Predicting Risky Taxpayers in Kenya Using Machine Learning

Cynthia Jemutai Cheboi

Adm. No. 151145

Supervisor

Dr. Rogers Ochenge

Submitted in Partial Fulfilment of the Requirements of the Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences and iLab Africa

Strathmore university

Nairobi, Kenya

April 2024

DECLARATION AND APPROVAL

Declaration

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

Student Name: Cynthia Jemutai Cheboi

Admission Number: Adm No. 151145

Student Signature: _____ Date: 05/04/2024

Approval

The dissertation of Cynthia Jemutai Cheboi has been reviewed and approved by Dr. Rogers Ochenge.

Supervisor Signature: _____ Date: 05/04/2024

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Rogers Ochenge for their invaluable guidance, unwavering support, and encouragement throughout the entirety of this dissertation journey. Their expertise, patience, and constructive feedback have been instrumental in shaping this research and pushing me to strive for excellence.

I am also profoundly thankful to my family for their unconditional love, understanding, and unwavering belief in me. Their constant encouragement and sacrifices have been the cornerstone of my academic pursuits, and I am forever indebted to them for their enduring support.

To my classmates and friends, I extend my heartfelt appreciation for their camaraderie, intellectual discussions, and moral support. Their diverse perspectives and shared experiences have enriched my academic endeavors and made this journey more rewarding.

ABSTRACT

Taxation is a fundamental tool for governments to raise revenue and fulfill their responsibilities to society. It is an essential component of modern governance, facilitating economic development, social welfare, and the provision of goods and services for the benefit of the public. Conversely, tax evasion poses a pervasive challenge impacting both advanced and emerging economies globally. In Kenya, addressing tax evasion is a significant hurdle, with the government estimating substantial annual revenue losses as a result leaving the government to seek debt financing for its programs. This study used machine learning models to classify taxpayers according to certain attributes and predict those who are most likely to evade. The study explored 24 such attributes. The target output variable was the payment time. The dataset was trained using six supervised machine learning algorithms including the Decision Tree, Logistic Regression, Random Forest, XGBoost, Support Vector Machines and Stacking. Among the trained models, the Random Forest classifier exhibited the optimal performance with a precision score of 90% and recall score of 86%. This suggests that the model can effectively predict the risky taxpayers to be subjected to a tax audit with likelihood of high returns. The study identified the top five crucial features influencing optimal tax evasion prediction as installment tax paid, total liabilities, credit brought forward, withholding value added tax credit and total expenses. Accordingly, adjusting these parameters within specified ranges is anticipated to result in an increased accuracy of the prediction of taxpayer classes. These results offer valuable insights for understanding determinants of tax compliance and enhancing the accuracy of predicting risky taxpayers towards optimizing resource allocation for better tax revenue mobilization outcomes.

Keywords: Taxation, Tax evasion, Risky taxpayers, Tax Audit, Machine Learning

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ANN	Artificial Neural Networks
API	Application Programming Interface
AUC-ROC	Area Under the Receiver Operating Characteristics
CIT	Corporate Income Tax
CRISP-DM	Cross-Industry Standard Process for Data Mining
EU	European Union
GBM	Gradient Boosting Machine
GDP	Gross Domestic Product
IT	Income Tax
IRS	Internal Revenue Service
KRA	Kenya Revenue Authority
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbor
Kshs	Kenya Shillings
ML	Machine Learning
MLP	MultiLayer Perceptron
OECD	Organization for Economic Cooperation and Development
PAYE	Pay As You Earn
NN	Neural Network
RF	Random Forest
RRP	Return Review Program

SDGs	Sustainable Development Goals
SVM	Support Vector Machines
VAT	Value Added Tax
XGBoost	eXtreme Gradient Boosting

DEFINITION OF TERMS

Taxation - This refers to the government's act of imposing taxes on individuals, businesses, or organizations to finance public services, pay off debt, and actively support the accomplishment of a nation's development goals (Murorunkwere et al., 2022).

Tax evasion – This entails a taxpayer deliberately providing inaccurate or incomplete information to tax agencies in an effort to lower their tax liability (Abedin et al., 2021).

Risk management – This comprises the ongoing steps of identifying, analyzing, assessing, prioritizing, searching for solutions, and assessing models related to risks (Ruzgas et al., 2023).

Tax Audit - This involves scrutinizing the taxpayer's financial records and organizational procedures to evaluate adherence to tax regulations (Chalye, 2020).

Machine Learning - This is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention (Chalye, 2020).

CHAPTER ONE

INTRODUCTION

1.1 Background

The World Bank and other international lenders have taken a keen interest in countries' ability to optimize the mobilization of financial resources from their own internal sources. According to World Bank statistics, roughly 40% of global firms fulfill their tax obligations, while 60% fail to pay their taxes (World Bank Group, 2020). It is noted that these unpaid amounts may not be recuperated in the coming tax years. The data also highlights a global trend of increasing rates of tax default. Taxation continues to be the most consistent and dependable means of generating state revenues to support the financing of the Sustainable Development Goals (SDGs) (UNDP, 2022). It is a crucial component of public finance and is used to redistribute income and control economic activity.

Tax evasion, however, continues to be a substantial barrier to reaching optimal tax collection, leading to severe repercussions for individuals, businesses, and society overall (Kifordu, 2021). Individuals and businesses whose income is subject to taxation commonly use tax avoidance strategies to elude or lower taxes, despite the government's persistent stance against tax avoidance and tax evasion (Mughal and Akram, 2012).

Murphy notes that the tax shortfall within the European Union (EU), primarily stemming from domestic tax evasion, could amount to €825 billion annually, as per 2015 data (Murphy, 2019). A research conducted by Schneider (2015) cites that the cumulative tax losses arising from shadow economy activities across all 28 EU member states amounted to 450.8 billion euros in 2011, equivalent to 3.6% of the EU-28-GDP. The amount increased to 457.3 billion euros in 2012, or 3.5% of the GDP of the EU-28, and to 454.2 billion euros in 2013, or 3.4% of the GDP. The term "shadow economy" denotes economic activities that escape taxation and regulation, frequently involving instances of tax evasion.

Tax audits play a vital role in enforcing tax regulations globally. They continue to serve as a primary tool in the effort to address tax fraud and evasion (Baghdasaryan et al., 2022). Increasing the frequency of audits can contribute to a reduction in tax avoidance (Ruzgas et al., 2023). Current methodologies often utilize traditional statistical techniques instead of

employing advanced data analytic approaches, such as cutting-edge Machine Learning (ML) methods (Abedin et al., 2021). ML can predict actions by taxpayers that violate tax laws, including criminal activities against the tax system. Therefore, leveraging ML enhances the effectiveness of fiscal audit strategies, contributing to the Government's success (Lozano et al., 2021).

The Organization for Economic Cooperation and Development (OECD) reports that in 2019, 38% of the member nations integrated artificial intelligence or ML in their tax administration, while an additional 34% were in the development stage. These technologies were used for both enforcement and taxpayer services (OECD, 2021). The Internal Revenue Service (IRS) has already made progress in this area with the introduction of the Return Review Program (RRP) in 2017 (Holtzblatt and Engler, 2022). This program combines conventional methods with ML to detect suspicious refunds on individual income tax returns.

All facets of operations – taxpayer service, enforcement actions, and different internal processes can be continuously improved by utilizing analytics. As a result, we will be able to optimize our data and test learning (IRS, 2018). The IRS considers updating its technology infrastructure as a critical component of its long-term strategy to increase agency productivity (Holtzblatt and Engler, 2022).

The task of addressing tax evasion is complicated by the fact that tax authorities have restricted resources at their disposal and rely on conventional tax auditing methods that are both laborious and time-consuming (Wu et al., 2012). Currently audits done by the Kenya Revenue Authority (KRA) are highly dependent local knowledge (Chepkwony, 2017). This underscores the need for the adoption of advanced and robust techniques in tax fraud detection (Murorunkwere et al., 2022).

There is potential for establishing a mechanism to categorize taxpayers based on their behaviour. This classification could assist in determining specific areas for audit focus so as to enhance revenue collection. Additionally, there is currently a move towards digitization in KRA which will result in incredible amounts of data collected from various platforms. This will spur the explosive growth in the use of analytics for decision making (Kasibwa and Allela, 2023).

1.1.2 Tax Evasion in Kenya

In numerous developing nations, the percentage of tax collection in relation to their Gross Domestic Product (GDP) sits between 15-20%, significantly falling short of the necessary levels to support essential services (UNDP, 2022). As per the OECD, Kenya's tax-to-GDP ratio in 2021 stood at 15.2%, which was marginally below the average of 15.6% for the 33 African countries, with a difference of 0.4 percentage points (OECD, 2023). Notably, this has been declining over the years from a high of 22% in the Financial Year 2015/16 to a low of 15.2% in 2021. The tax-to-GDP ratio serves as a crucial indicator for comparing tax systems globally, representing the proportion of tax revenue a country generates relative to the size of its economy, as measured by the GDP. Global patterns indicate a correlation between the amount of taxes collected and the development of a nation.

The expectation is that annual revenue targets are achieved consistently. Nonetheless, despite coordinated efforts, KRA has faced difficulties in meeting the revenue targets set by National Treasury. Various factors, including widespread tax evasion and revenue leakage, have impeded the tax authority's objective of revenue collection. There is an ongoing annual loss of billions in Kenya due to tax evasion schemes (Mburu, 2022). In the financial year 2021/2022, 1,309 individuals and companies, estimating a tax loss of around Kenya Shillings (Kshs) 259 billion were identified, examined and slated for additional investigations, and legal action for non-compliance.

1.2 Problem Statement

While the importance of fulfilling income tax obligations is widely recognized, taxpayers have been devising innovative, inherently challenging methods of tax evasion (Murorunkwere et al., 2022). This escalating issue results in substantial depletion in resources that could be directed toward crucial areas such as addressing climate concerns or improving healthcare and education quality (UNDP, 2022).

Tax compliance entails paying the correct amount of tax when it is due. Every month, there are taxpayers who pay their taxes late or fail to pay, resulting in a revenue loss of Ksh. 20 billion annually. Knowledge of these taxpayers is required to facilitate preventive measures, which include enforcement actions, tax audits, taxpayer education, stakeholder engagement, and reminders.

Currently, taxpayers are monitored manually, and compliance officers address the issue after the taxpayer has failed to pay. This is because there is no means of early warning before the event crystallizes. On average, a staff monitors 1,000 taxpayers, whereas the capacity for sending reminders, for instance, is for 50 taxpayers based on the perceived levels of non-compliance. Without a prediction model, determining the 50 taxpayers to engage with is a challenge.

Additionally, tax authorities have leveraged on the experience of compliance officers regarding risky cases. A new staff member thus has no historical advantage. A ML model that can identify high-risk taxpayers in real-time and minimize the authority's cost impact is crucial. Analytics shows promise as a means of improving tax administration efficiency, potentially streamlining over 50% of the activities associated with taxpayer supervision (Pijnenburg et al., 2017).

When officers manually choose cases to follow up based on their expertise in taxpayer behaviour, they may miss certain patterns of non-compliance embedded in historical data. Data-driven case selection operates on objective criteria, eliminating the need for subjective judgment from tax officials. This approach employs computational methods to extract valuable insights from taxpayer behaviour data (FiscalisRiskAnalysisProjectGroup, 2016).

The detection of tax fraud through ML approaches has emerged as a highly active area of interest in the research community (Murorunkwere et al., 2023). Mamo (2013) conducted a study focusing on detecting fraud using data mining, specifically within the context of customs in Ethiopia. The study revealed the value of data mining techniques in identifying tax fraud. In contrast, the current study delves into domestic taxes.

Chepkwony (2017) sought to develop a model to be used by auditors in selecting cases to evaluate taxpayers' compliance. The focus of the study was on corporate tax. It was concluded that the model is both effective and suitable for selecting cases. The current study focuses on VAT payment default. This research endeavors to contribute to combating tax evasion by predicting risky taxpayers using ML.

1.3 Research Objectives

The primary goal of this research was to predict risky taxpayers in real-time using a ML model thereby enhancing the audit case selection process. Specifically, the following objectives are explored;

1. To identify features that are important for predicting tax evasion.
2. To apply, test and evaluate a ML model for case selection in Kenya.
3. To deploy the ML model using a web application to predict taxpayer risks for case selection.

1.4 Research Questions

These research questions were the focus of the study:

1. What are the features that are important for predicting tax evasion?
2. How can a ML model be applied, tested and evaluated for case selection in Kenya?
3. How can a web application be developed to utilize the ML model to predict taxpayer risks for case selection?

1.5 Justification

Tax evasion will inevitably happen, and the tax administration must be prepared to handle it when it does. The implementation of the classification model will lead to a lot of benefits. These include early detection whereby officers will be able to identify risky taxpayers in their early stages. This allows the tax authority to implement mitigation strategies before tax default occurs. In addition, there is reduction of operating costs through simplification of the process of identifying high-risk taxpayers. Utilizing an automatic risk prediction system can alleviate the workload of tax administrators and improve overall administrative efficiency. The ML model will facilitate tailored treatment whereby suitable treatment will be rendered to each taxpayer according to their behaviour. Ultimately, these efforts will go a long way in fostering voluntary compliance and enhancing taxpayer satisfaction. Taxpayers will feel assured that instances of egregious non-compliance will be addressed. Notably, tax administrations have limited resources and therefore cannot address all compliance risks. The model will thus aid in optimizing revenue collection by enabling the focusing of resources on the highest risk to the tax base.

1.6 Scope of the Study

This research used taxpayer history data extracted from KRA databases. This included taxpayer demographic details, Income Tax declarations, VAT declarations and Pay As You Earn

(PAYE) declarations for a period of five years from 2018 to 2022 covering VAT registered taxpayers. The dataset was split into two subsets: the first 70% was used to train ML models, and the remaining 30% was allocated for evaluating the models' performance. The dissertation focused on feature selection to identify features that are most appropriate in distinguishing between payment defaulters and non-defaulters. Thereafter, different ML models used in classification of taxpayer behaviour were explored. Finally, the selected models were applied to the testing dataset where different metrics were used to evaluate their performance.

CHAPTER TWO

LITERATURE REVIEW

The structure of this chapter is as follows: Theoretical literature pertaining to taxpayer compliance behaviour is outlined in Section 2.1. Section 2.2 considers features that have been used to classify taxpayers. Section 2.3 delves into the efforts of past researchers who utilized ML models for predicting tax evasion. Ultimately, Section 2.4 provides a brief overview of the literature.

2.1 Theoretical Framework

Understanding taxpayer behaviour and employing strategies to influence it are crucial for achieving optimal levels of compliance (OECD, 2004b). The BISEP model in Figure 2.1 enables tax administrations to understand the factors that impact different compliance behaviours and assists in determining suitable interventions to undertake.

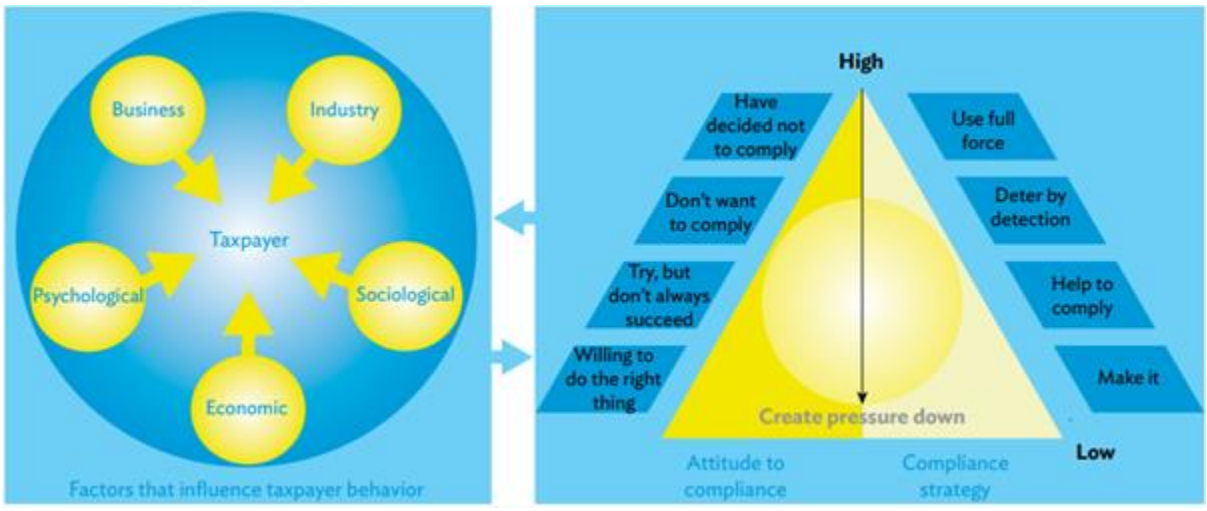


Figure 2.1: Compliance Model
Source: OECD 2004, Compliance Risk Management Guidance Note

These factors include business which considers the type, size, organizational structure of the business (whether it is a sole proprietorship, partnership or company) as well as location, whether it operates locally, nationally or internationally. Industry-specific factors include competitive dynamics, seasonal variations, profit margins, regulatory intensity, and infrastructure. Sociological factors delve into group norms, demographics such as age, gender, ethnicity, education level, and personal relationships. Economic factors focus on aspects related

to inflation, interest rates, tax system/rates and government policies. Lastly, psychological factors encompass taxpayer attitudes towards risk, trust, fairness perceptions, and their compliance history (Morris and Lonsdale, 2005).

