

DSpace Institution

DSpace Repository

<http://dspace.org>

Information Technology

thesis

2020-08

Developing Tax Payer Fraudulent Risk Level Prediction Model for Auditing Using Hybrid Machine Learning Techniques

Chalye, Zeryhun Taryku

<http://ir.bdu.edu.et/handle/123456789/12744>

Downloaded from DSpace Repository, DSpace Institution's institutional repository



BiT

Bahir Dar Institute Of Technology

ባሕር ዳር ቴክኖሎጂ ሊንስቲትዩት

Bahir Dar University

ባሕር ዳር ዩኒቨርሲቲ

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

FACULTY OF COMPUTING

**Developing Tax Payer Fraudulent Risk Level Prediction Model for Auditing
Using Hybrid Machine Learning Techniques**

By

Chalye Zeryhun Taryku

Program: MSc. Information Technology

August 2020

Bahir Dar; Ethiopia

**DEVELOPING TAX PAYER FRAUDULENTRISK LEVEL PREDICTION MODEL
FOR AUDITING USING HYBRID MACHINE LEARNING TECHNIQUES**

CHALYE ZERYHUN

A MSC thesis submitted to the school of Research and Graduate Studies of Bahir Dar
Institute of Technology, BDU in partial fulfillment of the requirements for the degree of
MASTERS in the Information Technology in the faculty of computing.

Advisor: Gebeyehu Belay (Dr. of Eng & Ass. Prof)

August , 2020

Bahir Dar, Ethiopia

DECLARATION

This is to certify that the thesis entitled “**Developing Tax Payer Fraudulent Risk Level Prediction Model for Auditing Using Hybrid Machine Learning Techniques**”, submitted in partial fulfillment of the requirements for the degree of Master of Science in Information Technology under **Faculty of Computing**, Bahir Dar Institute of Technology, is a record of original work carried out by me and has never been submitted to this or any other institution to get any other degree or certificates. The assistance and help I received during the course of this investigation have been duly acknowledged.

CHALYE ZERYHUN

August 2020

Name of the candidate

signature

Date

Advisor Name: Gebeyehu Belay (Dr. of Eng & Ass. Prof)

Advisor’s Signature: _____

@2020

CHALYE ZERYHUN

ALL RIGHTS RESERVED

Bahir Dar University

Bahir Dar Institute of Technology

School of Research and Graduate Studies

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

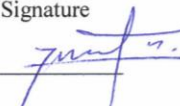
Approval of thesis for defense result

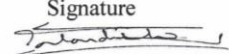
I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of Student CHALYEZERYUHN Signature  Date August 2020

As members of the board of examiners, we examined this thesis entitled “Developing Tax Payer Fraudulent Risk Level Prediction Model for Auditing Using Hybrid Machine Learning Techniques” by CHALYEZERYUHN. We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science in “Information Technology”.

Board of Examiners

Name of Advisor Signature Date
Gebeyehu Belay (Dr. of Eng & Ass. Prof)  27/12/2020

Name of External examiner Signature Date
Wondwosen Mulugeta(PhD)  August 2020

Name of Internal Examiner Signature Date
Seffi Geberehu (Ass-prof)  August 27/2020

Name of Chairperson Signature Date
Esubalew Alemneh (PhD)  27 Aug 2020

Name of Chair Holder Signature Date
Derejau L.  Aug. 27 - 2020

Name of Faculty Dean Signature Date
Belete B.  21/12/2012 e.c



DEDICATION

I would like to dedicate this thesis work to my mother Enat Mitku, my wife Wubrist Yalew myson Zekaryas chale.

ACKNOWLEDGEMENTS

First I thank the Almighty of God for all His blessings & helping me to complete this thesis. Then I would like to thank Dr. GebeyehuB at Bahir Dar University, that advise me for the success of this thesis. And thanks to my wife Wubrst Yalew initiate me to work this paper. Besides, thanks to my entire friend that put your idea for the success of this thesis .Also, thanks to Mr. TewodrosWorku, lecturer Bahir Dar University that helps me in providing relevant information to write this thesis.

ABSTRACT

Auditing a business or an individual taxpayer is not today's activity; it lasts for a long period. It was done by randomly selecting some of the taxpayers or by using a tax audit administration's attitude. In the past, various researches have been carried out to select fraudulent taxpayers for audit. In one organization, the tax audit department is required to audit some or all of its taxpayers to check the evasion of tax and ensure compliance. There are various tax audit computational techniques that have been performed for fraud detection in recent years by tax audit administration. Auditing all taxpayers are not a very welcome procedure for the tax administration. Therefore Tax administration agencies must use their limited resources very wisely to achieve maximal taxpayer compliance, minimum intrusion, and minimum costs (Gupta, 2014) Tax compliance refer here to the taxpayers fulfilling their registration, filing, and reporting and payment obligations correctly and at the right time. In reverse noncompliance is a taxpayer that did not fulfill the above obligations. This research proposed to design and develop a hybrid machine-learning model (non-supervised plus supervised machine learning) to predict taxpayers based on their fraudulent risk level. Specifically, hybrid techniques have shown their superiorities over single techniques. A machine-learning algorithm is the best cost-effective option to make taxpayer risk-based classification and effective by developing a standardized model to classify their status. A prediction model is built using historical taxpayer data and different attributes of the taxpayer. these experiment conducted by using the KDD process model, machine learning techniques such as K-means clustering and SVM are used for clustering and classification to detect fraudulent categorical events. First, by using SVM before clustering has accuracy of 97% and, in k-means clustering, 54.63% of the records correctly clustered. Second, we got an accuracy of 99% by using SVM after clustering.

Keywords:KDD process model, SVM, k-means clustering

TABLE OF CONTENTS

DECLARATION	iii
ACKNOWLEDGEMENTS	vi
ABSTRACT.....	vii
ABBREVIATIONS	xiii
CHAPTER ONE	1
INTRODUCTION	1
1.1. Background.....	1
1.2. Statement of the problem	3
1.3. Objectives of the study.....	5
1.3.1. General objectives	5
1.3.2. Specific objective	5
1.4. Scope and limitation of the study.....	6
1.5. Significance of the study.....	6
1.6. Organization of the Thesis	7
CHAPTER TWO	8
LITERATURE REVIEW	8
2.1. Machine learning concepts	8
2.1.1. Classification of Machine Learning Techniques	10
2.1.1.1. Supervised learning.....	10
2.1.1.2. Unsupervised learning	11
2.1.2. Hybrid approach.....	12
2.1.3. Application of machine learning	13
2.2. Research process models	15

2.2.1. The KDD process Model	15
2.3 The tax system in Ethiopia.....	16
2.3.1 Tax overview	16
2.3.2. Tax audit	18
2.3.2.1. Tax audit types.....	19
2.3.3.Tax Compliance	19
2.3.3.1. Determinant factors of tax compliance.....	21
2.4. Conceptual framework.....	22
2.5. Related works.....	24
CHAPTER THREE	26
RESEARCH DESIGN AND METHODOLOGY	26
3.1. Introduction.....	26
3.2. Methods and Algorithms of the fraudulent Level Prediction System.....	27
3.2.1. Data collection	27
3.2.2. Data Set Preparation	28
3.2.2.1. Data description	28
3.2.2.2. Data cleaning	31
3.2.2.3. Data transformation	31
3.2.2.4. Data reduction.....	31
3.2.2.5. Attribute selection.....	31
3.2.2.6. Data formatting	33
3.2.2.7. Threshold values	33
3.2.3. Clustering methods.	35
3.2.3.1. K-Means clustering algorithm	35

3.2.4. Classification methods	36
3.2.4.1. Support Vector Machine (SVM).....	37
3.3. Tool selection.....	39
3.4. The architecture of taxpayer risk prediction system.....	40
3.4.1 Row data	42
3.4.2. Preprocessing	42
3.4.3. Processed data	42
3.4.4. Training.....	42
3.4.5. Model building.....	43
CHAPTER FOUR.....	44
RESULTS AND DISCUSSION	Error! Bookmark not defined.
4.1. Introduction.....	44
4.2. Data Set Preparation	44
4.2.1. Data cleaning	44
4.2.2. Data transformation	44
4.2.3. Attribute selection.....	46
4.2.4. Data formatting	46
4.3. Experimental Setup.....	47
4.3.1. k-means clustering	47
4.3.2. SVM classification model building	51
CHAPTER FIVE	58
CONCLUSION AND RECOMMENDATION.....	58
5.1. Conclusion.....	58
5.2. Recommendation.....	59

References	60
APPENDICES	65
Appendix 1 Partial View of the Initial Collected Sample Data	65
Appendix II Sample code to load data	67
Appendix III train the data sets	68
Appendix IV Rescaling the data set	68
Appendix V: Train test split the dataset	69
Appendix VI Confusion matrix for cluster1	70
Appendix VII Confusion matrix for cluster 2	70

LIST OF TABLES

Table 2.1 Tax description (taken from MOR)	17
Table 3.1. Distribution of collected data concerning the taxpayers' risk level	27
Table 3.2. Data description	28
Table 3.3. The threshold value for risk level	34
Table 4.1 confusion matrix of classification with SVM	50
Table 4.2 classification result for experiment -1-	50
Table 4.3. Confusion matrix for experiment -2-	51
Table 4.4 clustering result for experiment -2-	51
Table 4.5 confusion matrix of classification for cluster -1- with SVM	54
Table 4.6. Classification result for cluster-1-	54
Table 4.7. confusion matrix of classification for cluster 2 with SVM	56
Table 4.8 classification result for cluster -2-	56
Table 4.9 confusion matrix of classification for cluster 3 with SVM	57
Table 4.10. Classification result for cluster 3	57

LIST OF IMPLEMENTATION CODES

Code 4.1 implementation for Data transformation	45
Code 4.2: Implementation confusion matrix with SVM.....	49
Code 4.3: Implementation confusion matrix with SVM.....	53
Code 4.4: Implementation confusion matrix with SVM.....	56

LIST OF FIGURES

Figure 2.1 Classification of Machine Learning Techniques (Simulink, 2018).....	10
Figure 2.2The KDD process Model (Mariscal, 2009)	15
Figure 2.3 Conceptual framework	Error! Bookmark not defined.
Figure 3.1 Architecture of taxpayer risk level prediction system	42
Figure 4.1 optimal cluster numbers	48

ABBREVIATIONS

ANN: Artificial Neural Network

CA: Comprehensive Audit

DA: Desk Audit

DT: Decision Tree

ERCA: Ethiopian Revenue Custom and Authority

FA: Full Audit

KDD: knowledge discovery from data

MOR: Ministry of Revenue

SIGTAS: Standard Integrated Government Tax Administration System

SVC: Support Vector Classifier

SVM: Support Vector Machine

TA: Tax Audit

CHAPTER ONE

INTRODUCTION

1.1. Background

Over the centuries and now a day the authority audits the business or an individual taxpayer by randomly select some cases among all taxpayers. And sometimes uses auditor's attitude for taxpayer selection. This procedure is not advisable and successful and harms the taxpayer and tax authority itself. It costs human resources, material resources, and times to the authority. The advancement of technology and research in the fields of data mining, machine learning, and artificial intelligence brings the need to automate taxpayers a risk-based selection and risk-level prediction for audit purposes.

Tax audit is an activity or a set of activities performed by Tax auditors to determine a taxpayer's correct tax liabilities for a particular accounting or tax period, by examination of a taxpayer's organization procedures and financial records to assess compliance to tax laws and verifying the true, fair, reliable, and accuracy of tax returns and financial statements.

Tax compliance is a problem facing many revenue authorities and has thus become the major focus of their operations. Enforcing taxpayers to obey tax laws is not always an easy task. Also, tax laws are not always precise and as a result, the state and taxpayers may have different interpretations of it. Moreover, taxpayers can dispute the meaning of the tax law depending on several factors, including their basic willingness to comply with a tax system (Chepkwony, 2017).

Tax fraud occurs when an individual or business entity willfully and intentionally falsifies information on a tax return to limit the amount of tax liability. Tax fraud essentially involves cheating on a tax return in an attempt to avoid paying the entire tax obligation.

