PREDICTIVE MODELLING OF TAXPAYER COMPLIANCE BEHAVIOUR USING MACHINE LEARNING AT NAMRA

By

Sifani Edwin Sifani

Student number: 222074299

Submitted in partial fulfilment of the requirements for the degree of

Master of Data Science

In the

Department of Computing and Informatics

At the

NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY

Supervisor: Dr. R. Maliwatu

Date of Submission: 31/07/2025

METADATA

TITLE:	Predictive modelling of taxpayer compliance behaviour using machine learning at NamRA.
STUDENT NAME:	222074299
SUPERVISOR:	Dr. R. Maliwatu
DEPARTMENT:	Informatics
QUALIFICATION:	Master of Data Science
MAIN KNOWLEDGE AREA:	Data Science
SPECIALISATION:	Machine Learning
TYPE OF RESEARCH:	Applied Research
METHODOLOGY:	Quantitative Research
KEYWORDS:	Artificial Intelligence, Machine Learning, Tax administration, Voluntary Compliance, Revenue Collection, NamRA
STATUS:	Thesis
SITE:	Namibia University of Science and Technology (NUST)
DOCUMENT DATE:	31/07/2025
SPONSOR:	N/A

DECLARATION

I, Sifani Edwin Sifani hereby declare that the work contained in the proposal for the Master Degree in Data Science project, titled: "Predictive modelling of taxpayer compliance behaviour using machine learning at NamRA" is my original work and that I have not previously in its entirety or part submitted it at any university or other higher education institution for the award of a degree. I further declare that I will fully acknowledge any sources of information I will use for the research by the institution's rules.

Signature:  Date: 31/07/2025

ABSTRACT

This study explored the application of machine learning techniques for predictive modelling of taxpayer compliance behaviour at the Namibia Revenue Agency (NamRA). Multiple classification algorithms were systematically optimised using 5-fold cross-validated Grid Search and Randomised Search, as implemented in the scikit-learn library (v1.2), to enhance predictive accuracy. The hyperparameter search spaces were tailored to each model's architecture; for instance, Random Forest optimisation included the number of estimators and maximum depth, while Gradient Boosting models emphasised learning rate and structural parameters. The optimisation process yielded notable improvements, with cross-validated accuracy scores ranging from 64% to 68%. The best-performing model, an optimised Random Forest classifier, achieved an accuracy of 68%. These findings demonstrated the efficacy of hyperparameter tuning in improving model performance and underscore the potential of machine learning to support data-driven compliance management at NamRA.

The use of SHAP and LIME in this study provided valuable interpretability of taxpayer compliance predictions, highlighting key factors such as income group, the COVID-19 period, taxpayer registration office, and marital status. These insights align with existing research and reveal how financial capacity, macroeconomic disruptions, and administrative or demographic variables influence compliance. SHAP offers a global view of feature importance, while LIME provides personalised explanations, enhancing trust and communication. Despite modest predictive accuracy, the interpretability benefits support targeted policy interventions and suggest future improvements through richer data and fairness assessments.

DEDICATION

This thesis is dedicated to my caring fiancée Sibongile Netha and my two children Derbby Sifani and King Sifani AKA Chris Hani. I would like to express my heartfelt gratitude for their unwavering support, inspiration, prayers and encouragement. Their support was truly invaluable throughout this academic journey.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my ancestors and Almighty God for enabling me to come this far. My ancestors' guidance provided me with the much needed intelligence, understanding, and expertise to compete this course. Secondly, I would like to thank Dr. Richard Maliwatu, my main supervisor and Dr. Gloria Iyawa as the co-supervisor for their immense backing and guidance during my study. In addition, I will fail in my duty if I do not recognise the wisdom and encouragement from Mbaunguraije Tjikuzu, a highly appreciated brother. Furthermore, I am grateful to Prof. Muyingi for the mentorship and my peers from the Master of Data Science class of 2022. Last but not least, I would like to thank my children and fiancée for their prayers and unending love. I could not have asked for a better family!