Taxpayers are expected to comply with tax obligations, but not all do so. Taxpayer behaviour impacts compliance levels. Grouping seeks to differentiate and characterize groups of taxpayers, discern the specific behavioural traits of each group, and implement the most effective targeted tax enforcement measures tailored to each particular tax group (Ruzgas et al., 2023). Due to resource constraints, tax administrations cannot address all compliance risks.

There are four taxpayer compliance attitudes and behaviours. Willing compliers are committed to meeting obligations and self-regulate compliance. (Shipton, 2015). The strategy for the tax administration in this case would be to make it easy for the taxpayers to comply through utilization of relationship management and stakeholder engagement. Those attempting compliance but facing challenges require assistance. The strategy for this group would be to help them to comply. Treatments would include reminders of upcoming filing and payment deadlines, pre-population of returns/ simplified tax processes, taxpayer education and amnesty programs.

Active non-compliers resist the self-regulatory system and try to avoid meeting their compliance obligations (Shipton, 2015). The strategy in this case would be to deter through detection by utilizing taxpayer education, penalties, compliance reviews, and tax audits. Finally, those deciding not to comply no longer want to participate in the system and will not take any steps to change this situation. The tax administration should use the full force of the law in such cases through sanctions, prosecutions, audits and investigations. Consequently, these strategies encourage taxpayers to shift toward the lower levels of the compliance pyramid.

It is crucial to strengthen and broaden the capacities of the tax administration by employing advanced scientific, econometric, and statistical methodologies for research. This enhancement will enable the modeling of taxpayers' behaviour and the evaluation of the effectiveness of strategies and measures in mitigating risks (Ruzgas et al., 2023). Taking into account the particular problem or risk and the taxpayer's behaviour, the tax administrator strives to implement tax compliance strategies that maximize tax adherence at the most cost-effective rate.

2.2 Determinants of Tax Compliance

Business

In a Rwanda-based study conducted by Murorunkwere et al. (2022), various factors, such as the nature of businesses (cross-border or domestic), the operational duration of the business, its size (both small and corporate), the classification of a taxpayer or business variable as either individual or non-individual (company) were identified as particularly relevant to the detection of income tax fraud. Similarly, Murorunkwere et al. (2023) note that certain factors contribute to a higher susceptibility to tax fraud. Specifically, businesses with a significant time difference between their start and audit dates, domestic businesses, taxpayers involved in importing and exporting goods, those without recorded losses, businesses situated in the eastern province, and companies registered under VAT and withholding categories are identified as more prone to tax fraud.

Tax administrative data, which is usually available in the form of income declarations or yearly tax returns, can be used to obtain tax data. Essentially, this process involves extracting taxpayer history information from databases and data warehouses (Mabe-Madisa, 2018). An analysis using a logistic regression model to assess the taxpayer risk identification criteria at KRA revealed the significance of nine variables in determining payment compliance. These are business ownership, attributes of tax agents, performance targets, unpredictable sector performance, business nature, financial aspects of the taxpayer such as profitability and liquidity, the structure of the company, and the frequency of claims for investment deductions (Sirengo, 2018).

In a study conducted by Chepkwony (2017) in Kenya, several factors that influence the probability of a taxpayer undergoing an audit were established. These comprise the type of business, including whether it is a public commission or works on government projects, as well as its organizational structure—sole proprietorship, partnership, or company. The presence of oversight bodies in the business sector, utilization of tax agents/auditors for tax return preparation, and their standing with licensing authorities, as well as the frequency of changes in these agents, are additional considerations. Non-submissions, late submissions, and nil submissions of returns also played a role.

Industry

Abedin et al. (2021) states that financial indicators extracted from financial statements serve as predictors of tax default risk. It is noteworthy that, in similar business domains, financial statement data are thought to be the most critical aspects determining default risk (Serrano-Cinca et al., 2019). The primary factor distinguishing non-default taxpayers from their default counterparts was found to be the equity ratio, serving as a crucial early-warning signal for tax default. Liquidity was also highlighted as another important indicator, with the quick and current ratios, as well as the debt-to-sales ratio, all playing key roles. Companies with lower levels of debt, higher current assets, and greater liquid reserves were observed to be less prone to tax payment failures. Conversely, the inventory turnover ratio had the least impact on the RF's model prediction accuracy (Abedin et al., 2021).

Rahimikia et al. (2017) note that tax returns encompass valuable information that can be utilized to identify instances of tax evasion. The objective was to identify instances of corporate tax evasion. The variables 'Net income,' 'Total liabilities (TL),' and 'TL/ (TL+Shareholders Equity(SE))' were specifically chosen for analysis within the food sector. Conversely, the variables 'Net profit,', 'Net income/total assets (ROA), and 'Retained earnings/total assets were exclusively chosen for examination within the textile sector. It is important to note that distinct financial variables could be linked to tax evasion behaviour in each sector.

Battaglini et al. (2022) note that the primary indicators of tax evasion are the various accounts that report income (such as net income, gross income, and business income excluding salary income), turnover, net production, and VAT credits. When predicting detectable and recoverable tax evasion, the most significant cost entries are purchases and imports, since they are essential in establishing the VAT tax base. Additionally, demographic characteristics such as the years of activity and business size (measured by the number of employees and its logarithm) prove to be valuable in predicting evasion.

Other pertinent factors involved the existence of branches, the frequency and magnitude of investment deduction claims, the taxpayer's financial risk, discrepancies between declared and paid taxes, trends in profitability and liquidity ratios, deviations in gross and net profit margins from industry standards, disproportionate increases in taxable income relative to turnover, variations in profitability, taxpayers group structure, variations in VAT import declarations,

temporary exports, export diversions, fluctuations in the declared value of imports, and changes in the declared value of exports.

Sociological

Vellutini (2011) argues that tax compliance is influenced by various individual-specific elements, including gender, age, and education, as well as firm-related factors like industry type, company size, and financial standing. In developing a robust case selection methodology, obtaining information from external sources is crucial. Third-party data can validate details found in tax returns, historical cases, and general profiles of taxpayers or business sectors (OECD, 2004a)**.**

Psychological

According to Baghdasaryan et al. (2022) taxpayers undergoing intricate audits are identified through a risk identification system. Different criteria are employed to classify taxpayers into high, medium, and low-risk categories. Profitability, external economic activity, outcomes of previous audits, and the diversity of economic activities encompass a range of these factors. It was observed that two important indicators in the fraud model are the historical fraud occurrence rate and the administrative costs percentage in relation to total taxpayer expenses.

Notably, the data found in VAT returns, specifically, the percentage of sales that are exempt and zero VAT turnover, contributes significantly to the categorization of fraudulent activities. Profitability and productivity metrics were also established. Total revenues per employee served as a proxy for productivity, while the ratio of taxable profit to total revenues served as a measure of profitability. The data accessible for each business taxpayer included annual reports on corporate income tax, quarterly reports on VAT, monthly aggregated information on tax receipts (sent to final consumers) and/or invoices received/ transmitted (quantities and amounts), and information on the number of employees at regular intervals, and the nature of economic activity.

2.3 Utilization of ML Models in Predicting Tax Evasion

Abedin et al. (2021) employed feature transformation based ML to predict tax default in China. The findings indicated that XGBoost and a systematically constructed forest of several

Decision Trees surpass other ML techniques in terms of various classification performance metrics and accuracy.

Rahimikia et al. (2017) effectively combined evolutionary feature selection with data analytic methods (SVM, logistic regression, and Multi-Layer Perceptron (MLP)) to predict potential tax evaders. Their findings reveal that the MLP-based hybrid algorithm outperforms other methods in the context of Iranian corporate tax evasion within the food and textile sectors.

Pérez López et al. (2019) investigated tax fraud detection in Spain through the utilization of MLP Neural Network (NN) models. The results demonstrated the capability of NNs in segmentation of taxpayers and the computation of the likelihood of an individual taxpayer attempting tax evasion. With an efficiency rate of 84.3%, the selected model outperformed other models that were employed to identify tax fraud. It is noted that these models can aid tax offices in making decisions on the best course of action to take in order to successfully combat tax fraud.

Similarly, Murorunkwere et al. (2022) employed Artificial Neural Networks (ANN)s to detect indicators of tax evasion within income tax records in Rwanda. The findings indicate that ANN demonstrate effective performance in detecting tax fraud, achieving an accuracy rate of 92%, precision of 85%, recall score of 99%, and an Area Under the Receiver Operating Characteristics (AUC-ROC) of 95%. Murorunkwere et al. (2023) used supervised ML models to predict tax fraud in Rwanda. According to various evaluation metrics, the most robust model for tax fraud prediction was ANN.

Mamo (2013) investigated the possibility of using data mining technologies to create models that can recognize and predict fraudulent tax claims in Ethiopia. To build a predictive model, the study used classification methods and a clustering algorithm. Tax claims submitted by taxpayers were divided into two categories: those that were fraudulent and those that weren't using the K-means clustering technique. The classification model was developed using Naïve Bayes methods and a J48 decision tree algorithm. At 99.98%, the model developed using the J48 decision tree technique had the highest classification accuracy. Using a dataset of 2,200 examples, the model was tested and its prediction accuracy was 97.19%. This underscores the effectiveness of data mining methods in identifying tax fraud, and future studies in this field ought to concentrate on creating a workable system in this area.

Didimo et al. (2020) presented the MALDIVE approach for tax risk assessment in a research in Italy. It integrates data analytical techniques such as logistic regression, MLP, SVM, and RF with social network visualization obtained from taxpayer data. Notably, the RF technique demonstrated superior performance, achieving an accuracy of 74.3% in discerning both good and negative results of a fiscal audit.

Chepkwony (2017) used decision tree algorithm and historical data from taxpayer audit to construct a model to categorize taxpayers' compliance status in a case-selection application prototype in Kenya. The analysis focused on a limited set of taxpayer data for a single year. The results revealed that the model achieved an accuracy rate and prediction efficiency of 65%, specifically in detecting non-compliant taxpayers. This proved that the model works well and is appropriate for case selection. The researchers also highlighted the potential for improved accuracy and predictive efficiency by incorporating additional sources of taxpayer information and expanding the volume of data.

Geremew (2017) conducted a data mining analysis on e-filing data to predict fraud in Ethiopia. The research utilized several data mining algorithms, such as J48 Decision Tree Modeling, Experiment RF, and NN, incorporating both a 10-fold cross-validation and percentage split mode. The results indicated that J48 Decision Tree Modeling emerged as the most effective method for fraud prediction, achieving an accuracy rate of 94.719% with 10-fold cross-validation.

Chalye (2020) developed a hybrid ML model, combining both non-supervised and supervised ML techniques in Ethiopia. The primary objective was to predict the degree of fraudulent risk related to taxpayers. Notably, hybrid models, integrating multiple techniques, have demonstrated greater effectiveness compared to single approaches. The prediction model makes use of different taxpayer traits as well as historical taxpayer data. In these tests, the Knowledge Discovery in Databases (KDD) process model was used, and ML techniques like SVM for classification and K-means for clustering were performed to find instances of categorical events that are suggestive of fraud. 54.63% of the records were correctly clustered using K-means clustering, compared to 97% accuracy utilizing SVM prior to clustering. Second, an accuracy rate of 99% was obtained by employing SVM after clustering.

2.4 Summary

The literature demonstrates the increasing adoption of the application of ML algorithms to detect tax evasion. These ML approaches present a practical resolution to the difficulties encountered with conventional methods. Notable benefits encompass the capacity to handle extensive datasets and adapt to new information. However, there are certain research gaps identified in existing studies. In the study by Chepkwony (2017), a model was developed for auditors to use in selecting cases to evaluate taxpayers' compliance, employing the decision tree algorithm. In contrast, the current study will utilize various models in addition to the decision tree. While the focus of the study was on corporate tax and limited to historical data from taxpayer audits for a single year, the current study concentrates on VAT payment defaults, utilizing data spanning a five-year period. Didimo et al. (2020) presented the MALDIVE approach for tax risk assessment, conducted in Italy. In contrast, the current study is conducted in Kenya.

CHAPTER THREE

METHODOLOGY

This research focuses on designing, validating, evaluating and implementing a ML model with an objective of predicting risky taxpayers for audit case selection. The research employs Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, an industry proven approach to guide data mining efforts. The CRISP-DM methodology involves an iterative process encompassing six phases (Hotz, 2018). The first phase, business understanding, involves understanding the goals, requirements, and how data mining can provide value to the business. Data understanding is the second phase which entails closely examining the available data. This entails accessing and exploring the data through tables and visuals, allowing for the assessment of data quality. This stage is crucial for preventing unforeseen issues in the subsequent phase of data preparation.

The third phase, known as data preparation, encompasses pre-processing and transforming the data. This stage includes various tasks such as merging datasets or records, selecting a sample subset of data, aggregating records, creating new attributes, sorting the data for modeling purposes, and dealing with missing values by either removing or replacing them. This phase is important because the effectiveness of the data mining models is directly influenced by the quality of the data that has been prepared.

Modeling is the fourth phase which involves selecting and applying various modeling techniques to the prepared data. In the evaluation phase, the models' performance is assessed against the business goals. The final phase, deployment, involves integrating the models into existing systems, monitoring their performance, and ensuring they provide ongoing value to the organization. By adhering to the CRISP-DM methodology, researchers can unearth valuable insights from data, construct precise and dependable models, and ultimately provide information for business decisions and strategies that contribute to success.

3.1 Business Understanding

The research study commenced by determining the key stakeholders/ users who stand to gain from the research findings. The primary industry expected to be impacted by the proposed

research is the public sector, specifically the tax administration. The stakeholders that will benefit from the outcomes of the study include taxpayers, tax authorities, and the Government.

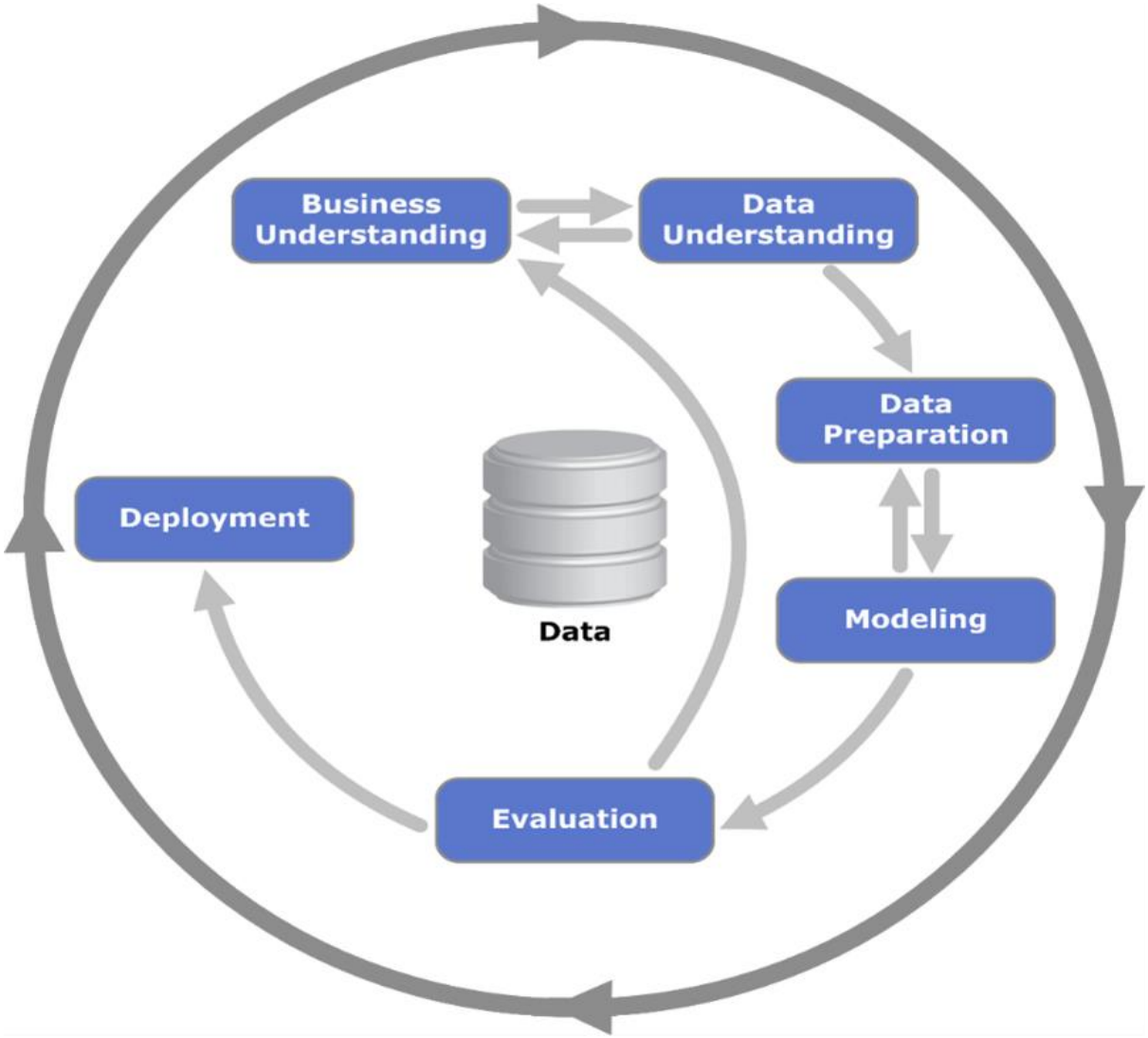


Figure 3.1: CRISP-DM Methodology

Source: Chapman, 2000

Taxpayers will benefit from a fairer and more consistent tax system. The research seeks to provide better treatment decisions of non-compliance. This makes the taxpayers feel supported in fulfilling their tax obligations and know that instances of gross non-compliance will be vigorously addressed. On the other hand, through identification of patterns in taxpayer behaviour that may indicate non-compliance, ML will enable the tax authority to effectively target the risky taxpayers thus fostering better allocation of scarce resources. This study gives a holistic view of taxpayers' non-compliance hence will facilitate treatment of the entirety of taxpayers' non-compliance and not just parts.