There are different types of fraud performed by taxpayers, tax avoidance, and tax evasion. Tax evasion is one of the most common economic crimes and has been present since the

introduction of the tax. People don't like paying taxes, and they follow many ways to reduce their payment, some styles are legal and others are illegal. Evasion of Tax takes place when the people report tax dishonestly which includes declaring fewer gains profits, or income than what has been actually earned, and they even go for overstating deductions. The Evasion of Tax level depends on certain factors such as fiscal equation which means that people's tendency to pay fewer tax declines when the payment due from taxes becomes obvious. The level of Tax Evasion is also dependent on the tax administration's efficiency and corruption levels. (Chen, 2010)

Tax avoidance refers to an attempt to reduce tax payments by legal means. It is the legal usage of the tax rule in a single territory to one's advantage to reduce, modify and lower the amount of income tax owed the amount of tax that is payable by means that are within the law. This is generally accomplished by claiming the acceptable deductions and credits. (FriedrichSchneider, 2001)

Now a day many kinds of researches are released for taxpayer compliance status prediction using data mining and machine learning algorithms. However, all have their limitations. Some researchers faced the problem due to the limited set of taxpayer data and others use a single algorithm for fraud detection and classification. Currently, in the ministry of revenue, there is no standardized compliance status classification model that the risk assessment department uses. As a result, the risk department faced problems of compliance status identification. Due to this, it is necessary to develop a standardized taxpayer risk-based classification model to solve the above problem. Researchers have been conducted to develop a standardized taxpayer compliance status prediction model to solve the problems of audit and to design efficient and effective taxpayer auditing.

Nowadays interests of researchers are growing in the development of a model to the problems in the tax audit case selection. Therefore, this study aims to develop a taxpayer fraudulent risk-based prediction model using hybrid machine learning techniques that could improve the workflow and taxpayer auditing by solving the above-mentioned problems. To build this model support vector machine and k-means clustering are used for classification and clustering respectively, and python3 is used as a development tool. Then evaluates the

effectiveness of machine learning algorithms and methods. Machine learning is a method of data analysis that automates analytical model building. Machine learning provides methods techniques and tools, which help to learn automatically and to make accurate predictions based on past observations. Machine learning is popularly being used in areas of business like data analysis, financial analysis; stock market forecast, etc. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention. Machine Learning algorithms are used in data mining applications to retrieve hidden information that may be used in decision-making.

Machine learning helps the agency refine its traditional audit selection strategies to produce more accurate results. In the classification method, many attributes are involved. The reasons for selecting a subset of attributes instead of the whole dataset are; It is easier to measure only a reduced set of data of selected dataset, Prediction accuracy may be improved through the exclusion of redundant and irrelevant attribute, The predictor to be built is usually simpler and potentially faster when fewer input data are used and Knowing which attributes are relevant can give the accurate result of the prediction problem and allows a better understanding of the final classification(N.Radha, 2011).

The basic idea of machine learning is that a computer can automatically learn from experience. Although machine-learning applications vary, its general function is similar throughout its applications. The computer analyzes a large amount of data and finds patterns and rules hidden in the data.

1.2. Statement of the problem

The government laying and collecting taxes to fulfill the need of society, like education, health, and roads. And to full fill these needs the required amount of tax should be collected. Since all businesses and individual taxpayers are not compliant and not pay the expected amount of tax, auditing is mandatory. The tax authority in each year collects a large amount of money through audit investigation. Due to the trend of IT, and the reasons stated above, automated, standardized, and well organized fraudulent business groups and individual

taxpayer selection is required. In the past few years, studies have been conducted to develop fraudulent taxpayer classification model using different algorithms like artificial intelligence, data mining techniques, and machine learning to solve the problem and to improve the efficiency of tax audits.

In the MOR Bahir Dar branch, there are lots of taxpayers so it is difficult to audit all the taxpayers in one year because there are limited resources and a limited workforce. Besides, no need for auditing all taxpayers because it is time-consuming. So the tax department should select the taxpayer for audits based on their risk level, the one with high risk should be audited first. The tax department uses different attributes to identify risks. Currently, the risk department of MOR uses z-score, which is the statistical measurement of scores relation to the mean in a group of scores to select taxpayers to be audited. However this techniques is not effective and efficient, up to 20percent of the taxpayers are incorrectly classified. And it leads resources and time consuming to the tax authority.

In the context of ERCA, there are some attempts made byDanielM(mamo, 2013)to classify importers and exporters based on their fraudulent level. Here the researcher does not include Inland Revenue taxpayers.

BerisG(beris, 2017) attempts to classify taxpayers based on their risk level. But the researcher focuses only on E-filer taxpayers. In Ethiopia, E-filers are only limited to Addis Ababa revenue collecting offices.

BeleteB(Belete, 2011)also try to segment taxpayers based on their risk level using different clustering and classification techniques. The researcher uses importer and exporter data for segmentation.

However there is no work done to select taxpayers based on their fraudulent for audit purposes. Today there is an increase in transactions of goods in the country. So, the organization has to revisit its audit case selection strategy and should support it with new technology to satisfy its customers. And also In MOR, now an integrated system or model is being applied to classify customers based on their risk level.

Hence to solve the above problem develop a taxpayer classification model is better by using hybrid machine learning techniques. This will increase the accuracy of the classification.

The study will, therefore, answer the following research questions:

- ✓ How to identify appropriate tools and algorithms to develop taxpayer fraudulent prediction models?
- ✓ How to identify determinants factors that make taxpayer fraudulent?
- ✓ What are the approaches to design a taxpayer prediction model?
- ✓ How to evaluate the performance of the developed model?

1.3. Objectives of the study

1.3.1. General objectives

The general objective of this study to design and develop the Tax Payer risk prediction model for Auditing using hybrid machine learning that helps the organization to easily identify the risk level of the customers to audit based on their priority. This is done by clustering and classification customers' data according to their risk level.

1.3.2. Specific objective

With the above general goal in mind, the specific objectives the researcher will fulfill are described as the following.

- ✓ to identify appropriate tools and algorithms to develop taxpayer fraudulent prediction models?
- ✓ To identify determinant factors for non-compliance taxpayers.
- ✓ To analysis, the taxpayer risk prediction model design approaches.
- ✓ To evaluate the performance accuracy of the developed model.

1.4. Scope and limitation of the study

Different researchers perform taxpayer compliance prediction for audits at different times by using data mining and machine learning techniques. The tax departments have their own rules in each country. The scope of this research is limited to model development to classify tax payers based on their fraudulent risk level for the ministry of revenue and evaluate the developed model accuracy. The data source is limited to the ministry of revenue of Ethiopia. The limitation of this study is that we didn't develop a prototype due to a shortage of time. The developed model is will be used only for private limited company taxpayers.

1.5. Significance of the study

If the above-mentioned general and specific objectives are achieved the research will have the following contributions:

- To automate taxpayer compliance status prediction procedure: now the tax audit officer performs the audit procedure manually by using different taxpayer attributes and statistical measurement. By developing this model the research will make the taxpayer compliance status classification fully automate.
- To have standardized taxpayer compliance status prediction to all tax departments in the ministry of revenue of Ethiopia.
- To simplify the risk assessor officers work. Currently, it is a challenging task for the tax department officers to identify high-risk taxpayers using statistical measurement. So by developing this model it will solve their problem and will simplify their work.
- To increase the amount of tax to be collected in a given period. Because the more we audit the more revenue generation is occurring.
-

1.6. Organization of the Thesis

This thesis is organized into five chapters. The first chapter discusses the background of the study and statement of the problem. It also presents general and specific objectives of the study, the methodology used to accomplish the study, scope, and limitation of the research and application of the investigated results.

The second chapter reviews the different machine learning algorithms and techniques, such as clustering and classification with their respective algorithms, are reviewed.

Chapter three describes the proposed architecture of the taxpayer risk prediction model and the methods and algorithms of the taxpayer risk prediction model.

Chapter Four presents the experimentation phase of the study, which mainly discusses the different stages of the experiment to build the fraudulent prediction model and interprets the results of the clustering and classification experiments.

The final chapter, Chapter Five, presents the conclusion of the result of the study and provides recommendations based on the investigation of the research.

CHAPTER TWO

LITERATURE REVIEW

The tax administration is required to audit some or all its taxpayers to check the evasion of tax and ensure compliance. Tax administration agencies must, therefore, use their limited resources very wisely to achieve maximal taxpayer compliance, minimum intrusion and minimum costs. Now a day with the advancement of technology the tax department uses machine learning, data mining, and artificial intelligence to detect tax fraud. The development of the machine-learning model is motivated for the prediction of compliant and non-compliant taxpayers.

2.1. Machine learning concepts

Machine learning is sometimes defined as a subset of data mining – meant as the computational process of discovering patterns in large data sets – which itself is a subset of data analytics. It can be defined as the field of computer science that uses algorithms coming from the discipline of statistics to give computers the ability of "learning". This happens through the analysis of data and leads to the progressive improvement of the algorithm's performance on a specific task without the need to be explicitly programmed. Machine learning defined by the different researcher as follow: Simon(Simon, 1983)was defined as machine learning any process where a system improves its performance. A few years after Mitchell(Mitchell, 1997)defined a machine-learning algorithm as "any computer algorithm that improves its performance at some tasks through experience." In addition to data mining, machine learning is also considered to be close to the field of pattern recognition, which as the same word explains, focuses on the recognition of patterns meant as regularities in data. In these terms, this is a radical change in addressing IT problems. In the conventional approach, software programs are hard-coded by developers with specific instructions for the tasks that need to be executed. This can work well in most of the cases, but it has big limitations. It

assumes that human programmers can imagine every scenario and code instructions for any possible state of the world. However, if the environment is in an unpredicted state, the hard-coded software will not work well anymore and will stop working. By contrast, the idea of the machine learning approach is that ideally, it is possible to create algorithms that "learn" from data automatically. Thus, in case of changes in the environment, they can adapt to the new circumstances without needing to be explicitly programmed by human programmers. The idea is to give these algorithms "experiences" (training data) and a general strategy for learning, and finally let them identify patterns, associations, and insights from the data. In short, machine learning systems are trained instead of programmed.

2.1.1. Classification of Machine Learning Techniques

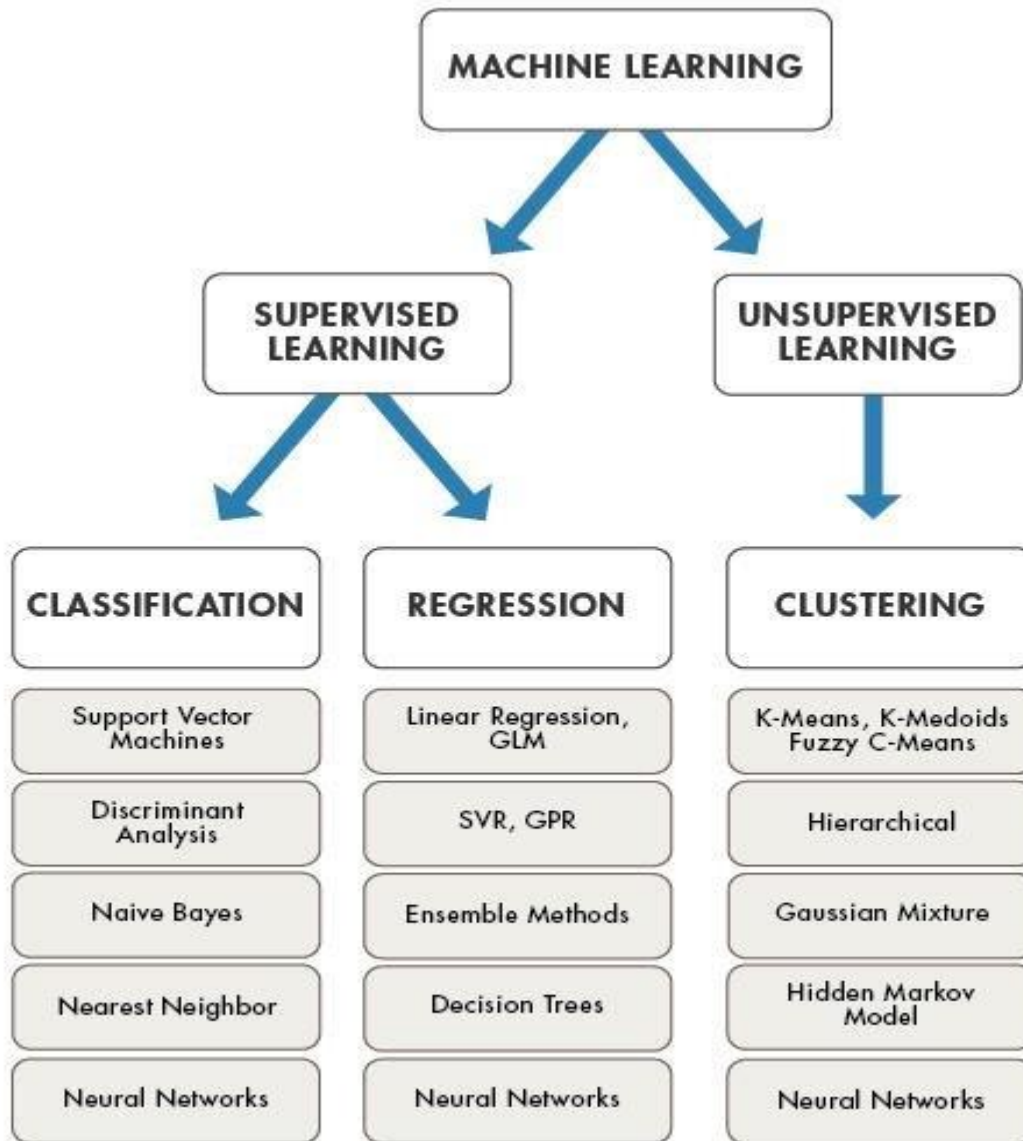


Figure 2:1 Classification of Machine Learning Techniques (Simulink, 2018)

2.1.1.1. Supervised learning

Supervised learning became an area for a lot of research activity in machine learning. Many of the supervised learning techniques have found application in their processing and analyzing a variety of data. One of the main characteristics is that supervised learning has the ability of

annotated training data. The so-called labels are class in the classification process. There is a variety of algorithms that are used in the supervised learning methods(Nasteski, An overview of the supervised machine learning methods, 2017)Classification is a standout amongst the most generally utilized strategies for mining in medicinal services organization(Tomar, 2013). Classification charges the set of examples being mined, which is separated into fundamentally unrelated and thorough sets. These sets are known as the preparation set and the test set. The arrangement procedure is correspondingly separated into two stages: preparing, when a characterization show is worked from the preparation set, and testing when the model is assessed on the test set. There are several algorithms used for classification in data mining such as KNN, ANN, Naïve Bayes, Decision tree, Support Vector Machine, Weighted Associative Classifier (WAC)(Neelamegam, 2013).

