USING A DATA-DRIVEN MODEL TO PREDICT TAXPAYERS FILING FALSE RETURNS: A CASE OF ZAMBIA REVENUE AUTHORITY

BY

MUBANGA MUBANGA

2018245571

A DISSERTATION SUBMITTED IN PARTIAL FULFUILMENT OF THE REQUIREMENT OF A DEGREE OF MASTERS' IN COMPUTER SCIENCE

THE UNIVERSITY OF ZAMBIA

SCHOOL OF NATURAL SCIENCES

LUSAKA

DECEMBER 2023

COPYRIGHT

DECLARATION

I, MUBANGA MUBANGA, hereby affirm that this dissertation is solely my original creation and has not been presented for consideration at any other college, institution, or university apart from the University of Zambia.

Name: …….

Sign: ……

Date: ……

APPROVAL

This dissertation, by MUBANGA MUBANGA, has been approved as partial fulfillment of the requirements for the award of Masters' in Computer Science by the University of Zambia.

Examiner 1

Name: …….

Signature: …….

Date: ……..

Examiner 2

Name: …….

Signature: ……..

Date: ……..

Examiner 3

Name: …….

Signature: ……..

Date: ……..

Chairperson (Board of Examiners)

Name: …….

Signature: ……..

Date: ……..

Supervisor

Name: …….

Signature: ……..

Date: ……..

ACKNOWLEDGEMENTS

I would want to express my appreciation to the University of Zambia, namely the School of Natural Sciences and the Department of Computer Science, for providing the fundamental basis that facilitated this research. A heartfelt thank you goes to Prof. Jackson Phiri, my research supervisor, for his invaluable guidance and unwavering support throughout this endeavor. His insights and assistance have been indispensable, and I deeply appreciate the time he has dedicated to providing feedback on my work.

I am also immensely grateful to the Zambia Revenue Authority, specifically the Innovation and Project Management team, and the Information Technology team, for their collaboration and assistance during the course of my research. Their contributions have been instrumental, and I extend my sincere thanks to them.

In conclusion, I would like to thank everyone who has assisted me in my academic journey in many ways. Your collective support has been pivotal in my advancement, and for that, I am genuinely thankful.

DEDICATION

This study is devoted to my mother, Lillian Nkaka Mubanga, whose unwavering support has been the bedrock of my life. Her encouragement to stay focused and diligent has provided me with a solid footing from which to flourish and succeed. I am profoundly grateful for her love and guidance. Additionally, I extend this dedication to my sister, Temweni, and my close friends, Jokiwe, Tasha, Francis, Jehaph, and Natasha. Your selflessness and unwavering presence have been invaluable in assisting me to complete this work. Thank you for standing beside me, for your steadfast support, and for the multitude of ways in which you have contributed to this achievement.

ABSTRACT

Tax fraud remains a global issue, with significant economic setbacks for many countries, including Zambia. Traditional methods of tackling this challenge often hinge on labelled datasets, which are scarce due to the slow nature of tax audits and the inherent biases in sample selection. To address this data scarcity and offer a more immediate solution, this study introduces an unsupervised approach utilising K-means clustering alongside anomaly detection techniques. Using an extensive dataset of VAT declarations and associated refund transactions spanning several years, we demonstrate the potential of this method for efficiently identifying potential tax fraud cases. The significance of this paper is twofold: it introduces an innovative approach to a persistent issue and applies it specifically to the context of Zambia. By bypassing the need for exhaustive labelled data, our methodology offers a promising direction for enhancing tax fraud detection capabilities, ensuring a more resilient fiscal landscape for Zambia.

Keywords: K-means clustering, Anomaly detection, VAT, Tax fraud detection

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

VAT	Value Added Tax
ZRA	Zambia Revenue Authority
RRA	Rwanda Revenue Authority
TPIN	Taxpayer Identification Number
BI	Business Intelligence
DM	Data Mining
KDD	Knowledge Discovery in Database
FD	Fraud Detection
ML	Machine Learning
AI	Artificial Intelligence
SMOTE	Synthetic Minority Over-sampling Technique
RF	Random Forest
GA	Genetic Algorithm
HMM	Hidden Markov Model
AD	Anomaly Detection
SOM	Self-Organising Map
ATTE	Affiliated-Transaction-based Tax Evasion
CRISP-DM	Cross-Industry Standard Process for Data Mining
EDA	Exploratory Data Analysis
RDBMS	Relational Database Management System
DB	Database
LOF	Local Outlier Factor
CSV	Comma-Separated Value
UML	Unified Modelling Language

1 INTRODUCTION AND BACKGROUND

1.1 Introduction

In the modern economic landscape, taxation remains a cornerstone of national development, with Value-Added Tax (VAT) being a crucial component, especially in countries like Zambia. [1] Value Added Tax (VAT), a form of indirect taxation on the consumption of goods and services, constitutes a substantial source of revenue for the Zambian government. However, the effectiveness of this tax system is continually undermined by pervasive issues of tax fraud and non-compliance. These challenges not only impede revenue collection but also compromise the fairness and efficiency of the taxation system. [2] Traditional methods of combating tax fraud, such as targeted and random audits, although essential, are increasingly proving to be insufficient in addressing the complexity and volume of modern tax evasion strategies. [3] The Zambia Revenue Authority (ZRA), in its 2021 annual report, acknowledged the limitations of these conventional approaches, citing the high costs and time required to conduct effective audits. [4] With the shift towards digital platforms, like the Tax online system, there's an unprecedented opportunity to harness the vast volumes of data generated for a more robust fraud detection mechanism. This research introduces a novel approach to enhancing VAT fraud detection in Zambia through the application of machine learning, particularly focusing on unsupervised prediction models like k-means clustering. Our proposed model seeks to examine the intricacies and subtleties inherent in VAT transactions, detecting patterns and irregularities that could suggest fraudulent behavior. Unlike conventional approaches that heavily depend on auditors' expertise and intuition, our approach leverages the objective analysis of data, reducing potential biases and increasing the efficiency of fraud detection.

1.2 Background

Most Governments worldwide, including Zambia, levy taxes as a primary source for national development. However, tax fraud and non-compliance are escalating challenges, dampening revenue collection potential. Tax fraud involves deliberately providing false information on a tax return to illegally gain financial advantages and reduce tax obligations [5]. Research shows that the onus of identifying and addressing taxpayer discrepancies falls on the tax auditors–a time-consuming and resource-intensive task [5] [6] Tax audits play a pivotal role in tax collection by serving as the primary mechanism for rectifying errors, enhancing compliance, and upholding equity within the taxation system. [7] Specifically, the value-added tax (VAT),

which is an indirect tax imposed on the consumption of goods and services, represents a notable revenue stream for the Zambian government. The Value Added Tax (VAT) functions through a simple mechanism where businesses registered for VAT are involved in two main tax-related activities: collecting and paying VAT. When a VAT-registered business sells goods or provides services, it charges VAT to its customers, known as the 'output tax.' On the other hand, when the same business purchases goods or services, it pays VAT, referred to as the 'input tax.' At the end of every tax period, the company deducts the input tax (VAT paid on purchases) from the output tax (VAT collected on sales) to ascertain the net VAT payable. If the output tax exceeds the input tax, the company owes the disparity to the Zambia Revenue Authority (ZRA). Conversely, if the input tax outweighs the output tax, the company is entitled to a VAT refund from the ZRA. VAT serves as a system to ensure taxation occurs solely on the added value at each stage of production or distribution, thereby averting double taxation. [8]

The VAT taxation system faces many challenges when dealing with fraudsters, including reducing their 'output taxes' to pay fewer taxes and increasing their 'input taxes' to receive refund payments. Tax experts are tasked with identifying instances of fraud or anomalies in transactions by conducting costly and time-consuming audits. Such investigations include checking the taxpayers' account records, physical verifications with clients and suppliers, bank statement audits, etc. [9] Recent studies have shown that ZRA is still heavily reliant on traditional methods of fraud detection, namely targeted and random audits, undercover operations, and whistleblowing. [6] Moreover, conducting audits manually is neither economically advantageous nor practically feasible. Through implementing the Tax online system, ZRA aims to improve compliance and maximise revenue. [4] Taxpayers can conveniently access their records and submit their VAT tax return declarations. This means large volumes of data are generated by the online system database, which can used to develop data mining models using a learning machine that can used to address challenges faced by tax administrators. Machine learning not only speeds up the time on building fraud detection models, but it can also remove bias against certain taxpayers if done correctly. Auditors usually rely on instincts, experience, and rule-based conditions to find taxpayers to audit, which can typically miss clues hidden in data and can lead to focusing mostly on experience audits. [10] Obtaining labelled data (instances of fraud or not fraud) to develop machine learning is a hard task that requires costly and time-consuming audits. According to the ZRA annual 2021 report [4], 1009.99 VAT audits were conducted for the year 2020. This highlights the challenges faced by the tax administration. In this paper, we introduce an unsupervised prediction model using

k-means clustering to identify similarities in existing data and anomaly detection methods specific to VAT declarations to help tax auditors quickly identify hidden patterns in data focused on fraud detection. Interviews conducted by this research highlight the tedious process that tax audits are faced with identifying any occurrences of tax fraud. Tax auditors encounter the significant hurdle of identifying instances of tax fraud using traditional approaches like manual case selection, reported case selection, and rule-based case selection. However, these methods are greatly limited by their dependence on human resources and the knowledge of financial and tax experts.

1.3 Statement of the Problem

In Zambia, the widespread problem of tax fraud presents a significant threat to economic stability, siphoning essential funds from the nation's fiscal reserves. The traditional approach to detecting tax fraud relies heavily on examining a history of past audits—data that is not only limited but also tainted by selection biases, making it an unreliable foundation for identifying fraud comprehensively. This challenge is compounded by the lengthy duration required to accumulate and label sufficient data, which creates a significant lag in the response to emerging fraudulent activities. The problem is further exacerbated in the Zambian context, where resources for extensive audit activities are often constrained, leading to a sparse dataset that fails to reflect the current and complete picture of tax compliance behaviour. Traditional methods of fraud detection are often slow, complex, difficult, and resource-intensive, prompting the need for more efficient approaches.

1.4 Aim of the Study

To create a model that tax auditors can utilize to forecast and detect taxpayers who submit fraudulent returns.

1.5 Research Objectives

i To develop a data mining model that can be used by auditors as a recommender system to identify taxpayers filing false returns based on the assessment.

ii To evaluate and validate the performance and accuracy of this model.

iii To develop a web-based prototype that uses machine learning to automatically help auditors predict taxpayers as filing false returns.

1.6 Research Questions

i What factors or datasets will be required to develop the data mining model and How to develop a data mining model to help achieve the objective(i)?

ii What methods can be used to determine and evaluate this model with the highest accuracy?

iii To what extent can we develop a web-based prototype that helps auditors predict taxpayers' filing false returns?

iv

1.7 Significance of the Study

The findings of this study will help ZRA auditors improve the use of resources, i.e., time and money, in detecting non-compliant taxpayers who are submitting false returns by under-reporting their profits to reduce taxes quickly and easily, thus improving productivity. An automated and data-driven strategy utilizing machine learning methods, including K-means clustering, is suggested as a solution to this issue. Data complexity, resource limitations, newly developing fraud techniques, and timeliness are important facets of the recommended strategy that involves deploying K-means clustering and other machine-learning techniques for detecting tax fraud. Tax authorities can proactively identify and address tax fraud by examining transaction data and spotting clusters of possibly fraudulent actions. The goals consist of creating a data pre-processing pipeline to prepare VAT datasets for analysis, the use of K-means clustering and other machine learning techniques to identify groups of potentially fraudulent transactions or taxpayers, and the implementation of anomaly detection techniques to flag outliers as potential fraud cases. In addition, it will help reduce tax evasion and tax fraud significantly.

1.8 Scope of the study

This study aims to solely detect fraud within the realm of tax administration at the Zambian Revenue Authority (ZRA), with a specific emphasis on VAT tax type using taxpayer's data for sales and refund declarations in a period of four years (2019-2023). Data mining will be used as a tool to mine data specifically the classification method as the most dominant in fraud detection, Given the constraints in availability, K-means Clustering was employed to forecast transactions as either fraudulent or non-fraudulent of labelled datasets which are used to train datasets. For ethical considerations, the dataset was encoded and stored in a secure database location.

1.9 Organisation of the Dissertation

The dissertation unfolds across five distinct chapters:

In Chapter One, the groundwork is laid with an introduction delving into the dissertation's context. It encompasses a thorough discussion of the problem statement, aims, objectives, research inquiries, scope, and significance of the study. Chapter Two embarks on a comprehensive exploration of existing literature by various scholars about the subject matter. It meticulously identifies their discoveries while pinpointing any gaps in the ongoing research. Moving forward, Chapter Three delineates the methodological approach adopted for the study. This encompasses a detailed exposition of the research design, data collection methodologies, analysis techniques, proposed research methodologies, hypotheses, and ethical considerations. Chapter Four takes center stage by offering an in-depth analysis of the amassed data. It rigorously scrutinizes and presents the outcomes of the hypotheses delineated in Chapter Three. Furthermore, it engages in the interpretation, discussion, and conclusion drawing based on the research findings. Lastly, Chapter Five circles back to the study questions introduced in Chapter One. It synthesizes the discourse by furnishing conclusive remarks and recommendations predicated upon the study's discoveries.

1.10 Chapter Summary

This chapter offers an overview and presents a problem statement regarding the detection of tax fraud in Zambia, along with an overview of the current methods employed by the tax administration to address this issue. The focus of the study was highlighted, followed by the outline of objectives designed to tackle the research questions. Finally, the importance of the study was detailed, identifying who would benefit from the research.

2 LITERATURE REVIEW

2.1 Introduction

This chapter begins by emphasizing the issues and challenges associated with fraud in the tax administration sector, specifically in Zambia, and proceeds to discuss the economic impact it has in Zambia; we look at the traditional methods used for fraud detection and introduce the concept of data-driven models and how they can be used to improve the efficiency of the traditional methods. Furthermore, we review the theoretical concepts of the data-driven DM methods applied in this work, interpolate key ideas related to this work, and talk about the methodology and technologies applied in data mining concerning fraud detection as a solution to this problem. We also examine current research on how data mining applications are being developed and integrated as systems in different sectors. Lastly, we review related academic works in areas of application fraud detection and present key findings and identify discrepancies among related studies, then provide a summary of the chapter.

2.2 Background to the Study

This section provides a review of the literature concerning the concept of tax fraud detection within the Zambian context, exploring the current methods employed by the Zambia Revenue Authority (ZRA) to identify tax fraud cases, the challenges encountered in this process, and examining current literature on methods and technologies available in addressing these challenges.

2.2.1 Tax Fraud in Zambia

Tax fraud in Zambia poses a significant challenge to its economic integrity and development. As a developing country with pressing social and economic needs, Zambia relies on tax revenues to finance its public sector and development initiatives. However, the prevalence of tax fraud significantly undermines these efforts, affecting the country's economic health and its prospects for growth. [11]

In Zambia, tax fraud typically manifests through false invoicing, underreporting of income, smuggling, and corruption. False invoicing is common in cross-border trade, where importers and exporters declare incorrect values to reduce tax liabilities. Underreporting on income, particularly by businesses and self-employed individuals, is widespread, as it is often difficult for the Zambia Revenue Authority (ZRA) to accurately assess income levels in a largely informal economy. Smuggling, especially of high-duty items such as tobacco and alcohol, is another form of tax evasion, leading to significant revenue losses for the government.

Additionally, corruption at various levels can facilitate tax fraud, with officials sometimes complicit in the activities of tax evaders. [12] The motivations for tax fraud in Zambia are complex. Economic hardship is a significant factor; individuals and businesses may resort to fraud to survive in a challenging economic climate. Furthermore, the perception of a high tax burden and the complexity of tax laws may drive some taxpayers towards evasion as a means of reducing perceived inequities. The administrative capacity of the ZRA also plays a role. Limited resources for enforcement and the difficulty of tracking informal economic activity can make tax evasion seem low risk, thus more enticing. Additionally, societal norms and the lack of robust legal repercussions contribute to a culture where tax fraud is not always seen as a severe offense. [13] [11]

Research indicates that from 2001 to 2010, Zambia experienced a loss of US$8.8 billion due to illicit capital flows, including tax evasion [14] [6]. The economic impact of tax fraud in Zambia is profound. Revenue losses from tax fraud constrain the government's ability to invest in essential public services, such as health, education, and infrastructure. These sectors are critically underfunded, which hampers human capital development and economic growth. Moreover, tax fraud creates an uneven playing field where honest businesses are at a disadvantage against those evading taxes. This distortion affects investment decisions, discourages new entrants into the market, and undermines overall economic efficiency. [15] As Zambia continues to engage with international investors and donors, the presence of pervasive tax fraud can deter investment and affect international relations. Donor countries and organisations demand transparency and fiscal accountability, which are compromised by widespread tax evasion. [16]

In the Zambian context, traditional methods of tax fraud detection largely echo global practices, albeit tailored to local conditions and regulatory frameworks. The Zambia Revenue Authority (ZRA) employs audits as a primary tool, which involves scrutinising the financial statements and accounts of individuals and businesses to ensure compliance with tax laws. [17] These audits may be random or based on specific criteria, such as businesses in sectors with a high risk of non-compliance or those with large, complex transactions. [18] Investigations in Zambia are typically more targeted and may stem from tips, whistle-blower information, or discrepancies unearthed during audits. The ZRA's investigations might also focus on cross-border activities, given that smuggling and under-invoicing in trade are prevalent issues. [6]

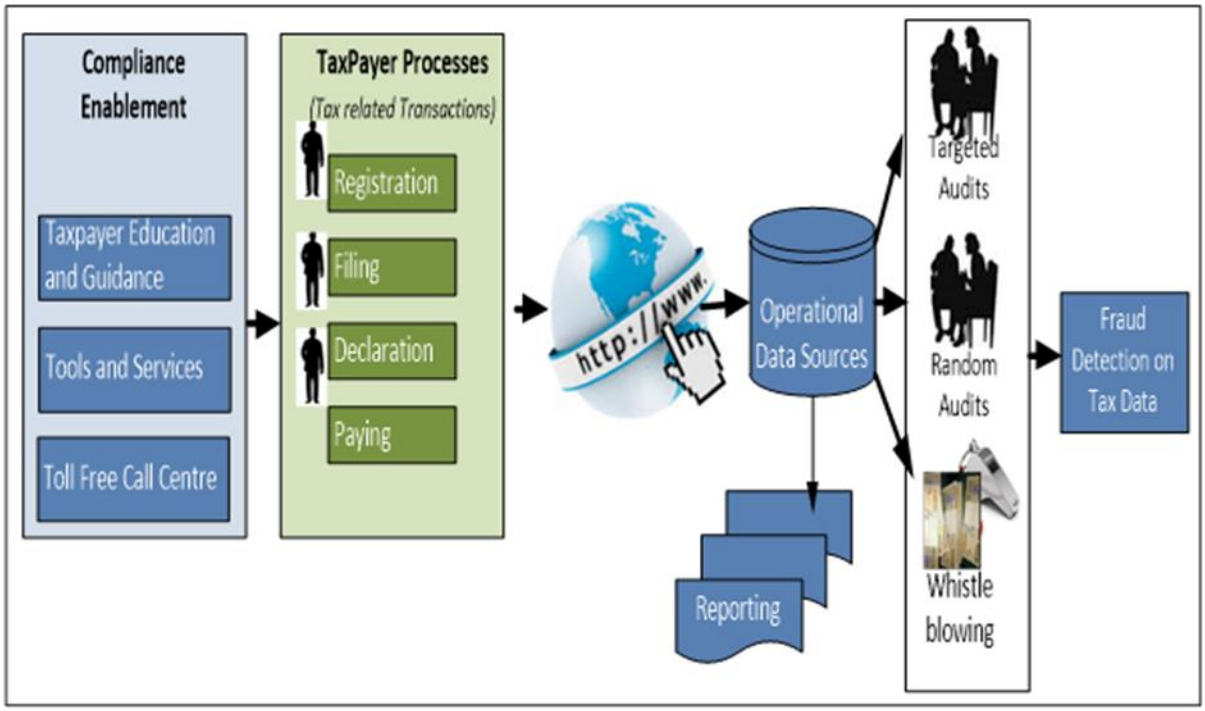

Figure 1: Traditional Method of Tax Fraud Detection [6]

The ZRA also conducts sector-specific inspections, particularly in mining, a sector that contributes significantly to the Zambian economy, to detect and deter transfer pricing abuses and other forms of tax avoidance or evasion. [19] However, these traditional methods are sometimes hampered by constraints such as limited resources, the need for specialised training, and the evolving sophistication of evasion strategies, all of which require ongoing adaptation by the ZRA. [20] This can be seen in Fig 2, by the number of audits conducted by the tax authorities each year in this case 2014, only 6,724 audits were conducted.

Table 14: Audit Activity in 2015

Large Taxpayer Office: Mining					
Type of Audit	**Number of Audits**	**Tax collected (K' Million)**	**Deferred (K' Million)**	**Total (K' Million)**	**Penalties (K' Million)**
Income tax	23	2.57	0	2.57	0.08
PAYE	0	0	0	0	0
VAT	198	2.80	86.80	89.60	0.26
Integrated[1]	42	731.25	0	731.25	314.28
Sub Total	**263**	**736.62**	**86.80**	**823.42**	**314.62**
Large Taxpayer Office: Non-Mining					
Type of Audit	**Number of Audits**	**Tax collected (K' Million)**	**Deferred (K' Million)**	**Total (K' Million)**	**Penalties (K' Million)**
Income tax	73	29.62	19.80	49.42	10.26
PAYE	22	1.97	0	1.97	1.04
VAT	325	52.82	14.86	67.68	15.75
Integrated	64	247.35	1.30	248.65	90.53
Sub Total	**484**	**331.76**	**35.96**	**367.72**	**117.58**
Medium Taxpayer Office					
Type of Audit	**Number of Audits**	**Tax collected (K' Million)**	**Deferred (K' Million)**	**Total (K' Million)**	**Penalties (K' Million)**
Income tax	121	32.42	33.86	66.28	23.79
PAYE	20	1.68	0.22	1.90	0.84
VAT	3814	359.98	63.27	423.25	14.36
Integrated	261	29.65	0.30	29.95	5.68
Sub Total	**4216**	**423.73**	**97.65**	**521.38**	**44.67**
Small Taxpayer Office					
Type of Audit	**Number of Audits**	**Tax collected (K' Million)**	**Deferred (K' Million)**	**Total (K' Million)**	**Penalties (K' Million)**
Income tax	1750	3.34	0.00	3.34	2.75
PAYE	21	8.32	0.00	8.32	0.23
VAT	0	0.00	0.00	0.00	0.00
Integrated	0	0.00	0.00	0.00	0.00
Sub Total	1771	11.66	0.00	11.66	2.98
Grand Total	6734	1,503.77	220.41	1,724.18	479.85

Figure 2: Audits Conducted by ZRA [21]

The ZRA tax administration must use its limited resources to maximise tax compliance. Therefore, the proposal is to use data mining applications to achieve maximum tax compliance. Data mining may be an effective tool for enhancing the efficiency and effectiveness of the detection of illegal tax evasion [22]

2.2.2 Big Data and ZRA Tax Administration

The Zambia Revenue Authority is responsible for collecting revenue on behalf of the Government of the Republic of Zambia. Among its primary duties are:

1. Accurate assessment and timely collection of taxes and duties.
2. Ensuring proper accountability and banking of all collected funds.
3. Furnishing the Government with statistical revenue data.

To achieve its goals, the ZRA has implemented automation in various processes. [6]

In 2013, the Zambia Revenue Authority (ZRA) launched the Taxonline System, a digital platform for tax administration. [23] This initiative aimed to replace an outdated and disjointed system that handled tax registrations, returns, and payments for only three tax types manually, relying on paper forms. The implementation of these systems was geared towards reducing the burden of tax compliance for taxpayers, improving service delivery, and enhancing tax compliance, ultimately leading to increased revenue collection. [24] [25] Automated systems have demonstrated their ability to significantly enhance efficiency in business processes, leading to heightened revenue collections and mitigating instances of tax evasion and avoidance. [26] [27] The advantages of minimizing tax evasion and avoidance become significant when data analytics are employed as additional resources for tax data analysis and detecting tax omissions. [28]

Given the presence of such electronic tax administration comes the vast availability of data [29]. Large amounts of data are generated on a day-to-day basis which are stored on the system's databases [30] In an age where organizations possess abundant data, the real value is found in the capability to gather, organize, and analyze this data to extract actionable business intelligence (BI). [31] Big data is increasingly viewed as a novel resource, serving as assets in business operations and as a major driver of productivity growth in the future. [32]

The handling of data in the realm of Big Data revolves around the principle of the five 'V's [33] which are described as:

i. Variety: Big data originates from a diverse array of sources and is categorized into three categories: structured, semi-structured, and unstructured. Structured data is organized within a data warehouse with predefined tags, facilitating easy sorting. Unstructured data, on the other hand, lacks organization and poses challenges for analysis due to its randomness. Semi-structured data doesn't adhere to rigid fields but includes tags to delineate data elements. [34]

ii. Volume: Refers to the significant quantity of data generated, often stored in terabytes and petabytes [35]

iii. Velocity: Because of the extensive volume and variety of data, all processing must be agile to produce the necessary information.

iv. Value This pertains to the "useful information" that can be derived from the data. [36]

v. Veracity: This directly concerns the reliability of the data.

