Data Mining for Detection of Tax Evasion: The Case of Tax Payers in Addis Ababa

A Thesis Presented

By

Etsegenet Mekonnen

to

The Faculty of Informatics

of

St. Mary's University

In Partial Fulfillment of the Requirements
For the Degree of Master of Science

in

Computer science

June 2021

ACCEPTANCE

Data mining for Detection of Tax evasion: The Case of Tax Payers in Addis Ababa

By

Etsegenet Mekonnen

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

Name	Signature	Date
Asrat Mulatu (Ph.D.) Internal Examiner	___________________	_04/05/2021_
Temtim Assefa, PhD External Examiner	___________	04/05/2021

____

Dean, Faculty of Informatics

Name	Signature	Date
Dr. Getahun Semeon	___________	__________

February, 2021

DECLARATION

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

Etsegenet Mekonnen

Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor.
Dr. Getahun Semeon

Signature

Addis Ababa

Ethiopia

February 17, 2021

Table of Contents

Acknowledgments vii

List of Figures viii

List of Tables ix

Abstract x

CHAPTER ONE: INTRODUCTION 1

1.1 Background and Motivation 1

1.2 Statement of the Problem 3

1.3 Research Questions 4

1.4 Objective of the Study 4

1.4.1 General Objective 4

1.4.2 Specific Objectives 4

1.5. Research Methodology 5

1.6 Significance of the Study 5

1.7 Scope of the Study 6

1.8 Limitation of the Study 6

1.9 Organization of the Study 6

CHAPTER TWO: REVIEW OF RELATED LITERATURES 7

2.1 Concepts and Definitions 7

2.1.1 Data Mining 7

2.1.2 Data Mining for Fraud Detection 8

2.1.3 Data Mining for Tax Evasion 9

2.2 Techniques of Data Mining 9

2.2.1 Classification 9

2.2.2 Clustering 11

2.2.3 Regression 12

2.3 Data Mining Algorithms 12

2.3.1 Decision tree 12

2.3.2 Naive Bayes 13

2.3.3 Logistic regression 13

2.3.4 Random Forest 13

2.4 Data Mining Tasks 14

2.5 Data Mining Models 15

2.5.1 KDD Process Model 15

2.5.2 CRISP-DM Process 17

2.5.3 Cios model 19

2.6 Related Works 21

CHAPTER THREE: RESEARCH METHODOLOGY……………………………………25

3.1. Introduction 25

3.2. Research Design 25

3.3 Understanding the Problem Domain: 26

3.4 Data Understanding 26

3.5 Data Preparation & Preprocessing 29

3.6 Model Building 32

3.7 Evaluation of the Discovered Knowledge 33

3.8 Using Discovered Knowledge 33

CHAPTER FOUR: EXPERIMENTATION 34

4.1 Introduction 34

4.2 Model Building 34

4.2.1 Cluster Modeling 35

4.2.2 Classification... 43

4.3 Evaluation 54

4.4 Deployment of the Result 55

4.5 Sample prototype 57

CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS…………………..58

5.1 Conclusions... 58

5.2 Recommendations... 59

5.3 Recommendation for Future Studies 59

5.4 Limitations of the Study.. 59

REFERENCES 60

Appendix.. 64

Acknowledgments

First of all, I would like to thank the almighty God and his Mother for giving me the strength, peace of my mind, and good health. Second, I would also like to express my deepest gratitude to my advisor Dr. Getahun Semeon, for his unreserved follow-up, invaluable comments, and constructive guidance throughout conducting this study.

Finally, I would also like to express the deepest gratitude to my family and friends who have been providing their advice and encouragement always including those hard times.

List of Figures

Figure 2-1KDD process 17

Figure 2-2 The CRISP-DM process.. 18

Figure 2-3 Cios model 20

Figure 4-1Preprocessing Data in Weka 3.8.4 (Source: Result Weka, 2020).................... 35

Figure 4-2 Threshold Curve of Cluster 1 46

Figure 4-3 Threshold Curve of Cluster 2 46

Figure 4-4 Threshold Curve of Cluster 1 Naïvebayes algorithm...................................... 48

Figure 4-5 Threshold Curve for Cluster 2: Naïvebayes.. 48

Figure 4-6 Threshold Curve for Cluster 1: Neural Network... 50

Figure 4-7 Threshold Curve of Cluster 2: Neural Network.. 50

Figure 4-8 Threshold Curves for Cluster 1: Random Forest 53

Figure 4-9 Threshold Curve Cluster 2: Random Forest.. 54

Figure 4-10 sample no tax evasion detector 57

Figure 4-11 sample tax evasion detector 57

List of Tables

Table 3. 1 Description about Attributes in Raw Data 28

Table 3. 2 Attribute Discretization labels 31

Table 4. 1 Cluster Modeling Experiment I 37

Table 4. 2 Cluster Modeling Result of Experiment II 39

Table 4. 3 Cluster Modeling Experiment III Summary 41

Table 4. 4 Comparison of Clustering Models 42

Table 4. 5 Summary of J48 classification experiments 44

Table 4. 6 Summary of Confusion Matrix of J48 algorithm 45

Table 4. 7 Result of Naïve Bayes Classifier Experiment 47

Table 4. 8 Result of Multilayer Perceptron Algorithm 49

Table 4. 9 Summary of Random Forest Experiment 51

Table 4. 10 Summary of Performance of Classifiers 52

Abstract

The Tax has a high contribution to an economy; the government uses tax revenue for different government expenditure. Businesses and privates have obligations to pay tax from their income to the government. Despite this importance and responsibilities, corporates and individuals are involved in tax evasion. In Ethiopia Specifically in Addis Ababa, this problem is severe that about 50% of companies are involved in tax evasion. This study is conducted to develop tax evasion detecting techniques by using data mining procedures. It has used data about taxpayers in Addis Ababa and collected from the ministry of revenue at different tax payer's branch offices in Addis Ababa. The study has followed the KDD method of data mining. The study has conducted two main procedures for model development; cluster modeling and classification modeling. The cluster modeling was conducted by using the K-mean algorithm and classification modeling was conducted by implementing different classifiers; J48, Naïvebayes, Neural Network, and Random Forest. Finally, the tax evasion detecting model was developed by using the Random Forest algorithm after making the comparison with other classifiers implemented. Besides, the decision rule construction was conducted by using the J48 algorithm. Finally, the study indicated that tax evasion practices with related to the liability of companies, expense, and amount of tax.

Keywords: Tax Evasion, Clustering, Classification, Model Development, Decision Rule

CHAPTER ONE

INTRODUCTION

1.1 Background and Motivation

In recent years, data mining has attracted a great deal of attention in the information industry and society, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. According to [1] DM is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data to discover meaningful patterns and rules. DM combines techniques from machine learning, pattern recognition, statistics, database theory, and visualization to extract concepts, concept interrelations, and interesting patterns automatically from large corporate databases. The primary goal of data mining is to extract knowledge from data to support the decision-making, planning, and problem-solving process. Primarily data mining is used for prediction and description. Prediction is identifying unknown values/relationships/patterns from known values by using classification; and the description is an interpretation of a large database based on clustering, pattern discovery, and deviation detection [2].

DM methodology can improve traditional statistical approaches to solving business solutions. It can easily predict using models that show patterns with reduced time and increased accuracy. Models that predict relationships and behaviors more accurately lead to greater returns with reduced costs. Data mining produces important information and knowledge for decision-making. The information and knowledge obtained through data mining can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [3].

Fraud detection is among the main concerns of researchers and companies when using data mining. Fraud encompasses a wide range of illicit practices and illegal acts involving intentional deception or misrepresentation. Fraud is an illegal act characterized by deceit, concealment, or violation of trust. Frauds are committed by parties and organizations to secure personal or business advantage through the unlawful act to obtain money, property, services, or to avoid payment or loss of services. Different types of frauds are frequently enacted for personal and business advantages. Tax evasion is a very common fraud executed by tax authorities [2].

Tax is one of the most necessary financial resources of a government for accomplishing specific goals. Taxes are the source of government earning that is mainly used for infrastructure development [1]. However, some businesses often attempt to evade their payment of correct taxes. Tax evasion and tax fraud have been common issues for tax administrations, especially pertaining to developing countries. Tax evasion is the illegal evasion of taxes by individuals and corporations. It is the intentional act of lying on a tax return form with the intent to lower one's tax liability. Under-reporting is one of the most common types of tax fraud, it consists of filing a tax return form with a lesser tax base. As a result of this act, fiscal revenues are reduced, undermining public investment [1].

Tax authorities have to bear the costs of the detection and prevention of illegal tax evasion activities. If the government cannot effectively detect illegal tax evasion activities, public investment would be negatively affected due to the budgetary shortage resulting from the loss of tax revenues [4]. Tax evasion not only leads to significant revenue losses, but also to a considerable increase in administrative costs used to detect the illegal tax evasion activities [1]. Indirectly underreported taxable income from the business, is often directly accompanied by underreported sales revenues.

Tax authorities have often relied on the sampling method and the personal judgment of tax auditors in selecting suspicious tax reports to audit for potential tax evasion activities. Auditing and Tax inspection are important and effective but checking all records is time-consuming. The large volume of data is a challenge for traditional data mining methods [5]. As a result, effective ways to detect related tax evasion activities have always been an important and challenging issue for tax authorities in any country. Since finding tax evasion in a large database is difficult, the data mining approach helps to find out tax evasion patterns from a large database and identifying suspicious groups of tax evasion with reduced time and increased accuracy by using machine learning.

Therefore, this study will be conducted with the purpose of identifying scientific approach to improve productivity and performance of tax audit in the detection of tax evasion by taxpayers in Addis Ababa.

1.2 Statement of the Problem

Tax evasion is highly practiced throughout the world irrespective of the level of economic development of a country. But the problem is highly observed in developing countries [1]. Ethiopia is not an exception from the tax fraud problem. In Ethiopia, annually about 11.4 billion tax evasion is estimated by a tax authority. Although tax evasion is practiced by businesses throughout the country, about 72% is in Addis Ababa. The amount of tax evasion in the country is equivalent to the budget of big public organizations. This suggests that the country is losing big investment due to tax evasion. The problem becomes very high due to the behavior of taxpayers and lack of an appropriate system to detect the problem [6]. Currently, the tax authority is mainly using manual auditing and inspection for tax evasion that makes detection of tax fraud challenging and the audit finding unreliable. This implies the existence of serious tax fraud and high importance of automated detection for the tax evasion problem.

Data mining through machine learning techniques is highly important for tax evasion detection that detecting tax fraud is one of the main priorities of tax authorities which are required to develop cost-efficient strategies to tackle the problem [4]. Continued tax evasion results in undermining public investment and negatively affecting the economy as a whole [1]. Data mining may be an effective tool for enhancing the efficiency and effectiveness of the detection of illegal tax evasion. Data mining techniques are important for detecting suspicious tax evasion reports and thereby recoup unpaid taxes [5].

Different studies were conducted to while developing tax evasion detection performance. but these studies have some limitations in the factors that result on tax evasion. for example Roung-Shiunn et.al [7] failed to include attributes that indicate financial aspects of the tax payers. although there are some external factors that encourage tax evasion, mainly tax evasion comes from internal aspects of the tax payers [1]. Therefore, this study mainly focuses on internal factors that are mainly related to tax payers. therefore, this study highly contributes by identifying internal causes for tax evasion.

Recently different studies were conducted to detect a problem of tax fraud and recommended tax authorities to use machine learning techniques by mining data of businesses [5] [8] [3] [9]. These studies have followed different strategies in developing tax evasion models that they reached different conclusions suggesting that the tax evasion detection models vary from area to area. In addition, these studies were conducted in a single category of taxpayers. In addition to this gap, there are no studies conducted about tax evasion detection at taxpayers in the context of Addis Ababa where 72% of tax evasion is being practiced. Therefore, this study will be conducted to mine data of taxpayers in Addis Ababa to detect tax evasion by using machine learning for different levels of taxpayers.

1.3 Research Questions

The research questions include;

- What are the characteristics of taxevading and non-evading taxpayers in Addis Ababa?
- What data mining model can best predict tax evasion by taxpayers in Addis Ababa?

1.4 Objective of the Study

1.4.1 General Objective

This study will be conducted with a main objective of building a predictive model that detects tax evasion by tax payers in Addis Ababa.

1.4.2 Specific Objectives

The following specific objectives were addressed in order to attain the stated general objective

- To identify characteristics of tax payers in Addis Ababa;
- To identify appropriate algorithms, techniques and tools for analyzing data on tax evasion
- To develop a prediction model for detection of tax evasion;
- To evaluate tax evasion detection models that can examine tax evasion practices;
- To develop decision rule that enables to detect tax evasion by tax payers in Addis Ababa; and
- To report the result and forward recommendations for policy interventions and further studies.

1.5. Research Methodology

This study was conducted by following experimental design and the data was collected from the tax administration branches in Addis Ababa. The branch offices store data with various features and information that are not related with tax evasion. The study conducted data preprocessing to reduce noise and handle missing values, to remove irrelevant attributes, and to conduct data transformation for normalization. Therefore, the study has followed data mining process of data cleaning, data reduction, data transformation, data formatting, and attribute selection.

In addition, the data mining tasks was descriptive and predictive that the study intends to describe the tax payers in the category of tax evading and non-evading. Further, prediction was conducted to associate the features of the tax payers. Thus, the study has followed knowledge discovery through clustering and classification.

For these activities, the study used J48, Naïve Bayes, Neural Network and Random Forest algorithms by using recent and stable version of WEKA 3.8.4 computer software.

The study has used audited financial reports of tax payers from small tax payers to large tax payers in Addis Ababa. According to Revenue Authority (2020) there are 21,087 reports of tax payment. Some of these reports are audited tax reports. During the data collection the researcher identified that there are 7,272 reports of tax payment. Thus, this study has used this dataset training and testing purpose in conducting the study experiment.

1.6 Significance of the Study

This study was conducted with main objective of developing tax evasion detecting model by taking tax payers in Addis Ababa. The finding of the study will help to detect tax fraud and make auditing activities efficient. Therefore, the study will be highly important to Tax Authorities and branch offices in Addis Ababa and other part of the country who face tax evasion problem from tax payers. Ministry of Revenue can include final work of this study as a strategy and monitoring mechanism. In addition, tax administration offices can follow rules provided by this study and easily can detect suspicious companies as a preliminary audit.

Further, this study will have contribution to studies in the area of tax fraud detection.

1.7 Scope of the Study

This study was conducted to detect tax evasion committed by tax payers in Addis Ababa. Therefore, geographically, the study was scoped to tax payers in Addis Ababa because the tax administration is not integrated at national level. The tax administration branch offices hold record of tax payers only under their follow-up. The study was scoped to data about the tax payers. Further, the study was conducted by using only secondary data about the tax payers.

1.8 Limitation of the Study

There might be effect of tax administration practices on tax evasion in the selected area. There may be corruption practices in tax administration and information recorded will be wrong. In addition, existence of the corruption will not be indicated in the data. Therefore, due to lack of information about corruption practice, the study will not include the role of tax administration on developing strategy for tax detection. Another limitation of the study will be excluding primary information from tax payers for tax avoidance.

1.9 Organization of the Study

This study was organized into five chapters; the first chapter will be the introductory chapter that briefs about background of the study, statement of problem, research questions, objective of the study, research methodology, significance of the study, and scope and limitations of the study. The second chapter presents review of related literature that mainly includes theoretical grounds and related works for the study. The third chapter was about methods used in conducting the study. The fourth chapter focus on detailed experimentation and analysis as well as interpretations of experimental results; and the final chapter was about conclusions and recommendations based on the findings of the study.

CHAPTER TWO

REVIEW OF RELATED LITERATURES

2.1 Concepts and Definitions

2.1.1 Data Mining

Different definitions are provided for data mining. Berry and Linoff (2000); Han and Kamber (2006) defining Data mining as a process of extracting or mining knowledge from large amounts of data in order to discover meaningful patterns and rules. It is valuable to discover implicit, potentially useful information from huge data stored in databases via building computer programs that sift through databases automatically or semi-automatically, seeking meaningful patterns (Berry and Linoff, 2000; Han and Kamber, 2006).

In addition to definition provided by Berry and Linoff (2000); Han and Kamber (2006), Guo (2003) has explained data mining as an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, data visualization, optimization, information retrieval, high end computing, and others. Data mining is the process of extracting or mining knowledge from large data sets. Data mining is the filed in which useful outcome that is being predicted from large database.

Data mining has similar or a bit different meaning with different terms, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Data mining also considered as an exploratory data analysis. Generally, Data mining uses advanced data analysis tools to find out previously unknown (hidden), valid patterns and relationships among data in large data sets. It is the core field for different disciplines such as database, machine learning and pattern recognition. It is a common practice to refer to the idea of searching applicable patterns in data using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing [3].

Similar definitions are provided to data mining and machine learning. But there are studies that differentiate data mining from machine learning. Clifton (2016) defines machine learning as exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules by enabling machines to learn without programming them explicitly. It enables machines to make predictions, perform clustering, extract association rules, or make decisions from a given dataset. Machine learning also enables data exploration and analysis without any specific hypothesis in mind, as opposed to traditional statistical analysis, in which experiments are designed around a particular hypothesis. While this openness adds a strong exploratory aspect to machine learning projects, it also requires that organizations use a systematic approach in order to achieve usable results.

On the development of tools for expressing domain expertise, translating it into a learning bias, and quantifying the effect of such a bias on the success of learning is a central theme of the theory of machine learning. Roughly speaking, the stronger the prior knowledge (or prior assumptions) that one starts the learning process with, the easier it is to learn from further examples. As it is described in, generally, machine learning is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories.

2.1.2 Data Mining for Fraud Detection

Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretenses. Fraud is typically an act which involves many repeated methods; making searching for patterns a general focus for fraud detection. It is committed in different organizations. Government organizations are places where frauds are frequently committed. Government fraud is committing fraud against federal agencies such as the departments of Health and Human Services, Transportation, Education, or Energy. Types of government fraud include billing for unnecessary procedures, overcharging for items that cost much less, providing old equipment when billing for new or reporting hours worked for a worker that does not exist.

Data analysts can prevent fraud by making algorithms to detect patterns and anomalies. Fraud detection can be separated by the use of statistical data analysis techniques or artificial intelligence (AI). Statistical data analysis techniques include the use of; Calculating statistical parameters, Regression analysis, Probability distributions and models and Data matching. AI techniques used to detect fraud include the use of: Data mining that classify group and segment data to search through up to millions of transactions to find patterns and detect fraud; Neural networks that learn suspicious-looking patterns, and use those patterns to detect them further; Machine learning that can automatically identify characteristics found in fraud; and Pattern recognition that detect classes, clusters, and patterns of suspicious behavior.

2.1.3 Data Mining for Tax Evasion

Lehmann and Coleman (1994) and Asprey and arsons (1975) defines tax evasion as a criminal falsification or non-disclosure as a means of reducing tax and have always been regarded as unacceptable at law. Tax evasion can occur where taxpayers employ fraudulent methods to evade the payment of taxes. Tax evasion activities are in contravention of the law whereby a person who derives a taxable income either pays no tax or pays less tax than he would otherwise be bound to pay. Tax evasion includes the failure to make a return of taxable income or a failure to disclose in a return the true amount of income derived. According to Alm and Vazquez (2001) and Chiumya (2006) tax evasion is Illegal activities or practices which are adopted by a taxpayer to escape him/her-self from taxation. For this purpose, taxable income/profits liable to tax or other taxable activities are concealed, tax-reducing factors like expenditures, exemptions, or other tax credits are knowingly and willfully overstated and the amounts received or the source of income misrepresented.

