Tax Fraud Prediction Using Machine Learning Models in Kenya

Onyango Calvince Ogutu

149481

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Data Science and Analytics Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June 2025

Declaration

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

Name: **Onyango Calvince Ogutu**

Signature: 

Date: May 22, 2025

Approval

The dissertation of Onyango Calvince Ogutu was reviewed and approved by the following:

Dr. Evans Otieno Omondi

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Prof. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

Abstract

With the rapid technological advancements in Kenya, the tax base has expanded, resulting in an increase in tax fraud cases. Detecting and preventing tax fraud has become a priority for tax agencies to maximize revenue and ensure compliance. This study focused on developing a robust tax fraud prediction model using machine learning techniques. Our approach involved training and evaluating multiple models, including Logistic regression, which was deployed as the baseline model for prediction. Key features such as age, business turnover, total turnover, and total financing expenses were identified and engineered to enhance predictive accuracy. The Random Forest model demonstrated superior performance in identifying fraudulent transactions, achieving notable precision and recall rates of 0.96 and 0.77 respectively. Additionally, exploratory data analysis (EDA) revealed significant patterns that contributed to the understanding of tax fraud behavior. This study highlights the effectiveness of machine learning in accurately detecting tax fraud and provides insights into the most influential factors driving fraudulent activities. Our findings support the application of predictive models for improving fraud detection efficiency in tax systems.

KEY WORDS: *Tax Fraud, Fraud Prediction, Machine Learning Algorithms, Classification.*

Table of contents

List of tables

List of abbreviations

GDP	Gross Domestic Product	KNN	K-Nearest Neighbor
VAT	Value Added Tax	PAYE	Pay As You Earny
AD	Anomaly Detection	ML	Machine Learning
CCTV	Closed-Circuit Television	SVD	Singular Value Decomposition
LLE	Locally Linear Embedding	MDS	Multidimensional Scaling
MLP	Multi-Layer Perceptron Neural Network	SRCNN	Spectral Residual Convolutional Neural Network
LOC	Locally Outlier Factor	FWAD	Fixed Width Anomaly Detection
GANs	Generative Adversarial Networks	CART	Combination For Regression Tree
CHAID	Chi-Squared Automatic Interaction Detector	PIT	Personal Income Tax
INTA	Iranian National Tax Administration	AI	Artificial Intelligence

Acknowledgement

I would like to express my sincere gratitude to the following individuals and organizations whose support and guidance have been invaluable in the preparation and development of this research dissertation:

I am deeply grateful for the unwavering support and expert guidance provided by Dr. Evans Omondi. His insightful feedback and encouragement has been instrumental in shaping the direction of this research.

I extend my thanks to Strathmore University for providing the necessary resources and facilities essential for conducting this research. The academic environment has been conducive to intellectual growth and exploration.

Lastly, I want to express my deepest gratitude to my family and friends, for their unwavering support, understanding, and encouragement throughout this research endeavor. Their belief in my abilities has been a constant source of motivation.

Dedication

To the infinite potential of data science, driving my curiosity and dedication to unraveling its complexities.

Chapter 1

Introduction

1.1 Background of the study

Tax fraud occurs when an individual or business entity deliberately provides false information on a tax return with the intention of minimizing their tax liability (Hasseldine and Morris, 2013). Essentially, tax fraud involves attempting to deceive the tax system by submitting inaccurate information in order to evade the full tax obligation. Instances of tax fraud include falsely claiming deductions, categorizing personal expenses as business-related, employing a fake Social Security number, and neglecting to report income (Mackevičius and Kazlauskienė, 2009).

In the past, addressing issues related to fraud or failures in taxpayer compliance predominantly involved regulatory enforcement and audit-centric methods. However, recent years have seen revenue administrators acknowledge the diverse and intricate determinants of taxpayers' compliance behavior in specific risk areas (Devos, 2013). It has become clear that successfully dealing with these factors necessitates a more nuanced and comprehensive strategy. Relying solely on regulatory enforcement actions, such as manual audits, is inadequate, given the complexity of compliance issues, which often requires a more multifaceted approach. Manual audits have faced challenges in efficiently analyzing large volumes of transactions and data. Human auditors have found it difficult to identify complex patterns and relationships between various entities and transactions (Singh and Best, 2016). Manual audits are time-consuming, and the results may not be available in real-time. Furthermore, they are resource-intensive, demanding significant time and effort from human auditors, who may introduce subjectivity and inconsistencies into the audit process.

The aim of this study therefore was to apply machine learning models in predicting tax fraud. Machine learning models offer solutions to the challenges associated with manual tax audits by handling large volumes of data, identifying complex patterns, providing real-time detection, adapting to changing patterns, reducing false positives, improving resource efficiency, and ensuring consistency and objectivity in the audit process.

This chapter provides an introduction to the study commencing with a discussion on the background and context, Subsequently, it addresses the research problem, outlines the research aims, objectives, and questions, delves into the significance of the study, and concludes by acknowledging the limitations.

It is a paradox that while tax rates are rising in Kenya, Tax to GDP ratio has declined from 17.4% in 2017 to 15.3% in 2022. Taxation serves as an essential method for generating revenue, enabling governments to furnish essential goods and services (Mabe-Madisa, 2018). These taxes include individual income taxes imposed at the national level and corporate taxes imposed on companies. Adjustments are made to account for specific deductions and credits permitted by the law, resulting in the computation of income tax for corporate entities (Cobham and Janskỳ, 2018). Although tax regulations vary with time and changing political administrations, the law is a constant in the society (Politou et al., 2019). Additional tax categories include input and output taxes, such as Value Added Tax (VAT), as well as employment or payroll taxes like PAYE, Withholding, and Excise taxes.

Fraud takes place in many types of financial transactions, including taxation (Schneider, 2013). According to Section 97 of the Tax Procedures Act, tax fraud involves actions such as excluding amounts that should be reported on a tax return, claiming unwarranted relief or refunds, providing inaccurate information that impacts tax obligations, creating or altering financial records dishonestly, or intentionally failing to meet any legal tax-related requirements.

While cases of financial fraud have received significant media attention, their scale and impact are dwarfed by the concealed and almost tolerated expense associated with tax fraud (Garner et al., 2016). Tax fraud encompasses a wide array of intricate strategies that human creativity can devise (Faccia and Mosteanu, 2019). These strategies are employed by individuals to

gain an unfair advantage over others through false representations or the concealment of the truth. It includes tactics that involve surprises, cunning maneuvers, trickery, and any dishonest means used to deceive government (Albashrawi, 2016). The core components of fraud are the presence of deception and the intention to deceive.

1.2 The research problem

The increasing sophistication of fraudulent activities poses significant challenges to traditional fraud detection methods, primarily characterized by their reliance on static rules and predefined patterns (Buoni, 2012). These methods struggle to adapt to evolving fraud tactics, resulting in high false positive rates and an inability to effectively identify novel or unknown patterns of fraud. Additionally, the manual-intensive nature of traditional processes leads to delays in detection and response.

While the Tax Procedures Act outlines different categories of fraud, it does not guide the methods for uncovering it. Fraud Detection involves a wide range of approaches, both direct and indirect. Gathering evidence is a fundamental step in establishing the occurrence of fraud, and the evidence's quality and quantity should align with the investigator's objectives. Machine learning can serve as a valuable tool to assist tax professionals, accounting firms, and government entities in their efforts to uncover fraudulent activities. Tax authorities have traditionally depended on manual case selection and batch-applied, risk-based models at the financial year's close to detect discrepancies. With the advent of machine learning, there's now an exciting potential to shift to almost real-time detection and reporting of irregularities in tax data to the government

1.3 Research objectives

1.3.1 General objective:

The study aimed to build and apply machine learning algorithms to detect and predict cases of fraud using a dynamic and adaptive approach.

1.3.2 Specific objectives:

1. To Conduct a comparative analysis of predictive accuracy between the Logistic regression classifier and the Random Forest, Decision Trees, KNN, Gaussian Naive bayes and Multi layer Perception classifier on fraud prediction, aiming to determine the best performing algorithm.
2. To develop a fraud detection model using the best-performing algorithm, tailored for use in the Kenya Revenue Authority's taxation system..
3. To optimize the performance of the machine learning algorithms to improve predictive accuracy.

1.4 Significance of the study

While existing literature on fraud detection using machine learning has made significant strides, there are limited industry specific Studies that focus on fraud detection and prediction in the tax administration. Many existing studies focus on general applications of machine learning in fraud detection, but there is a need for more industry-specific research. Different sub-sectors in the finance industry face distinct challenges and require tailored approaches. This study will contribute to the body of knowledge on tax fraud detection and prediction by building models that are robust and adaptable to the rapidly and constantly changing fraud mechanisms. This study will help address the current research gap in this domain, offering tangible value to revenue administrations navigating dynamic environments of fraud.

1.5 Limitations of the study

Limited Generalization: Models trained on specific datasets obtained from KRA itax systems may struggle to generalize well to other revenue agencies in other parts of the world, where the data structures are different. This limitation can affect the model's applicability to diverse fraud detection scenarios. Secondly, fraudulent instances are often rare compared to legitimate transactions, resulting in imbalanced datasets. Imbalances can lead to biased model training, where the model might prioritize the majority class and perform poorly in identifying instances of fraud.

1.6 Outline of the study

In the first chapter, the study's context has been introduced, outlining the identified research objectives and questions while highlighting the significance of such research. Additionally, the study's limitations have been addressed. Moving to the second chapter, an examination of existing literature will be undertaken, to provide a comprehensive overview of the current research landscape in tax fraud detection using machine learning, offering insights into methodologies, challenges, and emerging trends in the field.

In the third chapter, the research methodology choices will be detailed. Specifically, the methodology section provides a comprehensive and transparent overview of the research process in tax fraud detection using machine learning, enabling other researchers to replicate and validate the study's findings. The Key areas will include: research design, data collection, model development, and evaluation process.

Chapter 2

Literature Review

2.1 Introduction

All governments, regardless of scale, are affected by tax fraud, which occurs when individuals or entities knowingly misrepresent information on tax returns to lower their tax obligations (Yamen et al., 2023). This fraudulent activity reduces government revenue and consequently limits public expenditures. A common form of tax fraud is under-reporting, where a taxpayer reports a reduced tax base on their return. To detect such instances, supervised machine learning is frequently applied. This approach, relying on labeled data, is currently the most popular among researchers in audit-supported cases (Vanhoeyveld et al., 2020). As noted by Choi et al. (2018), the success of anti-fraud efforts in taxation will rely on the development of advanced machine learning algorithms and effective stakeholder engagement.

2.2 Supervised machine learning models

Researchers have employed various fraud detection methods based on the dataset used, with Logistic Models, Bayesian Networks, Artificial Neural Networks, and Decision Trees being the most commonly utilized approaches in tax fraud detection. These methods have proven effective in addressing the challenges associated with identifying and categorizing false data, as highlighted by (Vasco et al., 2021). A review by Abdallah et al. (2016) on the significance of data mining techniques across various financial sub-areas for tax fraud detection concluded that supervised ML Models, such as Logistic Regression, Support Vector Machines, Decision Trees, Neural Networks, and Bayesian Networks, consistently yielded optimal classification results.

Henke and Jacques Bughin (2016) conducted a comprehensive seven-year study characterizing taxpayers with false invoices using data mining techniques. Their research employed clustering methods to identify groups with similar behavior, followed by the application of neural networks, Bayesian Networks, and Decision Trees to uncover factors associated with tax fraud, identify behavior patterns, and assess the comparison with available information. The study demonstrated the feasibility of classifying and identifying taxpayers utilizing fake invoices based on information regarding their tax payment, business characteristics, and past performance. The correct identification rate for tax fraud cases was 86%, with a corresponding rate of 84% for large and medium enterprises. Additionally, the research suggested that taxpayers audited multiple times without any detected fraud are less likely to commit tax fraud in the future.

Numerous studies have explored the application of supervised machine learning for income tax fraud detection, incorporating comprehensive historical and socio-demographic information. Notably, a combined approach utilizing naïve Bayesian Network and Decision Tree for income tax prediction on a large dataset demonstrated effectiveness in tax fraud prediction(Mabe-Madisa, 2018). Comparisons of various machine learning models have been employed by researchers to identify and predict suspicious activities in tax and other domains, with variable contributions observed, highlighting the robustness of specific models (Hilas et al., 2015). Additionally, some studies leverage clustering and classification algorithms to construct fraud profiles, aiding in audit planning (Görtler et al., 2022). Researchers argue that a supervised machine-learning approach is practical for tax fraud detection, recommending the application of different machine learning methods and selecting the most robust one for large datasets encompassing various tax types (Pérez López et al., 2019).

