2021-06-17

CONSTRUCTING PREDICTIVE MODEL FOR TAX CLAIM FRAUD DETECTION USING DATA MINING TECHNIQUES: THE CASE OF AMHARA REVENUE AUTHORITY

Asmero, Lakiyaw

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND GRADUATE

STUDIES COMPUTING FACULTY

CONSTRUCTING PREDICTIVE MODEL FOR TAX CLAIM FRAUD DETECTION USING DATA MINING TECHNIQUES: THE CASE OF AMHARA REVENUE AUTHORITY

By:Asmero Lakiyaw

June 17, 2021

BAHIR DAR, ETHIOPIA

Constructing Predictive Model for Tax Claim Fraud Detection Using Data Mining Techniques: The Case of Amhara Revenue Authority

Asmero Lakiyaw

A thesis submitted to the school of Research and Graduate Studies of Bahir Dar Institute of Technology, BDU in partial fulfillment of the requirements for the degree of masters of Science in the information technology in the computing faculty

ADVISOR:

Tesfa Tegegne (Dr)

June 17, 2021
Bahirdar, Ethiopia

DECLARATION

This is to certify that the thesis entitled “Constructing Predictive Model For Tax Claim Fraud Detection Using Data Mining Techniques: The Case Of Amhara Revenue Authority”, submitted in partial fulfillment of the requirements for the degree of Master of Science in Information Technology under computing faculty ,Bahir Dar Institute of Technology , is a record of original work carried out by me and has never been submitted to this or any other institution to get any other degree or certificates. The assistance and help I received during the course of this investigation have been duly acknowledged.

Asmero Lakiyaw Bezabih 16/11/2013EC

Name of the candidate signature Date

BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
FACULTY OF COMPUTING

Approval of thesis for defense result

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of Student Asmero Lakiyaw Bezabih Signature _____ Date 12/11/2013 E.C

As members of the board of examiners, we examined this thesis entitled "*constructing predictive model for tax claim fraud detection using data mining techniques: the case of amhara revenue authority*" by *Asmero Lakiyaw Bezabih.* We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science in "Information Technology".

Board of Examiners

Name of Advisor	Signature	Date
Tesfa Tegegne (Dr)		JULY-19-2021
Name of External examiner	Signature	Date
Kindie Biredagn (Dr)		**July-17-2021**
Name of Internal Examiner	Signature	Date
Belete Biazen		16/11/2013 E.C
Name of Chairperson	Signature	Date
Alemu K.		July-19-2021
Name of Chair Holder	Signature	Date
Derejaw Lake		JULY-19-2021
Name of Faculty Dean	Signature	Date
Belete Biazen		16/11/2013 E.C

ACKNOWLEDGEMENT

First and foremost, I want to express my gratitude to God and St. Mary for their blessings and for providing me with the strength and knowledge to complete my thesis.

I'd like to express my gratitude to my advisor, Dr. Tesfa Tegegne, whose support and supervision enabled me to complete my thesis. He has shown a genuine interest in my work and has always been available to assist me. I am grateful for his patience, curiosity, and vast understanding.

I'd also like to thank those who helped me create training questions for this thesis by hand: Ayalew Atinafu, Zelalem Tsegaye, Dereje Mamo, and Muluken Birhan. Without their good participation and input could not have been successfully performed.

The Amhara Revenue Authority, especially tax education and communication directorate and tax assessment and collection directorate and tax audit and investigation directorate, thank you for your cooperation in supplying me the required data used in the preparation of dataset.

Last of all, I would like to thank my family for their continuous love and respect, support and encouragement to make my dream become real especially to my wife Yeshareg. Thank you. God bless you!

ABSTRACT

Revenues are vital sources of public infrastructure. The presence of collective consumption of properties and facilities necessitates putting some of our income into government hands. Nevertheless, gathering of tax is the main source of income for the government; it is challenging difficulties with fraud. Fraud encompasses one or more persons who deliberately act in secret to deprive the government income and use for their own benefit. Deception can take an unlimited variety of different forms. Fraudulent claims account for a significant portion of all dues received by auditors, and cost billions of birrs annually.

In this study, experiments were performed by succeeding the six step Cios et al. (2000) KDD process model. It starts from a business understanding in the AMRA tax policy system and fraud, specifically taking place audit data set. By compelling the data from database of AMRA and understanding of the data with the help of domain expertise and literature. In data preprocessing; missing values, inconsistencies, outliers and related issue handled properly. Afterward that, construction of models and analysis of the result done to facilitate decision making in the business risk analysis.

To gather the data, the researcher used interview and observation for primary data and database analysis for secondary data.

To conduct this research, we used a total of 12000 records for training the classifier model. Experiment taking place on different classification algorithms such as Naïve Bayes, random forest and J48 algorithms were done. the result of the various models have compared to find the best model using percentage split (66/34%) and 10-fold cross validation evaluation methods.

The study discovers that J48 classification algorithm achieves with superior accuracy than other testing mechanisms. J48 recorded an accuracy of 99% were 11880 instances are correctly classified out of 12000 test cases. As a result, further research directions are suggested in order to come up with a workable system in the subject field..

Key word: fraud, classification, j48, Data Mining.

List of Abbreviation

AMRA: Amhara Revenues Authority

ARFF: Attribute Relation File Format

BI: Business Intelligence

CRISP-DM: Cross Industry Standard Process for Data Mining

CSV: Comma Separated Values

DM: Data Mining

GIRS: Regional Inland Revenue Service

ICTD: Information Communication Technology Directorate

KDD: Knowledge Discovery in Databases

SIGTAS: Standardized Integrated Government Tax Administration System

TIN: Taxpayer Identification Number

WEKA: Waikato Environment for Knowledge Analysis

TABLE OF FIGURES

LIST OF TABLES

TABLE CONTENTS

CHAPTER ONE

INTRODUCTION

1.1 Background

In contrast to traditional statistical analysis, where experiments are built around a specific hypothesis, data mining allows for data exploration and analysis without a specific hypothesis in mind. While this openness gives data mining projects a strong exploratory element, it also necessitates a structured approach in order to provide useable results. (Elkan, 2001) . Data mining uses statistics, database theory, machine learning, pattern recognition, and visualization approaches to automatically extract concepts, concept interrelations, and intriguing patterns from huge corporate databases. (Lindgreen, 2004).

The term "fraud" here refers to the improper use of a profit organization's system that does not necessarily result in immediate legal consequences. Fraud can become a business-critical concern in a good environment if it is exceedingly frequent and the protective methods are not fail-safe. As part of comprehensive fraud prevention, fraud detection automates and helps reduce the manual aspects of the screening/checking process. This is one of the most well-known applications of data mining in industry and government.

In a variety of disciplines, the rise of information technology has resulted in a large number of databases and massive amounts of data. Data mining is a logical process that is used to search through enormous amounts of data in order to identify usable data, as a consequence of database and information technology research. The purpose of this methodology is to discover previously discovered patterns. Once these patterns have been discovered, they may be leveraged to make specific decisions for the growth of their enterprises.

The step by step process are

- Exploration

- Pattern identification
- Deployment

Exploration: Data exploration begins with the cleaning and transformation of data into a new format, followed by the identification of relevant factors and the type of data based on the problem.

Pattern Identification: The second phase is to develop pattern identification when data has been discovered, refined, and defined for the specified variables. Identify and select the patterns that make the most accurate predictions.

Deployment: Patterns are used to get the desired result.

Data mining (also known as data or knowledge discovery) is the process of analyzing data from multiple sources and distilling it into valuable information in order to improve income, reduce costs, or do both. It enables users to examine data from a variety of perspectives, categorize it, and effectively summarize the links discovered. It is the technique of determining relationships or correlations among a large number of fields in a relational database. (Palace, 1996).

Today, (R., 2012) tax authorities around the world are under increasing pressure to collect additional tax revenues, uncover underreported individuals, and predict the irregular behavior of non-filing taxpayers. Most tax authorities require collecting tax data from a number of independent sources and performing data matching and checking with other sources to find cases of non- compliance. As a result, tax evasion detection performance has been rather limited in the absence of information technology tools..

The amount of business data that is generated has risen steadily every year (Kanakalaksmil, 2019) and more and more types of information are being stored in unstructured or semi structured formats. Traditional data mining has no power anymore to deal with the huge amount of unstructured and semi structured written materials based on natural languages. Amhara Revenue Authority is a quasi- governmental organization which is mandated to collect revenue on behalf of the Government of the Republic of Ethiopia. Some of its main responsibilities include;

a) To assess and collect taxes and duties in a timely manner.
b) Provide guarantee that all monies collected are properly accounted for and banked.

c) To provide the necessary support to woredas with a view to harmonizing federal and regional tax administration

AMRA is home to various vital corporate databases that store massive amounts of data usually referred to as taxes and taxpayer information. Information is spread across multiple databases in several Divisions and Departments.

1.2 Problem Statement

African governments and associated community sector agencies are under increasing pressure to operate more professionally and effectively in all parts of the continent. Previously, the traditional approach to risk management suited many authorities well, but more modern technologies, such as Business Intelligence, Data Warehouse, and Data Mining, are now required to combat fraud, mistake, and waste at the Tax Administration Office (Atos, 2015).

The Internal Revenue Service (IRS), which is in charge of taxation in the United States, has used data mining techniques for a variety of purposes, including assessing the risk of taxpayer compliance, detecting tax evasion and illegal financial activities, detecting housing tax abuse, and detecting fraud by taxpayers who receive income from the government (Berry, M & Linoff, G, 1997) .

A database system is used for control of commercial transactions of taxpayers by the Amhara Revenue Authority. The main problem during tax collection is to get the exact taxpayer income report in the tax collector's offices. If the tax collection offices are unable to collect tax on the basis of the income of the taxpayer, the problem will fall into the annual government budget. Government annual expenditure depends on its revenue (Atos, 2015). The annual reports of the authority and several documents show that AMRA can't effectively collect revenue generated by the economy.

Finding out potential non-filers for tax liabilities, identifying potential taxpayers who are under-reporting, improving the submission of tax dedicators, identifies non-compliance in services and is important for an effective investigation. (Atos, 2015). So, detecting fraud using normal audit procedures is a difficult task ((AMRA), 2019). First of all, there is a lack of knowledge on fraud features. Secondly, most auditors

lack the knowledge to identify it. Finally, the auditors are deliberately tried by financial managers and accountants (Atos, 2015).

The use of the taxpayer database data collection tool is important, as massive information from each contributor is collected daily to the authority which cannot be managed easily in traditional ways. AMRA uses a tax monitoring database system and makes it easier for taxpayers to trade. The main issue is to get the exact taxpayer income report to the tax collector's offices. If a company uses a database system to declare what it wants online, including zero and annual tax credit, then it can be used efficiently. This database system is intended for all contributions and also implemented by the authority soon. On this situation, the following research questions are addressed by AMRA that implements data mining tools and technical needs to solve the problem to this end.

- What are the main determinant factors (attributes) of fraud from the taxpayer data?
- Which techniques of data mining are effective in developing a fraud detection model?

1.3 Objective of the Study

1.3.1 General Objective of the Study

The main objective of this research is to devise a predictive model to examine fraud suspect or non-fraud suspect of tax liabilities in order to improve Amhara Revenue Authority's effective tax collection

1.3.2 Specific Objective of the Study

The following specific objectives are formulated to achieve the general purpose of this research:

- ✓ Understand the business environment to identify and know fraudulent transactions.

- ✓ Identify and collect the necessary taxpayer data or profile from the Authority database.
- ✓ Construct AMRA target dataset used for data mining tools following major data preparation and pre-processing steps.
- ✓ Select a suitable model for identifying and revealing the factors behind taxpayers' compliance and non-compliance.
- ✓ Examine patterns that reveal taxpayer status relationships with other variables.
- ✓ Conduct experiments to evaluate system accuracy.
- ✓ Draw Recommendation on the basis of the study findings.

1.4 Limitation and Scope of the Study

The study was discovered the applicability of DM for fraud claim prediction and detection in Amhara Revenue Authority (AMRA) focused on the audit process owner. This research is focused on data from domestic taxes only.

The authority categorized its taxpayers in to three. Category „A" taxpayers annual income is more than 1,000,000 Birr. Category „B" taxpayer's annual income is between 500,000 and 1,000,000 Birr. Category „C" taxpayers annual income is less than 500,000 The scope of this research is limited to analysis of the category of "A" and category „B" taxpayer's with a total of 12000 record data of the year 2008-2011 E.C . Since Category „A" and category "B" taxpayers are large in number when compared to other categories, and they are more diverse; it is believed that the researcher of this paper gets sufficient data od category A and B for the purpose of the research. Due to the absence of sufficient data , it is difficult to cover all taxpayers who are found in other categories

To achieve the objective of this study, DM technique is used; i.e classification technique to develop prediction model, which helps to identify tax fraud and non -fraud suspicion.

The approach follow in this research is rule based data mining for fraud prediction in AMRA taxpayer data set.

1.5 Significance of the Study

Fraud detection is an important part of the modern finance industry. In this research, we have investigated the current practices in revenue fraud detection using data mining techniques. This research is important to AmRA because it announces the use of the data mining technology, which can extract good new knowledge and good knowledge from the data repository and predict other situations which are used in pro-active prevention techniques to take acceptable decisions about fraud protection.

Two important contributions will be made in this study. The first part of the study shows that the Amhara (AmRA) tax administration has problems when it comes to detecting fraud.

Our second point is the respective authorities may use the results of this study to make optimal use of resources in the field of fraud prevention, auditors and other specialists in fraud prevention or investigation. In addition, the study output is used to develop appropriate training and fraud prevention programs 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.