Table of contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI Artificial Intelligence

DCNN Deep Convolutional Neural Networks

ERS Eswatini Revenue Authority

GDP Gross Domestic Product

ICT Information Communication Technology

IRAS Inland Revenue Authority of Singapore

ITAS Integrated Tax Administration System

KRA Kenya Revenue Agency

ML Machine Learning

MLP Multilayer Perceptron

MoF Ministry of Finance

NamPost Namibia Post Limited

NamRA Namibia Revenue Agency

NUST Namibia University of Science and Technology

OECD Organisation for Economic Co-operation and Development

OMA's Offices/Ministry/Agencies

SDG Sustainable Development Goal

SDGs Sustainable Development Goals

SSC Social Security Commission

TIN Tax identification number

VAT Value Added Tax

1. Chapter One: Introduction

1. Introduction and Background

Primarily, tax administrations around the globe are tasked with collecting revenue for public service purposes. This research provided an overview of a transformed tax administration in Namibia and how it can leverage the use of Machine Learning to become a data-driven tax administration by exploring the individual salaried returns in gauging whether taxpayer are filing their returns on time. Namibia Revenue Agency (NamRA) is the nation's tax collection agency. According to Puyeipawa, (2021), NamRA was enacted by an act of Parliament and officially opened its operation in April 2021, taking over the functions of the tax administration from the Ministry of Finance. NamRA aspires to be a leading revenue authority, driven by dedication to enhance the well-being of every Namibian. NamRA' mission is to administer and enforce Namibia's tax and customs laws consistently, fairly, efficiently, and effectively, addressing the specific concern of every taxpayer and trader. As emphasised by Liuhong, (2022), data has gone from being a simple transactional object to a fundamental resource. Moreover, apart from being a strategic pillar, data is pervasive, for example, International Monetary Fund, (2024), further highlighted that data plays a pivotal role in driving evidence-based decision-making in Tax administration and has become a daily strategic topic in many tax administrations.

1.1 Namibia Revenue Agency

NamRA was stablished to replace the tax administration roles previously handled by the Ministry of Finance. NamRA operates as a semi-autonomous body to ensure tax efficiency, facilitating trade and promoting compliance. The tax system of Namibia is characterised by a diverse array of tax categories, including Value Added Tax (VAT), customs duties, excise duties, and taxes on both individual and corporate income. The Namibia Revenue Agency (NamRA) serves as the principal agency responsible for collecting tax revenue on behalf of the Namibian government. Its mandate encompasses not only the collection of revenue but also the allocation of these funds towards essential public services such as employment creation, infrastructure development such as building roads and hospitals, and improving educational facilities, thereby supporting the nation's broader developmental objectives. Tax revenue is defined as the total income generated from various taxes, including those on income and profits, social security contributions, taxes on goods and services, payroll taxes, property taxes, and other levies (OECD, 2016).

The structure of the tax system in Namibia is influenced by several factors, including the income levels of individuals and businesses, which determine the applicable tax rates. The complexity of the tax system, combined with issues of non-compliance, particularly in income tax, has emerged as a significant concern. Despite the importance of understanding the extent of tax non-compliance and its financial implications, there remains a notable gap in comprehensive studies addressing these issues in Namibia. Between 2016 and 2019, tax revenue accounted for approximately 30% to 34% of Namibia's GDP, contributing over 90% to the national budget, with the remainder sourced from non-tax revenues (Mwafongwe, 2020).

In response to the challenges faced in tax administration, the Integrated Tax Administration System (ITAS) was launched in January 2019 by the Ministry of Finance (MoF). This system was designed to facilitate the electronic filing of taxes, aiming to enhance efficiency and compliance among taxpayers. The MoF had set an ambitious target of 90% (jvanzyl, 2019) of the taxpayer population to register as e-filers by June 2020, reflecting a significant shift towards the automation of tax processes. The overarching goal of ITAS was to improve service delivery to taxpayers by transitioning from traditional manual filing interactions to a streamlined online filing system, thereby providing numerous benefits, including 24/7 access to tax accounts, self-service facilities, online filing of tax returns, and expedited processing of tax forms with real-time notifications.

