2017

Auditees case-selection model for evaluating taxpayer corporate tax compliance in Kenya

Caroline Chepkurui Chepkwony
Fuculty of Information Technology (FIT)
Strathmore University

Auditees Case-Selection Model for Evaluating Taxpayer Corporate Tax Compliance in Kenya

Chepkwony Caroline Chepkurui

Master of Science in Information Technology [MSc.IT]

2017

Auditees Case-Selection Model for Evaluating Taxpayer Corporate Tax Compliance in Kenya

Chepkwony Caroline Chepkurui

089597

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Information Technology at Strathmore University

Faculty of Information Technology

Strathmore University

Nairobi, Kenya

June, 2017

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Chepkwony, Caroline Chepkurui

....

8th June 2017

Approval

The thesis of Chepkwony, Caroline Chepkurui was reviewed and approved by the following:

Dr. Vincent Omwenga (PhD)
Senior Lecturer, Faculty of Information Technology,
Strathmore University

Dr. Joseph Orero (PhD)
Dean, Faculty of Information Technology,
Strathmore University

Professor Ruth Kiraka
Dean, School of Graduate Studies,
Strathmore University

Abstract

Tax compliance rate in Kenya is estimated to be approximately below 65%. It is important for the government to place measures that ensure improved tax compliance rate comparable with benchmark countries like Sweden, whose tax compliance rate stand at 93%. One measure implemented in Kenya Revenue Authority has been to conduct scrutiny assessments on the taxpayer fraternity. However, success in scrutiny assessments in addressing payment and reporting compliance is largely dependent on the cases selected for audit. A major challenge has been in the possibility of selecting of an honest taxpayer and failure to take up the potential under-reporter, scenarios which are both costly to the tax administration. Whereas the honest taxpayer will feel unfairly selected for scrutiny, under-reporters escape the purview of the authority. This study presents a data mining based approach aimed at addressing the case-selection challenge. A classification model built using historical taxpayer audit data and decision tree algorithm was used to predict the compliance status of taxpayers in a case-selection application prototype. Experimental results using limited taxpayer data for the period year 2014/2015 indicate that the model is effective and fit for case-selection with an accuracy rate of 65% and prediction efficiency of 65% in identifying non-compliant taxpayers. Moreover, with more sources of taxpayer information and increased quantity of data, the accuracy and prediction efficiency is expected to improve significantly. It is recommended that Kenya Revenue Authority adopts this approach to improve the traditional case-selection by auditors' for corporate tax as well as other tax obligations such as Individual income tax, VAT, and custom duties administered by the Kenyan government.

Table of Contents

List of Figures

List of Tables

Acknowledgement

I would like to thank my Supervisor Dr. Omwenga for the continuous support he has offered me so far. He has consistently allowed this study to be my work, but directed me in the right direction whenever he thought I was heading to the wrong direction. I would also like to thank my family, friends and classmates for the support they have offered to me in different ways.

Dedication

This dissertation is dedicated to almighty God, family, friends, and colleagues for all the support they have given me so far. I also dedicate it to my Supervisor Dr. Omwenga and Strathmore university fraternity for the guidance accorded to me in developing this dissertation.

Chapter 1: Introduction

1.1 Background to the Study

Tax is an important aspect for governments as it is their main revenue stream and drives many projects. As a result, most revenue authorities in administering tax for their governments are keen in maximizing revenue collection and efficiency in the tax administration. For instance, tax administrations undertake key changes in tax policy and embrace technological solutions with a view of widening the tax base and/or reducing the compliance burden Musau (2015).

Tax compliance broadly refers to the degree in which taxpayers obey with the tax laws in their country. According to James and Alley (2002), tax compliance viewed in terms of tax gap, tax evasion and tax avoidance. Tax gap refers to discrepancy between the amount of revenue collected and that which would be collected if there were 100 percent compliance, tax avoidance to the legal measures to reduce tax liability, and tax evasion to the illegal measures.

Tax compliance is a problem facing many revenue authorities and has thus become a major focus of their operations. Persuading taxpayers to obey the tax laws is not always an easy task. In addition, 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 a number of factors, including their basic willingness to comply with a tax system (Hallsworth et al., 2014).

In the Kenyan context, tax compliance refers to complying with the tax law in four aspects; registration, filing, reporting, and payment of taxes in accordance with taxpayers obligations. Registration compliance refers to the proportion of taxpayers registered with the tax authority. Filing compliance refers to the proportion of registered taxpayers that submit tax returns to the tax authority. Reporting compliance refers to the accuracy of declared taxable income information. Payment compliance refers to the proportion of taxes paid by the deadline (Kenya Revenue Authority, 2015).

According to 2015/2016-2017/2018 KRA's 6th corporate plan, Kenya's tax compliance rate falls below 65% and thus is one of the internal factors affecting the ability of the Kenyan government to raise direct tax revenues to meet its recurrent and development expenditure. Compliance risk areas include; miss-declaration of income/goods, Falsification of customs documents, non-compliance with Electronic Tax Register (ETR) requirements, Non-filers, and Diversion of transit and export cargo into the domestic markets (Kenya Revenue Authority, 2016).

1.1.1 Tax Structure in Kenya

The taxation system in Kenya is administered by Kenya Revenue Authority. The Authority was established in 1995 by an Act of Parliament, Chapter 469 of the Laws of Kenya with the mandate of collecting revenue for the government. As a result, all entities ranging from individuals, partnerships, companies, and corporations are required to register for a personal identification number (PIN) and applicable tax obligations with the authority. In addition, they are expected to remit taxes to the authority in accordance with tax laws. Kenya's taxation system majorly covers income taxes, value-added tax (VAT) and Customs and excise duty each of which are governed by independent legislations which include; Income Tax Act (Cap. 470), Value Added Tax Act (Cap. 476), Customs and Excise Act (Cap. 472) and the East African Community Customs Management Act (EACCMA) respectively.

Tax laws apply across the board. For instance, under the Income Tax Act (Cap 470), Corporate entities are subject to tax on their accounting profit/loss less allowable deductions; partnerships on partnership's earnings at the end of each year of income; and employees on employment income. Also, any individual whose business income is between kshs. 500,000 and kshs. 5 m is subject to turnover tax. Governed still by the income tax Act, advance tax and withholding tax are allowed whereby the former is payable in respect of every commercial vehicle and the latter imposed on certain services and deductible on payment made to service providers. In addition, taxpayers are required to carry out individual self-assessments and file their returns with KRA by June the 30 of the following year.

VAT is governed by the Value Added Tax Act (Cap. 476) .It covers all imports, supplies, manufactured goods and services provided in Kenya. Consequently, any person who supplies taxable goods or services with a value of Kshs. 5 million or more in a one year period is required to register for VAT obligation and remit the same to the authority by the 20th of the following month. However, whereas goods are taxed at the standard rate of 16%, some are exempted from this tax and thus taxed at 0%. These exempted goods are listed under the schedules of the VAT Act which are reviewed annually.

Governed by the Customs and Excise Act (Cap. 472), custom duties are those payable by importers at the point of importation. These duties include; import duty, excise duty, VAT, import declaration fee and railway development levy among others. Import duty is charged at the rate of 25% the custom value of the goods which is based on customs valuation, tariff classification and rules of origin. On the other hand, excise duty is imposed on specified imported or locally manufactured goods and services listed under Customs and Excise Act. Examples of goods liable to excisable goods and services include; duty include wines and spirits, beer, and cigarettes, mobile and wireless phone services, fees on money transfer services and fees charged by financial institutions. Import declaration fee is charged at the rate of 2% the value of the goods while the railway development levies at the rate of 2% the value of goods.

1.1.2 Auditing in Kenya

Auditing generally has two objectives which include; giving assurance that financial statements give a true view of the Company’s state of affairs and the detection and prevention of Frauds and errors (Kumar and Mohan, 2016). Tax audit however, aims at examining an organization's or individual's tax return to verify the correctness of financial information reported. The need of increasing government revenue, failure to meet targets, need to generate revenue to pay debts owed by the government, and the widening budget deficits have increased the need for taxpayer audits.

Paramount to any Audit strategy is the selection of audit subjects (Hsu et al., 2015). Depending on the objective of audit various methodologies have been devised. For instance, Random selection is used in situations where equal chance needs to be given to members of a population. Yet another technique is based on information and procedural non-compliance. Both these techniques have drawbacks, whereas random selection gives equal treatment to both compliant and non-compliant taxpayers, information-based presupposes some symptoms of non-compliance which may be due to other factor such as changing economic times. Recent advances in big data have seen introduction of case selection based on data mining. In conducting tax audits, the taxman uses information from tax returns to strategically pick audit subjects. As a result, the probability of audit is variable based on the behavior of the taxpayer.

Kenya Revenue Authority uses a risk-based strategy based on observations and local knowledge to select cases for audit. Each revenue department carries out its risk assessment to be used for selection audit subject. Domestic taxes audit office classifies risk into two groups; group risk and individual risk. Taxpayers are then assessed for their group and individual risk to obtain their overall risk rating. Those with high risk ratings are possible subjects for audit. For Customs, goods are profiled for risks before they are released. Those regarded as highly risky are subjected to full verification. Importers and clearing agents are also profiled based on their compliance levels with non-compliant once being subjects of possible audit. This approach is insufficient in that it is not automated and risk profiling has to be done regularly with large data requirements. Moreover, each department handles its own risk profiling despite them handling the same taxpayer. As a result, taxpayers may tailor their behavior to being compliant or not depending on the effectiveness or not of the various revenue departments.

1.2 Problem Statement

Many tax administrations focus on tax gap as a means of measuring tax compliance. For instance, Kenya revenue authority uses this concept in measuring payment and reporting compliance. Sirengo (2016) argues that countries with a narrow tax gap are believed to have high rates of tax compliance and thus more revenue from tax collection. Consequently, many tax authorities are increasingly devising ways of dealing with the tax gap in their countries. One strategy has been through the use of audits or scrutiny assessments of taxpayers. According to Hsu et al. (2015) this strategy has the advantage of generating additional revenue as well as enforcing a deterrent measure on the population towards non- compliance.

However, success in scrutiny assessments in addressing payment and reporting compliance is largely dependent on the cases selected for audit. A major challenge has been in the possibility of selecting of an honest taxpayer and failure to take up the potential under-reporter, scenarios which are both costly to the tax administration. Whereas the honest taxpayer will feel unfairly selected for scrutiny, under-reporters escape the purview of the tax administration (Kumar and Rao, 2015). There is also biasness on the part of case selectors on selecting the same cases for audit year in year out based on their knowledge and inherent subjectivity. In addition, using auditors to select cases among millions is like looking for a needle in a haystack, a tiresome process. Da Silva et al. (2016) suggests that advances in analytics and data mining techniques will help solve this problem and many others facing tax authorities.

Kenya Revenue Authority needs a solution that would not only automate its human resource- intensive risk based case selection but also apply predictions of non-compliance in the case selection process to prioritize good audits. The solution would leverage on already available vast amounts of data and data mining algorithm, specifically decision trees to evaluate taxpayer compliance and identify cases for scrutiny assessment. This will impact the authority in terms of objective case selection, increased revenue from tax audits, improved voluntary compliance, and efficient deployment of audit resources. The taxpayer will benefit from a fair case selection system and non-biasness.

1.3 Objective

The purpose of this study is to develop an auditee case-selection model to evaluate taxpayer corporate tax compliance in Kenya.

1.3.1 Specific Objectives

i. To analyze the determinants of corporate tax audit case-selection used by Kenya Revenue Authority
ii. To review existing tax audit case-selection methodologies used by tax authorities
iii. To develop a model for auditee case selection for Kenya Revenue Authority
iv. To validate the model using Taxpayer data in Kenya

1.4 Research Questions

i. What are the determinants for corporate tax audit case selection in Kenya's tax system?
ii. What are the existing methodologies used for tax audit case selection used by tax authorities?
iii. How will the auditee case-selection model for Kenya Revenue Authority be developed?
iv. How will the auditee case-selection model for Kenya Revenue Authority be validated?

1.5 Justification for the Study

This study builds upon E-government innovations in Kenya aiming to revamp the public service. Kenya revenue authority as a public service entity will benefit through an efficient and effective tax case selection system and as a result collect the right taxes and reduce tax evasion rate.

In the recent years, advanced analytics have also become key tools in creating new opportunities and informing important decisions. Tax administrations are using these techniques to observe patterns of non-compliance and hence inform their compliance frameworks.

1.6 Scope

This research is limited to the Kenyan tax system, specifically corporate tax. The study will employ a predictive model based on historic analysis of corporate tax compliance and decision tree as a classification algorithm.

Chapter 2: Literature Review

2.1 Introduction

This chapter begins with the theoretical framework where an overview of corporate tax in Kenya is provided and several theories pertaining to taxpayer compliance behavior are discussed. The chapter then focuses on discussing the determinants of corporate tax audit case- selection, methodologies used for identifying tax audit subjects and related works on audit case selection. Additionally, the chapter will describe data mining concept and focus on discussing decision trees as a classification data mining algorithm. Finally, the chapter will define graphically the conceptual structure of the proposed solution.

2.2 Theoretical Framework

Kenya's corporation tax refers to tax charged on corporations on income derived or accrued from within the country. This tax is imposed on the taxable income, which is the accounting profit/loss adjusted for allowable and disallowable expenses. The deductibility of expenses is premised on the fact that they were wholly and exclusively incurred in the generation of taxable income Companies which operate branches outside the country are required to report all their income in the country and claim a relief of any tax paid in foreign countries if there is a double tax agreement in place between Kenya and the other country (Kenya Revenue Authority, 2016).

The corporation tax rate for resident companies is 30%. Non-resident companies with a permanent establishment (PE) in Kenya are taxed on the income earned or derived from within the country at the rate of 37.5%, with some restrictions on deductible expenses. Non-residents without a PE in the country are taxed under the withholding tax system in cases where the payments made are eligible to withholding tax. Resident companies or non-resident companies with a PE in Kenya are allowed to offset their taxable losses against their taxable income in the year in which they occur and in the next four succeeding years of income.

A partnership is taxed at the partner’s level and not the entity level, whereby the partners are subject to tax on the partnership’s earnings for each year of income irrespective of whether they are distributed or not (Kenya Revenue Authority, 2016).

The Income Tax Act provides a provision for the exemption of the income of certain entities upon satisfying the following criterion; it is established solely for purposes of the relief of poverty or distress of the public; or it is established for the advancement of religion or education. This is upon satisfying the commissioner that the income is to be expended either in Kenya or in circumstances in which the expenditure of that income is for purposes which result in the benefit of the residents of Kenya (Kenya Revenue Authority, 2016).

All taxable income is assessed in the fiscal year in which the company’s accounting year ends. Consolidated returns are not permitted; each company must file a separate return. The self-assessment and compensating tax returns must be filed within six months of the end of a company's accounting period. Tax installments are due within 20 days of the end of each quarter (except the first installment, which is due in the fourth month of the period), based on the relevant proportion of the estimated current tax or 110% of the tax for the previous year, less previous installments paid and withholding tax deducted at source; the balance of tax, if any, is due four months after the company's year-end. Agricultural companies make their first installment payment 20 days after the end of the third quarter (Kenya Revenue Authority, 2016).

An employer is required to submit quarterly Pay as You Earn (PAYE) returns before the 10th day of the month following the end of each quarter, in respect of emoluments earned in each of the three months and the tax deducted. Late payments of self-assessed tax are subject to a 20% penalty, plus a 2% penalty per month. Late filing is subject to a 5% penalty on any amount still owed four months after the company's year-end(Oxford Business Group, 2016). Despite, there being written rules, provisions and penalties, the decision on whether to comply or not is largely dependent on individual taxpayers. Several theories have been put forward to explain taxpayers’ behavior regarding tax compliance. These include; economic deterrence theory, fiscal exchange theory, optimal tax theory, and political legitimacy theory (Fjeldstad et al., 2012).

2.2.1 Economic deterrence theory

The economic deterrence theory states that taxpayer's behavior is dependent on factors such as complexity of the tax system, probability of receiving audit coverage, penalties for non-compliance, and tax rates among others (Allingham and Sandmo, 1972). This implies a 'cost-benefit' approach whereby it is argued that some taxpayers weigh the benefits of successful evasion against the risk of detection and possible penalties.

Consequently, when the likelihood of detection or penalties is high the likelihood of tax evasion is low and vice versa. There is evidence to support use of this theory by tax administrations in addressing non-compliance. For instance, Chauke and Sebola (2016) in their paper conclude that the deterrence theory is the most applicable in municipalities and the South African Revenue Service revenue collection strategies as taxpayers do not pay taxes willingly but coerced. This study uses this theory to impose a deterrent measure on the taxpayer population by increasing the probability of detection in the event of tax evasion.

2.2.2 Fiscal and Social Psychology Model

This approach focuses on the psychological variables such as moral values, and perception of the fairness of the tax system and tax authorities. The fiscal exchange theory suggests people's perception about the government may motivate compliance (Moore, 2004). For instance, if the citizens are happy with what they get directly from their taxes, they tend to be more compliant. However, if the tax system is perceived to be unfair, tax evasion may be justified by taxpayers as an attempt to get even with the government.

There exists empirical evidence in support of this theory as pointed out by Nikiema (2016). Nikiema carried out a survey in 29 sub-Saharan African countries and concluded that individual's attitude towards paying tax was directly dependent on the quality of institutions. Muralidharan et al., (2017) carried a similar study in India, and concluded that the low quality of education explained the annual tax cost of close to 1. 5 billion.

According to OECD (2004b) tax audits continue to play an important role in enhancing compliance for most revenue authorities. For instance, as seen in the compliance pyramid in Figure 2.1**,** audit is the strategy that allows administrations to exercise effective sanctions against those on top of the compliance pyramid i.e. those that do not want to comply. In addition, besides having a corrective effect that encourages customers to move towards the bottom of the compliance pyramid, audit has a deterrent effect that encourages customers in their groups to be more compliant.