Recently, a study Murorunkwere et al. (2022) focused on detecting income tax fraud in Rwanda using Artificial Neural Networks achieved an impressive accuracy rate of 92%. The findings indicated the presence of income tax fraud in both domestic and cross-border businesses. Notably, the study underscored the importance of business operating hours as a key factor in identifying tax fraud, showing that smaller businesses are more frequently implicated in tax fraud than larger ones. Furthermore, businesses located in certain districts

of Kigali city were found to have a higher risk of income tax fraud. The study suggested the development of an automated model along with a dashboard to aid in detecting income tax fraud (Murorunkwere et al., 2022).

2.3 Unsupervised machine learning models

Deep learning can be broadly classified as either unsupervised or supervised learning, depending on whether the data includes labels. Labeled data come with specific tags or targets that denote their features, attributes, or classifications (Karayiannis and Venetsanopoulos, 1992). Unlabeled data, in contrast, lack such tags and only consist of descriptive features. Supervised learning models are trained with labeled data to predict the correct labels for new, unseen data. Key applications of supervised learning include sentiment analysis on Twitter using word embeddings (Gulmeher and Aiman, 2023), object detection in CCTV footage with convolutional neural networks, and predicting students' academic outcomes through neural networks (Giannakas et al., 2021). In unsupervised learning, however, the model lacks predefined tasks and instead autonomously identifies important patterns and latent structures within the data.

While labeled data offer distinct advantages over unlabeled data, they also require extensive human effort for annotation, converting raw data into labeled datasets and demanding considerable resources (Adadi, 2021). Unlabeled data, on the other hand, are widely available, less costly to obtain, and easier to source. For example, identifying non-compliant taxpayers using data on property acquisitions involves labeling records as "fraudulent" or "non-fraudulent"—a process that requires tax experts, such as auditors, to conduct reviews, making labeling time-consuming and resource-intensive (Algan and Ulusoy, 2021).

Unsupervised learning offers a valuable approach for analyzing property acquisition data, with clustering and dimensionality reduction as two primary methods. Clustering analysis is effective when grouping tax returns filed by property owners with similar traits (Verbeeck et al., 2020). On the other hand, when the focus is on identifying unusual cases within a large dataset of tax returns, dimensionality reduction is particularly useful. This technique

reduces the dataset to a lower-dimensional form while preserving the core variability (Beutel et al., 2015). This lower-dimensional form, often referred to as "distilled" information, retains significant details while discarding unnecessary aspects, allowing for efficient data compression. It can then be expanded back into the "reconstructed" original dataset (Zhu et al., 2018). Within property acquisition data, cases with uncommon attribute combinations—like a high-end commercial building with an unusually low purchase price—are likely to produce the highest reconstruction errors, whereas more typical cases are expected to have smaller errors (Jensen et al., 2014).

Various algorithms support dimensionality reduction, such as principal component analysis (PCA), singular value decomposition (SVD), locally linear embedding (LLE), multidimensional scaling (MDS), Isomap, and autoencoders (Abonyi and Feil, 2007). PCA and SVD are linear methods that transform data by linearly combining original variables, while LLE, MDS, Isomap, and autoencoders provide nonlinear transformations that capture more complex patterns. Nonlinear dimensionality reduction is often preferred for real-world datasets due to inherent nonlinear relationships (Sumithra and Surendran, 2015).

Reconstruction errors are widely used in identifying anomalies across various fields (Lee, 2022). For instance, Imoniana et al. (2013) leveraged reconstruction errors in image analysis to identify outlier pixels against a background, while Sumithra and Surendran (2015) utilized them in video surveillance to detect rare or unexpected events within specific contexts.

In one investigation, tax returns obtained from the Indian tax authority were employed to identify VAT fraud. The study employed various supervised methods, including classification trees, logistic regression, discriminant function analysis, and a hybrid approach, with the hybrid method demonstrating superior performance (Vanhoeyveld et al., 2020). Pérez López et al. (2019) focused on personal income tax returns in Spain and utilized Multi-Layer Perceptron neural network (MLP) models. The research incorporated a diverse set of features from personal income returns. His results showed that (MLP) model had an efficiency performance rate of 84.3% According to (Alsadhan, 2023), Indian tax returns were utilized alongside random forests to identify deceptive entities such as bill traders or shell companies. The researchers estimated the potential revenue-saving impact of their model by testing it on

historical data unseen by the model, reporting an estimated recovery of US$40 million in revenue.

Unsupervised models, particularly anomaly detection (AD) techniques, show significant promise in the realm of tax fraud detection. These models identify outliers characterized by features that significantly deviate from the majority of the population (Al-Hashedi and Magalingam, 2021). AD techniques have the advantage of utilizing the entire population, as opposed to a small and often biased labeled subset. This enables the identification of novel fraud methods that exhibit behavior distinct from legitimate activities. However, the effectiveness of AD techniques relies on certain assumptions used to distinguish outliers from regular data, and different AD approaches operate based on distinct assumptions (Sanni, 2019). Typically, two fundamental assumptions are crucial for the proper functioning of AD approaches: First, it is assumed that the number of anomalies is much lower than the number of normal occurrences, a condition met in the case of tax fraud. Second, anomalies are expected to exhibit behavior in terms of feature representation that sets them apart from normal instances. While certain fraud tactics may attempt to conceal their behavior and mimic lawful actions, this requirement seems to hold in the context of tax fraud (Daho and Chikh, 2015).

Anomaly detection (AD) methods have found applications in diverse domains, including electricity consumption fraud, financial statements analysis, and disease-drug relationships. In a study by Edjabou and Smed (2013), focusing on electricity consumption fraud, researchers examined data from electricity consumption meters, identifying reading errors and manipulated data by consumers. They employed two anomaly models, namely Spectral Residual-Convolutional Neural Network (SRCNN) and an anomaly-trained model based on martingales, to detect unusual usage trends for further inspection during on-site visits by utility providers. Another study by Demirhan (2024), utilized the Mahalanobis distance to assess anomalies in the financial statements of Vietnamese companies, calculating the distance between each data point and the distribution's centroid. In the medical field, Li et al. (2022) quantified disease-drug relationships, applying an anomaly detection method using a

fully connected neural network with sparse convolution and introducing a focal-loss function to address data imbalance.

In the context of tax fraud detection, AD methods have been less commonly employed. A study by Zheng et al. (2023) used Belgian VAT declarations to conduct sector profiling and applied two AD methods, namely Local Outlier Factor (LOC) and Fixed-Width Anomaly Detection (FWAD). Another study in the same domain utilized a statistical outlier detection strategy, outperforming the supervised model of the Kazakhstani tax administration.

In Baghdasaryan et al. (2022) the authors employ taxpayer-specific features for tax fraud detection, exploring the impact of taxpayer network data on fraud identification. Utilizing data from the Armenian tax authority, they suggest that the network relationships between suppliers and buyers can contribute to fraud prediction. In Xavier et al. (2022), researchers predict taxpayer compliance levels by assessing multiple indicators using Support Vector Machines (SVM). Despite the positive outcomes observed in fraud detection in the mentioned literature, supervised algorithms still face challenges related to overfitting and generalization. Unsupervised methods, on the other hand, are limited by their inability to incorporate audit outcomes, diminishing their effectiveness. To address these limitations, ensemble methods are proposed to enhance tax fraud detection performance by mitigating the drawbacks associated with both supervised and unsupervised approaches.

This study aimed to distinguish normal events from fraudulent ones, making it a binary classification task. Unlike other classification types, this research does not explore multiclass or multilabel classification. Multiclass classification involves assigning one of several possible classes, while multilabel classification applies when an event can belong to multiple categories simultaneously. Multilabel classification builds on single-label classification, encompassing both binary and multiclass setups where each event is uniquely assigned a single label.

2.4 Autoencoder and the apriori algorithms

Future identification of fraudulent cases can employ various advanced approaches. For instance, He et al. (2021) applied clustering techniques paired with probability distributions within clusters to detect tax fraud, while Cheng et al. (2021) used a random walk algorithm and social network analysis to expose telecom fraud. In telecom fraud, perpetrators typically defraud victims by persuading them to transfer money into fraudulent accounts. Recent studies indicate an increasing focus on deep learning-based methods to improve anomaly detection, with notable approaches including generative adversarial networks (GANs), restricted Boltzmann machines (RBMs), and autoencoders.

Some studies have utilized the autoencoder algorithm to identify fraudulent cases, leveraging its design as a data compression model comprising an encoder and a decoder. The encoder compresses the input into a smaller, more efficient representation, while the decoder reconstructs this encoding back to its original form (Schreyer et al., 2017). To evaluate the autoencoder's effectiveness in representing the data, a loss function, such as mean squared error or cross-entropy, is applied to measure the reconstruction loss. This loss penalizes the model for producing outputs that significantly differ from the input (Muhammad et al., 2020). By compressing the data, the encoder prioritizes key information, discarding less relevant details before the decoder restores the compact encoding to approximate the original input.

The autoencoder has found valuable applications across fields such as image generation, information retrieval, and anomaly detection (Lee, 2022). Unlike traditional dimensionality reduction methods like principal component analysis (PCA), which generally perform less effectively in data compression (Hasseldine and Morris, 2013), autoencoders provide a substantial advantage. PCA, which uses a linear transformation to project data onto a lower-dimensional space, often falls short of capturing the complex nonlinear relationships present in real-world data. In practice, input variables rarely have purely linear interactions. Thus, autoencoders excel in compressing data into a low-dimensional latent space by effectively modeling nonlinear structures within the data (Hilas et al., 2015).

Another benefit of utilizing the autoencoder is its capability to address the challenge of highly unbalanced samples, characterized by a skewed distribution of labels. This issue is exemplified in the classification of fraudulent and non-fraudulent cases, as seen in Louppe (2014) examination of credit card fraud detection. In this analysis, the dataset included a total of 284,807 credit card transactions, with only 492 identified as fraudulent (18%). The substantial imbalance in labels is a common occurrence in tasks involving the detection of fraudulent cases. Even with significant resource investment to obtain labels, supervised learning algorithms relying on labeled data are prone to suboptimal performance due to the highly skewed distribution of labels. Supervised learning algorithms typically achieve optimal results when operating on a balanced dataset, ideally with a 50:50 proportion for each case. The autoencoder circumvents this extreme imbalance problem, making it well-suited for identifying extremely rare outliers among a large number of normal cases.

A key challenge emerges when using an autoencoder to calculate reconstruction error—a metric that can indicate whether a case is likely normal or fraudulent. Specifically, when true labels are absent, there are limited tools to verify if the computed reconstruction errors are reasonable. Research has taken two approaches on this front. In the first type, studies had access to true labels at the outset, allowing them to evaluate the results of unsupervised learning by applying validation metrics like the Rand index or purity, particularly with clustering algorithms, a common unsupervised approach (Lee, 2022). However, such metrics are only feasible when true labels are available. The second approach applies to studies that initially lack labels but acquire them later through resource-intensive efforts, such as annotation through crowdsourcing (Lee, 2022). Still, previous research has not proposed an efficient method to confirm the reliability of reconstruction errors. This study aims to fill this gap by utilizing the Apriori algorithm, a widely used technique in data mining.

Niksa-Rynkiewicz et al. (2020) introduced the Apriori algorithm, designed to identify frequent item sets within databases. Initially, it pinpoints frequently occurring individual items and progressively expands them into bigger item sets. Mostly applied in marketing, the Apriori algorithm is employed to discover products likely to be bought together, with associated products often jointly promoted in marketing channels (Hegland, 2007). In the context of this

study, the Apriori algorithm is employed to identify frequent items, specifically properties frequently filed for acquisition tax. For instance, if the Apriori algorithm identifies any common (frequent) property in the property acquisition data and presents it to the autoencoder, the resulting reconstruction error will be small. Conversely, introducing an uncommon property to the autoencoder will yield a significant reconstruction error, as the autoencoder struggles to reconstruct rare (infrequent) properties. This study distinguishes itself from prior research by validating the reasonableness of reconstruction errors through this approach.