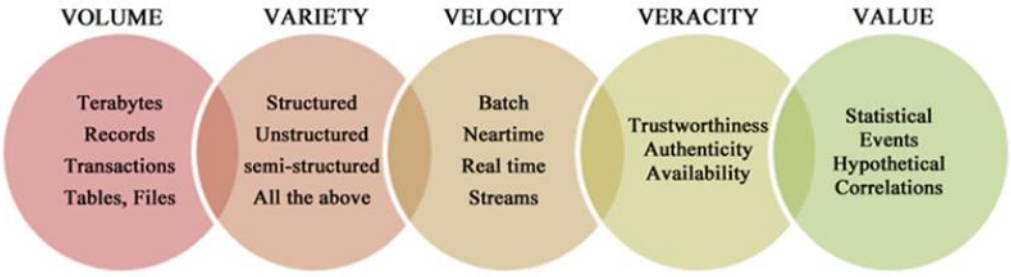

Figure 3: Five 'V' of Big Data [35]

The importance of big data and its influence on organizational effectiveness and business intelligence lies in its potential to be transformed into easily interpretable and shareable information. This information can then be utilized for decision-making processes concerning the organization's performance. [37]. Furthermore, at the business environment level, big data is distinguished by its capacity for collection, analysis, and conversion into various graphics and patterns. These capabilities greatly aid in decision-making processes. [38] [39]

"Big data and business intelligence converge on the notion that business intelligence facilitates the organization and analysis of vast datasets, presenting them in an understandable format that fosters sharing and, consequently, facilitates strategic decision-making. Moreover, it assists in the continual updating of both financial and non-financial data, leading to more informed decision-making." [40]. Business Intelligence (BI) incorporates software functionalities including the processes of Extraction, Transformation, and Loading (ETL), data warehousing, database queries and reporting, multidimensional/online analytical processing (OLAP) for data analysis, data mining, and visualization. [41] which are considered components of BI models and are described as:

i. Data Warehousing:

Data warehousing entails consolidating data from diverse sources into a centralized repository. This step is pivotal for Business Intelligence (BI) as it guarantees data integrity and

consistency. [42] In the context of ZRA, a data warehouse could consolidate information from taxpayer filings, payment records, and other relevant data sources, creating a comprehensive view of the tax landscape.

ii. Data Mining:

Data mining involves uncovering predictive insights from large databases. It is a potent tool for identifying patterns and relationships that would otherwise remain undiscovered. ZRA could utilise data mining to detect anomalies in tax filings, identify potential cases of tax evasion, and understand taxpayer behaviour more deeply. Which is the focus of this study.

iii. OLAP (Online Analytical Processing):

OLAP, or Online Analytical Processing allows business users to navigate data efficiently using advanced tools, allowing them to slice and dice information based on various dimensions like time or hierarchies [43]. It facilitates rapid, uniform, and interactive analysis of data from diverse viewpoints. [44] For ZRA, OLAP can enable the analysis of tax data across various dimensions, such as geographical regions, taxpayer categories, and periods, thereby aiding in strategic planning and decision-making.

iv. Reporting and Visualisation:

This involves the translation of complex data into easily understandable reports and visual formats, such as graphs and charts. Effective reporting can help ZRA communicate insights to stakeholders and support transparency in tax collection and administration.

v. Advanced Analytics:

This encompasses techniques like predictive analytics and machine learning. They can forecast future trends based on historical data. For ZRA, this could mean predicting future tax revenue, identifying sectors with the highest risk of non-compliance, and optimising audit selections.

2.2.3 Data Mining

Data mining, a component of BI, in tax administration like that of the Zambia Revenue Authority (ZRA), transforms the extensive databases generated by online tax systems into actionable insights. The essence of data mining lies in its capacity to sift through immense quantities of data to identify patterns, anomalies, and correlations that might not be evident at a cursory glance or through traditional analysis methods. [45]

The application of data mining by tax authorities, such as ZRA, heralds a revolution in tax administration. With the transition to the 'TaxOnline' system, not only has the process of tax assessment, remittance, and collection become streamlined, but it has also democratised tax compliance, enabling taxpayers to meet their obligations efficiently and at their convenience. [46] [47]The collateral benefit of this digital shift is the accumulation of vast datasets within the tax authority's databases. These datasets hold within them a treasure trove of information that, if analysed effectively, can offer profound insights into taxpayer behaviours, compliance patterns, and potential areas of fraud. [28].

Data mining employs sophisticated algorithms and analytical tools to delve into this rich data repository. It involves several key steps: data preparation, where data is cleaned and transformed; pattern discovery, where algorithms are used to detect structures within the data; and knowledge deployment, where the results are translated into operational strategies. [48]Through this process, ZRA can identify trends and patterns in taxpayer data that could indicate compliance issues or opportunities for improving tax policies.

Data mining involves extracting necessary data from a vast database. It represents a new interdisciplinary field within computer science, focused on automating extraction and generating predictive insights from extensive databases. Data mining aims to uncover hidden information and patterns within repositories.[48] The data mining process employs a range of analytical tools to discern relationships among data within a large database and involves a range of technical approaches, including machine learning, statistics, and database systems. The goal of data mining is to derive insights from large databases and translate them into a format that is comprehensible to humans. Data mining and knowledge discovery play crucial roles in its decision-making process.[49]

The process for Knowledge Discovery in Databases (KDD), as outlined by [50], encompasses the following stages:

1. Familiarizing with the application domain involves acquiring appropriate prior knowledge and understanding the application's objectives.
2. Establishing a target database entails selecting a dataset for conducting the discovery process.
3. Data cleaning and integration involve eliminating noise and inconsistent data, as well as addressing missing data where necessary.

4. Evaluation and presentation encompass selecting relevant data from the dataset by identifying useful features pertinent to the analysis task. This involves dimensionality reduction and transforming data into a suitable format for mining operations.
5. Data mining involves applying intelligent techniques capable of uncovering insights of interest within a specific dataset and generating potentially useful models.
6. Interpretation involves elucidating the discovered patterns by interpreting the interestingness score of each pattern into user-understandable terms based on relevant measures.
7. Knowledge representation involves visualizing and representing the discovered knowledge and performance derived from data mining results, enabling informed actions based on the acquired knowledge. Fig 4 illustrates the KDD stages.

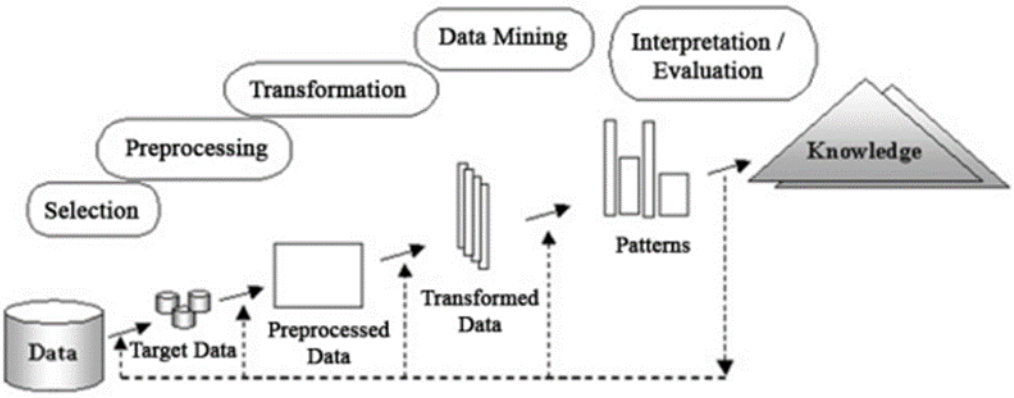

Figure 4: Knowledge discovery in Database [45]

2.3 Machine Learning

This section explores the literature surrounding the concept of machine learning and the approaches to ML involved, including supervised and unsupervised machine learning, an essential aspect of objective 1. We explore the complexities of utilizing machine learning to create data-driven models, emphasizing advancements in this field.

Machine learning is a learning method that automates the acquisition of knowledge, in this case, the mining of data. [51] At its simplest form, machine learning denotes any computer program capable of self-learning, without requiring explicit programming by a human. [52] Machine learning, a vital branch of artificial intelligence, enables computers to learn from data and autonomously make decisions without explicit programming for each new scenario. Its

essence lies in enabling autonomous knowledge acquisition, making it particularly significant in the field of data mining. [52]

To understand how machine learning relates to data mining, we must first recognise the distinct roles they play. Data mining is the process by which patterns are discovered within extensive sets of data. It involves extracting these patterns and transforming raw data into useful information. Machine learning employs algorithms to analyze data, learn from it, and then decide or prediction about something in the world. Therefore, while data mining seeks patterns, machine learning uses those patterns to influence decisions and learn from outcomes. [53] The relationship between the two is symbiotic. Data mining sifts through massive datasets to find connections and uncover patterns that may not be immediately apparent. These findings then become the foundation for machine learning models. With the insights gleaned from data mining, machine learning algorithms improve their predictive accuracy. [54]

In machine learning, "learning" refers to the ability of systems to enhance their performance on a task progressively with experience. [55] Consider, for instance, an algorithm that recommends products to online shoppers. Initially, its recommendations may be based on simple rules drawn from observed consumer behaviour. However, as it accumulates more data about preferences and purchasing habits, the machine learning model can refine its recommendations, becoming more personalised and accurate. The process of machine learning can be broken down into several types:

2.3.1 Supervised Machine Learning

Supervised learning, where the algorithm learns from a training set of labelled data, is analogous to a student learning under the guidance of a teacher. The teacher provides examples from which the student can learn and make predictions. In supervised learning, models are trained on labelled data, equipping them to predict outcomes based on prior knowledge. These models, which encompass methods like decision trees, support vector machines, and neural networks, thrive on classification and regression tasks.

i. Classification tasks involve algorithms tailored to handle classification tasks, where the output variable is categorical, such as yes or no, true or false, among others. This approach is practically applied in activities such as spam detection and social media filtering.[56]

ii. Regression tasks, on the other hand, employ algorithms designed for regression problems characterized by a direct correlation between input and output variables. These

algorithms forecast continuous output variables, with examples including weather prediction and market trend analysis. [56] They serve as the backbone for deciphering structured data, predicting future events, or categorizing data into various categories. [57]

2.3.2 Unsupervised Machine Learning

Unsupervised learning is like a student learning through observation without guidance, clustering information based on the trends and relationships identified within the input data. Unsupervised learning deals with the unknown. It takes unlabelled data and finds hidden structures within it, such as clustering algorithms that group data based on inherent similarities. [28] Dimensionality reduction is another unsupervised technique, simplifying the complex multi-dimensional data into more manageable and insightful forms. Consequently, unsupervised machine learning is divided into two types.:

i. Clustering entails assembling objects into groups based on certain criteria, like the similarities or distinctions among the objects. For instance, this can involve categorizing customers according to their buying patterns. [56]

ii. Dimension reduction: These techniques primarily focus on reducing the variable space by either selecting variables or constructing new variables as combinations of the original ones. [58]

Fig 5 gives a highlight of Machine learning's two major category types, where each of these categories has subtypes and the algorithms under these subtypes.

2.3.3 Semi-Supervised Machine Learning

Semi-supervised learning blends these approaches, merging a small quantity of labeled data with a greater amount of unlabeled data to build more accurate models. This method is particularly valuable when labelling data is expensive or time-consuming, and it capitalises on the structure found in unlabelled data to strengthen the learning model. [28]

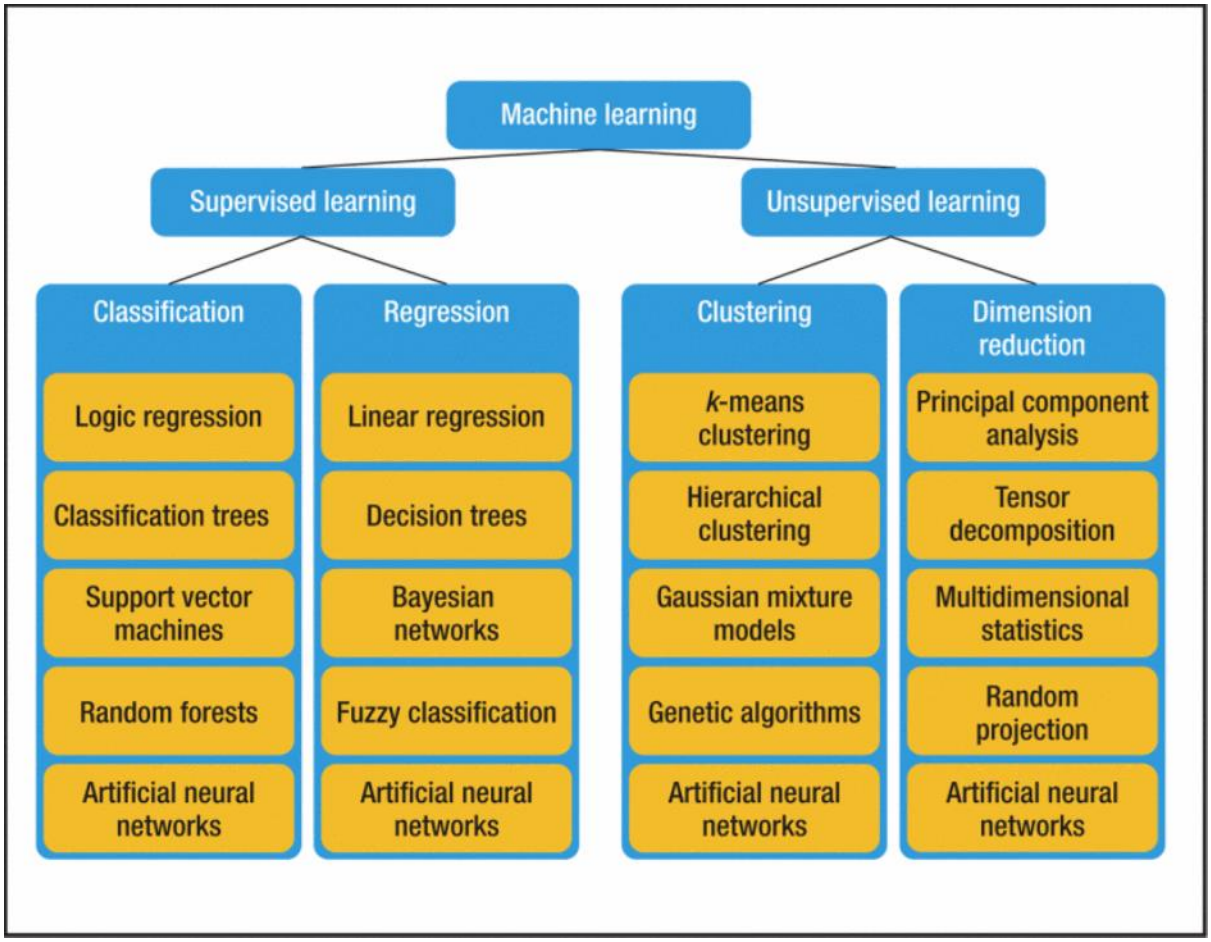

Figure 5: Machine learning and its classifications [28]

Machine learning algorithms are also used to automate complex decision-making processes. For example, in financial services, machine learning models are trained to detect fraudulent transactions by recognising patterns that deviate from legitimate behaviour. These models are continually refined as they are exposed to more transaction data, learning to discern between false positives and genuine instances of fraud with greater precision. Moreover, machine learning extends beyond mere data analysis to predictive analytics. Predictive analytics leverages machine learning for analyzing both present and past data to forecast future events or other unknown outcomes. Here, data mining extracts the relevant patterns and input variables from the data, and machine learning takes over to predict outcomes, adjusting its predictions as more data becomes available. Furthermore, machine learning and data mining are instrumental in handling big data – enormous datasets that exceed the capabilities of traditional data-processing software due to their size and complexity. These fields enable the analysis of big data in ways that reveal trends, patterns, and relationships, especially relating to human behaviours and interactions, which can be invaluable for businesses and researchers. [59]

2.4 Performance Evaluation of Machine Learning Models

This section explores the literature surrounding the performance evaluation of supervised and unsupervised machine learning models, an essential aspect of objective 2. It delves into the intricacies of how these models are assessed in various environments, highlighting the challenges and advancements in the field.

2.4.1 Performance Metrics and Evaluation in Supervised Learning

The capacity of supervised machine learning models to accurately anticipate or categorize data using labeled training data is usually used to assess the model. Numerous metrics, such as accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC-ROC), are commonly cited as employed in this context in the literature. [60]

This study referenced as, [61] centers on evaluating the efficacy of supervised machine learning models in classification, a fundamental technique within data mining. Various classification algorithms are applied to diverse datasets to comprehend and enhance their performance. Prominent algorithms like ID3, Naive Bayes, Multilayer Perceptron, and K-Nearest Neighbor are examined. The paper underscores the use of two primary criteria for algorithm performance assessment: metrics such as the confusion matrix and receiver operating characteristic (ROC) curves. The confusion matrix aids in gauging prediction accuracy by detailing the count of correct and incorrect classifications, including metrics like True Positive Rate, False Positive Rate, True Negative Rate, and False Negative Rate. Additionally, parameters like Precision, F-measure, and Error Rate derived from the matrix offer deeper insights into model performance. ROC curves offer both visual and quantitative methods for evaluating classifier output quality by plotting True Positive Rate against False Positive Rate, offering a comprehensive understanding of sensitivity and specificity trade-offs within the models.

In another study referenced as [62] researchers conducted an empirical comparative analysis to assess evaluating the performance of three data mining classification algorithms—Decision Tree, Multi-Layer Perceptron (MLP) Neural Network, and Naïve Bayes—using two critical performance indicators: classification/prediction accuracy and training time. The methodology involves experimentation with the three selected algorithms across simulated datasets of varying sizes. Performance metrics, comprising model construction time (training time) and accuracy rate (correct classifications), were compared. The findings reveal that Naïve Bayes demonstrates the shortest training time, although it displays lower accuracy compared to the

MLP and Decision Tree algorithms. Furthermore, the results demonstrate a balance between accuracy and training time among these algorithms.

2.4.2 Performance Metrics and Evaluation in Unsupervised Learning

Evaluating unsupervised learning models poses unique challenges, as these models seek to uncover hidden patterns or structures in data without labelled outcomes for guidance. The literature indicates a reliance on methods like silhouette scores, the Davies-Bouldin index, and the Calinski-Harabasz index to assess clustering performance. [63] In this section, we review some of this literature:

In their comparative analysis of clustering algorithms, [64] focused on evaluating the performance of K-Means, Farthest First, and Hierarchical clustering algorithms, using three distinct datasets: Wine, Haberman, and Iris. The primary criterion for performance evaluation is the algorithm's ability to accurately form class-wise clusters. They implement this analysis through the WEKA interface, assessing the proportion of correctly versus incorrectly clustered instances. The study uniquely considers both scenarios with and without the application of Principal Component Analysis (PCA) as a filter, providing a dual perspective on algorithm effectiveness. This approach enables a thorough examination of each algorithm's clustering accuracy, quantified by the percentage of error in clustering instances and the root mean square error. The outcomes of their investigation reveal varying degrees of performance across the algorithms, with K-Means outperforming Farthest First and hierarchical clustering in terms of lower error rates and more accurate cluster formation. The inclusion of PCA as a comparative parameter further enriches the analysis, illustrating its impact on enhancing the clustering performance of these algorithms. This study's comprehensive approach to evaluating clustering algorithms provides significant insights into their applicability and effectiveness in different data mining contexts, highlighting the critical role of accurate cluster formation in the broader field of data mining.

In another research, [5] the validation process in the study on detecting under-reporting tax declarations involved a two-step approach. First, the quality of constructed clusters was reviewed, ensuring statistical differences in the features of declarations within each cluster. They confirmed this using a two-sample Kolmogorov-Smirnov test. Second, an expert review by a tax auditor was conducted. Ten building construction declarations were evaluated, with a

focus on identifying potential under-reporting. This expert review sought to evaluate the practical efficacy of the model in identifying suspicious tax declarations. The validation demonstrated the model's potential in aiding auditors to prioritise investigations for potential under-reporting. The method used for detecting under-reporting tax declarations was chosen due to the challenges in traditional supervised analysis, especially the lack of labelled data for training. Tax authorities often don't have enough audited (labelled) cases, making it hard to apply conventional supervised learning techniques effectively. Therefore, an unsupervised methodology was proposed to screen suspicious tax declarations without relying on prior auditing data. This approach involves clustering similar tax declarations, predicting the probability distribution of stated tax bases within each cluster, and then identifying outliers as potential under-reporting cases. This method allows tax authorities to prioritise audits more effectively by focusing on the most suspicious cases identified through this data-driven approach.

2.5 Machine Learning in Application Systems

The adaptability of machine learning is showcased by its applications in various domains including Healthcare, Financial services, e-commerce, and Manufacturing. In this section, we look at existing literature on systems in these various domains that have been developed using machine learning models which can help us meet objective 3.

2.5.1 The Application of Machine Learning in Health Care

In this paper [65] they present a study on developing a machine learning-based web system for heart disease prediction, specifically the K-Nearest Neighbours (KNN) algorithm, with an emphasis on improving prediction accuracy through data pre-processing techniques like Z-score normalisation. The research highlights heart disease as a leading cause of death in Indonesia and stresses the importance of early detection and prevention. They developed the system using the Flask micro-framework, a lightweight framework based on Python, chosen for its efficiency and resource management capabilities. The paper underscores the significance of machine learning in improving health services and its role in medical diagnostics. The KNN algorithm’s suitability for this application is highlighted, considering its simplicity, effectiveness with large datasets, and compatibility with the Flask framework.

In a separate study, referenced as[66] the emphasis lies in creating machine learning (ML)--based classifier models to detect diabetes early, leveraging clinical data. The research assesses a range of supervised machine learning algorithms, such as Decision Tree (DT), Naive Bayes (NB), k-nearest Neighbor (KNN), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), and Support Vector Machine (SVM). Highlighting the significance of effective pre-processing, the research utilizes methods like label encoding and normalization to improve model precision. Additionally, feature selection methods are utilized to identify and prioritize risk factors. Subsequently, the most accurate ML algorithm is selected for integration into a web application developed using the Python Flask framework. This web application aims to provide accessible diabetes prediction to users.

2.5.2 The Application of Machine Learning in Finance

In this paper [67] they present a study on stock market trend prediction using machine learning algorithms, with a focus on reducing investment risks. The research compares the performance of three machine learning models – Linear Regression, Random Forest, and K-Nearest Neighbours (KNN) – in forecasting the movement of stocks in four market groups: diversified financial, petroleum, non-metallic minerals, and basic metals. They develop a web application for stock market prediction operates by integrating machine learning algorithms to analyse and predict stock market trends. Users interact with the application through a web interface where they can input or select parameters related to the stocks, they are interested in. Using the user's input and the model's analysis, the application anticipates future stock market trends or prices.

Likewise, another study referenced as [68] the objective is to develop a website application for stock price prediction in Indonesia, focusing on the top five banks by market capitalization. The research employs the Long and Short-Term Memory (LSTM) deep learning algorithm, renowned for its capability to handle complex data and capture long-term dependencies. The LSTM algorithm is integrated into a website using the Flask framework, selected for its adaptability and resource efficiency. The research focuses on stocks from Bank Central Asia, Bank Rakyat Indonesia, Bank Mandiri, Bank Negara Indonesia, and Bank Syariah Indonesia. The website showcases prediction outcomes in an accessible format, with the LSTM models achieving a Mean Absolute Percentage Error (MAPE) below 10%, indicating high accuracy.

2.5.3 The Application of Machine Learning in E-commerce

In this study referenced as [69] the research delves into the creation of a Movie Recommender System, combining progress in Data Science, Machine Learning, Deep Learning, and Artificial

Intelligence. Utilizing datasets from Kaggle, the system employs diverse filtering algorithms including demographic filtering, content-based filtering, and emotion/mood-based filtering facilitated through web scraping. The development of the system unfolds through several stages, commencing with dataset acquisition from Kaggle, followed by the generation of general recommendations for all users, and subsequently offering personalized, mood-based recommendations. The system's framework relies on a Flask Micro Web framework written in Python, linking the recommendation logic to a web interface. Moreover, it transitions into a mobile application via React JavaScript, with a focus on enhancing User Interfaces/User Experiences (UI/UX).

In another paper, [70] introduces a recommender system for online rental property search. Utilizing a preference-based search approach integrated with example-critiquing, the system allows users to input preferences, which are then used to create a tailored model. This model generates a list of properties matching these preferences. Developed as a web application using Ruby on Rails, this system addresses the challenges faced by urban Kenyan dwellers in finding rental properties, enhancing the efficiency and relevance of property searches.

2.5.4 The Application of Machine Learning in Agriculture

This paper [71] addresses the challenge of plant identification faced by botanists, who often need to recognise and classify a vast variety of plants. Traditional methods rely on physical traits like flowers, fruits, and leaves, but the sporadic appearance of flowers and seeds can make identification difficult. The study focuses on developing a system for automatic plant recognition using leaf structure, which is more reliable due to leaves' accessibility and abundance. To build this system, the Mendeley dataset, containing high-quality images of various species, is utilised. The research involves creating a web application, employing the Flask framework, enabling users, such as farmers and gardeners, to upload images of plant leaves. This application then classifies the leaf image using a CNN model and provides the user with information about the plant.

In another paper, [72]they developed a web application as a tool for farmers or farm managers, primarily in olive farming. Its primary purpose is to provide predictive insights into crop yields, leveraging historical data and environmental factors. Key aspects of this tool include its automatic updating of meteorological data, the need for users to input only harvest data, and its user-friendly interface. The application utilizes the Oracle Database Management System 19c, Oracle Data Mining, and Oracle Applications Express (APEX) for its creation. These

technologies enable the swift development of web applications suitable for both desktop and mobile devices.

In summary, the adaptability of machine learning is not only evident in specific case studies, but also its broad applications across different business areas. From healthcare, where machine learning algorithms predict disease outbreaks and enhance diagnostic accuracy, to e-commerce with sophisticated recommendation systems that personalise user experiences. Financial services benefit greatly from machine learning through the fortification of systems against fraud and the enhancement of credit scoring processes. In manufacturing, predictive maintenance and supply chain management is optimised through machine learning algorithms, leading to increased efficiency and reduced operational costs. Agriculture, too, has seen transformative changes with the introduction of machine learning, where precision farming and yield predictions have become more accurate and reliable. This widespread applicability of machine learning demonstrates its adaptability and learning capabilities, which are central to its definition and power. The capacity of machine learning to handle extensive datasets and learn from them allows for continuous improvement and adaptation in various applications. Table 1 offers a comprehensive overview of these applications, demonstrating their varied and transformative impact on machine learning across different business sectors.