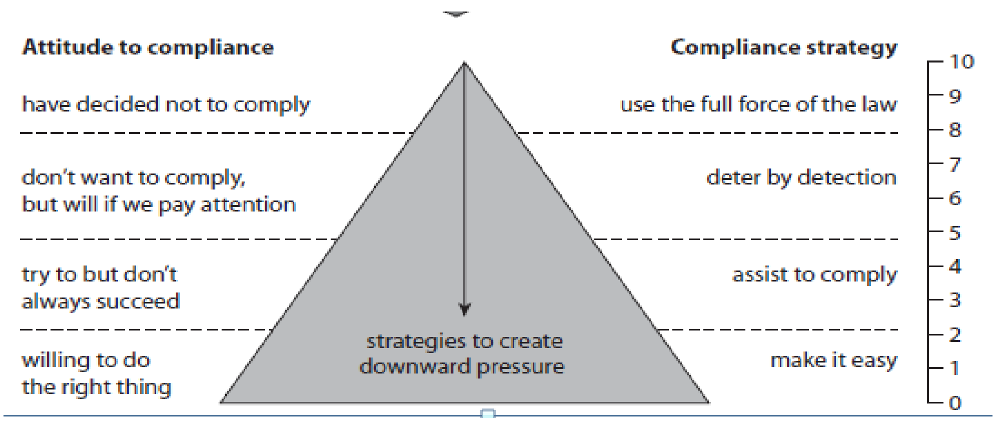


Figure 2.1 Compliance Pyramid adapted from "Compliance Risk Management: Audit case Selection Systems, Centre for Tax Policy and Administration" by OECD, 2004b

This study uses principles of this theory to create a just and fair system in an attempt to lower levels of tax evasion and promote compliance.

2.2.3 Optimal tax theory

This theory suggests that as the government raises a given amount of revenue through taxation, it creates a distortion of economic choices. As a result, how taxes are set and implemented play a big role in reducing this inefficiency and distortion (Bordignon et al., 1997). According to Emmanuel (2012) this theory is used by tax administration in the implementation of a tax system aimed at achieving optimal tax levels. This theory is appropriate in informing this study in analyzing how the implemented tax system affects taxpayer behavior across the rich –poor divide.

2.2.4 Political legitimacy theory

The political legitimacy theory states that political institutions with a higher level of legitimacy lead to higher tax compliance (Kirchler et al., 2008). This implies that when people trust institutions such as the revenue authority, then tax compliance will increase. According to Persson (2008), successful African countries are those that emphasized nationalism as opposed to ethnic identity upon attaining independence. This theory is used to support one of the proposed impacts of the solution which is increased trust in Kenya revenue Authority as a government institution.

2.3 Determinants for Tax Audit selection

Alm, Blackwell and McKee (2004) explored the selection rule for Sales tax in the US and its impact on tax compliance. Focusing on Gross Receipt tax in New Mexico, the authors estimated the process through which firms are selected for audit. Results indicated that returns were selected based upon a systematic, even if informal, audit rule. In addition, firms that exhibit greater variation in deductions, provide services, miss filing deadlines, and have an out–of –state mailing address have a lower compliance rate

According to Gupta and Nagadevara (2007), possible variables in an Audit Selection Strategy include dealer profile (e.g. new registrant , deals in high-tax-rate items, any other business operating from the same address, any other business having the same telephone number) ; Return compliance(non-filling, delay, nil returns); Returned values and ratios(Tax to turnover, gross profit, exempt sales to turnover, inventory to turnover, purchases to sales, refund claimed); Variations in returns across tax periods(tax growth, turnover growth, variance of turnover across periods); and Benchmarking vis-à-vis dealers of similar trade or industry.

Another key factor in creating an effective case selection methodology is information from third parties that can confirm details on tax return, historic cases, and generic taxpayer/ business sector profiles (OECD, 2006). This information should be accessible in a flexible manner and from a wide group as possible in order to identify patterns of non-compliance. Information technology is viewed as an enabler in not only conducting analyses but also has the capability of dealing with large and disparate sources of data that require risk identification.

Vellutini (2011) states that tax compliance depends on individual specific factors such as gender ,age, education; firm factors such as type of industry, firm size, financial situation; perceived fairness of the tax administration, use of public funds, treatment of taxpayers, and perceived compliance culture.

Sirengo (2016) assessed the risk identification criteria at Kenya revenue authority using a logistic regression model. He identified nine important variables which include; business ownership, characteristics of tax agents, performance targets, erratic performance of the sector, and nature of business, financial performance of the taxpayer in terms of profitability and liquidity, company structure and frequency of investment deduction claims. The author concluded that cultural and behavioral factors, control of complex transaction, financial performance, history of taxpayer compliance and erratic factors are significant determinants of payment compliance.

Naibei and Siringi (2011) examined the impact of Electronic Tax Registers (ETRs) on Value Added Tax (VAT) compliance. The empirical results showed that the use of an ETR as well as the frequency of inspection significantly influenced VAT compliance.

In an assessment of studies concerning factors which shape tax compliance behavior, Batrancea et al. (2012) summarized the factors into socio-psychological factors; attitudes, norms, fairness perceptions and motivational postures, Political factors; tax complexity, economic factors; audit probabilities, fines, tax rates and income

Barbutamisu (2011) also reviewed factors influencing tax compliance. The study identified the most important determinants to be economic factors such as the level of income, audit probabilities, tax audit, tax rate, tax benefits, penalties, fines and other non-economic factors such as attitude toward payment of tax, personal, social and national norms and perceived fairness, among others.

Torgler (2003) argues that taxpayer who engages a tax agent is more likely to be compliant than one who does not. He further categorized taxpayers into four classes namely social, intrinsic, and honest and evader. Whereas a social taxpayer will comply on moral grounds, an intrinsic taxpayer will always feel obliged to pay taxes without coercion. An honest taxpayer will also always comply and not attempt to search for loopholes for evading tax. However, a tax evader will react to tax rates and make decisions based on expected personal benefits of evading versus probability of detection.

Pritchard and Khan (2005) developed and applied a logistic regression model to 314 United Kingdom taxpayers to test the relationship between their personal attributes and their non-compliance behavior. Attributes used included source of income, whether the individual is a partner in business, a director of a company, and the nature of business. Results show that the taxpayer’s age, type of business, annual income and location of residence are significant determinants in taxpayer non-compliance. Taxpayers aged between 60 and 70 years are more likely to generate a high yield from an offshore tax avoidance investigation and a sole- proprietorship or partner in a business is more likely to evade tax as well as businesses that have huge volumes of cash transactions.

2.4 Methodologies for Identifying Tax Audit Subjects

Various methodologies exist currently for identifying tax audit subjects. These include but not limited to screening, random selection, risk-based audit selection, Statistical analyses, data matching and data mining.

2.4.1 Screening

Screening involves selection of auditees by auditors based on their knowledge of taxpayer’s behavior and environment. This technique has the advantage of creating less case worker resistance as they would be familiar with the cases. The downside for this technique is that it relies on a limited data set with little or no reference to other data sources. There is also an opportunity cost in asking auditors to undertake screening as they are the same ones who undertake substantive intervention. There may be a gap between those selecting cases and those who will be working on them (OECD, 2004c)

2.4.2 Random Selection

Under this technique, taxpayers to be audited are selected randomly from the overall population of taxpayers. It comes in two flavors; simple random selection or using stratified sampling. While simple random selection gives all taxpayers an equal chance of being audited, stratified sampling groups taxpayers into groups based on the basis of criteria such as size, industry, type of tax to be paid etc. A random sample is then drawn from each stratum (OECD, 2004c).

2.4.3 Risk-Based Audit Selection

OECD (2006) argues that most tax administrations have developed audit strategies focusing on taxpayer noncompliance risks. These selection techniques are inspired by the need to target non-compliant taxpayers only and those that would result in high yields of audit adjustments. An example is a risk-scoring system whereby a score is given to each taxpayer, based on certain attributes (size, industry, compliance history) and (knowledge acquired during previous audit campaigns (whatever the selection strategy).

However, this strategy comes at a cost for in terms of data and Information technology. It requires a significant amount of quality data (internal or external to the tax administration) on both past audit cases as well as current taxpayer attributes. In addition, IT systems capable of processing huge volumes of data and providing scores are needed.

2.4.4 Statistical Analyses

These are base case selection methodologies on the results of statistical analyses. Examples include;

Linear regression analysis: this is the most common predictive statistical technique used when the dependent variable is continuous. Audit case selection relies on well-known results and techniques.

Logistic regression: Hastie et al. (2001) describes logistic regression as a widely used technique to predict the likelihood of binary or more categorical outcomes like good or bad clients and compliant or noncompliant. This model is widely used in banking to estimate credit scores, but less so in tax administration.

Discriminant analysis: The U.S. Internal Revenue Service has used this method for the past 40 years, for predicting certain classes of tax return that fall into high, low categories. The most common linear discriminant function is called the Fisher function or model (Torrey, 2008).

2.4.5 Data Matching

This technique entails checking the consistency of tax returns with other data from customs, bank and insurance company records, and other taxpayers' returns. It does not seek to predict tax evasion but rather to track down events of non-compliance which have already occurred. This is effective for specific tax instruments.

2.4.6 Data Mining

Data Mining can be referred to as a process of discovering patterns in data (Witten et al., 2011).Data Mining is the center of the Knowledge Discovery of Data(KDD) process, involving the deducing of algorithms that investigate the data, build up the model and find unknown patterns (Maimon and Rokach, 2010). There are numerous techniques for Data Mining utilized for various purposes and objectives. Examples include decision trees, neural networks, Bayesian networks, Support Vector Machines, and instance based learning. Figure 2.2 presents Data Mining techniques as classified by Maimon and Rokach (2010).

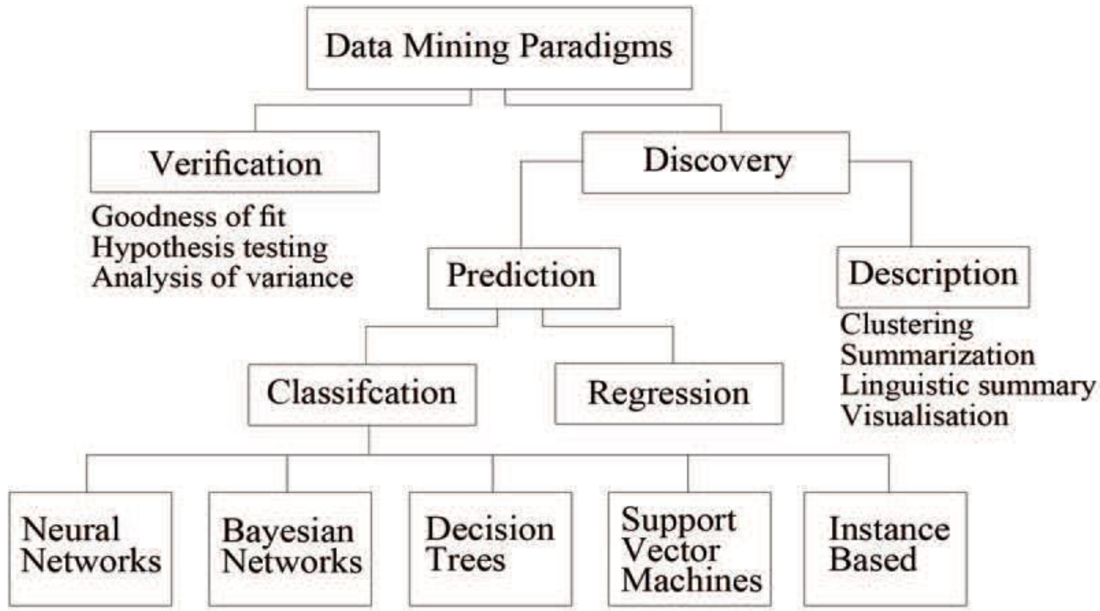


Figure 2.2 Data Mining Techniques adapted from "Data Mining and Knowledge Discovery Handbook" by Maimon and Rokach, 2010

Due to the numerous data mining techniques, choosing the right data mining approach can be a daunting task. Gibert et al. (2010) came up with a classification of the most common data mining methods in a conceptual map making it easier to select an efficient selection. The authors also provided an intelligent data mining assistant whose orientation is to suggest to the user the most suitable techniques for a given problem.

Many different data mining techniques have been used for auditee selection by various researchers. For instance, Hsu et al. (2015) presented a case study pilot project at Minnesota department of revenue that used data mining approach (classification models) in audit selection. Results showed that the data mining based approach as compared to manual screening achieved an increase of 63.1% in efficiency.

Gupta and Nagadevara (2007) in their paper concluded that data mining algorithms are the best cost effective options to make audit selection effective and efficient. The authors came up with eight models and tested them first against each other and then with random selection. All the models were found to be better than random selection whereas the classification tree and hybrid model (classification and logical regression) gave strike rates of 70% and 86%. They recommended that with more input variables, the performance of the models would increase.

Manasan (2003) developed industry benchmarks to aid in selection of VAT returns for audit. These are the ratio of VAT liability to VAT output for industry groupings. The authors concluded that excessive claims for VAT credit are a major source of revenue leakage in Philippines and that tax administrations can predict these claims using input-output ratio and industry benchmarks.

Shao et al. (2002) developed a fraud detection model named Intelligent Eyes for Qingdao customs port at China to give decision rules to the custom officials for inspection of goods on the basis of past transaction data, with the objective of improved hit rate.

Kumar (2005) developed a predictive risk assessment model using classification tree algorithm. The model is expected to detect anomalies in selective customs examinations in Indian customs. The model has been developed using classification tree algorithm and is expected to detect over 90% of the total duty short declarations with mere 30% of the original examination effort.

2.4.7 Critique of the Methodologies used for Identifying Tax Audit Subjects

Selection of auditees by auditors has the advantage of enabling case acceptability on the part of case workers. However, it is limited to a data set with little or no reference to other data sources. In addition, case workers can significantly miss some aspects of non-compliance which they are not familiar with.

Random selection on the other hand is perceived a fair strategy in that all taxpayers have an equal chance of being audited. It also prevents the risk of corruption or arbitrary selection. The downside of random selection is the possibility of selecting an honest taxpayer resulting in wastage of time and resource on both the taxpayer and tax administration.

Statistical analyses such as linear regression seem to be an effective technique though limited to a continuous dependent variable. Logistic regression however, provides a remedy to this limitation by having the capability of handling categorical.

Risk-based techniques such as a risk-scoring system provides a mechanism of focusing on high risk taxpayers or sectors based on their attributes as defined by a tax administration. However this technique requires a significant amount of quality data both internal and external to the tax administration as well as IT systems capable of processing huge volumes of data.

With advanced analytics, data mining techniques have proven to be effective solving many difficult and complex problems in today's world including in tax administration. However there is need to select an algorithm that works best for a given problem in that they all have advantages and disadvantages. For instance, decision tree algorithm results in fast learning, fast prediction, understandable rules, and lower memory requirements. However, it has a replication problem where each category requires multiple branches; a limited rule representation where attributes are assumed to be locally independent and difficulty handling numeric attributes.

On the other hand, neural networks are advantageous in that they act as general purpose learner and allow fast prediction. However, neural networks trains slowly and all inputs have to be translated into numeric inputs. In addition, learning might result in a local optimum. Bayesian networks have an efficient inference mechanism, readable structure; are relatively easy to design and have mechanisms for learning network structure. The disadvantage comes in trying to build the network automatically. Bayesian network does not also handle sequence information.

In conclusion, there is no "best" prediction approach. A prediction approach that is suitable for a particular problem would best suffice. It would be in the best interest of the tax administration to understand the attributes and rules derived from a given algorithm. Additionally, since taxpayer attributes are both categorical and numerical, an algorithm that handles these attributes easily will best suit the problem in this thesis. As a result, decision tree algorithm due to its suitable advantages was chosen for use in developing the model for selecting tax audit subjects.

2.5 Decision Trees

This research proposes to use decision trees for classifying taxpayers as compliant/non-compliant. Decision tree known as Iterative Dichotomized, is one of the most well-known and used classification algorithms since 1970. Improvements on this algorithm has occurred over the years. For instance, Breiman et al. 1984 introduced a Classification and Regression Trees (CART) which was utilized to produce binary decision trees. Quinlan (1993) and Han et al. (2012) later introduced C4.5 algorithm which has turned into a benchmark to which recent supervised learning algorithms are regularly compared.ID3, CART, and C4.5 use a greedy approach in which decision trees are constructed in a top-down recursive divide-and-conquer way (Han et al., 2012). However, C4.5 manages continuous attributes and handles missing values even though they are a bit slower.

2.5.1 How to develop a Decision Tree

Decision tree is a directed tree that obtains its structure by recursively separating the set of observations. It consists of a root with no incoming edges, internal or test nodes with exactly one outgoing edge for each, and leaves which represent the decision node and have no outgoing edges (Maimon and Rokach, 2010). The decision tree development algorithm is a greedy algorithm which is a top-down recursive divide-and-conquer in nature. The algorithm is represented below (Kargupta et al., 2008):

Algorithm 1: Generate-Decision-Tree (samples, att-list)

1: **Input:**

2: Samples: training samples

att-list: set of candidate attributes
Create a node N // represent the training samples
If samples are all of the same class, C **then**
return N as a leaf node labeled with class C;

If att-list is empty **then**
return N as a leaf node labeled with the most common class in samples;

Select test-attribute, the attribute among attribute-list with the highest
information gain based the Entropy;
Label node N with test-attribute;

for each known value a_i of test-attribute **do**
Let s_i be the set of samples for which test-attribute= a_i;
If s_i is empty **then**
attach a leaf labeled with the most common class in samples;
else attach the node returned by Generate-Decision-Tree (s_i,att-list)
end if
end for

To simplify the decision tree, pruning algorithms were introduced. Pruning is a measure against over fitting and impacts the tree size as well as accuracy in that it results in improved accuracy (Witten et al., 2011).Using Decision Trees, taxpayers can be classified as compliant or non-compliant.

2.5.2 How to Select Tree Root

We need to figure out which attribute can fill in as a root of a tree given an arrangement of training vectors. Information gain gives the significance of specific attribute and criticality of certain trait element vectors. Information gain helps choosing helps choosing the ordering of attributes in the nodes of a decision tree (Han et al., 2012).

$$\text{Information Gain} = \text{E (Parent)} - \text{AE (Children)} \qquad \text{Eq. 2.1}$$

$$\text{Entropy} = \sum_{i} -pi \log2 pi \qquad \text{Eq. 2.2}$$

E, AE and pi are the entropy, average entropy, and the probability of class i respectively. Entropy comes from information theory where higher entropy implies greater information content. For example, given training information set in Table 2.1, the table has three components f1, f2 and f3 and the two classes A and B. Assuming that f1 as the best characteristic to split the table, this node would be further split. Thus, the entropy of children and the gain can be computed as follows:

Table 2.1: Training Set

F1	F2	F3	Class
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

$$\begin{aligned} E_{child_1} &= -\frac{1}{3}log_2(\frac{1}{3}) - \frac{2}{3}log_2(\frac{2}{3}) \\ &= 0.5284 + 0.39 \\ &= 0.9184 \\ E_{child_2} &= 0 \end{aligned}$$

$$
\begin{aligned}
E_{parent} &= 1 \\
IG &= 1 - \frac{3}{4} \times (0.9184) - \frac{1}{4} \times (0) \\
&= 0.3112
\end{aligned}
$$

If we split using the feature f2, we get the following:

Echild1 = 0

Echild2 = 0

$$
\begin{aligned}
E_{parent} &= 1 \\
IG &= 1 - \frac{1}{2} \times (0) - \frac{1}{2} \times (0) \\
&= 1
\end{aligned}
$$

Splitting using feature f2 produces the best gain. The developed tree structure in this case can be presented as in Figure 2.3.