1.6 Organization of the Thesis

The research is organized by five sections. The first chapter is an overview of the thesis which contains background of the study, problem statement, objective, scope and limitation, the significance of the study. A literature review of data mining technology, data mining tasks, data mining tools (technique), and application of data mining and local related work is devoted in the second chapter. Research methodology, the tools used and the overall steps of data preparation are included in the third chapter. The fourth chapter deals with the result and finding more on experimentation with discussion. Finally, Chapter 5 presents the results of the study, findings and research recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.1 Concepts of Data Ming

In this chapter we try to give the research pictures with related work in this section and to define and explore various concepts of data mining.

2.1.1 Introduction

Data Mining is a system that utilizes a variety of tools to unlock heterogeneous and distributed historical information in big databases, warehouses and other massive repositories to identify data patterns that are valid, innovative, helpful, and comprehensible to the user (Ashour, Amira S, Dey, Nilanjan, & Dac Nhuong, 2014). It is also termed as Knowledge Discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

Clearly, massive numbers of data are generated in various fields of study more than ever. It is difficult to study the relationship, locate specific groups of data and collect information from this bulk of data. Based on the above reasons and the broad interest in the data collection, it becomes more important than ever for similar data to be grouped into one cluster and classified into the same label based on its predetermined class. Many techniques have been developed over the years to address this problem. However, there is a long way to go to make the most of what classification in data mining can do (D. Larose & C. Larose, 2014).

Data mining is an interdisciplinary approach that includes statistical tools and models, artificial intelligence, pattern recognition, data visualization, optimization, data collection, high-end and computing and others (Hajizadeh, Ardakani, D. , & Shahrabi, J, 2010). It is employed in various areas of study, including Biomedicine, genetic engineering function, DNA pattern analysis, disease diagnostics, retail data, telecommunications, sales, financial analysis and astronomy (D. Larose & C. Larose, 2014).

Many scientists write about data mining in order to find the knowledge that allows us to find the pattern in a massive collection of information which is very helpful in deciding the future using a computer program.

The analysis of data from databases is about solving problems. What about KDD? What about it? See the following paragraph.

2.1.2 Data Mining and knowledge Discovery

In Han (Han, J. & Kamber, M, 2006) the interactive as well as iterative nature of KDD emphasizes that many human decision-making processes and different steps are repeated frequently when the knowledge is refined.

The KDD process comprises 5 steps as shown in Figures 1.

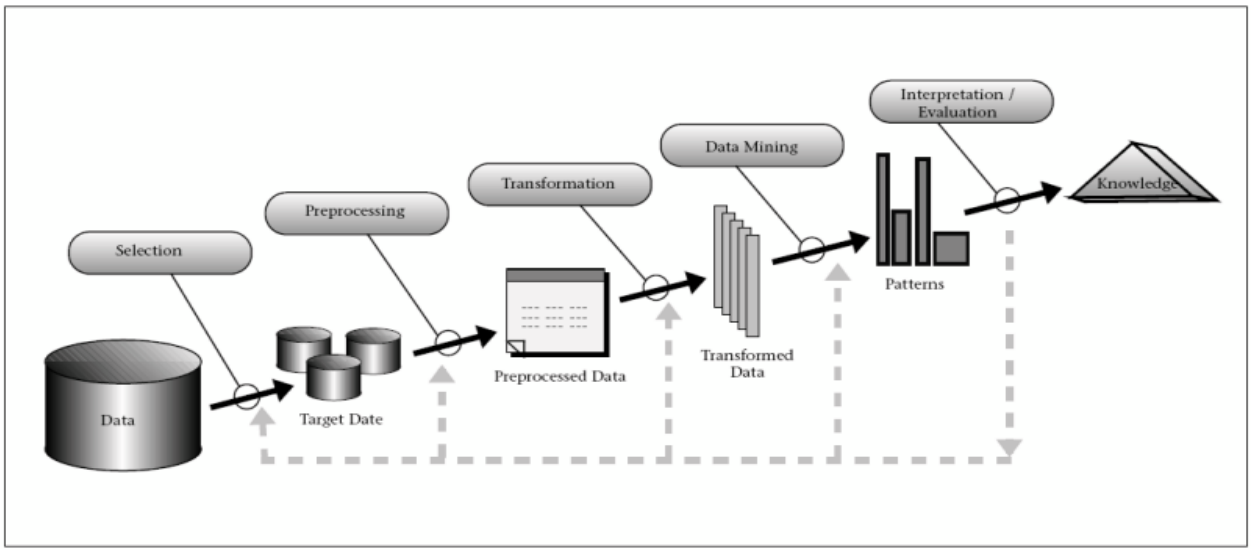


Figure 1 Knowledge discovery Process (Hendrickx, Cule, Meysman, & Stefan, 2015)

Data mining is a logical process which is used to find useful knowledge through a large number of data. The objective of this technique is to find previously unknown patterns. After these patterns are established, they can be further used to decide on their business development.

Data mining is a new method to discover knowledge from mass data. At present, researchers prefer to use special tools of data mining, such as Clementine and DB Miner,

to perform it (J.Han & M.Kamber, 2004) (Zheng Xinqi & Li Xinyun) but rarely use MATLAB as the mining tool.

Breaking the gap in analyzing large data volumes and extracting information and information that is useful for decision making has emerged in new years from the new generation of computerization techniques called data-mining (DM) or data-base knowledge discovery (KDD). Data mining is a process used to identify patterns and relations in data using a variety of data analysis tools to make effective predictions (Chung, Gray, & Mannino, 2005). Information is essentially extracted from the large data size. Data mining is a method of fraud detection, risk assessment and retailing of products (Jeffrey, 2006). Data mining requires the use of data analyses, tools for detecting patterns and relations that were previously unknown and valid in large data sets. Data mining is currently being used to identify various types of crimes such as illicit money transfers and tax fraud. The use of data mining is common in various organizations such as supermarkets, banks, insurance companies and research institutions. DM is a multidisciplinary approach involving statistical tools and models, artificial intelligence, pattern recognition, viewing, optimization, retrieval, high and computing information and other methods (Guo, 2003). Data mining (DM), a process where progress is defined by means of automatic or manual methods, is an iterative process. DM is most useful in the scenario of exploratory analysis where no predetermined ideas exist about what is an interesting result (Kantardzic M. , 2002). The use of financial classification data mining techniques is a fertile field of research. Many law enforcement agencies and special enquiries with the aim of identifying fraudulent activities have also successfully exploited data mining.

2.2 Data mining process

The discovery of knowledge is an iterative process, since you may need to proceed back and forth to obtain the necessary data knowledge. The process of knowledge finding is pictured and shows that data mining is only a part of the whole process (Sowan, 2011).CRISP-DM (CRoss Industry Standard Process for Data Mining) (D. Larose & C. Larose, 2014) proposed a complete process model for carrying out data mining projects. This process model is both industry-independent and technological independent. It is one

of the most widely used models for the discovery of knowledge. It has already been recognized and used relatively widely in the fields of research and industry in particular. The CRISP-DM main goal is to deliver an efficient process to generate knowledge from the extensive data repositories. The value of such vast amounts of data needs to be unlocked by the use of data mining in order to utilize its potential since it is not enough just to store data without using it to learn from previous occurrences. The Data mining phase can be divided into the following (U. Keshavamurthy & H.S. Guruprasad, 2014);

2.2.1 Business understanding

The first phase ensures that all participants understand the project objectives from a corporate or business viewpoint. These business goals are then integrated into a detailed project plan and the problem definition for data mining. Tasks and functions performed by a tax audit agency would understand the information collected and managed, as well as specific problems relating to improving audit efficiency. This information would form part of the data mining problem's definition and project plan.

Cross Industry Standard Process for Data Mining (CRISP-DM)

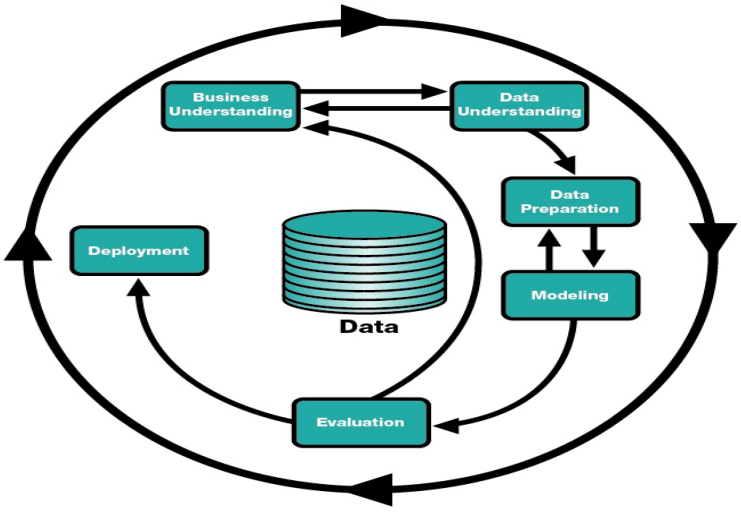


Figure 2 Generic process of Knowledge Discovery

2.2.2 Data understanding

The second phase aims to evaluate data sources, quality and features. This primary study can also provide insight into the project's focus. The result is an in-depth understanding of key data elements used for model development. For tax agencies with many data sources, this phase may take some time, but it is essential for research.

2.2.3 Data preparation

The next step consists of placing the data in a building model format. In the business understanding step, the expert uses the company objectives to determine which types of data and data-mining algorithms to use. This phase also resolves uncovered data problems, such as missing data, in the data understanding phase.

It should be noted that in fact although stored in the database, many problems affecting business and its processes are associated with bulk Tax data. These can generally be divided into two groups: incorrect data and inconsistent data. (H. Singh & S. Bhagat,, 2015). Efforts must be made for all organizations to identify data problems before the requirement arises.

2.2.3.1 Defective Data

Data defects of many kinds exist. Default data is defined as inaccurate, incomplete, unavailable or outdated data. For instance, manually entered data is often complete with typographical mistakes and spelling errors. At times they simply neglect to enter correct data in the wrong fields or to fill in all the fields. Data failure may also occur when a system is moved from one platform to another, when an old application is substituted for a new application, or when data is moved automatically and scheduled from one application to another. Or indeed, data from various sources, such a scenario as the one addressed by this study, must be integrated (H. Singh & S. Bhagat,, 2015).

The problem with defective data is that when it enters the system, it is difficult to find out. It is best to prevent the entry of defective data first of all so that companies can invest in systems that validate and fix data on the source when they enter the system or move between systems via an application interface. The company must devote data profiling and cleaning tools to clean and validate data sets before they are uploaded to

data stores for further use in BI to fix defective data in the system already (H. Singh & S. Bhagat,, 2015) (Kumari, 2013).

2.2.3.2 Inconsistent Data

Another problem found on data is inconsistency. This occurs when the data is duplicated or removed over time. For example, over time customer data deteriorate when people marry, divorce, die, move or modify names (H. Singh & S. Bhagat,, 2015) (Kumari, 2013).

2.2.4 Modeling

The modeling phase involves the development of algorithms for data mining that reveal interesting data patterns. There are various techniques for data mining; they can be used to determine a certain type of knowledge. For example, a tax authority will use classification or regression models to identify features of more productive fiscal audits. Each technique requires special data types that may require a return to the data preparation phase. The modeling stage produces a model or a series of models which contain the knowledge discovered in a suitable format.

2.2.5 Evaluation

We focus in this phase on assessing the model's quality. Algorithms for data mining can detect an unlimited number of patterns; however, many of these may be meaningless. This phase helps to identify the models that are useful to achieve the business goals of the project. A predictive audit result model is assessed against a target series of historical audits for which the results are known in the context of audit selection.

2.2.6 Deployment

The organization incorporates the results from data mining into the daily decision-making process in the deployment stage. Based on the significance of the results, only minor changes may be necessary, or a major overhaul of processes and decision-making support systems may be necessary. In the deployment phase, a recurred process for model improvements is also created. Over time, tax laws will probably change. To update the models accordingly and deploy new results, analysts require a standard process. Proper

results presentation ensures that decision-makers actually take advantage of the information.

This can be as simple as reporting or as complex as implementing a company-wide reproducible data mining process. Project managers must from the outset understand the actions they will have to take to use the final models..

Each data mining project comprises the six phases described. Each phase is important, however, and the sequence is not rigid; you might need to move forward between phases in some projects. The next phase or the next task in one phase depends on the results of each phase. The inner arrows show the main and most frequent phase dependencies. The external circle symbolizes the cyclical nature of data mining projects, namely that learners can trigger new and focused business issues following implementation and in the course of a data mining project. Consequently, subsequent data mining projects benefit from previous experience.

The analyst uses a collection of techniques and tools typically to create the models. Machine learning, statistical analysis, pattern recognition, signal processing, evolutionary computing, and patterns visualization are all disciplines in the field of data mining technology.

2.3 Data Mining Techniques

The study examined some of the available data extraction techniques, including the detection of fraud. The techniques are described as;

2.3.1 Classification

Classification constructs (from the training set) to use a model (from the target set) to predict unknown object categorical labels to distinguish objects of different classes. These are predefined, discreet and uncontrolled category labels (Sharma, Anuj & Kumar Panigrahi, Prabin, 2012). The research literature defines classification or prediction as the method by which a set of common characteristics (patterns) are identified and models are proposed that describe and differentiate data classes or concepts. Neural networks, the Naïve Bayes Techniques, Random Forests, Decision trees as well as support vector machines are common methods of classification. Such classification tasks, among other

types of fraud, are used for detecting credit card, health care and car insurance. Classification is one of the largest common learning models for the use of fraud detection in data mining. Naïve Bayes, decision tree, neural network, genetic algorithm are used in the classification.

Decision Tree

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

A decision tree model is a collection of rules for breaking down a big heterogeneous population into smaller, more homogeneous groups based on a specific objective variable. (Hajizadeh, Ardakani, D. , & Shahrabi, J, 2010). The decision tree model is used to either calculate the probability that a given record belongs to each of the categories, or to classify the record by assigning it to the most likely class. Decision tree can also be used to estimate the value of continuous variable.