Table 1: Machine Learning in Application Areas

Application Areas	Application	Specifics
Healthcare	Disease Outbreak Prediction	Using past and current health records to forecast and prevent potential disease outbreaks. [73]
E-commerce	Recommendation Systems	Analysing customer data to provide personalised product recommendations. [74]
Financial Services	Fraud Detection, Credit Scoring, Stock market predictions	Detecting irregular transaction patterns and assessing lending risks with historical data. [75]
Manufacturing	Predictive Maintenance, Supply Chain Optimisation	Foreseeing potential machine failures and managing inventory and production plans efficiently. [76]
Agriculture	Precision Farming, Yield Prediction	Employing satellite and sensor data for informed farming decisions and resource management. Each area benefits distinctly from the predictive analytics and adaptive learning capabilities of machine learning, illustrating its transformative impact across sectors. [77]

2.6 Related Works and Gaps in the Literature

Numerous publications discussing the application of machine learning in fraud detection (FD) have been reviewed. Various researchers have allocated a remarkable amount of interest to studying Fraud Detection in some domains such as banks, insurance, finance, and tax.

2.6.1 Fraud Detection in Financial Institutions

Various studies have been conducted within financial institutions to identify financial fraud employing data mining methods.

This paper [78] discusses the growing challenge of bank fraud due to advancements in communications and computing. It explores various fraud types, such as credit card fraud, money laundering, and insurance fraud, and the potential of data mining tools for early detection. The paper advocates for the use of Support Vector Machines (SVM) in creating models to distinguish between normal and abnormal customer behaviours, thereby identifying fraudulent transactions. A hybrid approach combining binary and single-class SVM methods is proposed to enhance detection efficiency. They validate the effectiveness of these techniques through credit card transaction databases, demonstrating their superiority in fraud detection compared to other methods. The review also highlights the growing nature of fraud tactics, emphasising the need for adaptive, intelligent systems in banking security.

Mousa [79] provides an overview of different data mining algorithms that can be used in a wide area of financial applications to researchers and organisations interested in this research. The paper presents 65 articles determining which methods are commonly employed for financial fraud detection and the methods that yield the best results. For example, according to this research, the logistic data mining model can help in detecting financial fraud in automobile insurance, corporate insurance, financial statements, and credit cards. From this research, it is noted that there is a large advancement in financial fraud detection using data mining techniques, which shows that data mining is a large contributor to financial fraud detection, which is of benefit to researchers and organisations.

Anuj Sharma [80] specifically provides an overview of the data mining algorithms and techniques that can be used to detect fraud in financial accounting. From their findings, they determine that the regression analysis algorithm is widely used for financial accounting fraud detection because of its great explanation ability. The paper further suggests that using only financial statement data may not be sufficient for the detection of fraud. It recognises the importance of data mining applications in financial accounting fraud detection

2.6.2 Credit Card Fraud Detection

In this research, [81] the paper recognises the pressing need for advanced strategies to detect anomalous transactions and advocates for the integration of ML-based classifiers in financial institutions. It suggests that the implementation of these technologies is vital for the security of banking systems, benefiting stakeholders ranging from bankers to end-users, ultimately leading to safer financial transactions and a reduction in credit card fraud. The paper sets forth the foundation for future research to further enhance the security measures in the banking sector.

This paper [82] investigates credit card fraud detection using a European cardholders' dataset from September 2013, containing a small proportion of fraudulent transactions. Due to the imbalanced nature of the dataset, the Synthetic Minority Over Sampling Technique (SMOTE) is used to improve model accuracy. Various machine learning models are tested after applying a Genetic Algorithm (GA) for feature selection. Results show that the Random Forest (RF) classifier combined with GA-based feature selection achieves up to 99.98% accuracy, outperforming other models and demonstrating the effectiveness of GA in optimising feature selection for fraud detection tasks. The study concludes with plans to further validate the framework using additional datasets.

This study [83] discusses a novel method for detecting credit card fraud through the utilization of data mining techniques and machine learning algorithms. The core methodology employs a Hidden Markov Model (HMM) to model the sequence of credit card transactions. They trained this model on a cardholder's normal transaction behaviour to establish a baseline pattern. If a transaction does not fit this pattern with a high enough probability, they flagged it as potentially fraudulent results suggest that the application of this improved algorithm can aid banks, organisations, and governmental centers in managing transactions and preventing losses due to fraud. The process also includes steps to ensure security by collecting detailed user profiles and setting up security questions. In summary, this paper highlights the innovative use of HMM in conjunction with an enhanced K-Means algorithm for detecting fraud in credit card transactions. This approach emphasises the proposed method's potential for improving security, the systematic approach to pattern recognition, and the direction for future research to further refine these techniques.

2.6.3 Fraud Detection in Tax Administration

Daniel [5]solves the problem of taxpayers who tend to under-report earnings to reduce their tax liabilities in Bogota, Columbia. Because traditional supervised analyses, which are most efficient with vast amounts of data, are not feasible due to the complexity and high cost associated with labelling taxpayers as fraudulent or non-fraudulent, the paper proposes an unsupervised approach for identifying and scoring taxpayers who may be under-reporting. The core assumption is that tax declarations with similar features should result in roughly equivalent tax payments. The proposed model evaluates tax declarations characterised by multiple variables in particular construction and seeks outliers within the declared tax bases.

Makani [6] gives a detailed insight into the methods used by the Zambia Revenue Authority to detect tax anomalies and fraud. Based on a survey, they revealed that ZRA heavily relies on traditional methods, such as random sampling and targeted audits to detect fraud. With this finding, a data-driven tool using an Outlier algorithm method was developed that relies on continuous monitoring of distances and distance-based outlier queries. Underpayments and overpayments based on the business rules were marked as outliers and thus could be considered fraudulent.

Similarly, Vanhoeyveld's [9] the paper presents a systematic method for identifying potential fraud cases among Belgian companies, aiming for efficiency and accuracy in flagging entities for further scrutiny by the tax administration. It focuses on solving the lack of labelled data challenge, meaning supervised learning techniques cannot be used directly to detect fraud. Instead, they use Anomaly Detection (AD) techniques to find companies that deviate significantly from the norm. Others proposed using a hybrid unsupervised method to detect fraud by combining k-means clustering and outlier detection, which is validated by cross-referencing each other. To further improve the accuracy of this method, it allows for the integration of user-inputted domain knowledge from the tax auditors.

Castellón [84] provides a sophisticated approach to tax evasion analysis by integrating detailed tax data with behavioural and historical performance metrics. They offer insight into an analysis of tax information to detect instances of fraud, especially concerning the use of false invoices. Initially, they group companies according to similarities in their behavior, tax variables, and other pertinent characteristics using the self-organizing map (SOM) clustering technique. Subsequently, they identify instances with or without fraud, utilizing Decision Trees, Neural Networks, and Bayesian Networks to identify taxpayers employing fraudulent invoices. In their case study, Neural Networks demonstrated superior performance compared

to the other algorithms, accurately identifying 92% of the cases of fraud. Thus, the study integrates both supervised and unsupervised learning methods to detect fraud.

In a different research study cited as [85] the examination centered on the utilization of Machine Learning (ML) for predicting tax crimes in the municipality of São Paulo. Leveraging a labeled dataset sourced from a comprehensive collection of fiscal audits, various ML techniques were explored, including Neural Networks, Naive Bayes, Decision Trees, Logistic Regression, random forest, and ensemble Learning. Among these methods, the "Random Forests" approach yielded the most favorable outcomes. To facilitate this analysis, they utilized a tool called KNIME. The researchers meticulously selected pertinent features from the dataset to ensure the accuracy of their predictions. They trained their model using historical data and subsequently evaluated its performance on more recent data from 2018. The findings demonstrated promising results, indicating the potential of ML in assisting governments in identifying individuals not adhering to tax regulations. Nonetheless, the study acknowledged certain limitations, such as not considering the impact of time on their results, which could enhance the predictive accuracy. Additionally, they discussed the prospect of employing alternative methods in the future to enhance their research outcomes.

To address the issue of tax evasion, Zheng's study [86] advocated for the implementation of a visual analytic system designed specifically for detecting suspicious affiliated-transaction-based tax evasion (ATTE) groups. This technique involves constructing a network of taxpayer interactions that captures both economic behaviors and complex social connections among taxpayers. The study utilized the random forest-supervised learning algorithm to pinpoint suspicious activities within these ATTE groups. The tool integrates coordinated visualization views and interactive exploration functionalities to facilitate the interpretation of the model's outcomes.

This paper [87]emphasises the importance of detecting tax fraud patterns to prevent them by proposing a systematic approach starting with data pre-processing, feature selection, clustering using K-means, classification via decision trees, and a multilayer feed-forward neural network for training on classified data. However, the paper doesn’t give full details of the data modelling and evaluation of the model.

Claudine’s paper [88] examines VAT fraud in Rwanda, a critical issue affecting the country’s economy. The Rwanda Revenue Authority (RRA) struggles with detecting fraudulent activities among vast numbers of VAT returns. The research proposes a data mining model utilising

Naïve Bayes, K-neighbours, and Decision Trees to distinguish between legitimate and fraudulent transactions with high accuracy, particularly Naïve Bayes at 98%. These findings suggest a promising tool for RRA to enhance audit efficiency and compliance. Despite technological measures, VAT fraud persists, highlighting the need for robust data analysis tools to manage and analyse tax data more effectively for fraud detection. This research paper [89] investigates how Machine Learning (ML) and Multilayer Perceptron (MLP) neural networks can be utilized to identify tax fraud in personal income tax returns in Spain. This research addresses the significant issue of tax evasion in the country, which exceeds 20% of GDP. Traditional methods have struggled to curb this trend, prompting the exploration of artificial intelligence, particularly neural networks, for their capability to handle large databases and complex algorithms. The study utilises accurate and complete data from the Spanish Institute of Fiscal Studies (IEF) and the Spanish Revenue Office. The MLP model, characterised by its feedforward structure with input, hidden, and output layers, aims to classify taxpayers by their likelihood of committing fraud and calculate fraud probability. Implementing this model required advanced software and hardware provided by IBM, to handle the complexity and volume of data. The model was trained on 70% of the data, approximately 2 million records, and tested on the remaining 30%. The MLP model demonstrated an 84.3% efficiency rate in identifying fraudulent and non-fraudulent taxpayers, as evidenced by the sensitivity analysis using the ROC curve. The study concludes that MLP networks offer significant advantages in tax fraud detection, efficiently classifying taxpayers, and determining fraud probability without strict statistical assumptions. This methodology's potential extends to other types of tax fraud detection, indicating a broad application scope. In this study referenced as [90] The paper presents a critique on the application of both supervised and unsupervised learning techniques. Consequently, introduced a novel tax fraud detection framework that integrates elements of both supervised and unsupervised models. This framework is devised to leverage the entirety of tax returns, thereby augmenting the efficacy of fraud detection. The suggested system consists of four modules: a supervised module that uses a tree-based model to extract data knowledge, an unsupervised module for calculating anomaly scores, a behavioral module that assigns compliance scores to each taxpayer, and a prediction module that integrates outputs from the earlier modules to predict the probability of fraud for each return. Although there were challenges in its development, the model demonstrated encouraging outcomes.

In another investigation referenced as [91] the focus is on addressing the significant challenge of tax evasion, particularly in the context of Value-Added Tax (VAT) evasion. The paper

advocates for the utilization of Business Intelligence (BI) and data mining as potent tools to bolster VAT evasion detection efforts. Specifically, the study applies data mining's association rule to VAT databases, to uncover patterns conducive to identifying problematic instances of tax evasion. The overarching objective is to augment tax auditors' capacity to pinpoint and scrutinize suspicious VAT activities more adeptly, consequently enhancing the recovery of tax revenue losses.

Similarly, [92] this paper addresses the importance of detecting fiscal fraud, especially in income tax evasion. The paper specifically explores the application of data mining techniques to develop a model for detecting and predicting fraud in personal income tax data based on information about deductible expenses and reductions. The fundamental techniques used in financial fraud detection include logistic models, artificial neural networks, Bayesian belief networks, and decision trees. The paper introduces a classification model that begins with a principal component analysis (PCA) for dimension reduction, followed by using Multilayer Perceptron (MLP) to assign fraud probabilities, and finally employs decision trees for segmentation. This approach aims to enhance the productivity of tax auditors in identifying and handling cases of tax evasion more efficiently and effectively.

In another research, [93]they focus on detecting income tax fraud using Artificial Neural Networks (ANNs). The study’s primary objective is to identify factors contributing to tax fraud in income tax data in Rwanda. It demonstrates that ANNs can effectively identify tax fraud, achieving an impressive accuracy of 92%, precision of 85%, recall of 99%, and an AUC-ROC of 95%. Factors such as business type, operation period, and size are identified as relevant in detecting income tax fraud.

In this study referenced as, [94] Big Data methods are explored for tax fraud detection, with a specific focus on a novel approach termed HUNOD (Hybrid Unsupervised Outlier Detection). The research addresses the pervasive challenge of tax evasion and avoidance, which is estimated to incur substantial economic costs, amounting to 3.2% of GDP in OECD countries. The HUNOD method represents an innovative unsupervised learning strategy that integrates two outlier detection methodologies grounded in clustering and representational learning. This unique approach enables the integration of pertinent domain knowledge to identify outliers relevant to specific economic contexts. Notably, the method enhances result interpretability by training surrogate models designed for explanation over the outcomes of unsupervised outlier detection methods. The paper's experimental assessment, conducted on datasets sourced from the Tax Administration of Serbia, demonstrates that HUNOD can internally validate between

90% and 98% of outliers, contingent on the clustering configuration and the application of regularisation mechanisms for representational learning. This empirical finding holds significance as it underscores the efficacy of HUNOD in pinpointing tax evasion risks, potentially facilitating the recovery of lost revenue. Consequently, the adoption of HUNOD could support the implementation of more equitable fiscal policies and resource allocation strategies.

In the study referenced as, [55] the focus is on constructing machine learning models utilizing open data to identify tax evasion. The research underscores the novelty of employing open data, as opposed to relying on sensitive datasets prevalent in this field. Incorporating graph neural networks (GNNs), the study juxtaposes their performance against conventional machine learning methodologies such as Random Forest and Multilayer Neural Networks in the context of tax evasion detection. The approach involves defining the identification of tax evaders as a binary classification problem, where a company's inclusion on the state's active debt list serves as an indicator of "evader" behavior. Employing supervised learning techniques, the research compares multiple methodologies including Random Forest, Multilayer Neural Networks, and GNNs. Notably, the results indicate that GNNs did not yield the most optimal outcomes.

2.7 A Summary of the Related Works

The results showed that data mining models are effective in detecting tax fraud. Table 2 highlights the literature review and gaps in the research.

Table 2: Literature Review and Gaps

	Article	Author	findings	Gap
1	"Tax fraud detection for under-reporting is declarations using an unsupervised machine learning approach." [5]	Daniel de Roux, Boris Perez, Andres Moreno, and Maria del Pilar Villami	They use spectral clustering to detect taxpayers' under-reporting declarations. The tax experts were to evaluate the model to a small degree.	The use of Clustering without the combination of other models limits the performance and usefulness of the model.
2	"Fraud detection on bulk tax data using business	Memorie Makani, Jackson Phiri	Based on a survey, they revealed that	A detailed evaluation of the

	intelligence data mining tool: A case of the Zambia Revenue Authority." [6]		ZRA heavily relies on traditional methods, such as random sampling and targeted audits to detect fraud. A data-driven tool using an outlier algorithm is developed.	model's performance is not outlined.
3	"Value-added tax fraud detection with scalable anomaly detection techniques." [9]	Jellis Vanhoeyveld and David Martens	The use of Anomaly detection algorithms gave them promising results	Only an anomaly detection algorithm was used.
4	"Characterization and detection of taxpayers with false invoices using data mining techniques" [84]	Pamela Castellón González and Juan D. Velásquez	In their case study, Neural Networks outperformed other algorithms, successfully identifying 92% of the fraud cases.	Detailed evaluation of the clustering algorithm was not given.
5	"Tax crime prediction with machine learning: a case study in the municipality of São Paulo." [85]	Ippolito A, Lozano ACG	The random forest provided the best results	unsupervised machine learning models are not explored which makes the model not flexible or adaptable to changes in data
6	"ATTENet: Detecting and Explaining Suspicious Tax Evasion Groups." [86]	Qinghua Zheng, Yating Lin, Huan He, Jianfei Ruan, and Bo Dong	A hybrid tool is developed that inputs domain knowledge and integrates it with the supervised learning algorithm	The unsupervised learning model is not considered for enforcement learning
7	"Data Analytics-Based Approach to Tax Evasion Detection."	Yashashwita Shukla, Neena Sidhu, Akshita Jain,	Gives general importance to the	They did not provide the data modelling

		T.B. Patil, S.T. Sawant-Patil	use of data mining techniques.	techniques and evaluation of the model
8	"Value Added Tax Fraud Detection Using Naïve Bayes Data Mining Approach Case Study: Rwanda 2016-2019" [88]	Claudine Munezero	This paper tested and found the Naïve Bayes algorithm to have the highest accuracy of 98%	The use of Supervised learning is a more rigid approach that is not adaptable to changes in the Dataset.
9	"Tax Fraud Detection through Neural Networks: An Application Using a Sample of Personal Income Taxpayers" [89]	César Pérez López, María Jesús Delgado Rodríguez	Multilayer Perceptron (MLP) neural networks to detect tax fraud in personal income tax returns	The labelled data is not representative of the entire population, leading to sample selection bias in the supervised model.
10	"A Multi-Module Machine Learning Approach to Detect Tax Fraud" [90]	N. Alsadhan	Their model surpassed the performance of other models utilizing the original dataset, except for the recall measure for the 'fraud' category when using SVM, and a matching precision score for the 'not fraud' category. Moreover, they observed that, particularly with the original data, all models had difficulties with the recall for the 'not fraud' category.	The integration of supervised and unsupervised models creates a complex system that might be challenging to manage, optimize, and scale. Real-world application is difficult.

11	"Using Data mining technique to enhance tax evasion detection performance"	Roung-Shiunn Wu	They apply the association rule of data mining in this study	Only the Supervised learning model is explored to detect tax fraud
12	"Characterization and detection of potential fraud taxpayers in personal Income Tax using data mining techniques" [92]	Maria del Camino González Vasco, Sonia De Lucas Santos	Multilayer Perceptron (MLP) and Decision Trees are used to detect fraud	Only Supervised learning models are explored to detect tax fraud
13	"Fraud Detection Using Neural Networks: A Case Study of Income Tax" [93]	Murorunkwere, B.F. Tuyishimire, O.; Haughton, D. Nzabanita, J	They used artificial Neural Networks as the algorithm to detect fraud in this research	Only a supervised learning model is explored to detect tax fraud
14	"Tax Evasion Risk Management Using a Hybrid Unsupervised Outlier Detection Method"	Jasna Dusan Jakovetic Atanasijević,	The paper's experimental evaluation of datasets from the Tax Administration of Serbia shows that HUNOD can internally validate between 90% and 98% of outliers	The paper does not provide a detailed description of the dataset used to detect tax evasion and avoidance
15	"Tax evasion identification using open data and artificial intelligence" [55]	Otávio Calaça Xavier, Sandrerley Ramos Pires, Thyago Carvalho Marques, Anderson da Silva Soares	The research employs supervised learning techniques and compares various methods, including Random Forest, Multilayer Neural Networks, and GNNs. Random Forest was the best-performing model	The use of Supervised learning is a more rigid approach that is not adaptable to changes in the Dataset.

2.8 Chapter Summary

In this chapter, we have examined the literature provided by other scholars and researchers, and in line with Claudine's sentiments, "Data mining plays a vital role in fraud detection due to its performance ability to compute a vast data set. Besides, data mining approaches can cope with the complexity of the patterns within a vast amount since they require using of low-cost computation and giving a high-precision outcome since uncovering potential fraud presents a challenging and formidable undertaking." [88]

Our study distinguishes itself from these case studies in several ways. Primarily, we utilize VAT declarations and refunds as the dataset. Secondly, most studies on tax fraud have not incorporated the use of unsupervised machine learning algorithm K-means and anomaly detection as a model to predict tax fraud. Our work aims to contribute to the existing research using data-driven decision-making models in the Zambian context. In this research, we employ machine learning techniques and algorithms to predict taxpayers' filing false returns. As input, we use data from VAT and refund declarations. Our methods encompass the following steps: data collection from the database, data preparation and exploration, data modelling, and evaluation. The results of our case study highlight K-means and local outlier algorithm tax fraud prediction performance and its capability to adapt to new data. We are unaware of any prior research focused on predicting tax fraud within the ZRA tax system based on K-means and local outlier algorithms.

3. RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, the methodologies utilized in the study are deliberated. It encompasses the research design, techniques for data collection and analysis, the proposed research model, and research hypotheses, and addresses ethical considerations.

3.2 Research Design

We anchored the study within quantitative research methods to systematically address the specific objectives. Quantitative research is a systematic empirical investigation through computational, statistical, and mathematical tools. [95] This methodology was instrumental in the development of the model, facilitating the analysis of vast datasets to detect tax fraud efficiently. In the realm of quantitative analysis, we utilised a collection of robust statistical tools and software. The primary tools for statistical computations and data visualisation were the NumPy and pandas' libraries, which are integral to data analysis in Python. We employed the matplotlib library to generate a variety of plots and charts, aiding in the visualisation of complex data and findings.

3.3 CRISP-DM Model

To meet the first Objective, the research followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) to achieve a structured and methodical approach to model development. This six-phase process model offers a detailed framework for executing data mining projects and is widely acknowledged for its orderly and methodical progression.

[96] describes the 6 stages of CRISP-DM namely:

1. Business Understanding: This stage entails establishing an understanding of the research's objectives and requirements from a business perspective, then translating this comprehension into a data mining problem statement and an initial plan intended at fulfill those objectives.
2. Data Understanding: The emphasis in this phase was on gathering initial data, cataloging the data, examining its features, and evaluating its quality to identify possible issues.
3. Data Preparation: This phase often takes up the most time in a project. It encompasses all the necessary tasks to transform the initial raw data into the final dataset, including

selecting cases, managing missing data, constructing variables, and transforming data as needed.

4. Modelling: Various modelling techniques were considered and applied during this phase. Models were built and assessed to ensure they met the requirements of the business understanding phase.
5. Evaluation: Before the final deployment of the model, it was crucial to conduct a thorough evaluation to confirm that it effectively met the business goals established at the beginning.
6. Deployment: The final phase involved deploying the model to a real-world setting, where the end-users could utilise it to identify and act upon the finding

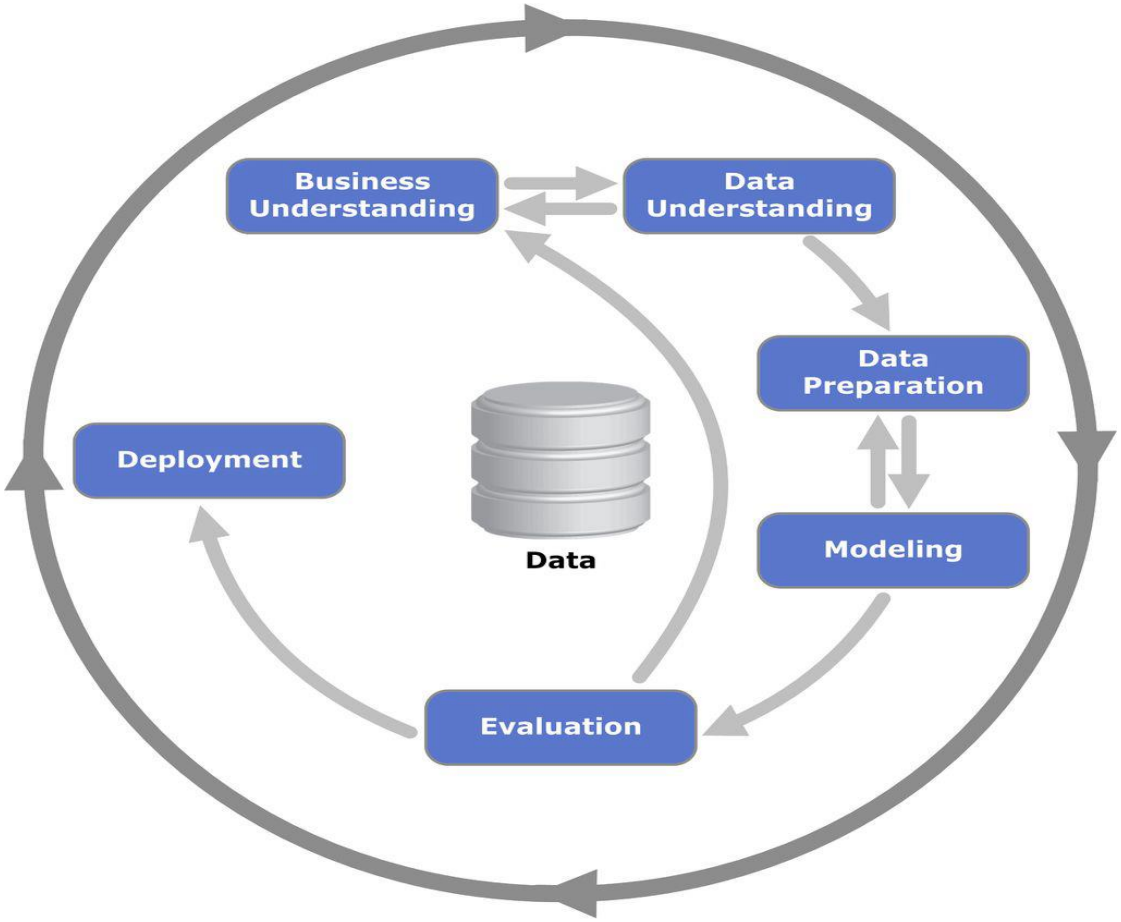

Figure 6: CRISP-DM model process [97] [98]

This research specifically benefits from the quantitative approach due to the need for precision and replicability in fraud detection. The structured nature of quantitative methods, coupled with

the robust framework provided by CRISP-DM, ensured that the developed model is both reliable and valid for application in practical settings.