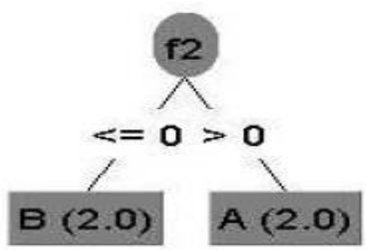


Figure 2.3 Tree structure adapted from "Special Interest Group on Knowledge Discovery and Data Mining" by Hall et al., 2009

2.6 Conceptual framework

Literature reviewed in section 2.4.6 points out the use of various data mining techniques in building predictive models. This study sought to develop a model for tax audit case-selection. This was achieved by first preparing a training data file by extracting a group of already profiled compliant and non-compliant taxpayers together with their corresponding attributes from the taxpayer database. The attributes used were those pointed out by literature in section 2.3 as determinants of tax audit case-selection as well as the knowledge of tax experts. The training file was used to train and build a decision tree classifier. The model was then used as an input to a case-selection application prototype for identifying cases for audit.

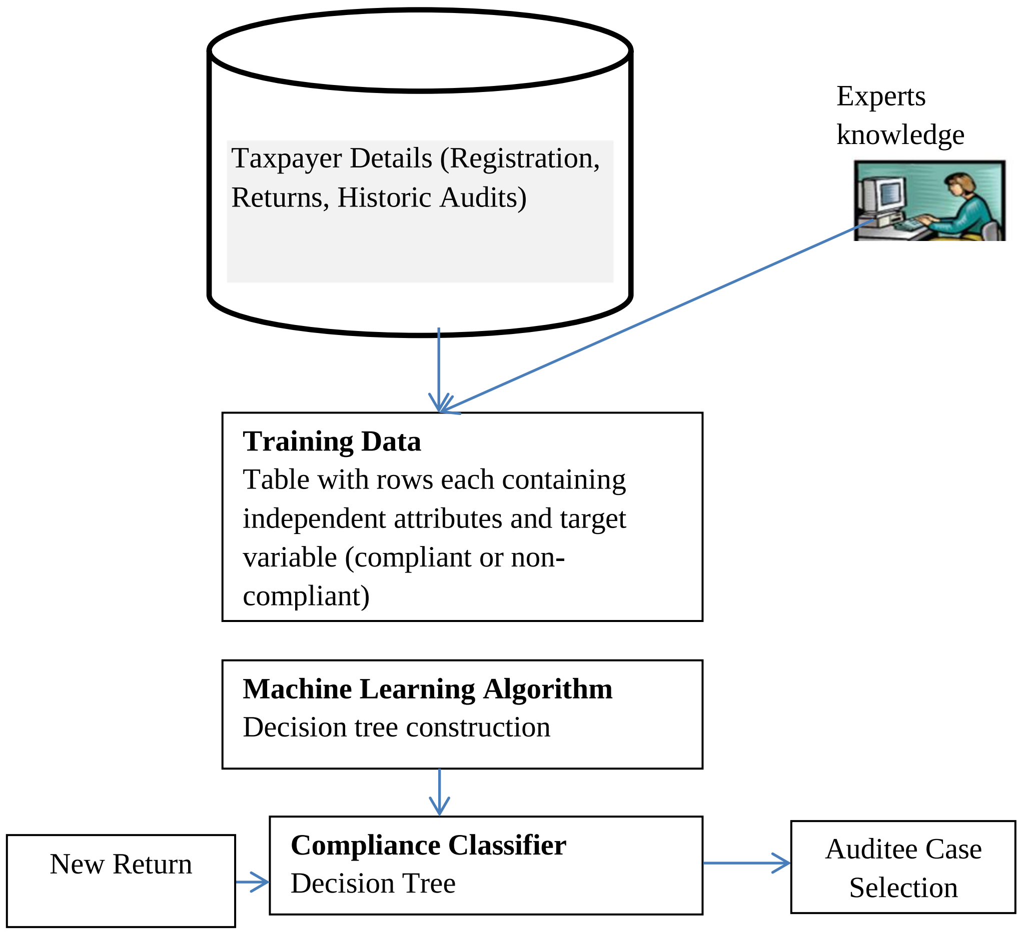


Figure 2. 4 Conceptual Model of Proposed Solution

Chapter 3: Research Methodology

3.1 Introduction

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 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).

This chapter discusses the research methodology that was used to conduct the research. It lays its focus on describing the research site, research design, population, sample design, data collection methods, data analysis methods and the research quality aspects. For each research method chosen, a justification for their viability is also provided.In addition, a description of the system development methodology used in developing the model is provided. Finally, this chapter provides a description of how the model was developed and evaluated.

3.2 Research Design

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).

The research adopted a positivism paradigm and was quantitative in nature. The research focused on accumulated theories, knowledge, methods, and techniques for the development and testing of a case-selection model to be used by tax experts in selecting audit subjects. This research also o employed survey method to get the needs and challenges of case-selection and experiment method for building the model. Nonetheless, the study also employed qualitative research to get factual information through truthful reporting, and firsthand experience of respondents carrying out case selection. This approach aided in the classification of taxpayers under various categories.

3.2.1 Prototype Development Methodology

Software development process describes an approach to building, deploying and possibly maintaining software. Various methodologies exist which include but is not limited to; waterfall, Rapid Application Development, Agile, and Prototyping (Larman, 2004). A particular methodology or a hybrid methodology is chosen based on the nature of the project, project objectives and time constraints.

3.2.1.1 Agile Unified Process Methodology

This research adopted the Agile Unified Process (AUP) Methodology for the development of the case selection application prototype. Agile Unified Process is a hybrid modeling approach that combines the Rational Unified Process (RUP) to agile methods (AM), Palaiologou et al, (2010) as cited by Edeki, (2013). Rapid Unified approach thus brings to AUP an iterative development approach that is able to provide high-quality software that meets the expectations of its users whereas values, principles and practices of software development are borrowed from the Agile approach. These principles are; adaptive planning, evolutionary development, early delivery, continuous improvement, and rapid and flexible responses to change.

The Agile Unified Process has four main phases: Inception, Elaboration, Construction and Transition, as shown in Figure 3.1 below. The **Inception** phase provides an unrefined ‘blurry’ vision of the system. The business case is developed- the scope of the project is determined using vague estimates. The **Elaboration** phase provides a refined vision, and an iterative implementation of the core architecture and resolution of high risks. The elaboration phase enables the identification of most requirements and scope, providing realistic estimates. The third phase, **Construction** provides an iterative implementation of the remaining lower risk and easier elements. It also prepares the model for deployment. Lastly, the **Transition** phase involves system testing and final deployment.

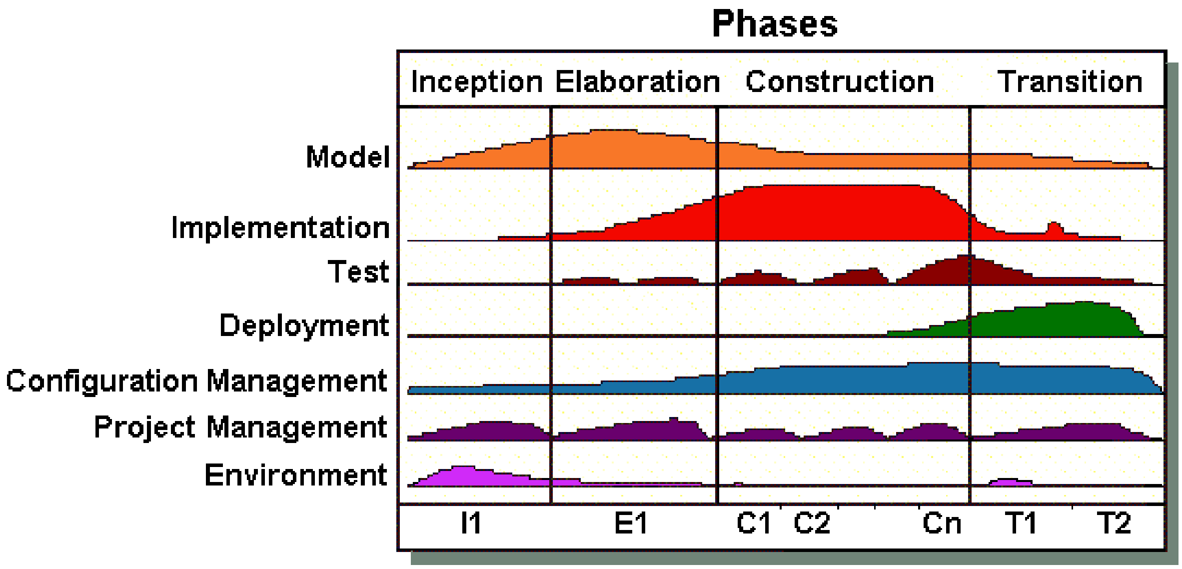


Figure 3.1 The Life Cycle of the Agile Unified Process adapted from "Agile Unified Process" by Ambler, 2005

The model thus begun with a simple implementation of a small set of software requirements, which was iteratively enhanced until the final prototype was implemented.The benefit of this iterative development was mitigation of high risks earlier on in the project. These risks revolved around the technical aspects, requirements, objectives or usability. In addition, users were able to see visible progress as early as with the first prototype, engaged at this stage and their feedback used to refine the final prototype. This approach also assisted in continual improvement of the final prototype due to managed complexity and learning within iterations.

3.2.1.2 Agile Knowledge Discovery in Databases

For the data mining exercise, Agile Knowledge Discovery in Databases, an approach that supports knowledge discovery in databases was used. The Agile Knowledge Discovery in Databases methodology is illustrated in Figure 3.2.

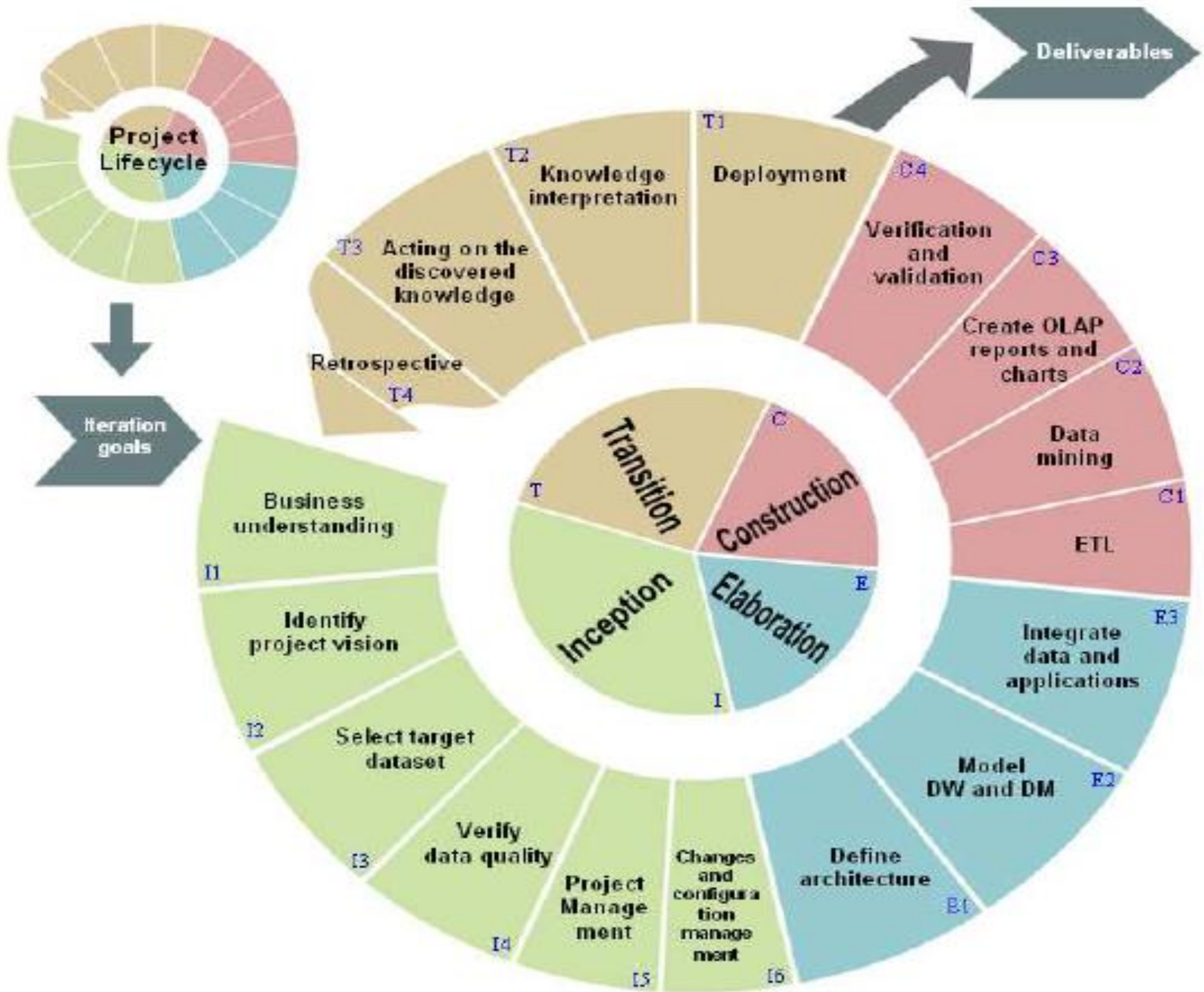


Figure 3.2 Agile Knowledge Discovery in Databases

3.2.2 Requirements Specification

Requirements elicitation was done through literature review, analysis of procedure manuals for case selection, observation and in-depth interviews with expected users of the system. An interview guide provided in **Appendix A** was used to guide the interview process. Requirements were categorized using the FURPS+ model, which defines the Functional, Usability, Reliability, Performance, Supportability, plus other sub-factors- the Implementation, Interface, Operations, Packaging and Legal Requirements.

A use case diagram was used to capture actors, their goals, and the boundary within which the system would operate. Several use cases were used to capture user's functional requirements with each use case capturing the sequence of success and failure scenarios in achieving a particular goal. This technique was used to ensure concentration on user needs and to suspend design for a later time to ensure that the system met user needs.

Additionally, a prototype demonstration was conducted to show the functions and features of the prototype and requirements further refined.

3.2.3 Prototype Analysis and Design

Object-oriented analysis and design, an approach that emphasizes the representation of objects was adopted for analysis and design. In addition, Unified Modelling Language (UML), a language for specifying, visualizing, constructing, and documenting software artefacts was used for this phase. A domain model was used to capture concepts and their interaction in the case-selection problem domain. An interaction diagram was used to capture the main use case of the prototype. A Design class diagram was used to capture software class definitions, their attribute types and methods. For modelling process control, a context diagram was used to capture the overall system architecture and level 1 Data flow diagram to capture data flow across the main processes of the prototype. These artifacts were created using Visual Paradigm Software.

3.2.4 Prototype Implementation

Implementation of the case selection prototype was iterative through the Elaboration, Construction and Transition phases. Interaction diagrams and Design class Diagrams generated in the design stage was used as input to the code generation process. The deliverables for this phase was a decision tree model, source code and executable file.

The coding environment comprised of ; Microsoft Windows 7 operating system as a platform; J48, a java version of C4.5 algorithm for the classifier; and Netbeans Integrated development environment. Netbeans Integrated development environment was chosen due to the researcher's familiarity and ease to use and troubleshoot errors. Java programming language was preferred as it was easier to translate designs already done in object-oriented approach to code. Windows 7 operating system was chosen as a stable platform to work with.

3.2.5 Prototype Testing

Testing of the model begun at the Elaboration phase, and was done again in the Construction phase.Testing was done to demonstrate the functionality of each module to ensure it conformed to user needs. Usability testing was also carried out to test usability aspects. Thus, users were presented with a system evaluation questionnaire to express their feedbacks towards the developed prototype. The evaluation questionnaire used is illustrated in **Appendix C.**

3.3 Population and sampling

3.3.1 Research Site

The research was conducted at Kenya Revenue Authority Headquarters office, Nairobi. This site was chosen as data for the study was easily obtained from this place and the location was also accessible to the researcher.

3.3.2 Population

The population refers to the entire group of people, events, or things of interest that the researcher wishes to investigate (Kothari, 2009).The population of study used in this research comprised of compliant and non-compliant taxpayers registered for Income Tax-Company tax obligation in the year 2014. According to Kenya Revenue Authority (2015), the number of taxpayers registered with the authority is 8.1 million, with only 1.6 million being active. For those registered for income tax-company tax obligation, their number stood at 400,000 as at 2015. It was difficult to get the exact number of non-compliant taxpayers from this population. However, it was assumed that 45 % of these are non-compliant.

3.3.3 Sampling Design

Since the population under study does not constitute homogeneous groups, the research used stratified sampling to obtain a representative sample (Kothari, 2009). Consequently the population was divided into two strata where the first strata constituted taxpayers who were found to be evading tax and penalties issued on them whereas the second strata constituted of taxpayers who did not evade tax.

The proposed sample size was 1500 records with a proposed proportional allocation of 55:45 whereby 55 % represented compliant taxpayers and 45% represented noncompliant taxpayers. As a result the sample sizes for the various strata were computed by the formula given below;

$$nh = P1 * n$$

Where n_h is the sample size, P_i the proportion of strata i and n the sample size. Therefore the sample size for compliant strata is (55/100* 1500) = 825 records and (45/100*1500) =675 records. For sampling within each stratum and extraction from the database, systematic sampling was used.

3.4 Data Collection Methods

This research used both primary and secondary data. Primary data was used to determine the need and challenges of tax audit case-selection whereas secondary data was used for data mining, to identify techniques used to select cases together with their limitations.

3.4.1 Primary data

This data was collected through survey method which will include interviews, observation and questionnaire.