The Kenyan Government will benefit from the findings of this research. There will be improved revenue collection as a result of enhanced tax administration enabled by availability of accurate and timely information on risky taxpayers. In turn, the Government is able to fulfill its responsibilities and obligations to society which include funding public services and programs, such as education, healthcare, infrastructure, transportation, defense, and social welfare. The study's results can also inform the formulation of policies to harness the potential of taxpayers in the different sectors.

3.2 Data Understanding

Data was collected from the KRA database on the taxpayer demographic details, Income Tax declarations, VAT declarations and PAYE declarations. During the process of extracting data, Standard Query Language (SQL) queries were employed to extract required information from various database tables. The dataset consisted of transaction records for a period of five years from 2018 to 2022 covering taxpayers with the VAT obligation. This dataset had 24,013 records and 31 variables. In this context, the target variable pertains to the payment time, whereas input variables denote the features that influence the payment time of the taxpayer. Details regarding these variables are outlined in Table 1 provided below.

For confidentiality purposes the Personal Identification Number (PIN) and name of the taxpayer were excluded during the data collection process. This data was a critical resource for training, validating and applying ML models. By leveraging on this data this research sought to develop a classification model to predict risky taxpayers. The variables of the dataset that were used were those highlighted by literature in section 2.2 as determinants of tax compliance as well as the knowledge of tax experts and data availability.

Figure 3.2 below illustrates the steps that were taken in the methodology.

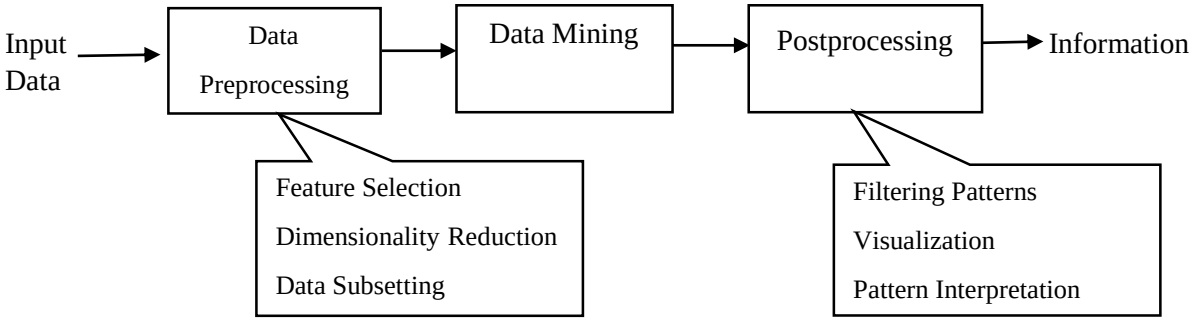


Figure 3.2: Knowledge Discovery in Databases

Table 1: A description of the Input Variables and the target Variable

Dataset	Description	Column Names
Taxpayer details	Specific attributes of taxpayers	Unique ID, Sector, Station, Region, Business Commencement date
Income Tax declarations	Information on specific income tax declarations	Gross turnover, Administrative expenses, Total expenses, Net profit, Current assets, Total inventory, Current liabilities, Share capital and reserves, Total debt, Capital allowance, Installment tax paid, Income Tax (IT) payable
VAT declarations	Information on specific VAT declarations	Zero rated sales, Exempt sales, Total sales, Total purchases, Output VAT, Input VAT, VAT payable, Credit balance from previous month, WHVAT credit, VAT advance paid, Debit adjustment voucher, Payment time (the target variable)
PAYE declarations	Information on specific PAYE declarations	Number of employees, PAYE payable

3.3 Data Preparation

Python programming language was used for this phase. A preview of the data was done to have a general understanding of the quality and completeness of the data. The preparation of the dataset for analysis involved various techniques, including addressing missing values (which are entries in the dataset that are empty, commonly referred to as null values), detecting and handling outliers, dealing with duplicates, and formatting the data to ready it for modeling. These are covered in the subsequent subsections.

3.3.1 Handling Missing Values

Missing data values are essentially unexpected empty or null entries in the dataset, which may have arisen from technical errors during data collection. These were checked using Python programming language with the .isnull() function. To address these particular missing data issues in the dataset, the following imputation techniques were applied:

a) Missing Completely at Random (MCAR), which occurs when the absence of data occurs randomly across variables and is independent of other features. Kenyhercz and Passalacqua (2016) encountered a similar issue within their dataset and resolved the missing values using K-Nearest Neighbor (KNN). This method was found to exhibit the highest accuracy, with the least discrepancy between imputed and actual values compared to the hot deck, iterative robust model and variable means. Actual values denote the complete data entries obtained during the data collection process, while imputed values represent entries calculated to fill the missing values in the collected dataset.

Essentially, the cluster-based method, employing an appropriate algorithm such as KNN, involves identifying the nearest neighbors of the record with missing data, imputing the values of those neighbors, and utilizing the outcome to fill in the null record. Consequently, this research embraced this approach for imputing missing values. The specific variable assessed using this technique was the business commencement date. This was implemented using KNNImputer from the sklearn library (Fadlil et al., 2022). On the other hand, the missing values under the variables sector, region and station were replaced with the category 'Other'.

b) Missing at Random (MAR) describes a situation where the absence of data in one variable can be explained by the presence or values of other variables within the dataset. For this study the variables imputed using this method include the gross turnover, administrative expenses, total expenses, net profit, current assets, total inventory, current liabilities, share capital reserves, total debt, capital allowance, installment tax paid and IT payable. These attributes were for taxpayers who were non-filers for the income tax obligation hence all the records under income tax were not existent in the database. To resolve this, the research imputed the missing values with zeros.

3.3.2 Outlier Treatment

An outlier refers to a data point that significantly deviates from the rest of the dataset, standing out as an extreme value. The examination of the dataset involved visually inspecting various plots like scatter plots and box plots to pick out potential outliers. Detecting and treating outliers is essential due to a number of reasons as they can significantly influence parameters and the performance of models, potentially leading to biased or inaccurate predictions. ML inherently relies on certain assumptions about the distribution and nature of data, but outliers have the

capability to violate these assumptions. Moreover, outliers can introduce noise during the training process, making the model less robust and affecting data normalization and scaling procedures.

Additionally, their presence can distort the interpretation of the model and consequently the findings derived from it. Outliers can also lead to overfitting. Therefore, addressing outliers is crucial to ensure the reliability and effectiveness of ML models. Once the columns with outliers were identified, the study applied the statistical Interquartile Range (IQR) method. This method entails computing the interquartile range (IQR) for each variable, which represents the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Any observations falling below Q1 - 1.5IQR or above Q3 + 1.5IQR were marked as possible outliers and handled. The technique employed was winsorization where the extreme values were replaced by the values at Q1 and Q3 in this case. This method has been used by other researchers, like Sharma and Chatterjee (2021), to address outlier data points.

3.3.3 Handling Duplicates

An inspection of the dataset using the .duplicated() function revealed that there were 31 duplicates. Typically, duplicates can introduce biases or inaccuracies into analyses or models hence it is crucial to address them. The research resolved these by removing them.

3.3.4 Data Formatting

This involved transforming data from one format to another to guarantee uniformity and suitability for analytical techniques. This was accomplished by recognizing incompatible data formats and opting for suitable conversion techniques. The business commencement date was converted from object type to datetime format.

3.3.5 Handling Class Imbalance

The issue of class imbalance occurs when there is a substantial disparity in the representation of different classes within the data, with one class (the majority class) having a much higher proportion than another (the minority class). Murorunkwere et al. (2023) had the same challenge with their dataset and utilized the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate it. In the utilization of SMOTE, each sample from the minority class is oversampled by introducing synthetic examples that resemble the k nearest neighbors in the

minority class. The selection of neighbors from the k nearest neighbors is done randomly, taking into account the desired amount of oversampling (Chawla et al., 2002).

3.4 Feature Engineering

Feature engineering plays a vital role in pre-processing by creating new features from existing data to enhance the predictive capability and interpretability of the analysis. Moreover, it can amplify the clarity of results, facilitating a more profound understanding of the underlying factors influencing the observed phenomena. The key new features generated are elaborated below**.**

Effective Tax Rate

This is a measure of the actual tax rate a company pays on their income or revenue after taking into account deductions, credits, and other adjustments. The effective tax rate calculated gives an insight into how much of the entity's revenue is going towards tax payments relative to its total VAT sales or gross turnover. As calculated in equation 1, one can assess the proportion of revenue that should have been allocated to taxes but was not, providing an indication of the severity of the default. It helps in understanding the tax burden relative to revenue and can provide insights into financial management and compliance. This approach is important since it enhances accuracy in predictive analytics, and it gives a more nuanced understanding of taxpayer behaviour thus improving the interpretation of payment default models.

$$Effective\ Tax\ Rate = \frac{\text{IT } payable + \text{VAT } payable + \text{PAYE } payable}{\max\left(Gross\ Turnover,\ Total\ Sales\right)} \quad (1)$$

VAT input output ratio

The second feature derived is the VAT input output ratio. This is a measure of the relationship between the input VAT and the output VAT generated by a business. It acts as a crucial measure for comprehending and predicting patterns in taxpayer behaviour. This ratio as seen in equation (2) indicates how much VAT a business is paying compared to how much it is collecting. A ratio greater than 1 suggests that the business is paying more VAT than it is collecting, while a ratio less than 1 indicates the opposite.

$$\text{VAT } Input\ Output\ Ratio = \frac{Input \text{ VAT}}{Output \text{ VAT}} \quad (2)$$

Profit margin

The profit margin measures the percentage of revenue that represents profit after accounting for all expenses. It serves as a key indicator of a taxpayer's profitability and operational efficiency. Notably, it empowers the tax authority to identify taxpayers with a low profit margin, who are potentially at greater risk of defaulting due to limited financial resources to fulfill their tax obligations. Analyzing patterns in profit margins can offer valuable insights into taxpayers' likelihood of default. By assessing average profit margins across different sectors, tax authorities can anticipate periods when taxpayers may be more prone to default and implement preventive measures to address this risk. The profit margin formula is given by equation 3 below.

$$ProfitMargin = \frac{NetProfit}{GrossTurnover} \quad (3)$$

Debt to sales ratio

The significance of the debt-to-sales ratio is a key factor. This metric is pivotal in evaluating the extent of a company's debt in relation to its sales revenue. A higher debt to sales ratio may suggest that a taxpayer is heavily reliant on borrowed funds, which could indicate a higher likelihood of payment default. Monitoring this ratio can serve as an early warning sign of financial distress, allowing tax authorities to intervene promptly to minimize default occurrences. Moreover, the debt-to-sales ratio serves as a valuable tool for identifying high-value taxpayers. Those with below average ratios are generally less prone to defaulting. A lower debt-to-sales ratio often correlates with lower default rates, as taxpayers exhibit stronger financial leverage and a better ability to manage their debt obligations. Equation 4 below provides the formula for calculating the debt-to-sales ratio.

$$Debt - to - SalesRatio = \frac{TotalDebt}{\max(GrossTurnover, TotalSales)} \quad (4)$$

Proportion of administrative costs to total expenses

This financial metric reveals the portion of a company's overall expenses allocated to administrative costs. As illustrated in Equation 5, this ratio provides insight into the efficiency of administrative cost management. This aspect holds significant importance in predictive models for payment default, as it sheds light on a taxpayer's ability to generate sufficient cash flows to cover their tax liabilities. If a taxpayer's administrative expenses are disproportionately high compared to their total expenses, their risk of tax payment default increases. By integrating this ratio into payment default prediction models, tax authorities can pinpoint such taxpayers and provide them with tailored tax education and stakeholder engagement. Such personalized approaches are crucial for enhancing taxpayer experiences and ultimately reducing default rates.

$$Proportion of Administrative costs to Total Expenses = \frac{Administrative Expenses}{Total Expenses} \quad (5)$$

Current Ratio

Another feature generated in this research is the current ratio, used to assess a company's ability to cover its short-term liabilities with its short-term assets. A comprehensive understanding of the current ratio aids the tax authority in acquiring valuable insights essential for understanding and mitigating payment default risks. For instance, a low current ratio may imply that the taxpayer lacks sufficient liquid assets to cover their short-term liabilities potentially leading to a higher likelihood of defaulting on tax payments. Analyzing current ratio can help the tax authorities to tailor interventions by identifying segments of taxpayers with high likelihood of defaulting. Equation 6 provides the formula for calculating the current ratio.

$$Current Ratio = \frac{Current Assets}{Current Liabilities} \quad (6)$$

Quick Ratio

The quick ratio, also known as the acid-test ratio, plays a pivotal role in default prediction by providing insights into a taxpayer's ability to promptly fulfill its short-term financial obligations without relying on inventory sales. A high quick ratio implies that the company is better positioned to manage its financial commitments, reducing the likelihood of tax payment

default. Conversely, a low quick ratio may indicate liquidity constraints, suggesting that the company may struggle to meet its tax obligations if it cannot quickly convert its assets into cash. By integrating the quick ratio into their default prediction models, tax authorities can pinpoint at-risk taxpayers and take proactive measures to mitigate default risk, consequently boosting revenue. Equation 7 outlines the formula for calculating the quick ratio.

$$QuickRatio = \frac{CurrentAssets - TotalInventory}{CurrentLiabilities} \tag{7}$$

Gearing Ratio

The gearing ratio, also known as the leverage ratio, measures the proportion of a company's debt relative to its equity or shareholder's funds. In the context of default prediction, understanding the ratio can provide insight into the extent to which a company relies on debt financing to fund its operations. A company with a high gearing ratio may face difficulties in generating sufficient cash flows to cover both its debt obligations and tax payments. If the company's financial position deteriorates further, it may be at greater risk of defaulting on tax payments.

$$GearingRatio = \frac{TotalDebt}{ShareCapitalandReserves} \tag{8}$$

Proportion of exempt and zero-rated sales to total sales

This feature refers to the percentage of sales revenue that is exempt from or subject to a zero VAT rate out of the total sales revenue generated by a business. This information serves as a valuable tool for identifying taxpayers at risk of default. Through analysis of this ratio, tax authorities can uncover underlying patterns in taxpayer behaviour associated with default tendencies. If a business heavily relies on exempt or zero-rated sales and faces financial difficulties, it may struggle to meet its VAT payment obligations. This could lead to tax payment defaults, particularly if the business struggles to generate sufficient taxable revenue to cover its VAT liabilities. The proportion of exempt and zero VAT turnover in total sales formula is given by equation 9 below.

$$Percentageexemptandzerototototalsales = \frac{Exemptsales + Zeroratedsales}{TotalSales} \times 100 \tag{9}$$

Total Liabilities

Total liabilities are commonly incorporated in default prediction models due to their significance in understanding a company's financial obligations to external parties, encompassing both short-term and long-term liabilities. This feature is crucial because of the influence it has on payment default. If a taxpayer has a high level of total liabilities and faces financial difficulties, they may struggle to meet their tax payment obligations. High levels of debt and other liabilities can strain the company's cash flow and financial resources, potentially increasing the risk of defaulting on tax payments. Analysis of total liabilities data can help the tax authorities to identify potential areas of concern and make informed decisions regarding tax audits, compliance reviews, enforcement measures and taxpayer education to mitigate this risk, thereby enhancing payment compliance.

$$Total\ Liabilities = Current\ Liabilities + Total\ Debt \tag{10}$$

Age

The age of a company typically refers to the length of time since its incorporation. A company's age may reflect its ability to withstand economic downturns, adapt to market changes, and maintain stability over time. Age plays a vital role in predictive modeling for defaults for several reasons. Firstly, established companies with longer operational histories may have more stable financial positions, making them less likely to default on tax payments. Newer companies may not have built up reserves, established credit lines, and developed financial management practices to handle tax obligations effectively, making them more prone to default.

Additionally, the age factor allows tax authorities to pinpoint taxpayers who are at risk of defaulting due to lack of engagement. By integrating age as a variable in default prediction models, tax authorities can target such taxpayers with tailored taxpayer education and reminders aimed at enhancing their understanding and awareness of tax obligations. Moreover, comprehending the age of taxpayers enables tax authorities to segment taxpayers effectively, facilitating the design and implementation of tailored mitigation strategies for different segments. This segmentation is pivotal for delivering personalized and impactful experiences to taxpayers, consequently reducing default rates. For this research, the age of the taxpayer was calculated by subtracting the business commencement date from the date when the data was collected.