As a data mining technique, the materials and methods are based on a database repository and data mining technique. Our work was structured around several key stages: understanding and collecting data, cleaning the data, integrating the data, preprocessing the data, creating data mining models, and evaluating the outcomes. In this study, the ZRA VAT fraud detection methodology was implemented using taxpayer's historical VAT and refund data which have been transformed into a suitable required format for our classifier algorithm with the help of data pre-processing together with data exploration to first discover the relationship of the independent variables and extracting useful features. The fraud detection model used VAT data from the operational database to be uploaded into data mining software where data mining techniques were applied.

3.3.1 Data Understanding and Collection

Data collection is crucial in studying Tax Fraud Detection using Unsupervised Machine Learning. Because we are focused on the VAT returns filed by taxpayers, it was important to understand how the VAT system works. Determining what data, we should collect was crucial to help us effectively develop the model. The VAT system allows VAT-registered taxpayers to self-declare their VAT returns. This supplies ZRA with all the necessary information to assess tax liability. The information can be categorized into four main groups: outgoing transactions (sales), incoming transactions (purchases), taxes owed, and deductible taxes. The VAT payable is determined by subtracting the total input tax from the total output tax. If the output tax exceeds the input tax, the difference is payable to the tax authority (ZRA). ZRA pays the taxpayer through the VAT refund process if the input tax is higher. We collected 229,000 VAT declaration records from the tax administration system database for four years based on business understanding. For tax fraud and anomaly detection, One aims to identify entities that exhibit behavior markedly different from what is typically expected within the examined population. The behavior of a company is captured through a set of manually defined variables (assumed indicators of fraud), which are expressed as tax ratios. Based on this data understanding and data extraction, we created 12 features. As we dived into this reservoir of data, ethical considerations were paramount. Taxpayer confidentiality is a serious concern; thus, we obfuscated any identifying information. We anonymised names, specific taxpayer IDs,

and other sensitive details to uphold privacy standards. The resultant dataset was a sanitised yet rich resource, securely stored with access stringently controlled.

This process of data understanding and collection was not merely a preliminary step, but the cornerstone of our study. It ensured that the ensuing analysis was built on a solid foundation of real-world tax declarations, framed within the practical workings of the VAT system. By melding this comprehensive understanding with sophisticated data mining techniques, the study aimed to advance the field of tax fraud detection, setting a precedent for how tax authorities like ZRA could leverage data to safeguard fiscal integrity.

3.3.2 Data Cleaning and Exploration

In the realm of data-driven decision-making, the preliminary steps often include an essential phase known as exploratory data analysis (EDA). This process serves as the groundwork for further analysis and model building. EDA involves a blend of techniques aimed at understanding the makeup of data sets by summarising their key characteristics, often using visual methods. The process started by importing the data into an analytical environment, Python with its libraries, followed by an inspection of the dataset's structure, size, and variable types. [99] [100]. A crucial part of EDA is to identify and handle missing values. Missing data can significantly affect the conclusions drawn from the data. Strategies for handling such gaps include deletion, imputation, or even leveraging the absence of data as an insightful feature itself. Following this, the EDA entails a thorough examination of summary statistics and the distribution of the data. This step can often unveil patterns, anomalies, or errors that warrant further investigation or rectification. [101]. Moreover, during the EDA, it's essential to validate the quality of the data, ensuring it's suitable for the intended analysis. This involves assessing the data for accuracy, completeness, reliability, and relevance. Data quality assessment can save time and resources in later stages and contribute to the credibility of the analysis. The iterative nature of EDA requires revisiting the analysis multiple times, as new insights or patterns may emerge with each review. It's also critical to document findings and insights as they develop. This documentation not only provides a trail of the analytical journey but also supports the reproducibility of the research. Through rigorous EDA, we formulated hypotheses regarding indicators of fraudulent behaviour. It also allowed us to prioritise which variables deserved our focus and how to design our data mining models for optimal performance. The insights drawn at this stage were invaluable—they informed the structure and parameters of the subsequent tax fraud detection models, paving the way for an efficient and precise mechanism to root out tax fraud.

3.3.3 Data Preparation

The data preparation stage is a foundational process in constructing a robust machine learning model, especially for complex issues such as tax fraud detection. A meticulously cleaned and consistent dataset not only facilitates more accurate models but also reflects real-world scenarios where the model is expected to perform. [102] [103]Here's an in-depth look at the processes that were involved at this stage.

3.3.3.1 Identification of Missing Values

The initial task in data pre-processing was the systematic identification of missing values across all columns of the dataset. Missing data poses a significant problem as it can lead to biased estimates, leading to less accurate models. We meticulously examined the dataset to map out the columns with missing values, considering the nature of each variable–whether categorical or numerical – and its significance within the dataset.

3.3.3.2 Handling Categorical Features

For categorical features, simple imputation methods like filling in missing values with the most common category might not always be appropriate, particularly if the data is sequential or time-based. Hence, we applied forward-fill or backward-fill techniques. Forward-fill carries the previous known value down to the missing entry, while backward-fill takes the next known value and assigns it to the missing slot. This method maintains the integrity of data sequences, ensuring that the continuity of taxpayer behaviour or transaction patterns is preserved.

3.3.3.3 Handling Numerical Features

Numerical features, on the other hand, were treated differently. Replacing missing values with the mean of their respective columns was the chosen approach here. This method is beneficial when the data is assumed to be missing at random, and the mean offers a central tendency measure that can substitute for the missing data without introducing significant bias.

3.3.3.4 One Hot Encoding Categorical Features

According to the Scikit-learn library [104] the process of one-hot encoding was pivotal in transforming categorical features into a machine learning-friendly format. It converts categories into a binary matrix, ensuring that categorical data is interpretable by algorithms. This step avoids the ordinal implications that come with numerical encoding, which could mislead the model into assuming a non-existent natural order within the categories.

3.3.3.5 Feature Normalisation and Scaling

Once encoded, all features underwent normalisation and scaling using the 'MinMaxScaler' in the Python libraries. Normalisation is a critical step in data pre-processing, especially in a dataset with features that vary in magnitudes, units, and ranges. By bringing every feature into the range [0,1], MinMaxScaler ensures that each feature contributes equally to the model's prediction capability. [105] This is essential in tax fraud detection, where differing scales of monetary values could otherwise skew the model's focus. Moreover, normalised data often speeds up the learning algorithm's convergence, leading to more efficient model training. Feature scaling is not just a technical requirement for many machine learning algorithms; it also has practical implications. Algorithms that rely on gradient descent, for example, will converge much quicker on scaled data, as the shape of the error surface will be more spherical. This means that the path to the minimum is more direct, and the algorithm doesn't have to navigate the long ravines that unscaled features can cause. Datasets in the real world frequently have features that differ in magnitude, range, and units. Thus, to enable machine learning models to analyze these features on a consistent scale, it is necessary to conduct feature scaling. [105]

Each pre-processing step underwent rigorous validation. Statistical measures were applied to confirm the validity of the imputations and transformations. Moreover, domain experts were consulted to ensure that the pre-processing aligned with real-world tax processing knowledge and that no critical information was lost during the transformation. By the end of this extensive data preparation process, we were left with a polished dataset–one that represents a truthful rendition of taxpayer behaviour, with features carefully sculpted to highlight potential fraud indicators. This process not only sets the stage for powerful model training but also ensures that the insights gained are as actionable and reliable as possible in the hands of tax authorities.

3.3.4 Data Modelling

To determine which machine learning algorithms should be used to help solve our problem with the available dataset at hand, we used the general flow chart design in Fig. 8 as a guide to determine which algorithm to use on our dataset. Thus, the K-means algorithm was employed to identify individuals who submitted fraudulent tax returns, and the Local Outlier Factor algorithm was utilized to improve the model's effectiveness. K-means Clustering and the Local Outlier Factor (LOF) are advanced data mining techniques used to discover anomalies, specifically cases of tax fraud, in this study. These approaches are especially valuable in situations when there is a limited amount of labeled data or no labeled data at all, which is

frequently the case in tax fraud detection. This section elucidates the methodologies behind K-means clustering and LOF and their application to the dataset under consideration.

3.3.4.1 K-means Clustering

An unsupervised learning technique called K-means divides a set of observations into a predefined number of clusters. The given core concept involves defining k centroids, each corresponding to a cluster, and then finally designating the closest centroid for each data point. In the context of K-means, the term 'means' refers to the calculation of the average value of data points, which is effectively the process of identifying the centroid. It operates through an iterative process where the primary goal is to minimise the within-cluster variance, effectively making the clusters as compact and distinct as possible. [106] Following a random centroid pick, the method iteratively carries out the following two steps: assign and update. Every data point is assigned to the nearest centroid during the assignment step; closeness is commonly calculated using the Euclidean distance. The centroids are computed as the average of all the points assigned to each cluster during the update stage. The iterations persist until the centroids reach a stable state and the allocation of data points to clusters remains unaltered. [107] Fig. 7 Illustrates the procedural steps involved in K-means clustering techniques.

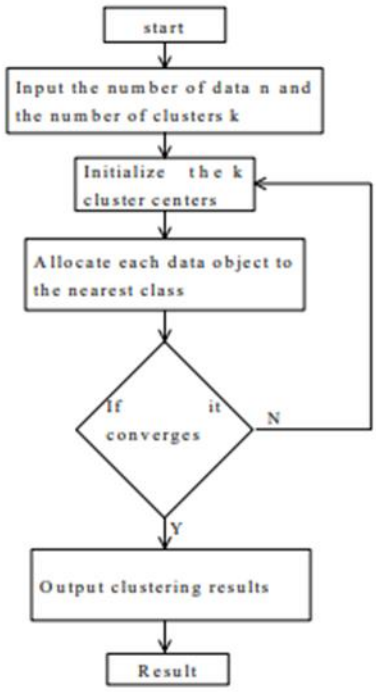

Figure 7: Process of K-means Clustering Algorithm [107]

In our study, K-means clustering was harnessed to discern patterns in the data indicative of fraudulent behaviour. We determined the number of clusters in advance using the elbow approach, which is a methodology that graphs the variance explained against the number of clusters and then we employed the algorithm in our pre-processed dataset. This helped us to aggregate similar data points together and potentially isolate unusual clusters that could represent fraudulent activity. [108]

3.3.4.2 Local Outlier Factor (LOF)

Another unsupervised method is the Local Outlier Factor (LOF) algorithm but with a focus on anomaly detection. Unlike global outlier detection methods, LOF looks at the local environment of a data point. The approach operates under the assumption that regular data points are clustered in a concentrated neighborhood, while anomalies are located far away from this cluster. It calculates a score reflecting the degree of isolation of a point concerning its surroundings, with higher scores indicating possible outliers. A point's density is assessed by the LOF algorithm concerning the densities of the points that surround it. A point is labelled as an outlier if the density around this point is significantly different from the density around its neighbours. This approach is effective in detecting outliers that may not be discernible from a global perspective, particularly in datasets with varied density regions. [109]

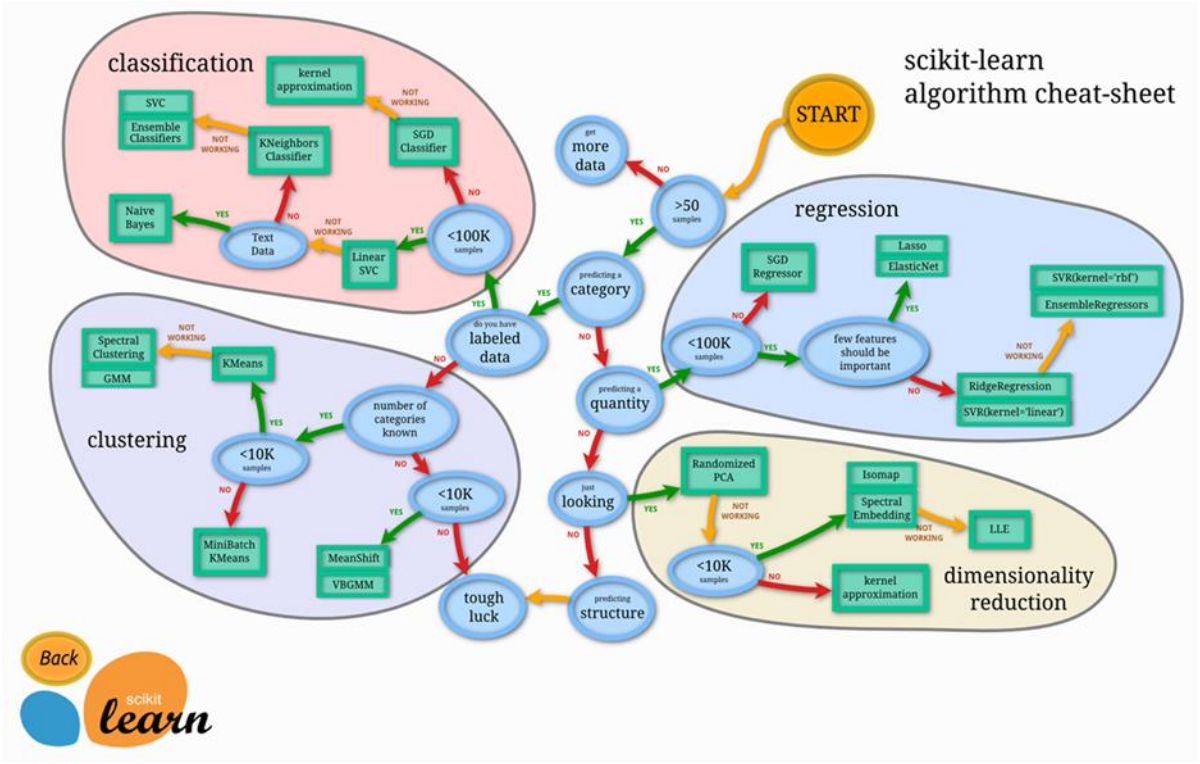

Figure 8: Choosing the right estimator [110]

Through the combined application of K-means clustering and LOF, the study was able to navigate the vast data terrain and flag areas warranting closer scrutiny for fraud. It is worth noting that while these techniques are powerful; they are not infallible. False positives can occur, and thus, the flagged instances require further investigation by tax experts to confirm the presence of fraudulent activity. Nevertheless, the methodologies provided a robust framework for streamlining the detection process and enhancing the overall efficiency of tax fraud surveillance systems.

3.3.5 Model Evaluation

To meet the second objective model evaluation was conducted. A critical stage in determining how well tax fraud detection algorithms, such as K-means clustering or other machine learning methods, perform is model evaluation. It aids in determining how effectively the model generalises to fresh data and achieves set goals. Data splitting, choosing appropriate evaluation metrics, threshold selection, performance visualisation, cross-validation, benchmarking, interpretability, real-world testing, iterative improvement, and reporting are important steps in the evaluation of models. [61]Model evaluation is an integral process of the model design aimed at finding the best model with generalisation accuracy on unseen data. The evaluation stage is where the data not seen by the trained model comes into play to give a picture of how the model will perform in real life. The performance measures to evaluate our classification model were:

Silhouette Coefficient: This measure assesses the consistency within data clusters by gauging the cohesiveness, or the resemblance of an object to its cluster compared to other clusters, as opposed to detachment. The formula for calculating the Silhouette Coefficient S for an individual sample is as stated below:

$$S= b – a / \max (a, b)$$

Equation 1: Silhouette Coefficient [111]

Where it is:

a denotes the mean distance between a data point and all other points within the same cluster,

b is the minimum mean distance from the point of interest to all points in any other cluster, which is minimized across clusters.

A higher score indicates that the object aligns well with its cluster and poorly with nearby clusters. The silhouette score ranges from -1 to +1. [111]

Domain Expert Knowledge: We provided the outcome set to domain experts for assessing the accuracy of the model and interpreting the cluster outcomes.

3.4 System Design and Implementation

3.4.1 Proposed Business Process

To develop a functional prototype of our predictive model, we propose the creation of a sophisticated website system. We meticulously designed this system to serve as an interactive interface for the prediction of potential fraudulent VAT returns. Making use of the K-means algorithm's resilient characteristics, the system is designed to function inside a sequential workflow, characterised by the following process:

1. User Input Reception: The core functionality of the website system revolves around its ability to receive input data from users. This data, pivotal in the prediction process, can encompass various parameters relevant to VAT returns. The interface is designed to be user-friendly, ensuring ease of data entry for users from diverse backgrounds.
2. Application of the K-means technique: The K-means clustering technique is used by the system once the user enters the data. This technique is a critical component of the system, responsible for analysing the input data to identify potential cases of fraudulent VAT returns. The K-means technique processes data to discern patterns or anomalies indicative of fraudulent activities, leveraging its data clustering capabilities.
3. Display of Prediction Results: After doing an analysis, the system quickly presents the forecast findings on the website. We generate these results based on the comprehensive analysis conducted by the K-means algorithm on the user-provided data. We present the results in a format that is clear and understandable, making it easy for users to interpret the outcome of the prediction.

3.4.2 Requirements Specification

In this section, we look at the Functional and non-functional requirements to design the proposed business process.

3.4.2.1 Functional Requirements

The functional requirement outlines the essential characteristics that a system should possess. [112] they focus on how the software must perform and specify the desired behaviour of the

system. In the Functional User Requirement specification for the proposed system, we have assigned a rating to each requirement as shown in Table 3;

D - Desirable, indicating that the action assessed is suitable.

M - Mandatory, indicating that the action assessed is necessary and required.

Table 3: Functional User Requirements

No	Action	System Functionality	D/M
1.	User Interface for Data Input	a) A web form where users can input or upload the data needed for the classification.	M
		b) Validation of user input to ensure it meets the expected format or criteria.	M
2.	Data Processing and Classification	a)Backend processing of the input data to format it appropriately for the classification model.	M
		b) A classification model that can process the input data and generate a classification result.	M
3.	Output Display	a) Displaying the results of the classification to the user in an understandable format.	M

		b) Options to download or save the results.	D
4.	Error Handling and User Feedback	a) Simple messages or feedback in case of invalid input, processing errors, or successful operations.	M
		b) Proper handling of exceptions to prevent application crashes.	M
5.	API Endpoints (if applicable)	a) RESTful API endpoints to receive input data and send back classification results.	D
6.	User Authentication (if applicable)	a) User accounts and authentication if the application needs to restrict access or provide personalised experiences.	M
7.	Security Measures	a) Protecting the application from Typical web vulnerabilities including SQL injection and Cross-Site Scripting.	M
		b) Ensuring user data privacy and compliance with	M

		relevant data protection regulations.	

3.4.2.2 Non-Function Requirements

Non-functional requirements prioritize the usability of a software system. [112]The structured format for outlining the key non-functional requirements for the Flask web application, categorising each as either Mandatory (M) or Desirable (D) is highlighted in Table 4.

Table 4: Non-Functional User Requirements

No	Requirement Type	System Feature	D/M
1.	Performance	a) The application should respond to user inputs or requests within a specified time frame.	M
		b) The system should possess the capability to manage a predetermined quantity of concurrent users without performance degradation.	M
2.	Scalability	a) The application should be able to scale horizontally or vertically to accommodate increasing load or data volume.	M
3.	Reliability & Availability	a) The application should have a high	M

		percentage of uptime (e.g., 99.9%).	
		b) The system should be fault-tolerant and continue operating properly in the event of component failure.	M
4.	Usability	a) The user interface should be intuitive and easy to navigate for users with minimal technical expertise.	D
		b) The application should be accessible to users with disabilities, complying with standards like WCAG	D
5.	Security	a) The application must protect sensitive data and resist common web security threats (e.g., SQL injections, and XSS attacks).	M

3.4.2.3 Hardware Requirements

i. Processor (CPU): The study recommends a multi-core processor minimum of corei7 with a processor speed of 1.6GHz or faster. We recommend this for the system to handle the computational demands of training the machine learning models efficiently.

ii. Memory (RAM): 8 GB minimum, to accommodate the demands of running complex models and simulations.

iii. Storage: Solid State Drive (SSD) with sufficient space to store test datasets and models. The recommended minimum storage size is 20GB.

3.4.2.4 Software Requirements

To develop the web-based application, we used the following software tools

i. Python: Flask is a Python framework, so Python (preferably the latest stable version) is a core requirement. Python's package manager, pip, for installing and managing Python packages. Python, as a powerful and friendly object-oriented programming language, will be used in data exploration and analysis with the use of its built-in Libraries such as pandas, Scikit-learn Matplotlib, etc. [113]

ii. Integrated Development Environment (IDE) or Text Editor: Visual Studio Code and Jupyter Notebook, which offer Python support and various helpful extensions, were used to develop the model

iii. Flask: Flask was used as the core framework for building the web applicationFlask, a micro-framework coded in Python, that features a minimalistic set of tools and libraries. [114]. Flask is utilized to enhance the efficiency of development.

iv. MySQL Database: MySQL open-source relational database was used to store the data. System data is stored on MySQL, a Relational Database Management System (RDBMS). Many systems employ MySQL because of its rapid data processing capabilities. The system development efficiency is supported by the inclusion of a user-friendly Structured Query Language (SQL).

v. Machine Learning Libraries: To develop the models, machine learning libraries were used which included, Scikit-learn libraries used for building and deploying machine learning models. Scikit-learn is one the most notable Python libraries that are helpful for data manipulation. It is mainly applied for feature selection, cross-validation, and construction of a confusion matrix during data modelling. [115] Pandas for data manipulation and analysis which is known in full as Python Data Analysis Library, and it provides functions used to organize data, enhance code legibility, provide speed during data processing, and allows SQL Databases, Microsoft Excel, and comma-separated values (CSV) files to be accessed, read and converted into a structured table format called Data Frames. [116] NumPy for numerical operations, which is a dynamic

Python library most useful for multidimensional arrays, numerical transformations, and computations, it also helps to store large amounts of data. [117] Matplotlib allows to creation of 2D data visualisation and plotting of a variety of graphs, such as bar, histogram, scatter, and line graphs, for a better result demonstration [118]

vi. Data Visualisation tools: The process of utilizing pictorial design to portray data graphically is called visualization. The objective is to provide a visually clear and aesthetically pleasing representation. This research employs common data visualization techniques, including scatter plots, bar charts, histograms, line charts, and pie charts. These visualizations are created using the Matplotlib and Seaborn libraries, or Plotly for generating charts and graphs. [118]

3.4.3 Design Specification

In this section, we highlight how The system was developed using Unified Modeling Language (UML) models to guide the development process. [68]Unified Modelling Language (UML) is a standard graphical language used to model software systems. It provides a set of diagrams that can be used to visualise, design, and document software systems. UML diagrams can also be used to document and communicate the design of the system to other developers or Parties with a vested interest in a particular project or organization. This can facilitate the establishment of a comprehensive comprehension of the system's framework among all participants in the development process and behaviour can help facilitate collaboration and feedback. [67]

3.4.3.1 System High-Level Overview Design

A web-based application is developed consisting of three layers, with each layer designed to handle specific functions within the system. This architecture, incorporating Flask, a K-means machine learning model, and a MySQL database, ensures a clear separation of concerns and enhances both maintainability and scalability.

1. Presentation Layer: This is the front-facing component of the application where user interaction occurs. Developed using HTML, CSS, and JavaScript, this layer presents a user-friendly interface for interaction. Flask, a lightweight and flexible Python web framework, is used to render the user interface in the web browser. It handles HTTP requests and responses, integrating seamlessly with the backend logic.

2. Business Logic Layer: This layer serves as the core of the application, where the main functionalities are processed. We again utilised here flask to manage the application's server-side logic. It interacts with the K-means machine learning model to process and analyse the data based on user input. The K-means algorithm, implemented in Python, is responsible for clustering and predictive analysis, forming the backbone of the application's predictive capabilities.
3. Data Layer: The last layer of the application is focused on data management and storage. MySQL, a popular relational database management system, is used for storing and retrieving data. It manages all the data transactions and storage needs of the application, including user data, the outcomes of the prediction analysis, and the input data for the machine learning model. The Flask application communicates with the MySQL database, ensuring data consistency and integrity. Fig.9 provides a high-level overview of the proposed system.

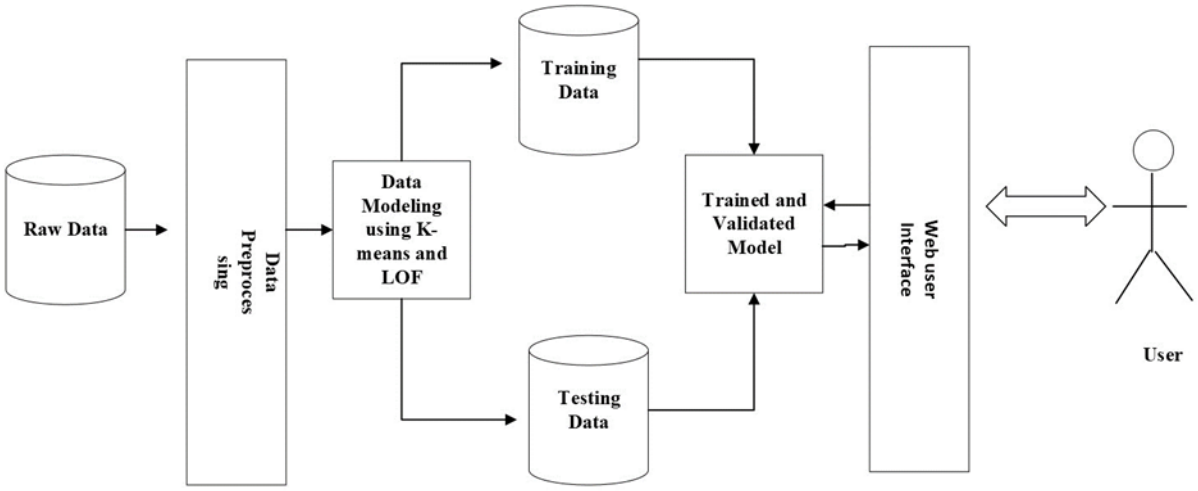

Figure 9: Overview of Proposed System

3.4.3.2 Use Case Diagrams

The interaction between users and the system is represented by the use case model. A Use Case diagram illustrates the specific functions that an actor plays in a given scenario [67]. System development is made easier with the help of this diagram. This system will involve two actors: the user and the admin. Users are knowledgeable tax professionals who possess expertise in VAT returns and tax files. They can conduct tests to assess the likelihood of fraud on a VAT return by inputting the necessary data into the system. An admin is the individual responsible

for overseeing and supervising the dataset. Fig.10 Illustrates the utilization of the proposed system through a use case diagram.