3.4.1.1Interviews

In-depth interviews were carried out with officers within the Compliance department in Kenya Revenue Authority. An interview guide with open ended questions was used to guide the interview. This is illustrated in **appendix A.** Interviews were chosen as they would best unearth details of the case-selection process as opposed to questionnaires**.** Depending on the respondent's response, additional questions were also raised and responses recorded**.**

3.4.1.2 Questionnaire

An evaluation questionnaire meant for obtaining information about importance and necessity of auditee case selection model and prototype was used to find the number of people who thought it was a good idea to come up with the model and prototype. This is illustrated in **appendix C.**

3.4.2 Secondary data

Secondary data for data mining constituted two samples of already profiled compliant and non-compliant taxpayers together with their corresponding attributes extracted from taxpayer database with the help of a Relational Database Management System RMDS and stored in a MySQL database. Additional data was also obtained from reputed journals, articles, websites and Kenya Revenue Authority procedure manuals.

3.5 Data Analysis

According to Kothari (2009), data analysis encompasses the processing and analysis of data. Specifically, processing implies editing, coding, classification and tabulation of collected data so that they are amenable to analysis. Analysis on the other hand refers to the computation of certain measures along with searching for patterns of relationship that exist among data-groups.

Accordingly, content analysis was used to analyze and make replicable and valid inferences by interpreting and coding textual material. On the other hand, data extracted from the database was tabulated into rows and columns and categorized into the two classes of compliant versus non-compliant. This data was pre-processed in WEKA software whereby missing values were filled or the records completely removed. Using WEKA, significant attributes that affect compliance/ non-compliance of tax were identified using the attribute selection feature. Those features found to be insignificant were dropped from the final table that was used as the training table/ file. For presenting the results of the data analysis phase, graphs and charts were used.

3.6 Research Quality

Research quality refers to the degree to which research was carried out correctly. To test research quality aspects validity and reliability were used.

3.6.1 Validity

According to Jopee (2000) as cited in Golafshani (2003) ,validity determines whether the research truly measures that which it was intended to measure or how truthful the research results are. Since the aim of this research was to develop a model for audit selection, validity would be to evaluate the model performance. Gupta & Nagadevara, (2007) in their paper, audit selection strategy for improving tax compliance used accuracy rate, prediction efficiency and strike rate in evaluating performance of the model.

This research uses similar performance evaluation measures to evaluate the model; accuracy rate was used to measures the proportion of cases that were correctly predicted by the model, prediction efficiency was used to measure the proportion of noncompliant cases which are correctly predicted by the model and strike rate was used to measure the proportion of noncompliant cases likely to be detected if predicted noncompliant cases are audited. In testing whether the model was of value to Kenya Revenue Authority, survey questions were sent to respondents and responses analyzed.

3.6.2 Reliability

Jopee (2000) as cited in Golafshani (2003) refers to reliability as the degree to which an assessment tool produces stable and consistent results. To test whether the model produced same results, various test scenarios were presented to it. For instance, independent tests using 10 fold cross- validation test and percentage split ratio of 66% was performed with the model producing similar accuracy rates.

3.7 Ethical Considerations

The researcher ensured that consent was obtained from the respondents before embarking on her survey and ensured that the data collected was solely used for the research purpose. Tax information is confidential, as a result, the researcher ensured to maintain the confidentiality of the data obtained. The researcher also properly cited the work obtained from other authors.

Chapter 4: System Design and Architecture

4.1 Introduction

The main purpose of this study is to develop a model for selecting cases for audit by effectively evaluating taxpayer tax compliance. Object oriented analysis and design were used in this research.

This chapter focuses on system analysis and design of the model. These two are discussed in detail through this chapter with system analysis focusing on data collected from proposed users of the system. Thus, a domain model, use case diagrams and system sequence diagrams are used in this phase. On the design phase, a design class diagram is used to define software classes for the application. Finally, for modeling the flow of data, data flow diagrams are used.

4.2 Results from Interview and Secondary Data

Data was extracted from taxpayer database and interview results are discussed below;

4.2.1 Secondary Data

A total of 1500 records were extracted of which 45% represented taxpayers who had been profiled to be non-compliant and 55% represented taxpayers profiled to be compliant. This is illustrated in Figure 4.1.

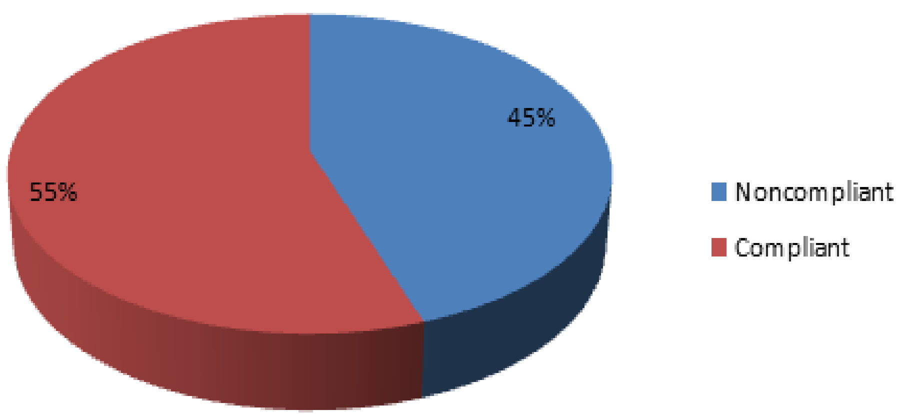


Figure 4.1 Distribution of Secondary data extracted

4.2.2 Challenges faced by compliance officers in identifying cases for audit

Respondents highlighted the challenges they faced in identifying cases for audit as; a time consuming exercise, the process is largely dependent on the knowledge of compliance officers and they could only cover few taxpayers majorly those with turnovers of over a billion Kenya shillings. This is illustrated in figure 4.2.

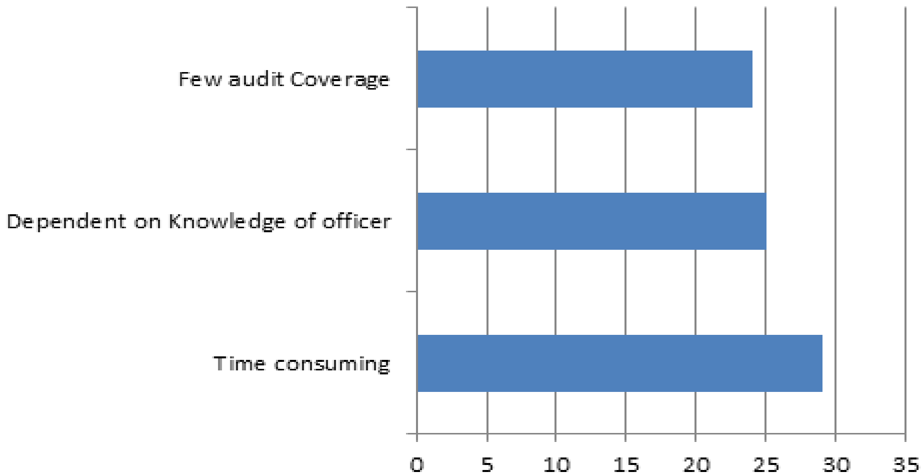


Figure 4.2 Challenges in identifying cases for audit

4.2.3 Need for Case-selection application

Majority of respondents identified the need for a case-selection application. This is illustrated in figure 4.3 where 94% responded with Yes and 6% responded with No.

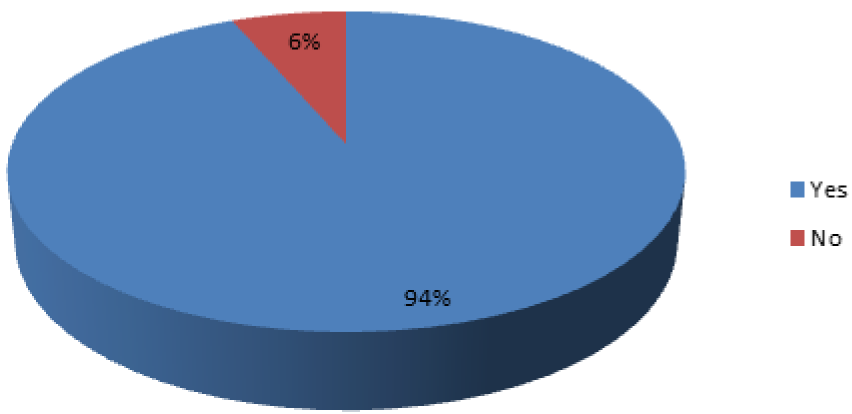


Figure 4.3 Need for case selection application

4.2.4 Features to be present in the case- selection application

Respondents highlighted various features to be present in the case- selection application. These are illustrated in Figure 4.4.

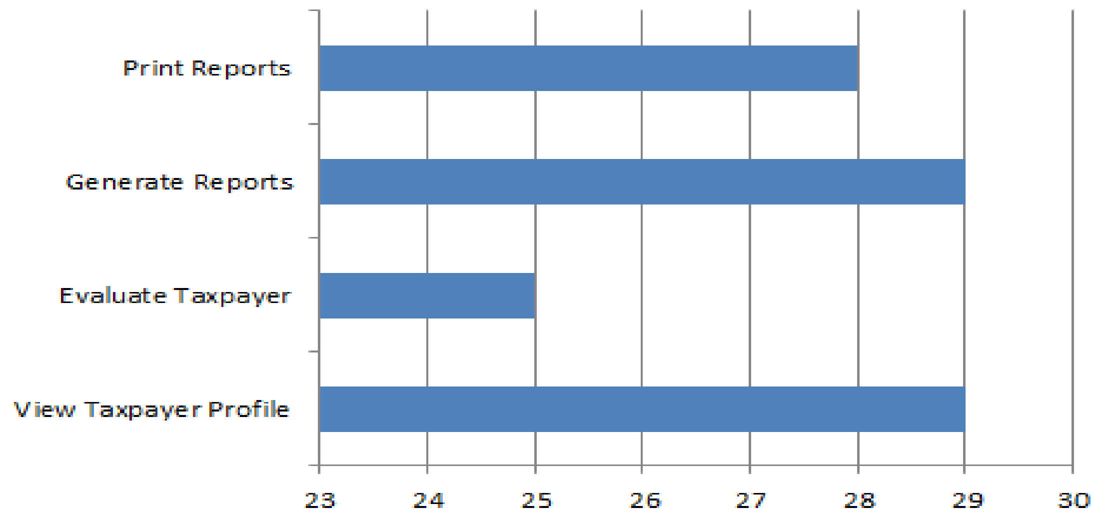


Figure 4.4 Features of Case selection application

4. 3 Requirement Analysis

This entails analysis of user needs regarding the proposed model. Interview feedback from Compliance Officers as illustrated in **appendix B**, use case diagram, use cases, domain model and system sequence diagram was used to carry out this task resulting in detailed descriptions of services, features and constraints to be addressed by the case-selection model. Moreover, these services and features were grouped into functional and non-functional requirements using the FURPS + model.

4.3.1 Functional Requirements

Functional requirements include the features, basic processes, security and capabilities that the implemented system should have. They include: Log in (Users and admin should be able to log in to the application using their credentials) Classify Taxpayer (System should be able to give taxpayer compliance status), View Classification (Users should be able to view the classification of one or more taxpayers), Generate Report

(Users should be able to generate reports), Print Report (Users should be able to print reports), Manage users (administrator can add, update or delete users)

4.3.2 Non-Functional requirements

The non-functional requirements include speed, security and usability aspects of the application

4.4 System Architecture

Data is extracted from the Taxpayer Database and fed into a data warehouse. Training, test and validation data are then obtained from the data warehouse and supplied as inputs to a classifier whose output is then used as an input to an Auditee Case-Selection application. Auditee case-selection application is a two tier application comprising a stand-alone application and a database. A Compliance officer logs in to the Auditee Case-Selection Application, selects a taxpayer or group of taxpayers to classify. The application classifies the taxpayer and displays the classification to the user. An administrator accesses the backend application for administrative tasks. This is illustrated in Figure 4.5;

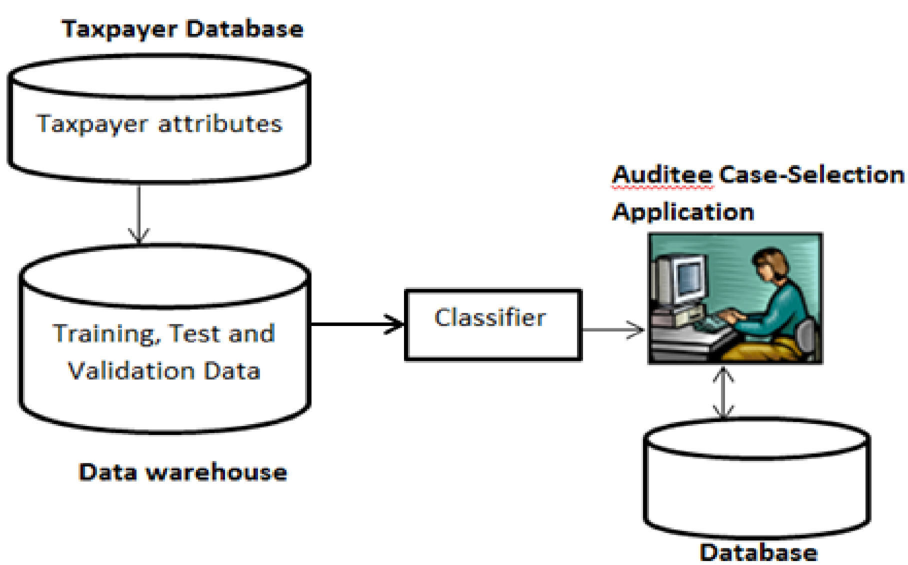


Figure 4.5 System Architecture

4.5 Domain Model

A domain model is used to visually illustrate meaningful conceptual classed or real-world objects in the domain of interest. The domain model consists of the concepts, the association between the concepts/objects and the attributes of the conceptual classes. The concepts of the case-selection application and corresponding attributes include Compliance Officer (name, email) Taxpayer (pin), Taxpayer description (taxpayer attributes), Classifier (model), Administrator (name, password), and Classification Report (ReportId). This is illustrated in the domain model in Figure 4.6;

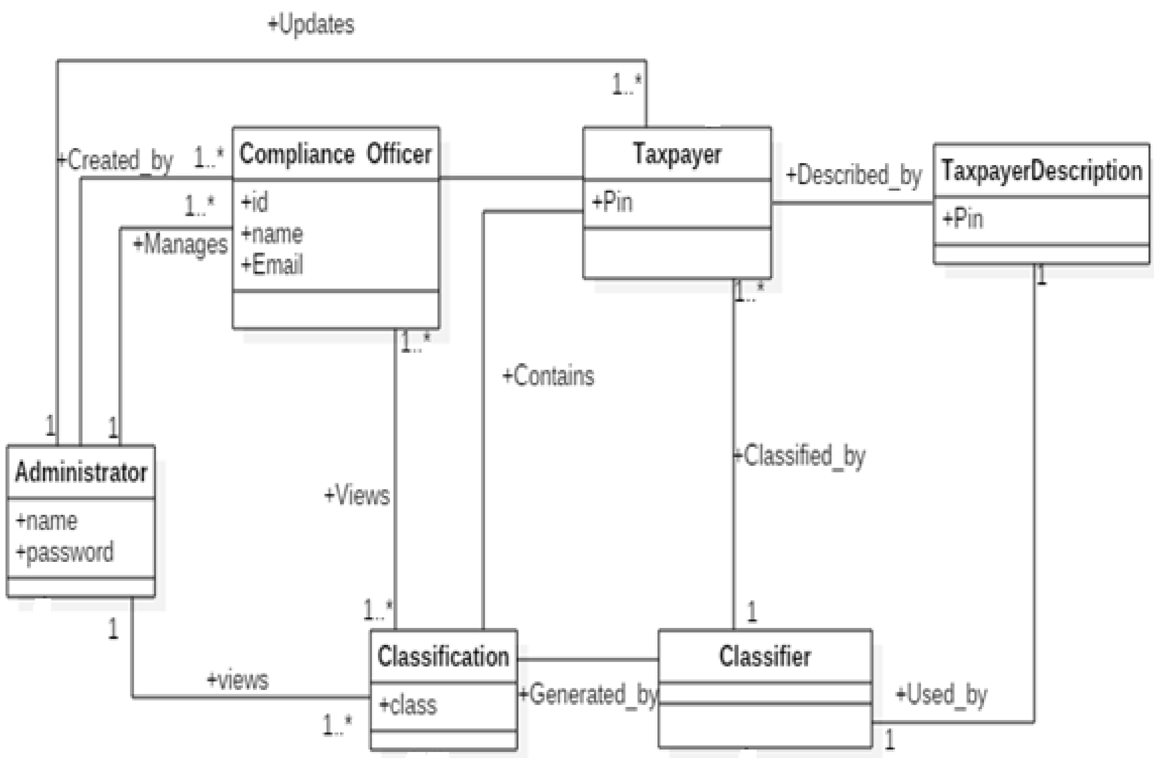


Figure 4.6 Domain Model

4.6 Use Case Diagram

Auditee Case-Selection Application has two actors i.e. Compliance Officer and System Administrator. The use case diagram in figure 4.7 illustrates the names of use cases (collection of success and failure scenarios), actors (something that interacts with the system) and the relationship between them. The boundary represented in the use case diagram is Auditee Case-Selection.