Labelling Initial default

The criteria used for labelling default is derived from domain expertise, industry best practices and availability of data. VAT is a monthly obligation and given that the data covered a period of 5 years there were 60 data points. The research utilized 30 data points of the payment time feature to come up with a statistical normal distribution. Any taxpayer who fell beyond the 75th percentile (Q3) of late payment was likely to have defaulted. The rationale behind this criterion is that taxpayers who have made payments on time multiple times (in this case, more than 22 months) demonstrate a certain level of tax compliance. However, if these taxpayers have not made payments on time for an extended period (less than 22 months), it may indicate a shift in their behaviour, suggesting that they have compliance, or awareness issues that need to be addressed by tax authorities through enforcement actions or educational efforts.

3.5 Exploratory Data Analysis

The research utilized an exploratory data analysis approach to comprehend the distribution, correlations, and interactions within the dataset.

3.5.1 The distribution of payment defaults per sector

The proportion of payment defaults across the different sectors was plotted using a stacked column chart. This visualization provides insight into the proportion of payments made on time versus those that are delayed within each sector. The results and discussion are provided in section 5.1.

3.5.2 The distribution of payment defaults across age

The distribution of payment defaults across the taxpayers' age was illustrated using a histogram. This provides valuable insights into the financial behaviour and risk profile of businesses over time. The findings and discussion are highlighted in section 5.1.

3.5.3 The distribution of payment defaults by number of employees

This was shown using a histogram. It provides insights into how the size of a company correlates with its likelihood of defaulting on payments. The results and discussion are depicted in section 5.1.

3.5.4 Correlation analysis

The research employed correlation analysis to explore the relationship between different variables. The findings and discussion are depicted in section 5.1.

3.6 Machine Learning Modeling

3.6.1 Feature Transformation

Feature transformation comprises methods aimed at converting variables into formats suitable for subsequent analysis. For instance, label encoding was integrated to convert the payment time variable from categorical data to numerical data to facilitate the learning process since ML models are not capable of handling plain text data in its raw form (Seger, 2018). Label encoding is typically used when there is ordinal data, meaning the categories have some kind of order or ranking. In this research, label encoding was applied using sklearn's LabelEncoder in Python.

Additionally, one-hot encoding was incorporated to convert the sector variable from categorical to numerical data for ML. One-hot encoding is used when there is no ordinal relationship between categories, or when there is no inherent order among the categories. One-hot encoding was executed in the study using the get dummies() function from the sklearn Python library.

3.6.2 Data Splitting

The data was divided into a 70:30 ratio where 70% of the data was allocated for training the models, and the remaining 30% served as the test set.

3.6.3 Training Machine Learning Models

This involved training or fitting six supervised ML models that are inbuilt in the sklearn library in Python. The models used for training were logistic regression, decision trees, random forest (RF), SVM, eXtreme gradient boosting (XGBoost) and stacking. The best model was then selected to classify the taxpayers based on their risk behaviour.

3.6.4 Logistic Regression

Logistic regression, as described by Russell and Norvig (2010), is a classification technique rooted in statistics. Its fundamental application involves categorizing data elements into two classes, a process known as binary classification. Central to logistic regression is the utilization of a logistic function, which employs the natural logarithm. This function produces values

ranging from 0 to 1, representing the likelihood of a data element belonging to a particular class. In practice, logistic regression adjusts weights associated with data features to minimize the disparity between the predicted class and the actual class (Lozano et al., 2021).

In the current study, the outcome Y, which indicates whether a taxpayer is a non-defaulter or defaulter, is represented as either one or zero.

$$Y = \{1, \text{non-defaulter } 0, \textit{defaulter}\} \tag{11}$$

for X = x1 + …+ xn, where x1 + …+ xn are the features under consideration. The likelihood of a taxpayer defaulting is expressed as (Shalev-Shwartz and Ben-David, 2014)

$$\pi(x) = E(Y|x_1, \ldots, x_n) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_n)} \tag{12}$$

Logistic regression is suited for making relatively broad assumptions: it does not require that independent variables be normal distributed; nor does it necessitate that regression errors follow a normal distribution, or require homoscedasticity (uniform variance) in the target variable to be assessed (Ruzgas et al., 2023). The logistic regression was employed in this research as a foundational or baseline model against which other models were evaluated and compared. The algorithm is chosen for its simplicity, efficiency, and interpretability.

3.6.5 Decision Trees

Classifiers denoted as h: X −→ Y, designed to forecast the category linked to an instance, such as variables, by traversing from the root node to a leaf (Yego et al., 2021). A decision tree is constructed in the form of branching fragments. It comprises leaf nodes and root nodes that correspond to class labels, while non-leaf nodes are indicated by the intermediate nodes. The root node is determined by selecting the data property with the most significance in decision making. The data values at each node in the decision tree dictate how the nodes split. The critical aspect of decision trees revolves around selecting the appropriate decision node (attribute) for branching out.

The goal is to minimize impurity or uncertainty within the data to the greatest extent possible (Mabe-Madisa, 2018). The heuristic aim is to select the attribute that offers the highest information gain. Entropy which given by equation 13 serves as a metric to gauge the level of

impurity or disorder within information. Learning occurs during the training phase, and the evaluation of effectiveness takes place in the testing phase. The dynamic impact on the classifier's performance and efficiency is influenced by the distribution and depth of training and test information within the decision tree (Naganandhini and Shanmugavadivu, 2019).

Described as tools for determining the most likely strategy to achieve a goal, decision trees find utility in addressing classification and regression problems. Their frequent adoption in ML is attributed to their notable speed and accuracy (Brownlee, 2019). In this study, decision tree is employed to predict tax payment default; the leaf nodes symbolize classification labels, indicating payment defaulter or non-defaulter. The rationale behind incorporating decision trees into this study is to evaluate them against the corresponding ensemble algorithms: RF, XGBoost and Stacking.

$$Entropy(S) = -\sum_{j=1}^{|c|} Pr(c_j) \log_2 Pr(c_j) \tag{13}$$

$$\sum_{j=1}^{|c|} Pr(c_j) = 1 \tag{14}$$

3.6.6 RF

The RF is comprised of numerous decision trees. In the practical implementation of this approach, sampling a random subset of input data features and randomly selecting a training set of data pieces for each tree is done. Based on these sets, each tree is grown without pruning. The objective is to minimize classification error, and to achieve this, the trees within the ensemble should be as dissimilar as possible, enhancing the collaborative effectiveness among them. Each tree predictor in the forest provides a class output, and the final prediction of the RFs is determined through a majority vote. The fundamental premise underlying this method is that a group of decision trees will yield superior performance compared to any individual tree considered in isolation (Lozano et al., 2021).

Using the developed model, the decision trees enable identification of objects (taxpayers, for example) based on their potential association to a specific classification group, assignment of entities (taxpayers, for example) to distinct categories, such as payment non-defaulter and defaulter groups and prediction of future events, like tax evasion (Ruzgas et al., 2023). RFs

offer several advantages, including handling non-linearity and capturing complex relationships in the data. The randomness introduced during tree construction helps mitigate overfitting, making the model generalize well to new data.

3.6.7 SVM

SVM operate by establishing a frontier hyperplane that effectively separates two classes by maximizing the margin between the closest points of each class within the training set, known as support vectors (Yego et al., 2021). Through the kernel trick, SVM can transform finite dimensional space features into higher-dimensional space, enabling linear separation despite the original dimensionality (Shalev-Shwartz and Ben-David, 2014). SVM have been widely utilized across various domains and have demonstrated high accuracy in numerous scenarios, including cancer diagnosis (Huang et al., 2018). In this study, support vectors facilitated the creation of the widest margin between taxpayers defaulting and those not defaulting, based on the provided features.

3.6.8 XGBoost

XGBoost is a type of supervised learning models based on trees. It utilizes an ensemble learning technique called boosting, where classification trees are trained sequentially to enhance the performance of each subsequent tree based on the errors of the previous one. The model employs a more formalized regularization compared to the Gradient Boosting Machine (GBM). This enhanced regularization in XGBoost improves control over overfitting, leading to superior model performance (Yego et al., 2021). This includes terms for tree depth, leaf nodes' weights, and the number of leaves.

XGBoost is an advanced implementation of gradient boosting designed for speed and performance. The algorithm's combination of gradient boosting, regularization, and optimization techniques results in a model that is robust, accurate, and resistant to overfitting. XGBoost also uses a technique called "pruning" during tree construction, removing branches that do not contribute significantly to the improvement of the model. This contributes to model efficiency.

3.6.9 Stacking

Ensemble learning, as described by Rokach (2010), involves amalgamating the outcomes of multiple learning algorithms with the aim of achieving superior performance compared to any of its constituent algorithms used in isolation. These techniques merge the outputs of a group of classifiers. This research utilized stacking which is a type of ensemble learning that combines multiple classifications or regression models via meta- classifier. The stacking involved three algorithms, which included the RF, GBM and logistic regression classifiers.

Stacking serves to reduce overfitting by averaging out individual predictions (variance reduction). Typically, assembling a heterogeneous committee of classifiers leads to improved results (Lozano et al., 2021). Nevertheless, its efficacy relies on the selection of base classifiers and their variation. Moreover, the computational burden can be significant due to the necessity of training and predicting with numerous classifiers.

3.7 Evaluation

To evaluate the models various metrics including precision, recall, F1-score, accuracy, and AUC-ROC were used to measure performance of the ML algorithms. The comparison offered insights into the capability of each model to predict the classes of payment defaulter and non-defaulter. Higher precision, recall, F1, accuracy and AUC-ROC scores indicate better model performance, whereas lower scores suggest poorer model performance.

The highest performing model would be used to predict the future cases that need to be subjected to the different interventions and subsequently compared to those from existing literature. In the performance measurement equations, TP denotes true positives, TN denotes true negatives, FP signifies false positives, and FN denotes false negatives.

Precision is the ratio of correctly predicted positive cases to the total predicted positive cases (Murorunkwere et al., 2023). It measures the classifier's capacity to accurately detect positive instances, which is particularly crucial in situations where false positives come with significant costs. Equation (15) gives the calculation of precision.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

The recall also known as true positive rate (TPR) or Sensitivity is the percentage of positive cases that will be accurately identified (Didimo et al., 2020), as illustrated in equation (16). This metric is of importance in addressing the payment default issue because incorrectly categorizing a taxpayer as non-defaulter when they are actually a payment defaulter carries substantial risks and costs.

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

The F1-score calculates a certain average of the precision and recall metrics in information retrieval (Murorunkwere et al., 2023). It proves especially beneficial in scenarios where the costs of both false positives and false negatives are substantial.

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{17}$$

A confusion matrix, serving as a table illustrating accurate predictions and different types of prediction errors, was utilized (Yego et al., 2021). It includes TP, FP, TN, and FN. AUC-ROC is a performance metric used to assess a classifier's capability to differentiate between positive and negative instances across various classification thresholds (Didimo et al., 2020). It plots the TPR versus the False Positive Rate (FPR), where the FPR represents the proportion of actual negative cases that are incorrectly classified as positive by the classifier. FPR can be obtained by equation (18).

$$FPR = \frac{FP}{FP + TN} \tag{18}$$

It can also be calculated by [1-(True Negative Rate (TNR)], that is 1-Specificity.

The TNR means when it is actually no, how often does the model predict no? (Yego et al., 2021). This is given by equation (19).

$$TNR = \frac{TN}{FP + TN} \tag{19}$$

Accuracy, defined in equation (20), is characterized as the ratio of the total number of correct predictions to the overall predictions made (Yego et al., 2021). This metric can lead to misleading outcomes when dealing with imbalanced datasets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

3.7.1 Ranking of Features

The impact of features, their individual contributions, and their respective attributions collectively elucidate the manner and, to a certain extent, the magnitude of influence each feature has on the model's prediction (Casalicchio et al., 2018). This study employed the built-in feature importance function within the sklearn Python library, to assess the significance of features in predicting the class of the taxpayer. For instance, equation (21) evaluates the significant attributes within the RF classifier model. Here, f, t, and m denote the features, trees within the RF classifier, and nodes, respectively. The *Gain* function is the predefined function utilized for computing the feature importance (Pedregosa et al., 2011).

$$\mathcal{I}mportance_f^t = \frac{\sum_{m \in \mathcal{M}_f^t} Gain_m}{\sum_f \sum_{m \in \mathcal{M}_f^t} Gain_m} \tag{21}$$

3.8 Deployment

The prediction model was deployed using a web application allowing the users to interact with the results of the study. The application provides the predictions for the taxpayer categories providing an overview of the risky taxpayers. Subsequently, users are able to select audit cases. This will guide decision making regarding resource allocation to specific areas, aiming for increased revenue.

Python was used for model development whereas the web application was built in the Streamlit Python framework. The deployment involved several steps, including building the model, saving the pre-trained model, creating a Streamlit application to showcase the model's functionality, and deploying it to a server. The Streamlit application was created in Python using a .py file. The pre-trained model was then loaded and predictions made using the model based on user input. These were displayed on the web application.

System Design and Architecture

This involves the analysis and design of the system model. It serves as a comprehensive exploration of the design principles, methodologies, and decisions made in the development of the software system or application. Figure 3.3 depicts the system design of the dissertation.

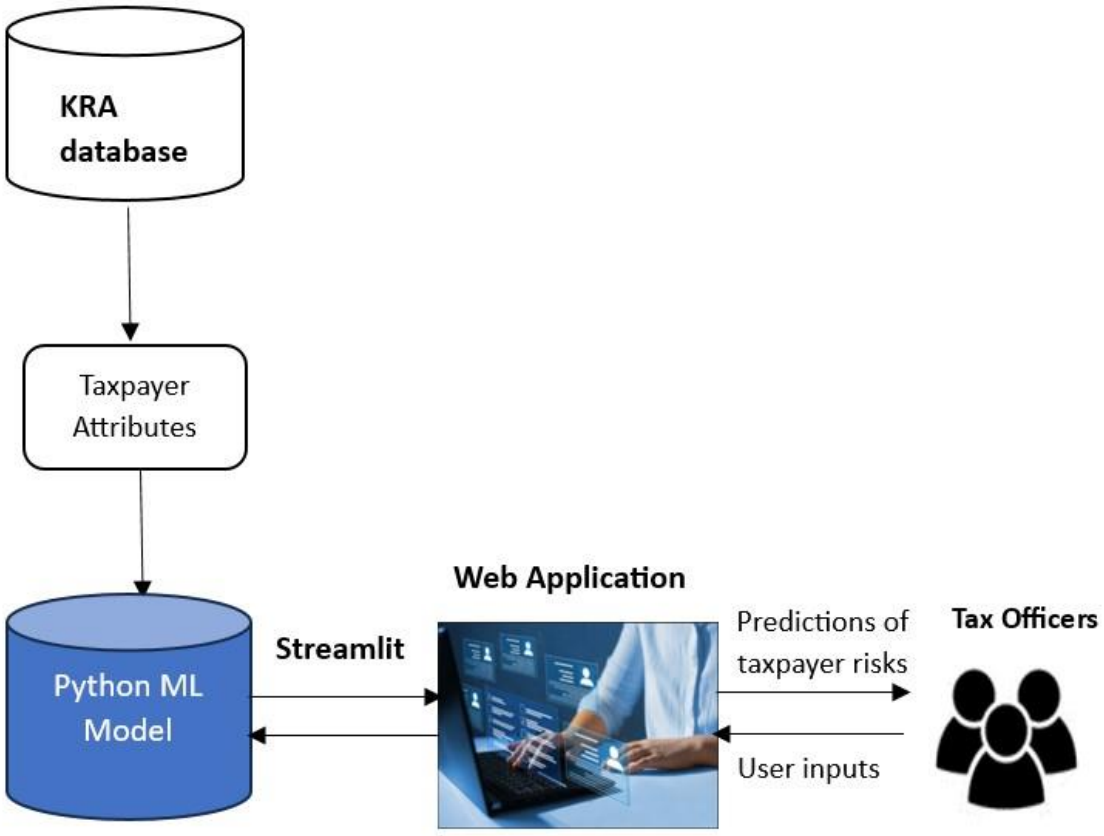


Figure 3.3: System Design

Data containing the key attributes for analysis is extracted from the taxpayer database. This data serves as inputs to a classifier, the output of which is then utilized by a web-based application for case selection. A compliance officer accesses the web application, chooses a taxpayer or a group of taxpayers for classification. The application then classifies the taxpayer(s) and presents the classification to the user.

Web Application User Interface

The user interface for the web application was built using the Streamlit framework. Streamlit is an open-source Python library used for building interactive web applications. It allows users to create web applications directly from Python scripts without requiring knowledge of web development languages such as HTML, CSS, or JavaScript. Streamlit was employed in this research due to its simplicity, flexibility, and integration capabilities with various data science libraries. It provides a simple and intuitive Application Programming Interface (API) that enables one to quickly prototype and deploy data-driven applications.

The tax form that takes in user inputs was created using the key attributes used in the study for the tax default prediction. Figure 3.4 illustrates the web application user interface.