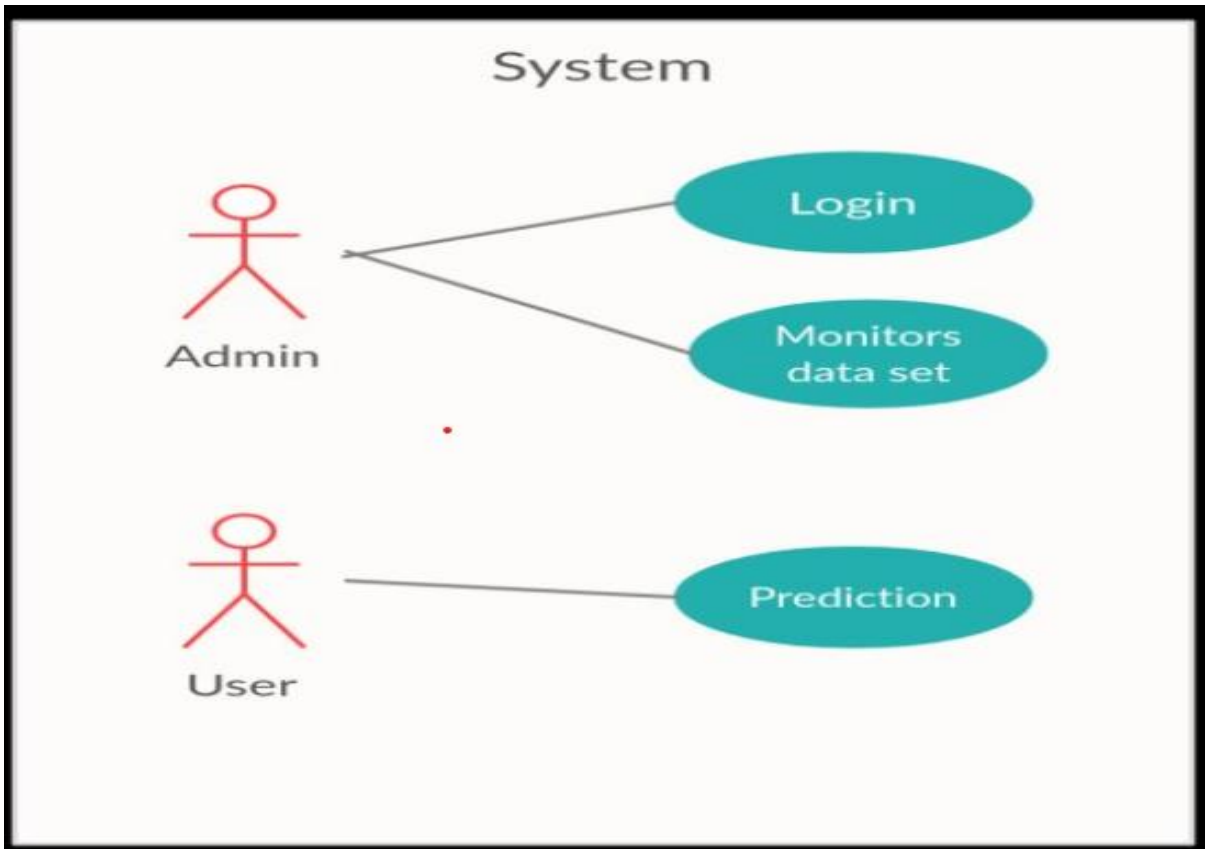

Figure 10: Proposed System Use Case [65]

3.4.3.3 Activity Diagrams

Figure 11's Activity Diagram illustrates how users must provide the information required by the system for it to do predictive analysis. If the data inputted is inaccurate or does not meet the specified requirements, an error message will be displayed, prompting the user to input the data accurately. Assuming the accuracy of the data, the system will generate forecasts, present the outcomes, and save them in the database.

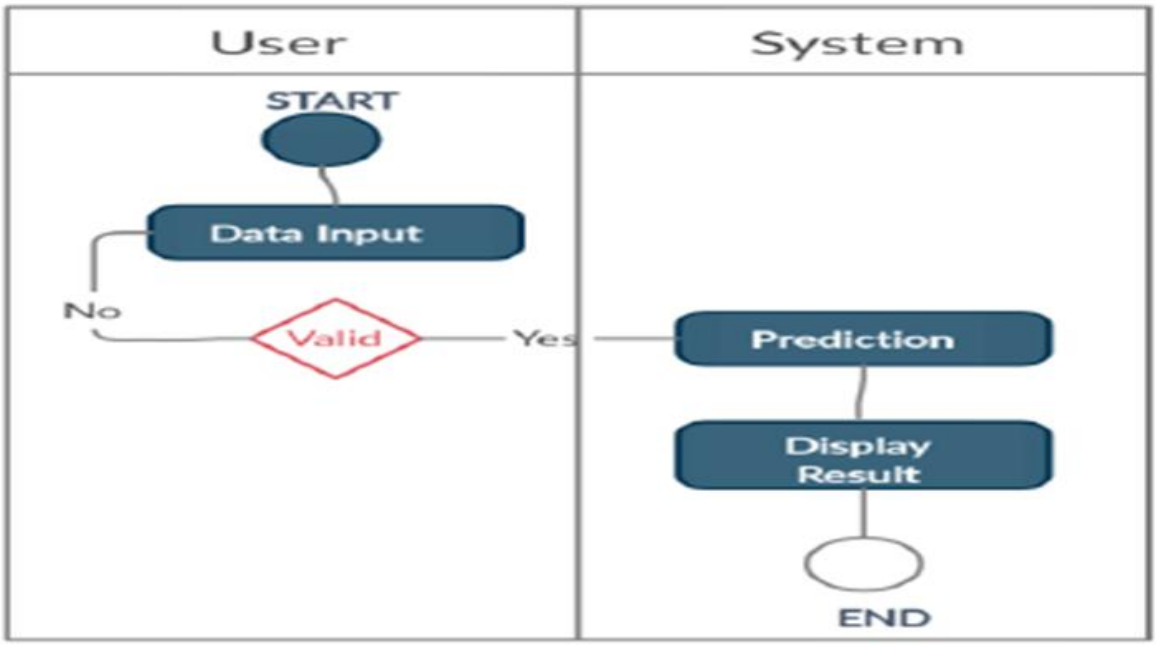

Figure 11: Proposed System Activity Diagram [65]

3.4.3.4 Class Diagram

The types of items in a system and their relationships are often described using class diagrams. Using design components like classes, packages, and objects, class diagrams simulate the contents and structure of a class. Class diagrams provide a comprehensive representation of a system's architecture from three distinct viewpoints: conceptual, specification, and implementation. All of these perspectives become apparent as the diagram is constructed and contribute to the consolidation of the design. [119]

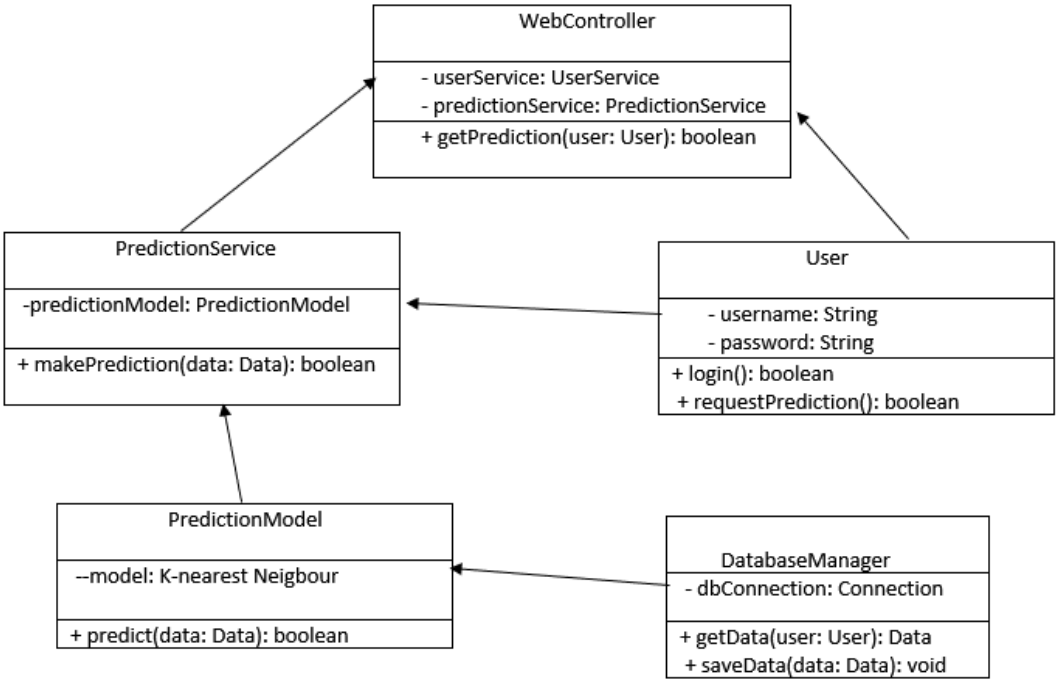

Figure 12: Class Diagram

3.4.3.5 Sequence Diagrams

One kind of interaction diagram that shows how and in what order processes interact with one another is a sequence diagram. The design is a Message Sequence Chart. Sequence diagrams are often referred to as event diagrams, event circumstances, and timing diagrams. [71] A sequence diagram illustrates the chronological order in which the items involved in the interaction take part. This comprises the temporal dimension (vertical) and the spatial dimension (horizontal) including various things. [120] The proposed system's sequence diagram is shown in Figure 13.

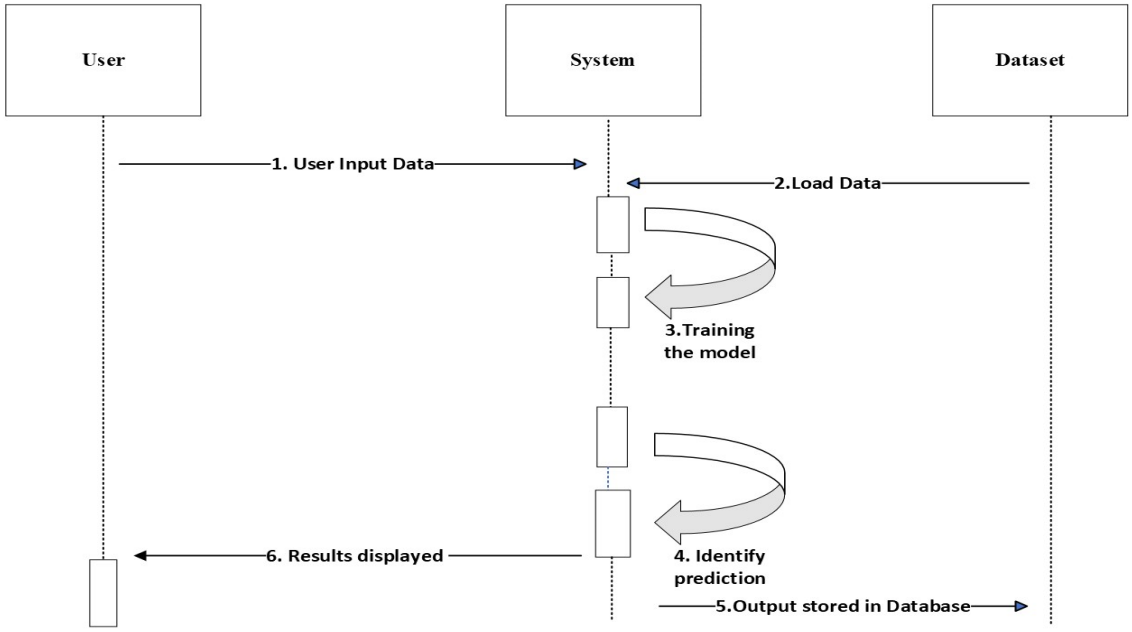

Figure 13: Sequence Diagram

3.4.3.6 Database Design

In the database design section, a MySQL database named VAT Dataset was established using MySQL Workbench 8.0. It was running on Port 3306 and the login user was set as root. Two tables, VAT_SALES and VAT_REFUNDS, were constructed specifically for this investigation. The features, descriptions, and data of the two tables are highlighted in Tables 5 and 6.

Table 5: VAT_SALES Table

Attribute Name	Description of Attribute	Datatype
RETURN_ID	*Taxpayers return filed identification number*	Number (11,0)
TPIN_OF_PURCHASER	*Taxpayer identification number for the purchaser*	Varchar2 (15 char)
TPIN_OF_SUPPLIER	*Taxpayer identification number for the supplier*	Varchar2 (15 char)
INVOICE_DATE	*Invoice Date of goods/services purchased*	Date
INOVICE_NUMBER	*Invoice Number of goods/services purchased*	Varchar2(50 char)
VAT_CHARGED	*VAT amount Charged on the goods/services*	Number (11,0)
DESCRIPTION_OF_GOODS_SERVICES	*Detailed description of the goods/services*	Varchar (255 char)
AMOUNT_BEFORE_TAX	*The cost of the goods/service minus the VAT charged*	Number (11,0)

Table 6:VAT_REFUND Table

Attribute Name	Description of Attribute	Datatype
REFUND_ID	*The taxpayer filed Refund identification number*	Number (11,0)
CLAIM_AMOUNT	*The amount claimed as a refund by the taxpayer*	Number (11,0)
APPROVED_AMOUNT	*The amount approved by the ZRA*	Number (11,0)
STATUS	*Status of the refund process*	Varchar2(100 char)
PERIOD_TO	*Tax period start date*	Date
PERIOD_FROM	*Tax period end date*	Date
TOTAL_AMOUNT_PAID	*The amount paid to the taxpayer as a refund*	Number (11,0)

An information system's numerous entities and their attributes are shown graphically in an entity-relationship diagram (ERD). The Entity-Relationship Diagram (ERD) of Fraud Detection, displayed in Figure 14, illustrates the VAT_SALES and VAT_REFUNDS entities and their respective characteristics.

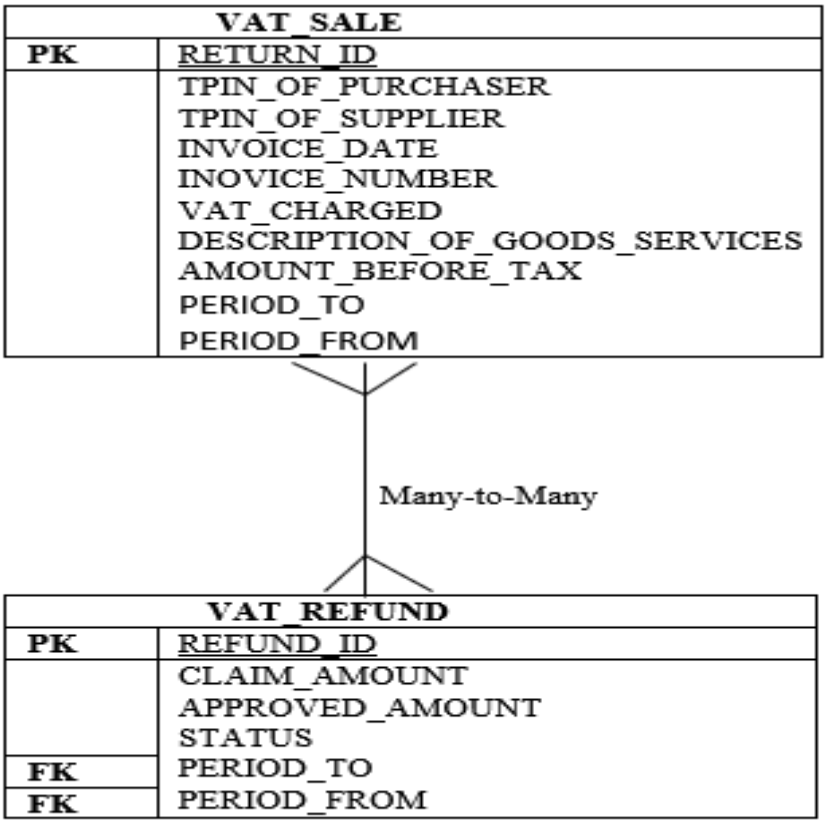

Figure 14: Entity Relationship Diagram

3.4.3.7 Security Design

The Security Design area encompasses the techniques, technologies, and protocols that are implemented to safeguard the system and its data from unauthorized access, use, disclosure, disruption, modification, or destruction. This section ensures that the system is resilient against various cyber threats and aligns with best practices and compliance requirements. The following key elements were included in the security design of the system:

1. User Authentication: We Implemented the multi-factor authentication (MFA) mechanisms, including username and password, to ensures that the user is who they claim to be.

2. Database Authorisation and Access Control: The protection of sensitive data was achieved by the implementation of role-based access control (RBAC), which allowed users to view and manipulate only the database data pertinent to their roles.
3. Database Activity Monitoring: Leveraging MySQL's auditing features enabled us to track and analyse database activities such as data creation, modification, or deletion.

For Application security, the following were implemented:

1. SQL Injection Attack Prevention: Sanitising all inputs received through forms to prevent SQL injection attacks.
2. Error Handling: Implementing comprehensive error handling to prevent exposure to sensitive information.
3. Encryption in transit and host authentication (SSL): To guarantee that the user connects with the intended host computer, all data—including usernames and passwords—is encrypted over the network using SSL and certificates.

3.5 System Implementation

3.5.1 System Development

This stage involves translating the design specifications into an actual software code. We utilize Flask, a Python-based microweb platform, to create the web application. Flask enables users to seamlessly include additional application functionality as if it were an inherent part of the framework. The following four program modules make up this system development process:

1. Model.pkl: The machine learning model to anticipate taxpayers filing bogus returns is contained in this software code file. By employing the K-means method with all the features, we will incorporate it as a predictive model into the overall model. Pickle file. The code in Fig. 15 provides a snippet of the K-means algorithm saved as model. pkl file.

```
import joblib

# joblib.dump(kmeans, "return_fraud_segementation_v3.pkl")

# joblib.dump(scaler, "scaler_v3.pkl")

spam_model = open('return_fraud_segementation_v3.pkl','rb')
model = joblib.load(spam_model)

spam_scaler= open('scaler_v3.pkl','rb')
model_scaler = joblib.load(spam_scaler)

model.predict([X_test_scaled[1]])
```

Figure 15: Code Snippet of Model.pkl

2. app.py: The Flask APIs in this package compute the projected value using our model and deliver the TPIN information that the user provides through Graphical User Interface or API requests. Figure 16 displays a code snippet of the app.py file.

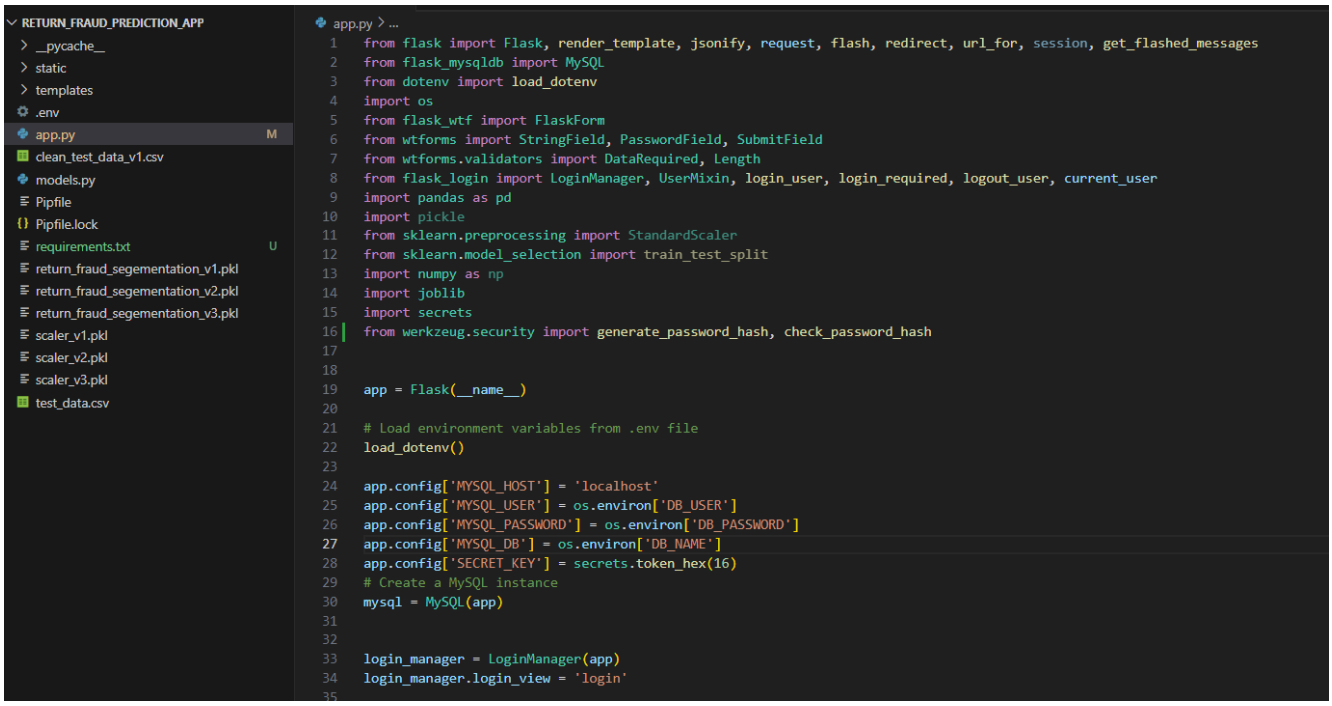

Figure 16:Code Snippet of app.py

3. Template- The index.html file in this directory has an HTML form that provides a template for users to input TPIN information. It also displays the predicted outcome of the model based on the input. Figure 17 provides a code snippet of the index.html file.

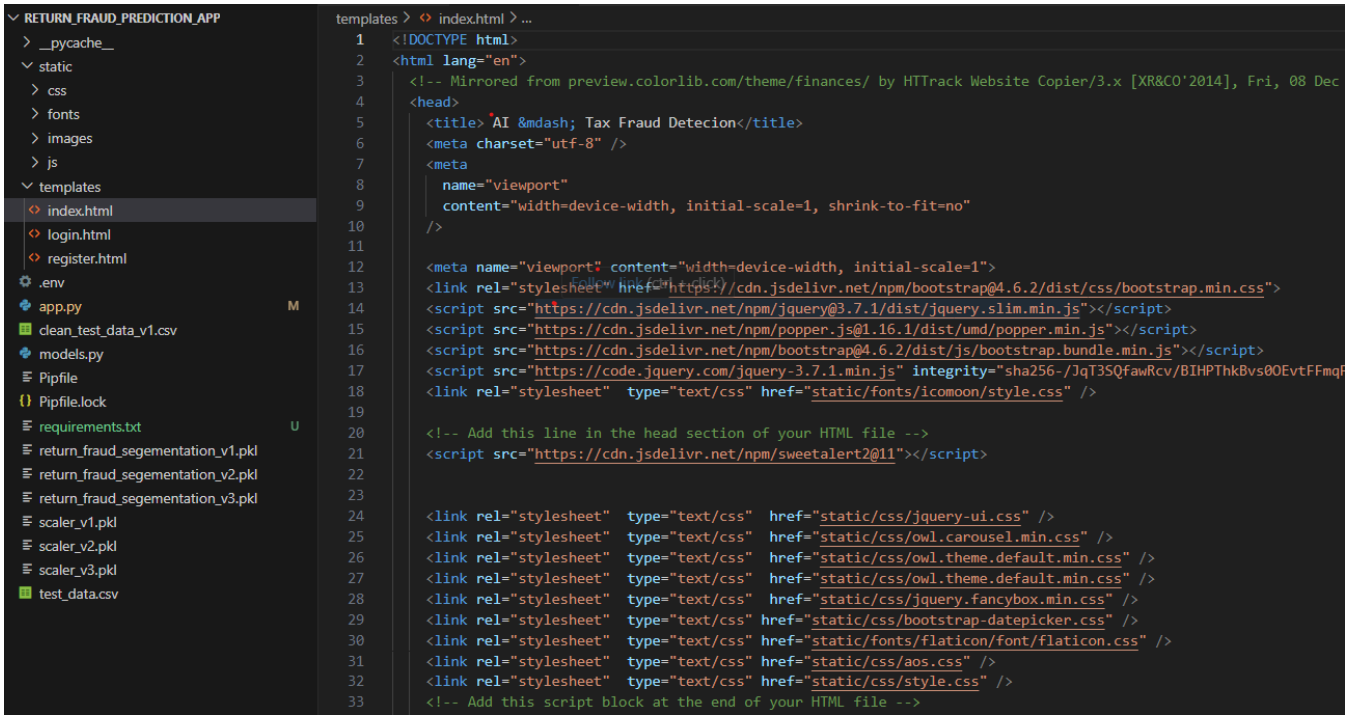

Figure 17: Code Snippet of index.html

4. Static: Our HTML form requires certain styling, and this folder provides the necessary JavaScript and CSS files. Figure 18 provides a code snippet of the static file.