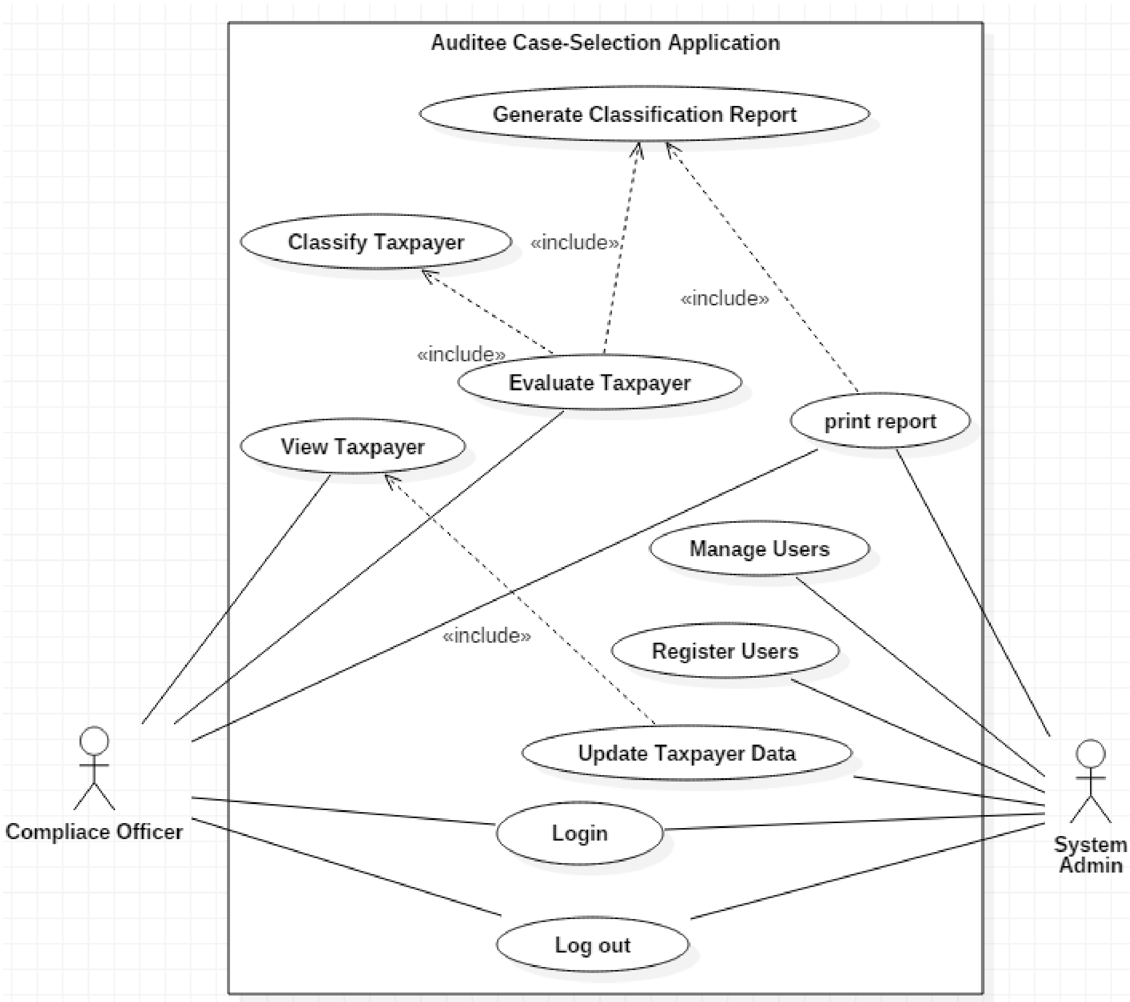


Figure 4.7 Use Case Diagram

The use case for carrying out Taxpayer evaluation is described below;

Use Case: Evaluate Taxpayer

Primary Actor:

Compliance Officer

Preconditions:

Evaluation underway

Post Conditions:

Correct Classification of Taxpayer

Main Success Scenario

1. Compliance officer searches for a taxpayer to evaluate
2. Compliance officer adds a taxpayer to evaluation list
3. Compliance officer requests for evaluation of taxpayer on evaluation list
4. System evaluates taxpayer
5. System displays taxpayer evaluation

Use Case: Classify Taxpayer

Primary Actor:

System

Preconditions:

Evaluation list not empty

Post Conditions:

Correct Classification of Taxpayer

Main Success Scenario

1. System classifies every item on evaluation list
2. System returns classification of evaluation list item

Use Case: Register User

Primary Actor:

System Administrator

Preconditions:

Administrator logged in

Post Conditions:

User registered in the System

Main Success Scenario

1. Administrator starts a new registration

2. System displays registration form
3. Administrator enters user details
4. Administrator saves user details
5. System saves record and displays feedback
6. System sends request to Classifier Service, requests for classification

Use Case: Manage User

Primary Actor:

System Administrator

Preconditions:

Administrator logged in

User already registered

Post Conditions:

User details updated

Main Success Scenario

1. Administrator starts a new update
2. Administrator enters user's name or id
3. System returns user details
4. Administrator enters user details to update
5. Administrator saves user details
6. System records user details and displays feedback

4.7 System Sequence Diagram

The main feature of the application is the evaluation of Taxpayers. The sequence diagram in Figure 4.8 shows the interaction of various entities to achieve the user goal.

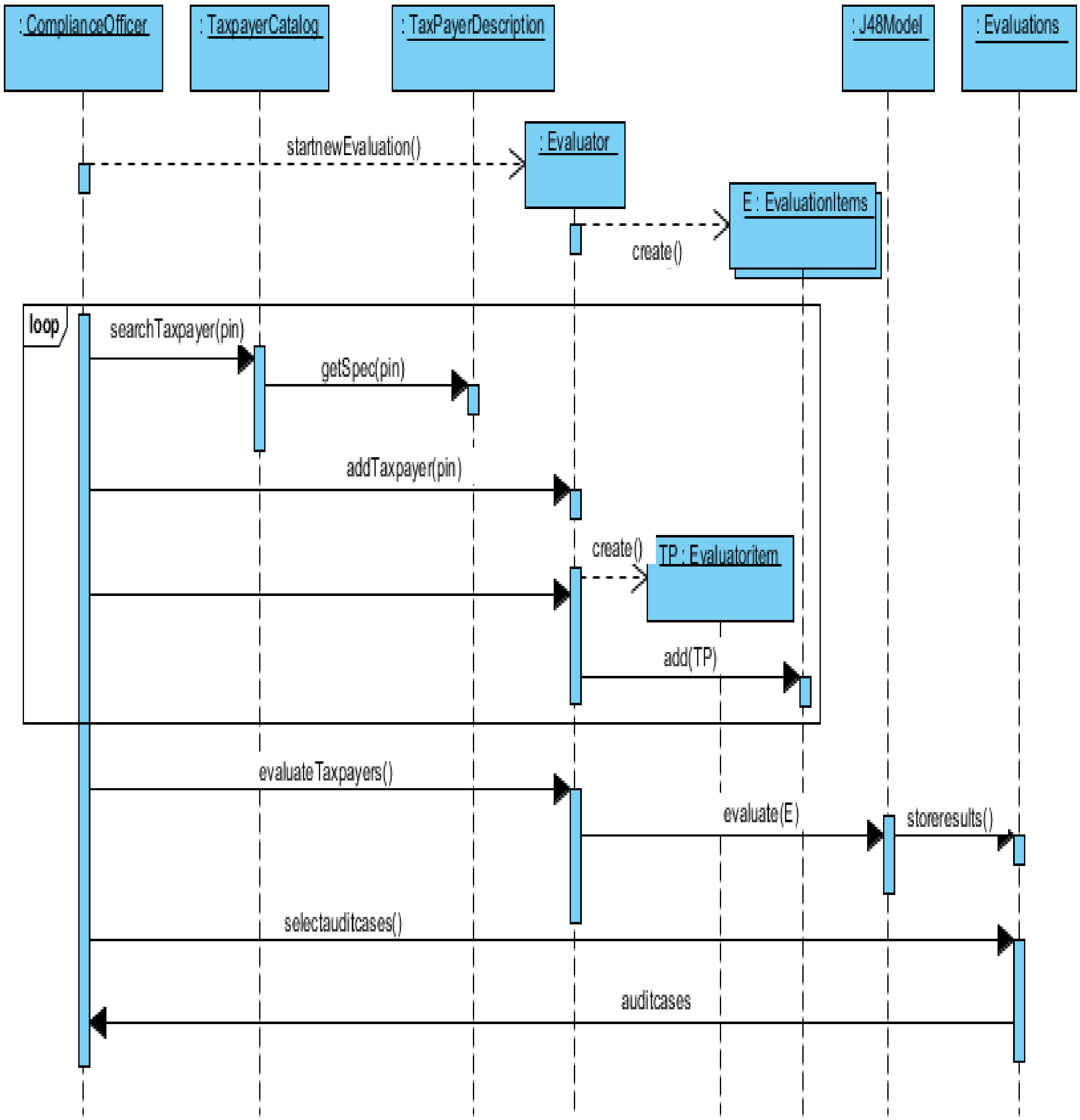


Figure 4.8 System Sequence Diagram for Evaluate Taxpayer Process

4.8 System Design

System design involved coming up with definitions of software classes as well designing the security aspects of the application. A design class diagram was used to capture the definition of all software classes, their attributes, methods and interactions. This is illustrated in Figure 4.9.

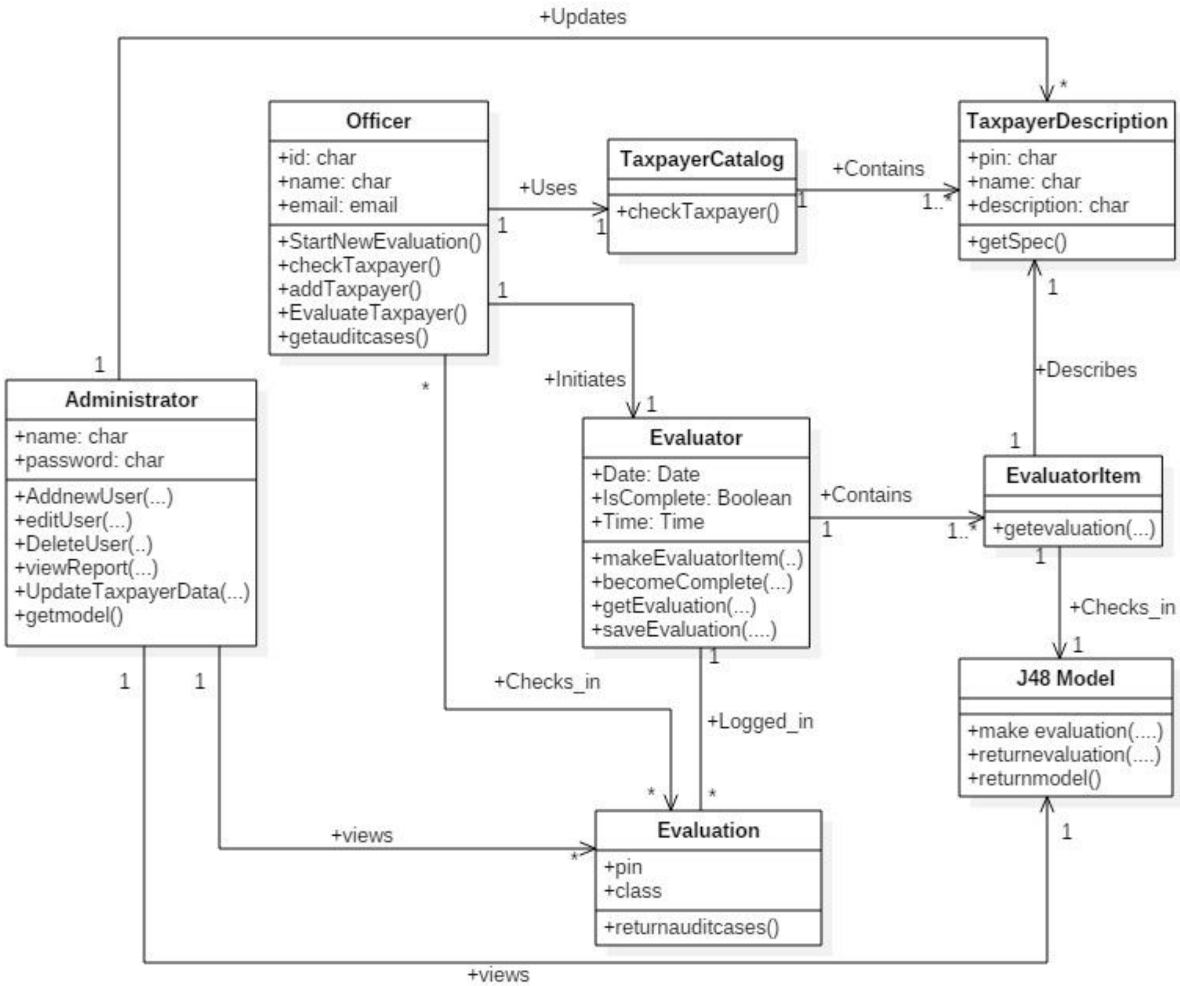


Figure 4. 9 Design Class Diagram

4.9 Security Design

Security design of the system and data was taken into consideration. To ensure data security, sensitive data such as administrator password was hashed in the application and in storage. The system also authenticated all users to ensure only authorized users accessed the application. The administrator functions were also separated from user functions on the interface.

4.10 Process Control

To map processes and flow of data from one entity to another and across processes, process to data flow diagrams were used. The entities interacting with auditee case selection application include Compliance officer, Integration Subsystem and Administrator. The integration system loads taxpayer data from registration database into the Auditee case-selection application. The compliance officer supplies taxpayer details for evaluation into the application and provides a report of the evaluation. The administrator performs administrative tasks such as registering end users and generation of reports. This is illustrated in the context diagram in Figure 4.10.

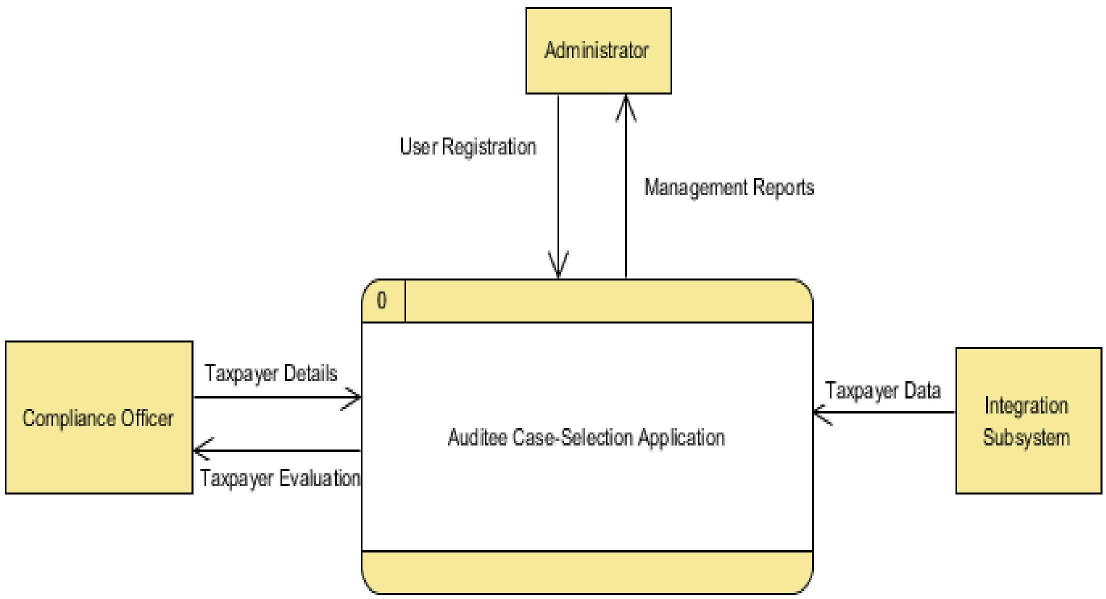


Figure 4.10 Context Diagram

There are 7 main processes in the application which include; view taxpayer, evaluate taxpayer, obtain taxpayer details, maintain taxpayer, create users and maintain users. The obtain taxpayer process receives data from integration subsystem and stores it in taxpayer table. The view taxpayer process and evaluate taxpayer process obtain their data from taxpayer table, processes it, and stores the output in an evaluation table as well as provide output to the compliance officer. Administrator supplies user details to the create user and maintain user process for registration or maintenance respectively after which it is stored in the user table. The prepare management report process obtains its authoritative data from user, evaluation and taxpayer tables for processing and provides output reports for administrator and compliance officers. This is illustrated in Figure 4.11.

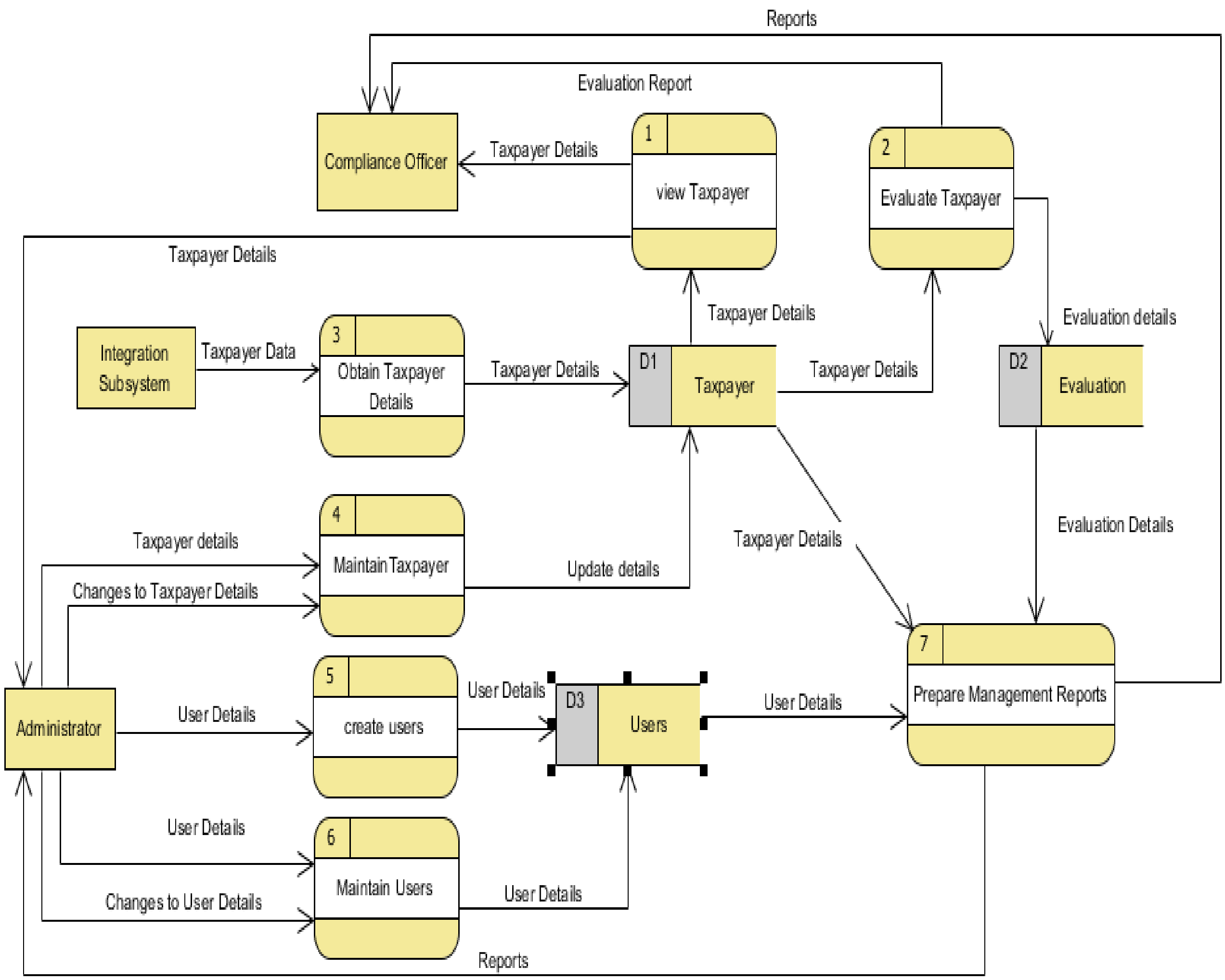


Figure 4.11 Level1 DFD

Chapter 5: Prototype Implementation and Testing

5.1 Introduction

This chapter focuses on the implementation and testing of the proposed application prototype. The chapter will begin with a description of the development environment, the experimental set up for building the classifier and then focus on the discussing the development of the application prototype. Finally, the chapter will lay its focus on functional and usability testing of the prototype.