2.2 Techniques of Data Mining

Data mining uses already built tools to get out the useful hidden patterns, trends and prediction of future can be obtained using the techniques. Data mining involves model to discover patterns which consists of various components; classification, clustering and regression.

2.2.1 Classification

Classification is one of the data mining techniques which is useful for predicting group membership for data instances. Classification is a supervised kind of machine learning in which there is provision of labeled data in advance. By providing training data the model can be trained and the future of data can be predicted. Prediction is in the form of predicting the class to which data can belong. Training is based on the training sample provided. Basically there are two types of attributes available that are output or dependent attribute and input or the independent attribute. In the supervised classification, there is mapping of input data set to finite set of discrete class labels [2]. Main data mining task is classification which has main work to assign each record of a database to one of the predefined classes. The next is clustering which works in the way that it finds groups of records instead of only one record that are close to each other according to metrics defined by user. The next task is association which defines implication rules on the basis of that subset of record attributes can be defined.

According to Cios [12], Classification is the best understood of all data mining approaches among all Predictive models. Classification is commonly characterized as with classification tasks such as supervised learning, categorical dependent variable and ability of assigning new data in to the set of well-defined classes. Classification is one of the classic data mining techniques used to classify each item in a set of data into one of predefined set of classes or groups [12]. Classification method makes use of mathematical techniques such as decision trees, support vector machine, neural network and Bayesian learning. In classification, software is developed that can learn how to classify the data items into groups [12].

Classification methods in data mining include decision tree induction, rule based, back-propagation, and Lazy learners. In the decision tree induction method, from the class labeled tuples the decision tree is build. Decision tree is tree like structure in which there are internal node, branch and leaf node. Internal node specifies the test on attribute, branch represents the outcome of the test and leaf node represents the class label. Two steps that are learning and testing are simple and fast. The main goal is to predict the output for continuous attribute but decision tree is less appropriate for estimating tasks. There may be errors in predicting the classes by using decision tree approach. Pruning algorithms are expensive and building decision tree is also an expensive task as at each level there is splitting of node.

Rule based classification is represented by set of IF- THEN rules. First of all how many of these rules are examined and next care is about how these rules are build and can be generated from decision tree or it may be generated from training data using sequential covering algorithm.

Backpropagation classification is a neural network learning algorithm. Neural network learning is often called connectionist learning as it builds connections. It is feasible for that application where long times training is required. The most popular neural network algorithm is backpropagation. This algorithm proceeds in the way that it iteratively performs processing of data and it learns by comparing the results with the target value given earlier.

Eager learner is the form in which generalization model is being developed earlier before new tuple is being received for classifying. In lazy learner approach when given a training tuple it simply stores it and waits until a test tuple is given. It supports incremental learning. Some of the examples of lazy learner are K-nearest neighbor classifier and case- based reasoning classifiers [10].

2.2.2 Clustering

It is an unsupervised classification. It is also known as exploratory data analysis in which there is no provision of labeled data. The main aim of clustering technique is to separate the unlabeled data set into finite and discrete set of natural and hidden data structures. There is no provision of providing accurate characterization of unobserved samples that are generated from by same probability distribution [12]. Broadly clustering has two areas: hard clustering and soft clustering. In hard clustering same object can belong to single cluster and in soft clustering same object can belong to different clusters.

The clustering process includes various steps and it is a step by step process in which the results can be verified. There are four main steps of clustering process; Feature selection or extraction, validation, relative indices, and result interpretation.

Feature selection or extraction is selecting distinguishing feature from set of candidates and extracting means which it utilizes in the transformation to generate the useful and novel features from original ones [14]. After feature selection, clustering algorithm is designed. Every clustering algorithm is affected by measures. Next is to optimize the clustering solutions. It has been very difficult to develop a unified framework for reasoning about it (clustering) at a technical level, and profoundly diverse approaches to clustering [15].

Validations of clusters are in the sense whether the groups formed are valid or not, the data is correctly identified according to groups. These all can be checked by main three indices which are known as testing criteria and these are External indices, Internal indices and Relative indices. These indices are defined on different clustering structures that are known as partitioning clustering, hierarchal clustering and individual clusters [16]. The final step of clustering process is Result interpretation that provide accuracy to user and provide a meaningful insight form original data so that efficient results can be provided.

There are various methods for clustering which act as a general strategy to solve the problem and to complete this, an instance of method is used called algorithm. Various clustering algorithms are compared based on parameters differentiating them like the algorithms supported, type and size of dataset supported.

Whether the algorithms can handle higher dimensionality of data and noisy data or not, broadly clustering methods can be divided into two main categories which have number of instances. These categories are hierarchical and partitioning based methods. In hierarchical based clustering, the data sets of n elements are divided into hierarchy of groups which has tree like structure. In partitioning based methods, the output is like k partitions of N dataset elements. Partitioning methods simply partitions the dataset into n objects.

2.2.3 Regression

Regression is another data mining technique which is based on supervised learning and is used to predict a continuous and numerical target. It predicts number, sales, profit, square footage, temperature or mortgage rates. All these can be predicted by using regression techniques. Regression starts with data set value already known. It is based on training process. It estimates the value by comparing already known and predicted values [20].

2.3 Data Mining Algorithms

No model is perfectly accurate; there are strengths and weaknesses to different types of models. A variety of machine learning algorithms are applied in the fraud detection in recent years. Combining multiple models together far outperforms any individual model. Here are some of the algorithms that can be employed in the process of data mining.

2.3.1 Decision tree

A decision tree is an approach using a tree data structure such as a chart or matrix of choices and its feasible results in order to forecast the ultimate choice. It is a pseudo code to approach evaluated objectives. These kinds of algorithms are very popular for interactive learning and have been used effectively for various assignments overseas. Similar to binary system, the Decision Support System (DSS) is categorized as regression and classification trees. The branch of decision tree follows a structure where there will be one root node and other will be leaf or child. The decision is taken on the basis of traversing of the flow. Based on the probability of occurred events, a decision tree learning method is to predict and select the relative optimum solution by comparing the solutions to be evaluated with probability calculation and tree-like graph. Generally, decision tree learning algorithms, such as ID3 or C4.5, use Entropy to measure Information Gain, and in some cases prune the tree based on Entropy to obtain better classification results.

2.3.2 Naive Bayes

In this approach all the features are categorized into parts. Such extracted features are classified in a way such no other cluster know about the other features. Further, the features are categorized as true or false fraud activity for the person. Naive Bayes is often considered a trivial model, but it contributes to producing accurate results in a large stack of algorithms, particularly while onboarding new customers with limited training data or providing reasons why and how a particular risk score is arrived.

There is variation between Naïve Bayes classifier and decision tree classifiers. Unlike to naïve Bayes classifier, decision tree is an approach using a tree data structure for decision supporting system. Thus, Naïve Bayes has weakness of excluding other decision supporting systems. It commits only the classification purpose.

2.3.3 Logistic regression

Logistic regression is a controlled technique for classifying binary count on a variable that estimates the probability of results with zero or one attributes, yes or no and false or true, based on the independent variable of the dataset, which is logistic regression. Regression of logistics is alike to linear regression, as the direct row is acquired in the linear regression, logistic regression indicates a curve. The forecast is counted on the use of one or more predictors or autonomous matrix, logical classification generates logistic equations that trace the numbers between null and 1. Logistic regression is particularly useful in cases where only a limited set of information is available for risk analysis; such as a case with sparse features (for example, a guest check-out experience while shopping). Logistic regression models provide easily interpretable results.

2.3.4 Random Forest

Random Forest is a classifying and regressive algorithm. In short, it's a decision-tab classification set. Spontaneous forests have benefited over the tree, as they actually correct only the practice of over fitting. A small subset of the training set is sampled completely randomly so that each tree is trained, then every node divides on a new feature that is chosen from a completely random subset of the entire feature set.

Random decision forests are a powerful, scalable, and intuitive model. They can model interactions among features, are relatively inexpensive to train, and are one of the most interpretable machine learning models around. Beyond that, random decision forests are highly accurate on our datasets, remove bias, and are widely used in large scale applications. Random Forests are more robust for a number of real-world problems such as missing data, noise, outliers, and errors. Random Forests also allow multiple types of data (numbers of different scales, text, Booleans, etc.) to scale immensely well and parallelize very easily. They are fast to train and score, and require less effort to achieve the best results. It is no surprise that Random Forests win many machine learning competitions.

2.4 Data Mining Tasks

According to Han and Kamber [12] Data mining tasks are used to specify the kind of patterns to be found in data mining tasks. Generally, data mining tasks are classified into two categories: descriptive and predictive.

A predictive model makes a prediction about values of data using known results found from different historical data. Prediction methods use existing variables to predict unknown or future values of other variables. The predictive mining perform inference to make predictions. It is constructed based on the analysis of the values of the other attributes or dimensions' describing the data objects (tuples) [12]. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is the process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to specific variable(s) you are trying to predict. It is a supervised learning because the classes are predefined before the examination of the target data. Predictive model includes classification, prediction, regression and time series analysis.

Descriptive mining tasks are characterized by generalizing properties of the data. A model is a high-level description, summarizing a large collection of data and describing its important features. Often a model is global in the sense that it applies to all points in the measurement space. The goal of a descriptive model is describing all of the data (or the process generating the data). Examples of such descriptions include models for the overall probability distribution of the data (density estimation), partitioning of the p-dimensional space into groups (cluster analysis and segmentation), and models describing the relationship between variables (dependency modeling).

According to Rokach [40], Descriptive data mining method can be defined as discovering interesting regularities in the data, to uncover patterns and find interesting subgroups in the bulk of data is normally used to generate frequency, cross tabulation and correlation. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the reasons for mobile call drops, not to predict new call drop reasons. Descriptive task encompasses methods such as Clustering, Association Rules, Summarizations and Sequence analysis. But data mining involves clustering and association rule discovery methods.

2.5 Data Mining Models

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

2.5.1 KDD Process Model

To analyze large amount of data, data mining came into picture and is also called as KDD process. To complete this process various techniques were developed. KDD will turn the low level data into high level data.

Knowledge discovery was coined as KDD to emphasize the fact that knowledge is the end product of a data-driven discovery and that it has been popularized in the artificial intelligence and machine learning fields. KDD refers to the overall process of discovering useful knowledge from data and data mining referring to a particular step in the process. Furthermore, data mining is considered as the application of specific algorithms for extracting patterns from data.

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

Nowadays, massive amount of data is produced and collected incrementally. The possibility of gathering and storing huge amount of data by different organizations is becoming true because of using fast and less expensive computers. When organizational data bases keep growing in number and size due to the availability of powerful and affordable database systems the need for new techniques and tools became very important. These tools are used for helping humans to automatically identify patterns, transform the processed data into meaning full information in order to draw concrete conclusions. In addition, it helps in extraction of hidden knowledge from huge amount of digital data [35].

In the private sector industries such as banking, insurance, medicine, telecommunication and retailing data mining is used to reduce costs, enhance research, and increase sales. Different organizations worldwide use data mining techniques for applying and locating higher value customers and to reconfigure their product offerings to increase sales. In the public sector, data mining applications initially were used as a means for detecting fraud and waste of materials, but it grown for different purposes such as measuring and improving program performance [34].

Generally, there are five steps in the KDD process (Two Crows Corporation 1999; Azevedo and Santos 2008):

1. Data selection: This stage consists on creating a target dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed. The data relevant to the analysis is decided on and retrieved from the data collection.
2. Data pre-processing: This stage consists on the target data cleaning and preprocessing in order to obtain consistent data
3. Data transformation: It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure. This stage consists on the transformation of the data using dimensionality reduction or transformation methods

4. Data mining: It is the crucial step in which clever techniques are applied to extract potentially useful patterns. It consists on the searching for patterns of interest in a particular representational form, depending on the DM objective.
5. Interpretation/Evaluation: This stage consists on the interpretation and evaluation of the mined patterns.

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

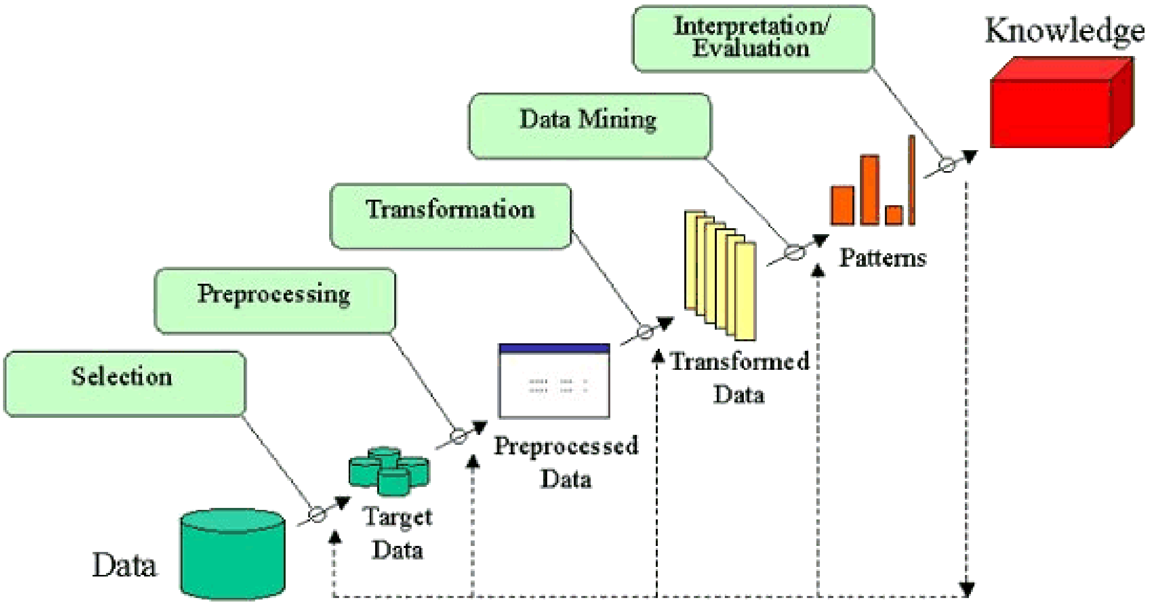


Figure 0-1KDD process

2.5.2 CRISP-DM Process

CRISP-DM (CRoss-Industry Standard Process for Data Mining) is a data mining project compromises a multi-step, iterative process. It consists on a cycle that comprises six stages (Chapman et al, 2000; Azevedo & Santos, 2008).

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

2. Data understanding- the data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
3. Data preparation- the data preparation phase covers all activities to construct the final dataset from the initial raw data.
4. Modeling- in this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
5. Evaluation- at this stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives.
6. Deployment- creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized presented in a way that the customer can use it.

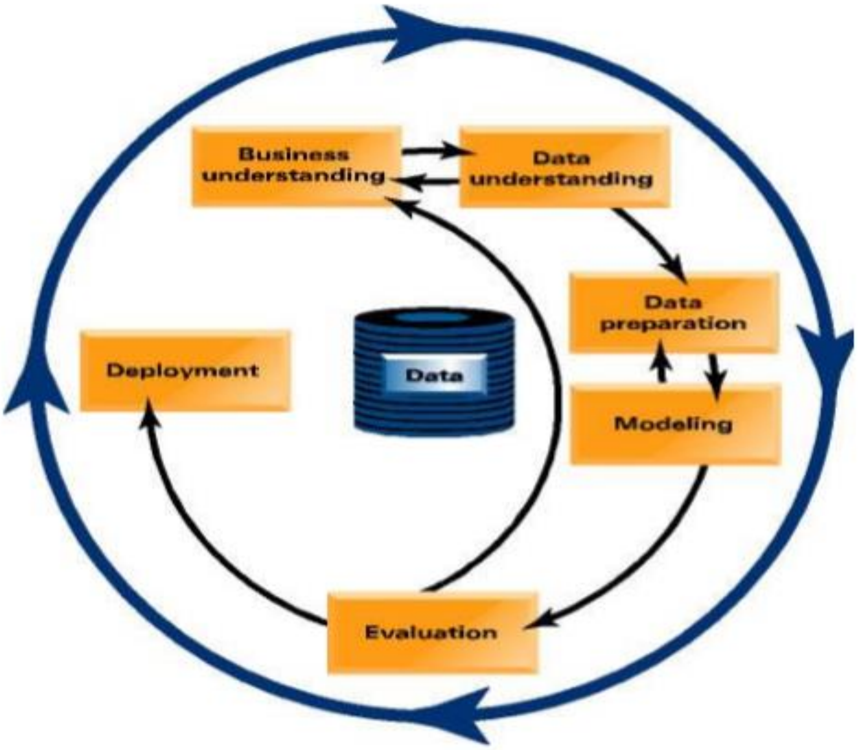


Figure 0-2 The CRISP-DM process

2.5.3 Cios model

The six step Cios model was developed, by adopting the CRISP-DM model to the needs of academic research community. The model consists of six steps [19].

1. Understanding of the problem domain: In this step one works closely with domain experts to define the problem and determine the research goals, identify key people, and learns about current solutions to the problem. A description of the problem including its restrictions is done. The research goals then need to be translated into the DM goals, and include initial selection of the DM tools.
2. Understanding of the data: This step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the DM goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.
3. Preparation of the data: This is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire research effort. In this step, which data will be used as input for DM tools of step 4, is decided. It may involve sampling of data, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be, further processed by feature selection and extraction algorithms (to reduce dimensionality), and by derivation of new attributes (say by discretization). The result would be new data records, meeting specific input requirements for the planned to be used DM tools.
4. Data mining: This is another key step in the knowledge discovery process. Although it is the DM tools that discover new information, their application usually takes less time than data preparation. This step involves usage of the planned DM tools and selection of the new ones. DM tools include many types of algorithms, such as neural networks, clustering, preprocessing techniques, Bayesian methods, machine learning, etc. This step involves the use of several DM tools on data prepared in step 3. First, the training and testing procedures are designed and the data model is constructed using one of the chosen DM tools; the generated data model is verified by using testing procedures.
5. Evaluation of the discovered knowledge: This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only

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

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

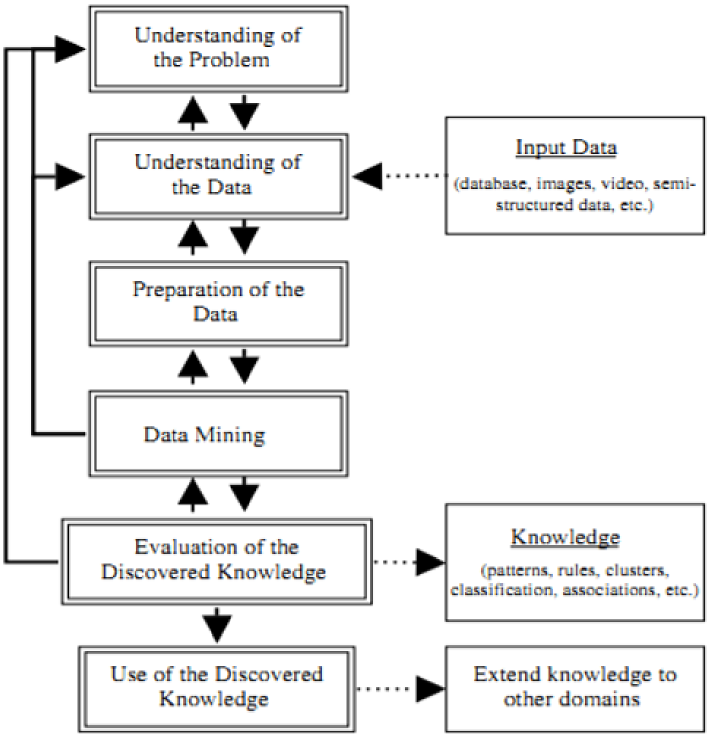

Understanding of the Problem
Understanding of the Data
Input Data
(database, images, video, semi-structured data, etc.)
Preparation of the Data
Data Mining
Evaluation of the Discovered Knowledge
Knowledge
(patterns, rules, clusters, classification, associations, etc.)
Use of the Discovered Knowledge
Extend knowledge to other domains

Figure 0-3 Cios model

The review of literature about the models of Data Mining shows that there are high similarities in the models. The unique part of the KDD process model is that it does not include business and data understanding. Thus, in this model there is no involvement of domain experts. But this study intended to include opinion of domain expert. Therefore, the KDD process model is not relevant for the study.