The learning process in a simple machine learning model is divided into two steps: training and testing. In the training process, samples in training data are taken as input in which features are learned by learning algorithms or learners and build the learning model. In the testing process, the learning model uses the execution engine to predict the test or production data. Tagged data is the output of the learning model which gives the final prediction or classified data.(Nasteski, 2017).

2.1.1.2. Unsupervised learning

Unsupervised learning algorithms aim at describing data and extracting knowledge from data without access to a labeled training set. The class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance. As a background for the research presented in this thesis, in this chapter, we describe the most classical type of unsupervised learning, namely clustering (Kamber, 2003). Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable class's .Clustering is also called unsupervised classification because the classification is not dictated by given class labels. There are many clustering approaches all

based on the principle of maximizing the similarity between objects in the same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

Clustering is typically based on the notion of similarity. Given a dataset and a measure of similarity, a clustering algorithm aims at identifying subsets (clusters), such that the similarities between pairs of data points from the same cluster are high and the similarities between pairs not from the same cluster are low. However, clustering algorithms are not necessarily based on this notion.

The combination of the facts that (i) the number of applications of clustering is large, (ii) clustering is both task and data-dependent, and (iii) clustering is subjective and no universal definition of a cluster exists, has resulted in a large number of different clustering algorithms. There are several possible ways to categorize clustering methods, for example, into hard and soft clustering algorithms. In hard clustering, each data point belongs completely to one cluster. On the other hand, in soft clustering, each data point potentially can belong to multiple clusters, i.e. the probability of being a member of a cluster can be non-zero for more than one cluster.

2.1.2. Hybrid approach

In the past lot of hybrid machine learning systems were developed to bring the best from the two different machine learning methods. For example, a hybrid machine learning system is created based on genetic algorithm and support vector machines for stock market prediction by Rohit and Kumkum (Rohit Choudhry, 2008). Nerijis, Ignas and Vida developed a hybrid machine learning approach for text categorization using decision trees and artificial neural network (Nerijus Remeikis, 2007). Sankar developed an integrated data mining approach for maintenance scheduling using case-based reasoning and artificial neural network (Sankar, 2017).

Hybrid Machine Learning Approaches: there are two types of commonly-used hybrid models, classification + classification and clustering + classification hybrid approaches. In most of the emerging applications, it is clear that a single model used for classification doesn't behave efficiently, so multiple methods have to be combined giving result to hybrid models.

2.1.3. Application of machine learning

Machine learning has been extensively applied in various application domains. Some of the most popular applications include medical diagnosis, credit risk analysis, customer profiling, market segmentation, targeted marketing, retail management, and fraud detection (George Tzani, 2014).

Medical diagnosis: -One application of machine learning in a healthcare context is digital diagnosis. ML can detect patterns of certain diseases within patient electronic healthcare records and inform clinicians of any anomalies (Gharagozyan, 2019).

Credit risk analysis: -The credit risk analysis is a major problem for financial institutions, credit risk models are developed to classify applicants as accepted or rejected concerning the characteristics of the applicants such as age, current account, and amount of credit. In the present investigation, we will apply four classification models to evaluate their performance and compare it with other previous investigations (Melendez, 2019).

Market segmentation: -Market segmentation has been described as a process of dividing the market into internally homogeneous groups that appear distinct concerning the other groups. In essence, the market segmentation approach recognizes that the total market demand for art offerings is essentially heterogeneous and, therefore, it can be disaggregated into segments with different needs and preferences (Tajtakova, 2009).

2.1.4 Machine Learning Common Challenges

In this section, the common problems of machine learning are provided: the bias-variance tradeoff, under/over fitting, high dimensionality, and big data. When deploying supervised learning algorithms, the error that an algorithm makes can be broken down into three components: bias, variance, and irreducible error(Huang, 2008). While the last component cannot be controlled, the first two can be influenced by tuning the algorithm parameters. Bias is about how consistently the model is "right" or "wrong," compared to the truth. On the other hand, the variance expresses how "smooth" the model is: larger variance indicates that small changes in the dataset can lead to radical changes in the outcomes.

Usually, to try to increase the accuracy of a supervised learning model, it is possible to reduce the bias, but this will also tend to increase the variance and vice versa. Thus, the ultimate goal is to find the optimal balance between the two variables of error. Over fitting is also related to the bias-variance trade-off within supervised learning and refers to the idea that a machine learning model can be trained "too much" so that its optimal performance on the training set may result in suboptimal performance on a separate test set and real-life data(Ivanovic,Radovanovic, 2015). This is because the model became overly complex compared to reality. It may be a consequence of a small or large number of training instances, noisy data, and/or high dimensionality, under fitting is the opposite extreme, where the derived model is too simple compared to the reality, and thus not able to accurately predict the right outcome in real situations.

The last common challenge when developing machine-learning systems is called high dimensionality. Often datasets have a large number of rows – representing the instances – and/or a large number of columns – representing the features of the model (Ivanovic, 2015). High dimensionality refers to the high number of columns since the rule of thumb is to have at least 5 training examples for dimension. Finally, big data and its processes are the main challenges for machine learning projects and in general big data analytics projects.

2.2. Research process models

There are several process models in research like; CRISP process model, KDD process model, six-step Cios model, and design science research process model. Among these in this research, we have used the KDD process model. Because this model is best fit with our research domain.

2.2.1. The KDD process Model

The KDD refers to the overall process of discovering useful knowledge from data. The basic problem addressed by the KDD process is mapping low-level data into other forms that might be more compact, more abstract, or more useful. It is defined in another way as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, 1996). The KDD as a process is defined as interactive and iterative, involving nine-steps with many decisions made by the user. These steps are shown in figure 2:2 below

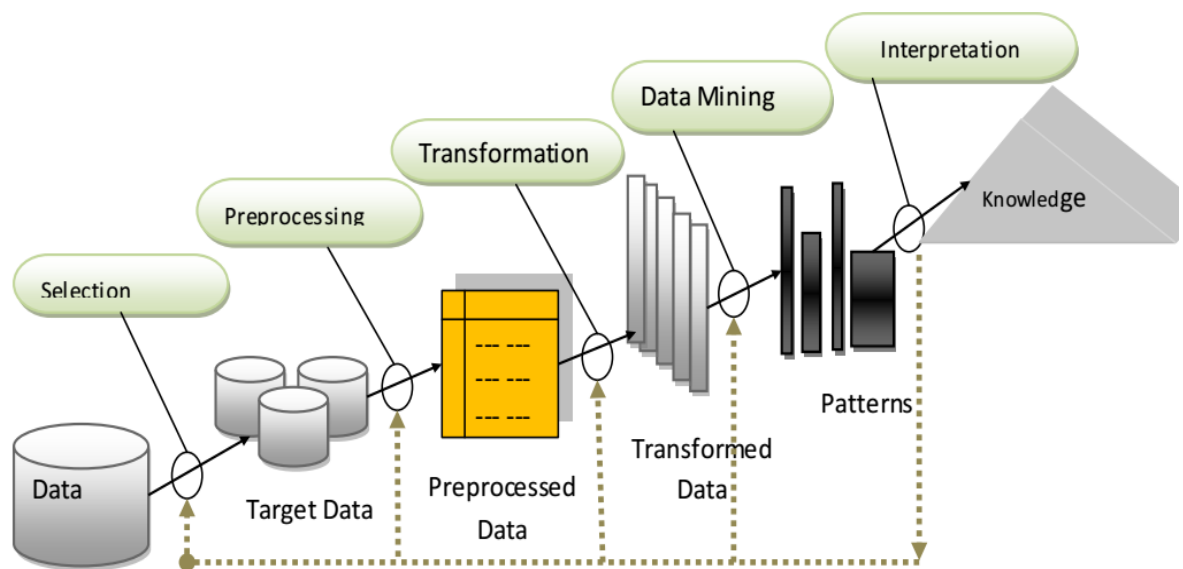


Figure 2:2The KDD process Model (Mariscal, 2009)

Developing and understanding the application domain: learning the relevant prior knowledge and goals of the application
Creating a target dataset: includes selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed
Data cleaning and preprocessing: this step includes operations such as removing noise and outliers, decide on strategies of handling missing field and deciding on database management issues.
Data reduction and projection: finding useful features to represent the data depending on the goal of the task and applying dimensionality reduction and transformation methods to reduce the data
.Choosing the data mining task: deciding on the function of the model derived (classification, clustering).

Choosing the data-mining algorithm: selecting methods to be used for searching the patterns
Data mining: searching for patterns by running the selected data mining algorithms on the prepared data.
Interpreting mined patterns: possible visualization of the mined patterns, removing redundant or irrelevant ones and translating the useful ones into terms understandable by the user
Consolidating discovered knowledge: includes incorporating the discovered knowledge into the system, taking actions based on it and resolving potential conflicts with the previous and new knowledge(Mariscal, 2009).

2.3 The tax system in Ethiopia

2.3.1 Tax overview

Tax: a tax is a mandatory financial charge or some other type of levy imposed upon a taxpayer (an individual or other legal entity) by a governmental organization to fund various public expenditures. A failure to pay, along with evasion of or resistance to taxation, is punishable by law. Taxes consist of direct or indirect taxes and may be paid in money or as its labor equivalent. MOR has divided the taxes as, direct and indirect tax. Direct tax means collected from the gain of income (net income). Indirect tax means collected from sales or production or value add-in product not collect from income (net income)(MelatA, 2016).

Direct taxes are:-

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

Indirect taxes are: -

- VAT
- Excise
- Turnover...

Tabular description of different taxes

Table 2:1 Tax description (taken from MOR)

No	Direct tax	Tax description
1	Business Profit tax	A tax is imposed on commercial, professional, or vocational activity or any other activity recognized as trade by commercial code of Ethiopia.
2	Personal Income tax	A tax is imposed on the employee on a monthly income or basic salary.
3	Rental Income	A tax that is imposed on the income from the rental of buildings.
4	Dividend	A tax imposed on shareholders when dividing the profit of a company.
5	Royalty	A tax imposed on the gain from innovation, patent, copyright, official document...etc
6	Technical service	A tax imposed on the gain from technical services when service provider company is not legally registered in Ethiopia.
7	Game of chance	Every person deriving income from winning at games of chance/for example, lotteries, tombola, and other similar activities.
No	Indirect tax	Tax description

1	Value-added tax	Tax on added value only (or compensate the paid amount that will be deducted before the payment).
2	Excise tax	Tax on production cost, consumption, and luxury goods.
3	Turnover tax	Tax on total sales the other name of turnover tax is sales tax.

2.3.2. Tax audit

Audit: Audit is defined as the structured examination of business-relevant commercial systems financial, non-financial records, physical stock, and other assets, internally generated data, and that produced independently of the business. A (tax) audit is a detailed exploration of the activities of a taxpayer to determine whether he/she has been correctly declaring the tax liabilities. Audits indirectly drive voluntary compliance and directly generate additional tax collections, both of which help tax agencies to reduce the ‘tax gap’ between the tax due and tax collected. Audit plays a pivotal role in the administration of tax and achieving the revenue objectives, ensuring the fiscal health of the country, and ensures a level playing field for an honest taxpayer.(Nagadevara, 2010)Develop a model through machine learning technique were better than random selection.