Furthermore, most of the studies mentioned above, predominantly focused computer vision (Lee, 2022), where unstructured datasets such as videos and images were extensively utilized. This study saught to enhance previous research by applying an autoencoder to the tax auditing business. In other words, unlike preceding studies, this research employs an unsupervised deep learning algorithm on structured data.

2.5 Other related works

Numerous publications discussing fraud detection (FD) through the application of data mining methods have been examined. Researchers have demonstrated significant interest in investigating FD across various domains, including banks, healthcare, finance, and taxation.

2.5.1 Tax fraud detection

According to Faccia and Mosteanu (2019) numerous tax authorities have embraced data mining methods to assess the tax compliance of their taxpayers. Despite being a subject of significant interest, various constraints have been identified for in-house projects. The taxpayer archive, being labeled and within the purview of internal auditors, is often treated as confidential information due to compliance risks. Malaszczyk and Purcell (2017) in their model for business tax fraud detection, introduced a colored network-based approach to address the limitations of traditional data mining-based tax evasion detection methods.

However, the developed method, incorporating heterogeneous network information, was noted for its time-consuming and labor-intensive nature.

Yang and Shami (2020) used hybrid algorithm to design a fraud detection for financial statements. The combination for regression trees (CART) with Chi-squared automatic interaction detector (CHAID) was adopted to identify the key variables. However, both had limitations as they were not able to compute the continuous numerical data. Murorunkwere et al. (2023) addressed collinearity and imbalanced class issues in statistical fraud detection, particularly modeling VAT fraud probabilities using logistic regression. The study utilized a dataset from the Norwegian Tax administration and tested logistic regression with ridge and elastic net. Ridge, not performing variable selection, used the highest number of covariates. Elastic net encountered challenges in determining a uniquely best value, introducing uncertainty and demanding more computational effort.

Vasco et al. (2021) implemented and estimated a model for Personal Income Tax (PIT) to identify and detect PIT evasion. They employed Multilayer Perceptron (MLP). A significant challenge in their approach was the presence of a large number of 11 multilinear independent variables, coupled with a considerably imbalanced distribution of the target variable. Wu et al. (2012) implemented a hybrid intelligent system using Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Logistic Regression (LR) for the Iranian National Tax Administration (INTA) with a focus on detecting corporate fraud. The proposed method primarily aimed at binary outcomes, but it emphasized the practical usage of probability outcomes. Pérez López et al. (2019) employed neural networks to create taxpayer groups and calculate the probability of a single taxpayer evading taxes. However, challenges arose during testing, preventing the inclusion of all available variables in the model, and the computational speed was identified as a significant concern.

2.5.2 The common frauds in tax administrations

The two most common tax fraud and evasion schemes which appear to be more widespread in their use include: under-declaration of income turnovers through sales suppression as well

as over-claiming of expenses through false and fictitious invoicing (Stiglitz, 1985). These schemes are relatively simple for tax evaders to achieve and can affect different economies regardless of their sizes (Desai and Dharmapala, 2009). These schemes can further be aided by a cash economy or/and a sharing economy. The magnitude of tax fraud and evasion scheme is huge, with anecdotal indicators showing that many billions of dollars of tax revenue are not collected by revenue authorities (Taylor and Richardson, 2012). As added by Dalu et al. (2012) a country facing an increasing amount of tax fraud and evasion is likely to experience lower investments in production, resulting into poor economic growth rate and adverse impact on the provision public goods and services.

Tax systems are designed with a primary focus on addressing the requirements of every segment of society aiming to achieve the following key objectives: Revenue generation: Taxation serves as a means to generate funds that are essential for financing public services, infrastructure development, and administrative functions. Redistribution of Wealth (Posner, 1971). Taxation seeks to mitigate disparities in both vertical (between individuals) and horizontal (between groups) income and wealth inequalities thus promoting a fairer and more equitable society (Elkins, 2006). Repricing Externalities: Taxation can be utilized to reprice certain activities that have adverse societal effects, such as the consumption of tobacco and the emission of carbon (Edjabou and Smed, 2013). By doing so, it aims to discourage such activities and promote more sustainable practices. Enhancing Representation: Taxation plays a role in fostering healthier democratic processes. A greater reliance on tax revenues for government spending is associated with improved governance quality and political representation (Kyriienko and Magnusson, 2022). Therefore, fraud in tax poses a significant threat to these objectives and undermines government efforts in multiple ways. It disrupts the equilibrium in all sectors of the economy, eroding public finances and hindering the government's ability to fund essential services and infrastructure. Moreover, tax fraud perpetuates economic imbalances and inequalities, further exacerbating disparities within society (Shafer et al., 2020). Therefore, addressing tax fraud is crucial to maintaining the integrity of tax systems and ensuring they effectively serve the needs of all citizens and society as a whole.

2.5.3 Fraud detection mechanisms

Traditional fraud detection mechanisms typically rely on rule-based systems with predefined patterns to identify potential fraudulent activities. These systems utilize established rules, thresholds, and triggers to flag transactions that deviate from expected patterns (Edge and Sampaio, 2012). Manual review by human analysts is often part of the process, where expertise is employed to assess flagged transactions for potential fraud. Behavioral analysis involves examining historical behavior to establish a baseline for normal activities and identify deviations. Additional components may include signature verification, geographical analysis (O'Sullivan et al., 2022), and the use of blacklists and whitelists. While some traditional systems may incorporate basic machine learning algorithms, their adaptability to evolving fraud tactics is limited. Despite their historical use, these mechanisms have drawbacks, leading to a growing shift towards more advanced technologies like machine learning and data analytics for more efficient and adaptive fraud detection in modern financial systems.

The main drawbacks with traditional fraud detection mechanisms include: Limited Adaptability: Traditional systems are typically static and may not adapt well to changes in fraud tactics. Fraudsters continually develop new strategies, and rule-based systems can become outdated quickly, leading to an increased risk of false negatives or missed fraudulent activities. High False Positive Rates: Rule-based systems may generate a high number of false positives, flagging legitimate transactions as potentially fraudulent (Malekian and Hashemi, 2013). This can result in increased operational costs and additional manual review efforts to distinguish between false positives and actual fraud. Inability to Detect Unknown Patterns: Traditional mechanisms are often unable to detect novel or unknown fraud patterns since they rely on predefined rules (Aziz and Andriansyah, 2023). As fraud tactics evolve, the system may not have the capability to identify emerging threats that do not match existing rule sets. Reactive Nature: Traditional fraud detection is often reactive, identifying fraud after it has occurred. This reactive approach may lead to delays in response time, allowing fraudulent activities to persist or escalate before detection. Difficulty in Handling Unstructured Data: With the growth of digital transactions, unstructured data such as text and images become

important for fraud detection. Traditional systems may find it challenging to analyze and extract meaningful insights from unstructured data sources. Human Error: Manual review processes, which are often part of traditional fraud detection, are prone to human error. Analysts may overlook crucial details or misinterpret patterns leading to both false positives and false negatives (Massa and Valverde, 2014). High Maintenance Costs: Maintaining and updating rule-based systems can be expensive. As fraud tactics change, frequent updates and adjustments to rules are necessary, contributing to higher maintenance costs over time. And, Rule-Based Systems: Traditional fraud detection often relies on rule-based systems that use predefined rules to identify suspicious activities. These systems may struggle to keep up with evolving fraud tactics.

The private investigators employ a range of methods to uncover instances of tax fraud. These techniques include surveillance, background checks, forensic accounting, and electronic data analysis (Manning, 2010). Surveillance entails closely observing the activities of individuals or businesses suspected of engaging in tax evasion. Background checks involve a thorough investigation into the financial history and assets of these entities. Forensic accounting is the process of scrutinizing financial records to identify irregularities and inconsistencies (Imoniana et al., 2013). Electronic data analysis involves the examination of electronic data, including emails and financial records, to reveal evidence indicative of tax evasion. The cost incurred on the private investigators are quite exorbitant. Secondly, private investigators, being human, are susceptible to errors, much like everyone else. There is a risk of misinterpreting information, overlooking crucial details, or being misled by false leads. As a result, the information provided by private investigators may not always be entirely accurate or reliable. In extreme cases, there is a risk that a private investigator might resort to fabricating information to meet a client's expectations. While this is not a widespread practice among professional investigators, it is a potential risk.

A range data analytics techniques and algorithms play a crucial role in effectively detecting and mitigating tax fraud risks. These methods assist agencies in sifting through large volumes of data to identify patterns, relationships, and anomalies that may signify fraudulent activities. Key algorithms commonly utilized include: Anomaly Detection Methods: These

methods focus on identifying instances that deviate significantly from expected patterns or behaviors (Zheng et al., 2023). Techniques may involve statistical measures like standard deviations or percentiles, as well as machine learning algorithms like one-class Support Vector Machines (SVMs) or Isolation Forests (Cheng et al., 2021). Detecting unusual transactions or data points helps organizations flag potential fraud cases for further investigation. Pattern Recognition: This technique involves analyzing data to identify recurring patterns, trends, or relationships indicating fraudulent transactions (Dutta et al., 2017). Techniques such as association rule learning or sequence mining assist in identifying common fraud schemes or behaviors that warrant closer scrutiny. Machine Learning Algorithms: This approach helps refine predictive models for fraud detection. Popular algorithms include clustering (e.g., K-means, DBSCAN), regression analysis (e.g., logistic regression), and neural networks (e.g., deep learning, recurrent neural networks). Leveraging these algorithms allows organizations to continually enhance their fraud detection capabilities and adapt to evolving risks.

The advancements in data analytics and artificial intelligence (AI) have shifted the focus from merely detecting fraud to proactively preventing it (Aziz and Andriansyah, 2023). Predictive analytics and machine learning techniques play a crucial role in recognizing potential fraud cases before they occur. Analyzing historical fraud data helps identify patterns indicative of deceptive behavior, enabling organizations to anticipate risks and target dubious activities proactively. Proactive Prevention: Predictive models facilitate the proactive identification of suspicious activities by incorporating variables like transaction volume, velocity, and customer behavior patterns to estimate fraud likelihood. This proactive approach minimizes fraud exposure and optimizes prevention strategies (Assylbekov et al., 2016; Baghdasaryan et al., 2022; Cobham and Janskỳ, 2018). Resource Allocation: With insights gained from predictive models, organizations can allocate resources more effectively. This allows them to take preventive measures and act to prevent fraudulent activities before they materialize, enhancing overall risk management strategies. In a nutshell, the integration of data analytics and AI models in fraud detection not only identifies potential fraud but also enables organizations to adopt proactive measures to prevent fraudulent activities, improving overall risk management and resource allocation strategies.

2.5.4 Credit card fraud detection

Behera and Panigrahi (2015) introduced an innovative algorithm for fraud detection, utilizing fuzzy c-means to identify patterns in cardholders' behavior in relation to their historical transactions. The designed approach effectively reduced misclassification scores based on feature occurrences; however, it did not assess the time gap between transactions. Behera and Panigrahi (2015) explored diverse techniques for credit card fraud detection, focusing on the implementation of the Hidden Markov Model (HMM). In this application, three classes (low, medium, and high) were established based on transaction frequencies in terms of amounts to categorize customer profiles. Probabilities related to monetary values were assigned to each client. Although the developed system demonstrated faster fraud detection, it did not provide experimental evidence for incoming transactions.

2.6 Proposed Tax fraud detection technique

2.6.1 Logistic regression classifier

Logistic regression is often considered ideal for detecting and classifying fraudulent taxpayers due to its inherent characteristics that align well with the nature of fraud detection problems (Fischer et al., 1992). Logistic regression is a binary classification algorithm, making it suitable for scenarios where the objective is to classify instances into two categories, such as fraudulent and non-fraudulent taxpayers. Moreover, logistic regression provides probabilities for the predicted outcomes, allowing for a nuanced understanding of the likelihood of fraud. This is crucial in fraud detection, where the consequences of false positives and false negatives can have significant implications. The algorithm's interpretability is another advantage, as it allows for the examination of the contribution of each feature in predicting fraud, aiding in the identification of key indicators. Logistic regression is well-suited for scenarios with a moderate to large number of features, making it applicable to tax-related datasets with multiple variables that may contribute to the identification of fraudulent activities (Bolton, 2009). Due to the aforementioned reasons, Logistic regression will be the used as the baseline

model in this study. Five other models will be deployed for comparative analysis of the model performance.