Different from the KDD model, the CRISP model includes business and data understanding. But under this model, the data mining project compromises a multi-step, iterative process. The main limitation of this model with regard to this study is including the process of the deployment. To deploy result of the data mining, Ministry of Revenue and tax administration offices have no formally established system. Therefore, the study cannot have the deployment stage.

Thus, this study has used followed the Cios Model that include business and data understanding and exclude the deployment of new system. Therefore, the study follows this model for the study purpose.

2.6 Related Works

Roux et.al (2018) indicated as 53% tax payers in Ghana involve in tax undervaluation. The study has used five year annual audited reports for 54,512 cases. The study has used an unsupervised learning techniques method for the detection of potential fraudulent tax payers by allowing the future use of supervised learning techniques. As a data mining tool, WEKA was used. PCA was sued to reduce attributes of the tax payers and selected five attributes; business type, amount of tax, size of the business based on amount of capital held, duration of tax payer in business and type of ownership. The model identifies under-reporting taxpayers on real tax payment declarations, reducing the number of potential fraudulent tax payers to audit. The study has implemented classification and clustering methods. As the classification method decision tree induction and rule based methods were implemented and the clustering was conducted based on hierarchical and partitioning mechanisms. Further, the study has implemented J48 algorithm. The obtained results demonstrate that the model doesn't miss on marking declarations as suspicious and labels previously undetected tax declarations as suspicious, increasing the operational efficiency in the tax supervision process without needing historic labeled data [11]. This study failed to include important attributes while conducting the data mining especially the characteristics of the tax payers.

Correa, Aouada, Stojanovic, and Ottersten (2016) has identified tax fraudulent behavior of large scaled companies in Latvia. The study has indicated tax evasion is highly common in large companies than small scaled companies in the country. The study indicated that 18% of tax is not collected because of tax evasion. For the study 2,981 large scaled companies were used. The study has used WEKA data mining tool. Neural networks and Bayesian classification algorithms were implemented for the tax fraud detection. The study has selected only four attributes of companies that are mainly focused on only financial behavior indicators. These attributes include asset, capital, current year capital expense and amount of liabilities. The study revealed that tax fraudulent behavior is related with financial capability of the companies [10]. This study was conducted in the country where there is high variation of characteristics of the tax payers. in addition, the revenue administration of the country has implemented strong system for tax evasion follow-up.

Wu et.al (2018) applied a data mining technique to enhance tax evasion detection performance in selected tax offices in Vietnam where the country lost 100 million dollars. 1,547 datasets were used for the study. The study has implemented python for the data mining. Using a data mining technique, a screening framework is developed to filter possible non-compliant value-added tax (VAT) reports that may be subject to further auditing. The study has implemented decision tree for classification purpose and hierarchical method for clustering. The results show that the proposed data mining technique truly enhances the detection of tax evasion, and therefore can be employed to effectively reduce or minimize losses from VAT evasion [4]. This study has focused only VAT payment associated factors. It failed to include other tax types in the study.

Dahee and Kyungho (2018) discussed a system using machine learning and artificial neural networks approach to detect fraud and process large amounts of financial data for 2,781 cases in Senegal small scaled companies where tax evasion is common practice and data mining was conducted by using python programing. The class imbalance problem was addressed and the usage of Synthetic Minority Oversampling Technique (SMOTE) and Random Under sampling (RUS) was applied. Data reduction was conducted by principal components method and five attributes of the tax payers were selected. The study has implemented J48 algorithm and indicated that data mining process can identify the tax fraud detection in the revenue authority [11]. This study has followed only J48 algorithm and it failed to include other algorithms that might result on better accuracy. In addition, the study was conducted only in the small scaled companies. Therefore, this stud has failed to include all relevant companies.

Diwakar et al. (2017) identified tax fraud in tax administration in Nigeria and mined tax payers' data by using association rule mining or APRIORI algorithm. The study indicated the country loses 25% of collectable tax due tax evasions. The study has sampled 6,345 cases for data training and experiment. WEKA data mining tool was implemented for experimenting. The frequent item set and the phish tank database are analyzed for fraud patterns. The frequent item sets are analyzed from the characteristics of tax payers and the anomalous patterns or outliers are found out based on the classification method used. The experiment was conducted by using hybrid algorithm from J48 and Bayes Naïve algorithms. The study has indicated that the tax evasion behavior of the companies is associated with amount of tax to be paid and their capital [12]. The main limitation of this study is excluding important attributes for the classification of the tax payers.

Shoubin Dong (2017) used data mining process to identify tax evasion practice in selected sectors in India. The study was conducted by using tax payers in manufacturing and construction sector where 12% and 18% companies involve in tax misspecification respectively. In the study 34,314 cases where included from both sectors. The data was collected from government revenue administration authority. The study has investigated behavior of the tax payers by using R-programing. Based on the classification and clustering mechanisms, 8 attributes were identified for data experiment. The attributes included in the experiment were time of establishment of the company, ownership of the company, debt history of the company, previously unpaid amount of tax, sector of the business, current tax amount, capital of the company and state where the company operates. In the study a novel bipartite graph-based propagation approach is adopted for fraud detection in in the tax payment system. The experiment was conducted by using J48 algorithm. The fraud detection problem in tax payment is analyzed to detect fraudulent tax payers and introduce the initial score learning model to a large tax payer by using bipartite graph propagation method for fraud detection. With the careful investigation of behavior patterns of tax payers, two key characteristics are identified: amount of payment and sector of the business [12].

Tian et.al (2016) identified fraudulent behavior tax evading companies in Bangladesh. The study has used large data from Ministry of Revenue and included 78,348 cases that included the audit report of 3 years and the data analytics was conducted by using python tool. The study investigated the classic tax evasion cases by employing a graph-based method to characterize their property that describes two suspicious relationship trails with a same antecedent node behind an Interest-Affiliated Transaction (IAT). Colored Network-Based Model (CNBM) was proposed for characterizing economic behaviors, social relationships, and the IATs between taxpayers, and generating a Taxpayer Interest Interacted Network (TPIIN). To accomplish the tax evasion detection task by discovering suspicious groups in a TPIIN, methods for building a patterns tree and matching component patterns were introduced and the completeness of the methods based on graph theory was presented. Then, an experiment based on real data and a simulated network was described. The experimental results show that the proposed method greatly improves the efficiency of tax evasion detection, as well as provides a clear explanation of the tax evasion behaviors of taxpayer groups [8].

Daniel (2013) [10] explored applicability of the data mining technology to develop models that can detect and predict fraud suspicious in tax claims at Ethiopian Revenue and Custom Authority (ERCA) by using total of 11,080 records of tax payers that were collected from ERCA and extracted from ASYCUDA database. The study has used characteristics of the tax payers as attributes that include Gross profit/loss, Net worth, Interest expenses, Net income/loss, Net tax due/Refundable amount, Liquid Cash, Non-operating income, Profit income tax, Repair and Maintenance expenses, Selling and Distribution expenses, Depreciation expenses, Net book value, Total expenses, and Total gross income. The study has applied clustering algorithm and classification techniques to develop predictive model. K-Means clustering algorithm was employed to cluster different tax claims as fraud and non-fraud. The classification was carried out by using J48 decision tree and Naïve Bayes algorithms. The experiments have been conducted following the six-step KDD process model. The model developed using the J48 decision tree algorithm and the study has shown that data mining techniques are valuable for tax fraud detection. But the study failed to include important attributes such as size of company and business activity that might be associated with tax evasions.

The review about the related works indicate that most of the studies are not relevant for administration practice Ethiopia and behavior of the tax payers. In addition, the studies have omitted very important attributes explaining tax evasion behavior of the tax payers. Therefore, this study includes details of tax payers that are unique to the tax payers.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1. Introduction

This study intends to identify tax evasion by companies in Addis Ababa by using data mining procedure. This chapter presents the methodology followed for achieving this objective. The first section of the chapter presents research design and is followed by the data mining process model employed for the study. The data mining process consists understanding the problem domain, data understanding, data preparation and pre-processing, modeling, evaluation of the discovered knowledge, and using discovered knowledge.

3.2. Research Design

This study was conducted by using secondary data from revenue authority about the tax payers in Addis Ababa. Thus, the research followed experimental research approach. Therefore, the experiment was conducted by using data of tax payers. The data mining process followed by this study was carried out following two-step process; clustering then classification data mining approaches. According to Koh and Gervais (2010) clustering and classification are mostly used data mining techniques for fraud detection. Clustering groups a set of physical or abstract objects into classes of similar objects. Thus, similar objects are collected within the same cluster and dissimilar objects are located at another cluster. Commonly, similarity is indicated by ‘how close’ the objects are in space based on distance function. Therefore, this study has used clustering to form similar groups of tax evading companies. The classification was applied to predict the occurrence of tax evasion. Both the clustering and the classification tasks were applied to training dataset.

In data mining, different methods are used. Han and Kamber (2006) classifies the algorithms as partitioning, hierarchal, Density-based, Grid-based, and model-based methods and expresses partitioning method is most commonly used method and conducted by using K-mean algorithm. K-mean algorithm has advantage over other clustering methods because it is applicable mean of cluster for numeric dataset. It is easiest algorithm and scalable and efficient in processing large datasets. Based on these importance, the clustering was conducted using the K-means algorithm.

There are various types of data mining process models in relation to the research approach selected. This study adopted the Cios model that was developed based on the CRISP-DM model by adopting it to academic research. The Cios model because it is advantageous over other models since it hybrids the academic and industrial purposes. Mainly, this model includes business and data understanding. In the study inclusion of domain experts is considered as main strategy while conducted the study.

3.3 Understanding the Problem Domain:

This study used data of the tax payers to predict the tax evasion. Therefore, the study used the secondary data from revenue authority at branch offices of tax payers. Further, interview was conducted with domain experts that are tax audit experts and managers.

In Ethiopia, it is estimated by tax authority that annually about 11.4 billion birr is lost due to tax evasion. Although tax evasion is practiced by businesses throughout the country, about 72% of it is committed by companies in Addis Ababa. The report of the authority shows that almost of 50% of tax payers commit tax shielding. Amount of tax evasion in the country is equivalent to budget of big public organizations. This suggests that the country is losing big investment due to tax evasion. The problem becomes very high due to behavior of tax payers and lack of appropriate system to detect the problem [6]. Currently, the tax authority is mainly using manual auditing and inspection for tax evasion that makes detection of tax fraud challenging and the audit finding unreliable. This implies existence of serious tax fraud and high importance of automated detection for the tax evasion problem.

Continued tax evasion results on undermining public investment and negatively affecting economy as whole. Thus, detecting tax fraud is one of the main priorities of tax authorities which are required to develop cost-efficient strategies to tackle the problem. Data mining may be effective tools for enhancing the efficiency and effectiveness of the detection of illegal tax evasion.

3.4 Data Understanding

Data understanding tasks should be carried out carefully to come up with good output in data mining. The reason is that the models that will be built mainly depend on these tasks. To understand the tax evasion, the researcher has reviewed reports of the revenue authority and interviewed experts in tax audit and management after explaining the objective of the study.

After understanding the problem to be addressed, the researcher conducted the process of understanding the data. This includes listing out attributes with their respective values and evaluation of their importance for this research and careful analysis of the data and its structure which was supported by tax professionals in the branch offices by evaluating the relationships of the data with the problem at hand and the particular DM tasks to be performed. Finally, the researcher verified the usefulness of the data with respect to the data mining objectives.

The information about the tax payers is stored in inspection department in each tax payer's branch office. It is stored in Excel format which is coded by the branch offices based on the audit finding. The information is mainly about the tax relevant items; income statement in addition to some descriptive attributes about the tax payers. The dataset extracted from the tax authority include 19 attributes that are further used for attribute selection by the study. These attributes in the original dataset include serial number of the tax payer, name of the company, TIN, location of the business, sector of the business, type of ownership, capital of the company, liability of the company, sales/turnover, operating expense, interest expense, repair expense, selling expense, depreciation expense, gross income, non-operating income, net tax due, profit tax, and net profit.

Different studies conducted in different area have used various attributes in predicting tax evasion. This variation in attribute inclusion is based on existence of the attribute in the original dataset. This suggests importance of the attribute varies from area to area. This study tries to include attributes used by different studies and they exist in the dataset of tax authority in Ethiopia that the organization considered they are important in managing tax evasion. Roux et.al (2018) selected attributes such as business type, amount of tax, size of the business based on amount of capital held, duration of tax payer in business and type of ownership; Correa, Aouada, Stojanovic, and Ottersten (2016) included asset, capital, current year capital expense and amount of liabilities; Shoubin Dong (2017) in predicting tax evasion used attributes such as time of establishment of the company, ownership of the company, debt history of the company, previously unpaid amount of tax, sector of the business, current tax amount, capital of the company and state where the company operates. In Ethiopia, study conducted by Daniel [10] used Gross profit/loss, Net worth, Interest expenses, Net income/loss, Net tax due/Refundable amount, Liquid Cash, Non-operating income, Profit income tax, Repair and Maintenance expenses, Selling and Distribution expenses, Depreciation expenses, Net book value, Total expenses, and Total gross income as an important attribute in prediction of tax evasion.

Description about these attributes is presented in Table 3.1 below. Table 3.1 presents about name of attribute, description and type of data.

Table 3. 1 Description about Attributes in Raw Data

Name of attribute	Description	Type of data
Sr.No	Serial number of the company in the audit list	Numeric
Name_company	Name of the tax paying company. it is trade name of the company	nominal
TIN	Tax payer's identification number	Numeric
Location	Geographic location where the company exists	nominal
Sector	Sector of the business of company	Nominal
Ownership	The type of ownership of the company (sole proprietorship, partnership and share company)	nominal
Capital_comp	Amount of the capital the company owns in value of Birr	numeric
Liability_comp	The amount of liability of company in value of Birr	numeric
Sales	Annual sales of the company in value of Birr	numeric
Operational_expense	The annual operational expense of the company in value of Birr	numeric
Interest_expense	The amount of annual interest expense of the company in value of Birr	numeric
Repair_expense	The Birr value of repair expense paid by the company annually	Numeric
Selling_expense	The total value Birr paid for selling activities by the company	Numeric
Depreciation_expense	The value of annual depreciation of assets of the company	Numeric
Nonoperating_income	The value of income collected from out of business activity	Numeric
Gross_profit	Total profit of the company before tax (in value of Birr)	Numeric
net tax due	The amount of tax refundable to the tax payer	Numeric
Profit tax	The value of amount of tax payable by the company	Numeric
Net profit	The amount of profit after tax deduction	Numeric

Source: Tax Payer's Branch Offices, 2020

3.5 Data Preparation & Preprocessing

Data Selection:

The study used audited financial reports of tax payers from small tax payers to large tax payers in Addis Ababa. There are 7,232 reports of tax audit. Therefore, this study used data set of 6,142 records for training and 1,090 records for testing purpose in conducting the study experiment.

The dataset includes also about the audit finding; amount of the tax fraud. The amount of appropriate tax and evaded tax. The companies were split into two; paid and evaded. Therefore, the records were split into two classes of tax evasion and no-evasion. But this study focuses on identifying model of only tax evasion.

Based on the available report, majority of the business companies do not involve in tax evasion when compared to business companies that evade tax [6]. The study used Synthetic Minority Oversampling Technique (SMOTE) to reduce the majority class in the dataset because the majority class can make the classification more directed to the majority class and the predictions biased. The number of companies that did not evade tax will be selected based on the number of companies that evade tax. After setting the number of companies that have no tax evasion report, the study used simple random sampling method to select the companies.

Data Cleaning

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

Among the variables presented in the dataset the study excluded attributes unique to the tax payers and similar to all tax payers. The variables unique to the tax payer include serial number, name of the company and TIN. As all companies selected are located in Addis Ababa and have similar values, the study didn’t include location of the companies.

Another important point in data cleaning is handling inconsistent data. The inconsistency of the data was identified with support of domain experts. This process has shown that there are no data represented in different ways that the organization uses uniform methods while handling the data. Thus, there is no duplicate attribute. Finally, the study has checked existence of noise data (outlier). The existence of outlier is identified with domain experts and no noisy data was detected. In addition, detection of noisy data was conducted by using WEKA. Based on this method also there were no noisy records identified. Therefore, all the records used for the study were used because they have records for removal because of noisy data.

Based on the data cleaning procedure, 8 attributes were selected to be used in the experiment. This attributes include sector of the duration in a business, ownership type, capital of the company, liability of the company, revenue, expense, receivables, and tax amount.

Data Transformation and Data Reduction

In data transformation, the data are transformed or consolidated into forms appropriate for mining. In the dataset, majority of the variables include continuous attributes. Thus, attributes were discretized (binned) to reduce the distinct values of the attributes so that it suites the mining tool and to obtain meaningful patterns. 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. 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 concise, easy to use, knowledge-level representation of mining results (Han & Kamber, 2006).

Discretization of continuous variable is presented in Table 3.2 below.

Table 3. 2 Attribute Discretization labels

Attribute			
Duration	Below 11	11-20	Above 20
	Low	Moderate	Old
Liability	≤1,000,000	1,000,000.01-10,000,000	≥10,000,000.01
	Low	Medium	High
Capital	≤1,000,000	1,000,000.01-10,000,000	≥10,000,000.01
	Small	Moderate	Large
Receivables	≤1,000,000	1,000,000.01-10,000,000	≥10,000,000.01
	Small	Moderate	Large
Revenue	≤1,000,000	1,000,000.01-10,000,000	≥10,000,000.01
	Low	Medium	High
Expense	≤1,000,000	1,000,000.01-10,000,000	≥10,000,000.01
	Low	Medium	High
Tax	≤100,000	100,000-500,000	≥500,000
	Small	Medium	Large

Source: Domain Experts, 2020

As shown in Table 3.2 above, nominal variables such as sector of the business and type of ownership are used with their original values but all continues variables are transformed to three discrete values; low, medium and high. These variables include liability, capital, revenue, expense, receivables, duration in business and amount of tax.

Attribute selection

The dataset includes number of attributes that include relevant and irrelevant variables for the study. Some important variables have large number of missing data and thus they were removed from the dataset. In addition, there were irrelevant variables for the study. Consequently, they are removed from the study data. Finally, the study has identified 8 important variables but the final version of the study attributes was selected by using WEKA data mining tool based on Principal Component Analysis (PCA). From these set of attributes from the initial dataset, information content of the attributes selected was evaluated and the final attributes selection was conducted with domain experts. The data mining tool selected all recommended variables by the study. Therefore, there are attributes removed from the recommended variables. The list of attributes for final attribute selection included in the dataset is about ownership, capital of the company, revenue, expense, tax, capital, liability and tax.