Decision tree is a group of nodes, organized as a binary tree. The leaves render decisions; in our case, the decision would be “likes” or “doesn’t like.” Each interior node is a condition on the objects being classified; in our case the condition would be a predicate involving one or more features of an item (Quinlan & J. R. C4.5, 1993).

To classify an item, we start at the root, and apply the predicate at the root to the item. If the predicate is true, go to the left child, and if it is false, go to the right child. Then

repeat the same process at the node visited, until a leaf is reached. That leaf is classified the item as liked or not.

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

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

- Traditional statistical forms of analysis (such as multiple linear regressions)
- A variety of data mining tools and techniques such as neural networks
- Newly developed multidimensional forms of reporting and analysis found in the field of business intelligence.

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

Decision tree can be implemented with several algorithms. Some of them are J48, ID3, C4.5, CART, etc. The decision tree C4.5 algorithm is a practical method for inductive inference (Peng, Jin, Lianwen, & Wu, Yaqiang, 2019). Connecting *J48* to the cross-validation fold maker in the normal way, but make the connection twice by first choosing training set and then choosing test set from the pop-up menu for the cross-validation fold maker. Amongst other enhancements (compared to the ID3 algorithm) the

C4.5 algorithm includes different pruning techniques and can handle numerical and missing attribute values. C4.5 rejects over fitting the data by determining a decision tree, it handles continuous attributes, is able to choose an appropriate attribute selection measure, and handles training data with missing attribute values and improves computation efficiency. C4.5 builds the tree from a set of data items using the best attribute to test in order to divide the data item into subsets and then it uses the same procedure on each sub set recursively. The main problem in decision tree is deciding the attribute, which will best partition the data into various classes (Meera & Srivatsa, SK., 2010). The ID3 algorithm is useful to solve this problem.

2.3.2 Clustering

Clustering is used to divide objects into previously unknown practically meaningful groups (i.e. clusters), with the objects in a cluster being similar to one another but very dissimilar to the objects in other clusters. Clustering is also known as data segmentation or partitioning and is regarded as a variant of unsupervised classification (Sharma, Anuj & Kumar Panigrahi, Prabin, 2012). Cluster analysis decomposes or partitions a data set (single or multivariate) into dissimilar groups so that the data points in one group are identical to each other and are as dissimilar as possible from the data points in other groups (Soni, Jyoti, Ansari, Uzma, & Sharma, Dipesh, 2011). It is recommended that data objects in each cluster should have high intra-cluster similarity within the same cluster but should have low inter- cluster similarity to those in other clusters (Sharma, Anuj & Kumar Panigrahi, Prabin, 2012). The most common clustering techniques are the K-nearest neighbor, the Naïve Bayes technique and self- organizing maps.

K-Means clustering algorithm

k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low (J.Han & M.Kamber, 2004). Cluster similarity is measured in respect to the mean value of the objects in a cluster, which can be viewed as the cluster's Centroid or center of gravity.

K-means algorithm works as follows. First, it randomly selects *k* of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects,

an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster.

K-means algorithm is simple, easily understandable and practically scalable, and can be easily modified to deal with streaming data. However, the drawbacks of k- means are the requirement for the number of clusters, *k*, to be specified before the algorithm is applied.

2.3.3 Prediction

Prediction calculates numeric and ordered future values based on a data set's patterns (Sharma, Anuj & Kumar Panigrahi, Prabin, 2012). It should be highlighted that the attribute for which the value being predicted is continuous-valued (ordered) rather than categorical is used for prediction (discrete-valued and unordered).). This property is referred to as the projected property (Sharma, Anuj & Kumar Panigrahi, Prabin, 2012).

2.4 Application of Data Mining

At present, the many advantages that DM brings are realized by various companies. And companies use data mining to manage all customer life cycle phases (acquiring new customers, increasing revenue from existing customers, and retaining good customers)
(Chung, Gray, & Mannino, 2005). In a wide variety of industries it provides a clear and competitive advantage by identifying potentially useful information from huge quantities of data collected and stored. It is currently mainly used by retail, financial, tax, communication and marketing companies with a strong consumer focus.

Data mining enables these companies to determine relationships among internal factors such as staff skills, product positioning, or price, and external factors such as customer demographics, competition, and economic indicators. Furthermore, it enables to determine the impact on sales, customer satisfaction, corporate profits by drilling down into summery of information (Famili & Turney, 1997)

2.4.1 Data mining in the Tax authority

As a general concept, Tax administration refers to the entire range of operations that a mandated government entity runs in order to implement and enforce the tax laws and

regulations. Tax administrations have varying mandates, structures and naming conventions across different countries.

A tax administration's core business is, generally speaking, to get the right tax at the right time from the right taxpayers, and to make the funds timely available for the right tax recipients (the state, municipalities, congregations, and others). Tax laws and regulations determine from whom, how much, and when, tax is due. The laws and regulations set forth certain registration, filing, reporting and payment obligations that the taxpayers must observe.

Tax laws and regulations are, however, not always simple and easy to comply with. On the other hand there are always citizens and organizations deliberately seeking ways to avoid or evade taxes.

Many fraud detection problems involve a large amount of information Says (Angell, 1938) (Castellón González, Velásquez, & Juan D., 2013) considering that Revenue Authorities tend to have vast amounts of data accrued over time from both active and inactive taxpayers. Considerable of this information is known only in terms of the amount of data it consumes within the tax administration storage systems. Most of the intelligence value of this data remains untapped within many revenue authorities and tax administrators as long as data mining techniques are not employed to mine the irregularities on this data (Atos, 2015).

Processing of these data in search of fraudulent transactions requires a statistical analysis which needs fast and efficient algorithms, among which data mining provides relevant techniques, facilitating data interpretation and helping to improve understanding of the processes behind the data. These techniques have facilitated the detection of tax evasion and irregular behavior in other areas such as banking, insurance, telecommunications, IT, money laundering, and in the medical and scientific fields, among others. Today, there are many fields and industries affected by this phenomenon.

In the operating context of a tax administration, Martikainen J. (Martikainen, 2012) states that two particularly prominent data mining uses are identified, and this is so: tax administrations can shape genuine workflows based on risk for the treatment of reports and payment transactions, registrations and filings, which are done or should be made by taxpayers. Each transaction is based on a comprehensive risk assessment based on data mining modeling so that all the relevant data available are used. Finally, risky transaction for

case-specific therapy can be flagged, while high risk companies can proceed to automated routine therapy.

Using data mining, Tax administrations can segment the taxpayers and identify segment specific compliance profiles in terms of diverse abilities and tendencies to comply. This supports tax administrations better design and target their services and submission actions.

Castellón P et al (Castellón González, Velásquez, & Juan D., 2013) reports that tax institutions started to use random selection audits to identify fraud beforehand or focused on those taxpayers who did not have previous audits in recent times and select cases based on auditors' experiences and knowledge..

Later methodologies were developed based on statistical analysis and construction financial or tax ratios (Castellón González, Velásquez, & Juan D., 2013) which evolved into the creation of rule-based systems and risk models (Castellón González, Velásquez, & Juan D., 2013). These transform tax information into indicators which permit ranking of taxpayers by compliance risk.

In recent years, the audit planning activities of various government organizations, with a main objective of detecting patterns of fraud or evasion, have been incorporated in the audit planning and have been used by the tax authorities for specific purposes.

2.4.2 Revenue and Tax Fraud Detection

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

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

As advised in the Fraud Control Guidelines, fraud against the Commonwealth is defined as dishonestly obtaining a benefit, or causing a loss, by deception or other means‟. A result of the Australian Institute of Criminology's is in 2007–08 Annual Reporting Questionnaire indicated that of the external fraud incidents, the focus of the highest number of activities was on entitlements. This group 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 (report, 2008).
Maximum entities that collect revenue or administer government payments conduct reviews across the various revenue and payment types. Based on previous experience, knowledge of their customers, and evidence from within their systems or from outside information, entities may undertake reviews that examine a recipient's circumstances where there is a perceived risk of fraud. The aim of such reviews is to detect a deliberate error, omission, misrepresentation or fraud on the part of a customer (Castellón González, Velásquez, & Juan D., 2013).

The opportunities to enhance tax administration and compliance functions through the use of data and advanced data Qualities 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 (report, 2008). As evasion and fraud schemes become more complex and pervasive, the need to leverage data and data analytics to 0unintentional is a critical core competency of tax administration.

Data Mining is an iterative process within which improvement is defined by discovery, either through manual or utomatic methods. DM is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an „„interesting‟‟‟ outcome (Kantardzic M. , 2002). The application of Data Mining techniques for financial classification is a fertile research area. Many law enforcement and special investigative units, whose mission is to identify fraudulent

activities, have also used Data mining successfully, however, as opposed to other well-examined fields like bankruptcy prediction or financial distress, research on the application of DM techniques for the purpose of management fraud detection has been rather minimal (Kirkos & Manolopoulos, 2007).

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

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

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

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

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

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

According to (Palshikar, 2002), as it does in many other spheres of our national and business life, data mining has many existing and potential applications in tax administration. For instance, predictive modeling will obviously assist the Nigerian FIRS to be in the best position to identify noncompliant taxpayers as well as ensure that tax auditing resources are channeled more appropriately on the accounts that will most likely yield the most desirable tax adjustments. There are many inherent advantages in this. First, the tax authority will be able to harness and manage its human resources more appropriately. Secondly, it will minimize the wastage of resources on compliant taxpayers. Thirdly, it will also enable the modification of the deployment of the hitherto traditional audit selection strategies for one that surely produces more efficacious outcomes.

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

Phua, et al. (Phua, C, Alahakoon, D. , & Lee, V., 2005) states that, it is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms. Evolved from numerous research communities, especially those from developed countries, the analytical engine within these solutions and software are driven by artificial immune systems, artificial intelligence, auditing, database, distributed and parallel computing, econometrics, expert systems, fuzzy logic, genetic algorithms, machine learning, neural

networks, pattern recognition, statistics, visualization and others. There are plenty of specialized fraud detection solutions and software which protect businesses such as credit card, e-commerce, insurance, retail, and telecommunications industries.

The Inland Revenue office of Australia (2009) stated that internationally, additional resources are being applied to address compliance issues similar to those highlighted above. For example the Australian Government will invest in excess of $600m over the next four years in the Australian Tax Office (ATO). This will be used to address key compliance areas such as the cash economy, abuse of tax havens, managing compliance risks to Australia's economic recovery and other public awareness campaigns. Additional investment would also be made in Inland Revenue's intelligence tools and processes, such as increased automated data-matching. This would increase revenue by ensuring that Inland Revenue can identify cases of non-compliance more quickly and accurately. This would allow us to intervene more quickly and appropriately. Inland Revenue would also invest additional funding in proactive compliance, which will increase revenue over time by encouraging taxpayer compliance through education, awareness and influencing social norms (Ashour, Amira S, Dey, Nilanjan, & Dac Nhuong, 2014) stated that selecting returns for Audits are like looking for a needle in a haystack. Every year, a large number of taxpayers fail to declare their tax liabilities correctly, and the Tax Administration is forced to tackle a tough task – to detect them (and enforce compliance from them) without increasing the compliance costs of the tax compliant taxpayers. It is not possible to identify the likely tax-evaders by simple query and reporting tools. Tax departments have access to enormous amounts of taxpayer data. However, it is impossible to ascertain the legitimacy of and intention behind a claim or a declaration in the tax return by simply looking at the return or a profile of a taxpayer. Given this reality, the best cost-effective option is to tease-out possible indications of fraudulent claims/declarations from the available data using data mining algorithms.

Based on (] Luciano, 2009). Explanation, once again, neural networks have been widely used. The main fraud detection software of the fraud solution unit of Nortel Network uses a combination of profiling and neural networks. Moreover, on the

continuous basis that made a research on user profiling by using neural network and probabilistic models and by other research which is call-based fraud detection by using a hierarchical regime switching model.

From the most influential research we can generalize that over time, the task of the tax administration with more opportunities for fraud will become even more difficult. The level of fraud is currently based on the financial report and income evasion.

Many approaches are taken to the detection of fraud. We emphasize the use of neural networks, Bayesian networks, expert system, regulatory systems and statistical outliers.

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

2.4.3 Local Related Works

Local research in the deferred data set and program for assessing DM application is being studied in the tax bureau. Danil.M (daniel, 2003) attempted clustering algorithm of the K-Mean algorithms used to identify the natural grouping of the various tax claims as fraud and non-fraud, followed by classification techniques for developing the predictive model, has shown the highest rating accuracy of 99.98 per cent by the j48 decision algorithm

In addition, Leul (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.

Further case studies concerning the use of DM in the various sectors were also conducted. For example, Tariku (Tariku, 2011) has tried, for Africa Insurance Company, to develop models to recognize and predict insured claims fraud. He attempted to apply the algorithm of the clustering to the development of a predictive model. K-means grouping algorithm is used to identify natural groupings as fraud and non-fraud the various insurance claims. Tilahun (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 (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. In addition, in the 2001 Ethiopian child labor survey, Helen (daniel, 2003) has also been trying to explore the use of DM technology to identify significant patterns in census or survey data. In order to identify relationships among attributes within the child labor study database for 2001 she used to clearly understand the nature of the child labor problem in Ethiopia, she applied the association rules DM technique and Apriori algorithm. The expectation maximizing-clustering algorithm has been used to classify the final selected datasets, aside from the Association rule technique.

Most researchers are used clustering with classification to detect fraud this is their drawback since the data are structured that means no prior grouping is used then using classification provides best result than clustering.