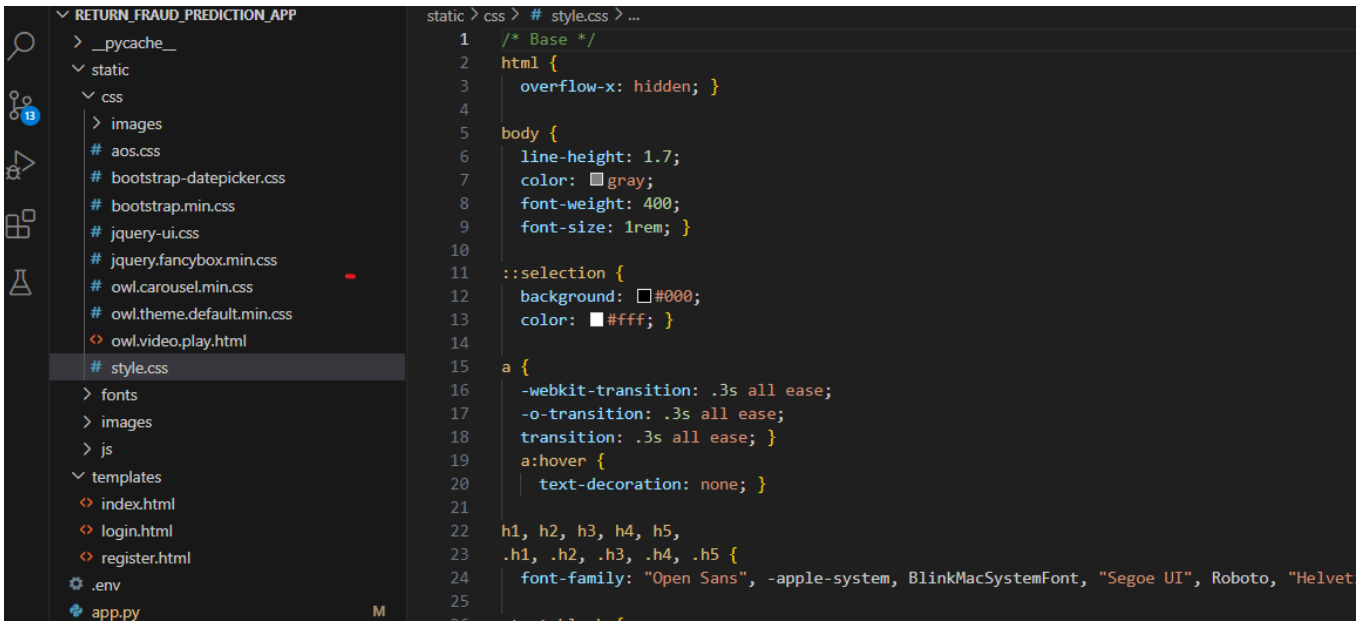

Figure 18: Code Snippet of static file

3.5.2 System Deployment

The deployment phase of a web application is a crucial process that transitions the software from a development environment, where it's built and tested, to a production environment, where it becomes accessible to end-users. In this case, the system's deployment involves a series of structured steps:

1. Hosting of the Application on Web Server: we deployed the web application onto a website for user access using a free web hosting service from Azure.
2. Integration Using Visual Studio Code: Visual Studio Code is a flexible Integrated Development Environment (IDE) or editor for text that integrates the many components of an application. These components include:

i. Model.pkl: The pickle-formatted Python machine learning model file.
ii. app.py: The primary Python file that contains the Flask application functionality.
iii. templates: The directory that holds HTML files that specify the web application's structure and design.
iv. static: A folder that includes static content such as CSS, JavaScript, and image files.

These modules are meticulously integrated and published to a web hosting server to create a cohesive web application.

3. Database Configuration with Apache Web Server: The application uses Apache, a popular open-source web server, to manage web requests and serve the application efficiently. Alongside Apache, we set up a MySQL database in the production environment. This database is essential for storing, retrieving, and managing the application's data effectively. Key database settings, including the username, password, and hostname, are meticulously configured in the system's configuration files. This step is vital to ensure secure and seamless database connectivity.

3.5.3 User Acceptance Testing

A crucial stage in the creation and execution of the program is User Acceptance Testing (UAT). It serves as the ultimate validation of the system's functionality and usability from the end-user's perspective. [121] In this research, UAT is not merely a procedural step; rather, it represents a pivotal moment where theory and practice converge, offering valuable insights into the practical efficacy of the developed application.

The methodology for conducting UAT in this research involved selecting a representative sample of users, creating test cases and scenarios, and collecting feedback. This approach ensures a comprehensive evaluation of the application from the user's standpoint, providing valuable insights that transcend technical performance metrics.

3.6 Ethical Considerations

Developing a model to detect tax fraud using machine learning entails serious ethical considerations. Privacy and security are paramount. To protect taxpayer information, we encoded the data headers into variables V1 to V12. We also stored the dataset in an external location with a requirement of password authentication.

3.7 Chapter Summary

The study's research technique was covered in this chapter. The research was conducted using the CRISP-DM approach. The CRISP-DM methodology was chosen due to the inclusion of a machine-learning model in the investigation. This chapter emphasizes the functional and non-functional requirements, as well as the design specifications utilizing UML modeling, that were utilized to create the Use Case Diagram, Activity Diagram, Sequence Diagram, and Logical Diagram for the proposed system development. In this chapter, we explained both the Database and Security design.

4 RESULTS

4.1 Introduction

The model development and experimentation outcomes are presented in this chapter. We present the outcomes of the data exploration conducted and elucidate the discoveries regarding the connections between VAT RETURNS and REFUNDS data variables through the use of data visualization. We also present the performance and evaluation of the data models LOF and K-means algorithm. Lastly, we describe the system automation and user acceptance testing results.

4.2 Data Patterns and Relationship Results

4.2.1 Feature Extraction

Our dataset contains 229,000 VAT instances, and 14 columns containing numerical and categorical data. For categorical data type, we had 2 features: (TPIN and refund status). Our dataset did not have any class labels. For numerical data type, we had twelve features: (input invoice,output_invoice,total_vat_charged,return_amount,refund_amount_claimed,refund_amount_paid, refund approved, credit_return, refund_created). Based on our main objective of detecting fraud from the historical VAT data, we had to dig deep and find hidden insight information that is related to taxpayer compliance classification. Hence, we needed to have a predefined hypothesis, and based on them; we could know whether a given transaction is fraudulent or not. The following conditions formulated our hypothesis:

i) Taxpayer ID identified by TPIN (Taxpayer Identification Number) and to know if a given taxpayer's total number of input invoices > output invoices

ii) To know if a given taxpayer has a high number of credit return amounts.

iii) To know if a given taxpayer has a high number of refund claims made

iv) To know if a given taxpayers' REFUNDS_APPROVED > REFUNDS_REJECTED

With the highlighted hypothesis, we extracted 12 features as summarised in Table 7.

Table 7: Feature Selection

Variables	Tax ratios Hypothesis	Description
V1	TPIN	*Taxpayer identification number*
V2	RETURN_AMOUNT	*Amount declared in the tax return*
V3	#_OF_INPUT_INVOICES	*The taxpayer's total count of input invoices filed.*
V4	#_OF_OUTPUT_INVOICES	*The taxpayer's total count of filed output invoices.*
V5	REFUND_AMOUNT_CLAIMED	*Amount claimed by a taxpayer as a refund.*
V6	REFUND_AMOUNT_PAID	*The actual amount refunded to the taxpayer*
V7	CREDIT_RETURN	*Indicates if the return amount was negative.*

V8	REFUND_CREATED	*Indicates if the taxpayer created a refund claim*
V9	REFUND_APPROVED	*If the authority approves the refund*
V10	INPUT_TO_OUTPUT_RATIO	*Determines the proportion between the overall quantity of input invoices and the overall quantity of output invoices.*
V11	CREDIT_TO_RETURN_RATIO	*Computes the ratio of the aggregate number of credits returned to the overall number of returns submitted by individual taxpayers.*
V12	REJECTED_REFUNDS	*If the authority rejects the refund*

4.2.2 Data Exploration

By using Exploratory Data Analysis as a tool, it helped us to understand our data and how they are related to each other. We reviewed the Categorical data to understand the distribution of classes, and in our case, we observed that most tax refund claims were initiated but not completed, See Fig 19. This posed a challenge due to the significant imbalance which could skew predictive models. During the exploration, we also engaged in correlation analysis and

outlier detection. The detection of outliers is pivotal, as these can be indicative of anomalies or errors in data collection.

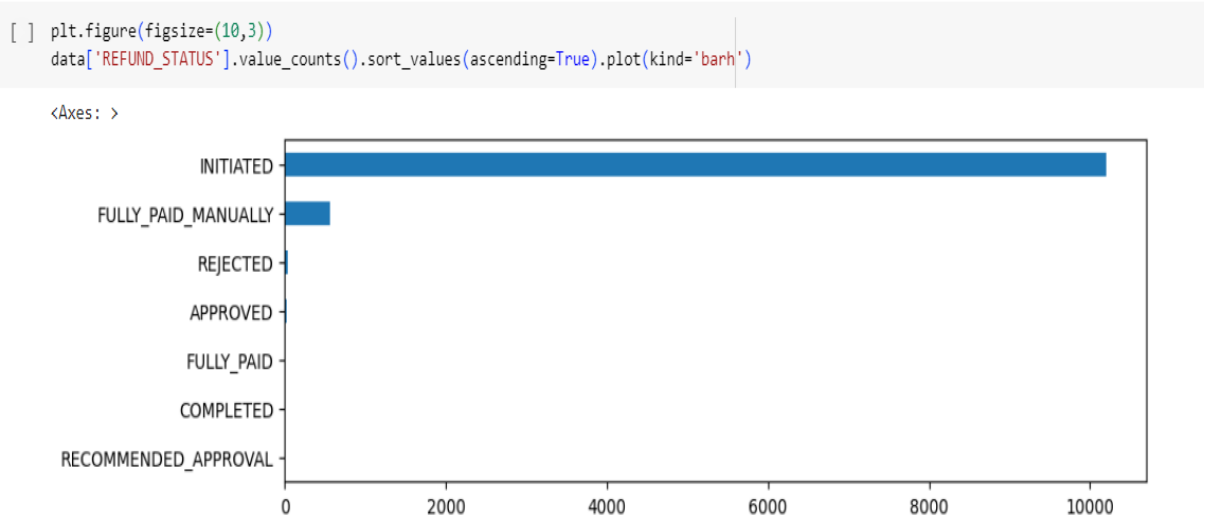

Figure 19: Distribution of categorical data

We also considered the interaction between features. Understanding the interplay between different variables is crucial for tax fraud detection, as it can point to irregularities and areas of fraudulent activities. As highlighted in our EDA, we identified a notable trend: there was a larger number of input invoices compared to output invoices among taxpayers, as depicted in Figure 20.

```
[ ]  plt.figure(figsize=(10,3))
     plt.title('Average number of input and output invoices',size=18)
     plt.scatter(data['NUMBER_OF_OUTPUT_INVOICES'],data['NUMBER_OF_INPUT_INVOICES'])
     plt.xlabel('Number of output invoices')
     plt.ylabel('Number of Input invoices')

     Text(0, 0.5, 'Number of Input invoices')
```

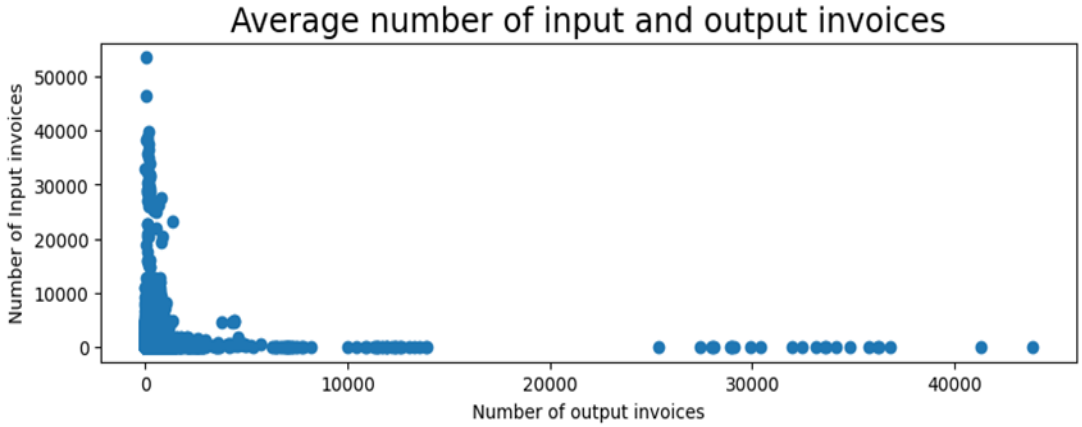

Figure 20: Average number of input and Output invoices

The exploration extends to using visual tools like heat maps to display the correlation between variables. In our data set, Figure 21 demonstrated a strong negative correlation between variables [V12] and [V9] and a positive one between [V11], [V7], and [V9]. Such insights from a heat map are indispensable as they lay bare the directional relationship between variables. Nevertheless, a heat map is merely a snapshot of correlation, not causation, and must be interpreted with domain knowledge and further statistical analysis to deduce why these relationships exist.

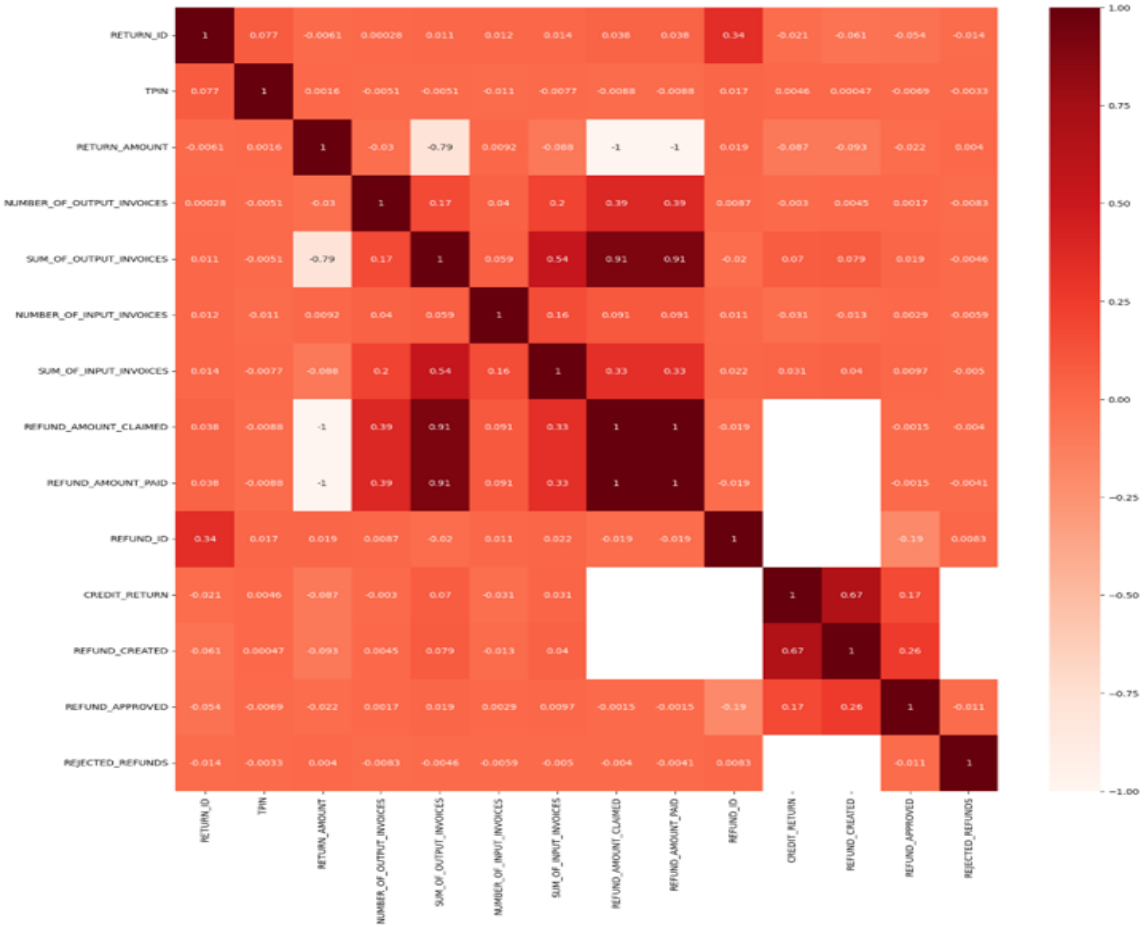

Figure 21: Heat Map

Subsequently, all features were standardized using the 'MinMaxScaler' technique to guarantee that each feature has an equal impact on the model's performance. The characteristics are also scaled and transformed into the range of [0,1] as shown in Figure 22, which displays the results of the scaled dataset. This scaling helped the machine learning algorithm to converge faster or perform better.

```
[ ] data.columns = data.columns.astype(str)
    scaler = MinMaxScaler()
    clean = scaler.fit_transform(data)
    train = pd.DataFrame(clean)
```

```
[ ] train.head()
```

	0	1	2	3	4	5	6	7	8	9	...	11	12	13	14	15	16	17	18	19	20
0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.000000	0.335082	0.832357	...	0.001741	0.000037	0.044029	1.901015e-03	1.900839e-03	0.183159	0.0	0.0	0.0	0.002866
1	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.000334	0.335450	0.832361	...	0.001741	0.000037	0.044023	1.901015e-03	1.900839e-03	0.183159	0.0	0.0	0.0	0.002866
2	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.000401	0.335064	0.832354	...	0.001746	0.000056	0.044032	1.901015e-03	1.900839e-03	0.183159	0.0	0.0	0.0	0.002866
3	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.000505	0.335001	0.832355	...	0.001741	0.000019	0.044025	1.901015e-03	1.900839e-03	0.183159	0.0	0.0	0.0	0.002866
4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.000545	0.336234	0.832353	...	0.001759	0.001981	0.044067	1.577550e-07	1.577690e-07	0.022271	1.0	1.0	1.0	0.000000

5 rows × 21 columns

Figure 22: Data Scaling

4.3 Data Modelling

LOF was applied to train a model that learned what ‘normal’ data looked like and could thus identify anomalies. With the LOF scores computed, they defined a threshold to flag outliers. The dataset contained 143,901 normal points and 21,700 outliers, suggesting a significant presence of abnormal patterns that could be outliers in the dataset. We removed these outliers from the dataset to improve the model’s prediction capabilities.

Figure 23: Local Outlier Factor Anomaly Detection Algorithm

After the application of the LOF anomaly detection algorithm and the removal of the outliers. The K-means algorithm was fed the normalised and encoded data to ensure equal weightage of features and was used to predict suspicious and non-suspicious transactions in the dataset. Post clustering, we analysed the distribution of data points across clusters to locate any groupings that seemed atypical or indicative of potential fraud. Figure 24 displays K-means clustering results.

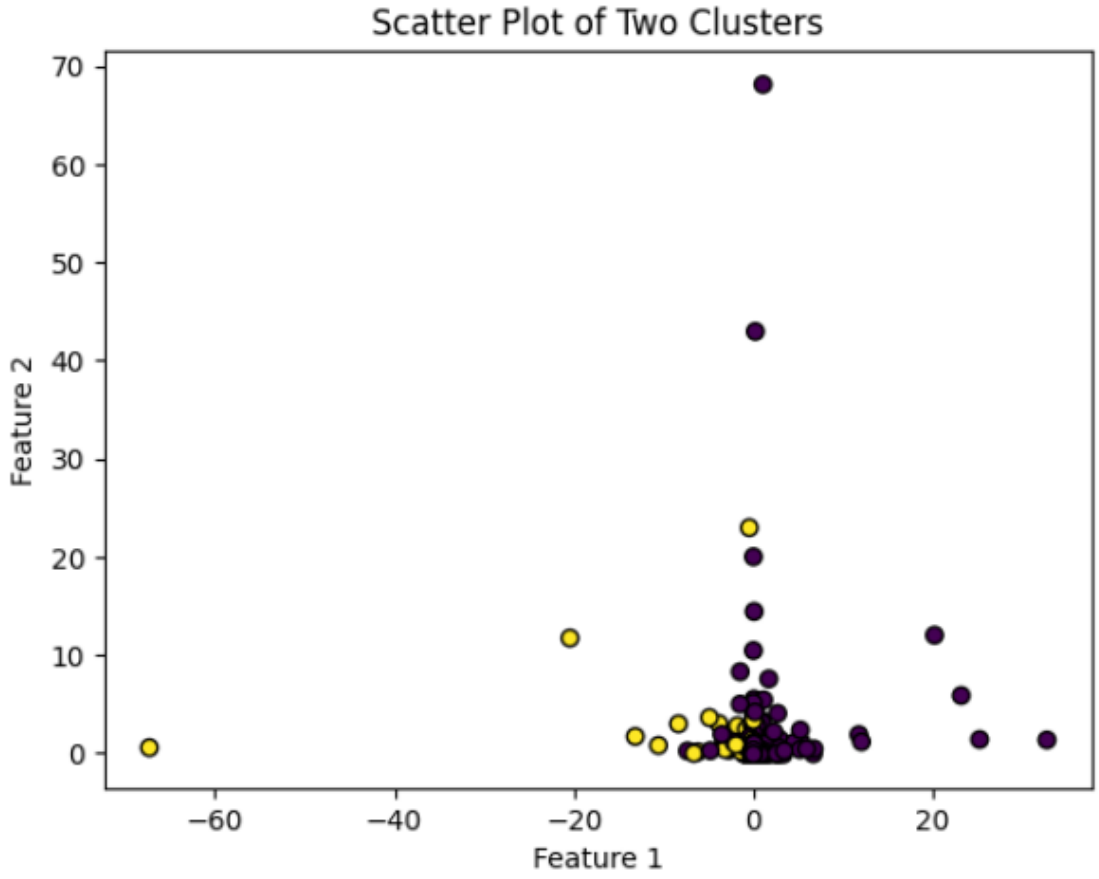

Figure 24: K-means Clustering Results

The Elbow Method was utilized to ascertain the ideal number of clusters. To find the "elbow" point at which the total squared distances is not considerably decreased by adding more clusters, we plotted the sum of squared distances against the number of clusters. Figure 25 shows that the Elbow Plot typically shows a curve that decreases rapidly at first (as “k” increases), and after that, the rate of decline begins to moderate. The point on the plot where the rate of decrease dramatically slows down and forms an elbow-like shape is called the "elbow point". This point suggests a good trade-off between capturing variance and avoiding overfitting. In our case, the sum_of_squared_distances for k values from 1 to 9 have been calculated, and the value we provided (3121038316928719.0) likely corresponds to Sum_of_squared_distances for k=9. The point where the elbow curve occurs is 2, which means two clusters are the optimal number for clustering our datasets. Therefore, developing our model using K-means 2 clusters were used.

