Total

Net_ptofit_marigin_after_tax

Test mode: evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree

---

gross_profit_margin = very_high

| SectorActivity = MINNING

| | Sales_Income = very_high: Fraud_suspected (6.0/2.0)

| | Sales_Income = high: not_suspected (6.0/1.0)

| | Sales_Income = medium: not_suspected (2.0/1.0)

| | Sales_Income = low: Fraud_suspected (14.0/1.0)

| SectorActivity = AGRICULTURE

| | Exp_margin = very_high: Fraud_suspected (37.0/2.0)

| | Exp_margin = high

| | | LossCFPrev = NO: not_suspected (9.0/3.0)

| | | LossCFPrev = YES: Fraud_suspected (3.0/1.0)

| | Exp_margin = medium: not_suspected (31.0/3.0)

| | Exp_margin = low

| | | FCredi_WithHldg_Total = very_high: not_suspected (0.0)

| | | FCredi_WithHldg_Total = high: not_suspected (2.0)

| | | FCredi_WithHldg_Total = medium: Fraud_suspected (2.0)

| | | FCredi_WithHldg_Total = low: not_suspected (8.0/1.0)

| SectorActivity = Hotel_and_tour_services: not_suspected (18.0)

| SectorActivity = SERVICE: not_suspected (181.0/9.0)

| SectorActivity = Bank_and_insurance: not_suspected (35.0)

| SectorActivity = Import_Export: not_suspected (75.0/4.0)

| SectorActivity = MANUFACTURING

| | Exp_margin = very_high

| | | FCredi_WithHldg_Total = very_high: not_suspected (2.0)

| | | FCredi_WithHldg_Total = high: not_suspected (3.0/1.0)

| | | FCredi_WithHldg_Total = medium: not_suspected (2.0)

| | | FCredi_WithHldg_Total = low: Fraud_suspected (19.0/4.0)

| | Exp_margin = high: not_suspected (26.0/4.0)

| | Exp_margin = medium: not_suspected (107.0/2.0)

| | Exp_margin = low: not_suspected (19.0/4.0)

| SectorActivity = CONSTRUCTION: not_suspected (115.0/6.0)

| SectorActivity = Wholsale_Trade_Other_Related

| | Exp_margin = very_high

| | | FCredi_WithHldg_Total = very_high: Fraud_suspected (1.0)

| | | FCredi_WithHldg_Total = high

| | | | LossCFPrev = NO: not_suspected (13.0/2.0)

| | | | LossCFPrev = YES: Fraud_suspected (3.0)

| | | FCredi_WithHldg_Total = medium: not_suspected (15.0/2.0)

| | | FCredi_WithHldg_Total = low

| | | | Sales_Income = very_high: not_suspected (4.0)

| | | | Sales_Income = high: Fraud_suspected (3.0/1.0)

| | | | Sales_Income = medium: Fraud_suspected (11.0/3.0)

| | | | Sales_Income = low: Fraud_suspected (31.0/2.0)

| | Exp_margin = high: not_suspected (42.0/2.0)

| | Exp_margin = medium: not_suspected (82.0/7.0)

| | Exp_margin = low: not_suspected (2.0/1.0)

gross_profit_margin = high

| SectorActivity = MINNING: Fraud_suspected (1.0)

| SectorActivity = AGRICULTURE: Fraud_suspected (18.0)

| SectorActivity = Hotel_and_tour_services: Fraud_suspected (1.0)

| SectorActivity = SERVICE: not_suspected (71.0/8.0)

| SectorActivity = Bank_and_insurance: not_suspected (1.0)

| SectorActivity = Import_Export: not_suspected (38.0/4.0)

| SectorActivity = MANUFACTURING

| | Exp_margin = very_high: Fraud_suspected (6.0)

| | Exp_margin = high: Fraud_suspected (6.0/1.0)

| | Exp_margin = medium: not_suspected (42.0/5.0)