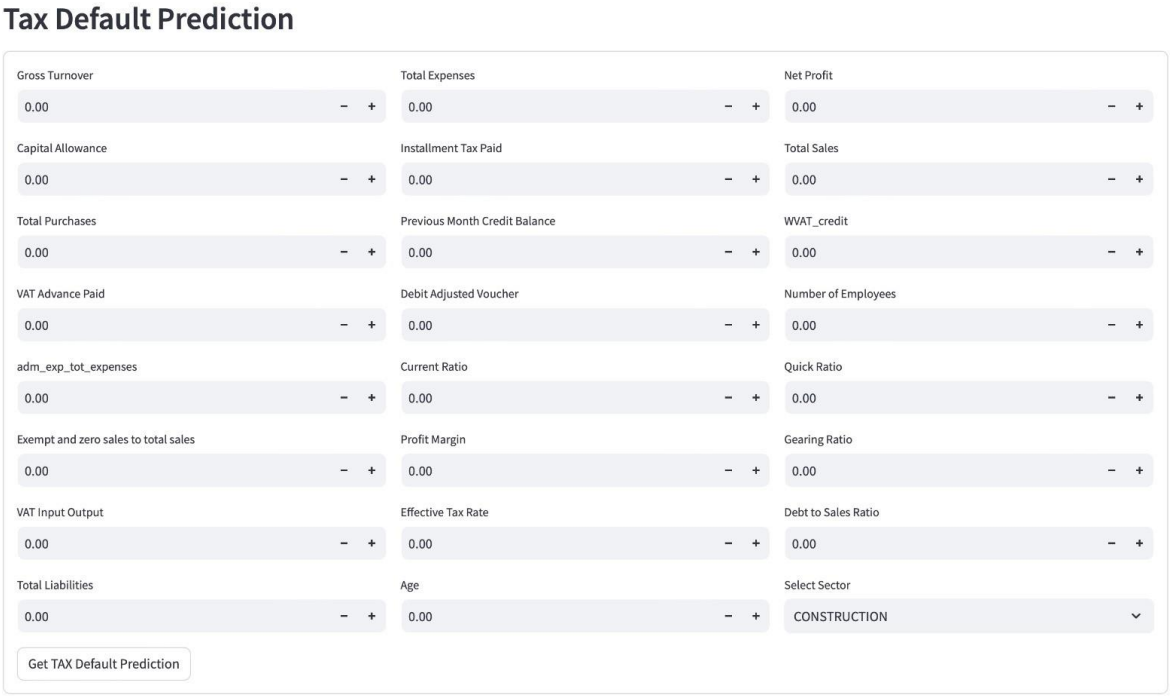


Figure 3.4: Web Application User Interface

Streamlit has a variety of built-in widgets such as sliders, buttons, dropdowns, text inputs, and date pickers. These widgets enable users to interact with the application, input data, and customize parameters dynamically. Users can adjust parameters, observe changes in real-time, and explore different scenarios. Customization options like themes, layouts, and styling can also enhance the user experience. The implementation and testing of the system is discussed in Chapter 4.

CHAPTER FOUR

SYSTEM IMPLEMENTATION AND TESTING

4.1 System Implementation

In this research, the trained RF model was saved, after which a web portal was built for a tax officer to interact with the model. The user does this by providing details about a taxpayer. The form validates the inputs and calls a function containing the pre-trained model. These functions process user inputs, pre-process the data if necessary, and feed it into the trained classifier model. The model predicts the class of the taxpayer and returns the classification output displayed as defaulter or non-defaulter to the tax officer through the web application user interface. This is possible because the pre-trained model is trained on historical taxpayer data with labeled default and non-default cases.

4.2 Testing

This involved testing individual components of the system, including data extraction scripts, classifier model, and web application functions, to ensure they performed as expected. Using the input fields, taxpayer details such as age, gross turnover, and effective tax rate were entered in the user interface. In this case, the data entered corresponded to a taxpayer known to be a defaulter. Upon clicking the “Submit” or “Get Prediction” button, the system processed the input data.

The trained classifier model predicted the taxpayer’s default status based on the input data. The system returned the prediction result, displaying a message on the interface: “The taxpayer is likely to be a defaulter.” Figure 4.1 shows the results of the user interface given a default case.

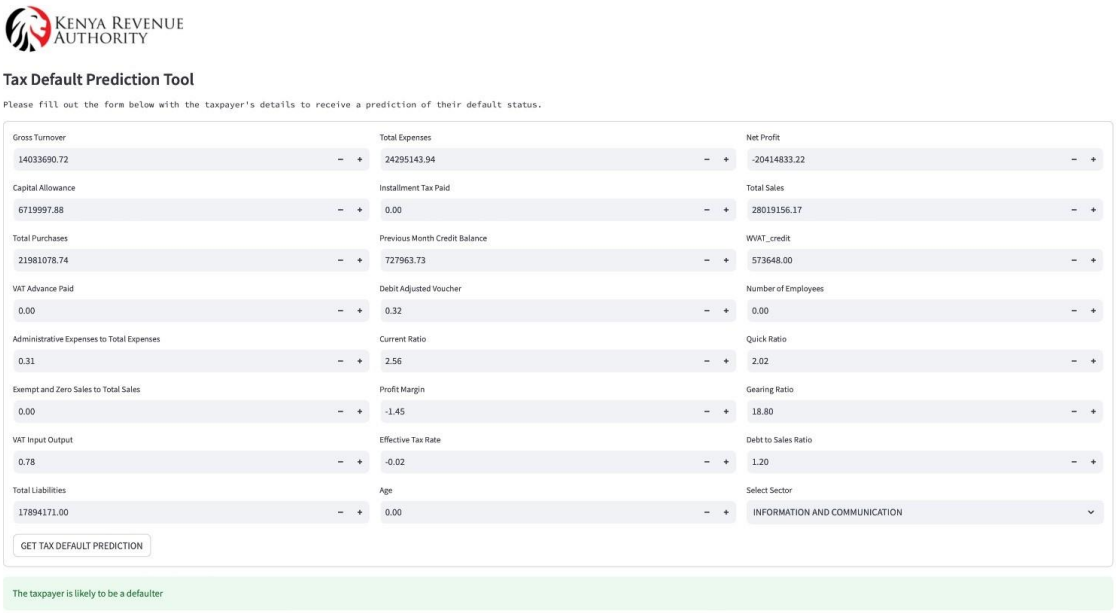


Figure 4.1: Defaulter Prediction

Likewise, the system was tested using data corresponding to a taxpayer known to be a non-defaulter and the process as illustrated above repeated. The classifier model made a prediction regarding the taxpayer's likelihood of default using the provided input data. The system then conveyed the prediction outcome through a message displayed on the interface, stating: "The taxpayer is likely to be a non-defaulter." Figure 4.2 illustrates the interface results in the scenario of a non-default case.

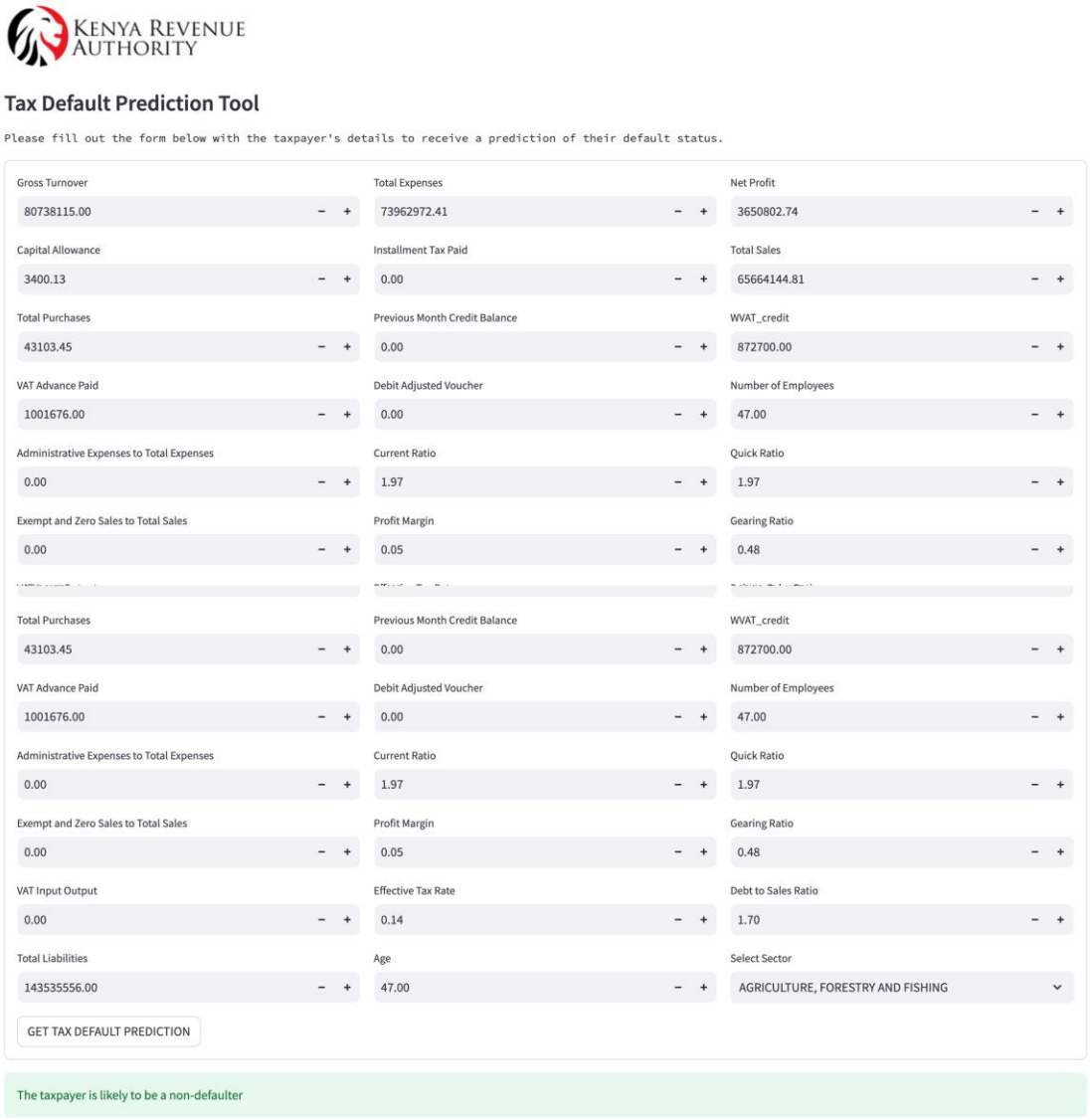


Figure 4.2: Non-defaulter Prediction

These testing scenarios validate the functionality of the system in correctly predicting the default status of a taxpayer based on the provided input data. They confirm that the classifier model, integrated with the web application, performs as expected in identifying default cases accurately. Such testing ensures the reliability and effectiveness of the system in assisting compliance officers in making informed decisions regarding potential tax default cases.

CHAPTER FIVE

DISCUSSION OF RESULTS

The results of this study offer valuable insights into default prediction, as well as the underlying structure of the dataset. These insights can be utilized in several ways to guide the tax authority's decisions, enhance mitigation strategies, and bolster revenue collected.

5.1 Descriptive Statistics

The dataset employed in the study exhibited a degree of imbalance, mirroring a common scenario encountered in real-world contexts. Figure 5.1 illustrates the proportional distribution of data balance with 82.5% (19,800 out of 23,982) of taxpayer data points labeled as on-time payment and 17.4% (4,182 out of 23,982) labeled as late payment. When the specific event of interest, particularly the underrepresented late payment (minority class), occurs, it tends to impede the accuracy of classification (Murorunkwere et al., 2023). This is because the ML algorithms may learn one class but not the other, leading to the introduction of bias and variance. Following the utilization of SMOTE (Yego et al., 2021), the resulting dataset comprised 19,800 data points.

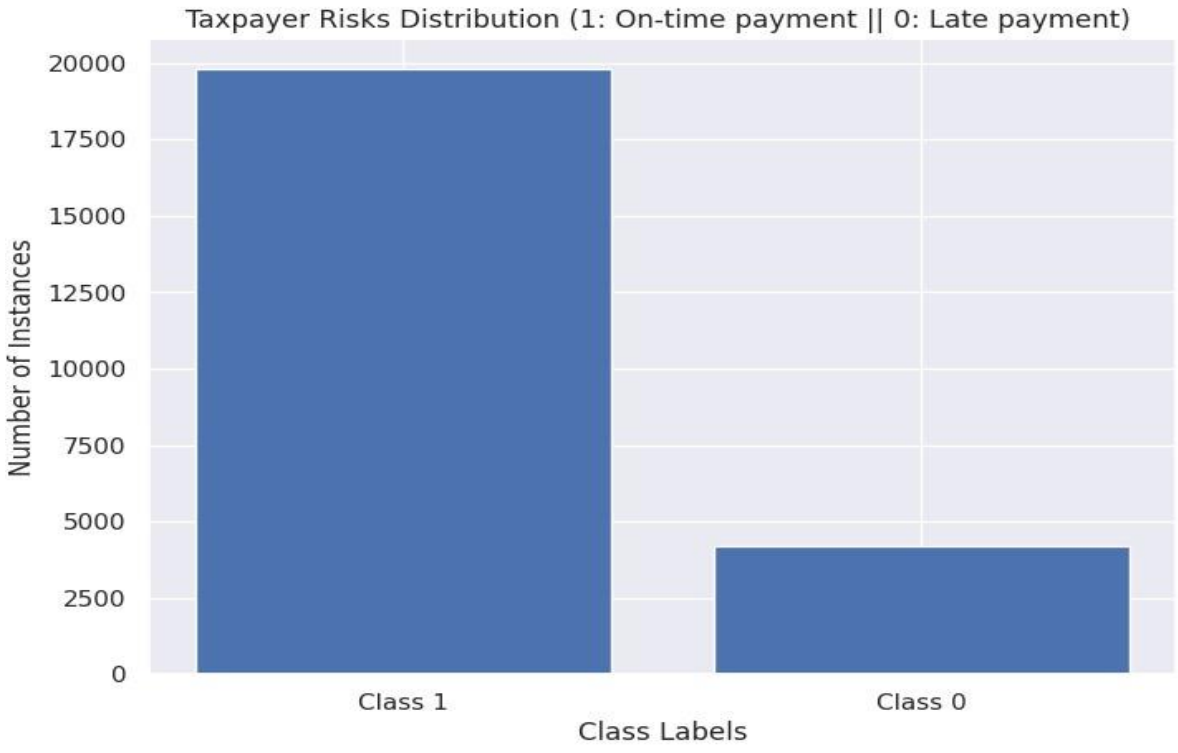


Figure 5.1: Class Imbalance

Figure 5.2 below depicts the distribution of payment defaults across the different sectors. The highest number of defaulters were in the Construction, Information, Communication and Technology (ICT) and digital economy sectors. This could be attributed to factors such as high cash transactions, complex business structures, and opportunities for under reporting income. For instance, from industry experience the Construction sector depicts a culture of non-compliance with wide scheme in fictitious claims. This corresponds with research conducted by Baghdasaryan et al. (2022), which identified external economic activity as one of the significant predictors of fraud.

Awareness campaigns and risk based audits need to be put in place for these sectors to assist in increasing their levels of compliance. In cases where companies are experiencing temporary financial difficulties, the tax authority can offer flexible payment arrangements. This can help companies manage their cash flow challenges while still meeting their tax obligations. Conversely, Real Estate and Financial/Insurance sectors have the least number of defaulters. This might indicate better compliance practices and stringent regulations due to the high-value transactions typically associated with these sectors.