The application of DM in various sectors such as aviation, banking, health care and customs is examined in local research. All researchers are mainly interested in investigating the applicability of data mining in the sectors mentioned above. Most researchers used k-means and decision tree algorithms for classification and clustering. Moreover, most research is carried out for a certain area of the domain

Similarly, this study will examine the applicability of DM in AMRA using the proposed DM techniques. The main purpose of this research study is to use data mining to create a

predictive model for establishing suspicious tax liabilities for fraud and non-fraud for the purpose of Amhara Revenue Authority development of effective tax collection. For this reason, the authority must use data mining technologies to protect fraud and improve loyalty to perform the task of collecting tax.

2.5 Summary

A complete overview of the background theory, concepts and technologies has been provided in this chapter. Examples of application areas and related works indicating where the fraud detection in different fields has taken place, and how the classification and classification techniques with k-means and decision tree algorithms are used.

CHAPTER THREE

METHODOLOGY

The purpose to specify methodology is to provide insight with fundamental knowledge of data modeling and design; the tools and techniques of data analysis using data mining technology; to inform beneficiaries with data mining concepts, and techniques; and to prepare data for further analysis in the database.

3.1. Research Design

The study employed a hybrid data mining process model approach to build predictive models using data mining techniques by applying a set of classifier algorithms on the revenue datasets. In this study, hybrid approach is built by combining knowledge discovery in databases (KDD) and cross industry standard process for data mining (CRISP-DM) process models. This method was selected for this specific study because of various reasons.

- It describes the steps more generally and in a research way.
- It underlines on the consistent aspects of the process, drawing experiences from the previous models, and
- It supports both academia and industrial data mining task standards.

In the selected data mining approach six main activities are considered for the development of the required model developed by Cios et al. (2007) is considered because this model combines both academics and industry aspects. These include, understanding the problem domain, understanding the data, preparing the data for mining, mining the prepared data, evaluating or testing the discovered knowledge on new datasets and using or applying the discovered knowledge for decision making.

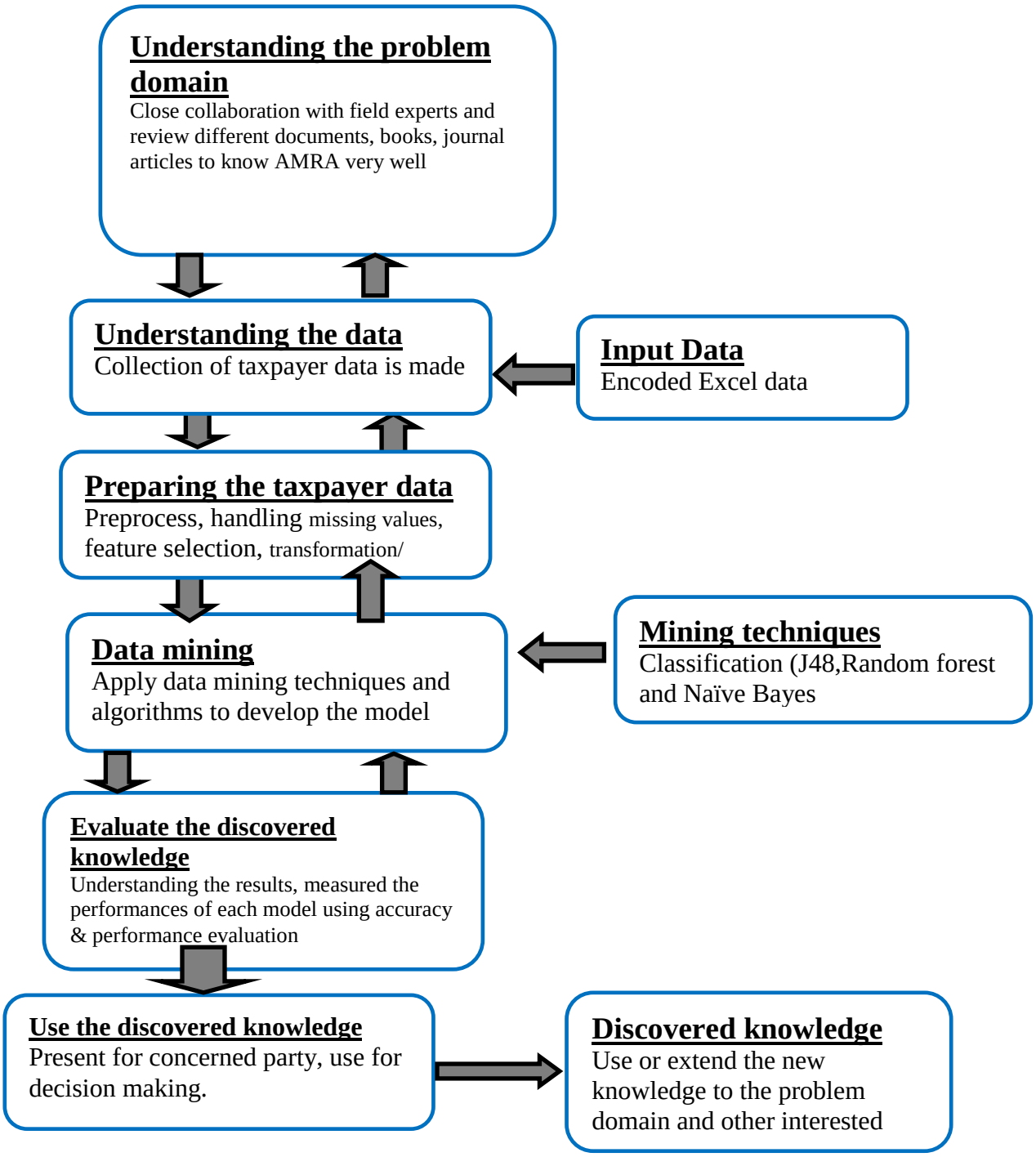


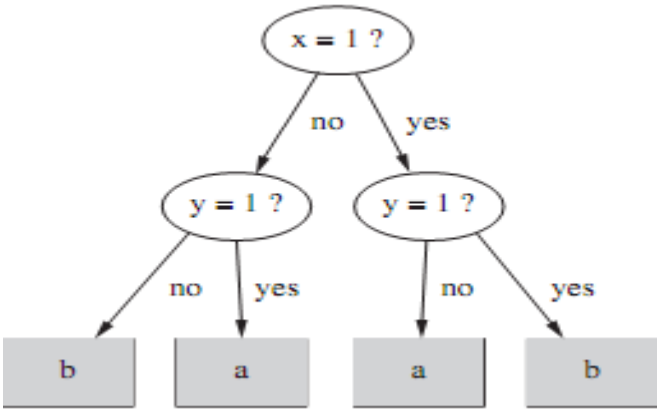
Figure 3 the Hybrid Models (description of the six steps follows)

3.2. Data mining techniques

3.2.1 Decision Tree Classification Technique

The problem of learning from a range of independent instances is naturally brought about by a "divide – and conquer" approach to a style of representation called decision tree. Nodes in a decision tree involve the testing of a specific feature. (Witten, H. Ian & Frank, E, 2005). In general, the node test compares a value with a constant attribute. However, some trees compare two attributes, or use some attribute function. Leaf nodes provide a classification applicable to all instances of a leaf or a series of classifications or a distribution of probabilities over all possible classifications. In order to classify the leaf as an unknown instance, the tree is routed down to values of the attributes tested in subsequent nodes, and the instance is classified according to the class of the leaf.

.



In all forms, trees can grow. These might be non-uniformly large binary trees, that is, every node has two children and the distance to the root of a leaf varies. The figure shows that each node is a "Yes" and "No," the answer determines which of the two paths a record leads to the next tree level.

Tree induction

Tree induction is the process of learning decision trees from class-labeled training tuples A decision tree is a tree structure that looks like a flow chart, with each internal node (non-leaf node) representing a test on an attribute, each branch representing the

test's outcome, and each leaf node (or terminal node) holding a class label. The root node is the topmost node in a tree. (Han, J. & Kamber, M, 2006).

Splitting: The tree-growing phase is an iterative process involving the division of the data into smaller sub-sets. The first is the root node containing all the data.
Subsequent iterations work on derivative nodes that will contain subset of the data. One important characteristics of splitting is that it is greedy, which means that the algorithm does not look forward in the tree to see if another decision would produce a better overall result.

Stopping criteria: Algorithms for tree building usually have a couple of rules to stop. These rules are normally based on several factors, including the maximum tree depth, the minimum number of node elements that must be in a new node, or are near equivalent. The user can change the parameters of these rules in most implementations. In fact, certain algorithms start to build up a tree as deep as possible. Whilst such a tree can predicate exactly all instances in the training set (with the exception of conflicting records), it more than likely fits into the data. This tree is a problem.
Pruning: When a decision tree is built many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over fitting the data. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data.

Different decision tree algorithms

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

Target variables: For most tree algorithms, the target or dependent variable is categorical. Such algorithms require the use of binned (grouped) continuous variables for regression purposes.

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

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

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

The decision tree classification was chosen for the following reasons:

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

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

3.2.1.1 The J48 Decision Tree Algorithm

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

J48 is the decision tree algorithm used in this study to classify tax claims as fraudulent or non-fraudlent suspicious. The J48 decision tree can serve as a model for

classification as it generates simpler rules and remove irrelevant attributes at a stage prior to tree induction. In several cases, it was seen that J48 decision trees had a higher accuracy than other algorithms (Witten, H. Ian, & Frank, E, 2005) J48 offer also a fast and powerful way to express structures in data.

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

J48 employs two pruning methods (Witten, H. Ian & Frank, E, 2005). the first one is Sub-tree replacement. This means that nodes in a decision tree may be substituted with a leaf basically along a certain path to reducing the number of tests. This procedure begins from the leaves of the fully formed tree, and works backwards toward the root. In this case, a node may be moved upwards towards the root of the tree, substituting other nodes along the way. Sub-tree rising often has a unimportant effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that sub-tree rising can be somewhat computationally complex.

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

To determine the specificity of the model there are several options. One powerful option is the minimum number of instances per leaf. This allows us to dictate the lowest number of instances that can constitute a leaf. The higher the number of

instances the more general the tree is. Lowering the number will produce more specific trees, as the leaves become more granular.

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

a) Choose an attribute that best differentiates the output attribute values.
b) Create a separate tree branch for each value of the chosen attribute.
c) Divide the instances into subgroups so as to reflect the attribute values of the chosen node.
d) For each subgroup, terminate the attribute selection process if:
    1) All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with specified value.
    2) The subgroup contains a single node or no further distinguishing attributes can be determined. As in (1), label the branch with the output value seen by the majority of remaining instances.
e) For each subgroup created in (c) that has not been labeled as terminal, repeated the above process. The algorithm is applied to the training data. The created decision tree is tested on a test dataset, if available. If test data is not available, J48 performs a cross-validation using the training data itself.

3.2.2 Naïve Bayes Classification Technique

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class (Han, J. & Kamber, M, 2006).

The Bayesian classification method is based on the Bayes theorem. Comparing classification algorithms has revealed that a simple Bayesian classifier known as the naive Bayesian classifier can outperform decision tree and neural network classifiers. When applied to large databases, Bayesian classifiers have also demonstrated high accuracy and speed.

Observations show that Naïve Bayes is consistent with a reduction in number of attributes before and after (Soni, Jyoti, Ansari, Uzma, & Sharma, Dipesh, 2011). Naïve Bayesian classificators assume that the effect on that class is independent of the values of the other attributes of an attribute value. The term class conditional autonomy is this assumption. Naïve Bayes analyzes the relationship between each input and the dependent characteristics, so that each relationship has the probability of being conditional. (Han, J. & Kamber, M, 2006)

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

Naïve Bayes works very well when tested on many real world datasets (Witten, H. Ian & Frank, E, 2005). By theory, this classifier has minimum error rate but it may not be the case always (Soni, Jyoti, Ansari, Uzma, & Sharma, Dipesh, 2011). However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. An advantage of Naïve Bayes algorithm over some other algorithms is that it requires only one pass through the training set to generate a classification model. In addition, Naïve Bayes can also obtain results that are much

better than other sophisticated algorithms. However, if a particular attribute value does not occur in the training set in conjunction with every class value, then Naïve Bayes may not perform very well. It can also perform poorly on some datasets because attributes were treated as though they are independent, whereas in reality they are correlated.

The Naïve Bayes Algorithm is discussed below

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

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

1. Let D be a training set of tuples and their associated class labels. Each tuple is represented by an n-dimensional attribute vector, $X = (x_1 , x_2,\ldots x_n)$, depicting **n** measurements made on the tuple from **n** attributes, respectively, $A_1, A_{2,\ldots} A_n$.

2. Suppose that there are **m** classes, $C_1, C_2,\ldots\ldots C_m$. Given a tuple, **X**, the classifier will predict that

   X belongs to the class having the highest posterior probability, conditioned on **X**. That is, the

   Naïve Bayesian classifier predicts that tuple **X** belongs to the class $\mathbf{C_i}$ if and only if

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

   Thus we maximize $P(C_i |X)$. The class C_i for which $P(C_i |X)$ is maximized is called the maximum

posteriori hypothesis. By Bayes" theorem $P(C_i |X) = P(X|C_i)P(C_i) P(X)$

3. As P(X) is constant for all classes, only $P(X| C_i)P(\bar{C}_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \ldots = P(C_m)$, and we

would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = |C_i,D|/|D|$, where $|C_i,D|$ is the number of training tuples of class C_i in D.

4. Given datasets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the Naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that are no dependence relationships among the attributes). Thus,

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

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

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

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

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

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

3.2.3 Random forest Algorithm

It works out and is used for unbalanced information. The prediction is a vote on the classification by aggregate majority and the regression average. Many studies say the

classifier shows a significant improvement in the performance over single tree like CATR and C4.5.This is the definite algorithm process (Kanakalaksmil, 2019)

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: Each record vote from the decision tree results shall determine the final grading. Usually, the techniques and the algorithms that are discussed before are used to conduct the experimentations of this research for developing the model used for predicting fraudulent and non-fraudulent tax claims.