Tax audit: a tax audit is an examination of whether a taxpayer has correctly reported its tax liability and fulfilled other obligations. It is often more detailed and extensive than other types of examinations such as general desk checks, compliance visits, or document matching programs. A tax audit may increase tax revenue in two ways: directly through assessment of additional taxes, and indirectly by improving taxpayer compliance with the tax laws and regulations. Tax audit results in increased tax revenue in two ways:

- Directly through assessment of additional taxes;
- Indirectly by discouraging underreporting of liabilities by all taxpayers.

The purpose of the tax audit is to check the evasion of tax and ensure compliance following the laws and regulations.(Mihret, 2011).Tax audit is a critical and significant component of the compliance activities of tax administration using proper use of enforcing tax laws; it is the conduct by audit staff for the appropriate verification of selected taxpayer’s whether

he/she has been correctly declaring the tax liabilities including a review of taxpayer's systems, books of account and other related information. It may include crosschecks of taxpayer's records with those of taxpayer's suppliers or with other government departments and agencies source of information and its effectiveness and efficiency must be guaranteed to utilize proper procedures and application of modern audit tools and techniques (SENBETA, 2018).

2.3.2.1. Tax audit types

Desk audit or verification: this type of audit can be conducted to the specific issue audits of a small enterprise or employee when the auditor is confident that all necessary information can be ascertained by accompanying the examination in the office.

Field Audit: It is a detailed examination of taxpayers' books and records to determine whether the correct amounts were reported on the tax returns. The auditor may also obtain information from other sources such as banks, creditors, and suppliers, to confirm items on returns.

Refund audit: Verifying the taxpayer is right to a refund before processing the refund. Usually undertaken for first refund claims as well as where the refund claim varies significantly from established patterns and trends. Refund audit carried out particularly for new registrants; also, it should emphasize only on the period covered by the claim.

Comprehensive or full audit: This audit covers all tax obligations over the number of tax periods, or extended to several years up to the limit provided for in the law. The objective is to determine the correct tax liability for a tax return as a whole. It typically entails a comprehensive examination of all information relevant to the calculation of a taxpayer's tax liability for a given period.

2.3.3. Tax Compliance

Tax compliance reflects the level of willingness of a community to meet its tax obligations following applicable regulations. Tax compliance is a major problem for many tax authorities and it is not an easy task to persuade taxpayers to comply with tax requirements even though

'tax laws are not always precise' (James and Alley (James, 2004)). The exact meaning of tax compliance has been defined in various ways. For example, Erard, and Feinstein (Andreoni, 1998) claimed that tax compliance should be defined as taxpayers' willingness to obey tax laws to obtain the economic equilibrium of a country. Kirchler (Kirchler, 2007) perceived a simpler definition in which tax compliance is defined as the most neutral term to describe taxpayers' willingness to pay their taxes. Song and Yarbrough (Yarbrough, 1978) suggested that due to the remarkable aspect of the operation of the tax system in the United States and that it is largely based on self-assessment and voluntary compliance.

Tax compliance should be defined as taxpayers' ability and willingness to comply with tax laws, which are determined, by ethics, legal environment, and other situational factors at a particular time and place. Similarly, several tax authorities also define tax compliance as the ability and willingness of taxpayers to comply with tax laws, declare the correct income each year, and pay the right amount of taxes on time. Besides, Jackson and Milliron (Milliron J. a., 1986) defined tax compliance as the reporting of all incomes and payment of all taxes by fulfilling the provisions of laws, regulations, and court judgments. Another definition of tax compliance is a person's act of filing their tax returns, declaring all taxable income accurately, and disbursing all payable taxes within the stipulated period without having to wait for follow-up actions from the minister. Compliance in pure administration terms, therefore, includes registering or informing tax authorities of status as a taxpayer, submitting a tax return every year (if required), and following the required payment periods.

Bhupalan and Somasundram (Somasundram, 2003) claimed that the wider perspective of compliance becomes a major issue in a self-assessment system since the total amount tax payable is highly dependent on the levels of tax compliance this perspective reveals, although it is inevitable that tax authorities will seek to 'influence' the areas taxpayers influence determining to reduce the risks of non-compliant behavior they face otherwise e.g. through continuously conducting tax audits of different sorts and other means such as various compliance influencing activities including tax education. Some authors have viewed tax compliance from a different perspective.

For example, Allingham and Sandmo (Allingham, 1972) described tax compliance as an issue of ‘reporting an actual income’ and also claimed that tax compliance behavior was influenced by a situation whereby taxpayers have to decide uncertainty Clotfelter (Clotfelter, 1983) i.e. either taxpayer would enjoy tax savings due to under-reporting income or have to pay tax on the undeclared amount at a penalty rate which is higher than they would have paid had the income been fully declared at the correct time.

Therefore, taxpayers’ tax knowledge and compliance behavior is an important issue for any government and revenue collecting minister to obtain knowledge and understanding of the taxpayer attitude and tax compliance behavior particularly in a self-assessment environment. Tax compliance is the multi-faceted measure and theoretically, it can be defined by considering three distinct types of compliance such as payment compliance, filing compliance, and reporting compliance. The issue of taxpayers’ tax knowledge and compliance behavior had received great attention around the world. A good understanding of taxpayers’ tax knowledge is important for the tax authority to improve the tax system and consequently encourage taxpayers’ compliance.

2.3.3.1.Determinant factors of tax compliance

Different researcher as follow defines different determinant factors.

According to Jackson and Milliron (Jackson, 1986) the main factors that have influenced tax compliance as argued by various researchers are age, gender, level of education, income, status, peers’ or other taxpayers’ influence, ethics, legal sanction, complexity, relationship with taxation authority, income sources, perceived fairness of the tax system, possibility of being audited and tax rate.

in addition to the above, researchers have enumerated factors that influenced tax compliance behavior such as demographics, income, compliance cost, and tax agents, in addition to moral or ethical factors (Singh, 2003).

Nurlis (Nurlis, 2015) in his article, examines the outcome of taxpayer awareness, knowledge, tax penalties, and service tax authorities on tax compliance a survey held on the individual taxpayer at Jabodetabek & Bandung in Indonesia. Study data were collected using the accidental sampling method that is questionnaires were distributed to those who visited tax offices of Jabodetabek & Bandung individual taxpayers". The result of the study indicated that awareness of the taxpayer has a positive and significant effect on individual taxpayer compliance. Besides, tax knowledge of taxpayers" also has a negative and significant relationship of taxpayer compliance. This indicates that the level of knowledge of good tax looking for gaps to avoid tax liabilities, tax penalties have a positive and significant relationship to the individual taxpayer compliance which performs at the tax office in the area.

Further, the study displays that the more effective application of tax penalties, the tax compliance rate will be higher. Service tax authorities have a positive and significant relationship of compliance with individual taxpayers" that performs at the tax office in the country.

Finally, the study proposes that the better the service tax authorities, the tax compliance rate will be higher Jackson and Milliron (Milliron, 1686) reviewed 46 tax compliance articles and identified eleven important factors that have been examined by researchers. These are tax system complexity, level of tax information services, withholding and information reporting, and tax return preparer responsibilities and penalties. Others include the probability of being audited, progressively and actual level of tax rates, penalties for non-compliance, age, gender, education, and income.

Low.J(Low, 2017) examined four tax compliance determinants Tax Rate, probability of being Audited, non-Complexity of the tax system, and the probability of detection.

Aemiro.T(Aemiro, 2014) defines this study examined the determinants of tax compliance behavior in Ethiopia particularly in Bahir Dar city administration. The data were collected using a structured questionnaire. The results revealed that perception on government spending; perception on equity and fairness of the tax system; penalties; personal financial

constraint; changes on current government policies; and referral group (friends, relatives, etc.) are factors that significantly affect tax compliance behavior.

Eshetu Y (ESHETU, 2018) defines the determinants of tax compliance. The results showed that tax knowledge, perception on fairness, and equity and role of tax authority were factors that positively significantly affect tax compliance behavior. However, the complexity of the tax system was negatively significantly affected tax compliance behavior of taxpayers.

Manaye M (Manaye, 2018) The study results from the survey conducted in the study area using 290 respondents indicate that tax compliance was influenced by the probability of being audited, financial constraints, and changes in government policy among thirteen potential determinants of tax non-compliance were examined in this study. The taxpayers should always be treated equally, involve taxpayers (which are conclusive stakeholders in this tax system) on each of the tax issue and should work jointly, the tax authority needs to be strong enough as well as should be perceived as powerful by the taxpayers and voluntary compliance is enhanced when the tax authority administers the law fairly and the authority needs to strengthen itself by educating and training its employees both at home and abroad were few of the recommendations and implications forwarded to the tax authority in the study area in particular and the country in general based the findings and discussions.

Melese K (kebede, 2018) defines the following determinant factors of tax compliance: Economic Factors (Economic factors with tax compliance refer to actions that are associated with the costs and benefits of performing the actions. The tax compliance determinants associated with economic factors such as tax rates, tax audits, and perceptions of government spending are explored in more detail. Social Factors: Tax compliance determinants from a social perspective relates to taxpayers' willingness to comply with tax laws in response to other people's behavior and their social environment.

2.4. Related works

Many works have been done in taxpayer compliance status prediction. Here we will look at some of the related works are discussed below.

The researchers Chepkwony and Chepkurui.C(Chepkwony C. C., 2017) developed a classification model using historical taxpayer audit data and decision tree algorithm to predict the compliance status of taxpayers in a case-selection application prototype. Experimental results using limited taxpayer data and one algorithm so the model should be improved by increasing a set of data and several algorithms.

The researchers Solon.Land Henrique de(Leon Solon da Silv, 2016) Showan essay of the Tax Administration on using Bayesian networks to predict taxpayers' behavior based on historical analysis of income tax compliance and Tried to improve a previous risk-based audit selection which detects a large number of taxpayers as high risk. Results are promising, considerably improve tax audit performance. However, in this research financial transactions and invoice data are not included as a variable.

Martikainen and Jani(Martikainen, 2012) have looked into the possibilities of using data mining, or more broadly analytics, in tax administrations to enhance tax compliance. Enhancing tax compliance is about making tax issues easy to deal with for the taxpayers, helping those who have difficulties, and ensuring that the taxpayers fulfill the registration, filing and reporting requirements, and pay their lawful share.

Beris.G(beris.G, 2017) performed mining e-filing Data for Predicting Fraud: The Case of Ethiopian Revenue and Customs Authority. He used different data mining algorithms to predict fraud in the E-filling system, the following algorithms are used for this; J48 Decision Tree Modeling, Experiment Random Forest, and Neural Network and test mode of 10-fold cross-validation and percentages split mode. Also, he concludes that J48 Decision Tree Modeling is best for prediction with an accuracy of 94.719% by using 10-fold cross-validation.

Belete.B (Biaz, 2011) attempted knowledge discovery for effective customer segmentation for ERCA by using the customs ASYCUDA database. He used the K-means clustering algorithm for clustering and J48 decision tree and multi-layer preceptor ANN algorithms for

classification. Using the J48 decision tree algorithm with default 10-fold cross-validation shows better performance, which is 99.95 percent of the overall accuracy rate.

Dani M. (Mama, 2013) attempted the clustering algorithm followed by classification techniques for developing the predictive model, K-Means clustering algorithm is employed to find the natural grouping of the different tax claims as fraud and non-fraud and developed using the J48 decision tree algorithm.

The above researchers use clustering and classification techniques in their specific domain and their accuracy percentage.

The main objective of this research study is to apply the hybrid machine learning techniques to classify taxpayers based on their risk level (low risk, medium risk, and high risk), For creating a classification model that, determines the compliance and non compliance behavior of taxpayers to develop an effective tax collection in the ministry of revenue.

CHAPTER THREE

RESEARCH DESIGN AND METHODOLOGY

3.1. Introduction

This chapter specifies the model that was used in the study, it also provides an overview of the methods that were used to collect analyses and interpret the data to generate findings necessary to answer the research questions.

Research design is the conceptual structure within which research is conducted. It constitutes the blueprint for the collection, measurement, and analysis of data. Research design is needed because it facilitates the smooth sailing of the various research operations, thereby making research as efficient as possible yielding maximal information with minimal expenditure of effort, time, and money (Kothari, 2009). In this study, experimental research is conducted.

A research methodology is a way to systematically solve the research problem and may be understood as a science of studying how research is done scientifically. Research the methodology outlines the various steps that are generally adopted by a researcher in studying his research problem along with the logic behind them (Kothari, 2009).

In this proposed research, we developed the Tax Payer fraudulent Risk Level Prediction Model using hybrid machine learning techniques. This system applied to Ethiopian taxpayers.