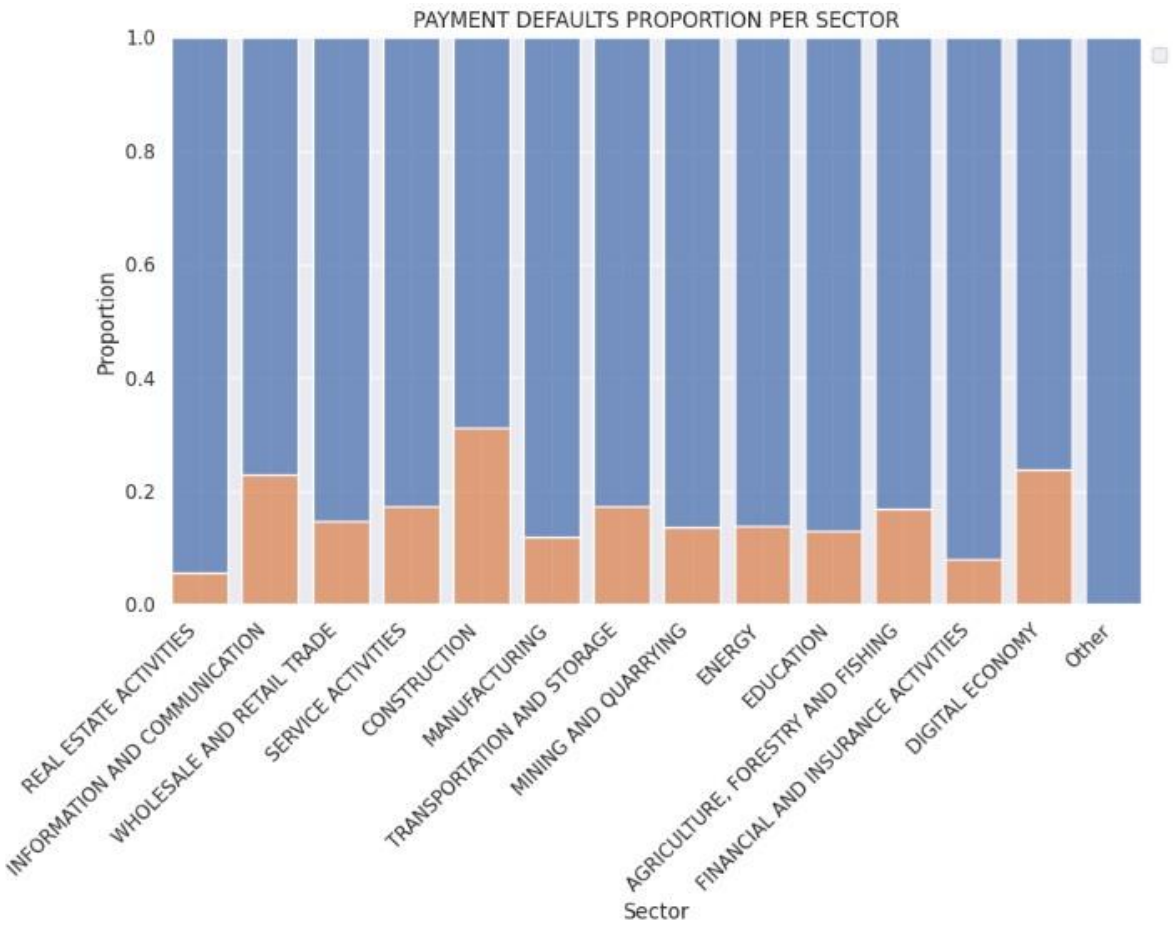


Figure 5.2: Distribution of Payment Defaults per Sector

Figure 5.3 below, depicts the distribution of payment defaults across the taxpayers' business lifespans. It was observed that the newer the company the more the defaults. This suggests that startups may face financial challenges and cash flow constraints in their initial stages, making it difficult for them to meet their tax obligations promptly. Additionally, new businesses may lack experience or knowledge regarding tax compliance requirements and procedures, leading to inadvertent defaults. This aligns with studies conducted by Murorunkwere et al. (2022), Battaglini et al. (2022), and Vellutini (2011), which emphasized the significance of a business's operational period in detecting income tax fraud.

The tax authority should implement proactive measures such as targeted taxpayer education and reminders. This will aid in promoting tax compliance awareness and assistance, helping startups navigate tax obligations effectively.

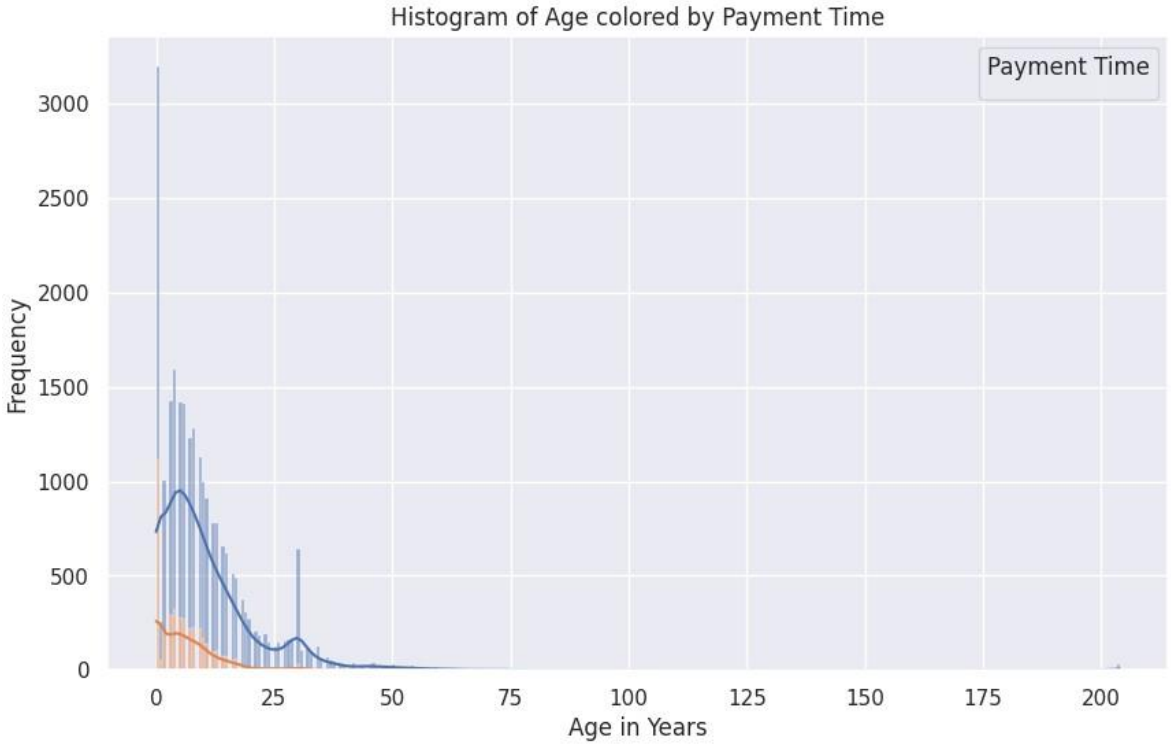


Figure 5.3: Distribution of Payment Defaults across Age

Figure 5.4 illustrates the distribution of payment defaults by the number of employees a taxpayer has. It was observed that the payment default rate increased as the number of employees increased. This pattern suggests that while larger companies may have more

resources and capabilities, their size and complexity can also introduce challenges that increase the risk of payment defaults if not managed effectively. This is consistent with other studies, such as Battaglini et al. (2022), which noted that the size of a business, as measured by the number of employees and its logarithm, is valuable in predicting tax evasion.

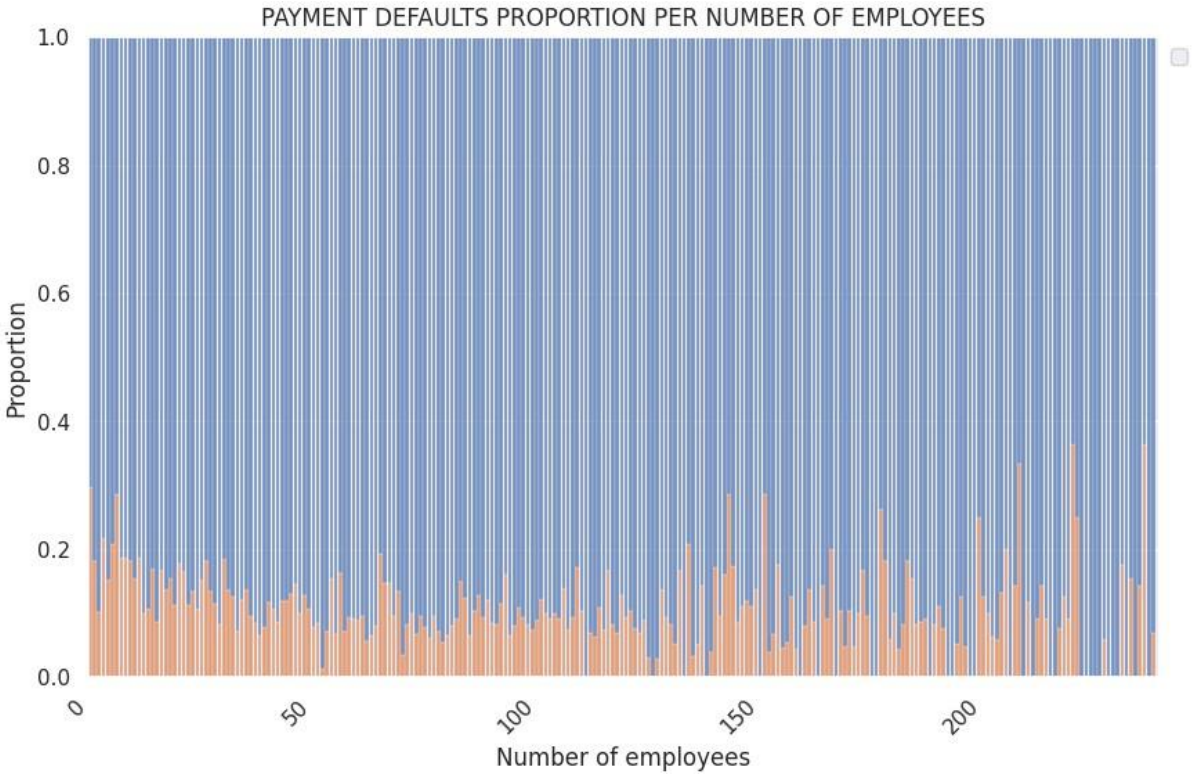


Figure 5.4: Distribution of Payment Defaults by Number of Employees

Tax authorities can take measures, such as conducting tax audits and compliance reviews, to mitigate the risk of payment defaults. Enforcement actions against companies that fail to meet their tax obligations, including imposing penalties and other sanctions can be put in place. This can serve as a deterrent to non-compliance and encourage companies to prioritize tax payments.

Taxpayer education and stakeholder engagement tailored to the needs of larger businesses also plays a key role to enhance awareness and understanding of tax obligations. Measures for instance, the tax amnesty program currently in place in KRA will go a long way address potential areas of non-compliance. By implementing these measures, tax authorities can help mitigate the risk of payment defaults and ensure compliance with tax laws and regulations, thereby safeguarding government revenue and promoting a fair and equitable tax system.

Focusing on the correlation matrix table (Figure 5.5), it is noted that a fairly high positive correlation exists between gross turnover and total expenses (0.8). This indicates that as the gross turnover of a business increases, its total expenses also tend to increase proportionally. This suggests that businesses with higher turnover tend to incur higher expenses, which is a common trend in various industries.

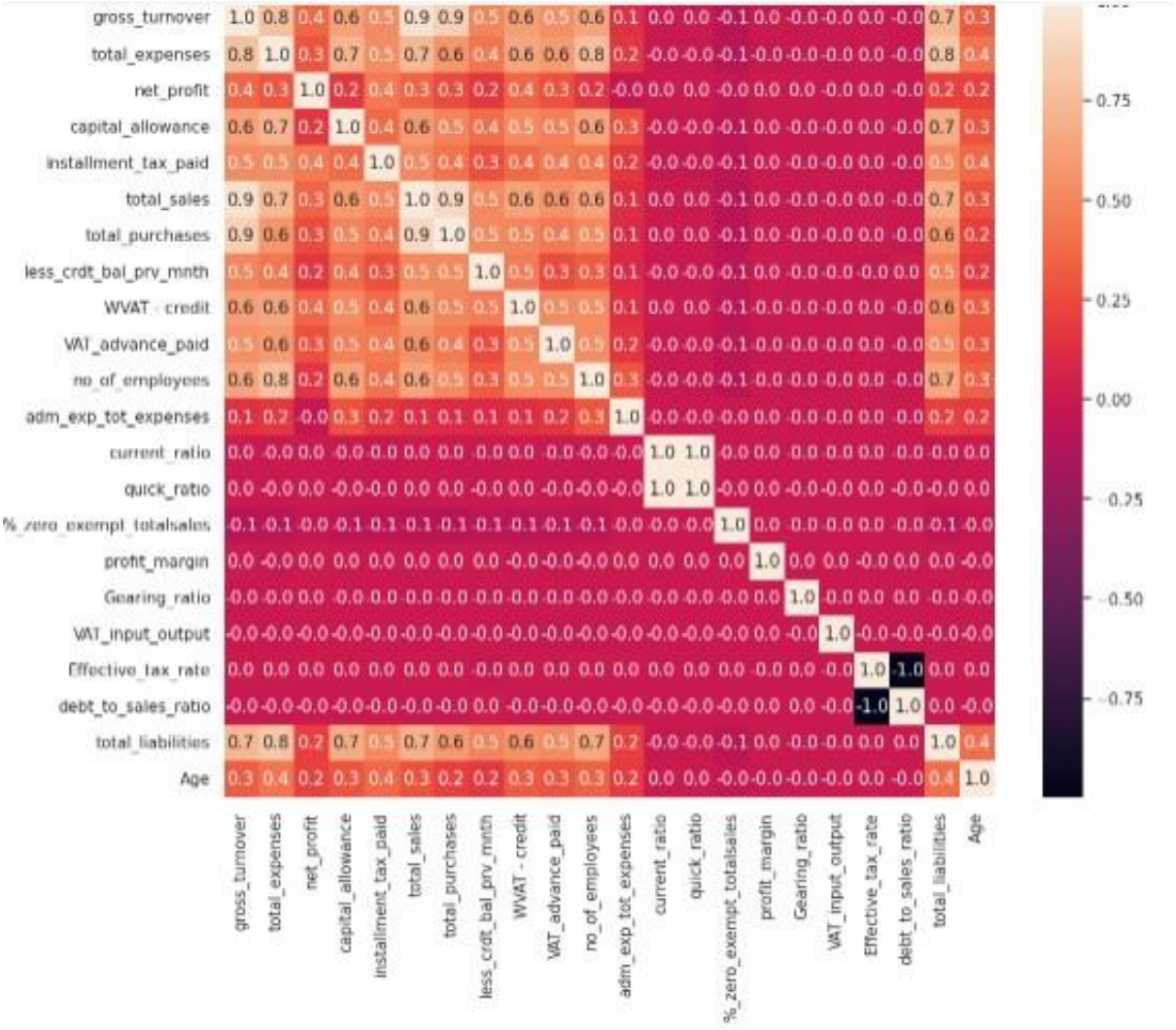


Figure 5.5: Correlation Analysis

Likewise, there is a strong correlation of (0.6) between gross turnover and number of employees which implies that as the gross turnover increases, businesses tend to employ more workers. The strong positive correlation of (0.7) between gross turnover and total liabilities suggests that as the gross turnover of a business increases, its total liabilities also tend to increase. This provides insight into a company's financial leverage and risk exposure. It suggests that higher levels of debt financing may amplify returns for shareholders during periods of growth but can

also increase financial risk, particularly if the company faces challenges in meeting debt obligations or sustaining revenue growth.

Additionally, a moderate correlation is observed between age and total expenses, as well as total liabilities (0.4), suggesting a relationship between the age of the taxpayer and their financial obligations. Older taxpayers may have higher expenses and liabilities compared to younger ones due to factors such as longer operating histories, accumulated debt, or larger operational scale. Understanding this correlation can help in segmenting taxpayers based on their financial profiles and risk factors.

These correlations provide valuable insights into the financial dynamics of taxpayers and can inform the development of predictive models for identifying risky taxpayers. In conclusion, the correlations within the dataset reveal that taxpayers with higher turnover and those that have been in business longer have a healthy business are financially stable thus are less likely to default on their tax payments.

5.2 Machine Learning Modeling and Performance Evaluation

The best performing model was chosen by evaluating four metric scores outlined in section 3.7. A comparison of the metrics for each model when applied to the dataset is presented in Table 2 below.

Table 2: Scores for the metrics of the machine learning algorithms

	precision	***recall***	***f1-score***	***accuracy***
Logistic Regression Classifier	0.66	0.68	0.67	0.67
Decision Tree Classifier	0.81	0.83	0.82	0.82
Random Forest Classifier	0.90	0.86	0.88	0.88
XGBoost Classifier	0.89	0.85	0.87	0.87
Support Vector Classifier	0.62	0.77	0.68	0.65
Stacking Classifier	0.89	0.86	0.88	0.88

.

The RF model had the highest precision (0.90) followed by Stacking and XGBoost (0.89), Decision Tree (0.81), Logistic (0.66), and the lowest was SVM (0.62). A high precision indicates a low rate of false positives. For recall scores, RF, and Stacking showed the highest (0.86) then XGBoost (0.85), Decision Tree (0.83), SVM (0.77) and Logistic had the lowest (0.68). A high recall indicates a low rate of false negatives. The RF and Stacking models had

the highest F1-score (0.88), followed by XGBoost (0.87), Decision Tree (0.82), SVM (0.68), and the lowest was Logistic (0.67). This is a useful metric when you want to consider both FP and FN. For accuracy, RF and Stacking showed the highest (0.88), then XGBoost (0.87), Decision Tree (0.82), Logistic (0.67), and SVM had the lowest (0.65). Higher accuracy is generally better, but it may not be the only metric to consider especially when dealing with imbalanced datasets.

Notably, the RF classifier consistently outperformed other models across all metrics, demonstrating its superiority in this task. Stacking also performs well, but RF exhibits slightly higher precision. The trend across different models suggests that ensemble methods like RF, Stacking and XGBoost tend to perform better than individual models such as Logistic Regression and Decision Tree. The findings corroborate those of Didimo et al. (2020) who asserted that RF demonstrated superior performance in identifying positive and adverse outcomes of a fiscal audit.

SVM and Logistic Regression consistently shows the lowest performance in all metrics, indicating potential limitations in their predictive capabilities on the given dataset. XGBoost and Decision Tree show decent performance, with scores falling between Logistic and RF. In line with expectations, the XGBoost, Stacking, and RF models exhibit higher accuracy compared to the Decision Tree classifier.

This observation suggests that for this kind of data, the ensemble tree methods, encompassing bagging, boosting and stacking techniques, demonstrate superior performance. The utilization of ensemble methods appears to enhance the predictive capabilities of the models, showcasing the effectiveness of combining multiple decision trees to handle the complexities present in the dataset. This supports the findings of Abedin et al. (2021), demonstrating that XGBoost and a methodically constructed ensemble of several Decision Trees outperform other ML methods in in accuracy and various classification performance metrics.