| | Exp_margin = low: not_suspected (10.0)

| SectorActivity = CONSTRUCTION

| | Exp_margin = very_high: Fraud_suspected (9.0)

| | Exp_margin = high: Fraud_suspected (27.0/6.0)

| | Exp_margin = medium

| | | Year = 2013: not_suspected (8.0)

| | | Year = 2014: Fraud_suspected (5.0/2.0)

| | | Year = 2015: Fraud_suspected (7.0/3.0)

| | Exp_margin = low: not_suspected (15.0)

| SectorActivity = Wholsale_Trade_Other_Related: not_suspected (0.0)

gross_profit_margin = medium

| Exp_margin = very_high: Fraud_suspected (89.0/2.0)

| Exp_margin = high

| | FCredi_WithHldg_Total = very_high: not_suspected (9.0/1.0)

| | FCredi_WithHldg_Total = high: not_suspected (24.0/1.0)

| | FCredi_WithHldg_Total = medium

| | | Sales_Income = very_high: Fraud_suspected (0.0)

| | | Sales_Income = high: Fraud_suspected (8.0/1.0)

| | | Sales_Income = medium: not_suspected (7.0/3.0)

| | | Sales_Income = low: not_suspected (2.0)

| | FCredi_WithHldg_Total = low: Fraud_suspected (86.0/1.0)

| Exp_margin = medium

| | SectorActivity = MINNING: not_suspected (0.0)

| | SectorActivity = AGRICULTURE: Fraud_suspected (1.0)

| | SectorActivity = Hotel_and_tour_services: Fraud_suspected (5.0)

| | SectorActivity = SERVICE: not_suspected (36.0)

| | SectorActivity = Bank_and_insurance: not_suspected (3.0)

| | SectorActivity = Import_Export: not_suspected (83.0/1.0)

| | SectorActivity = MANUFACTURING: Fraud_suspected (124.0/3.0)

| | SectorActivity = CONSTRUCTION: Fraud_suspected (29.0/2.0)

| | SectorActivity = Wholsale_Trade_Other_Related: not_suspected (60.0)

| Exp_margin = low

| | SectorActivity = MINNING: not_suspected (0.0)

| | SectorActivity = AGRICULTURE: Fraud_suspected (13.0/1.0)

| | SectorActivity = Hotel_and_tour_services: not_suspected (3.0)

| | SectorActivity = SERVICE: not_suspected (7.0)

| | SectorActivity = Bank_and_insurance: not_suspected (0.0)

| | SectorActivity = Import_Export: not_suspected (72.0/5.0)

| | SectorActivity = MANUFACTURING: not_suspected (26.0/2.0)

| | SectorActivity = CONSTRUCTION: not_suspected (9.0/1.0)

| | SectorActivity = Wholsale_Trade_Other_Related: not_suspected (11.0)

gross_profit_margin = low: Fraud_suspected (1053.0/4.0)

Number of Leaves : 79

Size of the tree : 100

Time taken to build model: 0.41 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.05 seconds

=== Summary ===

Correctly Classified Instances	2828	95.7346 %
Incorrectly Classified Instances	126	4.2654 %
Kappa statistic	0.9137	
Mean absolute error	0.0731	
Root mean squared error	0.1912	
Relative absolute error	14.8516 %	
Root relative squared error	38.5379 %	
Coverage of cases (0.95 level)	99.1537 %	
Mean rel. region size (0.95 level)	65.2167 %	
Total Number of Instances	2954	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.968	0.051	0.937	0.968	0.952	0.914	0.983	0.972	not_suspected
	0.949	0.032	0.974	0.949	0.962	0.914	0.983	0.986	Fraud_suspected
Weighted Avg.	0.957	0.040	0.958	0.957	0.957	0.914	0.983	0.980	

=== Confusion Matrix ===

a	b	<-- classified as
1252	42	a = not_suspected
84	1576	b = Fraud_suspected