3.2. Methods and Algorithms of the fraudulent Level Prediction System.

3.2.1. Data collection

In this research, the main sources of data are the MOR SIGTAS database. The researcher collected all the four-year transaction records, which was around 8140 records. To make the data confidential Bahir Dar University wrote an official letter for the ministry of revenue. For taxpayer privacy, the name of the taxpayer is omitted during data collection.

The number of records collected from the SIGTAS database is summarized in Table 3.1

Table 3:1. Distribution of collected data concerning the taxpayers' risk level

Risk level	No- of taxpayer
High	423
Medium	2083
Low	5634
Total	8140

3.2.2. Data Set Preparation

It is making the data ready for the next task, clustering, and classification. The Major Tasks in Data Preprocessing are:-

3.2.2.1. Data description

MOR Risk selection Criteria and their descriptions

Table 3:2 Data description

No.	Attributes	Weight	Risk Interval	Risk Level	Description	Data type
1	start date	2	>3 years	1	Date of the business starts	DATE
			>2<=3years	2		
			<=2 years	3		
2	sales	6	<=4,000,000 birr	1	Size of Business / Turnover of business	VARCHAR
			>4,000,000<=7,500,000 birr	2		
			>7,500,000 birr	3		
3	Industry of business	4	Agriculture, Hunting, Forestry and Fishing, Electricity, Gas and Water Supply, Community, Social and Personal Services, Private Household Exterritorial, Non-governmental Organizations Representatives of Foreign Governments and other activities not adequately defined	1	The types of business performed	VARCHAR
			Manufacturing, Wholesale and Retail Trade, Repair of Motor Vehicles, Motor Cycles and Personal and Household Goods; Hotels and Restaurants,	2		
			Import and Export, Financial Intermediation, Insurance, Real Estate, and Business Services, Construction, Mining and	3		

			quarrying			
4	no_of_branch	2	< 1	0		VARCHAR
			=1	2		
			>1	3		
5	foreign_branch	3	Not having foreign branch offices	0	Number of branch out of the country	VARCHAR
			Having foreign branch offices	3		
6	sister_company	5	<1	0	Number of sub branches	VARCHAR
			=1	2		
			>1	3		
7	penalty issues	4	Penalty <=3	1	Number of penalties due to late payment and late filling in the last two years	VARCHAR
			Penalty >3<=5	2		
			Penalty >5	3		
8	last audit year	4	<=2 year	1		VARCHAR
			>2<=3 years	2		
			>3 years	3		
9	assessment difference	4	<=5%	0	Assessment's Audit Case Assessment's Results: - % increase in on main tax audit findings	VARCHAR
			>=5%	1		
			>10%<=25%	2		
			>25%	3		
10	refund	4	<=5%	0	Return due to over payment	VARCHAR
			5%>x<=8%	2		
			>8%	3		
11	loss declaration	6	<1	0	Declare with null value	VARCHAR
			=1	2		
			>1	3		
12	profit margin	5	Deviation <=2%	0	Average % of Gross Profit per SIC) when the deviation is negative	VARCHAR
			Deviation >2% and <=5%	1		
			Deviation >5% and <=10%	2		
			Deviation >10%	3		
13	change_in_asset	4	Deviation <=5%	0		VARCHAR
			Deviation >5% and <=10%	1		
			Deviation >10% and <=20%	2		
			Deviation >20%	3		

14	change_in_liability	4	Deviation <=5%	0	Average % of the change in total liability from the previous year) when the deviation is positive	VARCHAR
			Deviation >5% and <=10%	1		
			Deviation >10% and <=20%	2		
			Deviation >20%	3		
15	change_in_sales	5	Deviation <=5%	0	Change in Sale/Turnover from the previous year) when the deviation is negative	VARCHAR
			Deviation >5% and <=10%	1		
			Deviation >10% and <=15%	2		
			Deviation >15%	3		
16	sales_and_expense	4	Deviation <=5%	0	Change in-Normal Average % of each individual expense compared to total sales/turnover per SIC	VARCHAR
			Deviation >5% and <=10%	1		
			Deviation >10% and <=20%	2		
			Deviation >20%	3		
17	tax_payable	2	Deviation <=5%	0	Change in Tax payable from previous year	VARCHAR
			Deviation >5% and <=10%	1		
			Deviation >10% and <=20%	2		
			Deviation >20%	3		
18	audit_opinion	2	Unqualified opinion	0	Audit Opinion from External Auditors	VARCHAR
			Qualified opinion	1		
			Adverse opinion	2		
			Disclaimer of opinion	3		
19	tax_holiday	2	<1 year	0	return without payment (nil, credit & loss) after tax-holiday time passes	VARCHAR
			=1 year	2		
			>1 year	3		
20	intelligence	6	If there is no finding	0	3rd party tax information	VARCHAR
			If there is any finding	3		
21	custom_profile	4	Green	1		VARCHAR
			Yellow	2		
			Red	3		

3.2.2.2. Data cleaning

Additionally known as data cleansing. It deals with error detection and removing from information to improve the quality of information. The data cleaning step includes activities such as removing unnecessary data values or attributes and filling missing values. Data cleaning has become necessary to improve the quality of data. Attributes, which have a low contribution for classification, are deleted.

3.2.2.3. Data transformation

Data transformation operations are additional procedures of data pre-processing that would contribute toward the successes of the classification process and improve our results. Some of the Data transformation techniques are Normalization, Differences, and ratios and Smoothing. There are many tools like java, python, Matlab, and others.

3.2.2.4. Data reduction

For large datasets, there is an increased probability that an intermediate, data reduction step should be performed before applying data mining techniques. While massive datasets have the potential for higher mining results, there is no guarantee that they will produce better knowledge than small datasets. Data Reduction obtains a reduced data set representation that is much smaller in volume; however, it produces constant analytical results. In this research we python to reduce the data.

3.2.2.5. Attribute selection

The attribute selection activities involve selection and identification of best variables or attributes for predicting risk level. There are many attribute evaluation methods to select the best attribute, GainRatioAttributeEval(evaluate attribute based on gain ratio), InfoGainAttributeEval, ChiSquaredAttributeEval(compute the chi-squared statistics of each

attribute concerning the class)The attributes used in this study were ranked in order of importance using ChiSquaredAttributeEval. Attribute selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested.

Having too many irrelevant features in your data can decrease the accuracy of the models.

Three benefits of performing feature selection before modeling your data are:

Reduces Over fitting: Less redundant data means less opportunity to make decisions based on noise.

Improves Accuracy: Less misleading data means modeling accuracy improves.

Reduces Training Time: fewer data means that algorithms train faster.

3.2.2.6. Data formatting

At this step, the researcher changes the data into a format, which was suitable for the machine learning algorithms.

3.2.2.7. Threshold values

These are values that determine the low, medium, and high-risk level of the taxpayer.

The risk is determined by using the weight of each attribute and the risk level assign for each attribute. If the sum of each features risk value is less than 66 it will be considered as low risk And if the sum is 66 greater than 66 and less than 106 it will be considered as medium risk and if it is 106 and greater than 106 it is high risk. This is taken from ministry of revenue audit department.

Table 3:3the threshold value for risk level

start_date	sales	industry_of_business	penalty_issues	Last_audit_year	refund	loss_declaration	Profit_margin	Change_in_asset	change_in_liability	change_in_sales	sales_and_expense	tax_payable	audit_opinion	Intel_ligence	custo_m_profile
=2, low risk	=6 low risk	=4 low risk	=4 low risk	=4 low risk	=0 low risk	=0 low risk	=0 low risk	=0 low risk	=0 low risk	=0 low risk	=0 low risk	=0 low risk	=0 low risk	=6 If there is findi ng and =0 if there is no findi ng	=4 Low risk
=4 Medi um risk	=12 Med ium risk	=8 Mediu m risk	=8 Mediu m risk	=8 Med ium risk	=8 Mediu m risk	=12 Medium risk	=10 Mediu m risk	=8 Mediu m risk	=8 Mediu m risk	=10 Mediu m risk	=10 Medium risk	=4 Medi um risk	=4 Medium risk		=8 Medi um risk
6 High risk	=18 High risk	=12 High risk	=12 High risk	=12 High risk	=12 High risk	=18 High risk	=15 High risk	=12 High risk	=12 High risk	=15 High risk	=15 High risk	=6 High risk	=6 High risk		=12 High risk

3.2.3. Clustering methods.

Clustering is a tool for data analysis, which solves classification problems. Its objective is to distribute cases (people, objects, events, etc.) into groups, so that the degree of similarity can be strong between members of the same cluster and weak between members of different clusters (et, 2010). In clustering, there is no pre-classified data and no distinction between independent and dependent variables. Instead, clustering algorithms search for groups of records (the clusters composed of records similar to each other). The algorithms discover these similarities. There are different types of clustering algorithms like K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Agglomerative Hierarchical Clustering. In our research, we have used k-means clustering to cluster the data set. We have used different Comparison Metrics like: (Chatterjee, 2019).

- ✓ size of dataset
- ✓ number of clusters
- ✓ type of dataset and type of software used
- ✓ performance of the algorithm
- ✓ accuracy of the algorithm
- ✓ quality of the algorithm

To select clustering algorithms. Based on the above criteria k-means clustering is the best fit for our research.

3.2.3.1. K-Means clustering algorithm

Clustering is one of the most common exploratory data analysis techniques used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups

in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different(Dabbura, 2018).

The k-means algorithm takes the input parameter, k , and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low (Han and Kamber, 2006) Cluster similarity is measured regarding the mean value of the objects in a cluster, which can be viewed as the cluster's Centroid or center of gravity. The k-means algorithm works as follows(Dabbura, 2018).

1. Specify the number of clusters K .
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e. assignment of data points to clusters isn't changing.
4. Compute the sum of the squared distance between data points and all centroids.
5. Assign each data point to the closest cluster (centroid).
6. Compute the centroids for the clusters by taking the average of all data points that belong to each cluster.

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

3.2.4. Classification methods

Classification phase is the decision making the phase of taxpayer risk level prediction this phase uses the features extracted in the previous stage for deciding the class membership. There many machine learning classification algorithms like Linear Classifiers: Logistic Regression, Naive Bayes Classifier, Nearest Neighbor, Support Vector Machines, Decision Trees, Boosted Trees, Random Forest, and Neural Network. In this work, we have

used Support Vector Machine (SVM) classifier. SVM is a very useful technique for data classification and has been widely applied in pattern recognition (Burgess, 1998) for classification we used SVM in this study. To select the classification algorithm we consider different parameters like data type, data set size, accuracy level, and time required to process (A. Lourdu Caroline, 2018). As an example, KNN takes time to process and performance is low at a large data set. In Bayes Classification Algorithm the accurate result of the algorithm is decreased when the data is low. To get a good result it requires a large amount of data. In decision tree the working process is slow in the data set and it is affected by over fitting. In Artificial Neural Network the input data is in the range of 1 and 0. It took a long time to understand the data. The processing time is high because of the large amount of data. SVM used in numeric prediction and classification. It is high in accuracy level and it is good for small data set. Based on the above parameters SVM is the best fit for our work.

3.2.4.1. Support Vector Machine (SVM)

A support vector machine is a machine learning technique that is well-founded in statistical learning theory. Statistical learning theory is not only a tool for the theoretical analysis but also a tool for creating practical algorithms for pattern recognition. As an application of the theoretical breakthrough, SVMs have high generalization ability and are capable of learning in high dimensional spaces with a small number of training examples. (Seema Asht, 2012) It accomplishes this by minimizing a bound on the empirical error and the complexity of the classifier, at the same time. SVM is a new type of hyper plane classifier, developed based on the statistical learning theory that was introduced in 1995 by Vapnik et al (L. Devroye, 1996), intending to maximize a geometric margin of the hyper plane, which is related to the error bound of generalization. Support Vector Machine (SVM) is a relatively new learning method used for binary classification. The basic idea is to find a hyper plane that separates the d-dimensional data perfectly into its two classes. Several published studies compare the paradigm of neural networks against the support vector machines. The main difference between the two paradigms lies in how the decision boundaries between classes are defined.

While the neural network algorithms seek to minimize the error between the desired output and the one generated by the network, the training of an SVM seeks to maximize the margins between the borders of both classes. SVM approach has some advantages compared to other classifiers. SVM is robust, accurate, and very effective even in cases where the number of training samples is small. SVM technique also shows a greater ability to generalize and a greater likelihood of generating good classifiers.