The logistic regression equation for binary classification is given by:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}. \tag{2.1}$$

where: P(Y=1|X) is the probability that the taxpayer is fraudulent given the features X. e is the base of the natural logarithm. β_0 is the intercept term, representing the baseline log-odds of fraud when all predictors are zero And $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, indicating the change in log-odds for a one-unit change in each predictor.

2.7 Research knowledge gaps

The present study aims to address gaps summarised in Table 2.1

Table 2.1: Research gaps

Study	Findings	Gaps
Pérez López et al. (2019)	The study focused on fraud detection using MLP on income tax. The findings of the MPL model enabled taxpayer segmentation as well as calculation of the probability concerning an individual taxpayer's propensity to attempt to evade taxes.	The selected model exhibited an efficiency rate of 84.3%. This study aims to improve the prediction accuracy to at least 90%.
Demirhan (2024)	The study on financial anomalies adopted a machine-learning approach using Mahalanobis distance to assess anomalies in financial statements. The findings showed that 89.61% of the data points were categorized as normal, while 10.39% were identified as anomalies.	While the study demonstrated the efficacy of machine learning algorithms in detecting anomalies in fraud, there is a need for further research to explore its application in the context of tax administration. This study, therefore, aims to develop a machine-learning model tailored specifically to tax focusing on fraud.
Baghdasaryan et al. (2022)	The study was about improving tax audit efficiency focusing on the role of the taxpayer network. The study utilized specific taxpayer features contained in the supplier and buyer network to detect fraud in tax. The research delved more into the importance of feature incorporation.	The study emphasized more on feature incorporation. However, there is a need to further investigate the impact of other features such as business turnover, chargeable income, and income declared.
Choi et al. (2018)	A study on the artificial intelligence approach to financial fraud underscored the need for technical advancement in predictive modeling and effective communication with all stakeholders to develop effective actionable strategies.	Despite the emphasis on effective communication by the study, a gap still exists in the literature on communicating technical findings to non-technical stakeholders. This study aims to communicate technical findings to stakeholders with non-technical expertise.

Chapter 3

Methodology

3.1 Introduction

The project creates a machine learning model to predict fraudulent transactions declared by taxpayers, employing a comprehensive methodology that covers data collection, data pre-processing, feature selection, model selection, evaluation and optimization. The primary model chosen for prediction is the Logistic Regression classifier, renowned for its efficiency in binary classification tasks. In addition to Logistic Regression. Five other models are deployed alongside logistic regression for comparative analysis. These other models include; Random Forest, Decision Tree, Gaussian Naive Bayes, Multi-layer Perceptron Classifier and KNN.

3.2 Data

The project used income declaration datasets that include information on income turnovers, VAT turnovers sales transactions and purchases transactions among other variables. Data pre-processing was done to ensure the quality and suitability of the data for machine learning models. The phase included thorough data cleaning, handling of missing values, and transformation of the dataset into a format conducive to machine learning analysis. The dataset was extracted from the iTax portal. The features of the dataset are as tabulated in Table 3.1.

Table 3.1: Feature description

Variable	**Variable Description**
Station Name	Name of the Station where the tax Payer registered
Obligation Name	Taxes Applicable (Income Tax)
Filing Period	A 12-month calendar year covered by a tax return
Gross tax payable	An individual's total earnings before taxes are applied
Net Tax	The total amount of revenue less amount of expenses
Off Credits	Any financial benefit applied to reduce a tax liability
Filing date	Date Tax Payer Files his annual returns in itax
Business Turnover	Total revenue generated by the business within a year
Farming Turnover	Total revenue by farming activities within a year
Rental Turnover	Total revenue generated from renting out properties
Interest Turnover	Total revenue generated from interests earned within a year
Other Turnover	Total revenue generated from any other activity not mentioned
Total Turnover	The sum of the total revenue generated from all turnovers.
Profit loss b4tax	Amount of profits earned before tax is applied
Profit or loss After tax	Amount of profits earned after tax is applied
Chargeable Income	Refers to the income on which an individual is liable to pay taxes
Income	Category of income earned by an individual
Unsued losses bf	Amount of losses in the previous period brought forward
Taxable Income	Portion of an individual's income that is subject to taxation
Instalment Tax	A method of paying taxes in multiple installments throughout the year,
Total Credits	Cumulative amount of credits
Fraud	Binary recodrs where 1 is fraud and 0 No fraud

3.3 Exploratory data analysis

Before applying machine learning models, comprehensive data analysis is performed to gain insights into the dataset's characteristics and inform the subsequent modeling steps. The analysis included the following aspects:

1. Exploring the Dataset's Dimensions: To examine the size and structure of the dataset.

2. Descriptive Statistics for the numerical variables.

3. Univariate Analysis: We used bar charts, scatter plots and Histograms to provide insights into the distribution, central tendency, and dispersion of the following variables: Income turnovers, VAT Turnovers and sales and purchases transactions.

4. Bivariate Analysis: This will establish relationships between variables such as the relationship between VAT Turnovers and Income tax turnovers.

5. Visualizing the distribution of turnovers across different regions in Kenya.

6. A pair plot to examine the relationship of multiple numerical variables to give insights on how different financial metrics relate to income levels.

3.4 Data pre-processing

3.4.1 Feature selection

Based on feature importance, seven variables were selected for the modeling phase. The variables included: businessturnover, farmingturnover, rentalturnover, totalturnover, profitlossb4tax, chargeableincome and Income. Income was the target variable while the rest were used as predictor variables.

3.4.2 Feature encoding

The first step in data pre-processing involves encoding categorical variables to make them suitable for machine learning models. The following categorical variables were encoded: Income – Income declared by the taxpayers, encoded as Small (0) and Medium (1).

3.4.3 Feature scaling

Feature scaling was performed to ensure that numerical features have consistent scales. StandardScaler was used to transform the data such that it has a mean of 0 and a standard deviation of 1. This process is particularly useful considering the features in the dataset have different scales. The formula for standardizing a feature X using a Standard Scaler is given by:

$$Z = \frac{(X - \mu)}{\sigma} \tag{3.1}$$

where: Z is the standardized value, X is the original value of the feature, μ is the mean of the feature and σ is the standard deviation of the feature. The standardized data has a mean of 0 and a standard deviation of 1 for the feature. The benefits of using a Standard Scaler include improved convergence for certain machine learning algorithms and better interpretability when comparing the relative importance of features.

3.5 Model development

3.5.1 Data splitting

The dataset is partitioned into training and test sets, utilizing an 80-20 split ratio. This configuration enables the model to undergo evaluation on novel data, serving as a preventive measure against overfitting. We deployed six models and evaluated their performance. The following models were selected: Logistic Regression, Decision Trees, Random Forest, Gaussian Naive Bayes, KNN and Multi-layer Perceptron Classifier.

3.5.2 Logistic regression

The logistic regression is represented as:

$$\text{Log-Odds} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \tag{3.2}$$

The logistic function is applied to obtain the probability (P(Y=1)):

$$P(Y=1) = \frac{1}{1+e^{-(\text{Log-Odds})}}, \tag{3.3}$$

Where: β_0 is the intercept term, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with each independent variable, $X_1, X_2, \ldots, X_n$ are the values of the independent variables and e is the base of the natural logarithm. And P(Y=1) = the probability of a fraudulent transaction

occurring. The logistic regression classifier model produces predicted probabilities of the dependent variable being in the positive class. To make a binary prediction, a threshold of 0.3 will be applied. If the predicted probability is above this threshold, the instance is classified as 1 (fraudulent); otherwise, it is classified as 0 (Not fraudulent).

3.5.3 Random forest classifier

For a binary classification task where the target variable is (Y)fraud The random Forest process can be summarized as:

$$Y_{\text{fraud}} = \text{RandomForest}(X_1, X_2, \ldots, X_n), \tag{3.4}$$

where Y is the target variable, and $X_1, X_2, \ldots, X_n$ are the predictors.

$$\hat{y} = \begin{cases} 1 & \text{if } \frac{1}{T}\sum_{t=1}^{T} h_t(x) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

Bootstrapped Sampling: Randomly sample (with replacement) from the original dataset to create multiple subsets. Each subset is used to train an individual decision tree.

Feature Randomization: At each node of a decision tree, a random subset of features is considered for splitting. This introduces diversity among the trees.

Decision Tree Training: Each subset of data is used to train a decision tree independently.

Voting or Averaging: For classification tasks, the final prediction is often determined by a majority vote among the individual trees.

3.5.4 Multi-layer perceptron classifier

The MLP has input layers, hidden layers and output layers. The Layers are represented as follows:

$$\textbf{Hidden Layer:} \quad Z^{(1)} = X \cdot W^{(1)} + b^{(1)}, \quad A^{(1)} = \text{activation}(Z^{(1)}) \tag{3.6}$$

$$Y_{\text{predicted}} \text{ is the predicted probability distribution across income categories.} \tag{3.7}$$

Here, $Y_{\text{predicted}}$ is the predicted probability distribution across income categories.

3.5.5 Model evaluation

The following metrics were used to assess and evaluate the performance of the models. Accuracy, precision, Recall (Sensitivity or True Positive Rate), F1 Score, Specificity (True Negative Rate), False Positive Rate (FPR), Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Area Under the Precision-Recall Curve (AUC-PR).

Accuracy is a measure of how often a classification model makes correct predictions across all instances. It represents the overall correctness of the model by considering both true positive and true negative predictions. Accuracy is given by:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}} \tag{3.8}$$

Precision focuses on the accuracy of positive predictions, indicating the proportion of instances predicted as positive that are positive. It helps minimize the occurrence of false positives, reducing the likelihood of incorrect positive classifications. Precision is given by:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{3.9}$$

Recall (Sensitivity or True Positive Rate): Recall emphasizes capturing as many positive instances as possible, measuring the model's ability to find all true positives among all actual positives. It aims to minimize false negatives, ensuring positive cases are not missed. Recall is given by:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{3.10}$$

F1 Score: The F1 score is a balanced metric that considers both precision and recall. It provides a harmonic mean of the two, offering a compromise between precision and recall. It is particularly useful when there's an uneven distribution between classes. F1 Score is given by:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3.11}$$

Specificity (True Negative Rate): Specificity focuses on the accuracy of negative predictions, indicating the proportion of correctly predicted negative instances among all actual negatives. It aims to minimize false negatives in negative predictions. Specificity is given by:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}. \tag{3.12}$$

False Positive Rate (FPR): FPR measures the rate of false alarms, indicating the proportion of incorrectly predicted positive instances among all actual negatives. It helps assess the model's performance in avoiding false positive classifications. False Positive Rate is given by:

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}. \tag{3.13}$$

Area Under the Receiver Operating Characteristic Curve (AUC-ROC): AUC-ROC evaluates a model's ability to distinguish between classes across different classification thresholds. The ROC curve illustrates the trade-off between true positive rate and false positive rate, and a higher AUC-ROC indicates better model performance.

Area Under the Precision-Recall Curve (AUC-PR): AUC-PR assesses a model's precision-recall trade-off, especially beneficial in imbalanced datasets. It quantifies the area under the Precision-Recall curve, with a higher value indicating better precision-recall performance.

3.5.6 Model optimization

GridSearchCV from scikit-learn was used for hyperparameter tuning on the models. GridSearchCV performed a grid search over specified hyperparameter values. A hyperparameter grid for the regularization parameter C with values ranging from 10^{-3} to 10^{3} defined. The grid search was conducted using 10-fold cross-validation on the training data, and the models were evaluated with different Cvalues. The best-performing model was determined by the highest average performance across cross-validation folds, retrieved and stored in best_lr_model.

3.6 Fraud reduction and revenue enhancement

The project will ascertain whether the created prediction models may be used to enhance revenue collection by detecting and predicting fraudulent transactions on a near real-time. When assessing the impact on revenue enhancement, the following variables will be considered: The models' ability to accurately predict fraudulent transactions and Secondly, the models' impact on fraud reduction.

3.7 Ethical considerations

Ethical considerations in the study were crucial for ensuring the well-being of participants, maintaining confidentiality, and conducting research with integrity and responsibility. Adhering to ethical principles was fundamental to the credibility and societal impact of the research. The dissertation was sent to Strathmore University for ethical review.