The implications of tax defaulters extend beyond mere financial losses; they adversely affect national development and the equitable distribution of resources. Tax defaulters face penalties as stipulated by NamRA which underscores the agency's commitment to enforcing compliance. Furthermore, there is a pressing need for NamRA to evolve into a data-driven, customer-centric organisation that leverages real-time data access to enhance taxpayer engagement and compliance. This transformation is essential as it shifts from a traditional tax administration model to a more modern approach that meets the needs of taxpayers and fosters voluntary compliance. The integration of data science techniques into NamRA's operations is pivotal for optimising the utilisation of taxpayer data and exploring collaborative opportunities with other institutions and key stakeholders.

The modernisation of tax administration systems, such as ITAS, is crucial for enhancing tax compliance, particularly among micro, small, and medium enterprises (MSMEs). According to the literature review in Chapter 2, literature, digitisation has been shown to significantly reduce compliance costs for businesses.. The effectiveness of such digital initiatives is underscored by the need for continuous improvement in tax administration, which can lead to better compliance rates and increased revenue generation for the government.

The Namibian tax system is multifaceted, with various categories of taxes contributing to national revenue. The role of NamRA as the collecting agency is vital for ensuring that tax revenues are effectively utilised for public services and developmental goals. The introduction of ITAS represents a significant step towards modernising tax administration, improving taxpayer compliance, and ultimately enhancing the efficiency of the tax system. However, ongoing efforts are required to address issues of non-compliance and to encourage a voluntary compliance culture among taxpayers, leveraging data-driven strategies and technological advancements to achieve these objectives.

1.2 Problem Statement

Tax revenue is the lifeblood of Namibia's public services, funding critical sectors such as healthcare, education, and infrastructure. The Namibia Revenue Agency (NamRA) is mandated to ensure taxpayer compliance, yet its efforts are hampered by incomplete, unclean, and non-real-time data. Despite a growing tax base, these data limitations, combined with the lack of advanced analytics and intelligence tools, restrict NamRA's ability to accurately identify, segment, and serve individual and business taxpayers.

For instance, as of 31 December 2024, NamRA reported 553,863 registered taxpayers on the ITAS portal, but could not determine how many were individuals or assess their compliance patterns. This impairs NamRA's ability to critically answer questions such as: How many individual taxpayers filed on time, filed late, or did not file at all? How can future compliance behaviour be predicted to support proactive service delivery and enforcement?

The financial implications of these limitations are significant. NamRA recently reported that N$61.8 million from the mass tax refund exercise could not be paid out due to unresolved taxpayer issues, highlighting the real cost of poor data quality and limited taxpayer intelligence. More broadly, late or non-filing of tax returns contributes to revenue leakage, which, when scaled across

the taxpayer base, can result in losses equivalent to an estimated 1% to 3% of Namibia's GDP annually, based on global benchmarks for tax non-compliance in developing economies.

Traditional spreadsheet-based tools have proven inefficient in addressing these challenges. Predictive modelling using machine learning offers a modern, data-driven solution for forecasting taxpayer compliance behaviour. This research sought to explore how such models could be developed and applied to NamRA's data to enhance voluntary compliance, improve service delivery, and support strategic decision-making.

1.3 Research Aims

This study's main aim was to develop and compare supervised machine learning models such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting to predict the filing behaviour of individual salaried taxpayers at NamRA. The models classify tax returns as either on-time or late, using historical individual tax return data to support data-driven decision-making and improve taxpayer compliance. The filing obligation deadline for NamRA is 30 June of each financial year.

Main objective

The main objective of this study was to explore the historical individual return data to predict individual taxpayers' return submission patterns at NamRA.

Sub-objectives

1. To explore data analysis of the individual taxpayer returns to gain a better understanding of the dataset to identify patterns and trends in the submission behaviour of individual taxpayers;
2. To enhance prediction accuracy through enhancement/feature engineering;
3. To assess the performance of ML models in predicting individual taxpayer submission patterns;
4. To analyse and recommend modern ways NamRA can leverage from using emerging technologies (ML/AI).