Data formatting

The datasets provided to WEKA software was prepared in a format that is acceptable for the software. Weka accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute-Relation File Format) (a file name with an extension of ARFF i.e. FileName.arff). To feed the final dataset into the Weka, the file was changed into other file format. The excel file was first changed into a comma-separated values(CSV) file format. After changing the dataset into a CSV format the next step was opening the file with the Weka DM software. Then this file was saved with ARFF (Attribute Relation File Format) file extension.

3.6 Model Building

The two primary goals of data mining tend to be prediction and description. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data mining activities into one of two categories: Predictive data mining, which produces the model of the system described by the given data set, or Descriptive data mining, which produces new, nontrivial information based on the available data set. This study will follow predictive modeling to produce model from the dataset.

Since this study intends to build a predictive model, the models intended to be addressed are classification models. This includes data mining tool selection and the algorithms used for modeling technique. The classification modeling technique has used the clustered dataset as an input and implemented using ensemble methods and mainly follows random forest algorithm that random forest models implement a level of differentiation based on different features rather than splitting at similar features at each node.

After the data was cleaned and prepared, it was analyzed using a data mining tool. There are different tools for data mining. This study used Weka data mining tool since the whole suite of Weka is written in java, so it can be run on any platform. In addition to this, the package has three different interfaces: a command line interface, an Explorer GUI interface which allows for preparation, transformation and modeling algorithms on a dataset, and an Experimenter GUI interface which allows to run different algorithms in batch and to compare the results.

Models were built by using the Weka data mining software version 3.8.4, and the proposed models were tested using test sets of data.

3.7 Evaluation of the Discovered Knowledge

The validity and performance of the model was tested to check its efficiency and effectiveness. The effectiveness and efficiency of the model is also computed in terms of recall (sensitivity) and precision (specificity). Confusion matrix was used to evaluate the accuracy and performance (time taken) of the model built with the decision tree algorithm. In addition, MAE, RMSE, and ROC curve analysis was used. The interpretation of the results was supported by the domain experts.

3.8 Using Discovered Knowledge

Integrating the newly developed model to the existing system was the main activity of data mining tasks. As it was shown in section 1.2, the company has manual auditing method. Therefore, the results of this study was used in supporting the manual auditing activities in the organization. Currently, the organization has no system for auditing but attributes of the tax payers are recorded in excel after tax audit. Since the organization have no auditing system, discovered knowledge will support the decision making in audit support. The organization uses the association rules identified by the study. Further, the study has developed classification model by using the classification algorithms.

CHAPTER FOUR

EXPERIMENTATION

4.1 Introduction

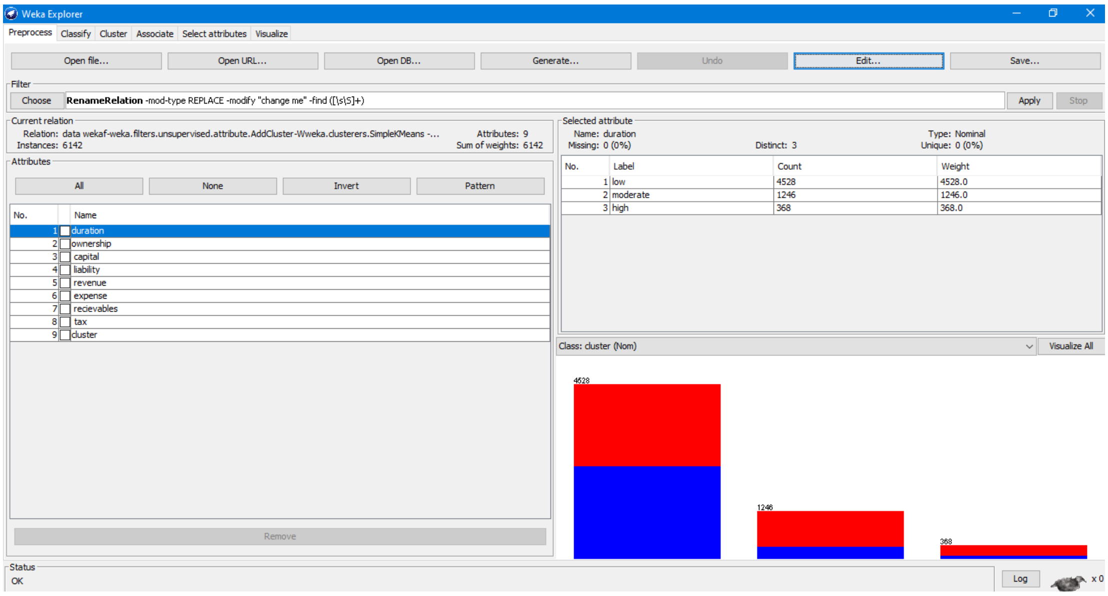
This study was conducted with an objective of developing predictive model for detection of tax evasion practices in Ministry of Revenue of Ethiopia. To achieve this objective the study followed data mining strategy. In line with the research objective, this chapter presented about procedure followed while conducting experiment for the data mining. The chapter discusses the result of experimentation about clustering, classification and model evaluation. Finally, discussion about the discovered knowledge was presented.

The study has used dataset of 7,232 instances that indicates audit report by the ministry of revenue. Among these instances, 85% (6,142) instances were used for the model training and the 15% (1090) instances were used for testing the model. The instances for the testing were randomly selected. The data mining process was conducted by using Weka data mining tool version 3.8.4 for Windows. For clustering the study used simple K-means clustering algorithm and 3 experiments were conducted. On the other hand for the classification, the study implemented different algorithms such as J48 decision tree, NaïveBayes, Multilayer perceptron (MLP) functions, and random forest tree algorithms. The best algorithm was selected based on classification accuracy.

4.2 Model Building

Model building process followed two steps modeling approach; cluster modeling and classification modeling. Thus, the study applied cluster modeling and classification modeling to conduct data mining from attributes of tax payers for identification of tax evasion behavior. The cluster modeling is intended to form segment of tax payers with tax evasion behavior and non-evasion behavior. On the other hand, the classification modeling is conducted to build tax evasion detecting model and form association from the attributes suggested.

Model building process of the study is conducted by using 6,142 training dataset. Based on the clustering strategy, clustered dataset is developed and used for training of classification model. The selection of clusters is decided based on intra-class similarity and judgment of domain experts. The model training for classification model development was conducted by using 10-fold cross-validation and percentage split. To test the prediction performance of the classification model developed, separately prepared testing dataset was used. The testing used 1,090 instances that are randomly selected from the original dataset. The study has used Weka version 3.8.4 for Windows 10 to conduct the process of data mining. This section of the study presented the result of experiment of clustering model and classification modeling. The preprocessing dataset in Weka is presented in figure 4.1 below for sample snapshot.



Filter: **RenameRelation** -mod-type REPLACE -modify "change me" -find ([\s\S]+)

Relation: data wekaf-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -...
Instances: 6142
Attributes: 9
Sum of weights: 6142

No.		Name
1	□	duration
2	□	ownership
3	□	capital
4	□	liability
5	□	revenue
6	□	expense
7	□	recievables
8	□	tax
9	□	cluster

Name: duration
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	low	4528	4528.0
2	moderate	1246	1246.0
3	high	368	368.0

Class: cluster (Nom)

Status
OK

Figure 0-1Preprocessing Data in Weka 3.8.4 (Source: Result Weka, 2020)

The study identified 3 labels for all attributes. The first attribute is duration of the companies in business activity and the first category of the attribute comprises 4,528 cases, the second category comprises 1,246 cases and the third category comprises 368 cases. This implies the number of cases at higher categories decreased suggesting that few companies have high experience (above 20 years) in their business.

4.2.1 Cluster Modeling

In line with the general objective, this study has constructed different specific objectives. Among these objectives, identifying the behavior of the tax evading businesses is among the most important objectives. Based on previous studies, behavior of the tax evading companies is indicated by different attributes. But some companies share the same attribute whether they involve in tax evasion or not. The researcher found that it is important to systematically identify the behavior of majority of companies that involve in tax fraud. This study has conducted experiments to develop best clustering model and the best model is used to develop the classification model. This section of the study presents the results of experiments conducted to develop clustering model.

The clustering experiments were conducted by using K-mean clustering algorithms. The experiments were conducted by altering distance function and seed size. The study has intended to classify the behavior of the tax payers into two; evading and non-evading. Thus, the K-value set to be 2 suggesting tax evasion suspicious and non-suspicious tax payers. Although the level of tax evasion varies, for the purpose of simplicity, the study has used only two clusters.

In addition to computational simplicity, domain experts were consulted in the authority and suggested that it is better to classify the tax payers into two; tax evaders and non-evaders that the audit finding reports whether the company involve in tax evading or not. Therefore, the researcher found that it is appropriate to use K = 2 (representing evading and non-evading companies). For the clustering result decision, the study has used three criterion; intra cluster similarity measure, number of iteration to conduct a convergence, and judgment of domain expert. The intra cluster similarity is measured by within cluster sum of squared error (the lower is better).

As presented in previous chapter, the study has finally selected 8 attributes. In addition, each attribute is represented in 3 labels. The attributes are mainly about financial level of the tax payers. As indicated by the domain experts, the tax payers mainly depend on these attributes while involving in tax evasion activity.

Experiment I

The first experiment was conducted by using the default values of the data mining tool; K = 2, EuclideanDistance as distance function and seed value of seed = 10. The result of experiment I is summarized in Table 4.1 below.

Table 4. 1 Cluster Modeling Experiment I

		Cluster 1	Cluster 2
Instances		2812 (46%)	3330 (54%)
Attributes	Duration	Low	Low
	Ownership	PLC	PLC
	capital	Medium	Large
	liability	Low	Moderate
	Revenue	High	High
	Expense	Medium	High
	Receivable	Small	Moderate
	Tax	Medium	Small
Rank		2	1

Source: Weka Clustering Result, 2020

Table 4.1 above shows the result of first experiment conducted to develop clustering model and presents instances in each cluster, attributes in the clusters and ranks of the clusters in explaining the tax evasion suspicious behavior of the tax payers. As shown in the Table 4.1 above, 2,812 (46%) of the instances are grouped in Cluster 1 and 3,330 (54%) of the instances are grouped in Cluster 2. The result of experiment about classification of attributes in each cluster shows that Cluster 1 is indicated by recently established companies, private limited companies, moderate capital amount holding, small amount of liability, high revenue collection, medium expense amount, small amount of receivables, and medium tax payers. On the other hand, the Cluster 2 represent attribute of tax payers that recently established, private limited companies, moderate liabilities, high revenue, high expense amount, moderate receivables and small tax payers. The result about attributes in the clusters suggests some attributes are commonly existing in both clusters. This attributes include duration in the business, ownership type of the companies, and revenue amount. But other attributes are differentiated in each cluster.

The cluster modeling is conducted to form clusters that indicate tax evasion and no tax evasion. The clusters created are used as dependent variable while the attributes are used as independent variables. Thus, conducting cluster modeling is mainly used to create input for classification model. Therefore, it is necessary to run cluster modeling that the development of final predictive model is conducted by using result of the cluster modeling.

To classify the clusters to the tax evading and non-evading group of companies based on the clustering result from the experiment, previous literatures and judgment of the domain expert are used. As shown in previous chapter, tax evasion behaviors are mainly indicated by the group of tax payer (small, medium, large), amount of expense reported and liability of the company. The judgment of the domain expert shows that small tax payers have high practice of tax evasion because the companies have low structured system that indicate inadequate income and expense reporting and they do not want to upgrade their group. In addition, the tax evading companies report exaggerated expense to lower the taxable amount of income. Further, the tax evading companies report higher liability than the actual to pay their liabilities in short period of time by reducing the tax payable. Thus, based on the justifications provided by the domain experts, Cluster 2 is better than Cluster 1 in describing tax evading behaviors. Consequently, Cluster 2 is ranked 1st and considered as behaviors of tax evasion.

The quality of this experiment is indicated that the instances are similar to report of the authority and suggestions of the domain experts. The report of the revenue authority indicated that about half of the audit reports indicate tax evasion. Based on this experiment, 54% of the cases are tax evasion suspicious. Therefore, the result of the experiment is closer to reality. But to explore other more realistic model than this model, additional experiments are conducted by altering the seed values (seed = 100) by holding other values to default and similar to this experiment.

Experiment II

This experiment is conducted to develop comparable clustering model in relation to model identified in Experiment I. similar to Experiment I, this experiment uses K = 2 and used EuclideanDistance as distance function. But the seed value is changed and implemented seed = 100. The result of Experiment II is summarized in Table 4.2 below and presents instances and values of attributes in the clusters and ranking of clusters in representing tax evasion behaviors.

Table 4. 2 Cluster Modeling Result of Experiment II

		Cluster 1	Cluster 2
Instances		4272 (70%)	1870 (30%)
Attributes	Duration	Low	Low
	Ownership	Plc	Plc
	Capital	Large	Medium
	Liability	Moderate	Moderate
	Revenue	High	Medium
	Expense	High	Medium
	Receivable	Moderate	Moderate
	Tax	Medium	Small
Rank		2	1

Source: Weka Result, 2020

As shown in Table 4.2 above, 4272 (70%) of the instances are clustered in Cluster 1 and the remaining 1,870 (30%) of the instances are clustered in Cluster 2.

Regarding the classification of the attributes, the result of this experiment shows that 4 attributes are not differentiated within the clusters and 4 attributes are adequately differentiated within the clusters. For the values of duration in business, ownership type of the business, amount of liability, and amount of receivables the clusters have similar values. On the other hand, attributes such as capital holding, revenue collected, amount of expense and group of tax payer are differentiated in the clusters. In both classes low duration in business, private limited companies, moderate liability amount and moderate receivables are reported. Regarding the differentiated attributes, large capital amount, high revenue, high expense and medium group of tax payers are reported in Cluster 1 and medium capital amount, medium revenue, medium expense and small group of tax payers are reported in the Cluster 2. Therefore, ranking is made by using capital, revenue, expense and group of tax payers.

According to the judgment of the domain experts, smaller companies engage in tax evasion than larger companies because smaller companies have more resource problems (including financial resources) than large companies. As a result, small companies report lower amount of tax than the actual amount when compared with large companies. This suggests Cluster 2 seems tax evading group of tax payers. In addition, lower revenue is reported to affect the amount of tax paid; low revenue, small tax. Thus, Cluster 2 suggests tax evasion. As shown in previous sections, expense is very important attribute to differentiate behavior of tax payers. Commonly, companies that report high amount of expense when other factors are constant, suggests practice of tax evasion. Based on this attribute the domain experts suggested Cluster 1 is suggesting tax evasion behavior. But among the important attributes to differentiate tax evasion is group of tax payers; small tax payers practiced more tax evasion than large or medium tax payers. This attribute suggests Cluster 2 represents tax evading companies. Based on the judgment of domain experts, three attributes suggest Cluster 2 and one attribute suggests Cluster 1 as a group of tax evading companies. Consequently, Cluster 2 is ranked 1st and Cluster 1 is ranked 2nd.

The weakness of this clustering model is the instances represented in the clusters highly varies from the reports of the authority and practices by the domain experts. The tax evading group comprises only 30% of the instances but 70% of the cases are non-evasion reports. This suggests the model is weak in clustering the tax evasion practices. Therefore, it is important to conduct additional experiment for comparison that represent the groups of tax evading and non-evading cases. As a result, third experiment is conducted by changing seed value and holding other procedures constant similar to Experiment I and II.

Experiment III

This experiment is conducted with K = 2, seed = 1000 and by using EuclideanDistance as distance function holding defaults of simple K means clustering method in the data mining tool. Result of the experiment is summarized in Table 4.3 below.

Table 4. 3 Cluster Modeling Experiment III Summary

		Cluster 1	Cluster 2
Instances		2030 (33%)	4112 (67%)
Attributes	Duration	Low	Low
	Ownership	SC	PLC
	capital	Medium	Medium
	liability	Moderate	Moderate
	Revenue	Medium	High
	Expense	Medium	High
	Receivable	Moderate	Moderate
	Tax	Medium	Medium
Rank		2	1

Source: Weka Result, 2020

As depicted in Table 4.3 above, 2,030 (33%) of the cases are grouped in Cluster 1 and 4,112 (67%) of the cases are grouped in Cluster 2.

As shown in Table 4.3 majority of the attributes were not differentiated in the created clusters. Among the attributes used 5 attributes have similar values in both clusters and the three attributes have differentiated values. Attributes such as duration of company in a business, amount of capital, liability, receivables and group of tax payers have the same values in both groups. On the other hand, ownership shown different values in the clusters that have similar values in the previous experiments. Domain experts have shown that private limited companies are engaged in more tax evasion than share companies because most of the time private limited companies are managed by owners or closer employees to the owners. But in the case of share companies, owners do not involve in the management and they do not involve in personal interest. Based on this attribute Cluster 2 seems tax evasion suspicious cluster. On the other hand, Cluster 1 includes group of cases with medium revenue and Cluster 2 includes high revenue cases. As suggested in previous sections, lower revenue reporting is behavior of tax evasion suspicious groups. As a result, Cluster 1 suggests tax evasion suspicious cases. The third important attribute differentiated in the clusters is amount of expense. As depicted in the Table 4.3 above, in the Cluster 2 high expense is indicated and in the Cluster 1 medium expense is indicated. This implies Cluster 2 suggests tax evasion suspicious instances.

Based on the values of attributes, in the experiment III, Cluster 2 is ranked 1st and Cluster 1 is ranked 2nd.

Regarding the relationship between number of instances and reports of revenue authority and opinion of domain experts, although the experiment III has shown better result than experiment II, the number of fraud cases are higher than the records of the authority. Experiment suggests larger cases than actual practices in the authority. This variation is accepted by domain experts and they suggested that there are tax evasions that are tracked during the audit. But the experts indicated the weakness of the model that attributes indicating the behavior of tax evasion status are weaker than the attributes indicated in the previous models in the Experiment I and Experiment II.

Comparison of Clustering Models

Three experiments were conducted to develop the cluster models by using simple K mean clustering algorithms. The experiments were conducted by using K = 2, EuclideanDistance as distance function and changing seed values. In the previous sections, the results of experiments were evaluated by domain experts and suggestions were provided. This section presents evaluation of the experiments and suggestion on best clustering model based on the clustering algorithm procedures. The performance of the best model is suggested based on number of iterations and within cluster sum pf squared errors. The performance measurement of within cluster sum of squared errors indicates intra and inter cluster similarity and it is main indicator of goodness of the clustering model. The lower values of within cluster sum of squared errors suggests good model than the higher values. Similarly, smaller number of iteration suggests better model and it indicates the algorithm has converged very soon.

The comparison of the models generated from the experiments conducted are summarized in Table 4.4 below based on the selected parameters.

Table 4. 4 Comparison of Clustering Models

Experimentation	Number of Iterations	Within cluster sum of squared errors
I	5	17627
II	3	18363
III	2	18570

Source: Weka Result, 2020

As shown in the Table 4.4 above, experiment I identified lowest number of within cluster sum of squared errors and highest number of iteration. In contrary to this, experiment III indicated highest number of within cluster sum of squared errors and lowest number of iterations. This indicates the model from experiment I is best clustering model but it has weak performance with regarding to convergence. In addition, the domain experts suggested as it is best model regarding to business problem. On the other hand, model from experiment III is best in convergence but it is the worst model. Similar to decision based on the within cluster sum of squared errors, the domain experts suggested it is bad model based scenario in the business problem. Therefore, this study reveals the model developed in the Experiment I is best clustering model and it is selected as final clustering model.