Figure 5.6 depicts the performance of the different models visually.

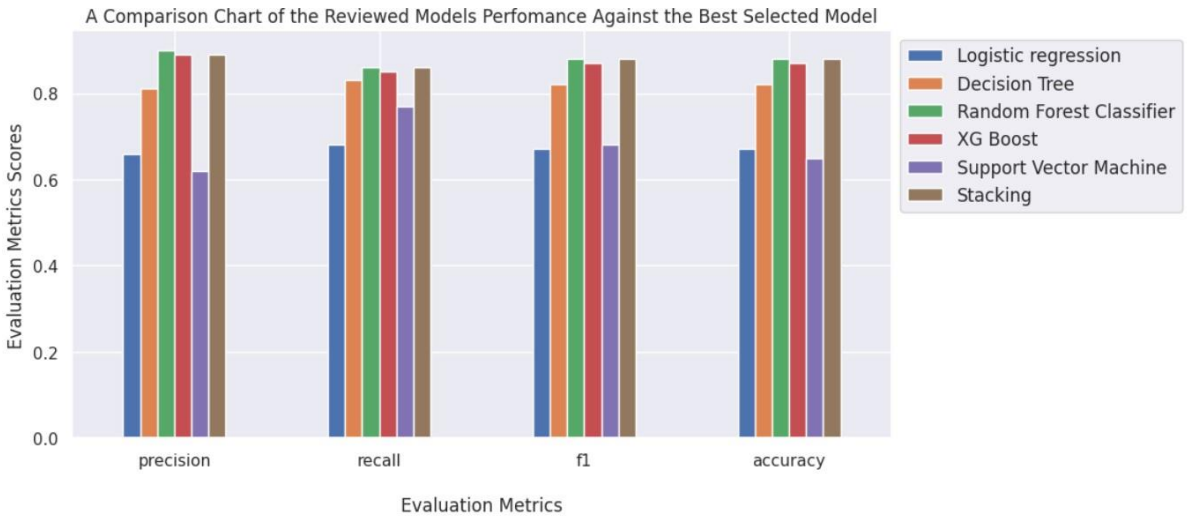


Figure 5.6: Visual representation of the performance metrics for the trained machine learning algorithms

Consequently, this study designated the best performing algorithm RF, exhibiting precision, recall, F1 and accuracy scores of 90%, 86%, 88%, and 88% respectively. This study is confident in the RF classifier's ability to discern and predict the classes of the taxpayers to be selected for audit. Further analysis and potential adjustments to the model, such as tweaking thresholds or feature engineering, could enhance its performance in predicting tax evasion across all risk levels.

The results of the evaluation were then displayed using a confusion matrix, as shown in Figure 5.7. The results are detailed as below.

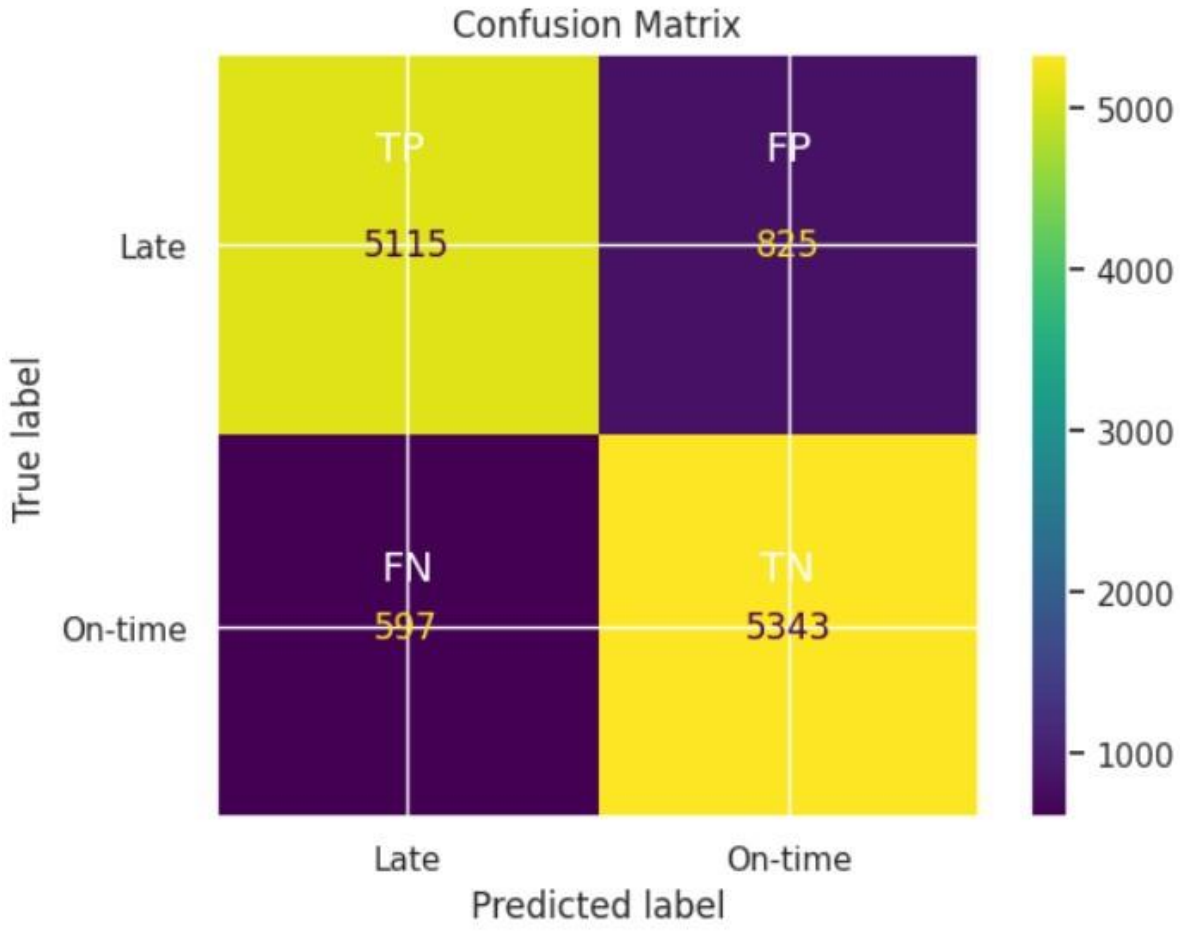


Figure 5.7 Confusion Matrix

Class 0 (Late):

True Positives (TP): 5115 - Instances of Class 0 that were correctly predicted as late payment. True Negatives (TN): 5343 - Instances correctly predicted as not belonging to Class 0. False Positives (FP): 825 - Instances not belonging to Class 0 but predicted as late payment. False Negatives (FN): 597 - Instances of Class 0 that were incorrectly predicted as on-time payment.

Class 1 (On-time):

True Positives (TP): 5343 - Instances of Class 1 that were correctly predicted as on-time payment. True Negatives (TN): 5115 - Instances correctly predicted as not belonging to Class 1. False Positives (FP): 825 - Instances not belonging to Class 1 but predicted as on-time payment. False Negatives (FN): 597 - Instances of Class 1 that were incorrectly predicted as late payment.

Additionally, the AUC-ROC for the RF model was drawn as seen in Figure 5.8 below having a score of 0.953. This depicts a highly accurate and reliable model for the given task. It suggests that ensemble tree-based models typically perform well for this type of data.

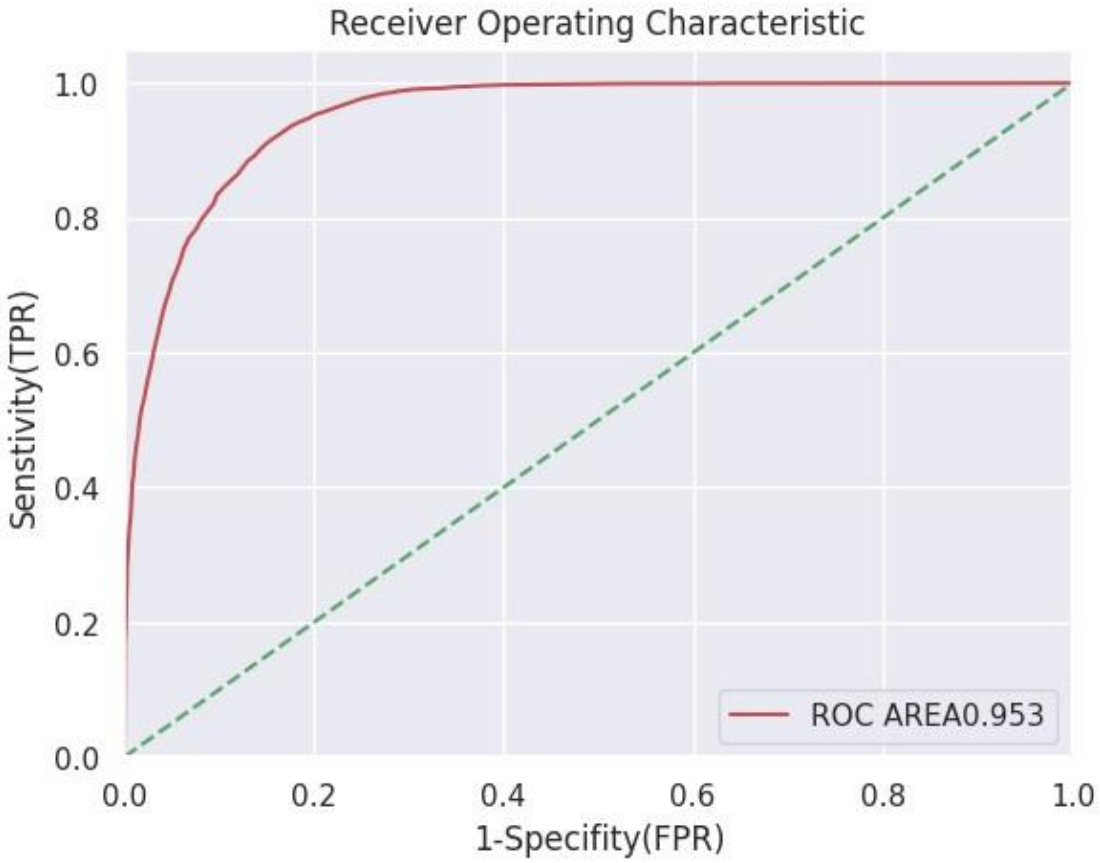


Figure 5.8: AUC-ROC

Similarly, Figure 5.9 below depicts the Precision-Recall Curve which has a AUC score of 0.91. This shows that the RF performs well, especially in situations where class imbalance is present.

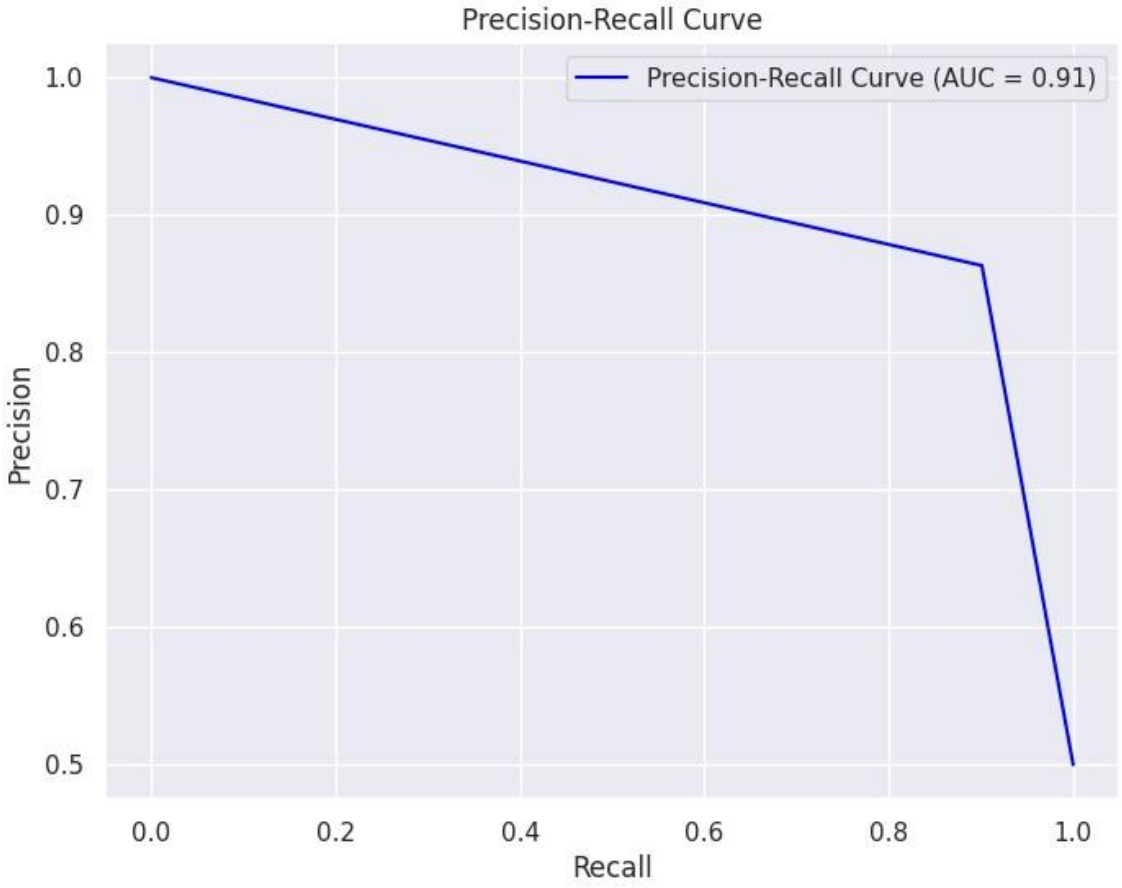


Figure 5.9: Precision-Recall Curve

5.2.1 Ranking of Features

Feature importance facilitated the examination and comparison of the significance of different features among diverse observations in the dataset. Based on the evaluation metrics, RF was the most robust model in taxpayer risk prediction. Therefore, this model was employed to identify the significance of each feature in predicting taxpayer risk. Table 3 illustrates the top 15 important features from the RF model in taxpayer risk prediction.

Every feature exhibits nonzero importance, yet their ranking varies from the most crucial to the least significant. These are installment tax paid, total liabilities, credit brought forward, WHVAT credit, total expenses, age, effective tax rate, VAT advance paid, total sales, service sector, wholesale and retail trade sector, net profit, total purchases, VAT input output and gross turnover. The trend seems to be decreasing from installment tax paid to gross turnover, with varying magnitudes.

Table 3: Feature Importance

Rank	***Feature***	***Importance***
1	Installment tax paid	0.087215
2	Total liabilities	0.056071
3	Credit brought forward	0.050981
4	WHVAT credit	0.050055
5	Total expenses	0.044703
6	Age	0.043853
7	Effective Tax Rate	0.043379
8	VAT advance paid	0.040157
9	Total Sales	0.039530
10	Service Sector	0.038579
11	Wholesale and Retail trade sector	0.037821
12	Net Profit	0.035437
13	Total Purchases	0.034227
14	VAT input output	0.033878
15	Gross turnover	0.035400

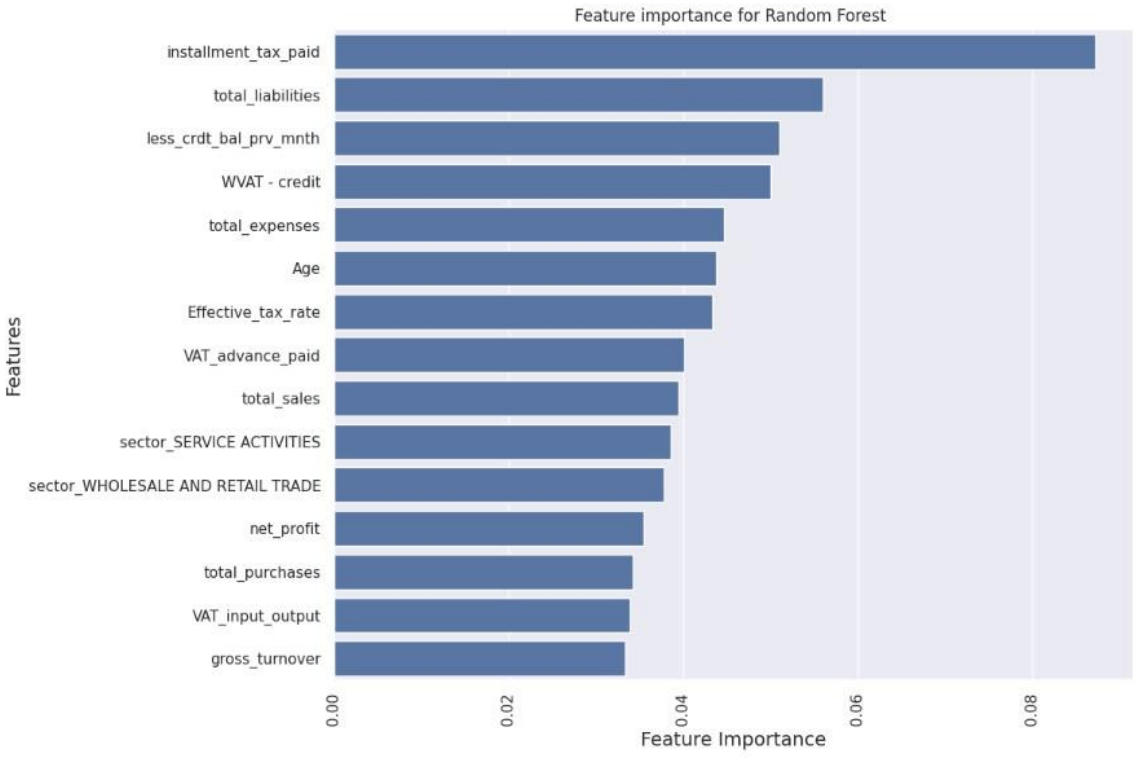


Figure 5.10: Important Features that predict taxpayer classes

The results suggest that the most important factor is installment tax paid. This implies that taxpayers who consistently make installment tax payments on time may have a lower risk of defaulting. The second most important feature is total liabilities. This implies that the amount of liabilities a taxpayer has plays a crucial role in taxpayer risk prediction. Higher total liabilities indicate financial strain and a higher risk of default. Indeed, research has pointed out that companies with lower levels of debt were observed to be less prone to tax payment failures (Abedin et al., 2021).