3.3. Business Understanding

3.3.1 AMRA tax vision and data

AMRA is the body responsible for collecting revenue from domestic taxes in Amhara region. In addition to raising revenue, AMRA is responsible to protect the society from adverse effects of smuggling. It seizes and takes legal action on the people and vehicles involved in the act of smuggling while it facilitates the legitimate movement of goods and people in the region. AMRA is established on July 1, 2000 E.C as a result of the merger of the Ministry of Revenues, the Ethiopian Customs Authority and the Federal Inland Revenues into one giant organization ((AMRA), 2019).

According to article 3 of the proclamation No. 74/2002, the Authority is looked upon as "an autonomous regional agency having its own legal personality". The Authority came into existence on July 1, 2000, by the merger of the Ministry of Revenue, Amhara Revenue Authority who formerly were responsible to raise revenue for the Regional government and to prevent contraband. Reasons for the merge of the foregoing administrations into a single autonomous Authority are varied and complex.

Some of those reasons include:

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

AMRA has its headquarters in Bahirdar. It is led by a Director General who reports to the president and is assisted by two Deputy Director Generals, namely D/Director General for Program Designing of Operation and Development Businesses; D/Director General of Enforcement Division; Change Management and Support Sector; and Enforcement Sector. Each deputy director general oversees at least four directorates. Both the Director General and the Deputies are appointed by the president ((AMRA), 2019).

AMRA categorizes taxes as either indirect or direct.

Indirect taxes are: -

- Turnover
- VAT
- Excise

Direct taxes are: -

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

The table below describes direct and indirect taxes.

Indirect Taxes	**Explanation**
VAT(value added tax)	As the name indicates It has 15% value which is added during buying or selling of goods

TOT (turn over tax)	It has 10% or 2% value which is added during buying or selling of goods or providing services
EXCISE TAX	It is an inland tax on the sale, or production for sale within a country
Direct Taxes	**Explanation**
WT (withholding tax)	It is deducted by government agents during buying goods and the final tax is reported to tax office
PIT (personal income tax)	Employment income or other private or non-governmental organization.
RT (rental tax)	A tax which is levied on revenue from building rental.
BPT (business profit tax)	A tax is imposed on commercial, professional or business activities or any activities recognized as commercial by Ethiopia's Commercial Code.

Table 1 Direct and indirect taxes

The following functions identify AMRA's key business drivers .

- ✓ Security and enforcement.
- ✓ Reliable statistics and data.
- ✓ Process oriented management
- ✓ Revenue collection.
- ✓ Facilitation of trading.
- ✓ exemplary governance

The main offices in Bahirdar comprise 182 Tax Centers. A Tax Center is a tax collection station that is administered by a branch office and is located near taxpayers. Although the above-mentioned six tasks are the key businesses of AMRA, this research focused on revenue collection. The computer system that allows AMRA to administer the aforementioned taxes is known as the Standard Integrated Government Tax Administration System (SIGTAS). The system enables AMRA to manage all aspects of most domestic taxes, such as registration, assessment, cashing, and auditing, in a single, simple-to-use integrated system. The system was implemented in Amra in December 2004 and is now operational at both the head office and branch offices. One of the

authority's primary functions is to audit the financial statements of its customers. The audit process and program development directorate collaborate closely with the Information Technology Management Directorate.

.

3.4. Data Understanding

The data itself is a prerequisite for any DM. The corporate data warehouse has been identified as a good source of data for DM purposes (Berry, M & Linoff, G, 1997). This is because the data is stored in a standard format with consistent definitions for fields and their values. To meet this requirement, raw data on tax payers was collected from AMRA's Information Communication Technology Directorate (ITMD) database, and careful analysis of the data and its structure was performed in collaboration with business (domain) experts, evaluating the data's relationship to the problem at hand and the specific DM tasks to be performed. The nature and organization of the data collected are described in the following sections.

3.4.1 Initial Data Collection

The central AMRA database in Bahir Dar was selected for data collection. All data of the top contributors are in the AMRA ICTD Task Process Owners database. Suburban tax offices levy taxes on small businesses. The tax offices of the lower towns are networked with the database of the owner of the AMRA ITMD task process, which is located around the medrok square.

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

The number of records collected from the ICTD database is summarized in Table 2

TAX CENTER (BRANCH)	Taxpayers category	Number of records collected	Total
Bahirdar city	A	1000	6000
Gonder city		1000	
Dessie city		1000	

Other tax centers		3000	
Bahirdar city	B	1000	6000
Gonder city		1000	
Dessie city		1000	
Other tax centers		3000	

Table 2 collected data for category A and B taxpayer

3.4.2 Description of Data Collected

As previously stated, the relevant data for the research is gathered from the AMRA ITMD database. These are TIN , Name ,sex Registration date , Total sales , Number of employees , Auditing time , Tax audit acceptance, Penalty amount , Total expenses, Net tax due , Number of refunds in a year , Vat refund amount , Vat reports, Gross profit, Net worth, Use of cash registration machine and Category. Though, the table has lots of attributes in the original dataset, the table below show initially selected attributes.

number	Name of Attribute	Data Type	Explanation
1	Registration date	Number	Registration day for a taxpayer.
2	Total sales	Number	Total amount of sales in birr.
3	Number of employees	Number	Total number of employees in accompany reach earnings before tax.
4	Auditing time	Number	Any Assessment linked to an audit case. Look at the most recent one.
5	Tax audit acceptance	Number	The most recent ‘Closed’ case is looked for
6	Penalty amount	Number	Number of penalties apply on tax payer for a year
7	Total expenses	Number	Expenses that are incurred for running business
8	Net tax due	Number	Amount of tax expected from a taxpayer
9	Number of refunds in a year	Number	Indicates the number of refunds asking for a year

10	Vat refund amount	Number	The amount of vat asked for refund
11	Vat reports	Number	The number of vat reports declared
12	Gross profit	Number	Total profit of business
13	Net worth	Number	Shows net resource of the company at hand.
14	Use of cash registration machine	Number	Proper use of cash registration machine
15	Loss carry forwarded	Number	Loss came from previous years
16	Loss carry balance adjusted	Number	Adjusted loss balance

Table 3 attribute description

3.5. Preparation of the Data

Data preprocessing can be time consuming, even though data mining is an important step in the knowledge discovery process. The preprocessing stage's goal is to clean up the data as much as possible and convert it into a format that can be used in the next steps. Before being used, the data extracted from AMRA's source database undergoes a number of transformations.

3.5.1 Data Selection

The DM task may not use the entire target dataset. Before beginning the actual DM function, irrelevant or unnecessary data are removed from the DM database. Initially, there were approximately 12000 records. Around 1000 of these records are chosen at random for testing purposes. The remainder of the dataset is used for training. To conduct this study, a total of 12000 datasets were used after removing irrelevant and unnecessary data.

Table 3 of the ICTD database contains the aforementioned 16 attributes. The first task was to remove from the database any fields or attributes that were unrelated to the current task. The original sets of attributes are listed below, and they are further preprocessed to select the final attributes used in the research. Number of employee, registration date, total sale ,auditing time ,tax audit acceptance penalty amount, total expense, net tax due, number of refunds in a year, vat refund amount, number of vat reports, gross profit, net worth ,use of cash registration machine were the initial set of attributes that are nominated.

3.5.2 Data Cleaning

Data cleaning is the process of filling in missing values, removing noise, and correcting inconsistencies in data. In most cases, real-world databases contain incomplete, noisy, and inconsistent data, which can confuse the data mining process (J.Han & M.Kamber, 2004). As a result, data cleaning has become a necessity in order to improve the quality of data in order to improve the accuracy and efficiency of data mining techniques.

This technique involves the removal from each column of incomplete, noise (invalid) data records. Data were removed because records of this kind are few and deletion of them will not affect the whole dataset.

The MS-Excel 2016 functionality is used to identify and fill missing values, for example by searching, replacement, filtering, auto-filling, and WEKA.

3.5.2.1 Missing Value Handling

Missing values mean the values in a data that does not exist for one or more attributes. Data are seldom complete in real world applications. It can also be especially pernicious. Especially if the dataset is a small number or the number of missing fields is large, the sample cannot remove all records with a missing field.

Furthermore, the absence of a value may be significant in and of itself. To calculate a substitute value for missing fields, such as the median or mean of a variable, a widely

used approach (Halkidi, M & Vazirgiannis, M, 2001) and searching and replaced by TIN is used manually. Accordingly, the researcher has been analyzed the taxpayers‟ dataset and identified missing values and take measure to solve the problem as follows. The total records are 12000 and the missing values are 3180. From the total amount the missing data contains 26.5%. Net worth, Vat refunds, Vat reports and Use of cash registration machine attribute which of them accounts of 10%,10%, 1.5% and 8% respectively. To handle the problem of missing values of numerical data type attributes were recommended to be replaced by the mean of that value a whole (Chung, , H.M.; Gray, , P.; Mannino, M., 2005) and manually search and replace method is used.

As a result, the researcher handles missing values based on the above principles, and WEKA preprocessing replace missing value techniques are also used. WEKA fills using the most common (model) value methods, which follows the same principle as the previous principle. Another technique used by the researcher is manually tracing and correcting the missing value.

As shown in table 4, four of the selected 14 attributes have missing values. As a result, the researcher reacts by taking the necessary steps to clean up the data.

No	**name of Attribute**	**% of missing value**	**Reason/Techniq ue applied**	**Data types**
1	Net worth	10.12	The mean of this attribute	Numeric
2	Vat refunds	10.25	The mean of this attribute	Numeric
3	Vat reports	1.5	searching and replace by TIN	Numeric

4	Use of cash registration	8	The mean of this attribute	Numeric

Table 4 missing value Handling

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

3.5.2.2 Handling Outlier Value

The data stored in a database can reflect a random variable measurement of noise, exceptional case, or incomplete data object. These erroneous attribute values may be due to data entry problems, incorrect data collection, and inconsistency in convention naming or technology limits (Gorunescu, 2011). Four basic methods to handle noise data have been explained by the authors. The method of binning, regression and combined computer and human inspection. These are the binning method.

The researcher identified and found the noise or surplus value from taxpayers' information. In this research. The identified outlier was manually corrected with the help of domain experts. The researchers and the domain experts have therefore combined to identify and correct the problem of incompleteness, noise or excess.

3.5.3 Data Construction

The other important step in preprocessing is to draw from the existing areas other fields. The addition of fields representing the relationships in data will probably be important to increase the chance of a useful result from the knowledge discovery process (Berry, M & Linoff, G, 1997). The following fields considered crucial for determining the fraudulence of claims have been derived from the existing fields in consultation with the domain experts at AMRA.

No.	Name of attribute	Data Type	Explanation
1	Registration date	Number	Registration day for a taxpayer.
2	Total sales	Number	Total amount of sales in birr.
3	Number of employees	Number	Total number of employees in accompany
4	Auditing time	Number	Any Assessment linked to an audit case. Look at the most recent one.
5	Tax audit acceptance	Number	The most recent 'Closed' case is looked for
6	Penalty amount	Number	Number of penalties apply on tax payer for a year
7	Total expenses	Number	Expenses that are incurred for running business
8	Net tax due	Number	Amount of tax expected from a taxpayer
9	Number of refunds in a year	Number	Indicates the number of refunds asking for a year
10	Vat refund amount	Number	The amount of vat asked for refund
11	Vat reports	Number	The number of vat reports declared in a year
12	Gross profit	Number	Total profit of a taxpayer before tax
13	Net worth	Number	Shows net resource of the company at hand.
14	Use of registration machine	Number	defines the existing price of the goods.
15	Risk level	nominal	Dependent attribute that shows the final class categories of the result

Table 5 Final List of Attributes used in the research

3.5.4 Data Integration

Before deriving the attributes, the data integration process was completed. As previously stated, the dataset for the database discussed above was available in multiple excel files. In order to prepare the data for the DM techniques to be used in this study, a data integration method for retrieving important fields from different files was used. The data integration process was carried out using the Oracle and SIGTAS databases. The research took a long time because of the data integration process. This

was due to the fact that when the various excel files were combined, the dataset's size increased by a factor of ten. This indicates that the records were being duplicated extremely. Table 5 shows fifteen attributes from the integrated files, including Net Worth, Liquid Cash, and registration date. Finally, all of the information is combined into a single Excel file.

3.5.5 Data Formatting

WEKA, like any other piece of software, requires data in a variety of formats and file types. The data provided to this software was formatted in a way that WEKA software can understand. WEKA accepts records with comma-separated attribute values saved in CSV format (comma separated value).

Initially, the integrated dataset was in the form of an excel spreadsheet. The file is converted to another file format in order to feed the final dataset into the WEKA DM software. First, the excel file was converted to a comma delimited (CSV) format. The next step was to open the file with the WEKA DM software after converting the dataset to CSV format.

3.5.6 Data Transformation and Concept Hierarchy

3.5.6.1 Data transformation

According to Han J and Kamber M. (2006) in data transformation; data are transformed or consolidated into forms appropriate for mining process. It involves smoothing, aggregation, generalization, normalization, discretization, and attributes construction. To make the dataset appropriate for this study data discretization technique is applied on numeric attributes to minimize distinct values of attributes, dimensionality reduction is also used to reduce the size of the dataset and attribute selection method is the last method applied to remove weakly relevant attributes.

3.5.6.2 Data Discretization

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute in to intervals (Han, J. & Kamber, M, 2006) 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 there by reduces and simplifies the original data. This leads to a brief, easy to use, knowledge-level representation of mining results.