A support vector machine is a computational algorithm that constructs a hyper plane or a set of hyper planes in a high or infinite-dimensional space. SVMs can be used for classification, regression, or other tasks. Intuitively, a separation between two linearly separable classes is achieved by any hyper plane that provides no misclassification on all data points of any of the considered classes, for example, all points belonging to class A are labeled as +1 and all points belonging to class B are labeled as -1. SVM is an approach where the objective is to find the best separation hyper plane, that is, the hyper plane that provides the highest margin distance between the nearest points of the two classes (called functional margin). This approach, in general, guarantees that the larger the margin is the lower is the generalization error of the classifier. If such hyper plane exists, it is clear that it provides the best separation border between the two classes, it is known as the maximum-margin hyper plane, and such a linear classifier is known as the maximum margin classifier. There are simple and advance SVM concepts.

Simple SVM

Linear SVM is the newest extremely fast machine learning (data mining) algorithm for solving multiclass classification problems from ultra-large data sets that implements an original proprietary version of a cutting plane algorithm for designing a linear support vector machine. Linear SVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time which scales linearly with the size of the training data set. Our comparisons with other known SVM models clearly show its superior performance when high accuracy is required. We would highly appreciate it if you may share Linear SVM performance on your

data sets with us (Kecma, 2009). In the case of linearly separable data in two dimensions, we use Simple SVM.

Non-Linear SVM

In the case of non-linearly separable data, the simple SVM algorithm cannot be used.

The most important parameters used to classify nonlinear SVM in python are

Kernel: the kernel type to be used. The most common kernels are:-

- ✓ RBF(this is the default value),
- ✓ Sigmoid.
- ✓ Linear
- ✓ polynomial

C: this is the regularization parameter described in the Tuning Parameters section.

Gamma: this was also described in the Tuning Parameters section.

Degree: it is used only if the chosen kernel is poly and sets the degree of the polinom

Probability: this is a Boolean parameter and if it's true, then the model will return for each prediction, the vector of probabilities of belonging to each class of the response variable. So basically it will give the confidences for each prediction.

3.3. Tool selection

Finding and selecting the most appropriate machine learning tool with better capability is one of the challenging tasks. For this study, the selection of an appropriate machine-learning tool was done by first listing certain criteria. Among the criteria used for the selection, the most important ones were:

- The speed and quality of the tool (performance).
- Tasks that the tool is intended for (clustering, classification).
- The algorithm supported by the tool.
- The number of records the tool can handle.

- User-friendliness.
- Availability of the tool on the internet.

The numbers of machine learning algorithms among these we will be select the best for our research.

3.3.1. Python 3

Python is a general-purpose and high-level programming language. We can use Python for developing desktop GUI applications, websites, and web applications. Reason for using python other than other programming languages, it is Readable and Maintainable Code, Multiple Programming Paradigms, Compatible with Major Platforms and Systems, Robust Standard Library, Simplify Complex Software Development.

3.4.The architecture of taxpayer risk prediction system

The proposed architecture of risk level prediction system is shown in figure 3.1.

Given the raw taxpayer data, they are first preprocessed for removing low valued attributes, empty data sets unnecessary values. After the document is preprocessed, the next step is classifying the data into training and testing data set.

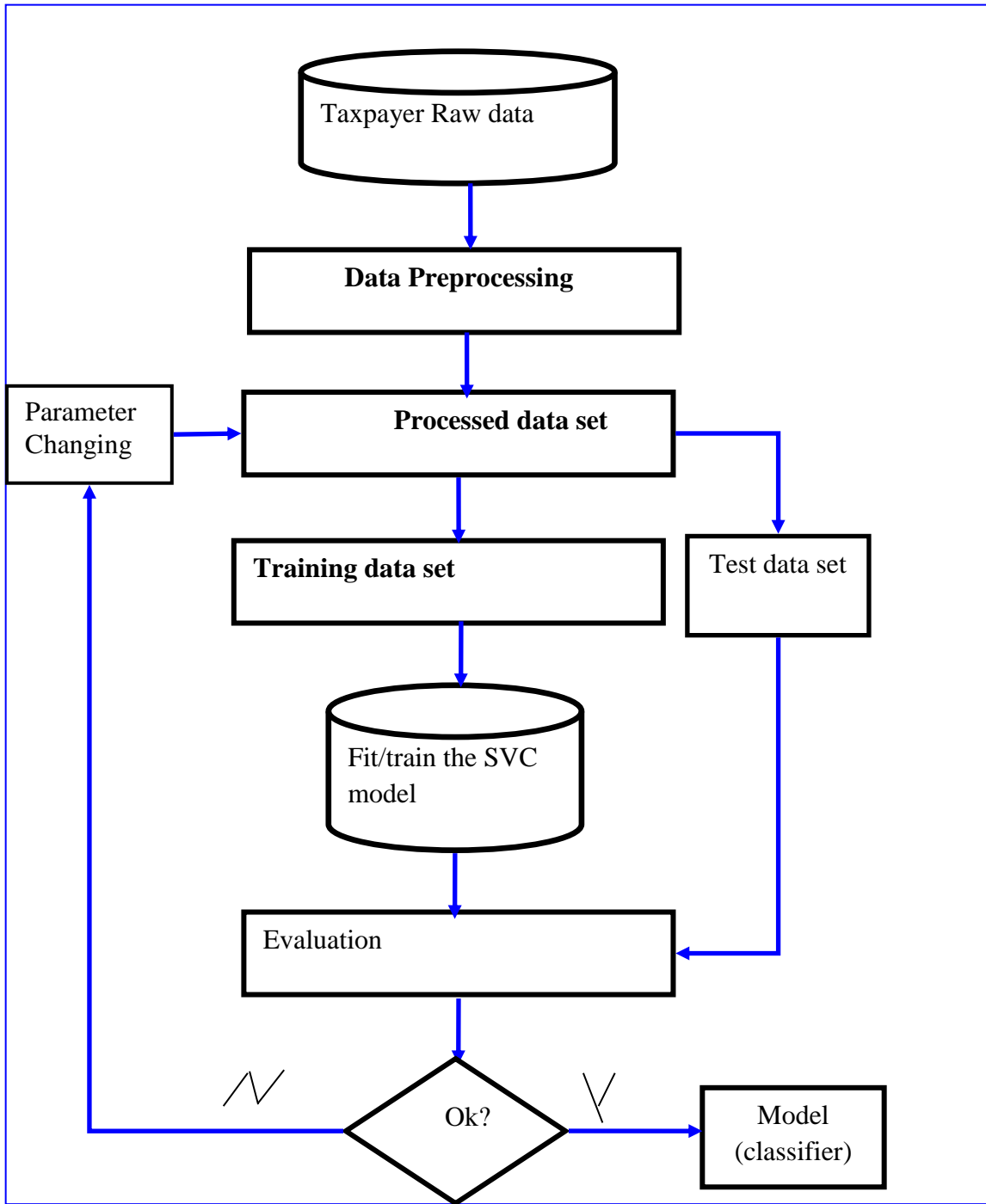


Figure 3:1 Architecture of taxpayer risk level prediction system

3.4.1 Row data

Raw data are primary data collected from a source. And also raw data refers to data that has not yet been processed. In our case row data are taken from MOR and are put in excel format.

3.4.2. Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format we use the following steps to preprocess the data.

- Import the libraries
- Import the data-set
- Check out the missing values
- Splitting the data-set into Training and Test Set
- Feature Scaling

3.4.3. Processed data

Processed data is the raw data, dealt with so it can be used more easily, or displayed to show a result or feature. And it is data ready for use.

Training set: -Training set is the one on which we train and fit our model basically to fit the parameters Training data's output is available to model.

Test set: -test data is used only to assess the performance of the model. Data is the unseen data for which predictions have to be made.

3.4.4. Training

You need to train the algorithm to create the model .Depending on the type of data and algorithm, the training process may be supervised, unsupervised, or reinforcement learning.

In our research, we use unsupervised and supervised training process. The total data split into the train-test split method the model will be train with the training data set and the performance of the trained model evaluated by using the testing data set. The goal of training is to create an accurate model that answers our questions correctly.

3.4.5. Model building

Building the model: in this step, we select and implement the appropriate machine learning tasks (association rules, serial pattern discovery, classification, regression, clustering, etc.), and the data processing algorithm(s) to create the model.

3.4.6. Evaluation

After the final model is created the accuracy of the model is evaluated. As the algorithm creates multiple models, either a human or the algorithm will need to evaluate and score the models based on which model produces the most accurate predictions. It is important to remember that after the model is operational, it will be exposed to unknown data. As a result, make sure the model is generalized and not over fit your training data.

3.4.7. Prediction

After deployment, start making predictions based on new, incoming data. Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome.

CHAPTER FOUR

EXPERIMENT AND IMPLEMENTATION

4.1. Introduction

In this study, an attempt is made to develop a taxpayer fraudulent risk level prediction model that predicts their risk level. Python is used to develop the prediction model, and excel is used for dataset preparation.

4.2. Data Set Preparation

4.2.1. Data cleaning

In our research we use MS-Excel 2010 is used to identify and delete the null value. Initially, the total data set contained 8200 records out of these 60 records have incomplete feature values . We have deleted these values because it may lead to the wrong prediction.

4.2.2. Data transformation

The data set values in our research are from 0 to 18 so to improve our classification result data is normalized by using python3 data analysis tools in the range of 0 and 1.

```

: #Dataset Preprocessing and prepare the data for machine learning
#Rescale data (between 0 and 1)
import pandas
import scipy
import numpy
from sklearn.preprocessing import MinMaxScaler
array = ex.values
x = array[:,0:18]
y = array[:, -1]
#ex = scaler.fit_transform(ex)
scaler = MinMaxScaler(feature_range=(0,1))
rescaldx = scaler.fit_transform(x)
#rescaldy = scaler.fit_transform(y)
numpy.set_printoptions(precision=3)
print(rescaldx)

[[0.    0.294 0.    ... 0.    0.    0.    ]
 [0.    0.647 0.286 ... 0.    0.    0.    ]
 [0.    1.    0.571 ... 0.    0.    0.    ]
 ...
 [0.    0.647 0.286 ... 0.    0.    0.    ]
 [0.    1.    0.571 ... 0.    0.    0.    ]
 [0.    0.294 0.143 ... 0.    0.    0.    ]]

```

Code 4:1 implementation for Data transformation

4.2.3. Attribute selection

The data has 21 attributes in total (20 dependent and one independent feature).

We have used which have high weight and have a significant impact on prediction. ChiSquaredAttributeEval(compute the chi-squared statics of each attribute concerning the class) for our research. Out of 21 attributes, 19attributes were found to be significant from this pool and used for clustering and 19 attributes (18independent and 1 dependent attributes) are selected for classification.

4.2.4. Data formatting

At this step, the researcher changes the data into a format, which was suitable for the machine learning algorithms. Since the data is in MS-Excel format; however, the selected machine-learning tool does not accept the data in Excel format. Therefore, the researchers first tried to convert the data in comma-delimited (CSV).

4.3. Experimental Setup

In this chapter real machine learning experiments conducted on the final dataset described above in detail. In this regard, preprocessed data containing 8140 records. The research problem is to predict taxpayer fraud levels for the case of the ministry of revenue. Initially, the data set target value is omitted for clustering purposes. The clustering methods are used for farther preprocessing the data. Machine learning algorithms for clustering are k-means clustering and SVM are used for classification, which is explained in chapter three. Here is the model building is started.

The experiment is done in three different scenarios. First with the full dataset classified and prediction is made second full dataset is clustered in different clusters. The last classification is done by using SVM in each cluster.

4.3.1. k-means clustering

Steps to implement k- means clustering

1. Specify the number of clusters (K) to be created (by the analyst)
2. Select randomly k objects from the data set as the initial cluster centers or means
3. Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid
4. For each of the k clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster. The centroid of a Kth cluster is a vector of length p containing the means of all variables for the observations in the kth cluster; p is the number of variables.
5. Iteratively minimize the total within the sum of the square that is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached determining the Optimal Number of Clusters.

There are different methods of Determining the Optimal Number of Clusters in python in our work we use an elbow method. This method does by looking at the total WSS as a function of the number of clusters. The optimal number of clusters can be defined as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters. Python 3 is used to implement the clustering process. Here it is the graphical representation of the elbow method.