4.2.2 Classification

In the previous section, development procedure of clustering model is presented. Following development of the clustering model, the classification modeling is developed since the clustering model cannot classify new instances. The classification model analyzes accuracy of classifiers while categorizing the tax reporting into specified classes. This section of the study presents the result of classification modeling.

The classification modeling is conducted by using different algorithms that enable to choose the best classification model. This study has used tree algorithms (J48), Bayes algorithms (NaiveBayes), function algorithm (Multilayer Perceptron), and ensemble algorithms (random forest and bagging). To test performance of the classification models, separate testing dataset was used. The classification modeling has used attributes selected for the cluster model building as independent variables and the clusters built by clustering algorithms are used as dependent variable.

J48 algorithm

Different experimentations were conducted to identify best classification modeling. The first experiment was conducted by using J48 algorithm that build decision tree model. J48 algorithm contains some parameters that can be changed for further improvement of accuracy of classification. The first experiment is by using J48 algorithm is built with the default parameter values. The experiments were conducted by changing confidence factor used for pruning (confidence factor (CF)) and minimum number of instances per leaf (minNumObj(MNO)). In addition, the experimentation was conducted by changing the test options; 10-folds cross validation and 66% percentage split. The summary of result of J48 classification experiment is presented in Table 4.5 below (see Appendix 1: A – L) for confusion matrix of the experiments of J48 algorithm).

Table 4. 5 Summary of J48 classification experiments

Experiment	Tuned Parameters		Test Mode	Number Of Leaves	Size of Tree	Time Taken For Build (seconds)	Accuracy
	CF	MNO					
1	0.25	2	10-Fold	55	82	0.16	99.6744 %
			Percentage split	55	82	0.05	99.8084 %
2	0.5	2	10-Fold	55	82	0.05	99.7395%
			Percentage split	55	82	0.03	99.8084%
3	0.75	2	10-Fold	57	85	0.39	99.7232%
			Percentage split	57	85	0.44	99.8084%
4	0.25	5	10-Fold	49	73	0.03	99.4139 %
			Percentage split	49	73	0.03	99.2816 %
5	0.25	10	10-Fold	39	58	0.05	99.365 %
			Percentage split	39	58	0.03	99.2816 %
6	0.25	15	10-Fold	37	55	0.02	99.2185 %
			Percentage split	37	55	0.02	98.7548 %

Source: Weka Result, 2020

As shown in Table 4.5 above, 6 experiments were conducted by changing CF and MNO values by changing the test mode. As depicted in the table, 3 values of confidence factors (0.25, 0.5, and 0.75) and 4 values for MNO (2, 5, 10, and 15) were used. The first experiment was conducted by using default parameters of the data mining tool by changing test mode. Based on the test mode of 10-Fold cross validation, 55 leaves and 82 trees were generated. To build the classification model, it took only 0.16 seconds. The predicting accuracy of the model is 99.67%. In addition to the 10-fold cross validation test mode, 66% percentage spilt was conducted by holding the values of CF and MNO constant. This test mode has created the same number of trees and leaves. Other performances of this mode are better than 10-fold cross validation. The time taken to build the model is only 0.05 seconds and the accuracy is 99.81%. Therefore, in the experiment 1, percentage split predicts better than 10-fold cross validation.

The second experiment is conducted by changing the CF value to 0.5 and holding MNO value similar to experiment 1. The number of leaves and size of tree are similar to the experiment 1. But the time taken to build the model was improved for both 10-fold and percentage split test modes. Under the 10-fold cross validation, it took 0.05 seconds and under the percentage split it took 0.03 seconds to build the predicting model. The accuracy of 10-fold cross validation test mode is improved and increased to 99.74%. But there is no change in level of accuracy under the model under the percentage split mode.

Since accuracy of the model is increasing and time to build the model is decreasing, additional experiment was conducted by changing value of CF (CF = 0.75) and holding other parameters similar to the previous experiments. Based on this specification, the model has generated 57 leaves and 85 trees suggesting that this model is weaker than models built from previous experiments. In addition, the model has lower prediction accuracy than the previous models. The accuracy of this model become 99.72% in the 10-fold cross validation and time to build is 0.39 seconds. Under the percentage split, similar level of accuracy is generated but the time to build the model increased to 0.44 seconds.

Another three experiments were conducted by changing the value MNO by holding CF constant. The results of the experiments suggested that increasing the value of MNO reduced the number of leaves and size of the trees. In addition, time to build the models is improved. But the weakness of the models is reducing the prediction accuracy in comparison with the previous three models. This suggests increasing the MNO decrease the size of trees and number of leaves that makes the model simple.

The result of the experiments conducted based on J48 algorithm, the best model is conducted by using CF = 0.5 and MNO = 2. The summary of classification of the instances is presented in Table 4.6 below for test options of 10-fold cross validation and 66% percentage split.

Table 4. 6 Summary of Confusion Matrix of J48 algorithm

Test Option	Time taken	Classified Instances	
		Correctly classified	Incorrectly classified
Cross validation	0.05	6126 (99.74%)	16 (0.26 %)
Percentage split	0.03	2084 (99.81%)	4 (0.19 %)

Source: Weka Result, 2020

As shown in Table 4.6 above, model developed based on cross-validation took 0.05 seconds and the percentage split took 0.03 seconds. Under the cross-validation option 99.74% of instances are correctly classified but 0.26% of the instances are incorrectly classified. On the other hand, in the percentage split test option 99.81% of the instances were correctly classified but 0.19% of the instances are incorrectly classified. This indicates 66% percentage split classified the instances better than 10-fold cross-validation.

The threshold curves of the model at test option of percentage split for cluster 1 and cluster 2 are shown in Figure 4.2 and Figure 4.3 respectively.

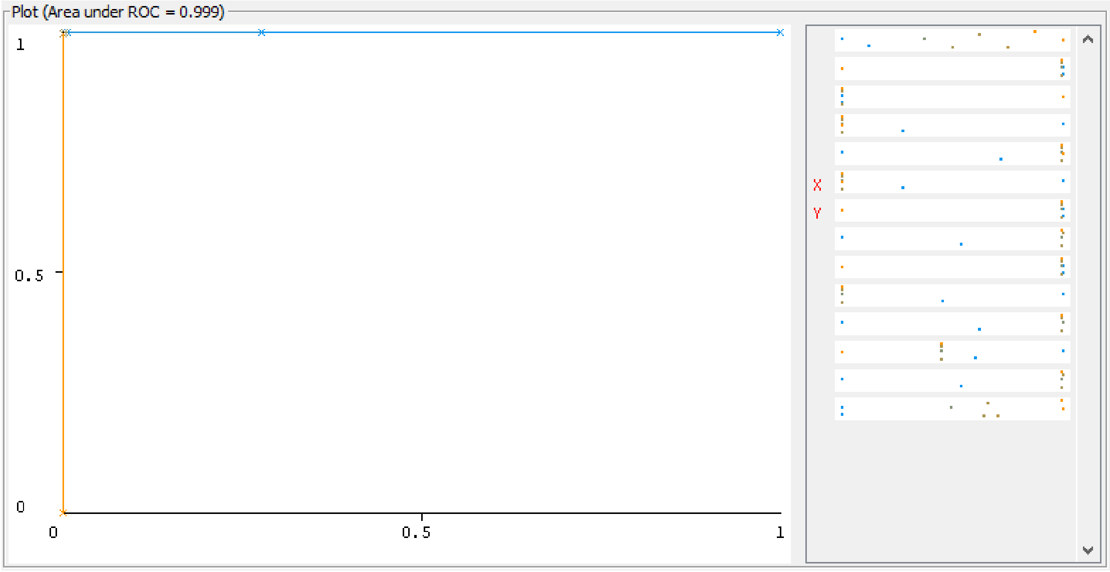


Figure 0-2 Threshold Curve of Cluster 1

Source: Weka Result, 2020

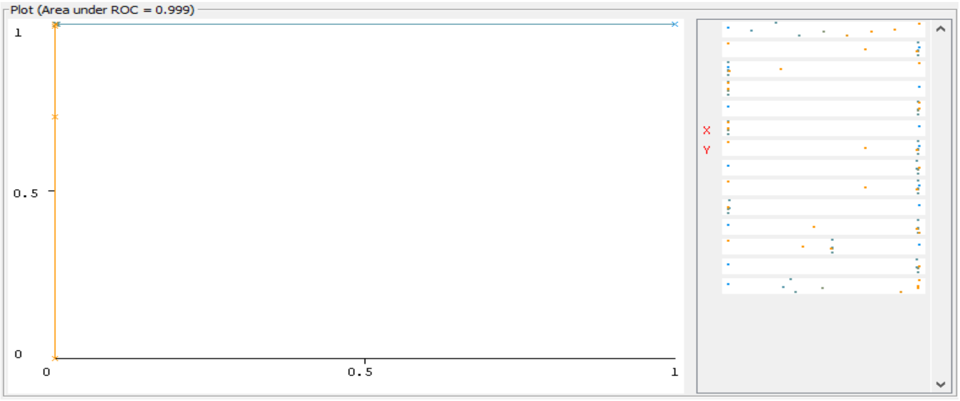


Figure 0-3 Threshold Curve of Cluster 2

Source: Weka Result, 2020

As shown in Figure 4.2 and 4.3 above, the ROC area covers 0.999 for each cluster that indicates the false positive covers very small. Therefore, the result suggests model accurately classify instances.

Naïve Bayes Classifier

Another group of experiment is conducted to explore classification model. This experiment followed Naïve Bayes Classifier algorithm. Naïve Bayes makes predictions using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes (Han and Kamber 2006).

The experiment by using Naïve Bayes classifier algorithm is conducted by using the default values of the data mining tool. Similar to previous classification modeling experiments, this experiment is conducted by using 10-fold cross validation and 66% percentage splits. The summary of result of this experiment is presented in Table 4.7 below (see Appendix 2-A snapshot of confusion matrix).

Table 4. 7 Result of Naïve Bayes Classifier Experiment

Test Option	Time taken	Classified Instances	
		Correctly classified	Incorrectly classified
Cross validation	0.01 seconds	5549 (90.3452 %)	593 (9.6548 %)
Percentage split	0.00 seconds	1883 (90.182 %)	205 (9.818 %)

Source: Weka Result, 2020

As shown in Table 4.7 above, under the cross validation test option, 90.35% of the instances are correctly classified but 9.65% of the instances are incorrectly classified. To build this model it took only 0.01 seconds. The second test option, percentage split 90.18% of the instances were correctly classified and 9.82% of the instances were incorrectly classified. The model building took 0.00 seconds. The experiment result of NaïveBayes Classifier shows that the model is very fast in predicting the business problem but it has weakness of incorrect prediction. Comparatively, 10-fold cross-validation test option has better performance than 66% percentage split.

The threshold curve of the model is presented in Figure 4.4 and 4.5 for Cluster 1 and Cluster 2 respectively.

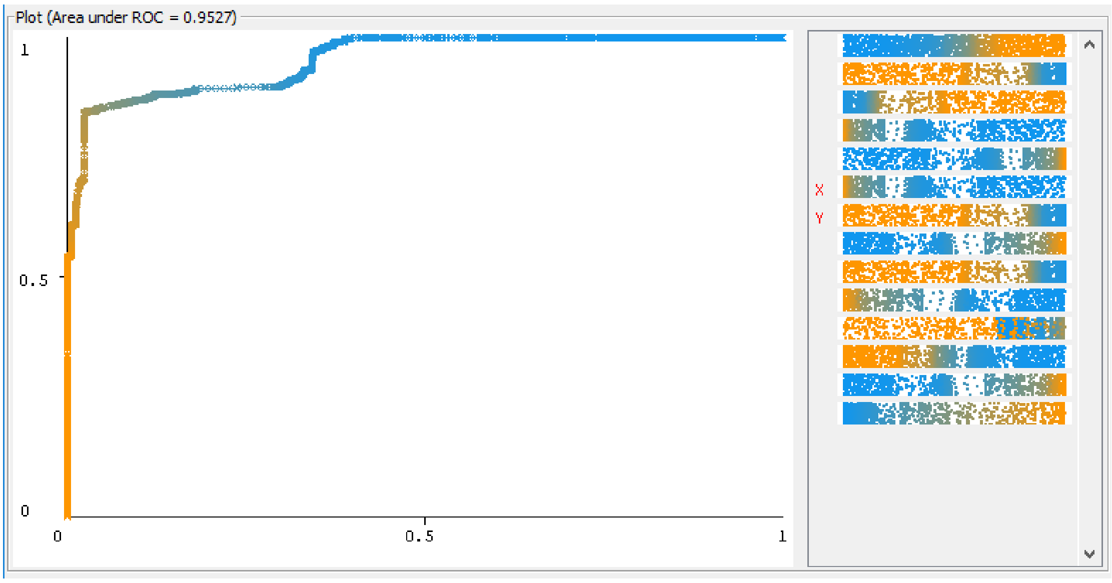


Figure 0-4 Threshold Curve of Cluster 1 Naïvebayes algorithm

Source: Weka Result, 2020

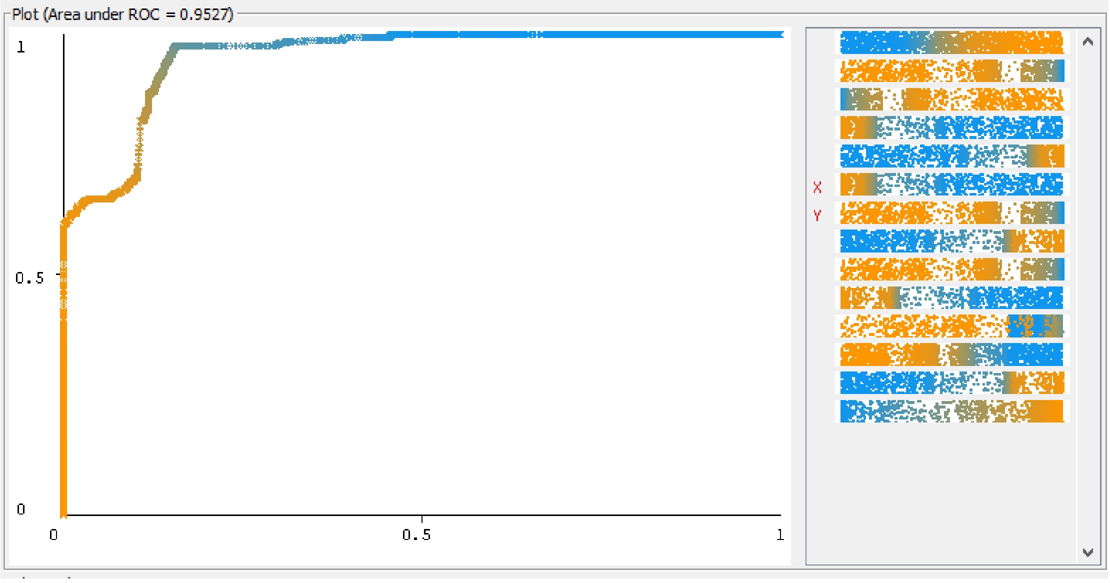


Figure 0-5 Threshold Curve for Cluster 2: Naïvebayes

Source: Weka Result, 2020

As shown in Figure 4.4 and 4.5 above, the ROC area covers 0.9527 for each cluster.

Neural Network Classifier

The third group of experiment to develop classification model is conducted by using neural network classifier. Therefore, this experiment used multilayer perceptron algorithms that uses backpropagation to learn multilayer perceptron to classify instance. This experiment is conducted by altering value of the Hidden Layer of neural network and using other default values of the algorithm. The first experiment of this classifier is conducted by using the default values of the data mining tool. The other experiments were conducted by changing hidden layers from the default. The learning rate = 0.3 and seed = 0. The experiments were conducted by using 10-fold cross-validation and 66% percentage split.

The results of the experiments neural network classifiers are summarized and presented in Table 4.8 below (see snapshot of confusion matrix Appendix 3: A – H).

Table 4. 8 Result of Multilayer Perceptron Algorithm

Hidden Layer	Test Option	Time (Seconds)	Classification	
			Correct	Incorrect
4	Cross validation	40.4	100%	0%
	Percentage split	40.55	100 %	0 %
3	Cross validation	13.02	100 %	0 %
	Percentage split	12.98	99.9042 %	0.0958 %
2	Cross validation	9.95	100 %	0 %
	Percentage split	9.64	99.9042 %	0.0958 %
1	Cross validation	6.36	100%	0%
	Percentage split	6.36	99.9042 %	0.0958 %

Source: Weka Result, 2020

As depicted in Table 4.8 above, the first experiment is conducted with value of hidden layer = 4 and using other parameters with default values. The result of the first experiment shows classification accuracy of 100%. In the other words, the model classified all instances in their appropriate clusters. This result is similar in cross-validation and percentage split. Regarding the time to build the model, cross-validation test option have smaller than the percentage split. This suggests the model built based on 10-fold cross-validation is more efficient than model built based 66% percentage split. In the all remaining experiments 10-fold cross validation has better classification accuracy than the percentage split. Results of all experiments shows 100% accuracy for cross-validation tests. The model building time is smaller at smaller values of hidden layer of neural networks. On the other hand, the predication accuracy is similar for percentage split test option for hidden layer values of 3, 2 and 1. To distinguish the best model of neural network classifier; for test option of cross-validation, the fourth experiment (hidden layer value = 1) is best model because it took smallest time (6.36 seconds) to build the model with the same level of accuracy. On the other hand, the best model built based on percentage split is at first experiment (hidden layer = 4) with time of 40.55 seconds and accuracy of 100%. In comparison with 10-fold cross-validation, model developed with percentage split takes higher time than model with 10 fold cross-validation. Therefore, best model built from neural network classifiers is developed by using hidden layer value of 1 and test option of 10-fold cross validation.