Discretization is the process of converting continuous valued variables to discrete values where limited numbers of labels are used to represent the original variables. The discrete values can have a limited number of intervals in a continuous spectrum, whereas continuous values can be infinitely many (Chung, , H.M.; Gray, , P.; Mannino, M., 2005). In this research, registration date, total sales, total expenses, and Vat refund amount attribute is re-binned into a new nominal attribute.

Discrediting the values of Use of cash registration attribute: The use of cash registration machine measures providing cash for customers; it describes current machine usage status. Final use of cash registration attribute discretized categories are 0 (Low) no penalty is given during usage, >1<2 penalty is given during usage (Medium) and >2 a above two penalty is given during usage(high).

Discrediting the values of total sales attribute: The values are 0 and l is less than 50,000 birr, 2 (medium) is equal to b/n 500,000 birr and 3(high)is equal to total sales above 500,000 birr. This is done by defining a portion of the values through explicit data grouping as presented in (table 3.1).

The rest discretized attributes are discussed on the following table 3.1

Table3.1: Data Discretization

Attribute name	Given criteria	Old value	New/replaced value

Registration date	<1 year	0	Low
	<2 year	1	
	>2<= 6year	2	Medium
	>6 year	3	High
Total sales	<25.000 birr	0	Low
	50.000 birr	1	
	500.000 birr	2	Medium
	>500,000 birr	3	High
Number of employees	**<4**	<4	Low
	>4<=14	>4<=14	Medium
	>14	>14	High
Auditing time	<1 year	0	Low
	<2 year	1	Medium
	>3 year	2	High
Penalty amount	**0**	1	Low
	< 50000	2	Medium
	>50000	3	High
Total expenses	0	Very low	Low
	<=10000	low	
	>10000 and less than 500000	medium	Medium
	>500000	High	High
Net tax due after audit	**0**	0	Low
	<=15%	<=15%	Medium
	>15%	>15%	High
Number of refunds in a year	**0**	0	Low
	>1and<=3	1	Medium
	>3	1	High
Vat refund amount	0	low	low
	<50000	medium	Medium
	>50000and<500000	high	High

	>500000	Very High	
Number of vat reports	**12**	0	Low
	>9<=11	1	medium
	<9	2	High
Net worth	**<5%**	0	Low
	>5%<=20%	1	medium
	>20%	2	High
Gross profit	**<100000**	1	low
	>100000<=200000	2	high
	>20000	3	Medium
Tax audit acceptance	1 time audited and accepted	0	Low
	2 times audited and accepted	1	Medium
	3 and above times audited and accepted	2	High

3.5.7 Target/Class attributes

Some data mining techniques need predefined classes in order to train and build classification models. In such cases, the data preparation task is unfinished until the target attributes are well-defined. In other words, the training set should be pre-classified so that the data mining algorithms know what the user is looking for. With respect to target attribute, as the purpose of this research is to identify the most influential factors of fraud among category A and B taxpayers, the target attribute selected in this study is risk level. This attribute primarily has three different values such as low risk, medium risk, and high risk. However, to make the interpretation of the final result much easier, like other independent variables, the original risk level is re-binned into a new structure of nominal attribute with values of fraud suspected (medium risk and high risk) and not fraud suspected. These new nominal attributes have been changed by using the same definitions of the 2005 AMRA documentation files.

In general, the target attribute used in this research work is risk level that has two values, namely, fraud suspected and not fraud suspected. This attribute is considered as

dependent variable whereas the rest of the attributes specified in Table 5 are the independent attributes for this research work.

3.6. Selecting of Models/Technique

This step involves usage of the planned data mining techniques, tools and selection of the new ones. Data mining tools include many types of algorithms, preprocessing techniques, and data mining elements. In the study, a classification technique is used to develop the model capable of answering the stated problems. The training and testing procedures are designed and the data model is constructed using the chosen data mining tools. The researcher used J48 decision tree, Naïve Bayes and Random Forest classification algorithms, data preprocessing and model development tools such as, MS excel, WEKA 3.9.3 and also documentation tools such as MS-Office packages.

This research is conducted by adopting three data mining classification algorithms, which includes Decision Tree (DT) with J48 classifier to generate rules sets, Random forest and Naïve Bayes algorithm to predict class membership probabilities were implemented.
Classification is one of the major data mining tasks. So, this task is accomplished by generating a predictive model of data and interpreting the model regularly to provide information for selective labeled classes in data. Each algorithm used in this study is described in the previous subsections. These algorithms were selected due to their popularity and usefulness in solving data mining classification problems. Moreover, their advantages such as easy to interpret/ understand and visualize the result, fast at classifying unknown records/new instances, and easily handles all types of attributes and missing values, implement and use and numeric attributes compared to other classification techniques are taken as other reason for selecting these classification techniques.

3.6.1 Validation techniques (Test Options)

All the identified algorithms experiment was tested with the option of using 10-fold cross-validation (the classifier evaluated using the number of folds that are entered in the folds text field). The default 10-fold cross validation was chosen and used for building

the model and testing the model performance. Due to its ability to carry out widespread tests on many datasets with different learning techniques, the researcher decided to use 10-fold test option, and 10 are the correct number of folds for best estimation of errors, as well as theoretical evidence to support that process.

The entire dataset is randomly split into 10 mutually exclusive subsets of approximately equal size in 10-fold cross validation. It is trained on nine folds every time and tested on the rest of the single fold. This testing technique was carried out to minimize the bias associated with the random sampling of the samples by repeating experiments ten times.

3.7. Evaluation of the discovered knowledge

The confusion matrix, which contains values of true positives (correct classifications) and false positives (incorrect classifications), was used to evaluate the various classification models

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

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 have been incorrectly classified as legitimate but are actually malicious.

To calculate the above performance measures such as accuracy, TP, FN, FP, TN, Precision, Recall, F-measure and ROC area the confusion matrix results are applied. For example, the following are the general formulas of each performance measures.

Accuracy: To gain the accuracy of a classifier is by dividing the total correctly classified positives and negatives instance by the total number of samples. These measures are defined as

Accuracy = $\frac{TP+TN}{TP+FN+FP+TN}$

TP (sensitivity): is defined as its ability to correctly identify actual cases, i.e. for this study. To classify the positive cases correctly is sensitivity.

Sensitivity =$\frac{TP}{TP+FN}$

TN (specificity): Specificity is the true negative rate is defined as ability to correctly identify negative cases which are correctly identified. These measures can be computed as: Specificity = $\frac{TN}{TN+FP}$

Precision: Can be thought of as a measure of exactness (i.e, what percentage of tuples as positive is actually positive.

Precision= Positive Predictive value= $\frac{TP}{TP+FP}$

Recall: Is a measure of completeness (how many positive tuples are labeled as positive). It is synonymous with sensitivity (or true positive rate). These metrics are defined as follows:

Recall = Sensitivity= TP Rate= $\frac{TP}{TP+FN}$

F-Measure: it is the inverse relationship between precision and recall, and calculated as harmonic mean between precision and recall.

F-measure= $\frac{2(Precision*Recall}{Precision+Recall}$

ROC curve: For different portions of the test set, the ROC curve allows us to visualize the trade-off between the true-positive (TP) rate and the false-positive (FP) rate at which the model can accurately recognize positive cases versus the rate at which it incorrectly identifies negative cases as positive. ROC curve for a given model shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) (H. Singh & S. Bhagat,, 2015) It is done by drawing a curve in two-dimensional space with vertical and horizontal axes, with the vertical axis representing true-positive rate and the horizontal axis representing false-positive rate. The area under the curve, which is a portion of the area of the unit square and has a value ranging from 0-1, can be used to assess the accuracy of a model.

The area of the model with perfect accuracy will be 1.0, indicating that the larger the area, the better the model's performance or the larger the values of the test result variable, the stronger the evidence for a positive actual state (1.00) (Castellón González, Velásquez, & Juan D., 2013).

CHAPTER FOUR

RESULTS AND DISCUSSION

The experimental activities and results are presented in this section. Labeled records are used in this study. The data is in an Excel spreadsheet from the AMRA data base. It has been preprocessed and is ready for the experiments described in the previous chapter. The main goal of this study is to find regularities in the taxpayers' dataset that can be used to predict and detect fraud suspicious and non-suspicious tax audit claims. The model building phase of this investigation's DM process is carried out using supervised classification techniques, which are implemented using the WEKA DM tool.

Using classification techniques such as J48 Decision Tree, Random Forest, and Nave Bayesian classification, which are widely applicable in solving the current problem and to create a predictive model.

4.1 Experiment Design

Experiments are carried out on a computer with the following specifications: Intel(R) Core(TM) 5 CPU 2.5GHz, 6 GB RAM, and Microsoft Windows 10 as the operating system platform. Filtering records is done with WEKA (3.9.3) and Microsoft Excel (2016). The steps used in this study experimentation are as follows:

- The researcher chose data mining software to conduct the experiment.
- The final records are converted from Microsoft Excel to CSV format, which is compatible with WEKA data mining software.
- The preprocessing tasks are completed using Microsoft Excel to resolve missing values (see chapter three section preprocessing) (incomplete data).
- The experiments are carried out using the WEKA data mining tool at this stage. For this study, the training models with the highest classification accuracy were chosen.
- Finally, the chosen model creates a rule for the chosen best model's deployment and future work.

4.2 Classification Modeling

The decision tree (in particular the J48 algorithm, Random Forest) and nave Bayes methods are used to conduct the classification modeling experiments. The model is trained by changing the default parameter values of the algorithms in order to classify the records based on their values for the given cluster index.

The training of the decisions tree classification models of the experimentation is done by employing the 10-fold cross validation and the percentage split classification models. The classification is analyzed to measure the accuracy of the classifiers in categorizing the tax claims into specified classes. When compared to actual classifications, accuracy refers to the percentage of correct predictions made by the model. (Baeza-Yates, R & Ribeiro-Neto, B., 1999). Each of these models' classification accuracy is reported, and their performance in classifying new instances of records is compared.

4.2.1 Model Building using J48 Decision Tree

A decision tree classifier is one of the most commonly used supervised learning methods for data exploration, approximating a function with piecewise constant regions and requiring no prior data distribution knowledge. In data mining, decision trees are frequently used to examine data and create the tree and rules that will be used to make predictions.

The J48 algorithm has a few parameters that can be tweaked to improve classification accuracy even more. The J48 algorithm's default parameter values are used to build the classification model at first.

The Experiment for J48 Decision Tree Modeling are described below

The decision tree model is constructed using the J48 algorithm. The experimentation's decision tree classification models are trained using a combination of 10-fold cross validation and percentage split classification models.

Experiment number	Total records	Number of attributes	No of leaves	Size of the tree	Test modes used	Time to construct a model (seconds)	Result(Accuracy)(%)
1	12000	14	109	163	10-fold cross-validation	0.11 second	99%
2	12000	14	109	163	Percentage split 66%	0.02 second	98.89%

Table 6 decision tree output with different test modes for J48 algorithm

In table 6 the result of j48 shows that the accuracy of 10-fold cross validation and percentage split (66/34%),are 99% and 98.89% respectively. The confusion matrix of all training set is:-

=== Confusion Matrix ===

```
   a    b   <-- classified as
6484   66 |    a = Fraud suspected
  54 5396 |    b = not fraud suspected
```

As shown in the resulting confusion matrix, the J48 learning algorithm tested using 10-fold cross validation scored an accuracy of 99%. This result shows that in the above table different result with the same parameter and with deferent testing mode.

4.2.2 Model Building using Naïve Bayes Algorithm

The Nave Bayes is the second DM technique used in the classification subphase. The WEKA software package is used to construct the Nave Bayes model, which employs the Nave Bayes Simple algorithm. The percentage split test option is used in conjunction with the 10-fold cross validation option, which is set by default.

The Experiment for Naïve Bayes Model is described below

The Bayes Theorem, which derives the probability of a prediction from the underlying evidence, is used by Nave Bayes to make predictions. The experimentation's Nave Bayes classification models are trained using a combination of 10-fold cross validation and percentage split classification models.

Experiment number	Total records	Number of attributes	Test modes used	Time to construct a model (seconds)	Results(accuracy) (%)
3	12000	14	10-fold cross-validation	0.02	97.04%
4	12000	14	Percentage split 66%	0.02 second	97.5 %

summarized output of the Naïve Bayes Simple algorithm

Table 7 summarized output of the Naïve Bayes Simple algorithm

Table 7 shows that the accuracy of 10-fold cross validation and percentage split (66/34 percent) is 97.04 percent and 97.5 percent, respectively, using the Nave Bayes Simple algorithm. All training set's confusion matrix:-

```
=== Confusion Matrix ===
   a    b   <-- classified as
6523   27 |    a = Fraud suspected
  28 5422 |    b =  fraud not suspected
```

The Nave Bayes Simple algorithm, as shown in the resulting confusion matrix. The results show that different results with the same parameter and different testing modes can be found in the above table.

The results of this experiment show that the model developed with the Nave Bayes Simple Algorithm performs poorly in classifying new tax claims to the appropriate class when compared to the decision tree model developed previously.

]

4.2.3 Model Building using random forest algorithm

The decision tree model is constructed using the random forest algorithm. The experimentation's decision tree classification models are trained using a combination of 10-fold cross validation and percentage split classification models.

Experiment number	Total Records	No of attribute	Test modes used	Time to construct a model (seconds)	Result Accuracy (%)
5	12000	14	10-fold cross-validation	0.58second	98.98%
6	12000	14	Percentage split 66%	0.07	98.1 %

Table 8 finalized output of decision tree for random forest

The accuracy of 10-fold cross validation and percentage split (66/34 percent), 98.98 percent and 98.1 percent, respectively, are shown in table 8 of the random forest results. All training set's confusion matrix:-

=== Confusion Matrix ===

```
   a    b   <-- classified as
6511   39 |    a = Fraud suspected
  31 5419 |    b =  fraud not suspected
```

The random forest learning algorithm, as shown in the resulting confusion matrix. This result demonstrates that different results with the same parameter and different testing modes can be found in the above table.