5.2 Development Environment

Suitable development environment had to be established to ensure that the implementation process runs smoothly. Tables 5.1 and 5.2 describe the software and hardware requirements for the development process.

5.2.1 Software Requirements

Table 5.1 describes the software requirements for the prototype.

Table 5.1: Software Requirements

Software	**Description**
Operating System	Microsoft Windows XP or higher, Linux
Relational Database Management System	MySQL 5.0.45 or higher
Programming languages	Java, Python
Machine Learning toolkit	Weka, Rapid Miner or Orange biolab
Internet Browser	Google Chrome, Mozilla, Internet Explorer
Integrated Development Environment	Netbeans IDE, Eclipse
WampServer	Apache Server with MySQL database

5.2.2 Hardware Requirements

Table 5.2 describes the hardware requirements for the prototype.

Table 5.2: Hardware Requirements

Hardware	**Description**
Processor	Intel Centrino 1.6 Ghz Processor or higher or other equivalent processors
Memory	At least 512 MB ,Recommended: 1GB or more
Hard Disk Space	At least 50MB
Others	Internet access

5.3 Experimental Setup

This thesis adopted C4.5 classification algorithm to develop a model for evaluating Taxpayer Compliance. The classifier was trained and tested using Waikato environment for knowledge analysis (WEKA). WEKA is a collection of machine learning algorithms used for data mining tasks. It is open source software and contains tools for data pre-processing, regression, classification, clustering, and association rules. It also has visualization (Hall et al., 2009).

Taxpayer Data extracted from Taxpayer Database at Kenya Revenue Authority was used to form the experimental data set. Since tax data is confidential, the Personal Identification Numbers (PINs) were masked and taxpayer names deleted from the initial raw extract. The experiment scenario will be explained in detail in the following subsections;

5.3.1 Experimental Dataset

Two samples of taxpayers were used to form the experimental data set. The first purposeful sample of taxpayers included those that had been audited and penalties issued against them. The other purposeful sample was of taxpayers who were audited and no penalties were issued on them. While the former represented noncompliant cases, the later represented compliant cases. The year chosen was 2014/2015 as this was the period year when audits were almost being completed and records were readily available.

Based upon domain expertise of Kenya Revenue Authority Tax department, reviewed literature in Chapter two, as well as the feasibility to extract the parameter from the database, the following input variables were identified initially for the data mining exercise;

Dealer Profile (new registrant? (Y/N), nature of business? (Y/N), any other business operating from same Address? (Y/N), any other business having same Tel No?); **Return Compliance** (any non-submission)? (Y/N), delay in filing returns? No of returns that are NIL return, no of late submissions ?); **Returned Values & Ratios** (Tax : Turnover, amount declared/ amount paid, amount assessed/ amount of self-declared tax, expense amount/turnover, additional assessment/ self-assessment amount, profit change/total profit, frequency of disposal of assets; **Variations in the returns across tax-periods (**Turnover growth (compared to last year), Tax Growth (compared to last year), Variance of Turnover across tax-periods; **Benchmarking vis-à-vis dealers of same industry, in respect of following parameters** (Tax: Turnover, Gross profit %,)

Many of the variables although technically possible, could not practically be calculated from the database because of non-availability of data, data inconsistency, and high programming and computer resource requirement. At the time of extraction, many of them were required to be dropped because of practical considerations. The final table from the extract had 1116 records each with 51 independent attributes and a dependent attribute of Compliant or noncompliant. Out of these 506 were classified as compliant and the remaining 610 classified as non-compliant.

5.3.2 Attribute Selection

The table in excel was loaded into WEKA for the purposes of selecting significant attributes. Attribute selection feature in the WEKA software, searches through all possible combinations of attributes in the data and finds which subset of attributes works best for classification. The attributes used in this study was ranked in order of importance using information gain. Information gain evaluates the worth of an attribute by measuring the information gain with respect to the class. The search method used was Ranker and full training set was used as attribute selection mode. Out of a total of 51 attributes, 16 attributes were found to be significant from this pool and used for model building. The initial attributes and those selected for building the model are illustrated in **appendix D.**

5.3.3 Model Building

The process of building the model followed the following steps: Loading the data file through selecting the file loader component, specifying the class attribute using the class assigner component, selecting the training and testing mode, and finally attaching these components to C4.5 classifier (J48).The classifier performance evaluator was then assigned to the classifier and results viewed via the text viewer.

The output of the C4.5/J48 classifier was a comprehensible tree graphically represented in **appendix E**. In order to get the tree small as possible, information gain was used. In addition, pruning, a process of reducing the tree size was used to get a smaller tree, reduce the classifier complexity as well as improve on prediction accuracy.

5.3.4 Model Evaluation

In order to check the performance of the developed model, this thesis explored set performance evaluation functions such as;

5.3.4.1 Accuracy

The accuracy (AC) is defined as the proportion of the total number of predictions that were correct. TP represents the true positive, TN represents the true negative, FP represents the false positive and FN represents the false negative in the equations that were used to measure performance. See equation below;

Accuracy (AC) = (TP+TN) / (TP+TN+FP+FN)

5.3.4.2 Recall ratio

The recall or true positive rate (TP) is the proportion of positive cases that will be correctly identified as shown in the equation below;

Recall ratio = TP/ (TP+FP)

5.3.4.3 Precision

Precision (P) is the proportion of the predicted positive cases that will be correct, as computed in the equation below;

Precision (P) = FP/ (FN+FP)

5.3.4.4 F_Measure

The F-measure computes some average of the information retrieval precision and recall metrics.

The confusion matrix is given below;

Table 5.3: Confusion Matrix

Predicted → Actual ↓	0 (COMPLIANT)	1 (NONCOMPLIANT)
0(Tax complying)	TN(True Negative)	FP(False Positive)
1(Tax evading)	FN(False Negative)	TP(True Positive)

5.4 Functionalities Implemented.

The application prototype has been developed in java source code integrated with WEKA Libraries. The application also has a MYSQL database hosted locally using Wampserver and accessible via PHPMyAdmin interface. The front-end application is accessible by normal users whereas the back-end is accessible by the system administrator. The model and prototype source code is provided in **Appendix F**.

The functionalities implemented include user authentication, evaluate taxpayer, data management and user management. This are described below and screenshots provided in **Appendix G.**

5.4.1 User Authentication

Users log in into the application with their credentials after they have been registered by the administrator. This function authenticates users of the application and allows only authorized users to access the application prototype.

5.4.2 Evaluate Taxpayer

This function evaluates a taxpayer and provides a prediction of a taxpayer's compliance status. It is integrated with the J48 model built in section 5.2.3 that carries out the actual evaluation. The model returns a compliance status for each of the taxpayers selected for evaluation. Those with compliance status predicted as noncompliant are automatically selected for audit.

5.4.3 Data Management

Various data management tasks can be carried out using the functions given below;

Load Current Data: This function enabled the system administrator to load data into the system for classification

View Taxpayer Data: Users and System administrator can view taxpayer details before and after classification

View Training data: System Admin can view the data that was used to train the model.

5.4.4 User Management

For carrying out tasks of user management, the following functions are used;

Add User: System administrator can register a new user of the system using this function.

Edit User: System administrator can update the details of a registered user

Delete User: System administrator can delete a user from the system.

5.5 Prototype users

The potential users of the application are Compliance Officers and the System Administrator. The function they performed is described below and screenshot of each home page provided in **Appendix G.**

5.5.1 Compliance Officer

The functions performs the following functions; Login, Logout, Search Taxpayer, Evaluate Taxpayer, and View Report.

5.5.2 System Administrator

The functions performed by the System Administrator includes; Login, Logout, Register Users, Manage Users, Generate Reports, Search Taxpayer, and View Reports.

5.6 Prototype Testing

Prototype testing involved functional testing to ensure prototype conformed to user needs and usability testing for testing usability aspects.

5.6.1 Functional Testing

Prototype testing involved white box testing where each module was tested separately to ensure it functioned as expected. Thereafter, black box testing was conducted where the modules were brought together and tested on their integration. Thus, the User, Admin and Classifier modules were first tested separately and later integrated where they underwent integration testing.

5.6.2 Usability Testing

Usability testing evaluated the ease with which system users were able to achieve their system goals. Usability testing looked at the following attributes for the various system modules: consistency, efficiency, navigability, ease to learn and use, easy to find content, user interface, user-friendliness, predictability, usefulness and responsiveness

Out of the 15 respondents who participated in the application usability testing, 9 rated navigability Excellent and 6 rated Very Good. On “Easy to learn” attribute 10 rated Excellent, 3 Very Good and 2 rated Good. On “Easy to find core functionality” attribute 9 rated excellent, 4 rated Very Good and 2 rated Good. On “User friendly attribute” 9 rated Excellent, 4 rated very Good and 2 rated Fair. On “Responsiveness” attribute 11 rated excellent and 3 rated Good.

Finally, on “Useful and satisfying” attribute 10 rated Excellent, 2 rated Very Good and 3 rated Good.

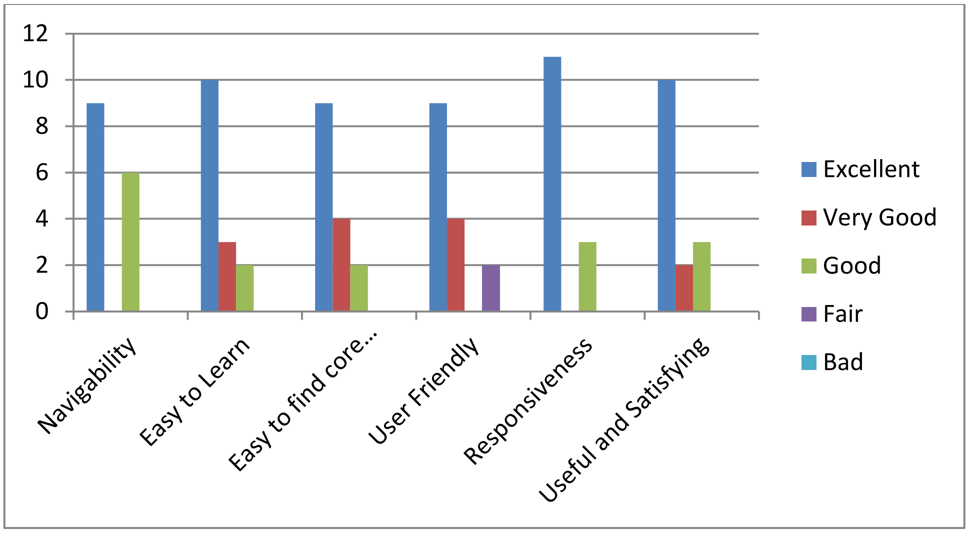


Figure 5. 1 Usability Testing

Respondents further recommended the application be implemented as a web application so that they can be able to access it from ubiquitously. These responses are in figure

5.7 Acceptance Testing

The main aim of this proposed solution was to select taxpayer audit subjects from a pool of taxpayers annually. Opinion from users was sought on whether the application prototype had the potential of successfully selecting audit cases. The Question was put forward on a Likert scale among the following choices: Strongly Agree, Agree, Do not Agree, Strongly Do not agree, Not Sure. Out of 30 respondents, 15 selected Strongly Agree, 10 selected Agree, 2 chose Do not Agree and 3 selected Not Sure as their response. Their response is given figure 5.9.

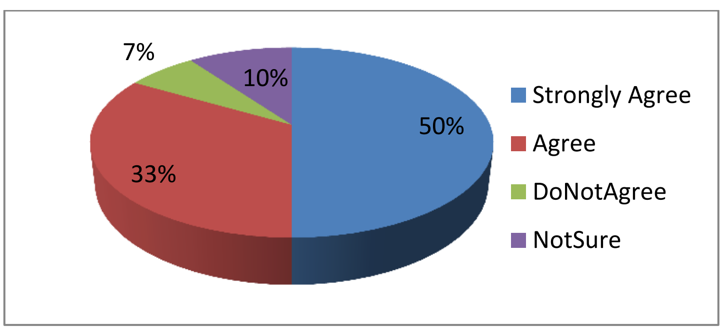


Figure 5.2 Acceptance Testing

5.8 Conclusion

This chapter has outlined the hardware and software requirements needed to build the model and develop the prototype. It has also discussed the data extraction, data pre-processing and ultimate building and evaluation of a decision tree classifier using this data. A detailed description of the development and testing of a prototype that makes use of the classifier has also been provided. Finally, usability and acceptance testing conducted on respondents indicate that the model and prototype as a whole is fit for purpose.

Chapter 6: Discussions

6.1 Introduction

The objectives of this thesis were four. The first objective was to analyze the determinants of corporate tax audit case-selection used by Kenya Revenue Authority; the second objective was to review existing tax audit case-selection methodologies used by Kenya Revenue Authority; the third was to develop a model for case-selection for Kenya Revenue Authority and the fourth was to validate the model using Taxpayer data in Kenya. This chapter analyzes the findings in relation to the research objectives and extent to what the findings agree with the literature review.

6.2 Determinants of Corporate Tax Audit Case-Selection used by Kenya Revenue Authority

Findings from this study points out various factors that are used by Kenya Revenue Authority to determine the likelihood of a taxpayer being audited. These factors include; nature of business, whether the business is a sole proprietorship, partnership or company, whether the company is a public commission , whether the company is involved with government projects, whether the business sector has an oversight body, whether the taxpayer engages a tax agent/auditor in preparing tax returns, whether the auditor/agent has been blacklisted by the licensing authority, the frequency of change of the tax agents/auditors ,return non-submissions, return late submissions, and return nil submissions.

Other factors include, difference in tax declared versus tax paid, whether the taxpayer has recorded declining profitability, declining liquidity ratio, whether the gross and net profit margins deviates from the industrial gross and net profits, whether the taxpayer has a disproportionate increase in taxable income in relation to turnover, whether there is a variation in profitability, taxpayers group structure, whether the taxpayer has branches, the frequency of investment deduction claim , the magnitude of investment deduction claim, taxpayer financial risk, variation in vat import declarations, temporary export, export diversion, variation in vat import declarations, variation in value of imports declared, variation in value of export declarations.

This is in harmony with literature reviewed in section 2.5 which highlighted factors that are likely to determine taxpayer tax compliance.

6.3 Tax Audit Case-Selection Methodologies used by Kenya Revenue Authority

Compliance officers are responsible for screening returns submitted by taxpayers and ear marking some for audits. Information from reliable third parties can warrant a taxpayer to be audited. A risk profiling framework is also used to profile taxpayers according to tax compliance risk levels and those with high scores selected for audit. This technique is a rather tiresome process and officers can only detect a few taxpayers to audit due to human limitation. Many of non-compliant taxpayers are yet to fall under the taxman's net.

6.4 Auditee Case-Selection Model for evaluating Corporate Tax Compliance in Kenya

A decision tree model was built using training data discussed in section 5.2.1 and incorporated into an application prototype developed in Java with a MYSQL database at the backend. Users were then able to select audit cases via the application prototype. The model is illustrated graphically in **Appendix E** and java code that integrates with the model provided in **appendix F**.

6.5 To Validate the Auditee Case-selection Model using Taxpayer data in Kenya

Literature review points out at accuracy, precision, recall ratio and confusion matrix in evaluating a model. The model was validated for accuracy, precision, recall ratio using the confusion matrix. Using percentage split of 66:33, 33% of the training data was used to test the model.725 out of 1116 instances presented to the network were correctly classified. This resulted to accuracy 65%.this is illustrated in Table 6.1

Table 6.1: Classification Output

Correctly classified instances	725	65%
Incorrectly classified instances	391	35%

The performance evaluation for the classification of tenders based on the precision, recall, F_measure rate is provided below as explained in section 3 is given below;

Table 6. 2: Detailed accuracy by class

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.677	0.677	0.383	0.68	0.679	compliant
0.617	0.323	0.613	0.617	0.615	Noncompliant

A confusion matrix that was obtained from the classification is illustrated in Table 6.3. The confusion matrix contains information on the actual and predicted classifications. There were a total of 1116 instances that were used to train and test the network. 610 instances of the compliant target were presented to the network. 413 instances were correctly classified as compliant while 197 were incorrectly classified as Noncompliant. There were a total of 506 noncompliant instances. 312 were correctly as noncompliant whereas 194 were incorrectly classified as Noncompliant.

Table 6. 3 Confusion matrix

		PREDICTED	
		Compliant	Noncompliant
ACTUAL	Compliant	413	197
	Noncompliant	194	312

Basing the evaluation on the above metrics of accuracy, precision, recall ratio and F-measure, the model is effective in identifying cases for audit.

Moreover, in testing the application prototype, users were satisfied with the functionality and usability of the application. As illustrated in section 5.6, 50 percent of the 30 respondents strongly agreed that the application had the potential of selecting cases for audit.

6.6 Contributions of the Model to Research

Considering the challenges compliance officers face in selecting cases for audit, the model offered an improved solution compared to the screening of tax returns by auditors along complex attributes. The model provided Kenya revenue authority with reliable results of compliant and non- compliant cases and would help reduce reliance on the knowledge of tax experts in identifying cases for audit. The model would also cover a broader scope and bring more taxpayers within the purview of the tax administration.

6.7 Limitations of the Study

The main limitation of this study is that it focused only on corporate tax compliance. Another limitation is on the small number of attributes used to build the model due to data unavailability.