Chapter 4

Data analysis results and interpretation

4.1 Introduction

This chapter presents the development and experimentation of predictive models aimed at assisting Kenya Revenue Authority (KRA) in detecting fraud in income Tax. It begins by presenting the results of a comprehensive data analysis in income tax, followed by the development of a robust fraud detection engine that classifies taxpayers based on their compliance behavior. Leveraging advanced data mining techniques, the models are tailored to the unique characteristics of Income tax fraud detection, with rigorous testing and evaluation ensuring their effectiveness in real-world scenarios.

4.2 Exploratory analysis

In Table 4.1, the model summary for fraud detection produced five signififcant predictors with their respective p values. From the summary, there were 27,498 data points. Out of these records, cases with fraud accounted for 5.7% while the rest were non-fraudulent. The results revealed that: Age was significantly associated with fraud, with p-value of 0.001. Individuals involved in fraud were slightly older, with a median age of 50 years compared to 48 years for those not involved in fraud. The p-value indicated a highly significant difference, suggesting that age may be a relevant predictor of fraud. Gross tax payable is significantly higher for those involved in fraud. The median value is substantially greater in the fraud group (Kshs. 68,961), and a highly significant p-value of less than 0.001, indicating that higher gross tax payable is associated with a greater likelihood of fraud. Business Turnover showed a somewhat counterintuitive pattern. While the feature was statistically significant ($p < 0.001$),

the median business turnover was slightly lower in the fraud group (Kshs. 1,258,043) than in the non-fraud group (Kshs. 1,800,000). Total turnover is also significantly higher for those involved in fraud. The median total turnover was slightly higher in the fraud group (Kshs. 2,924,500) than in the non-fraud group (Kshs. 2,764,675), with a highly significant p-value of less than 0.001. This suggests that total turnover is a strong predictor of fraud. Total financing expenses are markedly higher for those involved in fraud. The median financing expenses for the fraud group are significantly greater, suggesting a strong correlation between higher financing expenses and fraudulent activity.

Table 4.1: Model Summary of Feature Importance

Dependent: fraud	**No**	**Yes**	**Total**	**p**
Total N (%)	25917 (94.3)	1572 (5.7)	27489	
Age (Median [IQR])	48.0 (40.0 to 58.0)	54.0 (44.0 to 65.0)	48.0 (40.0 to 58.0)	<0.001
Gross Tax Payable (Median [IQR])	52405.0 (3487.0 to 296756.0)	68961.0 (2922.0 to 600548.8)	53000.0 (3465.0 to 309468.0)	<0.001
Business Turnover (Median [IQR])	1800000.0 (0.0 to 4471200.0)	1258042.5 (0.0 to 3178274.2)	1767009.0 (0.0 to 4391600.0)	<0.001
Total Turnover (Median [IQR])	2764675.0 (1601110.0 to 5578000.0)	2924500.0 (1662051.6 to 5679086.5)	2772876.0 (1605685.0 to 5585236.0)	0.0014
Total Financing Expenses (Median [IQR])	1146.0 (0.0 to 36400.0)	495646.0 (223486.0 to 1119470.5)	3085.0 (0.0 to 53612.0)	<0.001

4.3 Model formulation and evaluation

The project employed six machine learning algorithms to analyze the behavior of fraudulent taxpayers using various predictors. By leveraging these advanced algorithms, we can comprehend the complex relationships and underlying anomalies in the data to accurately identify fraud.

4.3.1 Logistic regression

A logistic regression model was developed to analyze the characteristics of both fraudulent and non-fraudulent cases, aiding in the classification and prediction of fraud. This logistic model served as the baseline for the project, against which the performance of the other five machine learning models was compared. The model was ideal due to its simplicity, interpretability, and effectiveness in handling binary classification problems, making it a robust starting point for understanding the key predictors of fraudulent behavior.

Logistic regression assumptions

1. **Independence of Observations:** It is important that each observation stand alone from the others. This implies that one observation's result shouldn't affect the result of another.

2. **Linearity of Logits:** There should be a linear relationship between the independent variables and the dependent variable's log-odds. This implies that the influence of a certain predictor on the outcome's log-odds should be linear.

3. **No Perfect Multicollinearity:** There shouldn't be perfect collinearity between the independent variables. It is impossible to assess the influence of each predictor independently when there is perfect multicollinearity, when one predictor variable is a perfect linear function of other predictor variables.

4. **Additivity:** The logit (log-odds) of the outcome is assumed to be a linear combination of the predictor variables. This means that the effect of each predictor is additive in the log-odds scale.

Logistic regression model results and coefficients

The results from the logistic regression model were presented in Table 4.2 The results encompass metrics such as precision, F1-score, recall, and support. Additionally, the tables provide the coefficients and p-values for each predictor variable. Accuracy is not included in

these results because it may be misleading in the context of imbalanced datasets, such as fraud detection, where other metrics like precision, recall, and F1-score offer more meaningful insights into the model's performance.

Table 4.2: Logistic Regression Model Performance and Predictor Coefficients

Model Performance		**Predictor Coefficients**		
Metric	**Performance**	**Covariate**	**Coefficient**	**P-value**
Accuracy	0.92	Intercept (Constant)	5.5269	0.0000
Precision	0.276173	Age	0.0486	0.0000
Recall	0.916168	business turnover	-0.0000	0.0000
F1-score	0.424411	total turnover	0.0000	0.0147
Support	167	total financing expenses	0.0000	0.0000
		transaction ratio	-0.0000	0.0000

The model performance as tabulated showed that: Precision (0.276), the proportion of true positive predictions (fraud cases correctly identified) out of all the predicted positive cases was 0.276. This implies that out of all the cases predicted as fraudulent, only 27.6% were actually fraudulent (low precision level). For recall, the proportion of true positive predictions out of all actual positive cases (all fraud cases in the test set), was approximately 0.916. This means that the model correctly identifies about 92% of all actual fraud cases. This is a strong result and indicates that the model is very effective at capturing most of the fraud cases. The F1-score, a statistic that provides a balance between precision and recall calculated as the harmonic mean of the two, An estimated F1-score of 0.42 indicates the precision vs. recall trade-off. This is a moderate overall performance given the low precision and high recall,

The logit prediction coefficient revealed the following: The intercept represented the baseline log-odds of fraud when all other variables were zero. A large negative intercept suggested that the baseline probability of fraud was quite low, meaning fraud was unlikely without considering other factors. The P-value of 0.0000 indicated that the intercept was statistically significant. The second variable, Age, had a coefficient of 0.0486 and a P-value of 0.0000. The positive coefficient indicated that as Age increased, the log-odds of fraud also increased, meaning older individuals or entities were more likely to engage in fraud. The effect was moderate, but statistically significant. The third variable, Business Turnover, had a coefficient of -0.0000 and a P-value of 0.0000. Although the coefficient was negative, its value was

effectively zero, showing that business turnover had minimal or negligible influence on fraud likelihood. However, the negative direction suggested a slightly lower chance of fraud with higher turnover. The P-value confirmed the variable was statistically significant, despite its small effect size. The fourth variable, Total Turnover, had a coefficient of 0.0000 and a P-value of 0.0147. Like business turnover, the coefficient was very small but positive, indicating that higher total turnover was slightly associated with an increased likelihood of fraud. Although the effect was marginal, the P-value showed it was statistically significant, though less so than the other variables. The variable Total Financing Expenses had a coefficient of -0.0000, which indicated a small negative effect on the likelihood of fraud. The Transaction Ratio variable, with a negative coefficient, suggested that a higher transaction ratio slightly reduced the likelihood of fraud. Both variables were statistically significant despite their minimal impact on the model.

4.3.2 Random forest

The model's performance was assessed using three metrics which yielded as follows: Precision: 0.962, which means that 96.2% of the cases identified as fraudulent (class 1) were indeed fraudulent. This high precision indicates a low number of false positives. Recall: 0.766, which shows that the model correctly identified 76.6% of the actual fraud cases. This high recall indicates a low number of false negatives. The model excelled in both precision and recall, as reflected by an F1-Score of 0.853, the harmonic mean of precision and recall. The model detected 442 instances of fraud.

Results obtained from the random forest on feature importance indicated that the most important feature for predicting fraud was total financing expenses 52%, followed by total turnover 20% and business turnover 15%. Lesser importance was given to transaction ratio and Age with 8% and 4% respectively. That was an indication that financial metrics play a larger role in fraud detection compared to demographic information. The results for the Random forest are tabulated in Table 4.3

Table 4.3: Random Forest Feature Importance

Feature	Importance
total financing expenses	0.521821
total turnover	0.199468
business turnover	0.146815
transaction ratio	0.088265
Age	0.043632

4.3.3 Decision Tree

The Decision Tree model had a precision of 0.772, with 77.2% of instances predicted as fraud actually being fraud. It had a recall of 0.832, correctly identifying 83.2% of the actual fraudulent cases. The strong balance between precision and recall resulted in a high F1-score of 0.801, indicating a good overall performance.

On feature importance, the tree indicated that the most important feature for predicting fraud was total turnover 42% followed by total financing expenses 40%, and business turnover 9%. transaction ratio and Age had 5% and 3% respectively. Similar to the Random forest, the results were indicative that financial metrics play a larger role in fraud detection. The results for the decision tree was tabulated in Table 4.4

Table 4.4: Decision Tree Feature Importance

Feature	Importance
total turnover	0.420631
total financing expenses	0.404421
business turnover	0.094422
transaction ratio	0.050979
Age	0.029547

4.3.4 Mlp, Naive Bayes and Knn

These models performed as tabulated in Table 4.5. The K-Nearest Neighbour (KNN) model achieved a precision of 0.963, recall of 0.715, and an F1-score of 0.821, indicating a

good balance between correctly identified fraud cases and false positives. After testing various values of K (from 1 to 30), the best performance was observed at K = 3, yielding the highest F1-score. There were 442 fraud instances in the dataset. The Naive Bayes model had a precision of 0.121, meaning only 12.1% of instances predicted as fraud were actually fraud—indicating many false positives. However, the model achieved a recall of 0.968, correctly identifying 96.8% of actual fraud cases. Despite the high recall, the very low precision led to a moderate F1-score of 0.215, reflecting poor overall performance in balancing false positives and false negatives.

The Multi-Layer Perceptron (MLP) model achieved a perfect precision of 1.0, meaning 100% of the instances it labeled as fraud were indeed fraud. It had a recall of 0.948, correctly identifying 94.8% of the actual fraud cases. The combination of high precision and high recall resulted in an excellent F1-score of 0.973, indicating exceptional overall performance

Table 4.5: Model Performance

Metric	**Precision**	**Recall**	**F1-score**	**Support**
K nearest neighbour	0.963415	0.714932	0.820779	442
Naive bayes	0.120836	0.968326	0.214859	442
Multi-layer perceptron	1.0	0.947964	0.973287	442
Decision tree	0.932018	0.961538	0.946548	442

4.4 Model Evaluation

In evaluating the models, Random Forest and MLP emerged as the top performers, with both achieving high precision (96.09% for Random Forest and 97.81% for MLP), high recall (76.06% for Random Forest and 76.40% for MLP), and strong F1 scores (84.84% for Random Forest and 85.71% for MLP). Decision Tree provided moderately good performance with balanced precision and recall at around 76%, while Logistic Regression, despite its high precision (95.78%), had a very low recall (12.34%), making it ineffective at detecting most fraud cases. Naive Bayes was the weakest model, with very poor recall (2.15%) and a low F1 score (3.98%), rendering it unsuitable for fraud detection.

Random Forest is preferable over MLP due to its lower computational complexity and greater interpretability. While both models perform similarly, Random Forest requires less computational resources, making it more efficient for large datasets and faster to train. Additionally, Random Forest's decision trees are easier to interpret, allowing for better insight into how fraud decisions are made, which is important for transparency and model explainability. MLP, being a neural network, can act more like a black box, making it harder to explain and justify predictions. Therefore, Random Forest strikes a better balance between performance and usability. Table 4.6 and Figure 4.1 shows the evaluation results.