4.3 Comparison of random forest algorithm, J48 Decision Tree, and Naïve Bayes Models

Selecting a better classification technique for building a model, which performs best in handling the prediction and detection of non-accepted tax audit claims, is one of the aims of this study. For that reason, the decision tree (particularly the J48 algorithm and random forest algorithm) and the Bayes (the Naïve Bayes Simple algorithm in particular) classification methods were applied for conducting experiments to build the best model.

Summary of experimental result for the three classification algorithms is presented in table 9 and table 10.

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

Experiment number	Classifier algorithm	Testing mode used	Result(Accuracy)	Ranking
1	J48	10-fold cross-validation	99%	1
2	Random forest	10-fold cross-validation	98.9%	2
3	Naïve Bayes	10-fold cross-validation	97%	3

Table 9 summarized 10-fold cross-validation experiment result

In table 10 the result of percentage split up data into 66/34%, 66%of the data for training and 34% for test.

Experiment No	Classifier algorithm	Testing mode	Accuracy	Ranking
1	J48	Percentage split 66/34%	98.89%	1
2	Random forest	Percentage split 66/34%	98%	2
3	Naïve Bayes	Percentage split 66/34%	98.89%	3

Table 10 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.

Generally, the first experiment that is conducted using the Naïve Bayes Simple algorithm, Random forest and J48, with the default 10-fold cross validation test option generates a better classification model with a better classification accuracy than the second one conducted training and testing percentage split test option.

Based on the result show in the above tables J48 is scores accuracy of 99% in 10-fold cross validation and scores accuracy of 98.89% in percentage split and best classifier than others algorithm.

4.4 Evaluation of the Discovered Knowledge

Data is required for DM task is the core of every process. Nevertheless, unfortunately the data required for effective Data mining is not readily available in a format that the DM algorithm required it. To make things worse some of the fields may contain outliers, missing values, inconsistent data types within a single field and many other possible anomalies. But this must be cleansed, integrated and transformed in a format suitable for the DM task to be undertaken. For that cause, the researcher has taken considerable time for the data-preprocessing task. Data cleaning (handling missing values and outlier detection and removal), and data integration tasks are carried out in a format suitable for classification techniques. The classification model developed using the J48 decision tree algorithm is chosen as the final model for this study.

From the decision tree developed in the above-mentioned experiment, it is possible to find out a set of rules simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node (Berry, M & Linoff, G, 1997). This generates rules that are clear in that it doesn't matter in what order they are executed. The following are some of the rules extracted from the decision tree.

RULE # 1

If Net worth = High and number of vat reports = High and total expenses = Low: then the tax payer is classified as **Fraud suspected**

RULE # 2

If Net worth = High and number of vat reports = High and total expenses =medium: then the tax payer is classified as **Fraud suspected**

RULE # 3

If Net worth = medium and total expenses= low: then the tax payer is classified as **not Fraud suspected**

RULE # 4

If Net worth = low and total expenses = high and vat refund amount= low and interest expense=high: then the tax payer is classified as **not Fraud suspected**

RULE # 5

If Net worth = low and Repair and Maintenance expenses = high and total expense= low and registration date=medium: then the tax payer is classified as **Fraud suspected**

RULE # 6

If Net worth = low and total expenses = low and gross profit= low and number of vat repors=low and vat refund amount =low then the tax payer is classified as **not fraud suspected**

RULE # 7

If Net worth = low and total expenses = low and gross profit= low and number of vat reports=low and vat refund amount =high then the tax payer is classified as **not fraud suspected**

RULE # 8

If Net worth = low and total expenses = low and gross profit= low and number of vat reports=low and vat refund amount =medium then the tax payer is classified as **fraud suspected**

Based on tax audit department experts statement the following is one of interesting rules, if companies declare high net worth means based on Rule#2 and the consultation of domain expert it is more fraud suspected than others. In Rule#5 if a company declares low Net worth and high total expenses these shows the company try to hide the profit (in other word Fraud suspected). Among those motivating rules and assessment of the model from the domain expert and audit dep't perspective was evaluated and Rule #8, Rule #1, and Rule #2 are interesting rules.

The rules that are presented above indicate the possible conditions in which a tax claim record could be classified in each of the fraud and non-fraud suspicious classes. 5 of the total fourteen variables are used for constructing the decision tree model. These attributes are claim net worth, total sales, penalty amount, vat refund amount and total expenses which are basis for building the decision tree. The generated decision tree has shown that net worth is the most determinant variable, as well as the model's top splitting variable.

Since the researcher conduct this research by his own without any sponsors, because of this reason the researcher has meet challenges related with budget shortage in performing this study.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The use of DM technology has gradually grown in popularity and has proven to be useful in a variety of industries, including finance, businesses, airlines, banking, and healthcare. DM technology has been useful for fraud detection, primarily in the tax authority. Tax fraud is, in fact, the most difficult problem in today's tax collection.

In this study, an attempt was made to use DM technology to aid in the detection and prediction of fraudulent tax claims in the tax authorities. While conducting the experiment, the six-step Cios et al. (2000) process model was strictly followed. This process model includes the phases of understanding the problem domain, understanding the data, preparing the data, DM, evaluating the discovered knowledge, and applying the discovered knowledge.

The data used in this study was obtained from the Amhara Revenue Authority's main database. After collecting the data, it was preprocessed and prepared in a format suitable for the DM tasks. This phase of the study consumed a significant amount of time. Classification techniques were used in the research.

The target class for this study was not included in the initial data collected from AMRA. The collected records are then sent to the classification module for model construction using the J48 decision tree algorithm. Different decision tree models were created using a 10-fold cross, a 66 percent percentage split, and the algorithm's default parameter values. With the Net worth as a splitting variable, the model developed using 10-fold cross validation with the default parameter values had a better classification accuracy of 99 percent on the training dataset.

In general, the findings of this study are very encouraging. The study has given away that it is possible to identify those fraud suspicious and non-suspicious tax claims and suggest tangible solutions for detecting them, using DM techniques.

5.2 Recommendations

This study is mainly accompanied for an academic purpose. Nevertheless, the results of this research are established promising to be applied to address practical problems of tax fraud. This research can help pave the way for a more comprehensive study in this area in the future, particularly in our region. The findings of this study also show that DM technology, specifically the J48 decision tree classification technique, is useful in detecting tax fraud.

As a result of the findings of this study, the following recommendations have been expanded.

- The awareness of fraud is not only a big issue in revenue sector, but also it is the core worry of most financial sector, as well all over the country. Furthermore, this study is worried about fraud in Amhara revenue authority; there are also other fraud activities that are occur in public zones, and others. These might leave a area to conduct further studies.
- The inclusion of many taxpayer attributes, as much as possible, should be given high attention, and more comprehensive models should be built by using large training and testing datasets taken from the relevant Amhara revenue authority databases of more than three-year tax payer profile in the organization.
- Only small subsets of all possible attributes, along with their values, are available in the Amhara revenue authority's database for this study. The database contains irregularities and missing values. There is no information about the number of employees in the companies, as well as their total material and currency assets. Because data is the most important component of DM research, the authority must create a data warehouse that can store both active and invalid data.
- Fraud can occur not only in revenue collection, but also within the authority, among professionals, auditors, and other employees. These can also be used as a starting point for further research.

REFERENCES

A, F., & p, T. (1997). Data preprocessing and Intelligent Data analysis. *nstitute of Information Technology, National research council*.

Bologna, & Lindquist. (1995). Fraud auditing and forensic accounting.

J.Han, & M.Kamber. (2004). Data Mining: Concepts and Techniques.

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

Witten, H. Ian, & Frank, E. (2005). Data mining practical machine learning tools andtechniques.

(AMRA), A. R. (n.d.). *BPR document*. Retrieved march 20, 2019, from www.amra.gov.et

] Luciano, A. (2009). Uses of Artificial Intelligence in the Brazilian Customs Fraud Detection system.

BPR document. (2005, may). (Amhara Revenue Authority (Amra))

Angell, M. B. (1938). Tax Evasion and Tax Avoidance. *Columbia Law Review, 38*(1), 80.

Ashour, Amira S, Dey, Nilanjan, & Dac Nhuong. (2014). Biological data mining: Techniques and applications. *Mining Multimedia Documents, 1*(4), 161-172.

Atos. (2015). tackling fraud with a big data approach. The common error of under-utilizing Revenue Authority data.

Baeza-Yates, R, & Ribeiro-Neto, B. (1999). Modelrn Information Retrieval.

Belete, B. (2017). The application of data mining techniques to support customer relationship management: the case of ethiopian revenue and customs authority. *Master of Science Thesis, School of Information Science Addis Ababa University*.

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

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

C, M. C. (2019). A concise study on Text Mining for Business Intelligence. *International Journal of Advanced Research in Computer and Communication Engineering, 6*(1).

Castellón González, Velásquez, P., & Juan D. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications, 40*(1), 1427-1436.

Chung, , H., Gray, , P., & Mannino, M. (2005). *Introduction to Data Mining and Knowledge Discovery.*

Chung, , H., Gray,, ,., & Mannino, , M. (2005). *Introduction to Data Mining and Knowledge Discovery* (3rd ed.). usa: Potomac, MD.

D. Larose, & C. Larose. (2014). *Discovering Knowledge in Data: An Introduction toData mining* (2nd edition ed.).

daniel, m. (2003). Application of Data Mining Technology to support fraud :acase of ethiopia revenue and custums authority. *Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia,*.

Directorate, I. T. (n.d.).

Elkan, C. (2001). Magical thinking in data mining. *Lessons from CoIL Challenge 2000.*, 426-431.

Gorunescu, F. (2011). Data Mining: Concepts, Models and Techniques.

Guo, L. (2003). Applying data mining techniques in property/casualty insurance. *Casualty Actuarial Society Forum*, 1-25.

H. Singh, & S. Bhagat,. (2015, january). Systems are developed and designed to help the organisatdions understand their Customs. *International Journal of Science, Technology & Management, 4*(1).

Hajizadeh, E., Ardakani, D. , & Shahrabi, J. (2010). Application of data mining techniques in stock markets:. *A survey, Journal of Economics and International Finance, 2(7)*, 09-118.

Halkidi, M , & Vazirgiannis, M. (2001). Evaluating the validity of clustering results based on density criteria and mulit-representatives. *A survey, Journal of Economics and International Finance, 2(7)*, 109-118.

Halkidi, M, & Vazirgiannis, M. (2011). Evaluating the Validity of Clustering Results Based on Density Criteria and Multi-representatives. *Greece*.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (second edition ed.). Elsevier inc.

Hendrickx, T., Cule, B., Meysman, P. N., & Stefan, L. (2015). Mining association rules in graphs based on frequent cohesive itemsets. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9078*(3), 637-648.

(n.d.). *Investigating Fraudulent Acts, university of Houston system administrative memorandum.*

J., H. (june,1996). Conference tutorial notes: data mining technique. *In Proceedings of ACMSIGMOD International Conference'96 on Management of Data (SIGMOD'96).* Montreal, Canada.

Jeffrey. (2006). 19th {IEEE} Computer Security Foundations Workshop. Venice, Italy: IEEE} Computer Society.

Julio Ponce, & Adem Karahoca. (2009). *Data Mining and Knowledge Discovery in Real Life Applications.* Vienna, Austria,: IN-TECH.

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

Kantardzic, M. (2002). Data mining: Concepts, models, methods, and algorithm.

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

Keshavamurthy, U., & Guruprasad, D. (2014). Learning Analytics: A Survey. *International Journal of Computer Trends and Technology, 18*(6), 260-264.

Kirkos , & Manolopoulos. (2007). *Applying Data Mining Methodologies for Auditor.* Selection.Expert Systems with Applications, Greece.

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

Kumari, N. (2013, june). Business Intelligence in a Nutshell. *Internation Journal of Innovative Research in Computer Communication Engineering, 1*(4), 969 - 975.

Lindgreen, A. (2004). Corruption and unethical behavior: report on a set of Danish guidelines. *Journal of Business Ethics, 51*(1), 31-39.

Martikainen, J. (2012). Data Mining in Tax Administration - Using Analytics to Enhance Tax Compliance. *Aalto University School of Business*, 74.

Meera, G., & Srivatsa, SK. (2010). Adaptive machine learning algorithm (AMLA) using J48 classifier for an NIDS environmen. *3(3)*, pp. 291-304.

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

Oluba, M. (2011). Nigeria's FIRS can continuously triple tax revenue through data mining.

Palace, B. (1996). What is data mining. *Anderson Graduate School of Management at UCLA*(0704).

Palshikar, G. (2002). Data Analysis Techniques for Fraud Detection.

Peng, D., Jin, Lianwen, & Wu, Yaqiang. (2019). *A fast and accurate fully convolutional network for end-to-end handwritten chinese text segmentation and recognition.* Proceedings of the International Conference on Document Analysis and Recognition, ICDAR.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *roceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103-119.

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

Pitas, , C., Chourdaki, K., Panagopoulos, A., & Constantinou, P. (2011). QoS Mining Methods for Performance Estimation of Mobile Radio Networks. *Measurement of Speech, Audio and Video Quality in Networks*(june), 20-23.

Quinlan, & J. R. C4.5. (1993). *Programs for machine learning.* usa: Morgan Kaufman Publisher.

Quinlan, J. R. (n.d.). Programs for machine learning.

report, A. a. (2008). *The Australian taxation office's use of data matching andanalytics in tax administration.* Canberra.

S, O. b.-I. (2012). Expert Systems with Applications Using data mining technique to enhance tax evasion detection performance. *An International Journa, 39*, 8769-8771.