1.4 Research Questions

The following research questions guided the research:

Main Question

RQ1: Can we explore historical taxpayer data for predicting individual taxpayer submission patterns at NamRA?

The main research question was subdivided into three sub-research questions to carry out comprehensive research.

RQ1.1 What is the present filing pattern of individual taxpayers at NamRA?

RQ1.2 How can ML models be used to achieve NamRA's objectives? Objectives such as improved voluntary compliance and Improved data management and analytical capability can be attained using emerging technologies such as ML and AI at NamRA.

RQ1.3 Which ML model achieves the highest accuracy/F1-score while maintaining interpretability? In this thesis, ML were used to forecast taxpayer filing behaviour based on the submission status.

RQ1.4 How can NamRA leverage emerging technologies such as ML and AI to enhance understanding of taxpayer behaviour (compliance) and data-driven decision making? By exploring historical returns using the due date variable to predict how many taxpayers filed before and after the deadline.

Alignment of research objectives with questions

Table 1.1 Alignment of research objectives with research questions

Research objectives	Research questions
An exploratory data analysis of the individual taxpayer returns to gain a better understanding (identify patterns) of the dataset.	What is the present filing pattern of individual taxpayers at NamRA?
Increase voluntary compliance by predicting taxpayer behaviour.	How can ML models be used to achieve NamRA's objectives?

To assess the performance of ML models in predicting individual taxpayer submission patterns.	Which ML model achieves the highest accuracy/F1-score while maintaining interpretability?
Analyse and recommend modern ways NamRA can leverage emerging technologies (ML/AI) to improve traditional methods.	How can NamRA leverage emerging technologies such as ML and AI to enhance understanding of taxpayer behaviour (compliance) and data-driven decision making?

1.5 Significance of the Study

This research is relevant as it will use Data Science techniques in improving tax management and developing a good trust in data management and analytics to drive the value of data using ML models for evidence-based decision making. By enhancing data-driven decision making and digital services and streamlining interactions, NamRA will foster greater compliance, transparency, and trust in the tax system through understanding of the taxpayers, detect tax-compliance risk and close the tax gaps. Predictive analytics has become a vital tool for strengthening financial compliance by proactively detecting potentially fraudulent activities (Paul-Emeka George et al., 2024).

The revenue collected through these improved processes will directly contribute to the achievement of the United Nations Sustainable Development Goals (SDGs), positively impacting the livelihoods of all Namibians. Through responsible tax administration and digital transformation, NamRA will play a pivotal role in national development and inclusive economic growth.

1.6 Delineation and delimitation of the study

The study was limited to the exploration of individual salaried persons and to evaluate the performance of different Machine models but did not make an in-depth prediction of fraud detection or tax evasion. It is also worth noting that not all machine models were trained in the current study, as the study only focused on the Random Forest classifier, Gradient Boosting Classifier, Extra Trees Classifier, Decision Trees Classifier, Bagging Classifier, XGB classifier, LGBM classifier, K-Nearest Neighbour classifier, AdaBoost classifier and logistic regression.

1.7 Organisation of the thesis

The remainder of this thesis is structured as follows:

i. Chapter 2 presents a review of previous research on the application of data-driven or Machine Learning technology in tax administration.
ii. Chapter 3 presents the approach and methodology used to build the model.
iii. Chapter 4 presents the implementation and results of the model, detailing the performance of different models experimented with.
iv. Chapter 5 presents the conclusion based on findings, future recommendations of the study and adds on with expounding on the contribution of the thesis to the body of knowledge.

2. Chapter Two: Literature Review

2. Introduction

This chapter is based on evaluations of earlier research and theories pertaining to tax administration. It commences with a review of relevant theoretical literature, followed by a discussion of the empirical findings of other studies. The review is focused on research done on tax administration to gauge its effects on tax revenue and compliance. Predictive modelling has emerged as a powerful tool in public sector analytics, particularly in tax administration, where it supports risk-based compliance strategies, fraud detection, and resource optimisation. Traditional approaches to taxpayer segmentation and compliance forecasting relied heavily on rule-based systems and manual analysis, which are limited in scalability and accuracy. The integration of machine learning (ML) offers a data-driven alternative capable of uncovering complex behavioural patterns and improving decision-making. Just like the bird inspired the construction of the aeroplane, this study was influenced by comparable research conducted in the modernisation of tax administration.