Table 4.6: Model Performance Metrics

Model	**Precision**	**Recall**	**F1 Score**	**Support**
Logistic Regression	0.958 ± 0.085	0.123 ± 0.042	0.216 ± 0.067	50.400 ± 6.135
Random Forest	0.961 ± 0.029	0.761 ± 0.047	0.848 ± 0.035	50.400 ± 6.135
Decision Tree	0.769 ± 0.037	0.767 ± 0.066	0.767 ± 0.045	50.400 ± 6.135
Naive Bayes	0.318 ± 0.274	0.022 ± 0.014	0.040 ± 0.026	139.400 ± 10.735
MLP	0.978 ± 0.029	0.764 ± 0.034	0.857 ± 0.020	50.400 ± 6.135

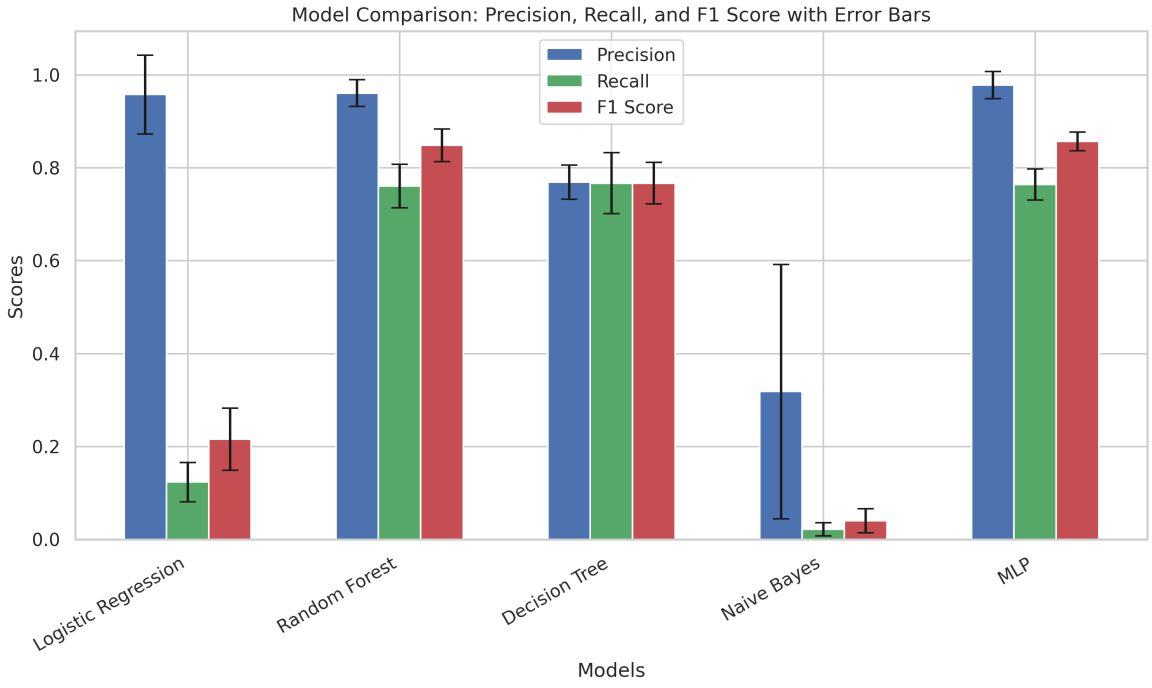


Figure 4.1: Model Performance Comparison

4.5 Optimization of the Random Forest Model

Optimization of the Random forest model achieved the following: Improved precision: After optimization, the precision improved from 0.96 to near perfect. Implying that all flagged cases were indeed fraudulent. On feature importance, the two most important features remain consistent in the two models, indicating that the core determinants of fraud prediction are robust.

4.6 Fraud threshold and prediction

Threshold After analyzing the ROC curve, the optimal threshold was determined to be 0.4, as this value effectively balances the true positive rate (recall) and the false positive rate. By setting the threshold at 0.4, the algorithm was able to maximize the identification of suspicious cases, ensuring a high detection rate for potential fraud. This threshold proved to be ideal because it significantly increases the number of cases flagged by the system without causing a substantial drop in precision. In other words, it strikes a favorable balance between catching fraudulent activity and minimizing the likelihood of incorrectly flagging legitimate cases. This makes the 0.4 threshold particularly suitable for scenarios where identifying fraud is a top priority while maintaining a reasonable accuracy in the flagged results.

Prediction From the prediction analysis, using a threshold of 0.4, the algorithm identified 845 returns as fraudulent, accounting for 3.1% of the total returns. As the threshold increases, the number of suspicious returns detected decreases. At a threshold of 0.1, 1,375 cases were flagged as fraudulent, At this high sensitivity threshold, the algorithm focuses on detecting more suspicious returns but with a higher risk of false positives. The project, therefore, opted for a balanced approach that flags a moderate number of suspicious cases while maintaining a reasonable false positive rate.

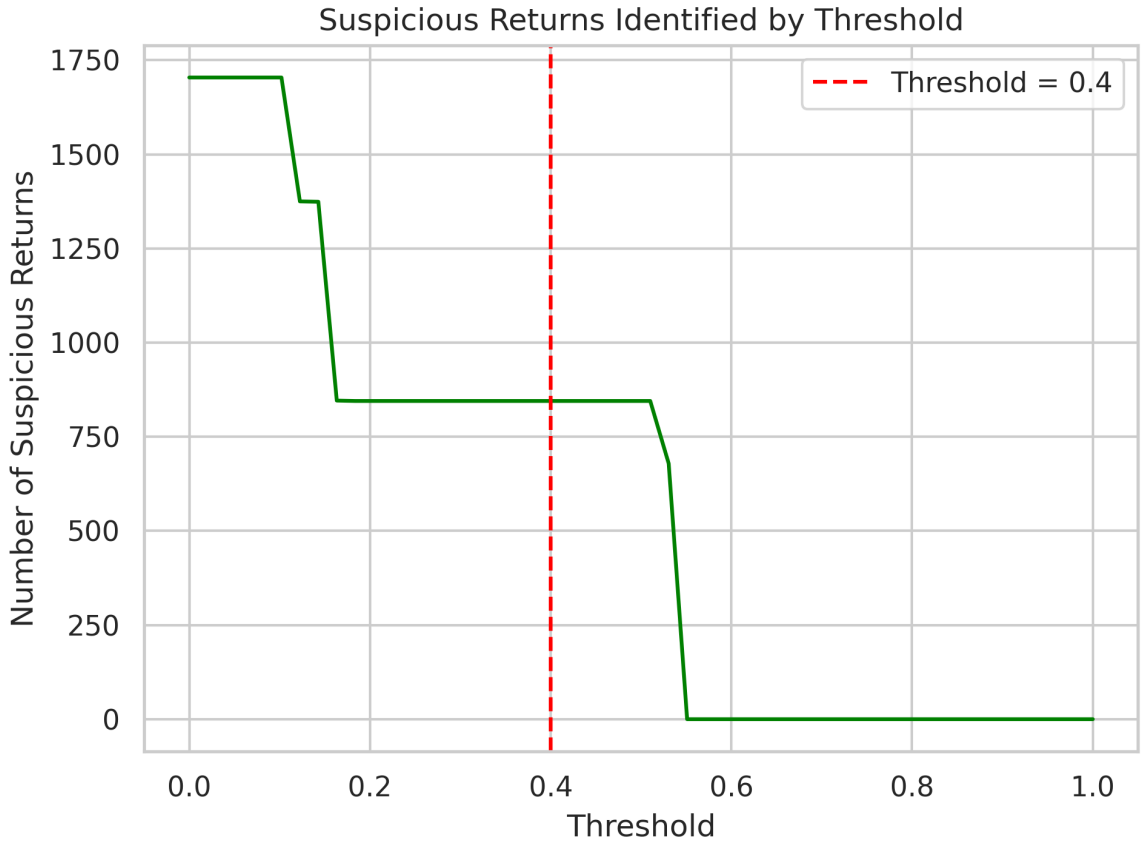


Figure 4.2: Distribution of Predicted Fraud Probabilities

In summary, the Random Forest outperformed the other models by effectively capturing complex nonlinear relationships between the features. Its lower sensitivity to class imbalance contributed to its superior performance compared to the other models. On the other hand, Logistic Regression struggled, showing a strong sensitivity to class imbalance, as reflected in the results.

Chapter 5

5.1 Introduction

This chapter explores the results presented in the previous chapter, comparing them with findings from recent studies on fraud detection, particularly in tax-related contexts. It also provides an in-depth analysis of these results, offering insights and practical recommendations for improving fraud detection methods. Additionally, the chapter identifies key areas where future research could further enhance understanding and effectiveness in this field.

5.2 Discussion

The primary goal of this study was to develop and implement machine learning algorithms to identify and predict fraudulent cases in a rapidly evolving environment with aggressive fraudulent activities. The results showed that machine learning models could effectively analyze large datasets from the Kenya Revenue Authority, detect unusual patterns, and distinguish between fraudulent and non-fraudulent taxpayers in near real-time. Among the models tested, the random forest algorithm performed the best, achieving a good overall performance.

Regarding the performance of the trained machine learning models, the results indicate that the Random Forest model outperformed the other models in the year of income sample. The MLP model also showed strong and consistent performance. These findings align with previous research studies by Abedin et al. (2020) and Beutel et al. (2015). This study provided evidence of the effectiveness of ensemble-based machine learning models in predicting tax fraud. It demonstrated that the Random Forest model is particularly adept at detecting tax anomalies in real-world scenarios, suggesting potential for significant cost savings and faster decision-making for stakeholders. Similarly, a study by Andrade et al. (2021) on machine

learning-based financial fraud detection in Brazil reported comparable results, where the Random Forest model outperformed others with a precision of 0.94, recall of 0.92, and an F1 score of 0.93. The current study demonstrates an improvement in accuracy, achieving a precision of 0.96.

Another significant finding from the study was the importance of features. Our results indicated that total financial expenses should be regarded as the most informative indicator for distinguishing between fraudulent and non-fraudulent tax returns. Additionally, total turnover and business turnover emerged as other crucial predictive indicators.These findings are consistent with those from alternative fraud prediction models by Febriminanto and Wasesa (2022) and Prolhac and Gaie (2023), which revealed that financial expenses accounted for more than 43% of the predictive importance, surpassing all other indicators.

In comparing findings among the models: Decision Tree, Random Forest, and Logistic Regression models, some clear patterns emerged regarding which factors play a key role in predicting fraud. The models placed the greatest emphasis on total financial expenses (0.4206) and total turnover (0.4044), suggesting that these financial metrics were the primary drivers in distinguishing fraudulent from non-fraudulent cases. Secondary factors like business turnover (0.0944) and transaction ratio (0.0510) contribute more modestly, while Age (0.0295) had the least impact, highlighting the dominant role that financial data plays in a decision tree framework for fraud detection.

Overall, the comparison across these models highlighted the consistent dominance of financial metrics, particularly total turnover and total financing expenses, in identifying fraud, while factors like Age and transaction ratio played lesser roles. The different modeling techniques revealed both the complexity of fraud detection and the nuanced ways in which various factors contributed to predicting fraudulent activities.

5.3 Recommendation

Based on the machine learning results and the comparative analysis of different models, The study makes the following recommendations to the Kenya Revenue Authority to enhance fraud detection and prediction:

5.3.1 Recommendations for further studies

Limited Feature Set: The current models primarily rely on structured financial data. Future studies should incorporate additional features, including non-financial data (e.g., taxpayer history, behavior patterns) or unstructured data (e.g., text or audit reports), to capture more complex fraud behaviors and improve model performance.

Static Model Assumptions: The models used in this study assume that future fraud patterns will mirror past ones, which may not hold true. Future research should explore adaptive learning models that continuously update based on new data and can detect emerging fraud schemes without historical bias.

Generalization to Other Domains: The study's findings are based on Kenya Revenue Authority data, which may not generalize to other tax environments. Future research could apply these models to international datasets or other industries to evaluate their broader applicability.

Limited Assessment of Temporal Patterns: The current models do not account for temporal changes in fraud patterns. Future studies should explore time-series analysis or sequential learning models like Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to enhance the early detection of dynamic fraud schemes.

Threshold Sensitivity: The models' performance is highly sensitive to decision thresholds, which might not be optimal in all cases. Future research could explore dynamic thresholding techniques or cost-sensitive learning to adjust thresholds based on different business or operational needs.

5.3.2 Policy recommendations

Leverage Financial Metrics for Early Detection: Total turnover and financing expenses were identified as the most critical features for fraud detection. Kenya Revenue Authority (KRA) should prioritize monitoring deviations in these metrics and develop systems that flag anomalies for early intervention.

Adopt Random Forest for Operational Use : The Random Forest model demonstrated superior performance, making it a reliable choice for detecting fraud with high accuracy and minimizing false positives. KRA should consider deploying this model in its fraud detection systems for handling large datasets.