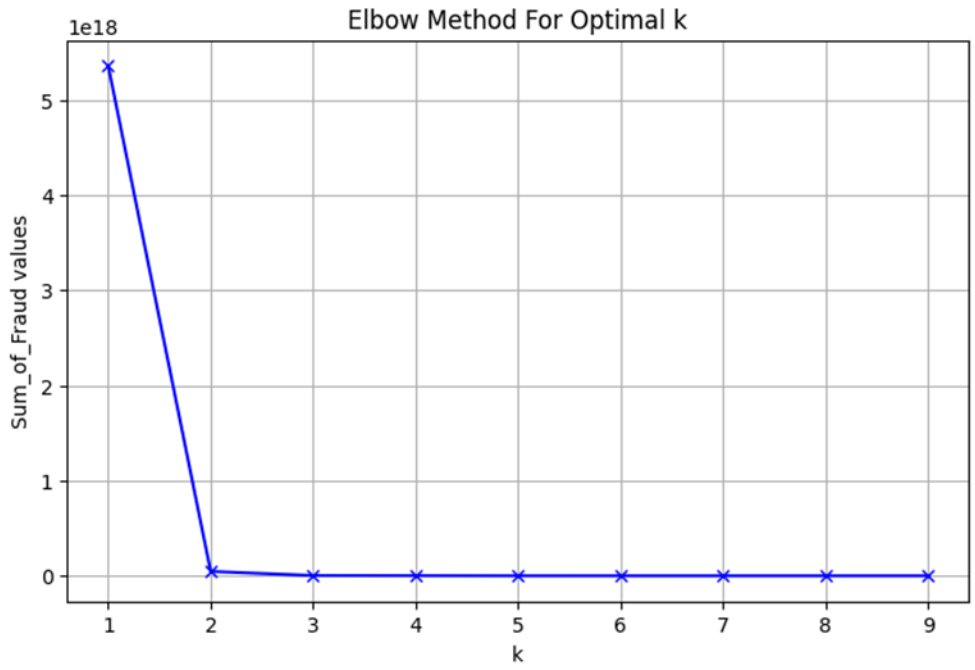

Figure 25: Elbow method for optimal K

Visualisations were an integral part of the analysis, providing intuitive representations of the results. Scatter plots with clusters represented by different colours illustrated how the data points were grouped and offered a visual assessment of the separation between typical and atypical transactions. These visual tools were vital for conveying complex, multidimensional data in a form that was easily interpretable by tax experts and stakeholders. The dataset's labels, indicating fraudulent or legitimate transactions, were compared to the predictions made by the LOF model. We constructed a box plot to visualise the LOF model's negative outlier factor score distribution. The horizontal box plot represents the distribution of LOF scores. LOF is an algorithm used for outlier detection. In this context, positive LOF scores indicate non-fraudulent data points (inliers), and negative LOF scores indicate potentially fraudulent or outlier data points (outliers). Figure 26 box plot shows the LOF scores' median, quartiles, and potential outliers. It helps visualise how well the LOF algorithm can identify potential outliers or fraud cases in the dataset. Outliers in the negative direction (left side of the plot) represent data points more likely to be fraudulent, while inliers will have positive LOF scores.

```
[ ] isfraud_.replace([0,1],[1,-1],inplace=True)
```

+ Code + Text

```
[ ] plt.figure(figsize=(10,2))
    plt.boxplot(lof.negative_outlier_factor_,vert=False)
    plt.show()
```

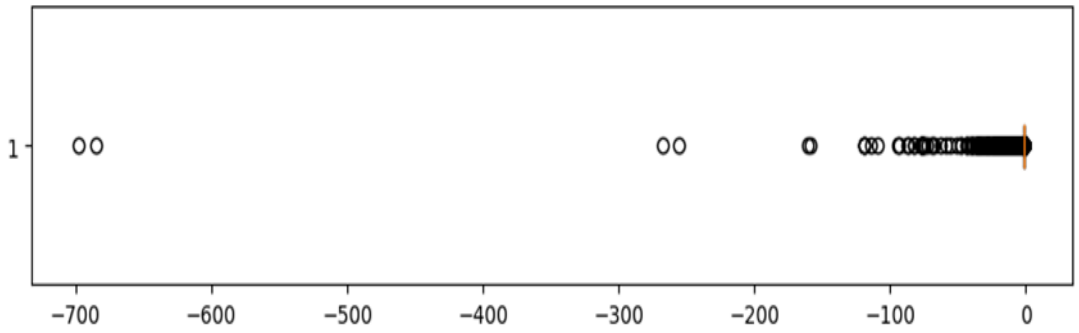

Figure 26: LOF Score

4.4 Model Evaluation

To evaluate the model's performance, we used two methods:

1. We might quantify the degree to which an object is more similar to its cluster (cohesion) than the other clusters (separation) by importing the Silhouette Coefficient from the Python libraries. The analysis revealed that the Silhouette Coefficient has a value of 0.482, as shown in Figure 27. In our case, a Silhouette Coefficient of 0.482 suggests that, on average, the samples are relatively well-clustered. The clusters are reasonably well separated from each other, and the data points are, on average, closer to the centroids of their clusters than to those of other clusters. This value indicates a reasonably good cluster configuration, but there might still be room for improvement.

```
[ ]  # Calculate Silhouette Coefficient
     from sklearn.metrics import silhouette_score

     kmeans = KMeans(n_clusters=3, n_init=10,random_state=42)
     kmeans.fit(train)

     sil_coeff = silhouette_score(train, kmeans.labels_)
     print("Silhouette Coefficient:", round(sil_coeff, 3))

     Silhouette Coefficient: 0.482
```

Figure 27: Silhouette Coefficient Output

2. To review the model further, we then submitted the result set of clusters in the declarations to an auditor of the ZRA. We provided the auditor with 100 VAT-filed return declarations so they could be assessed. Our model identified 40 of them as false. The auditor employed an approach to ascertain the veracity of the declarations by utilizing credibility metrics established within the tax system. Out of the declarations that were not fraudulent, the auditor did not identify any as suspicious. It would appear from this that our approach correctly flags statements as suspect. The auditor identified 35 questionable declarations among the false ones flagged by our model. Due to the need for auditing, which typically spans many months, it remains uncertain whether our model demonstrates a low type 1 error when determining the authenticity of these invoices. Nevertheless, the anticipated outcome was not surprising, as the proposed model incorporates a greater amount of information regarding the return declarations in contrast to the auditor's intuition. Our concept proposes a tax auditing priority.

4.5 System Automation and Implementation Results

This part presents the outcomes of automating the system and implementing the web application, as well as the findings from validating the system. This prototype can be implemented as an Application Programming Interface (API) integration into the taxation system, enabling auditors to access the tool and execute predictions for detecting taxpayers who file false returns.

4.5.1 System Implementation Results

The system automation process begins with a user signing up on the web application. In the reasearch, the users of the system are the tax auditors that use this system to detect tax fraud.

To successfully sign up, the tax auditor must provide a username, first name, last name, and password, see Figure 28.

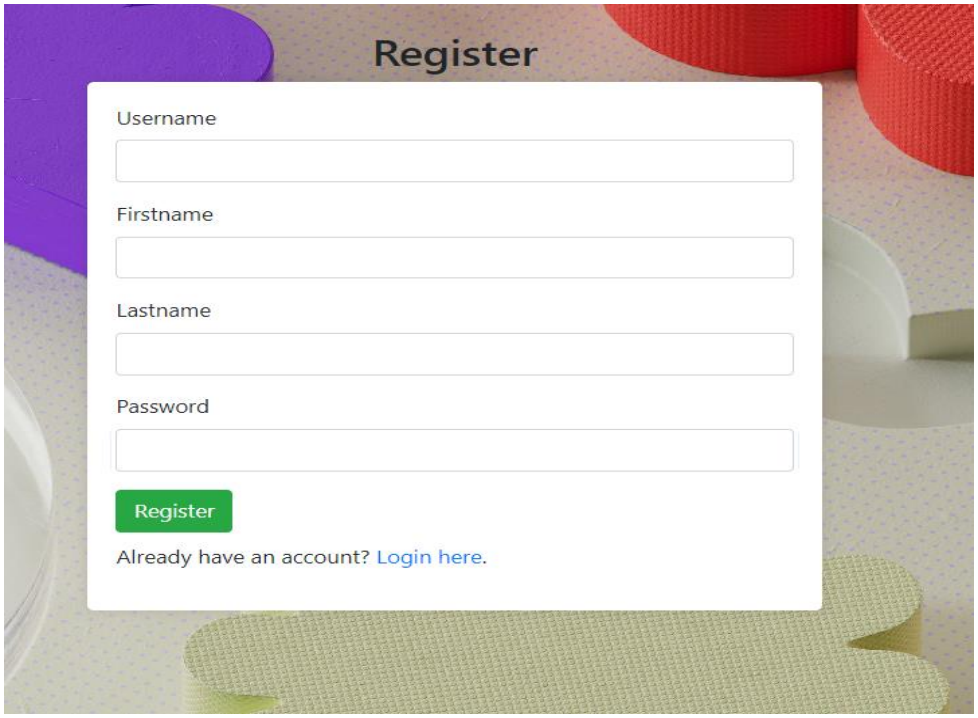

Figure 28: User Creation and Signup

Once the tax auditor has successfully signed up, the user can then log in by providing a valid username and password

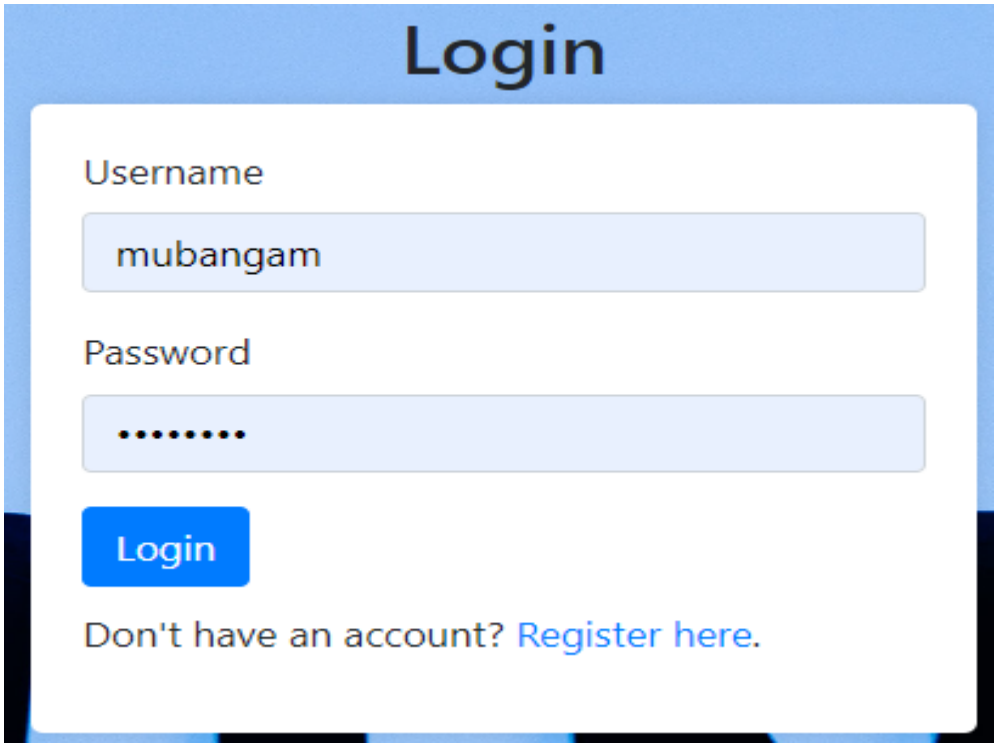

Figure 29: Login Page

On the landing page, we provide an input form that allows the tax auditor to input a TPIN that is used to predict if the return filed by this TPIN is suspicious or not. A valid TPIN must be provided to generate a result. Additionally, the input form comes with a link that allows the users to download and access sample returns from the database which can used to determine if a return filed by a particular TPIN is fraudulent or not. Fig.30 displays the home page of the system.

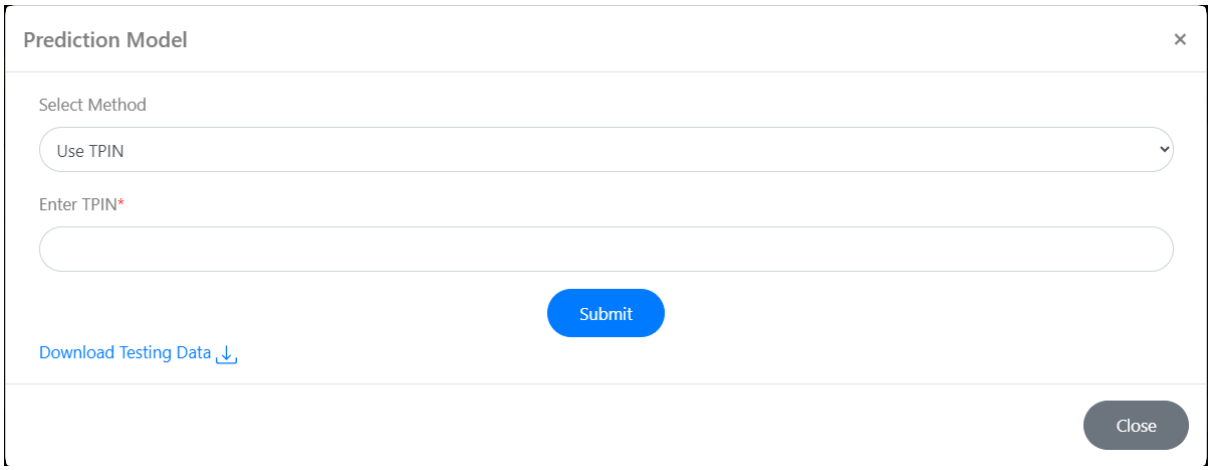

Figure 30: Home Page

The system prediction model displays the output information “This return data is not suspicious” to the user when no fraud is detected.

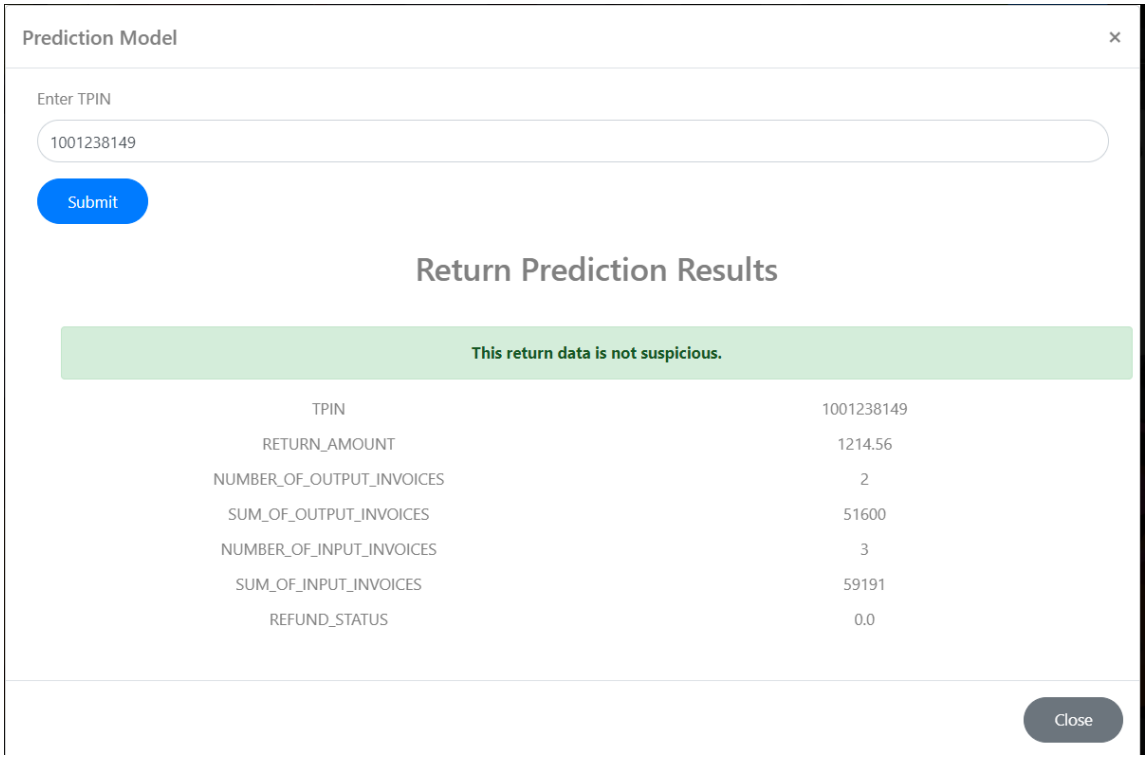

Figure 31: Prediction Display Results – No Fraud

Similarly, when fraud is detected, the system displays an output message to the tax auditor, stating “This return data is suspicious." As illustrated in Figure 32.

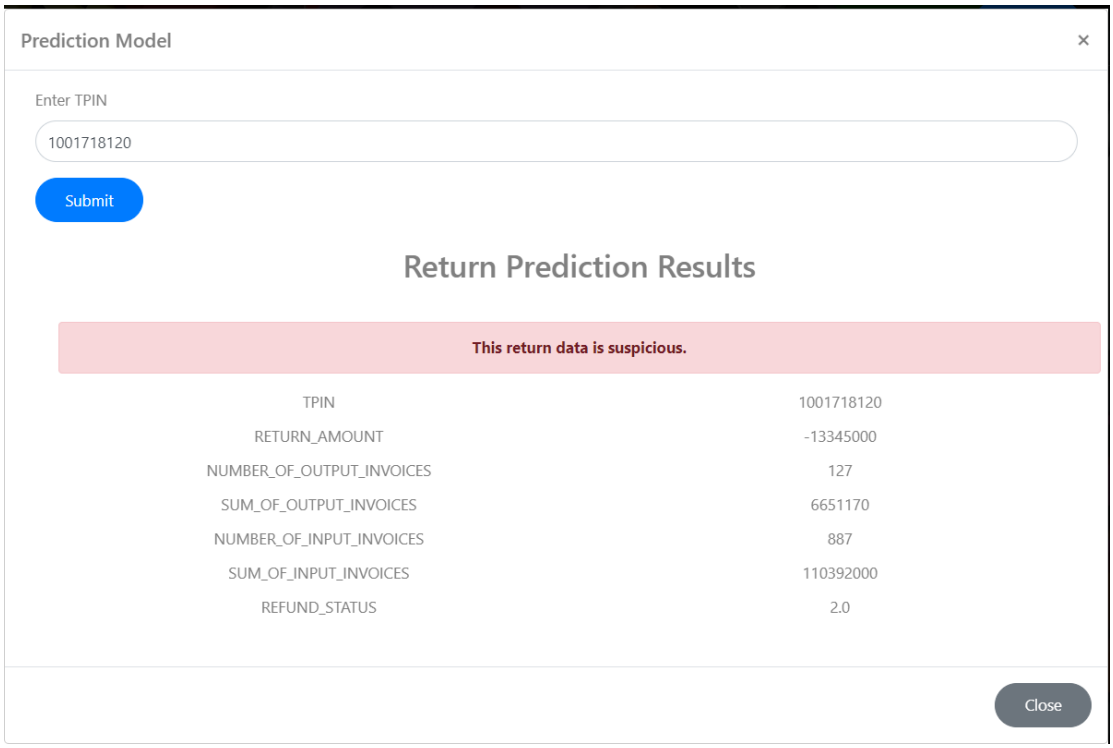

Figure 32: Prediction Display Results - Fraud

4.5.2 System Validation

The system validation process for the web-based application involved conducting a series of functional tests during development and user acceptance testing. This is to verify that various components of the system perform as expected. The validation covers essential functions like accessing the home page, user login, prediction capabilities, and informational content about heart disease and the system itself.

1. Home page: When a user accesses the website, the system validates by successfully displaying the home page, confirming the correct functioning of the initial access point of the application.
2. Login page: For the login page, there are two validation scenarios. Firstly, when a user inputs the correct username and password, the system validates by correctly navigating to the admin page and displaying the data relevant to the logged-in user. This demonstrates the system's ability to authenticate and authorise users accurately. The second scenario involves entering an incorrect email and password. When the user enters an incorrect email and password, the system confirms its validation by redirecting them back to the login page and displaying an appropriate error message.
3. Prediction: We input the user's data and watch the system's response to verify the prediction functionality of the system. The system successfully processes this input and displays the results of the prediction, confirming the validity and effectiveness of the prediction mechanism in the application. The tax auditor has two options to check whether a particular taxpayer's tax filing is fraudulent or not. Firstly, the auditor can submit a Taxpayer Identification Number (TPIN) in the provided field box of specific interest. Secondly, the auditor can click the provided link to download the Excel document containing the list of TPINs.
4. Information about the tax return: the system's validation includes testing access to informational content. The system correctly displays information about tax returns. The successful display of this content validates the system's ability to provide information to users.

4.5.2.1 User Acceptance Testing

In this section, we provide the user acceptance test results that were conducted during the deployment of the application. Table 8 displays the User Acceptance test results that were conducted.

Table 8: UAT Test Case

	Test Case	Test Conditions	Test Input	Expected Result	Actual Result	Status
1.	Usability	Check if the users can successfully register	Users submit personal information according to the rule form	Users register successfully	Users register successfully	Pass
		Check if users can input the TPIN number	The user inputs a TPIN number for prediction	The user inputs a TPIN successfully	The user inputs a TPIN successfully	Pass
2.	Integration of K-means prediction model (Functionality)	Check if the integrated model will predict if the TPIN's VAT return is suspicious or not.	An automatic fraud detection predictor using K-means	Accurately predict Output	Engineering Eligible	Pass with remarks
		Check if the prediction model generates an output		The system successfully displays the output of the prediction model to the user	Output successfully displayed	Pass
3.	Security	Check if the username and	Input correct information	Login successful	Login successful	Pass

		password are working	Input wrong information	Login failed	Login failed	Pass

4.6 Chapter Summary

This chapter showcases the outcomes of the feature selection and data exploration analysis. We go into the utilization of the Local Outlier Factor (LOF) method for detecting anomalies and the application of K-means clustering for prediction, specifically in the context of fraud detection. We also present the results of the model evaluation, system automation, and validation carried out.

5 DISCUSSION AND CONCLUSIONS

5.1 Introduction

The research questions from the first chapter are addressed in this chapter. We examine the offered solution for Fraud detection in ZRA by describing the produced model for Fraud detection and the employed assessment methods for this model. Subsequently, we derived a conclusion based on the study and provided a concise summary of the research investigation.

5.2 Discussion

The detection of tax fraud in Zambia currently depends largely on traditional methods, involving the domain of knowledge of tax experts. The time and resources required for this method are high. Moreover, there is a conspicuous lack of annotated datasets accessible for the identification of tax fraud in the nation. The purpose of this section is to analyze the results obtained from the research questions presented in the first chapter, to tackle the difficulties that have been discovered.

5.2.1 Objective 1 Discussion

The primary goal was to develop a model that could assist tax auditors in detecting taxpayers who file false returns, doing so more efficiently and effectively, while requiring less time and resources. Understanding the business procedures associated with the VAT tax type was first required to determine the requisite feature set to accomplish this goal. Following this business understanding phase, data from 229,000 returns filed by VAT-registered taxpayers were collected from the tax administration database. There is no text provided. A comprehensive exploratory data analysis was performed on this dataset to acquire a thorough comprehension of the data. However, because there isn't a historical dataset available, which would have served as a labelled dataset for predicting false returns by taxpayers, unsupervised machine learning was employed to address limited labelled dataset availability. During the pre-processing stage, we handled columns with missing values using forward and backward fill methods for categorical features, and by replacing missing values with mean values for numerical features. Given the nature of our problem, the K-means algorithm, and the Local Outlier Factor (LOF) algorithm were chosen as the models to predict taxpayer returns and to identify and remove outliers from the datasets, respectively. The LOF successfully identified 21,700 outliers, which were then removed to enhance the accuracy of the predictions. Subsequently, the K-means algorithm was applied as our classifier, dividing the dataset into categories of fraudulent and non-fraudulent return predictions.

5.2.2 Objective 2 Discussion

Evaluating the built model's accuracy in detecting fraudulent tax return files was the second goal of this study. The model's performance was assessed using two distinct methods, each offering unique insights into its effectiveness. Firstly, the Silhouette Coefficient, derived from Python libraries, was utilized to gauge the model's clustering efficacy. With a coefficient of 0.482, the model demonstrated a satisfactory level of cluster separation and cohesion. This indicates that, on average, the clusters are reasonably well-defined, with data points tending to be closer to the centroids of their clusters than to those of other clusters. While this result signifies a good cluster configuration, it also suggests potential areas for refinement. Secondly, the practical application of the model was tested through a collaboration with a ZRA auditor. By presenting a set of VAT-filed return declarations, including those flagged as fraudulent by our model, we sought to validate the model's real-world effectiveness. The findings were promising: all non-fraudulent declarations identified by the model were collaborated by the auditor's assessment, indicating a high level of accuracy in the model's non-fraudulent predictions. Among the fraudulent declarations identified by the model, the auditor confirmed a suspicious one. This outcome, while limited in scope, underscores the model's potential utility in prioritising tax audits. However, it is crucial to note that the full validation of the fraudulent predictions is a time-consuming process involving extensive auditing. Therefore, while the initial results are encouraging, the definitive efficacy of the model in reducing type 1 errors (false positives) remains to be thoroughly established in practice.

5.2.3 Objective 3 Discussion

The third objective of this research was centered on developing a Flask-based web application prototype designed to predict fraudulent VAT return filings. This application was conceptualized to aid in the automatic classification of VAT returns, focusing on effective fraud detection among taxpayers. To achieve this, the K-means clustering algorithm was seamlessly integrated into the web application, serving a crucial role in identifying potentially fraudulent transactions. The application, developed using the Flask framework, offers a user-friendly interface where users can enter a Taxpayer Identification Number (TPIN). Upon input, the application employs the K-means algorithm to analyse and classify the return transaction, subsequently providing the user with detailed insights into the nature of the return. This feature enhances the transparency and efficiency of fraud detection in VAT transactions. For data management and storage, the application utilises a MySQL database, where all VAT taxpayers' return details are securely stored. This robust database integration ensures the integrity and accessibility of data for analysis. Users of the application are required to first register and create

profiles, post which they gain access to the system platform. Enhancing the entire operation of the application, this phase guarantees a secure and customized experience for every user. By combining several technologies, we were able to construct the web application. HTML, CSS, and JavaScript were utilized for front-end web development, leading to the creation of an interactive and aesthetically pleasing user interface. Python, renowned for its efficiency in handling machine learning tasks, was utilised to implement the K-means clustering algorithm and other backend functionalities. We chose the Flask framework as the web application framework, known for its simplicity and ability to create lightweight yet powerful web applications. This combination of technologies ensured seamless integration of machine learning capabilities into a user-friendly web interface, fulfilling the objective of developing an effective tool for predicting and analysing fraudulent VAT transactions.

5.3 Conclusions

All things considered, this study has successfully shown how the model may be used as a cutting-edge instrument to improve the efficacy and efficiency of tax fraud detection in Zambia. This model offers a promising alternative to standard approaches that are more time-consuming and resource-intensive by incorporating unsupervised machine learning techniques such as K-means grouping and Local Outlier Factors. It offers a data-driven methodology, allowing tax authorities to prioritise their audits in a more informed and efficient manner, even in the absence of historically labelled data. The model, specifically designed for VAT tax-registered businesses, shows significant promise in its initial evaluations, suggesting that it could be a valuable asset in streamlining tax fraud detection processes. On the other hand, it is essential to acknowledge that, even though the model exhibits encouraging outcomes, it is not devoid of obstacles. Ongoing refinement and extensive real-world testing are essential to fully understanding and enhancing the model's accuracy and practical applicability. The initial positive evaluations underscore its potential, yet they also highlight the need for continuous improvement and thorough validation in practical scenarios to ensure its reliability and effectiveness in real-world tax fraud detection.

5.4 Recommendations

In light of the experimental results and the limitations identified in both the model and the prototype, this study proposes the following recommendations:

1. Improving the Quality of the Dataset: During the course of the data collection analysis, it was discovered that the database of the tax administration had a problem with the quality of the data. A significant imbalance in the data was observed, because of factors

such as key processes being managed externally to the database. For instance, the refund process, which is crucial for fraud detection, was found to be manually handled by the Zambia Revenue Authority (ZRA). This meant that many vital transactions were not automated and subsequently not stored in the database. Therefore, it is recommended that all processes be automated to enhance the efficiency of the fraud detection process.

2. Incorporation of More Features and Variables: To develop a highly effective model with minimal prediction errors, it is essential to continuously incorporate new features and variables. This approach will assist in identifying high-value features that are fundamental to developing high-quality models.