Sharma, Anuj, & Kumar Panigrahi, Prabin. (2012). A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *nternational Journal of Computer Applications, 39*(1), 37-47.

Soni, Jyoti, Ansari, Uzma, & Sharma, Dipesh. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. *3*(6), 2385-2392.

Sowan, I. B. (2011). Enhancing Fuzzy Associative Rule Mining Approaches for improving prediction acuracy. *Bradford*.

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

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

U. Keshavamurthy, & H.S. Guruprasad. (2014, December). "Learning Analytics: A survey. *International Journal of Computer Trends and Technology (IJCTT), 18*.

Witten, H. Ian, & Frank, E. (2005). *Data mining practical machine learning tools and.* usa: Morgan Kaufman.

Wodajo, Biruk, Appalabatla, Sreedhar, & Krishna, Telkapalli Murali. (2017). Predicting the Fraudulent Claims of Tax PayersA Case of Boditi Town Revenue Authority, SNNPR, Ethiopia. *8*(10), 1380-1384.

Wodajo, Biruk, Krishna, S., & Telkapalli Murali. (2017). Predicting the Fraudulent Claims of Tax PayersA Case of Boditi Town Revenue Authority, SNNPR, Ethiopia. *8*(10), 1380-1384.

Woldu, L. (2003, julay). the Application of Data Mining in Crime Prevention : the the Application of Data Mining in Crime Prevention : the Case of Oromia Police. 107.

Zheng Xinqi, & Li Xinyun. (n.d.). *Actural Characters and Developmental Tendencies of the Data Mining Software*. Retrieved from http://www.paper.edu.cn

APPENDICES

Appendix1: interview

1. How many times the tax payer file is audited in a year?
2. How the auditors are providing priority for the taxpayer?
3. How many taxpayers are audited by one auditor in a year?
4. What are the current activities used by the authority to protect fraud?

Appendix 2: *decision tree generated with all training set techniques*

Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: data two
Instances: 12000
Attributes: 15
registratin date
total sales
number of employes
auditing time
tax audit acceptance
penality amount in ayear
total expences
net tax due
refund amount
vat refund amout
number of vat reports
gross profit
Net worth
use of cash registration machine
class
Test mode: evaluate on training data

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

J48 pruned tree

Net worth = High
| number of vat reports = High
| | total expences = Low: Fraud suspected (5768.0)
| | total expences = Medium: Fraud suspected (434.0)
| | total expences = High
| | | gross profit = Low: not suspected (5.0)
| | | gross profit = Medium: Fraud suspected (0.0)

```
|  |  |  gross profit = High
|  |  |  |  penality amount in ayear = Medium
|  |  |  |  |  vat refund amout = Medium:  not suspected  (5.0)
|  |  |  |  |  vat refund amout = High: Fraud suspected  (0.0)
|  |  |  |  |  vat refund amout = Low: Fraud suspected  (5.0)
|  |  |  |  penality amount in ayear = Low: Fraud suspected  (5.0)
|  |  |  |  penality amount in ayear = High: Fraud suspected  (0.0)
|  number of vat reports = Medium
|  |  registratin date = Medium:  not suspected  (401.0)
|  |  registratin date = High:  not suspected  (0.0)
|  |  registratin date = Low
|  |  |  vat refund amout = Medium
|  |  |  |  number of employes = Medium: Fraud suspected  (6.0)
|  |  |  |  number of employes = High
|  |  |  |  |  total sales = Low: Fraud suspected  (4.0)
|  |  |  |  |  total sales = High: Fraud suspected  (0.0)
|  |  |  |  |  total sales = Medium:  not suspected  (4.0)
|  |  |  |  number of employes = Low
|  |  |  |  |  tax audit acceptance = Low: Fraud suspected  (15.0/4.0)
|  |  |  |  |  tax audit acceptance = Medium:  not suspected  (4.0)
|  |  |  |  |  tax audit acceptance = High:  not suspected  (17.0/4.0)
|  |  |  vat refund amout = High: Fraud suspected  (0.0)
|  |  |  vat refund amout = Low: Fraud suspected  (10.0)
|  number of vat reports = Low
|  |  refund amount = Low
|  |  |  net tax due = Low: Fraud suspected  (5.0)
|  |  |  net tax due = Medium
|  |  |  |  vat refund amout = Medium: Fraud suspected  (0.0)
|  |  |  |  vat refund amout = High:  not suspected  (5.0)
|  |  |  |  vat refund amout = Low: Fraud suspected  (5.0)
|  |  |  net tax due = High
|  |  |  |  penality amount in ayear = Medium:  not suspected  (30.0/10.0)
|  |  |  |  penality amount in ayear = Low:  not suspected  (25.0/10.0)
|  |  |  |  penality amount in ayear = High: Fraud suspected  (5.0)
|  |  refund amount = Medium:  not suspected  (5.0)
|  |  refund amount = High:  not suspected  (0.0)
Net worth = Medium
|  total expences = Low:  not suspected  (3438.0)
|  total expences = Medium
|  |  penality amount in ayear = Medium: Fraud suspected  (1.0)
|  |  penality amount in ayear = Low
|  |  |  auditing time = Low
|  |  |  |  refund amount = Low: Fraud suspected  (15.0)
|  |  |  |  refund amount = Medium: Fraud suspected  (0.0)
|  |  |  |  refund amount = High:  not suspected  (4.0)
|  |  |  auditing time = Medium:  not suspected  (4.0)
|  |  |  auditing time = High
```

| | | | refund amount = Low: not suspected (4.0)
| | | | refund amount = Medium: not suspected (0.0)
| | | | refund amount = High: Fraud suspected (2.0)
| | penality amount in ayear = High: not suspected (6.0)
| total expences = High
| | gross profit = Low
| | | registratin date = Medium: not suspected (4.0)
| | | registratin date = High: Fraud suspected (0.0)
| | | registratin date = Low: Fraud suspected (11.0/1.0)
| | gross profit = Medium
| | | use of cash registration machine = High: Fraud suspected (0.0)
| | | use of cash registration machine = Medium: Fraud suspected (13.0/3.0)
| | | use of cash registration machine = Low: not suspected (5.0)
| | gross profit = High: not suspected (5.0)
Net worth = Low
| total expences = Low
| | gross profit = Low
| | | number of vat reports = High: Fraud suspected (1.0)
| | | number of vat reports = Medium
| | | | use of cash registration machine = High: not suspected (0.0)
| | | | use of cash registration machine = Medium: not suspected (10.0/1.0)
| | | | use of cash registration machine = Low: Fraud suspected (7.0/1.0)
| | | number of vat reports = Low
| | | | vat refund amout = Medium: Fraud suspected (10.0)
| | | | vat refund amout = High: not suspected (297.0)
| | | | vat refund amout = Low: not suspected (1047.0)
| | gross profit = Medium
| | | total sales = Low: Fraud suspected (15.0/2.0)
| | | total sales = High: Fraud suspected (0.0)
| | | total sales = Medium: not suspected (6.0/1.0)
| | gross profit = High: not suspected (0.0)
| total expences = Medium
| | vat refund amout = Medium
| | | auditing time = Low: not suspected (6.0)
| | | auditing time = Medium: not suspected (0.0)
| | | auditing time = High: Fraud suspected (4.0)
| | vat refund amout = High: Fraud suspected (30.0)
| | vat refund amout = Low
| | | number of vat reports = High: Fraud suspected (12.0)
| | | number of vat reports = Medium
| | | | refund amount = Low: Fraud suspected (15.0)
| | | | refund amount = Medium
| | | | | use of cash registration machine = High: not suspected (0.0)
| | | | | use of cash registration machine = Medium: not suspected (14.0/1.0)
| | | | | use of cash registration machine = Low
| | | | | | auditing time = Low: Fraud suspected (4.0)
| | | | | | auditing time = Medium: not suspected (8.0)

```
|  |  |  |  |  |  auditing time = High: Fraud suspected  (8.0)
|  |  |  |  refund amount = High: Fraud suspected  (5.0)
|  |  |  number of vat reports = Low
|  |  |  |  use of cash registration machine = High: Fraud suspected  (0.0)
|  |  |  |  use of cash registration machine = Medium:  not suspected  (7.0)
|  |  |  |  use of cash registration machine = Low
|  |  |  |  |  auditing time = Low: Fraud suspected  (20.0/2.0)
|  |  |  |  |  auditing time = Medium
|  |  |  |  |  |  refund amount = Low:  not suspected  (6.0/1.0)
|  |  |  |  |  |  refund amount = Medium: Fraud suspected  (6.0/1.0)
|  |  |  |  |  |  refund amount = High: Fraud suspected  (0.0)
|  |  |  |  |  auditing time = High:  not suspected  (11.0/2.0)
|  total expences = High
|  |  vat refund amout = Medium:  not suspected  (10.0)
|  |  vat refund amout = High
|  |  |  refund amount = Low
|  |  |  |  penality amount in ayear = Medium: Fraud suspected  (5.0)
|  |  |  |  penality amount in ayear = Low:  not suspected  (4.0)
|  |  |  |  penality amount in ayear = High: Fraud suspected  (0.0)
|  |  |  refund amount = Medium: Fraud suspected  (0.0)
|  |  |  refund amount = High: Fraud suspected  (17.0)
|  |  vat refund amout = Low
|  |  |  total sales = Low
|  |  |  |  gross profit = Low
|  |  |  |  |  registratin date = Medium:  not suspected  (9.0/1.0)
|  |  |  |  |  registratin date = High:  not suspected  (0.0)
|  |  |  |  |  registratin date = Low
|  |  |  |  |  |  number of vat reports = High: Fraud suspected  (13.0/4.0)
|  |  |  |  |  |  number of vat reports = Medium: Fraud suspected  (3.0/1.0)
|  |  |  |  |  |  number of vat reports = Low:  not suspected  (9.0/3.0)
|  |  |  |  gross profit = Medium
|  |  |  |  |  registratin date = Medium: Fraud suspected  (10.0/1.0)
|  |  |  |  |  registratin date = High: Fraud suspected  (0.0)
|  |  |  |  |  registratin date = Low
|  |  |  |  |  |  use of cash registration machine = High:  not suspected  (0.0)
|  |  |  |  |  |  use of cash registration machine = Medium
|  |  |  |  |  |  |  auditing time = Low: Fraud suspected  (0.0)
|  |  |  |  |  |  |  auditing time = Medium: Fraud suspected  (22.0/9.0)
|  |  |  |  |  |  |  auditing time = High:  not suspected  (10.0/4.0)
|  |  |  |  |  |  use of cash registration machine = Low:  not suspected  (12.0/1.0)
|  |  |  |  gross profit = High: Fraud suspected  (4.0)
|  |  |  total sales = High: Fraud suspected  (14.0/1.0)
|  |  |  total sales = Medium
|  |  |  |  penality amount in ayear = Medium
|  |  |  |  |  refund amount = Low:  not suspected  (5.0)
|  |  |  |  |  refund amount = Medium
|  |  |  |  |  |  gross profit = Low
```

```
|  |  |  |  |  |  |  use of cash registration machine = High:  not suspected  (0.0)
|  |  |  |  |  |  |  use of cash registration machine = Medium: Fraud suspected  (5.0/1.0)
|  |  |  |  |  |  |  use of cash registration machine = Low:  not suspected  (12.0)
|  |  |  |  |  |  gross profit = Medium: Fraud suspected  (3.0)
|  |  |  |  |  |  gross profit = High:  not suspected  (0.0)
|  |  |  |  |  refund amount = High:  not suspected  (2.0)
|  |  |  |  penality amount in ayear = Low
|  |  |  |  |  number of vat reports = High: Fraud suspected  (0.0)
|  |  |  |  |  number of vat reports = Medium
|  |  |  |  |  |  use of cash registration machine = High: Fraud suspected  (0.0)
|  |  |  |  |  |  use of cash registration machine = Medium: Fraud suspected  (2.0)
|  |  |  |  |  |  use of cash registration machine = Low:  not suspected  (2.0)
|  |  |  |  |  number of vat reports = Low: Fraud suspected  (3.0)
|  |  |  |  penality amount in ayear = High:  not suspected  (0.0)

Number of Leaves  :   109

Size of the tree :        163


Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.06 seconds

=== Summary ===

Correctly Classified Instances       11930               99.4167 %
Incorrectly Classified Instances        70               0.5833 %
Kappa statistic                          0.9882
Mean absolute error                      0.0083
Root mean squared error                  0.0644
Relative absolute error                  1.6747 %
Root relative squared error             12.9411 %
Total Number of Instances            12000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.994    0.006    0.995      0.994    0.995      0.988    1.000     1.000     Fraud suspected
               0.994    0.006    0.993      0.994    0.994      0.988    1.000     1.000     not suspected
Weighted Avg.  0.994    0.006    0.994      0.994    0.994      0.988    1.000     1.000
```

=== Confusion Matrix ===

a	b	<-- classified as
6511	39	a = Fraud suspected
31	5419	b = not suspected

Appendix 3: result of confusion matrix for classification techniques

1. J48 algorithm Confusion matrix result of with 10-fold cross validation

=== Confusion Matrix ===

a	**b**	**<-- classified as**
6484	**66**	**a = Fraud suspected**
54	**5396**	**b = Fraud not suspected**

2. Naïve Bayes algorithm Confusion matrix result of with 10-fold cross validation

=== Confusion Matrix ===

a	**b**	**<-- classified as**
6213	**337**	**a = Fraud suspected**
18	**5432**	**b = not suspected**

3. Random Forest algorithm Confusion matrix result of with 10-fold cross validation

=== Confusion Matrix ===

a	**b**	**<-- classified as**
6491	**59**	**a = Fraud suspected**
60	**5390**	**b = fraud not suspected**