The credit brought forward which is the third factor could be taken to imply that taxpayers who consistently utilize available credit balances may have a lower risk of payment defaults. The fourth factor that is crucial in taxpayer risk prediction is the WHVAT credit. This implies that active claiming of the credits could depict low risk. This aligns with prior research findings indicating that VAT credits serve as one of the key indicators of tax evasion (Battaglini et al., 2022).

On the other hand, the fifth most important feature was the total expenses. This is a key indicator of a taxpayer's operations and, consequently, their risk level. Interestingly, the gearing ratio did not feature as a key factor impacting risk assessment yet it is an important aspect of financial management and can impact a company's profitability. This could imply that risk assessment tends to focus on a broader set of factors that directly affect the financial health and stability of a business.

As a decision support system, the algorithms can guide the tax authority to adjust input variables in the order of their importance for optimal case selection. The features with higher importance scores are crucial in predicting taxpayer risk. These could be used to develop a predictive model where a combination of these features influences the overall risk assessment. Understanding the impact of each feature allows for better-informed decisions in risk management and compliance monitoring.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

The study's findings offer valuable insights into tax compliance prediction, sector-wise default distribution, and the correlation between various financial indicators. Notably, ensemble methods like RF, Stacking and XGBoost demonstrated superior performance in predicting taxpayer risk, with installment tax paid, total liabilities, credit brought forward, WHVAT credit and total expenses identified as key features influencing risk assessment. Precise predictions of default behaviour enable better allocation of resources, allowing tax authorities to focus on taxpayers most likely to default. This leads to a more efficient use of funding, tax personnel, and other resources. Moreover, the findings from default prediction models can guide managerial decisions, such as system enhancement, policy changes and mitigation strategies. Understanding the factors that influence taxpayer default empowers tax authorities to make data-driven decisions, effectively addressing these issues and enhancing compliance levels. These findings underscore the potential of advanced analytics to inform tax authorities' decision-making processes. This is by enabling more targeted interventions to address compliance challenges and optimize revenue collection efforts, thus fostering a fair and equitable tax system.

6.2 Recommendations

Building upon the study's findings, tax authorities should prioritize awareness campaigns and risk-based audits targeting sectors with high default rates, such as Construction and ICT. The Government should also facilitate the Construction sector by paying pending bills which a pain area for the industry. This will go a long way in improving the sectors' performance. Tailored taxpayer education programs and flexible payment arrangements can assist startups in navigating tax obligations effectively, while enforcement actions and stakeholder engagement efforts can mitigate risks associated with larger companies. Additionally, leveraging ensemble methods in predictive modeling and incorporating key features identified in this study can enhance the accuracy and efficiency of tax compliance prediction systems. The RF classifier ranked the studied input variables in terms of importance, highlighting the installment tax paid

as the most crucial parameter to adjust. This implies that aligning input variables in accordance with this ranking is essential, as it contributes significantly to higher prediction.

6.3 Future Works

Future research endeavors could focus on refining predictive models by incorporating additional data sources, such as engagement in Government projects, business ownership – sole proprietorship, partnership, or company and socio-economic indicators. This will allow for a comprehensive exploration of how combined features impact the determination of taxpayer compliance, ultimately improving and optimizing the audit case selection process. Furthermore, exploring the impact of regulatory changes and economic factors on tax compliance dynamics could enrich our understanding of the underlying drivers of taxpayer behaviour and inform policy recommendations for promoting voluntary compliance. Longitudinal studies tracking taxpayer compliance over time could provide insights into the effectiveness of intervention strategies and identify emerging trends in tax evasion behaviour. Another avenue for further research in this paper is on utilizing techniques like deep learning for default prediction. This approach can assist tax authorities in tailoring their strategies and recommendations to suit the unique characteristics and behaviour of each taxpayer.

REFERENCES

Abedin, M. Z., Chi, G., Uddin, M. M., Satu, M. S., Khan, M. I., and Hajek, P. (2021). Tax default prediction using feature transformation-based machine learning. *IEEE Access*, 9:19864–19881.

Baghdasaryan, V., Davtyan, H., Sarikyan, A., and Navasardyan, Z. (2022). Improving tax audit efficiency using machine learning: The role of taxpayer's network data in fraud detection. *Applied Artificial Intelligence*, 36(1):2012002.

Battaglini, M., Guiso, L., Lacava, C., Miller, D. L., and Patacchini, E. (2022). Refining Public Policies with Machine Learning: The Case of Tax Auditing. NBER Working Papers 30777, National Bureau of Economic Research, Inc.

Brownlee, J. (2019). A tour of machine learning algorithms. machine learning mastery.

Casalicchio, G., Molnar, C., and Bischl, B. (2018). Visualizing the feature importance for black box models. *ArXiv*, abs/1804.06620.

Chalye, Z. T. (2020). Developing Tax Payer Fraudulent Risk Level Prediction Model for Auditing Using Hybrid Machine Learning Techniques. PhD thesis.

Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chepkwony, C. C. (2017). Auditees case-selection model for evaluating taxpayer corporate tax compliance in Kenya.

Didimo, W., Grilli, L., Liotta, G., Menconi, L., Montecchiani, F., and Pagliuca, D. (2020). Combining network visualization and data mining for tax risk assessment. *IEEE Access*, PP:1–1.

Fadlil, A., Herman, and M, D. (2022). K nearest neighbor imputation performance on missing value data graduate user satisfaction. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6:570–576.

FiscalisRiskAnalysisProjectGroup (2016). Risk Management Guide. For Tax Administrations (FRAPG). Technical report, European Commision.

Geremew, B. (2017). Mining e-filing data for predicting fraud: The case of ethiopian revenue and customs authority.

Holtzblatt, J. and Engler, A. (2022). Machine Learning and Tax Enforcement. Technical report, Tax Policy Center.

Hotz, N. (2018). What is CRISP DM? Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Huang, S., Cai, N., Pacheco, P., Narrandes, S., Wang, Y., and Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer genomics proteomics*, 15:41–51.

IRS (2018). Internal Revenue Service Strategic Plan FY2018-2022. Technical report, Internal Revenue Service.

Kasibwa, A. and Allela, A. (2023). Digital Transformation: The Emerging Use Of Artificial Intelligence.

Kenyhercz, M. and Passalacqua, N. (2016). Missing Data Imputation Methods and Their Performance With Biodistance Analyses, pages 181–194.

Kifordu, B. C. (2021). Machine learning in tax administration: A case study of Nigeria.

Lozano, A., Cezar, A., Lozano, G., and Ippolito, A. (2021). Tax crime prediction with machine learning: A case study in the municipality of São Paulo.

Mabe-Madisa, G. (2018). A decision tree and naïve bayes algorithm for income tax prediction. *African Journal of Science, Technology, Innovation and Development*, 10(4):401–409.

Mamo, D.(2013). Application of data mining technology to support fraud detection protection: the case of Ethiopian revenue and custom authority.

Mburu, G. (2022). Understanding Tax Evasion. Technical report, Kenya Revenue Authority.

Mughal, M. and Akram, M. (2012). Reasons of tax avoidance and tax evasion: Reflections from pakistan. *Journal of Economics and Behavioral Studies*, 4.

Morris, T., and Lonsdale, M. (2005). Translating the compliance model into practical reality. Wellington, New Zealand: Inland Revenue.

Murorunkwere, B. F., Haughton, D., Nzabanita, J., and Kipkogei, F. (2023). Predicting tax fraud using supervised machine learning approach. *African Journal of Science, Technology, Innovation and Development*, 15:1–12.

Murorunkwere, B. F., Tuyishimire, O., Haughton, D., and Nzabanita, J. (2022). Fraud detection using neural networks: A case study of income tax. *Future Internet*, 14(6):168.

Murphy, R. (2019). The European Tax Gap; A report for the Socialists and Democrats Group in the European Parliament. Technical report, Tax Reasearch LLP.

Naganandhini, S. and Shanmugavadivu, P. (2019). Effective diagnosis of alzheimer's disease using modified decision tree classifier. *Procedia Computer Science*, 165:548–555. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP TIV INNOVATION, 2019 November 11-12, 2019.

OECD (2004a). Compliance Risk Management: Audit case Selection Systems. Center for tax policy and administration, Organization for Economic Cooperation and Development.

OECD (2004b). Compliance Risk Management: Managing and Improving Tax Compliance. Center for tax policy and administration, Organization for Economic Cooperation and Development.

OECD (2021). Tax Administration 2021: Comparative Information on OECD and Other Advanced and Emerging Economies.

OECD (2023). Revenue Statistics in Africa 2023. Technical report, Organization for Economic Cooperation and Development.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,

Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Pijnenburg, M., Kowalczyk, W., andvanDijk, L.H. (2017). A roadmap for analytics in taxpayer supervision. *Electronic Journal of e-Government*, 15:19–32.

Pérez López, C., Delgado Rodríguez, M. J., and de Lucas Santos, S. (2019). Tax fraud detection through neural networks: An application using a sample of personal income taxpayers. *Future Internet*, 11(4).

Rahimikia, E., Mohammadi, S., Rahmani, T., and Ghazanfari, M. (2017). Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran. *International Journal of Accounting Information Systems*, 25:1–17.

Rokach, L. (2010). Ensemble-based classifiers. *Artif. Intell. Rev.*, 33:1–39.

Russell, S. J. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Pearson, 3rd edition.

Ruzgas, T., Kižauskienė, L., Lukauskas, M., Sinkevičius, E., Frolovaitė, M., and Arnastauskaitė, J. (2023). Tax fraud reduction using analytics in an east European country. *Axioms*, 12(3).

Schneider, F. (2015). Tax Losses due to Shadow Economy Activities in OECD Countries from 2011 to 2013: A Preliminary Calculation. CESifo Working Paper Series 5649, CESifo.

Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

Serrano-Cinca, C., Gutiérrez-Nieto, B., and Bernate-Valbuena, M. (2019). The use of accounting anomalies indicators to predict business failure. *European Management Journal*, 37(3):353–375.

Shalev-Shwartz, S. and Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Sharma S, Chatterjee S. (2021). Winsorization for Robust Bayesian Neural Networks. *Entropy*, 23(11):1546. https://doi.org/10.3390/e23111546

Shipton, D. E. C. (2015). Sitting on the bench: an exploratory study into Inland Revenue's industry benchmarking programme.

Sirengo, J. (2018). Assessing taxpayer compliance risk identification criteria at the Kenya Revenue Authority.

UNDP (2022). Tax for SDGs Initiative. Technical report, UNDP, New York, NY 10017, USA.

Vellutini, C. (2011). Key principles of risk-based audits. *Risk-Based Tax Audits*, page 13.

WorldBankGroup (2020). Paying Taxes 2020. Technical report, PwC,Washington, DC, USA.

Wu, R.-S., Ou, C., Lin, H.-y., Chang, S.-I., and Yen, D. (2012). Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*.

Yego, N. K., Kasozi, J., and Nkurunziza, J. (2021). A comparative analysis of machine learning models for the prediction of insurance uptake in Kenya. *Data*, 6(11).

APPENDICES

Appendix A: Ethical Approval

Strathmore UNIVERSITY

8th February 2024

Ms Cheboi Cynthia,
cynthia.jemutai@strathmore.edu

Dear Ms Cheboi,

RE: Predicting Risky Taxpayers in Kenya Using Machine Learning

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC2001/24.** The approval period is from **8th February 2024 to 7th February 2025.**
This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix B: NACOSTI permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **195425**

Date of Issue: **20/February/2024**

RESEARCH LICENSE

This is to Certify that Ms.. Cynthia Jemutai Cheboi of Strathmore University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in on the topic: Predicting Risky Taxpayers in Kenya Using Machine Learning for the period ending : 20/February/2025.

License No: **NACOSTI/P/24/33204**

195425

Applicant Identification Number

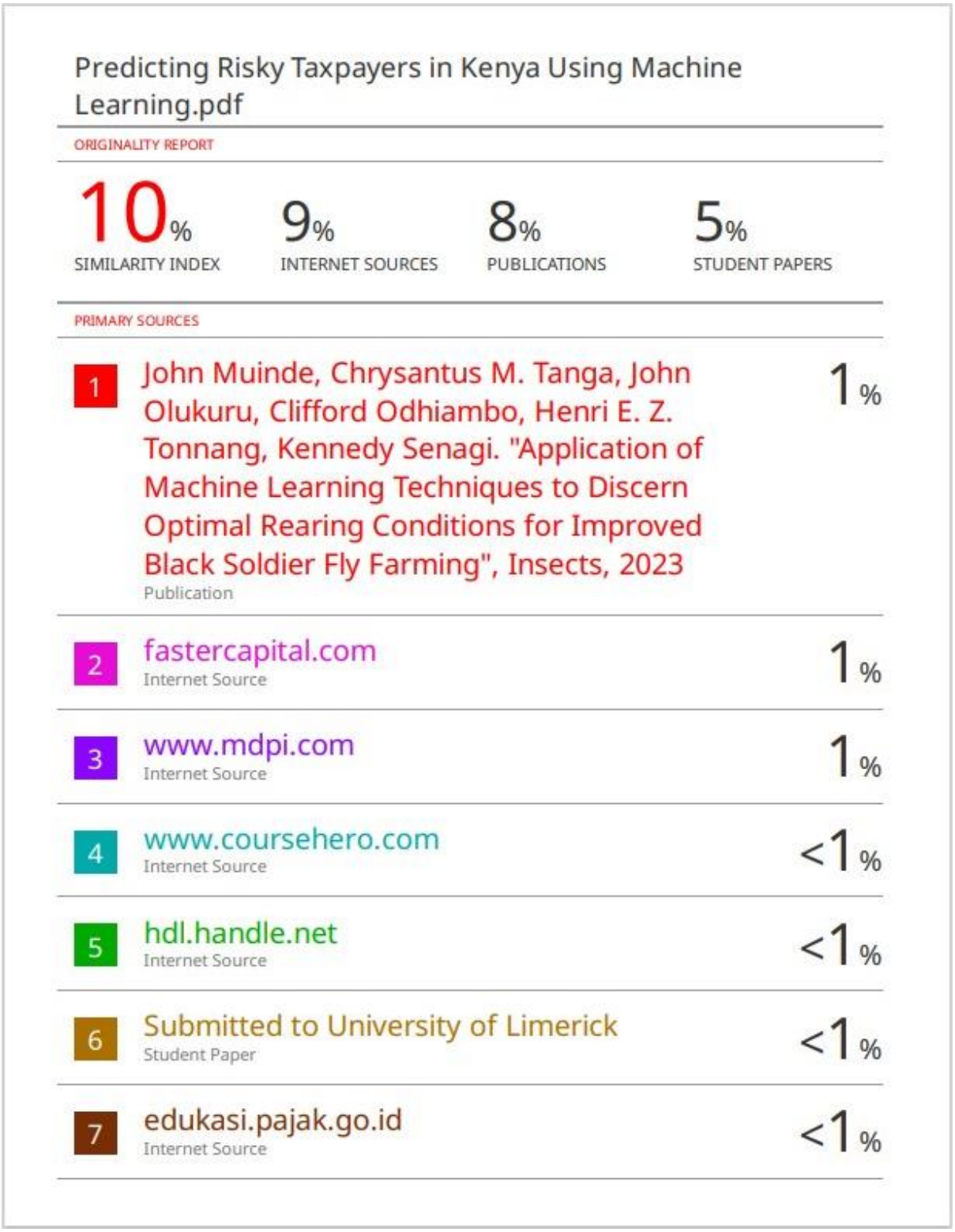
Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix C: Turnitin Report



Predicting Risky Taxpayers in Kenya Using Machine Learning.pdf

ORIGINALITY REPORT

10%	9%	8%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	John Muinde, Chrysantus M. Tanga, John Olukuru, Clifford Odhiambo, Henri E. Z. Tonnang, Kennedy Senagi. "Application of Machine Learning Techniques to Discern Optimal Rearing Conditions for Improved Black Soldier Fly Farming", Insects, 2023 Publication	1%
2	fastercapital.com Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	www.coursehero.com Internet Source	<1%
5	hdl.handle.net Internet Source	<1%
6	Submitted to University of Limerick Student Paper	<1%
7	edukasi.pajak.go.id Internet Source	<1%