Enhance Data Collection and Feature Tracking: Expanding data collection to include additional financial and transactional metrics can improve fraud detection accuracy. KRA should focus on capturing more comprehensive data from taxpayers to better identify fraudulent activities.

Regular Model Evaluation and Updates: Machine learning models require constant evaluation and updates to remain effective as fraud tactics evolve. KRA should implement a feedback loop for retraining models with new data to detect emerging fraud patterns.

Fraud Risk Segmentation: KRA could develop a risk-based segmentation system based on findings, prioritizing taxpayers with high total turnover and financing expenses for audits. This would allow for more efficient resource allocation and focus on high-risk cases.

Future Focus on Unstructured Data: To further enhance fraud detection, KRA should explore integrating unstructured data like emails, tax forms, or communication logs through text mining and natural language processing (NLP) techniques.

Collaboration with Other Institutions: By partnering with other financial and government institutions, KRA can access additional data sources, enhancing fraud detection accuracy through cross-referencing taxpayer information.

By implementing these recommendations, the tax agency can significantly enhance its fraud detection capabilities, improving both the accuracy and efficiency of identifying fraudulent activities in a dynamic tax environment.

5.4 Conclusion

The study successfully achieved its primary objective of developing and applying machine learning algorithms to detect and predict fraudulent cases in a dynamic environment, specifically within the context of the Kenya Revenue Authority's taxation system. By training and evaluating several machine learning models, including Logistic Regression, Random Forest, Decision Trees, K-Nearest Neighbors (KNN), Gaussian Naive Bayes, and Multilayer Perceptron (MLP), the research was able to conduct a thorough comparative analysis of predictive accuracy.

Among the algorithms tested, Random Forest emerged as the best-performing model, showcasing superior performance in detecting fraud, with a precision of 96.1%, a recall of 76%, and an F1-score of 85%. This underscores the model's robustness in handling complex data patterns and interactions, making it particularly effective in distinguishing fraudulent taxpayers from non-fraudulent ones in near real-time. Additionally, the study successfully developed a fraud detection model that can be applied within the Kenya Revenue Authority's taxation system. The model demonstrated the ability to handle large volumes of data and identify anomalies based on financial metrics, providing a scalable and practical tool for real-world applications.

Efforts to optimize the machine learning algorithms resulted in significant improvements in predictive accuracy, confirming that model tuning and feature importance analysis are critical to enhancing performance. The importance of features such as total turnover and total financing expenses was evident across multiple models, highlighting the relevance of financial indicators in fraud detection. Overall, the findings confirm that machine learning models, especially Random Forest, are well-suited for detecting fraudulent activities in dynamic tax environments. These results offer a foundation for future enhancements in

fraud prediction systems, making them more effective, adaptive, and applicable to real-world taxation systems.

References

Abdallah, A., Maarof, M. A., and Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68:90–113.

Abedin, M. Z., Chi, G., Uddin, M. M., Satu, M. S., Khan, M. I., and Hajek, P. (2020). Tax default prediction using feature transformation-based machine learning. *IEEE Access*, 9:19864–19881.

Abonyi, J. and Feil, B. (2007). *Cluster analysis for data mining and system identification*. Springer Science & Business Media.

Adadi, A. (2021). A survey on data-efficient algorithms in big data era. *Journal of Big Data*, 8(1):24.

Al-Hashedi, K. G. and Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, 40:100402.

Albashrawi, M. (2016). Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015. *Journal of Data Science*, 14(3):553–569.

Algan, G. and Ulusoy, I. (2021). Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771.

Alsadhan, N. (2023). A multi-module machine learning approach to detect tax fraud. *Comput. Syst. Sci. Eng.*, 46(1):241–253.

Andrade, J. P. A., Paulucio, L. S., Paixao, T. M., Berriel, R. F., Carneiro, T. C. J., Carneiro, R. V., De Souza, A. F., Badue, C., and Oliveira-Santos, T. (2021). A machine learning-based system for financial fraud detection. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 165–176. SBC.

Assylbekov, Z., Melnykov, I., Bekishev, R., Baltabayeva, A., Bissengaliyeva, D., and Mamlin, E. (2016). Detecting value-added tax evasion by business entities of kazakhstan. In *Intelligent Decision Technologies 2016: Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016)–Part I*, pages 37–49. Springer.

Aziz, L. A.-R. and Andriansyah, Y. (2023). The role artificial intelligence in modern banking: an exploration of ai-driven approaches for enhanced fraud prevention, risk management, and regulatory compliance. *Reviews of Contemporary Business Analytics*, 6(1):110–132.

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.

Behera, T. K. and Panigrahi, S. (2015). Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network. In *2015 second international conference on advances in computing and communication engineering*, pages 494–499. IEEE.

Beutel, A., Akoglu, L., and Faloutsos, C. (2015). Fraud detection through graph-based user behavior modeling. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1696–1697.

Bolton, C. (2009). *Logistic regression and its application in credit scoring*. University of Pretoria (South Africa).

Buoni, A. (2012). Fraud detection in the banking sector: a multi-agent approach.

Cheng, X., Liu, S., Sun, X., Wang, Z., Zhou, H., Shao, Y., and Shen, H. (2021). Combating emerging financial risks in the big data era: A perspective review. *Fundamental Research*, 1(5):595–606.

Choi, D., Lee, K., et al. (2018). An artificial intelligence approach to financial fraud detection under iot environment: A survey and implementation. *Security and Communication Networks*, 2018.

Cobham, A. and Janskỳ, P. (2018). Global distribution of revenue loss from corporate tax avoidance: re-estimation and country results. *Journal of International Development*, 30(2):206–232.

Daho, M. E. H. and Chikh, M. A. (2015). Combining bootstrapping samples, random subspaces and random forests to build classifiers. *Journal of Medical Imaging and Health Informatics*, 5(3):539–544.

Dalu, T., Maposa, V. G., Pabwaungana, S., and Dalu, T. (2012). The impact of tax evasion and avoidance on the economy: a case of harare, zimbabwe. *African Journal of Economic and Sustainable Development*, 1(3):284–296.

Demirhan, H. (2024). Financial anomalies and creditworthiness: A python-driven machine learning approach using mahalanobis distance for ise-listed companies in the production and manufacturing sector. *Journal of Financial Risk Management*, 13(1):1–41.

Desai, M. A. and Dharmapala, D. (2009). Corporate tax avoidance and firm value. *The review of Economics and Statistics*, 91(3):537–546.

Devos, K. (2013). *Factors influencing individual taxpayer compliance behaviour*. Springer Science & Business Media.

Dutta, I., Dutta, S., and Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems with Applications*, 90:374–393.

Edge, M. E. and Sampaio, P. R. F. (2012). The design of ffml: A rule-based policy modelling language for proactive fraud management in financial data streams. *Expert Systems with Applications*, 39(11):9966–9985.

Edjabou, L. D. and Smed, S. (2013). The effect of using consumption taxes on foods to promote climate friendly diets–the case of denmark. *Food policy*, 39:84–96.

Elkins, D. (2006). Horizontal equity as a principle of tax theory. *Yale L. & Pol'y Rev.*, 24:43.

Faccia, A. and Mosteanu, N. R. (2019). Tax evasion, information systems and blockchain. *Journal of Information Systems & Operations Management*, 13(1):65–74.

Febriminanto, R. D. and Wasesa, M. (2022). Machine learning analytics for predicting tax revenue potential. *Indonesian Treasury Review: Jurnal Perbendaharaan, Keuangan Negara dan Kebijakan Publik*, 7(3):193–205.

Fischer, C. M., Wartick, M., and Mark, M. M. (1992). Detection probability and taxpayer compliance: A review of the literature. *Journal of accounting literature*, 11:1.

Garner, S., Crocker, R., Skidmore, M., Webb, S., Graham, J., and Gill, M. (2016). *Organised fraud in local communities*. Police Foundation London.

Giannakas, F., Troussas, C., Voyiatzis, I., and Sgouropoulou, C. (2021). A deep learning classification framework for early prediction of team-based academic performance. *Applied Soft Computing*, 106:107355.

Görtler, J., Hohman, F., Moritz, D., Wongsuphasawat, K., Ren, D., Nair, R., Kirchner, M., and Patel, K. (2022). Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Gulmeher, R. and Aiman, U. (2023). A novel approach to unveiling employee attrition patterns using machine learning algorithms. *Journal of Scientific Research and Technology*, pages 234–241.

Hasseldine, J. and Morris, G. (2013). Corporate social responsibility and tax avoidance: A comment and reflection. In *Accounting Forum*, volume 37, pages 1–14. Taylor & Francis.

He, H., Wu, X., and Wang, Q. (2021). Forecasting urban mobility using sparse data: A gradient boosted fusion tree approach. In *1st ACM SIGSPATIAL International Workshop on the Human Mobility Prediction Challenge (HuMob-Challenge 2023) Nov. 13, 2023, Hamburg, Germany*, page 3628507.

Hegland, M. (2007). The apriori algorithm–a tutorial. *Mathematics and computation in imaging science and information processing*, pages 209–262.

Henke, N. and Jacques Bughin, L. (2016). The age of analytics: Competing in a data-driven world.

Hilas, C. S., Mastorocostas, P. A., and Rekanos, I. T. (2015). Clustering of telecommunications user profiles for fraud detection and security enhancement in large corporate networks: a case study. *Appl. Math. Inf. Sci*, 9(4):1709.

Imoniana, J. O., Antunes, M. T. P., and Formigoni, H. (2013). The forensic accounting and corporate fraud. *JISTEM-Journal of Information Systems and Technology Management*, 10:119–144.

Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., and Aanæs, H. (2014). Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413.

Karayiannis, N. and Venetsanopoulos, A. N. (1992). *Artificial neural networks: learning algorithms, performance evaluation, and applications*, volume 209. Springer Science & Business Media.

Kyriienko, O. and Magnusson, E. B. (2022). Unsupervised quantum machine learning for fraud detection. *arXiv preprint arXiv:2208.01203*.

Lee, C. (2022). Deep learning-based detection of tax frauds: an application to property acquisition tax. *Data Technologies and Applications*, 56(3):329–341.

Li, Y., Liang, W., Peng, L., Zhang, D., Yang, C., and Li, K.-C. (2022). Predicting drug-target interactions via dual-stream graph neural network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

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.

Mackevičius, J. and Kazlauskienė, L. (2009). The fraud tree and its investigation in audit. *Ekonomika*, 85:90–101.

Malaszczyk, K. and Purcell, B. M. (2017). Big data analytics in tax fraud detection. *Northeastern Association of Business, Economics and Technology*, page 233.

Malekian, D. and Hashemi, M. R. (2013). An adaptive profile based fraud detection framework for handling concept drift. In *2013 10th International ISC Conference on Information Security and Cryptology (ISCISC)*, pages 1–6. IEEE.

Manning, G. A. (2010). *Financial investigation and forensic accounting*. Routledge.

Massa, D. and Valverde, R. (2014). A fraud detection system based on anomaly intrusion detection systems for e-commerce applications. *Computer and Information Science*, 7(2):117–140.

Muhammad, G., Hossain, M. S., and Garg, S. (2020). Stacked autoencoder-based intrusion detection system to combat financial fraudulent. *IEEE Internet of Things Journal*.

Murorunkwere, B. F., Haughton, D., Nzabanita, J., Kipkogei, F., and Kabano, I. (2023). Predicting tax fraud using supervised machine learning approach. *African Journal of Science, Technology, Innovation and Development*, pages 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.

Niksa-Rynkiewicz, T., Landowski, M., and Szalewski, P. (2020). Application of apriori algorithm in the lamination process in yacht production. *Polish Maritime Research*, 27(3):59–70.

O'Sullivan, R., Schmidt, O., and Monahan, F. J. (2022). Stable isotope ratio analysis for the authentication of milk and dairy ingredients: A review. *International Dairy Journal*, 126:105268.

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):86.

Politou, E., Alepis, E., and Patsakis, C. (2019). Profiling tax and financial behaviour with big data under the gdpr. *Computer law & security review*, 35(3):306–329.

Posner, R. A. (1971). Taxation by regulation. *The Bell Journal of Economics and Management Science*, pages 22–50.

Prolhac, J. and Gaie, C. (2023). Providing an open framework to facilitate tax fraud detection. *International Journal of Computer Applications in Technology*, 73(1):24–41.