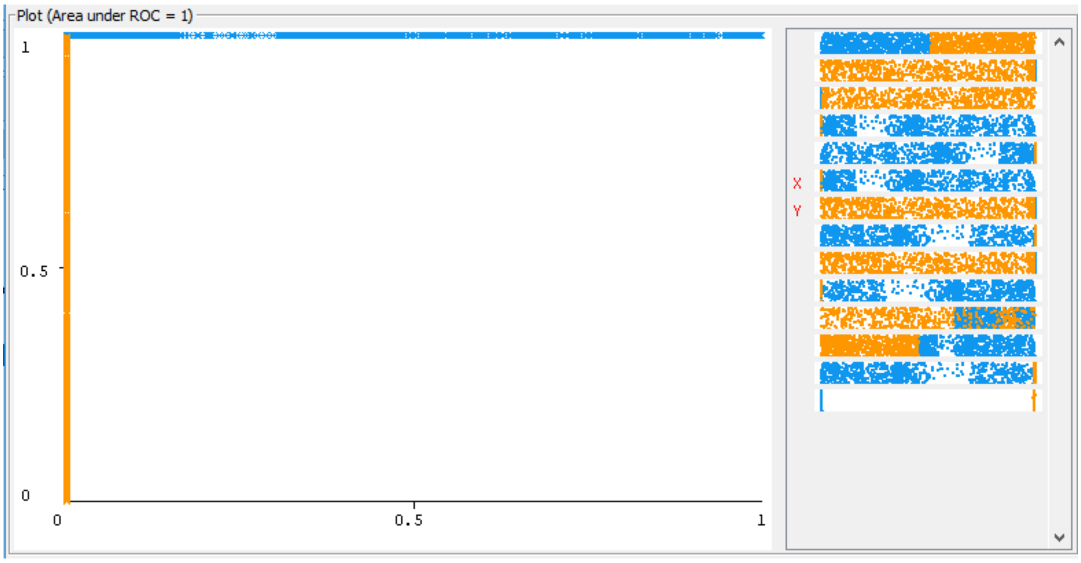


Figure 0-6 Threshold Curve for Cluster 1: Neural Network

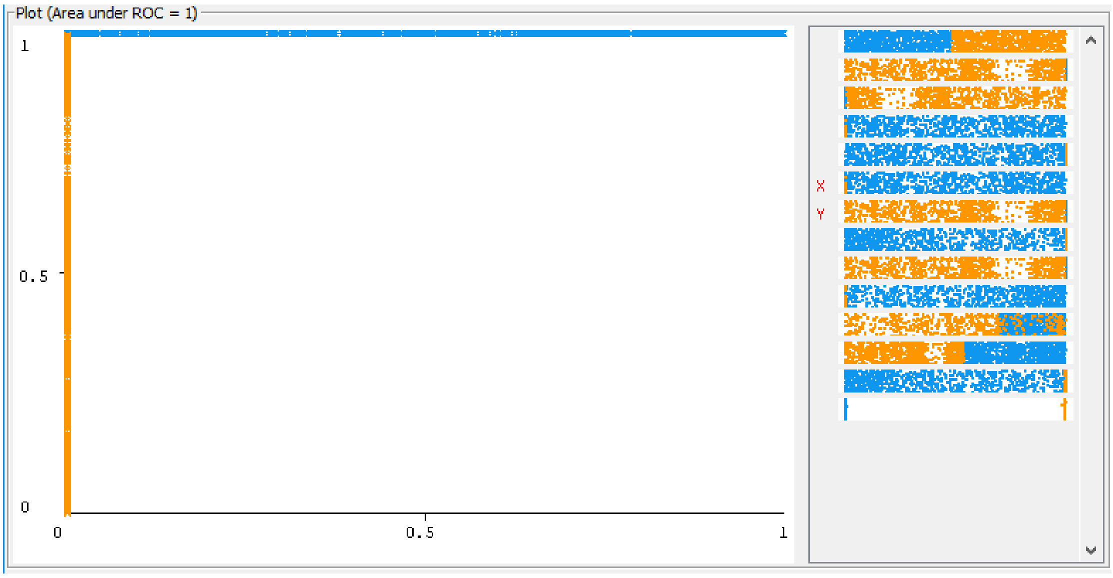


Figure 0-7 Threshold Curve of Cluster 2: Neural Network

The threshold curves of the clusters is presented in Figure 4.6 and Figure 4.7 above for Cluster 1 and Cluster 2 respectively. As shown in the figures, the ROC area covers 1 suggesting that the model has no incorrectly classified cases.

Random Forest

The last group of the experiments conducted to develop classification model was random forest algorithm. Random forest is a tree classifier algorithms used in ensemble learners. The experiment conducting by using random forest algorithm has used default values of parameters of the data mining tool. The result of the experiment is summarized in Table 4.9 below (see confusion matrix 4: A-B).

Table 4. 9 Summary of Random Forest Experiment

Test Option	Time taken	Classified Instances	
		Correctly classified	Incorrectly classified
Cross validation	1.22 seconds	6131 (99.8209 %)	11 (0.1791 %)
Percentage split	0.48 seconds	2084 (99.8084 %)	4 (0.1916 %

Source: Weka Result, 2020

As shown in Table 4.9 above, model built based on the 10-fold cross validation took 1.22 seconds and predicted 99.82% of the instances accurately. On the other hand, model built based on 66% percentage split took 0.48 seconds and correctly classified 99.81% of the instances. This suggests although the percentage split option took smaller time than 10-fold cross-validation, accuracy of model built based on cross-validation is higher than percentage split.

Comparison of the results

This study was conducted mainly to develop model to handle tax evasion in Revenue authority of Ethiopia. In line with this general objective study intends to identify best classification algorithm for detection of the tax evasion. In the previous section, different classification techniques were used while conducting experiments to develop best classification model. This section of the study presents comparison of classifiers used to build classification model.

As presented in Section 4.2.2 above, this study was conducted by using 4 classification algorithms. These algorithms include decision tree, Bayes, Neural Network and Random

Forest. J48 algorithm was used from the decision tree classifiers; NaïveBayes algorithm was used from Bayes classifiers; multilayer perceptron algorithm was used as neural network classifier; and random forest algorithms was used as.

Experiments for classification model development were conducted based on the test options; 10-fold cross-validation and 66% percentage split. Best model was selected based on prediction accuracy. Summary of performance of each best classifier is presented in Table 4.10 for comparison and selection of best classification model.

Table 4. 10 Summary of Performance of Classifiers

Classifier	Test option	Time (sec)	Accuracy	Precision	Recall	ROC	F-measure
J48	Percentage Split	0.05	99.81%	0.998	0.998	0.999	0.998
Naïvebayes	10-fold	0.02	90.35 %	0.905	0.903	0.953	0.903
MLP	10-fold	40.4	100%	1.000	1.000	1.000	1.000
Random Forest	10-fold	1.22	99.82 %	0.998	0.998	1.000	0.998

Source: Weka Result, 2020

As depicted in Table 4.10 above, Naïve Bayes algorithm has smallest accuracy with value of 90.35% and MLP has highest accuracy with value of 100%. This suggests MLP is best classifier to predict tax evasion practice of the tax payers. On the other hand, Naïve Bayes classifier is the least efficient to build the classification model. In addition, the study has used two additional classifiers; J48 and Random Forest algorithms. J48 and Random Forest algorithms classifiers built the classification models with similar level of accuracy; 99.81% and 99.82% respectively. But the J48 algorithm took smaller time than Random Forest algorithm with time of 0.05 and 1.22 seconds respectively. On overall, Naïve Bayes is least efficient model with accuracy and MLP is least efficient regarding to time taken to build the model. In addition, the domain experts have doubt about the classification result of MLP model that tax audit result on misclassification of tax payers. Consequently, there are some court cases that changed the audit findings. The study has used additional indicator for the classification performance of the models proposed. As a result, J48, MLP and Random Forest classifiers were proposed for further experiment and tested by using testing dataset prepared for the study. The experiment result by using the test dataset shows that J48 is the least accurate classifier and MLP and Random Forest classifiers have the same level of accuracy.

The accuracy of the J48 classifier is 99.34% and accuracy of MLP and random forest classifiers is 99.45%.

As shown in Table 4.10 above, MLP took very long time to build the classification model when compared to Random Forest classifier. In addition, the domain experts suggest Random Forest classifier to indicate the reality. Therefore, the study selected the random forest classifier as best algorithm to develop tax evasion detecting model. Random forest is ensemble learner that efficiently predicts fraud detection by following boosting strategies. Therefore, random forest and decision tree algorithm were used as final classification model builder.

In addition to developing the classification model, this study intends to develop decision rule for easier decision making and evasion prediction. As shown in Table 4.10 above, the J48 algorithm was good classifier to predict the classification of the tax evasion. Thus, J48 algorithm was used to generate the decision rules from the decision trees. Therefore, this study has followed hybrid strategy that the random forest is used for model building and J48 for decision rules.

The threshold curves of the clusters are presented in Figure 4.8 and Figure 4.9 below for Cluster 1 and Cluster 2 respectively.

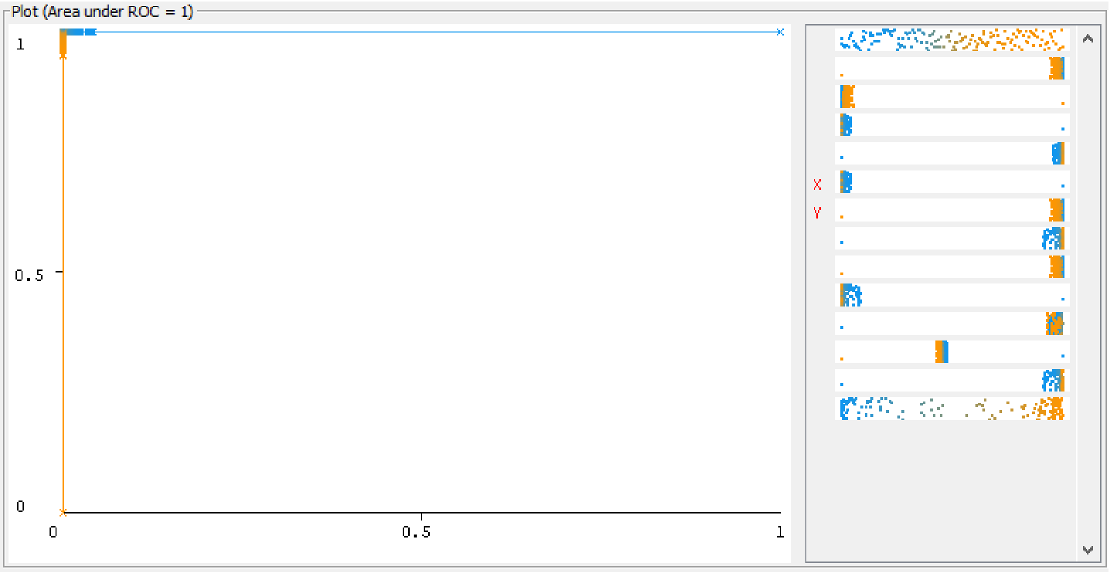


Figure 0-8 Threshold Curves for Cluster 1: Random Forest

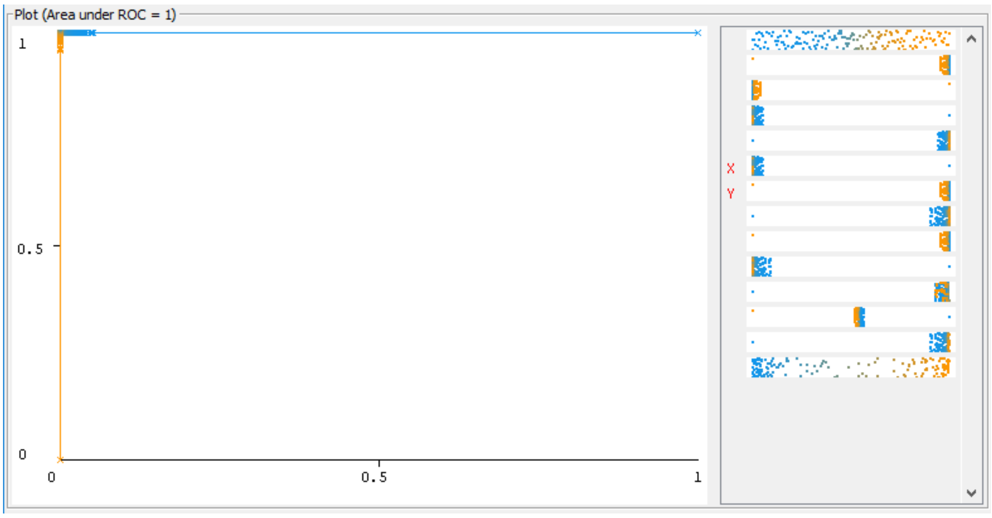


Figure 0-9 Threshold Curve Cluster 2: Random Forest

4.3 Evaluation

To reach at the research objectives, different experiments were conducted. These experiments were mainly conducted for cluster modeling and classification modeling. The cluster modeling was conducted with simple K-mean algorithm with K = 2 and varying the seed values. At the end, the clustering model was built at seed value of 10 by using Euclidean distance function. This model segmented 46% instances to cluster 1 and 54% instances to cluster 2. Based on the values of attribute grouped in the clusters, domain experts suggested Cluster 1 as the tax payers that do not engage in tax evasion and Cluster 2 as of the tax payers that involve tax evasion. Following clustering experimentations, the classification experimentations were conducted.

The classification modeling was conducted by using four classifiers; J48, NaiveBayes, neural network and random forest. The experimentation followed 10 fold cross-validation and percentage split test options. The best experiments were selected based on prediction accuracy and time taken to build the classification model. The J48 algorithm was conducted by changing CF and MNO values. Experimentation under this classifier, highest accuracy and lowest number of leaves and size of tree was generated at CF = 0.5 and MNO = 2. Under these specifications, with 10-fold cross-validation test option, the prediction accuracy is 99.74%; and under the 66% percentage split the prediction accuracy is 99.81%. The second type of algorithm implemented to build tax evasion detecting model naïve bayes classifier.

Classification model developed based on this classifier has prediction accuracy of 90.35%. The prediction accuracy of this model is smallest among the algorithms implemented for the study. The highest prediction accuracy was recorded by the model developed by using neural network and this model has predicted 100% of the cases accurately in the training dataset and 99.45% of the testing dataset. This algorithm was slowest classifier that took longest time to build the prediction model. Therefore, model developed by MLP is taken as final model classification model.

4.4 Deployment of the Result

As presented in Chapter Three, currently the organization is using manual auditing practices. The finding of this study has shown that tax evasion can be detected by using system. Thus, the study has developed an empirical model that can predict tax evasion practice. But this model could not be deployed as there is no structured system implemented to run tax auditing and handle tax evasion practices. This study has followed Cios model that the study could not deploy the system to existing system; instead it has developed knowledge. Therefore, to support decision making, the study has developed decision rules. For this purpose, decision part rules were used. Based on the PART rule, 23 rules were generated and best 12 rules are presented below. As aforementioned, Cluster1 represents non-suspicious tax reports (no tax evasion) and Cluster2 represents class of suspicious tax reports (tax evasion).

Rule 1 Liability = low AND receivables = small AND capital = medium : cluster1
Rule 2 expense = high AND capital = small AND liability = low : cluster1
Rule 3 expense = high AND receivables = moderate AND liability = moderate : cluster2
Rule 4 capital = small AND receivables = small : cluster1
Rule 5 tax = small AND expense = high : cluster2
Rule 6 capital = medium AND tax = medium : cluster1
Rule 7 receivables = small AND capital = medium : cluster1
Rule 8 tax = small AND liability = moderate : cluster2
Rule 9 expense = high AND liability = high : cluster2
Rule 10 receivables = high AND tax = medium AND liability = high : cluster1
Rule 11 capital = large AND liability = moderate AND receivables = moderate : cluster2
Rule 12 capital = medium : cluster1
Rule 13 tax = small : cluster2

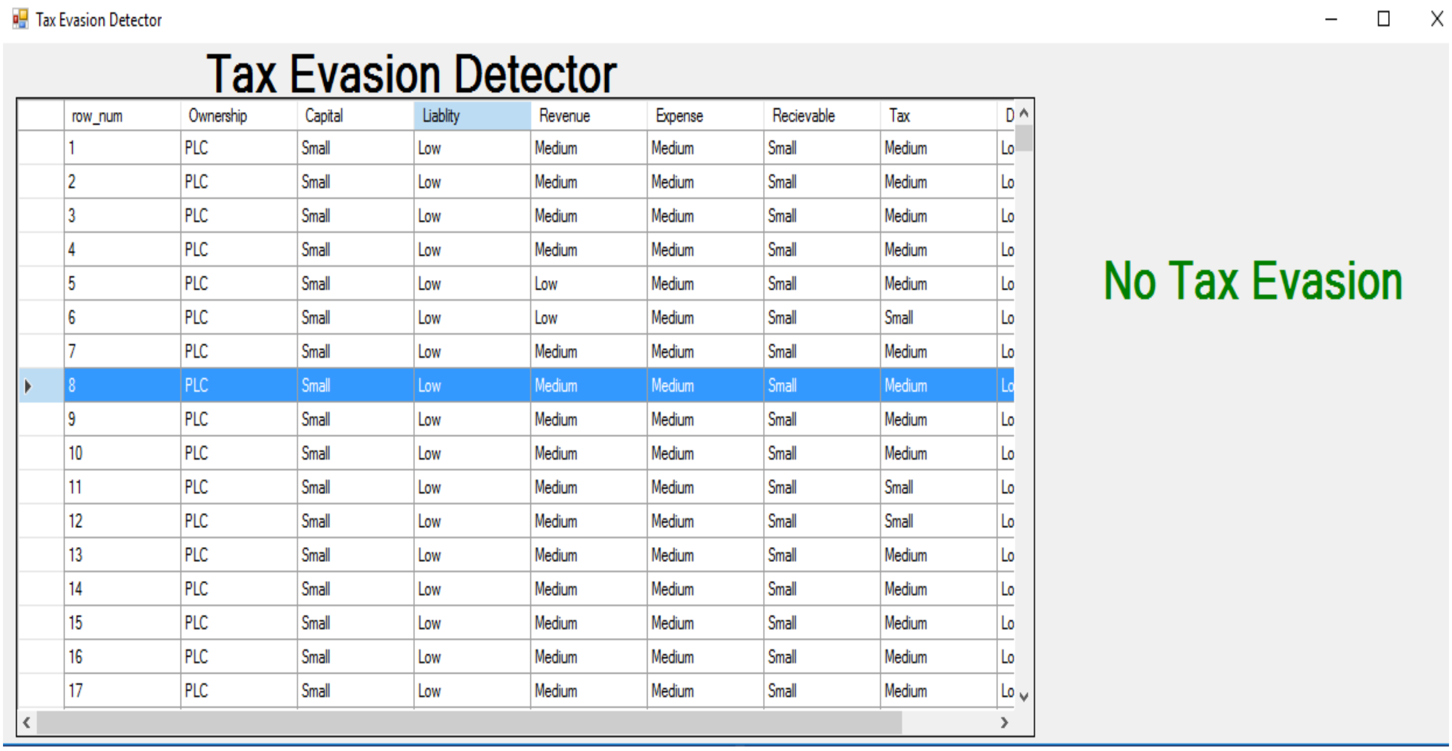
As discussed in previous sections, this study has used 8 attributes. But construction of the decision tree model used 5 attributes; liability, receivables, capital, expense and tax.

Among the attributes used while constructing the decision rule, tax payer group is the top splitting attribute, thus, it is most deterministic attribute. Small tax payers highly involve in tax evasion. The tax payers declare small amount of tax instead of actual amount of the tax. they want to be categorized in small tax payers that they have low amount of tax burden. As shown in Rule 5 and Rule 8, if small tax payers report high expense and moderate liability respectively, the tax payers are involving in tax evasion.

Tax evasion is highly suspected when high expense is reported that tax payers report high expense to reduce amount of profit and pay small amount of tax. The Part rule classification in Rule 3, Rule 5 and Rule 9 confirms the business problem. These rules are rated as best rules by the domain experts.

As shown Rule 11, if tax payers report high capital in combination with high liability and moderate receivables, they are suspicious that they are involving in tax evasion. The rule suggests that tax payers involve in tax evasion, when the companies increase capital by high payables and moderate receivables.