Many businesses are getting smarter globally by leveraging data science and other technology models such as Machine Learning (ML) and Artificial Intelligence (AI). Eswatini Revenue Authority (ERS) is aspiring to use data analytics for fraud detection, revenue optimisation and improved process efficiency (ATAF Communication, 2022). NamRA is not spared in using analytics to get smarter and make data-informed decisions, such as tax collection per region, assessment turnaround time, and e-filing system performance. Tax authorities worldwide are increasingly embracing digital technologies to enhance revenue collection through the digital transformation of tax systems. To fully harness the power of data analytics, there must be a fundamental shift in mindset regarding how data is managed, interpreted, and utilised (Gichohi, 2020). NamRA has a vast amount of data that operates in silos. To get good insights from its data, it is vital to have a data governance framework in place to manage the processing, storage, and sharing of its data. NamRA may use AI to examine all its internal as well as external data, openly accessible data, such as state revenue, number of audits conducted by jurisdiction, legal changes, and political focus, to model tax risks (Milner & Berg, n.d.). While researchers agree that prediction and forecasting are the gold of data science, herein we postulate that for a model to predict accurately, it requires quality data and understanding of the taxpayers' filing patterns.

Moreover, without a good governance framework in place, data quality issues would often surface. NamRA has a data strategy that focuses on establishing a structured approach for managing, governing and using data to support decision-making processes.

Currently, NamRA holds tax-related records dating back 20 years, with the data collected from the ITAS database from 2019 up to the 2023 financial year, when the e-filing system was commissioned.

2.1 Overview of the use of analytics and machine learning in tax administration

Much of the existing research on the use of advanced analytics in tax administration, particularly machine learning (ML) and artificial intelligence (AI), centres on two primary themes: understanding their current applications and examining the future potential of AI in the tax domain. AI technologies are widely used to support taxpayer compliance and reduce administrative burdens. At the same time, they improve operational efficiency by automating workflows, audits and decision-making processes. Studies exploring compliance factors have contributed to a deeper understanding of the variables influencing taxpayer behaviour. A 2021 OECD survey of 59 countries found that 80% employed data analytics and 75% used ML techniques, and nearly half implemented digital assistants, mainly chatbots (Bassey et al., 2022). In today's digital landscape, it is essential for tax administrators and policymakers to gain clearer insights into taxpayer behaviour to enhance the effectiveness of tax administration. It is of paramount importance to provide a comprehensive literature review on the use of machine learning (ML) in tax administration, emphasising its role in automating detection processes, enhancing risk assessment, and supporting predictive analytics for more efficient tax management.

2.2 Challenges faced by NamRA

The Namibia Revenue Agency (NamRA) continues to face significant challenges as a relatively new tax authority. Key among these is low voluntary compliance rates and limited taxpayer confidence in the Integrated Tax Administration System (ITAS). Furthermore, NamRA's constrained analytics and reporting capabilities, coupled with fragmented and inaccessible data, impede timely and evidence-based decision-making. The findings suggest that NamRA could benefit from adopting a comprehensive analytics approach that effectively leverages its existing data assets. Integrating machine learning (ML) and artificial intelligence (AI) into its analytical framework may enhance the institution's capacity to predict taxpayer behaviour, monitor compliance more effectively, and improve service delivery outcomes.

Such a transition toward a more data-driven and customer-oriented approach aligns with broader trends in public sector innovation and digital governance (Hayek & Noordin, 2024). Future research may further explore the operational and policy implications of embedding advanced analytics within tax administration systems.