5.5 Limitations

1. The accessible variables do not provide a thorough description of the VAT transactions and refunds that have been reported. 1. The feature set is not comprehensive. This makes some VAT returns look fraudulent, even though they are not.
2. Another limitation was the moderate amount of data. The tax system is a relatively new system, which makes the availability of data low, and because of privacy policies and concerns, incorporating information from financial institutions, which can be vital in this study, is unavailable.
3. Moreover, because our data is not labeled with "Fraud" or "not fraud" labels, it is not possible to directly evaluate the accuracy of our model for screening questionable tax reports. It was with the assistance of a tax auditor that a partial assessment was carried out. Because auditing is necessary to identify whether or not a tax declaration is fraudulent, this evaluation is only partial. The completion of such a process typically takes many months.

   Furthermore, there are two ways to enhance the clustering method using K-means that divide fraudulent from non-fraudulent returns in each cluster: either incorporate the tax auditors' intuition into the decision-making process after several well-verified case reviews or flag the left-most mode of the distribution as suspicious.

5.6 Future Works

For future work,

1. Combining unsupervised clustering and incorporating domain knowledge from the tax auditors into the unsupervised algorithms model in a user-friendly way would be more beneficial to the tax auditors.
2. Further, in this work, we have only used VAT taxpayer's data. In the future, it will be better to work with a vast amount of data by integrating more tax types and including more taxpayer details. Such as income tax, PAYE, and Turnover Tax.
3. Despite Ethical issues surrounding data analysis, integration, and access to financial accounts such as the Bank, access to bank transaction information would improve the Quality of the dataset.

5.7 Chapter Summary

The findings of the study were examined and summarized in this chapter. The chapter demonstrated the findings that were obtained from the study questions. The findings of the hypothesis testing were also given, and a conclusion was reached as a result of the findings.

REFERENCES

[1] M. Mwila, “The Taxation System in Zambia,” JCTR Repository Home, JULY 2020. [Online]. Available: https://pmrczambia.com/wp-content/uploads/2020/08/Taxation-in-Zambia-Infographic.pdf. [Accessed 07 December 2023].

[2] H. L. M. d. Carvalho, “TAX GOVERNANCE: A STUDY OF ITS EFFECTS ON TAX EVASION,” *Brazilian Business Review,* vol. 19, no. 4, pp. 454-474, 2022.

[3] Y. X. a. b. H. L. a. b. B. S. a. b. J. W. a. b. B. D. b. c. Qinghua Zheng a b, “A Survey of Tax Risk Detection Using Data Mining Techniques,” *Engineering,* 2023.

[4] ZRA, “ Annual report 2021,” ZRA , 2021. [Online]. Available: https://www.zra.org.zm/wp-content/uploads/2022/05/Annual-Report-2021.pdf. [Accessed 27 August 2023].

[5] B. P. A. M. a. M. d. P. V. Daniel de Roux, “Tax fraud detection for under-reporting declaring using unsupervised machine learn-ing Approach.,” *Applied data science track paper,* pp. 19-23, 2018.

[6] M. M. a. J. Phiri, “Fraud detection on Bulk tax data using business intelligence data mining tool: A case of Zambia Revenue Authority.,” *IJARCCE journal 2016,* vol. 5, 2016.

[7] C.-E. T. Trevor Chan, “Audit lead selection and yield prediction from historical tax data using artificial neural networks,” *PLoS One,* vol. 17, no. 11, 2022.

[8] ZRA, “zra.org.zm,” ZRA website, ‘VAT GUIDE,’ , 2020. [Online]. Available: https://www.zra.org.zm/wp-content/uploads/2020/07/VAT-Guide.pdf . [Accessed 23 August 2023].

[9] J. V. a. D. Martens, “Value-added tax fraud detection with scalable anomaly detection techniques.,” *Applied soft computing article,* vol. 86, January 2020.

[10] C. Lee, “Deep learning-based detection of tax frauds: an application to property acquisition tax,” *Data Technologies and Applications,* vol. 56, no. 3, pp. 329-341, 2021.

[11] I. BENUOGA, FACTORS INFLUENCING TAX EVASION IN ZAMBIA, AND THE ADVERSE EFFECT ON THE ECONOMY., Lusaka, 2020.

[12] Matthew Jenkins, “Corruption risks in tax administration, external audits and national statistics,” Transparency International, 2018.

[13] A. H. Shebo Nalishebo, “Uncovering the Unknown - An Analysis of Tax Evasion in Zambia.pdf,” zambia insitiute for policy analysis and research, lusaka, 2014.

[14] p. m. a. r. center, “THE ROLE OF TAX MORALE IN ENHANCED TAX COMPLIANCE: Increasing Domestic Revenue Collection through Improved Tax Morale,” PMRC, July 2014. [Online]. Available: https://www.pmrczambia.com/wp-content/uploads/2015/06/The-Role-of-Tax-Morale-in-Enhanced-Tax-Compliance-Taxation-Background-Note.pdf. [Accessed 8 November 2023].

[15] Y. X. a. b. H. L. a. b. B. S. a. b. J. W. a. b. Qinghua Zheng a b, “A Survey of Tax Risk Detection Using Data Mining Techniques,” *Engineering,* 2023.

[16] WaronWant, “How Zambia is losing $3 billion a year from corporate tax,” WaronWant, [Online]. Available: https://www.waronwant.org/sites/default/files/WarOnWant_ZambiaTaxReport_web.pdf. [Accessed 6 November 2023].

[17] ZRA, “TAX AUDITS,” Zambia revenue authority, 2020. [Online]. Available: https://www.zra.org.zm/wp-content/uploads/2020/01/Tax-Audit.pdf. [Accessed 8 November 2023].

[18] ZRA, “Tax AUdit,” Zambia revenue Authority, 2020. [Online]. Available: https://www.zra.org.zm/wp-content/uploads/2020/01/Tax-Audit.pdf. [Accessed 5 November 2023].

[19] A. Y. S. Musonda Kabinga, “Zambia VI: Tax administration,” taxjustice-and-poverty.org, [Online]. Available: https://taxjustice-and-poverty.org/fileadmin/Dateien/Taxjustice_and_Poverty/Zambia/Country_Report/CHAPTER_VI.pdf. [Accessed 5 November 2023].

[20] R. A. a. J. L. Munawer Sultan Khwaja, “Risk-Based Tax Audits: Approaches and Country Experiences,” The world bank, Washington DC, 2011.

[21] ZRA, “2015 Annual report,” ZRA, 2015. [Online]. Available: https://www.zra.org.zm/wp-content/uploads/2021/03/2015-Annual-Report.pdf. [Accessed 5 November 2023].

[22] C. O.-y. L.-I. C. C. Y. c. Roung-Shiunn Wu, “Using data mining technique to enhance tax evasion detection performance,” *Expert Systems with Applications,* vol. 39, no. 10, pp. 8769-8777, 2012.

[23] T. Dreisbach, “Creating an electronic tax administration system in Zambia,” Global Delivery Initiative; KDI School, Lusaka, 2019.

[24] J. P. Vernon Mukuwa, “The Effects of E-Services on Revenue Collection and Tax Compliance among SMEs in Developing Countries: A Case Study of Zambia,” *Open Journal of Social Sciences,* no. 8, pp. 98-108, 2020.

[25] H. KAFUSHA, AN EMPIRICAL CORRELATIONAL STUDY ON ZAMBIA REVENUE AUTHORITY (ZRA) ELECTRONIC-TAX SYSTEM AND ADOPTION LEVELS AMONG SMALL, LUSAKA, ZAMBIA: UNZA, 2022.

[26] Y. Y. Yuhan He, “Digitalization of tax administration and corporate performance: Evidence from China,” *International Review of Financial Analysis,* vol. 90, November 2023.

[27] M. A. Zhou Gideon, “Systems, Processes, and Challenges of Public Revenue Collection in Zimbabwe,” *American International Journal of Contemporary Research,* vol. 3, no. 2, 2013.

[28] B. C. Kifordu, Machine Learning in Tax Administration: A case study of Nigeria, Nigeria, 2021.

[29] J. Martikainen, Data Mining in Tax Administration - Using Analytics to Enhance Tax Compliance, Aalto University, 2012.

[30] S. S. a. D. SINANC, “Big Data: A Review,” *2013 International Conference on Collaboration Technologies and Systems (CTS),* pp. 42-47, 2013.

[31] N. Y. DR. A. N. Nandakumar, “A Survey on Data Mining Algorithms on Apache Hadoop platform,” *International Journal of Emerging Technology and Advanced Engineering,* vol. 4, no. 1, 2014.

[32] D. J. K. Jasna Atanasijević, “USING BIG DATA ANALYTICS TO IMPROVE EFFICIENCY OF TAX COLLECTION IN THE TAX ADMINISTRATION OF THE REPUBLIC OF SERBIA,” *researchgate,* 2018.

[33] A. Collosa, “Big Data in Tax Administrations,” Kluwer International Tax Blog, 16 July 2021. [Online]. Available: https://kluwertaxblog.com/2021/07/16/big-data-in-tax-administrations/. [Accessed 3 December 2023].

[34] P. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, New York, 2011.

[35] H. Smaya, “The Influence of Big Data Analytics in the Industry,” *Open Access Library Journal,* vol. 9, no. 2, 2022.

[36] M. B. L. K. Patrick Mikalef, “Big data analytics and firm performance: Findings from a mixed-method approach,” *Journal of Business Research,* vol. 98, no. 2, pp. 261-276, 2019.

[37] M. B. b. G. L. b. J. K. A. Patrick Mikalef, “Big data analytics and firm performance: Findings from a mixed-method approach,” *Journal of Business Research,* vol. 96, pp. 261-276, 2019.

[38] P. Akhtar, J. G. Frynas and K. a. U. S. Mellahi, “Big data-savvy teams’skills, big data-driven actions and business performance,” *British Journal of Management,* vol. 30, no. 2, p. 252–271, 2019.

[39] H. D. &. F. S. Hashem. T., “Role of big data analytics in increasing brand equity within the pharmaceutical industry,” *Academy of Entrepreneurship Journal,* vol. 28, no. 1, pp. 1-13, 2018.

[40] H. Al-Malahmeh, “Influence of Business Intelligence and Big Data on Organizational Performance,” *Journal of System and Management Sciences,* vol. 12, no. 5, pp. 193-212, 2022.

[41] J. RANJAN, “BUSINESS INTELLIGENCE: CONCEPTS, COMPONENTS, TECHNIQUES AND BENEFITS,” *Journal of Theoretical and Applied Information Technology,* vol. 9, no. 1, pp. 60-70, 2009.

[42] T. Maaitah, “The Role of Business Intelligence Tools in the Decision Making Process and Performance,” *Journal of Intelligence Studies in Business,* vol. 13, no. 1, p. 43–52, 2023.

[43] P. S. A. P. Girdhar, “Online Analytical Processing (OLAP),” *International Journal for Research in Computer Science,* vol. 2, no. 7, 2015.

[44] F. H. Zawaideh, “The impact of the Data-Warehouses and the Online AnalyticalProcessing in the risk management processes on the Jordanian insurance companies,” *International Journal of Business and Management Invention,* vol. 6, no. 5, pp. 66-75, 2017.

[45] T. S. A. A. D. KulthidaTuamsuk, “Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012,” *International Journal of Data Mining & Knowledge Management Process,* vol. 2, no. 5, September 2012.

[46] A. S. Vardan Baghdasaryan, “Improving Tax Audit Efficiency Using Machine Learning: The Role of Taxpayer’s Network Data in Fraud Detection,” *Applied Artificial Intelligence,* vol. 36, no. 1, 2021.

[47] T. Dreisbach, “Creating an Electronic Tax Administration System in Zambia,” Global Delivery Initiative, Lusaka, Zambia, 2021.

[48] P. a. S. R. P. Daniele Micci-Barreca, “Improving Tax Administration with Data Mining,” SPSS, 2009.

[49] J. A. S. A. S. S. Seema Sharma, “Machine Learning Techniques for Data Mining: A Survey,” *IEEE International Conference on Computational Intelligence and Computing Research ,* 2013.

[50] D. P. M. P. Neelamadhab Padhy, “The Survey of Data Mining Applications And Feature Scope,” *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT),* vol. 2, no. 3, 2012.

[51] X. T. a. Y. Gong, “Research on Application of Machine Learning in Data Mining,” *IOP Conference Series: Materials Science and Engineering,* vol. 392, no. 6, 2018.

[52] H.-D. Wehle, “Machine Learning, Deep Learning, and AI: What’s the Difference?,” *Steinbeis Transfer Center for Industrial Digitalization,* 2017.

[53] S. Arora, “Data Mining Vs. Machine Learning: The Key Difference,” Simplilearn, 11 August 2023. [Online]. Available: https://www.simplilearn.com/data-mining-vs-machine-learning-article#:~:text=Can%20machine%20learning%20be%20used,in%20the%20data%20mining%20process.. [Accessed 7 December 2023].

[54] S. Li, “Research on Data Mining Technology Based on Machine Learning Algorithm,” *Journal of Physics: Conference Series,* 2018.

[55] S. R. P. C. M. Otávio Calaça Xavier, “Tax evasion identification using open data and artificial intelligence,” *BRAZILIAN JOURNAL OF PUBLIC ADMINISTRATION,* vol. 56, no. 3, pp. 426-440, 2022.

[56] V. Kanade, “What Is Machine Learning? Definition, Types, Applications, and Trends for 2022,” spice works, August 2022. [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/. [Accessed 6 November 2023].

[57] V. L. S. CLIFTON PHUA, “A Comprehensive Survey of Data Mining-based Fraud Detection Research,” *2010 International Conference on Intelligent Computation Technology and Automation,* 2010.

[58] A. I. D. &. M. Y. Michel van de Velden, “Special feature: dimension reduction and cluster analysis,” *Behaviormetrika ,* p. 239–241, 2019.

[59] C. A. A. Lidong Wang, "Machine Learning in Big Data," *International Journal of Mathematical, Engineering and Management Sciences* , vol. 2, no. 1, p. 52–61, 2016.

[60] T. d. O. Barreto, "Artificial intelligence applied to analyzes during the pandemic: COVID-19 beds occupancy in the state of Rio Grande do Norte, Brazil," *Frontiers in Artificial Intelligence,* vol. 6, pp. 1-16, 2023.

[61] C. OPREA, "PERFORMANCE EVALUATION OF THE DATA MINING CLASSIFICATION," *Information society and sustainable development,* pp. 249-253, 2014.

[62] O. J. O. S. Olalekan Akinola, "Accuracies and Training Times of Data Mining Classification Algorithms:An Empirical Comparative Study," *Journal of Software Engineering and Applications,* vol. 8, pp. 470-477, 2015.

[63] S. Misra, "Unsupervised Learning: The Power of the Underdog," Journal of petroleum technology, 2022.

[64] S. G. Pallavi, "A Comparative Performance Analysis of Clustering Algorithms," *International Journal of Engineering Research and Applications,* vol. 1, no. 3, pp. 441-445.

[65] N. C. A. Dimas Aryo Anggoro, "Implementation of K-Nearest Neighbors Algorithm for Predicting Heart Disease Using Python Flask," *Iraqi Journal of Science,* vol. 62, no. 9, pp. 3196-3219, 2021.

[66] R. A. a. ,. M. M. I. a. ,. ∗. ,. M. A. U. Nazin Ahmed a, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering,* vol. 2, p. 229–241, 2021.

[67] A. P. K. N.Pavan Kumar, "Stock Market Trend Prediction Using Machine Learning Algorithms And Performing Comparative Analysis," *international journal of creative research thoughts,* vol. 4, no. 11, 2023.

[68] R. S. O. R. Kefas Bagastio, "Development of stock price prediction system using Flask framework and LSTM algorithm," *Journal of Infrastructure, Policy and Development,* vol. 7, no. 3, 2023.

[69] N. D, "FILM SAGA – A MOVIE RECOMMENDATION SYSTEM USING MACHINE LEARNING," *International Journal for Research Trends and Innovation,* vol. 7, no. 7, pp. 1263-1267, 2022.

[70] G. N. Bernard Shibwabo Kasamani, "A System for Recommending Rental Properties," *Journal of Systems Integration,* vol. 3, pp. 10-18, 2017.

[71] A. L. A. R. Regatte Sahithi Reddy, "PLANT IDENTIFICATION SYSTEM USING MACHINE LEARNING," *international journal of creative research thoughts,* vol. 11, no. 4, pp. 565-570, 2023.

[72] "A Machine Learning Model for Early Prediction of Crop Yield,Nested in a Web Application in the Cloud: A Case Study in an Olive Grove in Southern Spain," *agriculture,* vol. 12, p. 1345, 2022.

[73] A. R. B. R. V. K. M. D. K. Andre Esteva, "A guide to deep learning in healthcare," *naturemedicine,* vol. 25, p. 24–29, 2019.

[74] B. Smith and G. Linden, "Two Decades of Recommender Systems at Amazon.com," *IEEE Internet Computing,* vol. 21, no. 3, 2017.

[75] P. K. P. Anuj Sharma, "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," *International Journal of Computer Applications,* vol. 39, no. 1, 2012.

[76] "Optimizing supply chain management through the use of predictive analytics," stldigital, [Online]. Available: https://www.stldigital.tech/blog/optimizing-supply-chain-management-through-the-use-of-predictive-analytics-2/. [Accessed 5 november 2023].

[77] A. J. G. C. Abhinv Sharma, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," *IEEE access,* 2020.

[78] S. K. A. Djeffal Abdelhamid, "Automatic Bank Fraud Detection Using Support Vector Machines," *Proceedings of the International conference on Computing Technology and Information Management,* 2014.

[79] M. Albashrawi, "Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015," *Journal of Data Science,* pp. 553-570, 2016.

[80] P. K. P. A. Sharma, "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," *International Journal of Computer Applications,* vol. 39, p. 0975 – 8887, 2012.

[81] A. B. a. H. Garg, "An Efficient Techniques for Fraudulent detection in Credit Card Dataset: A Comprehensive study," *IOP Conf. Series: Materials Science and Engineering,* 2021.

[82] Y. S. &. Z. W. Emmanuel Ileberi, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," *Journal of Big Data,* 2022.

[83] S. K. Z. B. D. Viji, “An Improved Credit Card Fraud Detection Using K-Means Clustering Algorithm,” *International Journal of Engineering Science Invention,* pp. 59-64, 2018.

[84] P. C. G. a. J. D. Velásquez., “Characterization and detection of taxpayers with false invoices using data mining techniques.,” *Expert Systems with Applications,* vol. 40, p. 1427–1436., 5th April 2013.

[85] L. Ippolito A, “Tax crime prediction with machine learning: a case study in the municipality of São Paulo.,” *ICEIS ,* no. 1, p. 452–459., 2020.

[86] Y. L. H. H. J. R. a. B. D. Qinghua Zheng, “ATTENet: Detecting and Explaining Suspicious Tax Evasion Groups.,” *International Joint Conference on Artificial Intelligence (IJCAI-19),* 2019.

[87] N. S. A. J. T. P. S. S.-P. Yashashwita Shukla, “ Big Data Analytics Based Approach to Tax Evasion Detection.,” *IJERCSE ,* vol. 5, no. 3, March 2018.

[88] C. MUNEZERO, Value Added Tax Fraud Detection using Naïve Bayes Data Mining Approach 2016-2019, Rwanda, 2020.

[89] M. J. D. R. César Pérez López, “Tax Fraud Detection through Neural Networks: An Application Using a Sample of Personal Income Taxpayers,” *Future internet,* vol. 11, no. 4, 2019.

[90] N. Alsadhan, “A Multi-Module Machine Learning Approach to Detect Tax Fraud,” *Computer Systems Science and Engineering,* 2022.

[91] C. O. b. H.-y. L. b. S.-I. C. b. D. C. Y. Roung-Shiunn Wua, “Using data mining technique to enhance tax evasion detection performance,” *Expert Systems with Applications,* vol. 39, p. 8769–8777, 2012.

[92] M. J. D. R. D. L. S. Maria del Camino González Vasco, “Characterization and detection of potential fraud taxpayers in Personal Income Tax using data mining techniques,” *researchgate,* 2018.

[93] B. Murorunkwere, O. Tuyishimire and D. Haughton, “Fraud Detection Using Neural Networks: A Case Study of Income Tax,” *Future Internet,* vol. 14, no. 168, 2022.

[94] J. A. D. J. N. K. Miloš Savić, “Tax Evasion Risk Management Using a Hybrid Unsupervised Outlier Detection Method,” *Expert Systems with Applications,* vol. 193, no. 116409, 2022.

[95] P. J. Phiri, Writer, *RESEARCH METHODS FOR COMPUTER SCIENCE - Lecture notes.* [Performance]. UNZA computer science, 2020.

[96] M. H. C. &. İ. Ş. M. J. Nodeh, “Analyzing and Processing of Supplier Database Based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) Algorithm,” *The International Conference on Artificial Intelli-gence and Applied Mathematics,* 2022.

[97] D. L. Phiri, *Introduction to Machine learning lecturer notes.,* Lusaka, 2019.

[98] R. W. a. J. Hipp, “Crisp-dm: Towards a standard process model for data mining,” *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining,* p. 29–39, 2000.

[99] D. A. Dr. Tarak Hussain, “Visualization and Explorative Data Analysis,” *International Journal of Enhanced Research in Science, Technology & Engineering,* vol. 12, no. 3, March 2023.

[100] P. Patil, “What is Exploratory Data Analysis?,” Towards Data Science, 2018.

[101] Kaggle, “Intro to Exploratory data analysis (EDA) in Python,” Kaggle, 2020. [Online]. Available: https://www.kaggle.com/code/imoore/intro-to-exploratory-data-analysis-eda-in-python. [Accessed 8 november 2023].

[102] T. O.-A. G. R. O. D. C. W. I. Khalid K. Al-jabery, “Data preprocessing,” *Computational learning approach to data analytics in biomedical applications,* pp. 7-27, 2020.

[103] J. Brownlee, Data Preparation for Machine Learning: Data Cleaning, Feature selection and transforms in python, 2020.

[104] scikit-learn, “sklearn.preprocessing.OneHotEncoder,” scikit-learn library, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. [Accessed 07 december 2023].

[105] J. Chong, “What is Feature Scaling & Why is it Important in Machine Learning?,” Towards Data Science, 30 December 2020. [Online]. Available: https://towardsdatascience.com/what-is-feature-scaling-why-is-it-important-in-machine-learning-2854ae877048#:~:text=Feature%20scaling%20is%20the%20process,need%20to%20perform%20feature%20scaling.. [Accessed 7 December 2023].

[106] U. Riswanto, *K-mean Clustering for anomaly detection.*

[107] R. Li, “Detection of Financial Reporting Fraud Based on Clustering Algorithm of Automatic Gained Parameter K Value.,” *International Journal of Database Theory and Application,* vol. 8, no. 1, pp. 157-168, 2015.

[108] M. Cui, “Introduction to the K-Means Clustering Algorithm Based on the Elbow Method,” *Accounting, Auditing and Finance,* pp. 5-8, 2020.

[109] Scikit-learn, “Outlier detection with Local Outlier Factor (LOF),” Scikit-learn.org, 2023. [Online]. Available: https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html. [Accessed 8 November 2023].

[110] S.-l. library, “choosing the right estimator,” Scikit-learn, [Online]. Available: https://scikit-learn.org/sta-ble/tutorial/machine_learning_map/index.html . [Accessed 12 October 2023].

[111] C. Y. Wijaya, “Exploring Unsupervised Learning Metrics,” KDNuggets, 13 April 2023. [Online]. Available: https://www.kdnuggets.com/2023/04/exploring-unsupervised-learning-metrics.html. [Accessed 6 November 2023].

[112] S. V. P. Tejas Shaha, “A Novel Approach for Specifying Functional and Non-Functional Requirements using RDS (Requirement Description Schema),” *Procedia Computer Science,* vol. 79, p. 852 – 860, 2016.

[113] S. C. More, “Python Libraries and Tools for Data Science: A review,” *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT),* vol. 2, no. 9, 2022.

[114] N. S. M. V. A. P. A. &. B. G. Chauhan, “Implementation of database using python flask framework.,” *International Journal of Engineering and Computer Science,* vol. 8, no. 12, pp. 24894-24899, 2019.

[115] G. V. G. Fabian Pedregosa, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Researh,* vol. 12, pp. 2826-2830, 2012.

[116] W. Mckinney, “pandas: a Foundational Python Library for Data Analysis and Statistics.,” in *Python High Performance Science Computer,* 2011.

[117] S. V. Atharva Sapre, “Scientific Computing and Data Analysis using NumPy and Pandas,” *International Research Journal of Engineering and Technology,* vol. 7, no. 12, pp. 1334 - 1347, 2020.

[118] R. C. Arnav Oberoi, “Visualizing data using Matplotlib and Seaborn libraries in Python for data science,” *International Journal of Scientific and Research Publications,* vol. 9, no. 3, pp. 202-206, 2019.

[119] R. G. S. C. JEPPIAAR NAGAR, “PREDICTION OF MULTIPLE DISEASE USING MACHINE LEARNING TECHNIQUES,” INSTITUTE OF SCIENCE AND TECHNOLOGY, 2022.

[120] S. R. Y. A. S. Bhavyanshu Ghariya, “Brain Stroke Prediction,” *International Journal of Advances in Engineering and Management*, vol. 3, no. 10, pp. 813-819, 2021.

[121] O. D. I. Otaduy, “User acceptance testing for Agile-developed web-based applications: Empowering customers through wikis and mind maps,” *Journal of Systems and Software*, vol. 133, pp. 212-229, 2017.

APPENDICES

Appendix 1: Introduction letter

Appendix 2: Publications

One paper based on this study was peer reviewed and accepted. The paper addressed the first two objectives of this research.

Objectives 1 and 2: To develop a data mining model that can be used as a recommender to identify taxpayers filing false returns based on the assessment; To evaluate and validate the performance and accuracy of this model.

The data mining model was developed and evaluated, and I have written a book chapter under Springer Journal. The paper entitled,

'Using a data-driven model to predict taxpayers filing false returns: A case of Zambia Revenue Authority' was peer reviewed, presented, accepted, and awaiting publication