Chapter 7: Conclusions and Recommendations

7.1 Introduction

This chapter will lay its focus on giving conclusions about the findings of the research in regards to research objectives. It will also discuss recommendations from the researcher and suggest areas for future research.

7.2 Conclusions

The objectives of this thesis were to analyze the determinants of corporate tax audit case-selection used by Kenya Revenue Authority, to review existing tax audit case-selection methodologies used by Kenya Revenue Authority, to develop a model for auditee case selection for Kenya Revenue Authority and to validate the model using Taxpayer data in Kenya.

In the endeavor to achieve the objectives, the literature reviewed pointed at six categories of factors that determine taxpayer compliance. These are economic, social, psychological, demographic, institutional, political and erratic factors. Discussions by Tax Experts identified fifty one attributes that are used to determine cases for audit which fall broadly among these categories. However, this research found out that other attributes have a higher significance than others in determining compliance behavior of taxpayers. The ranking of those attributes used in building the model are shown in **appendix H**.

The literature review, both theoretical and empirical, pointed out to the use of a data mining techniques for identifying cases for audit. However, Kenya Revenue Authority depended heavily on screening of tax returns by auditors and informers in order to identify audit subjects. A model to select cases for audit was therefore developed using decision tree algorithm and incorporated into an application prototype. Results showed that the model was well specified and had an overall percentage of prediction efficiency of 65%. If adopted the model will ease the work of compliance officers, and unearth those taxpayers who escape the purview of the tax administration.

7.3 Recommendations

It is recommended that Kenya Revenue Authority build up a more robust database of Taxpayer compliance behavior to enable future research in this area. The authority can also implement this model for evaluating all taxes administered by the Kenyan government.

7.4 Future Work

Future work would be in the areas of building models using different data mining techniques, comparing them and picking the one that has the most accuracy and prediction rates since this research adopted only one data mining technique. In addition, there is need to build models with larger samples , more tax units and include more information about taxpayers since in this research only used historical audit data , tax returns and registry information. There is also need to build compliance scores after prediction so as to prioritize the outputs more for better decision making on the part of tax experts.

References

Allingham, M. G. & A. Sandmo. (1972). Income tax evasion: A theoretical analysis. Journal of Public Economics, 1, 323-338.

Alm, J., Blackwell, C., & McKee, M. (2004). Audit selection and firm compliance with a broad-based sales tax. *National Tax Journal*, 209-227.

Ambler, S. (2005). The agile unified process (AUP)

Barbutamisu N. (2011).A Review of Factors for Tax Compliance Annals of Dunarea de Jos. University of Galati Fascicle I. Economics and Applied Informatics.

Batrancea, L., Nichita, R., & Batrancea, I. (2012) Understanding the determinants of tax compliance behavior as a prerequisite for increasing public levies. The USV Annals of Economics and Public Administration, 12(1), 15

Bordignon, M., Manasse, P., & Tabellini, G. (1997).Optimal regional redistribution under asymmetric information.

Breiman, L., J. Friedman, C. Stone J., and R. Olshen (1984). Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series.

Chauke, K. R., & Sebola, M. P. (2016). Revenue collection in South Africa: a comparative analysis of South African Revenue Services and the South African municipalities. *Journal of Public Administration*, *51*(3), 423-435.

Craign, Larman. (2004). Applying UML and Patterns.

Da Silva, Leon Sólon, et al.,(2016) "Bayesian networks on income tax audit selection-a case study of brazilian tax administration." *Bayesian Modeling Application Workshop (BMAW)*. Vol. 37

Emanuele, C. (2012).Economics of taxation: Normative and positive theories. Hauppauge, NY, USA: Nova Science Publishers, Inc.

Fjeldstad, O. H., Schulz-Herzenberg, C., & Hoem Sjursen, I. (2012). People's Views of Taxation in Africa: A Review of Research on Determinants of Tax Compliance.

Gibert, K., Sanchez-Marre, M., & Codina, V. (2010). Choosing the right data mining technique: classification of methods and intelligent recommendation. International Environmental Modelling and Software Society.

Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research.

Gupta, M., & Nagadevara, V. (2007). Audit selection strategy for improving tax compliance -Application of data mining techniques. In *Foundations of Risk-Based Audits. Proceedings of the eleventh International Conference on e-Governance, Hyderabad, India, December* (pp. 28–30).

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The Weka data mining software: An update. *Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) 11*(1), 10–18.

Hallsworth, M. (2014). The use of field experiments to increase tax compliance. *Oxford Review of Economic Policy*, *30*(4), 658-679.

Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, *148*, 14-31.

Han, J., M. Kamber, and J. Pei (2012). *Data Mining: Concepts and Techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hsu, K.-W., Pathak, N., Srivastava, J., Tschida, G., & Bjorklund, E. (2015). Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue. In M. Abou-Nasr, S. Lessmann, R. Stahlbock, & G. M. Weiss (Eds.), *Real World Data Mining Applications* (Vol. 17, pp. 221–245). Cham: Springer International Publishing.

James, S., & Alley, C. (2002). Tax compliance, self-assessment and tax administration.

Joseph, Sirengo. (2016). Assessing Taxpayer Compliance Risk Identification Criteria at the Kenya Revenue Authority

Kargupta, H., J. Han, P. S. Yu, R. Motwani, and V. Kumar (2008). *Next Generation of Data Mining* (1 ed.). Chapman & Hall/CRC.

Kirchler, E., E. Hoelzl & I. Wahl (2008) Enforced versus voluntary tax compliance: The"slippery slope" framework. *Journal of Economic Psychology*, 29, 210-225.

Kothari, C. R. (2009).Research Methodology: Methods and Techniques. New Delhi: New Age International Publishers.

KRA (2015).Sixth Corporate Plan. Kenya Revenue Authority

Kumar, Anuj and V. Nagadevara. (2006) Development of Hybrid Classification Methodology for Mining Skewed Data Sets – A Case Study of Indian Customs Data, Proceedings of the 4thACS/IEEE International Conference on Computer Systems and Applications, Mar8-10, 2006, Sharjah, UAE.

Kumar, E. P., & Mohan, B. (2016). Origin And Development of Auditing. *PARIPEX-Indian Journal of Research*, *4*(9).

Maimon, O. and L. Rokach (Eds.) (2010). Data Mining and Knowledge Discovery Handbook, 2nd ed. Springer.

Manasan Rosario G. (2003) Estimating Industry Benchmarks for the Value Added Tax. Philippine Journal of Development, 55, (30), 1, 71-90.

Moore, M. (2004). Revenues, state formation, and the quality of governance in developing countries. International Political Science Review, 25, 297-319.

Muralidharan, K., Das, J., Holla, A., & Mohpal, A. (2017). The fiscal cost of weak governance: Evidence from teacher absence in India. *Journal of Public Economics*, *145*, 116-135.

Musau, N. (2015). Determinants of tax compliance among Small and Medium Enterprises in Kenya: A case of Nairobi County.

Naibei, K. I., & Siringi, E. M. (2011). Impact of Electronic Tax registers on VAT compliance: A study of private Business firms. African Research Review, 5(1).

OECD (2004a), Compliance Risk Management: Managing and Improving Tax Compliance, Centre for Tax Policy and Administration

OECD (2004b), Compliance Risk Management: Audit case Selection Systems, Centre for Tax Policy and Administration

Oxford Business Group (2016). Kenya's recent tax changes and regulations retrieved from https://www.oxfordbusinessgroup.com/overview/duties-detail-look-tax-regulation-and-recent-changes

Persson, A. (2008).The institutional sources of statehood-Assimilation, multiculturalism and taxation in Sub-Saharan Africa Doctoral Thesis, Gothenburg University.

Pritchard, D. and N. Khan (2005), Offshore Compliance: Using Regression and Risk Modelling to Selected Cases from Large Datasets, Inland Revenue, United Kingdom

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning.

Shao H, Zhao H, Chang G. (2002) Applying Data Mining to Detect Fraud Behavior in Customs Declaration, Proceedings of the International Conference on Machine Learning and Cybernetics, Vol 3, IEEE 2002, 1241-1244

Torgler, B. (2003). Tax morale: Theory and empirical analysis of tax compliance

Torrey, A. (2008). The Discriminant Analysis Used by the IRS to Predict Profitable Individual Tax Return Audits. *Honors Projects in Mathematics*, 1.

Vellutini, C. (2011), Key principles of Risk-Based Audits, In Risk-Based Tax Audits: Approaches and Country Experiences (eds) M. S. Khwaja, R. Awasthi and J. Loeprick, The World Bank, Washington, DC.

Witten, I. H., E. Frank, and M. A. Hall (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc.

Zhu, J., & Hastie, T. (2001). Kernel logistic regression and the import vector machine. In *NIPS* (Vol. 14, pp. 1081-1088).

Appendix A: Interview Guide

Auditee Case-Selection Model for Evaluating Corporate Tax Compliance

i. What are the determinants for corporate tax audit case selection in KRA?
ii. What are the methods used to select taxpayers for audit in the KRA?
iii. Are these methods effective in selecting the Taxpayers for audit?
iv. What challenges are faced by compliance officers in identifying cases for audit?
v. Would use of a Case –Selection application make the work of identifying cases for audit easier for compliance officers?
vi. What features would you like to be present in the proposed application?

Your assistance will be highly appreciated

Appendix B: Interview Feedback

Auditee Case-Selection Model for Evaluating Corporate Tax Compliance

i. What are the determinants for corporate tax audit case selection in Kenya's tax system?

The audit case selection process is informed by various taxpayer factors given below;

Business Information Indicators

- Type of business
- Nature of business,
- No of years in Business
- No of businesses operated by same directors
- Regulatory framework

Return Compliance Indicators

- Non-Filers
- Late Submissions
- Under-reporters
- Non-payments

Audit Indicators

- Never audited
- Bad audit history
- High risk auditors

Returned Values & Ratios Indicators

- Tax : Turnover
- Profit Change
- Expense amount/turnover amount)
- Payments Amount / Tax Declarations amount
- Payments Amount / Tax Declarations amount

- Additional Assessment amount / self-Assessment amount
- Deduction Claim

Variations in the returns across tax-periods Indicators

ii. What are the methods used to select taxpayers for audit in the KRA?

- Information from reliable sources requesting a probe into a taxpayer
- Station committees profile taxpayers under them and come up with a list of those to be audited

iii. Are these methods effective in selecting the Taxpayers for audit?

- The technique is effective in identifying the audit subjects. However, it is a rather tiresome process and officers can only detect a few taxpayers to audit due to the human limitation. Many of non-compliant taxpayers are yet to fall under the taxman’s net.

iv. What challenges are faced by compliance officers in identifying cases for audit?

- The huge customer base of taxpayers is overwhelming for officers to go screen each and every case.
- The current process is heavily dependent on the experience of compliance officers. A new officer may not be able to effectively identify audit cases as one who had gathered the knowledge and expertise over the years.

v. Would use of a Case –Selection application make the work of identifying cases for audit easier for compliance officers?

- Yes. That will really ease our work were it to be done by a computer.

vi. What features would you like to be present in the proposed application?

- A facility to automatically flag out the taxpayers suspected to be non-compliant, prioritized in order of most risky to least risky
- A facility to select one taxpayer and obtain their compliance status
- A facility to view taxpayer characteristics and their compliance statuses

- A facility to update taxpayer details as informed by reliable sources
- A facility for targeted profiling e.g. non –filers only or defaulters only
- A facility to generate taxpayer reports
- A facility to print reports
- A facility to check the analysis done by the system on a taxpayer and why the taxpayer has been classified as non-compliant or otherwise

Your assistance will be highly appreciated

Appendix C: Usability Questionnaire

Auditee Case-Selection Application Prototype Usability Testing

* Required

1. **Where you able to Log in?** *

   Mark only one oval.

   ⬭ Yes

   ⬭ No

2. **If your answer to the above is 'No', kindly list the problems encountered**

   ____

How would you rate the whole application prototype? Please mark each oval appropriately

3. **Ease of Use** *

   Mark only one oval.

	1	2	3	4	5	
Excellent	⬭	⬭	⬭	⬭	⬭	Poor

4. **Navigability** *

   Mark only one oval.

	1	2	3	4	5	
Excellent	⬭	⬭	⬭	⬭	⬭	Poor

5. **Responsiveness** *

   Mark only one oval.

	1	2	3	4	5	
Excellent	⬭	⬭	⬭	⬭	⬭	Poor

6. **Usefulness** *

   Mark only one oval.

	1	2	3	4	5	
Excellent	⬭	⬭	⬭	⬭	⬭	Poor

7. **User Friendly** *

   Mark only one oval.

	1	2	3	4	5	
Excellent	⬭	⬭	⬭	⬭	⬭	Poor

8. **Which feature/s did you like most in the application prototype?** *

9. **Any other comments about the application prototype**

Acceptability Tests

10. **Do you think the Application Prototype have the potential of successfully selecting cases for audit?** *
Mark only one oval.

	1	2	3	4	5	
Strongly Agree	⬭	⬭	⬭	⬭	⬭	Not Sure

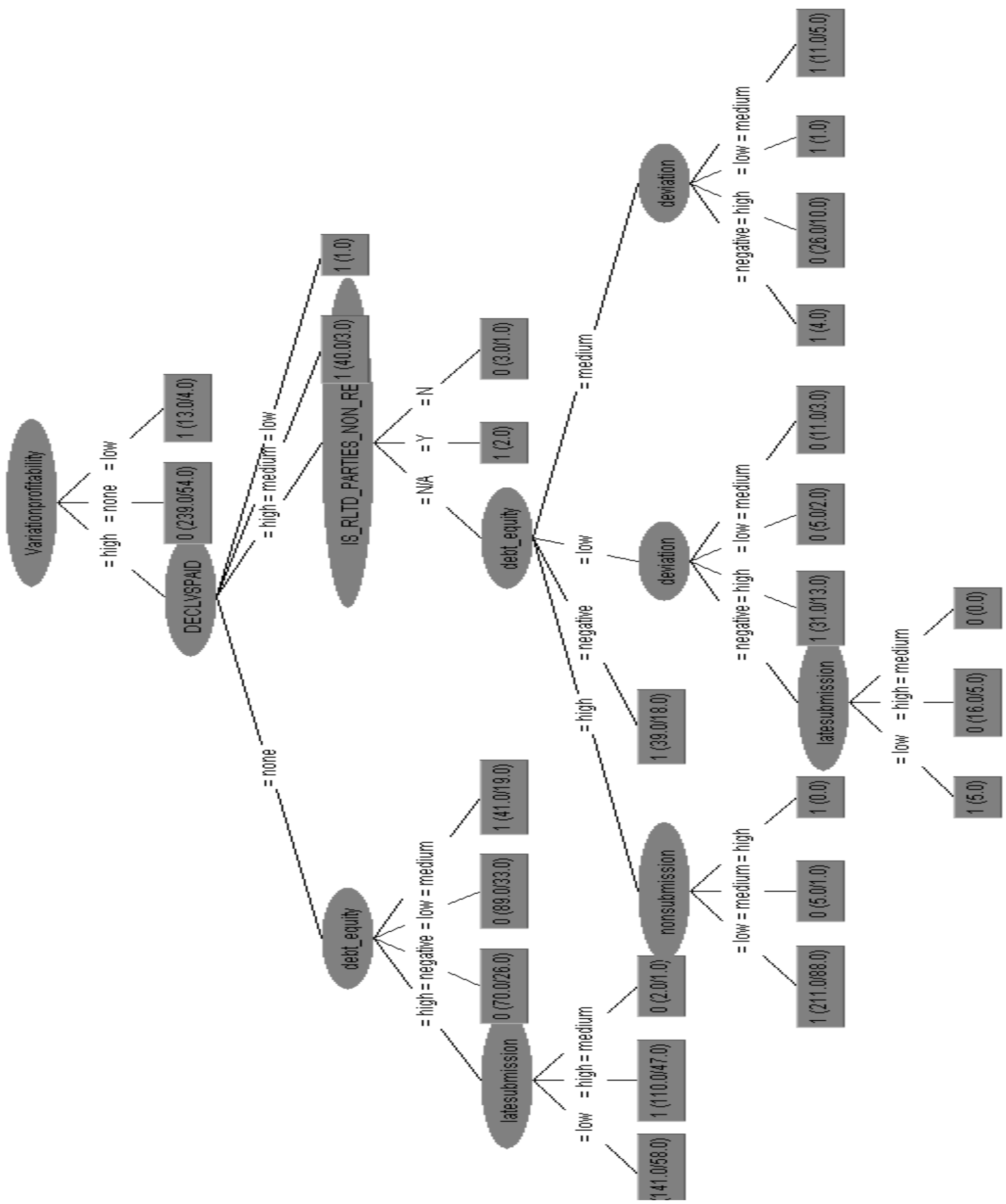
11. **Does the Application Prototype have all your useful features?** *
Mark only one oval.