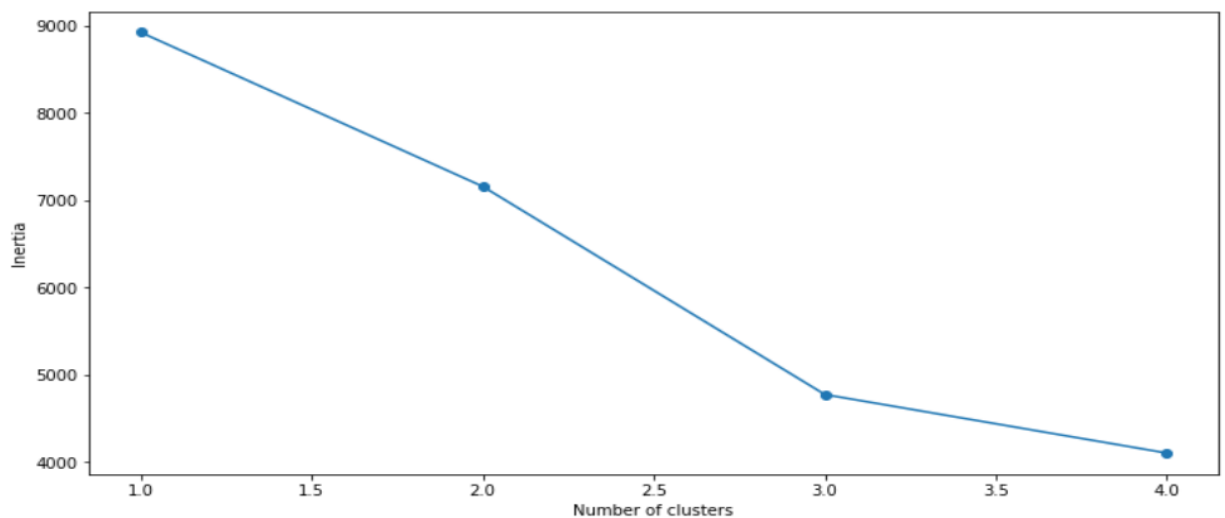


Figure 4:1 optimal cluster numbers

Based on the graph the location of a bend (knee) indicate at number 3 so we determine the optimal number of cluster is 3. The experiments are done based on the following scenarios: The first full data set containing 8140 records is clustered into three sub-clusters by omitting the target value.

Experiment -1-

Here we will classify the whole data set before clustering using the SVM algorithm. The Confusion matrix by using test data set(1628) =83(high risk)+406(medium risk)+1139(low risk) are selected for testing out of this 66+396+1131 are correctly classified which is 97.8% and 17+10+8 are incorrectly classified which is 2.2%.

```
#train the model with scaled dataset
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
svc_model = SVC(kernel='sigmoid',random_state=90)
svc_model.fit(rescaledx_train, y_train)
y_predict = svc_model.predict(rescaledx_test)
accuracy_score(y_test, y_predict)
```

```
C:\Users\good\Miniconda3\envs\tensorflow\lib\site-packages.py:193: FutureWarning: The default value of gamma will
to' to 'scale' in version 0.22 to account better for uns
et gamma explicitly to 'auto' or 'scale' to avoid this w
    "avoid this warning.", FutureWarning)
```

```
0.9785012285012284
```

```
pd.crosstab(y_test, y_predict)
```

col_0	1	2	3
row_0			
1	1131	8	0
2	10	396	0
3	0	17	66

Code 4:2 Implementation confusion matrix with SVM

Table 4:1 confusion matrix of classification with SVM

Actual	Predicted			total
	High risk	Medium risk	Low risk	
High risk	66	17	0	83
Medium risk	10	396	0	406
Low risk	0	8	1131	1139

The table below shows the classification results obtained from the implementation of the above code.

Table 4:2 classification result for experiment -1-

	Result	Percentage
Correctly classified	1593	97.8%
Incorrectly classification	35	2.2%

Experiment -2-

By considering the threshold value cluster 1 considered as a low risk taxpayer, cluster 2 = medium risk taxpayer and cluster 3 contains the property of the high-risk taxpayer.

Below the table shows the actual and the prediction values of clustering the whole data set.

From the total of 5634 in cluster 1 2937 are correctly clustered and the remaining 2654 and 43 are incorrectly clustered to cluster 2 and cluster 3 respectively.

From the total of 2083 in cluster 2 1103 are correctly clustered and the remaining 102 and 878 are incorrectly clustered to cluster 1 and cluster 3 respectively.

From the total of 423 in cluster 3 407 are correctly clustered and the remaining 3 and 13 are incorrectly clustered to cluster 1 and cluster 2 respectively.

Table 4:3 Confusion matrix for experiment -2-

	Prediction			total
Actual	Cluster 1	Cluster 2	Cluster 3	
Cluster 1	2937	2654	43	5634
Cluster 2	102	1103	878	2083
Cluster 3	3	13	407	423
Total				8140

Table 4:4 shows the percentage of correctly clustered and incorrectly clustered data sets by looking at table 4:3.

Table 4:4 clustering result for experiment -2-

	Result	Percentage
Correctly clustered	4447	54.63%
Incorrectly clustered	3693	45.37%

4.3.2. SVM classification model building

Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems. There are different parameters to train the SVM model in python (random state, kernel,max_iter---) among these we use sigmoid kernel function because we got best result by using this function and the other parameters are best in the default value. For classification, there are many train test split options like 66/34, 70/30, 80/20, 90/10 among these we use percentage split of 80/20. Because we got better accuracy at 80/20/ splits. From the total data set

of 8140, 1628 are selected as test data set. The implementation is done in Python's Scikit-Learn library.

Experiment -3-

In this experiment, we use each cluster obtained from k-means clustering as input for classification. We use a percentage split of 80/20 for each cluster. The total data set in cluster 1 is 5634. The total test data set is 1127.

Classification for cluster -1-

Confusion matrix by using test data set(1127) =5(high risk)+542(medium risk)+580(low risk) are selected for testing out of this 4+534+580 are correctly classified which is 99.2% and 1+8+0 are incorrectly classified. The number of test data sets in each level is determined by using cross-tabulation python code.

```

#train the model with scaled dataset
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
svc_model = SVC(kernel='sigmoid',random_state=90)
svc_model.fit(rescaledc11x_train, c11y_train)
c11y_predict = svc_model.predict(rescaledc11x_test)
accuracy_score(c11y_test, c11y_predict)

```

C:\Users\good\Miniconda3\envs\tensorflow\lib\site-packages\sklearn\svm_base.py:193: FutureWarning: The default value of gamma will change from 'auto' to 'scale' in version 0.22 to account better for feature scales. Set gamma explicitly to 'auto' or 'scale' to avoid this warning.

"avoid this warning.", FutureWarning)

Out[123]:

0.9920141969831411

In [124]:

```
pd.crosstab(c11y_test,c11y_predict)
```

Out[124]:

col_0	0	1	2
row_0			
0	580	0	0
1	8	534	0
2	0	1	4

Code 4:3 Implementation confusion matrix with SVM

Table 4:5 confusion matrix of classification for cluster -1- with SVM

Actual	Predicted			Total
	High risk	Medium risk	Low risk	
High risk	4	1	0	5
Medium risk	8	534	0	542
Low risk	0	0	580	580

Table 4:6 shows that the classification results obtained from table 4.5.

Table 4:6. Classification result for cluster-1-

	Result	Percentage
Correctly classified	1118	99%
Incorrectly classification	9	1%

Classification for cluster -2-

In this experiment also, we use each cluster obtained from k-means clustering as input for classification. We use a percentage split of 80/20 for each cluster. The total data set in cluster 2 is 2083. The total test data set is 417.

Confusion matrix by using test data set(417) =19(high risk)+222(medium risk)+176(low risk) are selected for testing out of this 19+219+176 are correctly classified which is 99.2% and 0+3+0 are incorrectly classified.

```
#train the model with scaled dataset
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
svc_model = SVC(kernel='sigmoid',random_state=300)
svc_model.fit(rescaldc22x_train, c22y_train)
c22y_predict = svc_model.predict(rescaldc22x_test)
accuracy_score(c22y_test,c22y_predict)
```

C:\Users\good\Miniconda3\envs\tensorflow\lib\site-packages\sklearn\svm_base.py:193: FutureWarning: The default value of gamma will change from 'auto' to 'scale' in version 0.22 to account better for different scales of input features. Set gamma explicitly to 'auto' or 'scale' to avoid this warning.

"avoid this warning.", FutureWarning)

Out[7]:

0.9928057553956835

```
pd.crosstab(c22y_test, c22y_predict)
```

Out [8]:

col_0	0	1	2
row_0			
0	19	0	0
1	0	219	3
2	0	0	176

Code 4:4 Implementation of confusion matrix with SVM

Table 4:7. Confusion matrix of classification for cluster 2 with SVM

Actual	Predicted			total
	High risk	Medium risk	Low risk	
High risk	19	0	0	19
Medium risk	0	219	3	222
Low risk	0	0	176	176

Table 4:8 classification result for cluster -2-

	Result	Percentage
Correctly classified	414	99.2%
Incorrectly classification	3	0.8%

Classification for cluster 3

Confusion matrix by using test data set(85) =84(high risk)+0(medium risk)+1(low risk) are selected for testing out of this 84+0+0 are correctly classified which is 98.8% and 1+0+0 are incorrectly classified.

Table 4:9 confusion matrix of classification for cluster 3 with SVM

Actual	Predicted			total
	High risk	Medium risk	Low risk	
High risk	84	0	0	84
Medium risk	0	0	0	0
Low risk	1	0	0	1

Table 4:10. Classification result for cluster 3

	Result	Percentage
Correctly classified	84	98.8%
Incorrectly classification	1	1.2%

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

The purpose of this study was to design a fraudulent risk level prediction model for the taxpayer. The research is done by using hybrid machine learning algorithms. The first classification is done by using SVM algorithms and then the result is evaluated using the confusion matrix. Second, the data is clustered into three clusters and each cluster is further classified to get a better accuracy rate. In this study, SVM is used as a classification technique and k-means clustering is used for clustering purposes.

In this research, the experiments have been conducted by following the KDD Process Model. The KDD as a process is defined as interactive and iterative, involving nine-steps with many decisions made by the user.

The data used in this research has been gathered from the ministry of revenue Bahir Dar branch. In this study, we used three-phase, First with the full dataset classified and prediction is made second full data set is clustered in different clusters. The last classification is done by using SVM in each cluster.

In general, in this study, we can conclude that by using a hybrid approach it is possible to increase the accuracy of prediction.

For this study, we used 8140 records to train and test the model. K-means clustering and SVM are used for clustering and classification techniques respectively. By using SVM, before clustering has an accuracy of 97.8% and using k-means clustering 54.6% of the records correctly clustered. In the second phase, we got an accuracy of 99.9% by using SVM after clustering.

5.2. Recommendation

Based on the findings of the current study we recommend the following as future research direction.

Involvements of multiple algorithms: In this study, we used k-means clustering for clustering purposes and SVM for classification. However by including other clustering algorithms like HierarchicalClusterer, DBScan, FarthestFirst, and FilteredClusterer, and classification algorithms like, decision tree, artificial neural network can improve prediction performance.

Increasing data set: the data set used in this study is taken only from the ministry of revenue Bahir Dar branch so it may affect the prediction. We recommend to include other branches of the ministry of revenue.

Increasing model developing phases: In this study, we used a two-level hybrid approach to develop the model. However, by increasing the numbers of levels it is possible to get better results.

References

- A.LourduCaroline1, D. D. (2018). Comparative study of Classification algorithms for Data Mining.
- Aemiro, T. (2014). Determinants of Tax Compliance Behavior in Ethiopia The Case of Bahir Dar City Taxpayers.
- al., H. e. (2010). clustering techniques in data mining.
- Allingham, M. a. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics*, 1(3-4), 323-38.
- Andreoni, J. E. (1998). Tax compliance. *Journal of Economic Literature*, 36, 818-60.
- beris.G. (2017). Mining e-filing Data for Predicting Fraud: The Case of Ethiopian Revenue and Customs Authority.
- Biaz, B. (2011). KNOWLEDGE DISCOVERY FOR EFFECTIVE CUSTOMER SEGMENTATION: THE CASE OF ETHIOPIAN REVENUE AND CUSTOMS AUTHORITY.
- Burges. (1998). "A tutorial on support vector machines for pattern recognition", *Knowledge Discovery and Data Mining*, Volume 2, Issue 2, pp 121-167, June 1998,.
- chatterjee, i. (2019). Comparative Study of Clustering Algorithms.
- Chen, S. (2010). Tax Evasion and Fraud Detection: A Theoretical Evaluation of Taiwan's Business Tax Policy for Internet Auctions.
- Chepkwony. (2017). Auditees case-selection model for evaluating taxpayer corporate tax compliance in Kenya.
- Chepkwony, C. C. (2017). Auditees case-selection model for evaluating taxpayer corporate tax compliance in Kenya.