4.5 Sample prototype



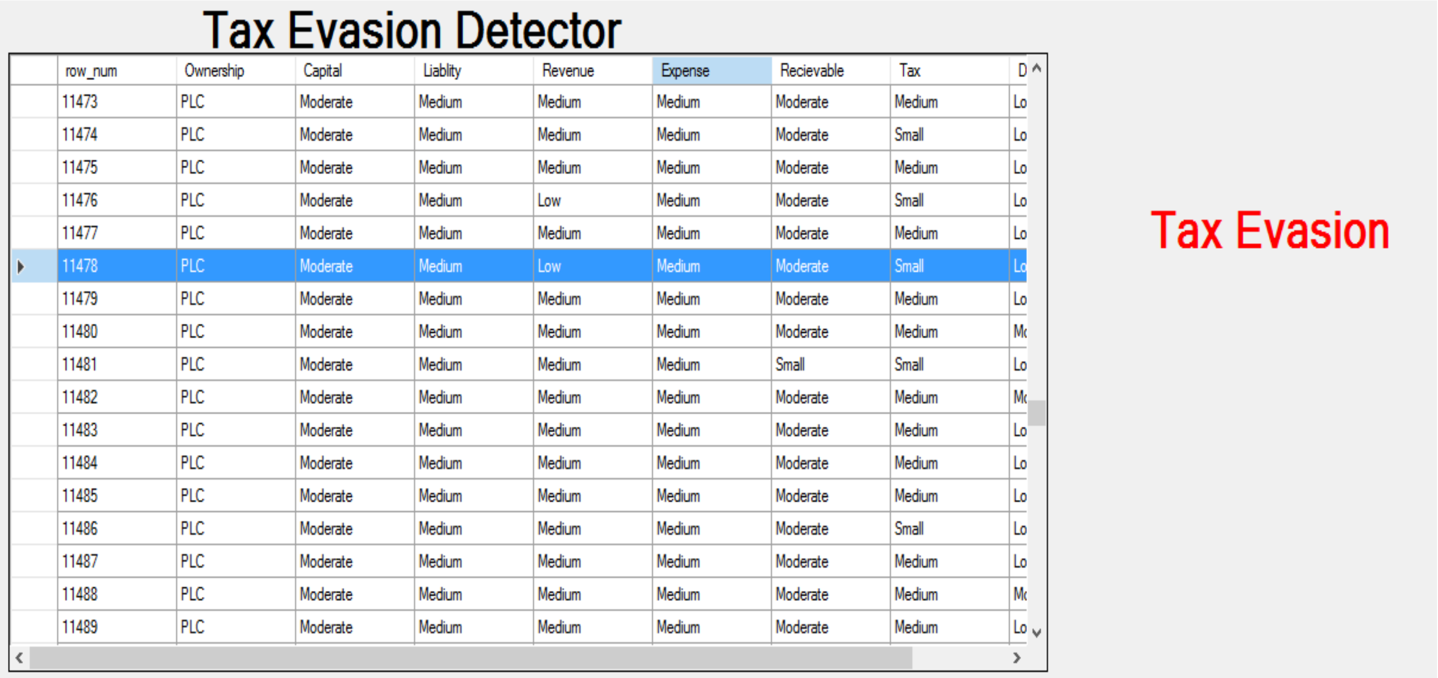
Tax Evasion Detector

Tax Evasion Detector

	row_num	Ownership	Capital	Liablity	Revenue	Expense	Recievable	Tax	D
	1	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	2	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	3	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	4	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	5	PLC	Small	Low	Low	Medium	Small	Medium	Lo
	6	PLC	Small	Low	Low	Medium	Small	Small	Lo
	7	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
▸	8	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	9	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	10	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	11	PLC	Small	Low	Medium	Medium	Small	Small	Lo
	12	PLC	Small	Low	Medium	Medium	Small	Small	Lo
	13	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	14	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	15	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	16	PLC	Small	Low	Medium	Medium	Small	Medium	Lo
	17	PLC	Small	Low	Medium	Medium	Small	Medium	Lo

No Tax Evasion

Figure 0-10 sample no tax evasion detector



Tax Evasion Detector

	row_num	Ownership	Capital	Liablity	Revenue	Expense	Recievable	Tax	D
	11473	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Lo
	11474	PLC	Moderate	Medium	Medium	Medium	Moderate	Small	Lo
	11475	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Lo
	11476	PLC	Moderate	Medium	Low	Medium	Moderate	Small	Lo
	11477	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Lo
▸	11478	PLC	Moderate	Medium	Low	Medium	Moderate	Small	Lo
	11479	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Lo
	11480	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Mc
	11481	PLC	Moderate	Medium	Medium	Medium	Small	Small	Lo
	11482	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Mc
	11483	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Lo
	11484	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Lo
	11485	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Lo
	11486	PLC	Moderate	Medium	Medium	Medium	Moderate	Small	Lo
	11487	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Lo
	11488	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Mc
	11489	PLC	Moderate	Medium	Medium	Medium	Moderate	Medium	Lo

Tax Evasion

Figure 0-11 sample tax evasion detector

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

This study was conducted by implementing the data mining techniques to detect and predict tax evasion practice by tax payers in Addis Ababa at Ministry of Revenue of Ethiopia. The study has followed Cios model that applies understanding the problem domain, understanding the data, preparation of the data, DM, evaluation of the discovered knowledge, and using the discovered knowledge. Based on the aforementioned objective the model development was conducted in two phases; cluster modeling and classification modeling. The cluster modeling was conducted by using simple K-mean algorithm to segment data in tax evasion and no evasion. Different clustering models were experimented by the study and the best clustering model was developed by using k = 2, seed = 10 and Euclidean distance function. This model clustered 46% of the instance into non-evading group and 54% into evading group.

After clustering model is developed, the classification modeling was conducted by using four classification algorithms; J48, Naïvebayes, Neural Network and Random Forest. The classification modeling was conducted by changing test options and default values of the data mining tool. Finally, the study has selected MLP algorithms for classification of tax evading and non-evading tax payers. the decision rule was conducted by using Part Rule. The classification model developed through MLP algorithm predicted the training data with an accuracy of 100% and the supplied testing data with an accuracy of 99.45% suggesting that the model correctly classify new instances with their right class.

The Part Rule was implemented to construct decision rules that support decision making about tax evasion practices by the tax payers. among the attributes used by the study, liability, receivables, capital, expense and tax are most important variables. Part Rule classifier created tax as most splitting attribute used in the study and the model suggests small tax payers are most likely to practice tax evasion. Companies that report high expense in combination with moderate receivables, and high and moderate liability involve in tax evasion.

This study identified promising results that enables to detect tax evasion practices. In addition, the study suggests solutions to solve the problems by using data mining techniques.

5.2 Recommendations

Based on the findings of the study, the following recommendations were provided.

- The predictive model developed constructed various rules. Therefore, to use the model developed effectively, ministry of revenue is recommended to design system that can solve the tax evasion practice instead of focusing on manual tax auditing practices. As the manual tax auditing is resulting on corrupted practices, it is suggested to use automated system for detection of tax evasion.
- As it is shown in the model construction process, the accuracy of the models built in the study process were decreased on test data. Therefore, further investigations are important by using other classification models such as meta classifiers and support vector machine.
- The organization has big problem of not adequately handling data about tax payers that makes tax evasion detection more difficult. Therefore, the organization is highly recommended to develop database to handle information about tax payers.
- Include other area of coverage(regional) to increase the population
- Apply network integrated system with cash register machine and devlop centralized database to handle information about tax payers.

5.3 Recommendation for Future Studies

- The study has not included descriptive attributes that might explain tax evasion behavior of the tax payers. the study has used mainly financial aspect of the companies. Therefore, this study suggests further studies to include these attributes such as management practice, type of business where a company involves, price, and location of business.
- In addition, future studies are recommended to include detailed algorithms and larger dataset by focusing on national level.

5.4 Limitations of the Study

The main limitation of this study inability to deploy the result of the study. In addition, the study has entirely focused on financial aspects and reports of the tax payers. Only secondary sources were used as inputs for the study except including domain experts. Specifically, tax payers were not involved in the study that provide opinions in tax administration.

REFERENCES

[1] W. Fox, L. Lunab and G. Schau, "Destination Taxation and Evasion: Evidence from U.S. Inter-State Commodity Flows," *Journal of Accounting and Economics,* vol. 57, no. 1, p. 43–57, 2017.

[2] R. Wu , C. Ou, H. Lin, S. Chang and D. Yen, "Using Data Mining Technique to Enhance Tax Evasion Detection Performance," *Expert Systems with Applications,* vol. 39, no. 10, p. 8769–8777, 2018.

[3] R. Komal and M. Rashmi, "More Focus on Tax Evasion Detection with Graph Based Approach," *IOSR Journal of Computer Engineering (IOSR-JCE)* , vol. 661, no. 4, pp. 5-7, 2018.

[4] Ministry of Revenue, "Ministry of Revenue of Ethiopia Annual Report," Ministry of Revenue, Addis Ababa, 2019.

[5] T. Tian, K. Lan, N. Chao, Q. Godwin, N. Zheng, F. Shah and F. Zhang, "Mining Suspicious Tax Evasion Groups in Big Data.," *IEEE Transactions on Knowledge and Data Engineering,* vol. 28, no. 10, p. 2651–2664, 2016.

[6] R. Jianfei, Y. Zheng, D. Bo, Z. Qinghua and Q. Buyue, "Identifying suspicious groups of affiliated-transaction-based tax evasion in big data," *Journal of Information Sciences,* vol. 477, no. 11, p. 508–532, 2019.

[7] D. Walter, G. Luca, L. Giuseppe, M. Fabrizio and P. Daniele, "A visual analytics system to support tax evasion discovery," *Decision Support Systems,* vol. 110, p. 71–83, 2018.

[8] D. Roux, G. Pérez and D. Roux, "Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach," vol. 221, no. 10, pp. 215-222, 2018.

[9] Gupta GK (2012) Introduction to data mining with case studies PHI, New Delhi

[10] D. Mamo, "APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT FRAUD PROTECTION: THE CASE OF ETHIOPIAN REVENUE AND CUSTOM AUTHORITY," *Unpublished Master's Thesis: Addis Ababa University,* 2013.

[11] Kumar R, Kapil AK, Bhatia (2012) A Modified tree classification in data mining. Global Journals Inc. 12, 12: 58-63

[12] Zhao Q, Fränti P (2014) WB-index: A sum-of-squares based index for cluster validity. Data & Knowledge Engineering 92:77–89

[13] Rui Xu, Donald CW II (2005) Survey of Clustering Algorithms. IEEE Transactions on neural Networks, 16: 645-678

[14] Kleinberg J (2002) An impossibility theorem for clustering. Conf. Advances in Neural Information Processing Systems, 15: 463–470

[15] Jain A, Dubes R (1988) Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall

[16] Abbas OA (2008) Comparisons between Data Clustering Algorithms. International Journal of Information Technology 5: 320-325

[17] Kotsiantis SB, Pintelas PB (2004) Recent Advances in Clustering: A Brief Survey. WSEAS Transactions on Information Science and Applications, 1(1): 73–81

[18] Jain AK (2010) Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters, 31(8): 651-666

[19] Rao GN, Nagaraj S (2014) A Study on the Prediction of Student's Performance by applying straight-line regression analysis using the method of least squares. IJCSE 3: 43-45

[20] Sansgiry SS, Bhosle M, Sail K (2006) Factors That Affect Academic Performance Among Pharmacy Students. American Journal of Pharmaceutical Education 70 (5) Article 104

[21] Kriegel HK, Borgwardt KM, Kröger P, Pryakhin A, Schubert M, Zimek A (2007) Future trends in data mining. Data Mining and Knowledge Discovery 15:87–97

[22] Radaideh Q, Nagi E (2012) Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance. IJACSA 3:144-151

[23] Vijiyarani S, Sudha S (2013) Disease prediction in data mining- A survey. IJCAIT (2).

[24] Velmurugan T (2014) Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data. Applied Soft Computing 19 pp.134–146

[25] Huang Z (1998) Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Acsys CRC, CSIRO

[26] Ngai EWT, Yong Hu, Wong YH,Chen Y, Sun X (2011) The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems 50:559-569

[27] André L.V. Coelho, , Everlândio Fernandes, Katti Faceli (2011) Multi-objective design of hierarchical consensus functions for clustering ensembles via genetic programming Decision Support Systems 51:794-809

[28] Aviad B, Roy G (2012) A decision support method, based on bounded rationality concepts, to reveal feature saliency in clustering problems. Decision Support Systems 54: 292–303

[29] Combes C, Azema J (2013) Clustering using principal component analysis applied to Autonomy – disability of elderly people. Decision Support Systems 55:578–586

[30] Sandeep, Priyanka, Bansal R (2014) Performance Comparison of Various Partition based Clustering Algorithms. IJEMR pp. 216-223

[31] Oyelade OJ, Oladipupo OO, Obagbuwa, IC (2010) Application of k-Means Clustering algorithm for prediction of Students' Academic Performance. IJCSIS 7: 292-295

[32] Rao GN, Ramachandra M (2014) A Study on the Academic Performance of the Students by Applying K-Means Algorithm. IJETCAS 14-180

[33] Adhikari A, Rao PR (2008) Efficient clustering of databases induced by local patterns. Decision Support Systems 44:925–943

[34] Lin PL, Po-Huang PW ,Kuo PH , Lai YH (2014) A size-insensitive integrity-based fuzzy c-means method for data clustering. Pattern Recognition 47:2042–2056

[35] Jacques J, Preda C (2014) Model-based clustering for multivariate functional data. Computational Statistics and Data Analysis 71:92–106

[36] Angelis LD, Dias JG (2014) Mining categorical sequences from data using a hybrid clustering method. European Journal of Operational Research 234:720–730

[37] Xiao FU, Fan C (2014) Data mining in building automation system for improving building operational performance. Energy and Buildings 75: 109–118

[38] Irpino A, Verde R, Francisco de A.T, Carvalho (2014) Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Systems with Applications 41:3351–3366

[39] Liu Y, Qianqian Li, Tang X, Ning Ma, Tian R (2014) Superedge prediction:What opinions will be mined based on an opinion supernetwork model. Decision Support Systems 64:118–129

[40] Romero C ,Ventura S (2007) Educational data mining: A survey. Expert Systems with Applications 33: 135–146

[41] Breese JS, Heckerman D, Kadie C (1998) Empirica Analysis of Predictive Algorithms for Collaborative Filtering Microsoft Research, Morgan Kaufmann Publishers, pp. 1-18.

[42] Padmaja S and Fatima SS (2013) Opinion Mining and Sentiment Analysis –An Assessment of People’s Belief: A Survey. International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) 4(1)

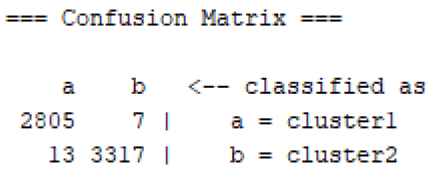
[43] Basili, R., Di Nanni, M. and Pazienza, M. T. (1999) Engineering of IE systems: an object oriented approach. In: Pazienza, editor, Information Exctraction, LNAI 1714, pp. 134–164

[44] Ferrucci D, Lally A (2004) UIMA: an architectural approach to unstructured information processing in the corporate research Environment. Natural Language Engineering 10:327 – 348

[45] Low Y, Gonzalez J, Kyrola A, Bickson A, Guestrin C, Berkeley UC (2010) GraphLab: A NewFramework For Parallel Machine Learning

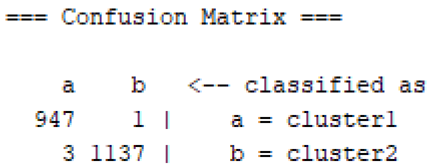
Appendix

Appendix 1-A: Confusion Matrix of J48 Algorithm with test option of Cross-validation at (default values)



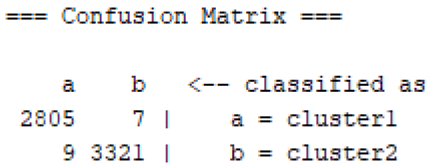
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2805    7 |    a = cluster1
   13 3317 |    b = cluster2
```

Appendix 1-B: Confusion Matrix of J48 Algorithm with test option of Percentage Split at (default values)



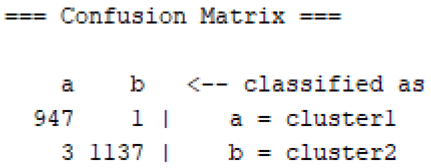
```
=== Confusion Matrix ===

    a    b   <-- classified as
  947    1 |    a = cluster1
    3 1137 |    b = cluster2
```

Appendix 1-C: Confusion Matrix of J48 Algorithm with test option of Cross-validation at (CF = 0.5 and MNO = 2)



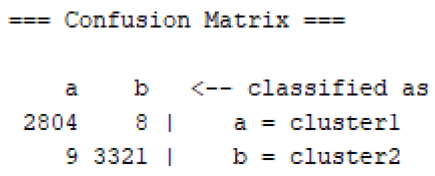
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2805    7 |    a = cluster1
    9 3321 |    b = cluster2
```

Appendix 1-D: Confusion Matrix of J48 Algorithm with test option of Percentage Split at (CF = 0.5 and MNO = 2)



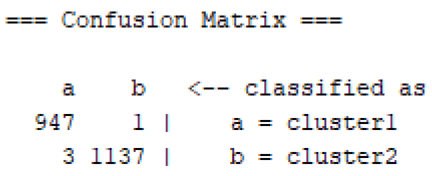
```
=== Confusion Matrix ===

    a    b   <-- classified as
  947    1 |    a = cluster1
    3 1137 |    b = cluster2
```

Appendix 1-E: Confusion Matrix of J48 Algorithm with test option of Cross-validation at (CF = 0.75 and MNO = 2)



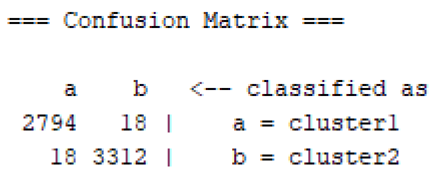
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2804    8 |    a = cluster1
    9 3321 |    b = cluster2
```

Appendix 1-F: Confusion Matrix of J48 Algorithm with test option of Percentage Split at (CF = 0.5 and MNO = 2)



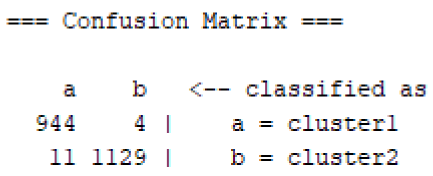
```
=== Confusion Matrix ===

    a    b   <-- classified as
  947    1 |    a = cluster1
    3 1137 |    b = cluster2
```

Appendix 1-G: Confusion Matrix of J48 Algorithm with test option of Cross-validation at (CF = 0.25 and MNO = 5)



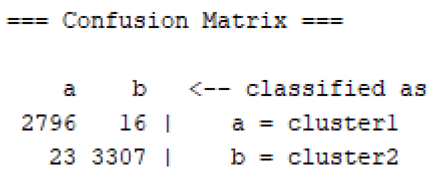
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2794   18 |    a = cluster1
   18 3312 |    b = cluster2
```

Appendix 1-H: Confusion Matrix of J48 Algorithm with test option of Percentage Split at (CF = 0.25 and MNO = 5)



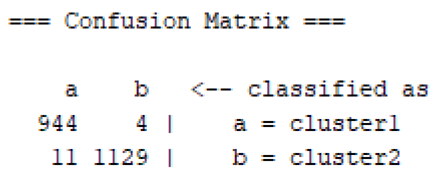
```
=== Confusion Matrix ===

    a    b   <-- classified as
  944    4 |    a = cluster1
   11 1129 |    b = cluster2
```

Appendix 1-I: Confusion Matrix of J48 Algorithm with test option of Cross-validation at (CF = 0.25 and MNO = 10)