Sanni, L. (2019). *An Informational Perspective and a Framework for Employee Fraud Detection*. PhD thesis, Université Paris 1 Panthéon Sorbonne, France.

Schneider, F. (2013). The financial flows of transnational crime and tax fraud in oecd countries: What do we (not) know? *Public Finance Review*, 41(5):677–707.

Schreyer, M., Sattarov, T., Borth, D., Dengel, A., and Reimer, B. (2017). Detection of anomalies in large scale accounting data using deep autoencoder networks. *arXiv preprint arXiv:1709.05254*.

Shafer, W. E., Wang, Z., and Hsieh, T.-S. (2020). Support for economic inequality and tax evasion. *Sustainability*, 12(19):8025.

Singh, K. and Best, P. (2016). Interactive visual analysis of anomalous accounts payable transactions in sap enterprise systems. *Managerial Auditing Journal*, 31(1):35–63.

Stiglitz, J. E. (1985). The general theory of tax avoidance. *National Tax Journal*, 38(3):325–337.

Sumithra, V. and Surendran, S. (2015). A computational geometric approach for overlapping community (cover) detection in social network. In *2015 International Conference on Computing and Network Communications (CoCoNet)*, pages 98–105. IEEE.

Taylor, G. and Richardson, G. (2012). International corporate tax avoidance practices: Evidence from australian firms. *The International Journal of Accounting*, 47(4):469–496.

Vanhoeyveld, J., Martens, D., and Peeters, B. (2020). Value-added tax fraud detection with scalable anomaly detection techniques. *Applied Soft Computing*, 86:105895.

Vasco, M. D. C. G., Rodríguez, M. J. D., and Santos, S. D. L. (2021). Segmentation of potential fraud taxpayers and characterization in personal income tax using data mining techniques. *Hacienda Publica Espanola*, (239):127–157.

Verbeeck, N., Caprioli, R. M., and Van de Plas, R. (2020). Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass spectrometry reviews*, 39(3):245–291.

Wu, R.-S., Ou, C.-S., Lin, H.-y., Chang, S.-I., and Yen, D. C. (2012). Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*, 39(10):8769–8777.

Xavier, O. C., Pires, S. R., Marques, T. C., and Soares, A. d. S. (2022). Tax evasion identification using open data and artificial intelligence. *Revista de Administração Pública*, 56:426–440.

Yamen, A. E., Mersni, H., and Ramadan, A. (2023). Tax evasion and public governance before and after the european "big bang": a red flag for policymakers. *Journal of Financial Crime*, 30(2):420–436.

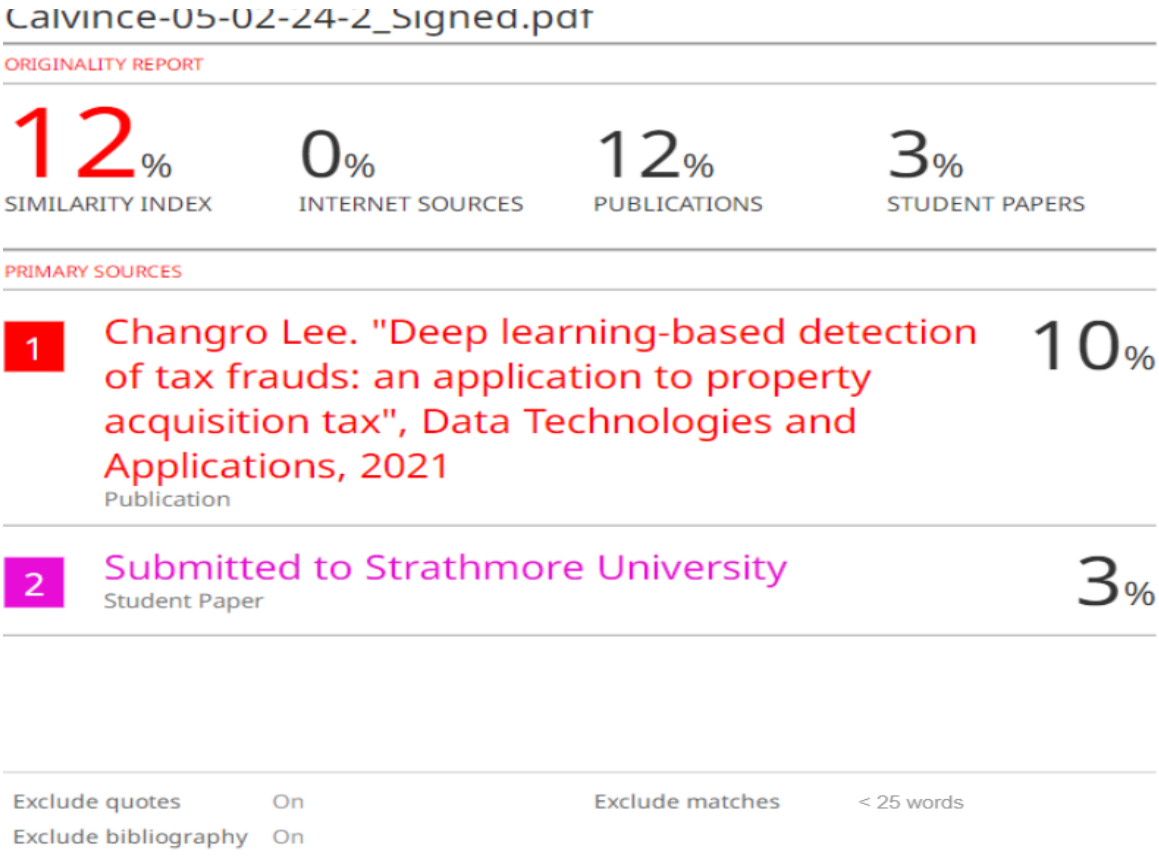
Yang, L. and Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316.

Zheng, Q., Xu, Y., Liu, H., Shi, B., Wang, J., and Dong, B. (2023). A survey of tax risk detection using data mining techniques. *Engineering*.

Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R., and Rosen, M. S. (2018). Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492.

Appendix A

Similarity Report



Appendix B

Python code link

https://colab.research.google.com/drive/1G6KctAKdN-6mwBAtfPcZ40fGn43J-t7j?usp=sharing

Appendix C

Deployment Tool Images

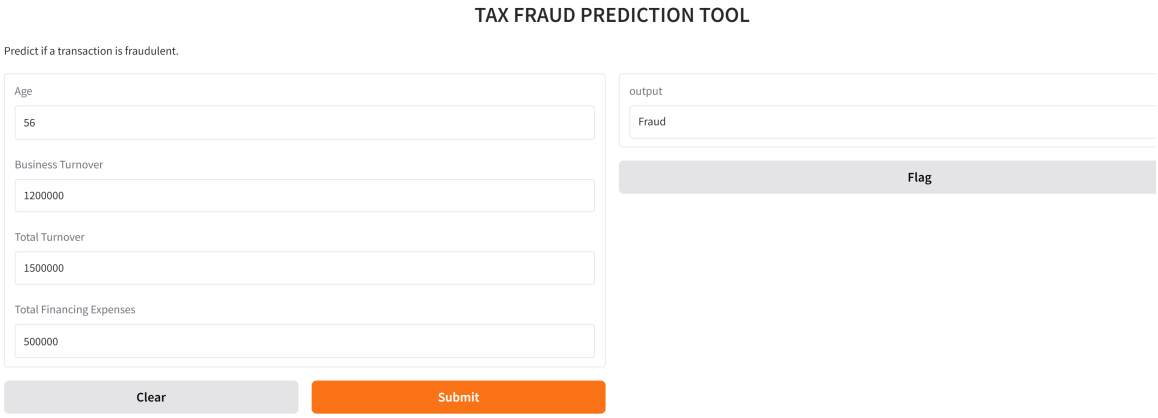


Figure C.1: Deployment Prediction tool

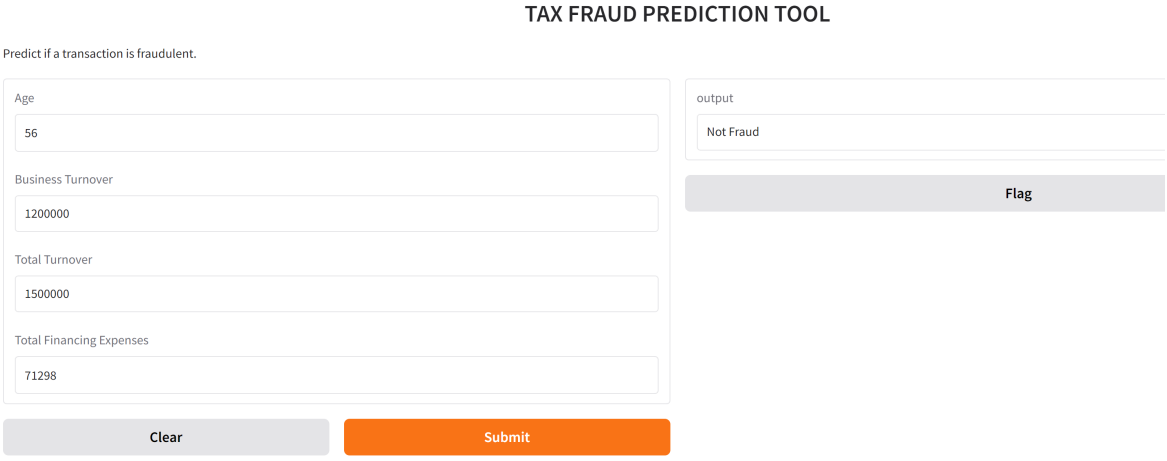


Figure C.2: Deployment Prediction tool

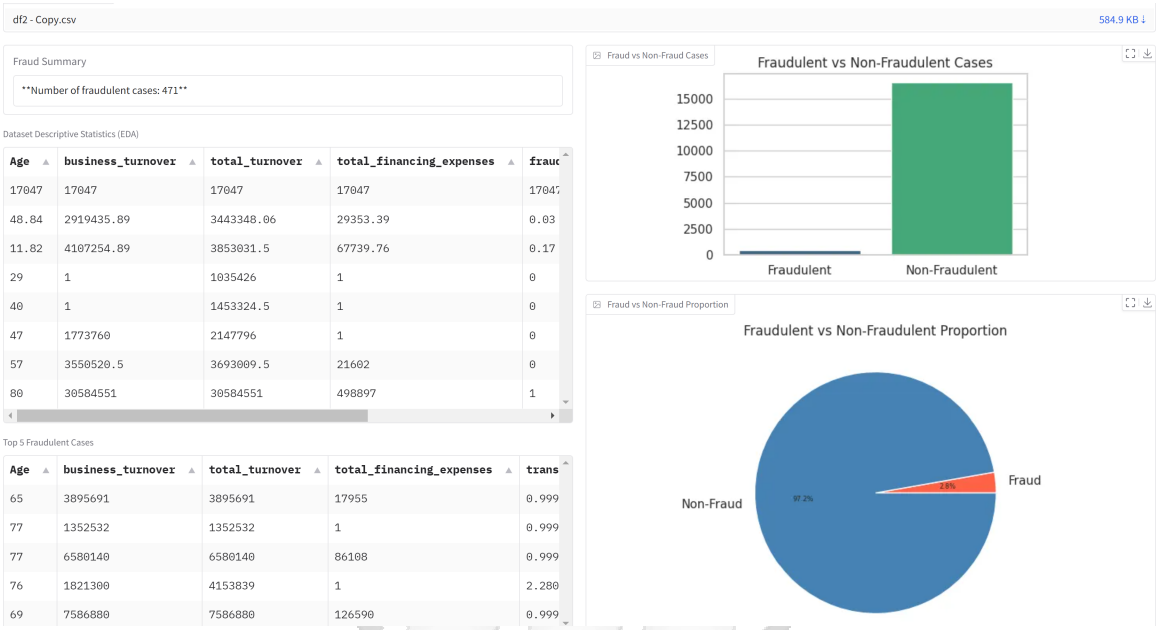


Figure C.3: Deployment Prediction tool

Appendix D

Ethics approval letter

Strathmore UNIVERSITY

24th May 2024

Mr Onyango Calvince,
calvince.onyango@strathmore.edu

Dear Mr Onyango,

RE: Tax Fraud Prediction using Machine Learning Models in Kenya

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2048/24.** The approval period is from **24th May 2024 to 23rd May 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