- Clotfelter. (1983). Tax evasion and tax rates: An analysis of individual returns. *The Review of Economics and Statistics*, LXV(3), 363-73.
- Dabbura, I. (2018). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.
- Dabbura, I. (2018). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.
- ESHETU, Y. (2018). DETERMINANTS OF TAX COMPLIANCE BEHAVIOR IN ETHIOPIA: A CASE STUDY OF WEST SHOWA ZONE.
- Fayyad, U. P.-S. (1996). P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, vol. 17, No. 3, 37.
- Friedrich Schneider, E. K. (2001). Tax avoidance, tax evasion, and tax flight: Do legal differences matter?
- George Tzani, I. K. (2014). *Modern Applications of Machine Learning*.
- Gharagyozyan, H. (2019). *A Practical Application of Machine Learning in Medicine Learning in Medicine*.
- Gupta, M. (2014). *Audit Selection Strategy for Improving Tax Compliance –Application of Data Mining Techniques*.
- Han and Kamber. (2006). *clustering techniques in the field of marketing*
- Huang. (2008). K-H.
- Ivanovic, R. 2. (2015). Nearest neighbors in high-dimensional data: The emergence and influence of hubs.
- Ivanovic,Radovanovic. (2015). DATA MINING CLASSIFICATION TECHNIQUES ON THE ANALYSIS OF STUDENT’S PERFORMANCE.

- Jackson, B. a. (1986). Tax compliance research: Findings, Problems, and Prospects, *Journal of Accounting Literature* 5, 125-135.
- James, S. a. (2004). Tax Compliance, self-assessment, and tax administration. *Journal of Finance and Management in Public Services*.
- Kamber, M. (2003). data mining concepts and techniques.
- Kamber, M. (2003). data mining concepts and techniques.
- Kebede, M. (2018). Determinants of Taxpayers' Voluntary Compliance with Taxation: The Case of Wolaita Sodo and Tercha Town in Dawuro Zone.
- Kecma, T.-M. H. (2009). Linear Support Vector Machine.
- Kirchler, E. (2007). *The Economic Psychology of Tax Behaviour*. Cambridge: Cambridge University Press.
- Kothari. (2009). *Research Methodology: Methods and Techniques*. New Delhi: New Age International Publishers.
- Kothari. (2009). *Research Methodology: Methods and Techniques*. New Delhi: New Age International Publishers.
- L. Devroye, L. G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer Verlag, Berlin, 1996.
- Leon Solon da Silv, H. d. (n.d.). Bayesian Networks on Income Tax Audit Selection - A Case Study of Brazilian Tax Administration.
- Low, J. a. (2017). Exploring Key Determinants of Tax Compliance Decision Among Individual Taxpayers in Sri Lanka.
- mama, D. (2013). APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT FRAUD PROTECTION.

- mamo, d. (2013). APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT FRAUD PROTECTION: THE CASE OF ETHIOPIAN REVENUE AND CUSTOM AUTHORITY.
- Manaye, M. K. (2018). Determinants of Taxpayers' Voluntary Compliance with Taxation.
- Mariscal, O. M. (2009). A Data Mining & Knowledge Discovery Process Model.
- Martikainen, J. (2012). Data Mining in Tax Administration - Using Analytics to Enhance Tax Compliance.
- Melata. (2016). FACTORS AFFECTING TAX AUDIT EFFECTIVENESS EVIDENCE FROM LARGE TAXPAYERS OFFICE OF ETHIOPIAN REVENUE AND CUSTOMS AUTHORITY.
- Melendez, R. (2019). Credit Risk Analysis Applying Machine Learning Classification Models Classification Models.
- Mihret, G. (2011). Tax Audit Practice in Ethiopia: The Case of the Federal government.
- Milliron. (1986). Tax compliance research: Findings, Problems, and Prospects, Journal of Accounting Literature 5, 125-135.
- Milliron, J. a. (1986). Tax compliance. Journal of Economics.
- Mitchell. (1997). annual review of information science and technology.
- N.Radha, R. L. (2011). Machine Learning Approach for Taxation Analysis using Classification Techniques.
- Nasteski, V. (2017). An overview of the supervised machine learning methods.
- Nasteski, V. (2017). An overview of the supervised machine learning methods.
- Neelamegam, R. (2013). DATA MINING CLASSIFICATION TECHNIQUES ON THE ANALYSIS OF STUDENT'S PERFORMANCE.

- Nerijus Remeikis, I. S. (2007). Hybrid Machine Learning Approach for Text Categorization.
- Nurlis. (2015). The Effect of Taxpayer Awareness, Knowledge, Tax Penalties, and Tax Authorities Services on the Tax Compliance: Survey on the Individual Taxpayer at Jabodetabek & Bandung, . .
- Rohit Choudhry, a. K. (2008). A Hybrid Machine Learning System for Stock Market Forecasting.
- Sankar. (2017). A developed case-based reasoning system for machine tool selection.
- Seema Asht, R. D. (2012). Pattern Recognition Techniques: A Review”, International Journal of Computer Science and Telecommunications, Volume 3, Issue 8, August 2012.
- SENBETA, T. B. (2018). TAX AUDIT IN OROMIA REVENUE AUTHORITY:PRACTICES AND CHALLENGES.
- Simon. (1983). knowledge management and data mining in medicine.
- Singh. (2003). The behavioral intention of tax non -compliance among sole proprietors.
- Somasundram, N. (2003). Tax evasion and tax investigation - a study on tax compliance management. Chartered Secretary Malaysia, July, 20-24.
- Tajtakova, M. (2009). Machine learning methods for the market segmentation of the audiences of the performing arts.
- Tomar, A. (2013). A survey on Data Mining approaches for Healthcare.
- Yarbrough, S. a. (1978). Income tax evasion: A theoretical analysis.

APPENDICES

Appendix 1: Partial View of the Initial Collected Sample Data

start_date	sales	industry_	no_of_br	penalty_	is last_audit_	refund	loss declar	profit_mar
2	6	4	0	4	12	0	18	15
2	12	8	0	4	4	0	18	0
2	18	12	0	4	12	0	18	15
2	6	8	0	4	12	0	0	15
2	4	4	0	4	12	0	0	0
2	6	8	0	4	12	0	0	15
2	6	8	0	4	8	0	0	10
4	6	12	0	4	12	0	12	15
2	4	4	0	4	12	0	0	5
2	6	8	0	4	12	0	0	15
4	6	6	0	4	12	0	0	10
2	4	4	0	4	4	0	0	15
2	6	4	0	4	12	0	0	0
2	4	8	0	4	8	0	0	10
2	4	6	0	4	8	0	0	5
4	6	6	0	4	12	0	0	10
2	6	4	0	4	8	0	0	0
4	6	4	0	4	12	0	0	0
4	6	4	0	4	8	0	0	0
2	6	4	0	4	12	0	0	10
4	6	4	0	4	12	0	0	15
4	4	4	0	4	8	0	0	5
2	6	8	0	4	4	0	0	0
2	6	4	0	4	8	0	0	15
2	6	4	0	4	12	0	0	0
2	6	8	0	4	12	0	0	0
2	6	4	0	4	12	0	0	5
2	4	4	0	4	8	0	0	15
2	12	8	0	4	12	0	0	0
2	6	4	0	4	12	0	0	5
2	6	8	0	4	12	0	0	0
2	6	6	0	4	12	0	0	0
2	6	4	0	4	8	0	0	10
4	6	4	0	4	12	0	0	10
2	4	4	0	4	12	0	0	0
2	4	4	0	4	12	0	0	0
2	6	4	0	4	18	0	0	0
2	6	4	0	4	12	0	0	0
2	6	4	0	4	12	0	0	10
2	6	4	0	4	12	0	0	0
2	4	4	0	4	12	0	0	0
2	6	4	0	4	12	0	0	0
2	6	4	0	4	12	0	0	5
2	4	4	0	4	12	0	0	0
2	6	4	0	4	12	0	0	0
2	18	4	0	4	8	0	0	15

2	6	4	0	4	12	0	0	0
2	18	4	0	4	12	0	0	0
2	6	4	0	4	12	0	12	15
4	6	6	0	4	8	0	0	0
2	6	8	0	4	4	0	0	0
2	18	8	0	4	12	0	18	15
2	6	4	0	4	12	0	0	0
4	6	4	0	4	4	0	0	0
2	18	8	0	4	12	0	0	0
2	18	4	0	4	8	0	0	0
2	6	4	0	4	12	0	0	0
2	6	4	0	4	4	0	0	0
2	6	4	0	4	4	0	12	0
2	18	8	0	4	4	0	0	0
2	6	4	0	4	12	0	12	0
2	6	8	0	4	12	0	0	0
2	6	4	0	4	4	0	0	10
2	6	8	0	4	4	0	0	10
2	6	4	0	4	12	0	0	5
4	6	8	0	4	12	0	0	5
2	6	4	0	4	4	0	0	5
2	6	4	0	4	12	0	0	0
2	6	8	0	4	4	0	0	0
2	6	4	0	4	12	0	0	0
2	18	12	0	4	4	0	0	0
2	18	4	0	4	12	0	12	15
2	6	12	0	4	12	0	0	0
2	6	4	0	4	4	0	0	0
2	6	12	0	4	12	0	0	15
2	6	8	0	4	8	0	18	0
2	6	4	0	4	12	0	0	0
2	6	8	0	4	12	0	18	0
2	6	4	0	4	12	0	0	5
2	12	4	0	4	4	0	0	0
2	6	8	0	4	12	0	12	15
2	6	4	0	4	12	0	18	0
2	12	4	0	4	4	0	0	15
2	6	4	0	4	12	0	0	0
2	6	8	0	4	12	0	0	0
2	12	4	0	4	12	0	0	0
2	6	4	0	4	12	0	0	0
2	18	8	0	4	12	0	0	15
2	6	8	0	4	8	0	18	0
2	6	4	0	4	12	0	0	0
2	12	4	0	4	4	0	0	15
2	6	4	0	4	12	0	0	0
2	6	4	0	4	4	0	0	15
2	6	4	0	4	12	0	0	0
2	6	4	0	4	4	0	12	0

Appendix II: Sample code to load data

```
import pandas as pd
import numpy as np
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
ex = pd.read_csv("D:\pls.csv")
X = ex.iloc[:,0:18] #independent columns
y = ex.iloc[:, -1] #target column
bestfeatures = SelectKBest(score_func=chi2, k=18)
fit = bestfeatures.fit(X,y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs', 'Score']
print(featureScores.nlargest(18, 'Score'))
```

Appendix III: train the data sets

```
#train the model with scaled dataset
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
svc_model = SVC(kernel='sigmoid',random_state=90)
svc_model.fit(rescaledcx_train, cly_train)
cly_predict = svc_model.predict(rescaledcx_test)
accuracy_score(cly_test,cly_predict)
```

C:\Users\good\Miniconda3\envs\tensorflow\lib\site-packages\sklearn\svm\base.py:193: FutureWarning: The default value of gamma will change from 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoid this warning.

"avoid this warning.", FutureWarning)

0.9901129943502824

Appendix IV:Rescaling the data set

```
[29]: #Dataset Preprocessing and prepare the data for machine learning
#Rescale data (between 0 and 1)
import pandas
import scipy
import numpy
from sklearn.preprocessing import MinMaxScaler
array = c2.values
c2x = array[:,0:18]
c2y = array[:, -1]
scaler = MinMaxScaler(feature_range=(0,1))
rescaldc2x = scaler.fit_transform(c2x)
#rescaldy = scaler.fit_transform(y)
numpy.set_printoptions(precision=3)
print(rescaldc2x)
```

```
[[0.    0.5   0.286 ... 0.    0.    0.    ]
 [0.    1.    0.571 ... 0.    0.    0.    ]
 [1.    0.    0.143 ... 0.    0.    0.    ]
 ...
 [0.    0.    0.    ... 0.    0.    0.    ]
 [0.    0.5   0.286 ... 0.    0.    0.    ]
 [0.    1.    0.571 ... 0.    0.    0.    ]]
```

Appendix V: Train test split the dataset

```
In [7]: #train test split with un scaled data
from sklearn.model_selection import train_test_split
import pandas as pd
#y=ex.z_score
xx=ex.drop('risk_level',axis=1)
#rescaldx = ex.iloc[:,0:18].values
#y = ex.iloc[:, -1].values
xx_train, xx_test, y_train, y_test = train_test_split(xx,y, train_size = 0.8)
y_train.shape, y_test.shape
xx_train.shape, xx_test.shape
```

```
Out[7]: ((6512, 18), (1628, 18))
```

Appendix VI:Confusion matrix for cluster1

```
In [124]: pd.crosstab(c11y_test,c11y_predict)
```

```
Out[124]:
```

col_0	0	1	2
row_0	<hr/>		
0	580	0	0
1	8	534	0
2	0	1	4

Appendix VII:Confusion matrix for cluster 2

```
In [136]: pd.crosstab(c22y_test,c22y_predict)
```

```
Out[136]:
```

col_0	0	1	2
row_0	<hr/>		
0	11	0	0
1	0	166	1
2	1	1	73