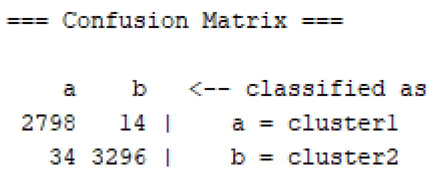
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2796   16 |    a = cluster1
   23 3307 |    b = cluster2
```

Appendix 1-J: Confusion Matrix of J48 Algorithm with test option of Percentage Split at (CF = 0.25 and MNO = 10)



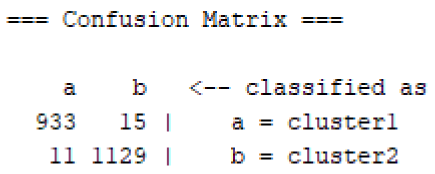
```
=== Confusion Matrix ===

    a    b   <-- classified as
  944    4 |    a = cluster1
   11 1129 |    b = cluster2
```

Appendix 1-K: Confusion Matrix of J48 Algorithm with test option of Cross-validation at (CF = 0.25 and MNO = 15)



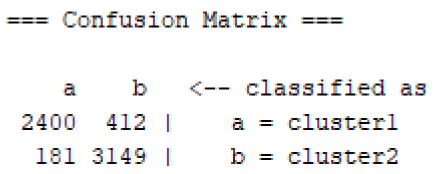
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2798   14 |    a = cluster1
   34 3296 |    b = cluster2
```

Appendix 1-L: Confusion Matrix of J48 Algorithm with test option of Percentage Split at (CF = 0.25 and MNO = 15)



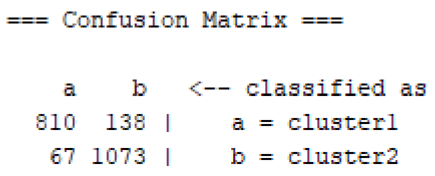
```
=== Confusion Matrix ===

    a    b   <-- classified as
  933   15 |    a = cluster1
   11 1129 |    b = cluster2
```

Appendix 2-A: Confusion Matrix of Naïve Bayes Algorithm with test option of Cross-Validation



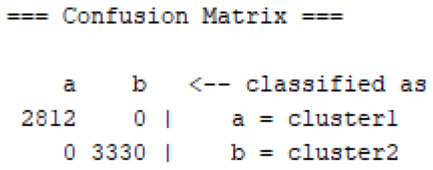
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2400  412 |    a = cluster1
  181 3149 |    b = cluster2
```

Appendix 2-B: Confusion Matrix of Naïve Bayes Algorithm with test option of Percentage Split



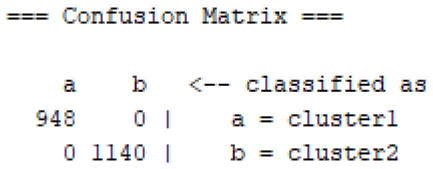
```
=== Confusion Matrix ===

    a    b   <-- classified as
  810  138 |    a = cluster1
   67 1073 |    b = cluster2
```

Appendix 3-A: Confusion Matrix of MLP Algorithm with test option of Cross-Validation (hiddenlayer = 4)



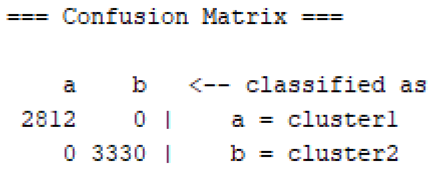
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2812    0 |    a = cluster1
    0 3330 |    b = cluster2
```

Appendix 3-B: Confusion Matrix of MLP Algorithm with test option of Percentage split (hiddenlayer = 4)



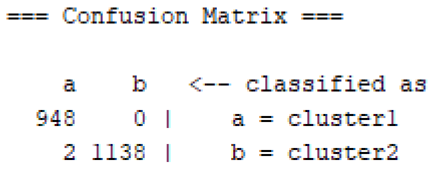
```
=== Confusion Matrix ===

    a    b   <-- classified as
  948    0 |    a = cluster1
    0 1140 |    b = cluster2
```

Appendix 3-C: Confusion Matrix of MLP Algorithm with test option of Cross-Validation (hiddenlayer = 3)



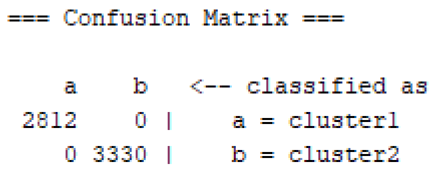
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2812    0 |    a = cluster1
    0 3330 |    b = cluster2
```

Appendix 3-D: Confusion Matrix of MLP Algorithm with test option of Percentage Split (hiddenlayer = 3)



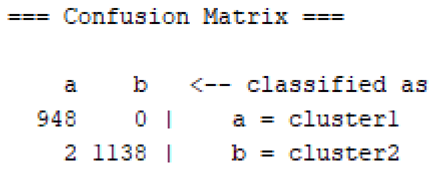
```
=== Confusion Matrix ===

    a    b   <-- classified as
  948    0 |    a = cluster1
    2 1138 |    b = cluster2
```

Appendix 3-E: Confusion Matrix of MLP Algorithm with test option of Cross-Validation (hiddenlayer = 2)



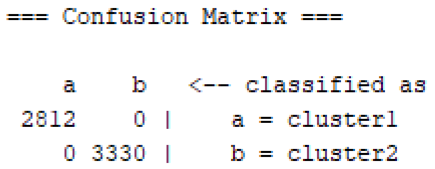
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2812    0 |    a = cluster1
    0 3330 |    b = cluster2
```

Appendix 3-F: Confusion Matrix of MLP Algorithm with test option of Percentage Split (hiddenlayer = 2)



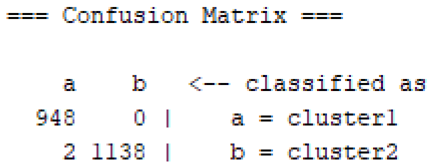
```
=== Confusion Matrix ===

    a    b   <-- classified as
  948    0 |    a = cluster1
    2 1138 |    b = cluster2
```

Appendix 3-G: Confusion Matrix of MLP Algorithm with test option of Cross-Validation (hiddenlayer = 1)



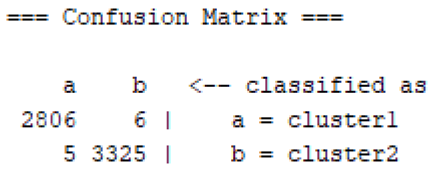
```
=== Confusion Matrix ===

    a    b   <-- classified as
 2812    0 |    a = cluster1
    0 3330 |    b = cluster2
```

Appendix 3-H: Confusion Matrix of MLP Algorithm with test option of Percentage Split (hiddenlayer = 1)



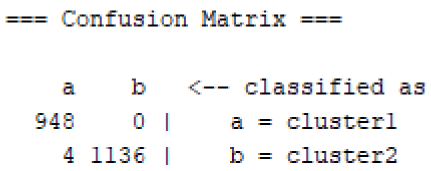
```
=== Confusion Matrix ===

    a    b   <-- classified as
  948    0 |    a = cluster1
    2 1138 |    b = cluster2
```

Appendix 4-A: Confusion Matrix of Random Forest Algorithm with test option of Cross-Validation

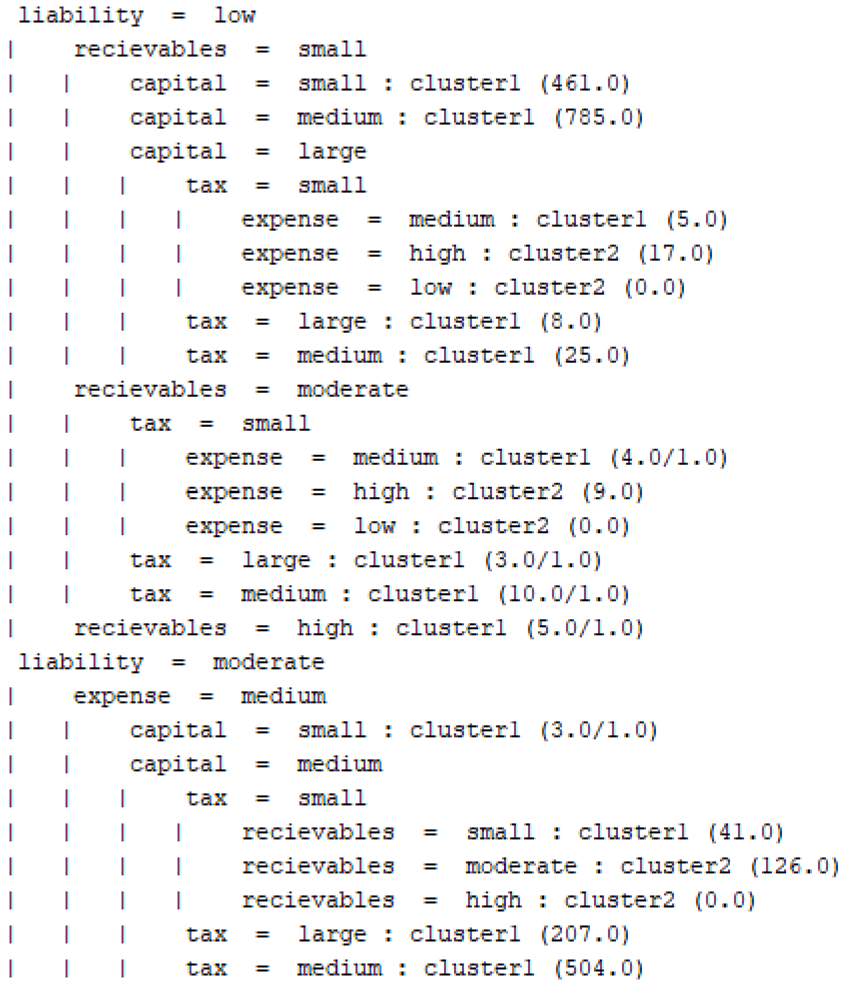


```
=== Confusion Matrix ===

    a    b   <-- classified as
 2806    6 |    a = cluster1
    5 3325 |    b = cluster2
```

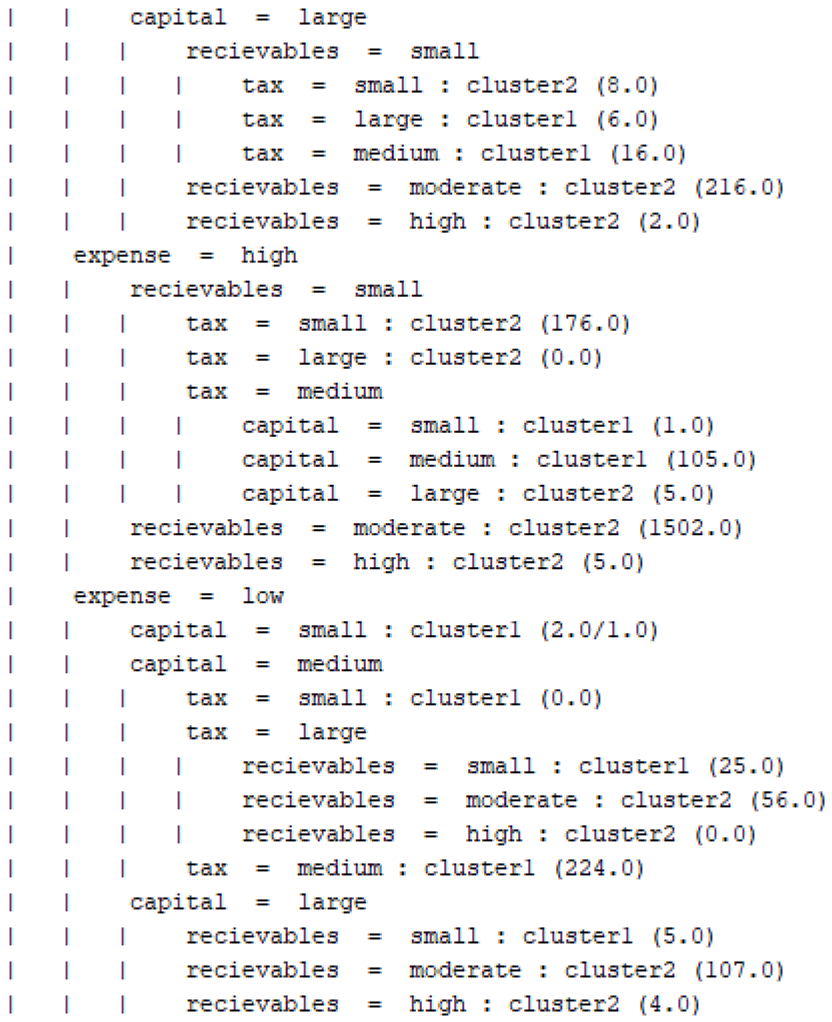
Appendix 4-B: Confusion Matrix of Random Forest Algorithm with test option of Percentage Split



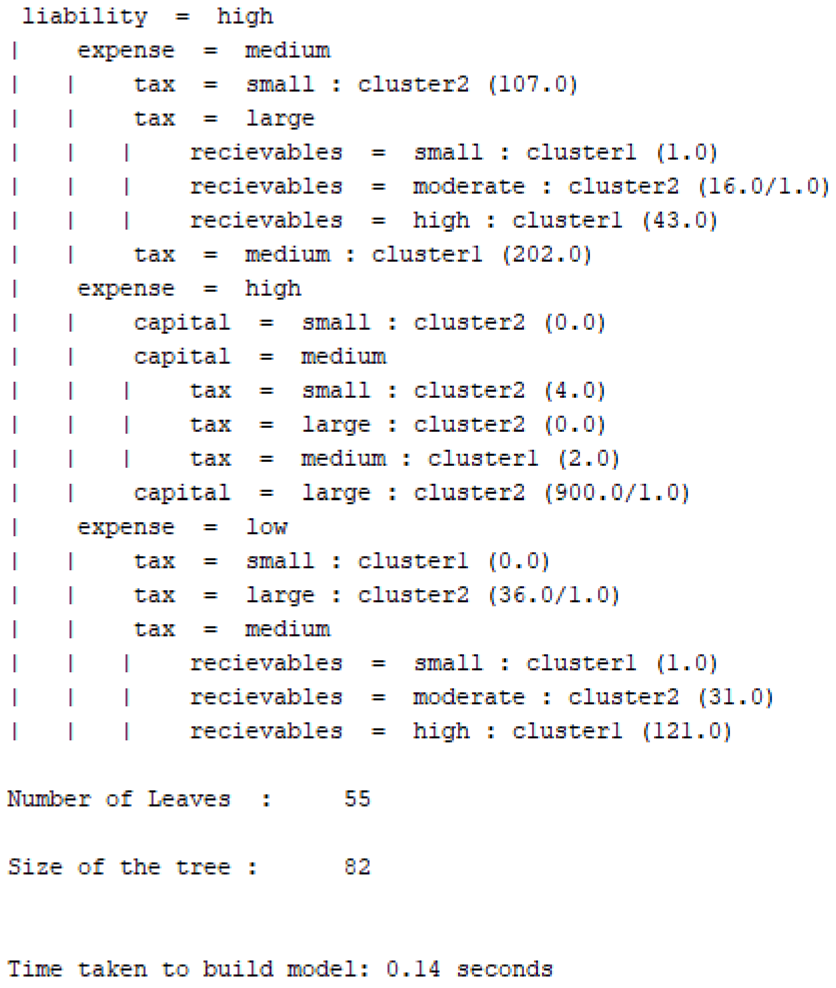
```
=== Confusion Matrix ===

    a    b   <-- classified as
  948    0 |    a = cluster1
    4 1136 |    b = cluster2
```



```
 liability  =  low
|    recievables  =  small
|   |    capital  =  small : cluster1 (461.0)
|   |    capital  =  medium : cluster1 (785.0)
|   |    capital  =  large
|   |   |    tax  =  small
|   |   |   |    expense  =  medium : cluster1 (5.0)
|   |   |   |    expense  =  high : cluster2 (17.0)
|   |   |   |    expense  =  low : cluster2 (0.0)
|   |   |    tax  =  large : cluster1 (8.0)
|   |   |    tax  =  medium : cluster1 (25.0)
|    recievables  =  moderate
|   |    tax  =  small
|   |   |    expense  =  medium : cluster1 (4.0/1.0)
|   |   |    expense  =  high : cluster2 (9.0)
|   |   |    expense  =  low : cluster2 (0.0)
|   |    tax  =  large : cluster1 (3.0/1.0)
|   |    tax  =  medium : cluster1 (10.0/1.0)
|    recievables  =  high : cluster1 (5.0/1.0)
 liability  =  moderate
|    expense  =  medium
|   |    capital  =  small : cluster1 (3.0/1.0)
|   |    capital  =  medium
|   |   |    tax  =  small
|   |   |   |    recievables  =  small : cluster1 (41.0)
|   |   |   |    recievables  =  moderate : cluster2 (126.0)
|   |   |   |    recievables  =  high : cluster2 (0.0)
|   |   |    tax  =  large : cluster1 (207.0)
|   |   |    tax  =  medium : cluster1 (504.0)
```



```
|   |   capital = large
|   |   |   recievables = small
|   |   |   |   tax = small : cluster2 (8.0)
|   |   |   |   tax = large : cluster1 (6.0)
|   |   |   |   tax = medium : cluster1 (16.0)
|   |   |   recievables = moderate : cluster2 (216.0)
|   |   |   recievables = high : cluster2 (2.0)
|   expense = high
|   |   recievables = small
|   |   |   tax = small : cluster2 (176.0)
|   |   |   tax = large : cluster2 (0.0)
|   |   |   tax = medium
|   |   |   |   capital = small : cluster1 (1.0)
|   |   |   |   capital = medium : cluster1 (105.0)
|   |   |   |   capital = large : cluster2 (5.0)
|   |   recievables = moderate : cluster2 (1502.0)
|   |   recievables = high : cluster2 (5.0)
|   expense = low
|   |   capital = small : cluster1 (2.0/1.0)
|   |   capital = medium
|   |   |   tax = small : cluster1 (0.0)
|   |   |   tax = large
|   |   |   |   recievables = small : cluster1 (25.0)
|   |   |   |   recievables = moderate : cluster2 (56.0)
|   |   |   |   recievables = high : cluster2 (0.0)
|   |   |   tax = medium : cluster1 (224.0)
|   |   capital = large
|   |   |   recievables = small : cluster1 (5.0)
|   |   |   recievables = moderate : cluster2 (107.0)
|   |   |   recievables = high : cluster2 (4.0)
```



```
 liability  =  high
|    expense  =  medium
|    |    tax  =  small : cluster2 (107.0)
|    |    tax  =  large
|    |    |    recievables  =  small : cluster1 (1.0)
|    |    |    recievables  =  moderate : cluster2 (16.0/1.0)
|    |    |    recievables  =  high : cluster1 (43.0)
|    |    tax  =  medium : cluster1 (202.0)
|    expense  =  high
|    |    capital  =  small : cluster2 (0.0)
|    |    capital  =  medium
|    |    |    tax  =  small : cluster2 (4.0)
|    |    |    tax  =  large : cluster2 (0.0)
|    |    |    tax  =  medium : cluster1 (2.0)
|    |    capital  =  large : cluster2 (900.0/1.0)
|    expense  =  low
|    |    tax  =  small : cluster1 (0.0)
|    |    tax  =  large : cluster2 (36.0/1.0)
|    |    tax  =  medium
|    |    |    recievables  =  small : cluster1 (1.0)
|    |    |    recievables  =  moderate : cluster2 (31.0)
|    |    |    recievables  =  high : cluster1 (121.0)

Number of Leaves  :     55

Size of the tree :      82


Time taken to build model: 0.14 seconds
```