	1	2	3	4	
Strongly Agree	⬭	⬭	⬭	⬭	Strongly Disagree

Appendix D: Attributes

Attributes	Attribute Usage in Model
PUBLIC COMMISSIONS	N
GOVERNMENT PROJECT	N
BUSINESS TYPE	Y
INDIVIDUAL - INCOME SOURCE	N
EMPLOYEES OF DIPLOMATIC ORGANIZATIONS	N
REGULATORY FRAMEWORK	Y
AUDITORS/ TAX AGENTS	Y
TAX AUDITORS WHO HAVE BEEN BLACKLISTED BY THE LICENSING AUTHORITY	Y
RISK GENERATED FROM BAD AUDITORS	N
DEVIATION OF TAXPAYER GROSS PROFIT MARGIN FROM THE INDUSTRY	Y
NATURE OF BUSINESS	Y
LATE RETURN SUBMISSION	Y
NIL FILER RETURN	Y
NON SUBMISSION OF RETURN	Y
DIFF IN TOTAL TAX DUE VS PAID	Y
ADDITIONAL ASSESSMENT ISSUED (INCOME)	N
ADDITIONAL ASSESSMENT ISSUED (TAX)	N
DISPROPPORTIONATE INCREMENT IN TAXABLE INCOME IN RELATION TO TURNOVER	Y
VARIATION IN PROFITABILITY	Y
DECLINING LIQUIDITY RATIO	Y
FREQUENCY OF DISPOSAL OF FIXED ASSETS	N
DEVIATION OF NET PROFIT MARGIN FROM IND. STD	Y
DEVIATION OF GROSS PROFIT MARGIN FROM IND. STD	N
TAXPAYERS GROUP STRUCTURE	N
TAXPAYERS WITH BRANCHES	N
FREQUENCY OF INVESTMENT DEDUCTION CLAIM IN FIVE YEARS	Y
MAGNITUDE OF INVESTMENT DEDUCTION CLAIM	Y
DISPROPPORTIONATE INCREMENT IN TURNOVER AGAINST FIXED ASSETS	N
FREQUENCY IN TAX AUDITOR CHANGE	N
FINANCING RISK(LEVERAGE)	N
VARIATION IN VAT IMPORT DECLARATIONS	N
VARIATION IN VALUE OF IMPORTS DECLARED	N
TEMPORARY EXPORT	N
EXPORT DIVERSION	N
VARIATION IN VALUE OF EXPORT DECLARATIONS	N

Appendix E: Decision Tree Model



Appendix F: Source Code

```
package auditee;
import java.awt.BorderLayout;
import java.io.BufferedReader;
import weka.classifiers.trees.J48;
import weka.core.Instances;
import weka.experiment.InstanceQuery;
import java.io.File;
import java.io.FileReader;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.PreparedStatement;
import java.sql.ResultSet;
import java.sql.SQLException;
import java.util.ArrayList;
import java.util.List;
import java.util.logging.Level;
import java.util.logging.Logger;
import javax.swing.JOptionPane;
import javax.swing.table.DefaultTableModel;
import net.proteanit.sql.DbUtils;
import weka.classifiers.Evaluation;
import weka.classifiers.trees.J48;
import weka.core.Attribute;
import weka.core.Instance;
import weka.core.Instances;
import weka.core.converters.ArffSaver;
import weka.core.converters.DatabaseSaver;
import weka.gui.treevisualizer.PlaceNode2;
import weka.gui.treevisualizer.TreeVisualizer;


 * @author caroline
 */
public class Classification {
    J48 tree;
    J48 model;
    Instances trainingdata;
    Instances testdata;
    InstanceQuery query;
    InstanceQuery query1;
    Attribute att;
   public Classification()
   {
       this.tree=null;
       this.model=null;
       this.trainingdata=null;
       this.query=null;
       this.query1=null;
       this.att=null;
       this.testdata=null;
   }
    public void buildtree() throws Exception
    {
     String [] options =new String[1];
     options[0]="-U";
     tree= new J48();
     tree.setOptions(options);
     tree.buildClassifier(trainingdata);
    }
```

```java
 public void settraining(String trainingset) throws Exception
 {
  query=new InstanceQuery();
  query.setUsername("root");
  query.setPassword("");
  query.setQuery(trainingset);
  trainingdata = query.retrieveInstances();
  trainingdata.setClassIndex(trainingdata.numAttributes() - 1);
 }


public void setpredictionSet(String predictionset) throws Exception
{
   query1= new InstanceQuery();
   query1.setUsername("root");
   query1.setPassword("");
   query1.setQuery(predictionset);
   testdata= query1.retrieveInstances();
    String classlabel=trainingdata.classAttribute().name();
    List classvalues=new ArrayList();
    for (int i=0;i<trainingdata.numClasses(); i++)
       {
       String classvalue=trainingdata.classAttribute().value(i);
       classvalues.add(classvalue);
       //System.out.println("Class Value is" + i + classvalue);
       }
    att = new Attribute(classlabel, classvalues);
   testdata.insertAttributeAt(att, testdata.numAttributes());
   testdata.setClassIndex(testdata.numAttributes() - 1);

}

public void Classify(J48 model) throws Exception
{
   for (int i = 0; i < testdata.numInstances(); i++)
{

  double clsLabel = model.classifyInstance(testdata.instance(i));
  testdata.instance(i).setClassValue(clsLabel);
}
```

```java
public Instances getpredicted()
{
return testdata;
}

public Instances gettraining()
{
return trainingdata;
}

J48 loadmodel(String url) throws Exception
{

model=(J48)weka.core.SerializationHelper.read(url);
return model;
}

public void savemodel(String url) throws Exception
{
weka.core.SerializationHelper.write(url,tree);
}
```

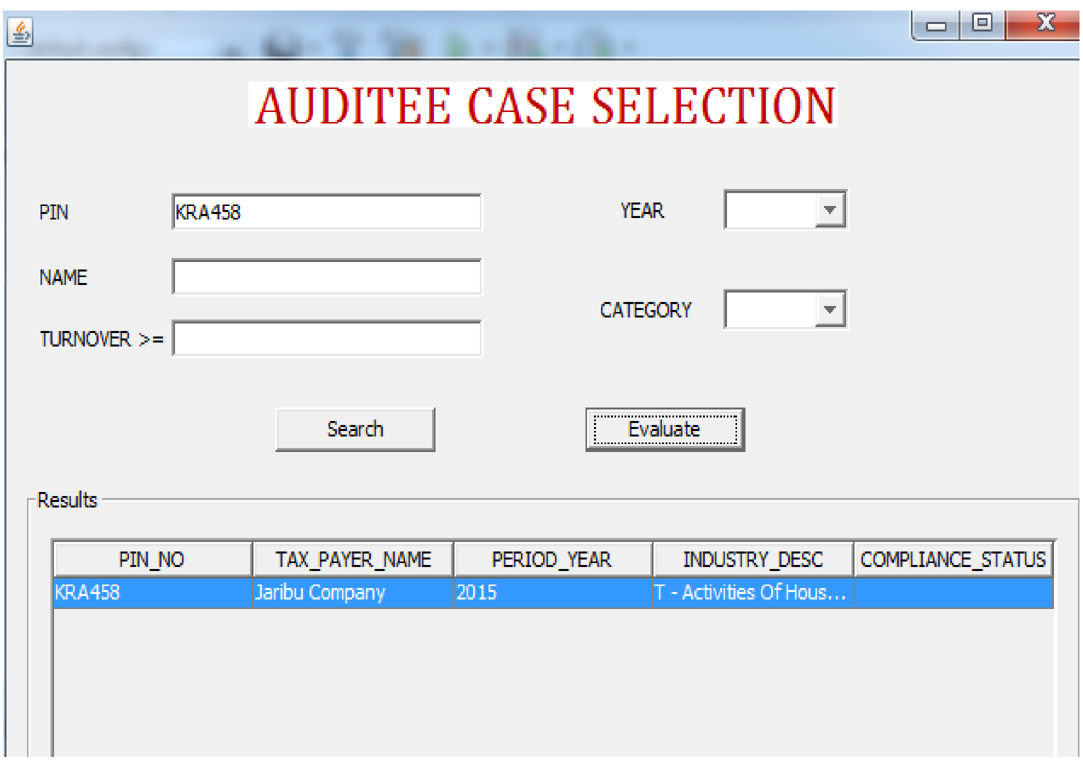
```java
public static void main(String arg[] ) throws Exception
   {
  Classification clas = new Classification();

  J48 tree= new J48();

  String url="select * from training2014";
  String url1="select * from trial3 ";
  String location="D:wekafiles/model1.model";
  clas.settraining(url);
  clas.buildtree();
  clas.savemodel(location);

BufferedReader reader = new BufferedReader(new FileReader("D:Test Files/training2014.arff"));

 Instances data = new Instances(reader);
 reader.close();
 data.setClassIndex(data.numAttributes() - 1);

 tree=clas.loadmodel(location);

 for (int i = 0; i < data.numInstances(); i++)
 {

   double clsLabel =tree.classifyInstance(data.instance(i));
   data.instance(i).setClassValue(clsLabel);
 }
   System.out.println(data);
```

```
 Evaluation eval=new Evaluation(data);
  eval.evaluateModel(tree,data);
  System.out.println(eval.toSummaryString("Evaluation Results\n",false));
  TreeVisualizer tv=new TreeVisualizer(null,tree.graph(), new PlaceNode2());
final javax.swing.JFrame jf= new javax.swing.JFrame("Classifier Visualization");
jf.setSize(1000,1000);
jf.getContentPane().setLayout(new BorderLayout());
jf.getContentPane().add(tv,BorderLayout.CENTER);
jf.setVisible(true);
tv.fitToScreen();
}
}
```

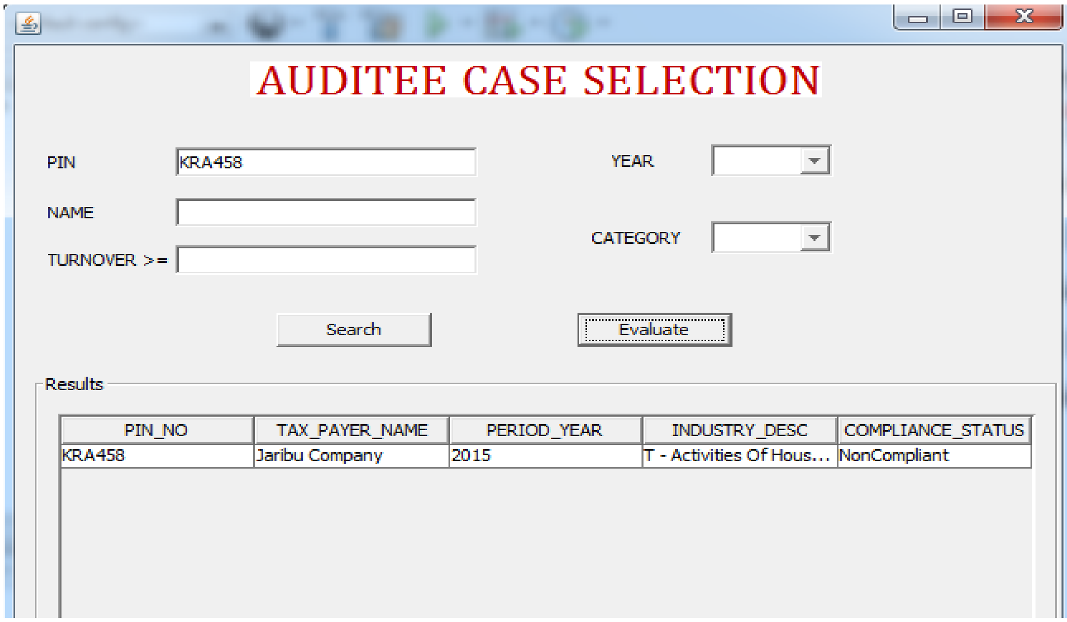
Appendix G: Screenshots

Login Page

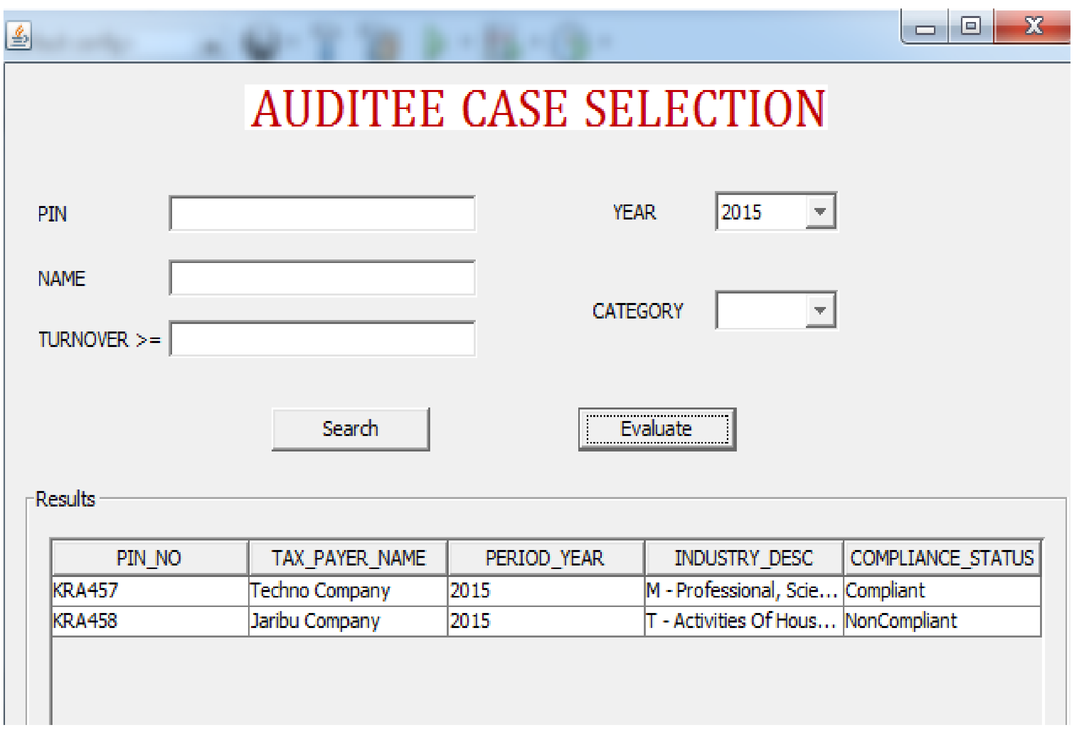
Search for Taxpayer to Evaluate



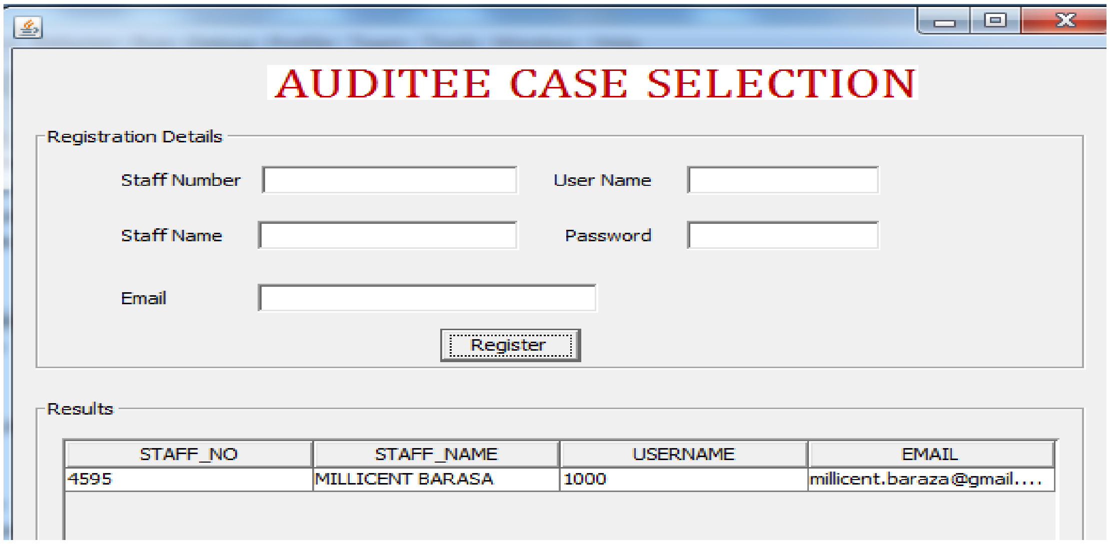
Results after clicking “Evaluate” button-Taxpayer Compliance status displayed



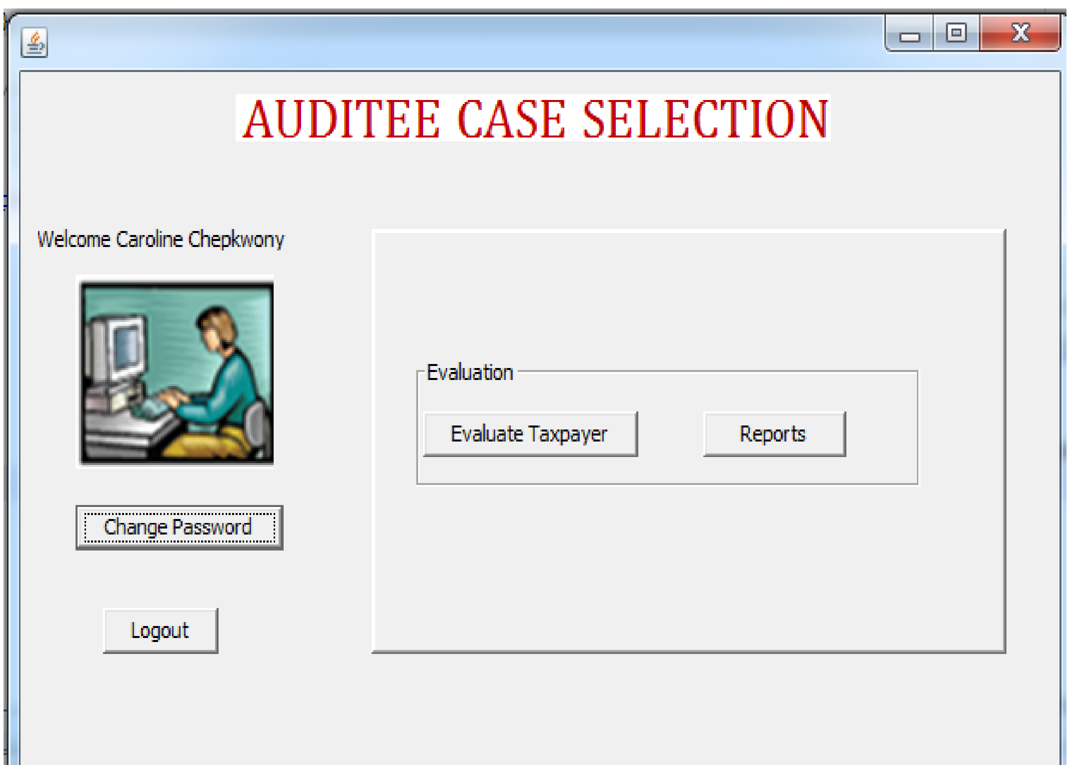
Evaluating a group of Taxpayers e.g. 2015



Adding a new user to the Application



Compliance Officer Home Page



Administrator homepage

Appendix H: Ranking of Attributes

```
Ranked attributes:
 0.0448213   8 DECLVSPAID
 0.0444344  10 Variationprofitability
 0.04331     3 ACTIVITY_CLASS_CODE
 0.0163061  13 debt_equity
 0.0084876  14 latesubmission
 0.0054545   2 INDUSTRY_TYPE_CODE
 0.0039242  11 LR_DECLINE
 0.0025668   9 deviation
 0.0020797  15 nonsubmission
 0.0015637  16 nilsubmission
 0.001124    1 BUSINESS_SUBTYPE_CODE
 0.0010232   6 OBLIGATION_STATUS
 0.0000684   5 IS_RLTD_PARTIES_NON_RESIDENT
 0.0000684   4 IS_MEMBER_OF_GRP
 0.0000183   7 AGENT_PIN
 0.0000113  12 idc_taxincome
```

Appendix I: Originality Report