2.3 Machine learning in tax compliance forecasting

Recent studies have demonstrated the effectiveness of supervised ML models in predicting taxpayer behaviour. For instance, Random Forest and Gradient Boosting have been widely used due to their ability to handle noisy, high-dimensional data and provide interpretable outputs through feature importance scores (Zhou et al., 2021). Support Vector Machines (SVM) have shown strong performance in binary classification tasks, particularly in identifying non-compliant taxpayers (Chen et al., 2020).

These models outperform traditional statistical methods by capturing non-linear relationships and interactions between variables, which are common in taxpayer datasets. Moreover, ensemble methods like Random Forest and Gradient Boosting are particularly effective in dealing with imbalanced datasets, a frequent challenge in compliance modelling, where the majority of taxpayers may be compliant, and only a small subset exhibits non-compliance.

2.4 Work most related to this study

Several tax authorities have undertaken similar initiatives to modernise tax administration and enhance voluntary compliance and revenue collection. For instance, the Kenya Revenue Authority (KRA) implemented a project titled *"Harnessing Big Data and Emerging Technologies to Strengthen Domestic Revenue Mobilisation."* The initiative yielded significant improvements in data quality, with the data maturity level rising from 2 to 3 within just one year. Additionally, over 90% of customer biodata were cleansed, resulting in improved traceability and a better understanding of taxpayer profiles. KRA's adoption of Business Intelligence (BI) tools and dashboard reporting has played a central role in its transition to a data-driven administration, enabling more effective decision-making through real-time insights and identification of operational gaps (Gichohi,2020). The methodology followed was aligned with DAMA International's *Data Management Body of Knowledge (DAMA-DMBOK)*, providing a structured framework for effective data governance.

Inland Revenue Authority of Singapore (IRAS)

The IRAS has demonstrated notable success in leveraging data-driven strategies and advanced technologies to enhance taxpayer services and operational efficiency (Mok, 2021). Its achievements are underpinned by a robust data governance framework and a sustained commitment to innovation. However, the scale and sophistication of IRAS's initiatives, such as the integration of Natural Language Processing and Knowledge Graphs, are resource-intensive and may not be directly transferable to contexts with limited technical capacity or infrastructure. For NamRA, the key lesson is the importance of establishing foundational data governance and quality management practices before pursuing advanced analytics. Incremental adoption of analytics, starting with descriptive and diagnostic tools, can yield substantial benefits without overextending resources. Furthermore, IRAS's use of explainable AI (XAI) tools to interpret model outputs provides a model for transparency and accountability, which NamRA should emulate to foster trust in predictive analytics.

The Rwanda Revenue Authority's (RRA) application of neural networks for fraud detection illustrates the potential of machine learning to inform targeted policy interventions and recover lost revenue (Murorunkwere, 2022). The research-driven approach ensured that model outputs were actionable and aligned with policy objectives. Nevertheless, the deployment of advanced models such as Deep Convolutional Neural Networks (DCNNs) requires significant data volumes and technical expertise, which may present sustainability challenges. For NamRA, a pragmatic approach would involve piloting simpler machine learning models and gradually building internal capacity. Integrating analytics outputs into operational dashboards and decision-making processes is essential to avoid the common pitfall of analytics projects remaining siloed from core business functions. Additionally, the adoption of XAI methods, such as feature importance analysis, can enhance the interpretability of model predictions and support evidence-based policymaking.

The SAT's adoption of cloud computing and advanced analytics has led to substantial improvements in processing efficiency and taxpayer convenience, notably reducing electronic invoice processing times (PricewaterhouseCoopers, 2018). However, the reliance on cloud infrastructure introduces data privacy and security considerations that must be carefully managed. For NamRA, the lesson is to pursue automation incrementally, prioritising high-impact, low-complexity processes such as pre-filling forms for common taxpayer segments. Before large-scale

cloud adoption, NamRA should assess regulatory, privacy, and security implications. Even basic analytics, such as identifying frequent errors in tax returns, can meaningfully improve taxpayer experience and compliance.

Research on Indonesia's tax administration highlights the transformative potential of AI for enforcement and compliance, but also underscores significant barriers, including regulatory ambiguity and a shortage of skilled personnel (Saragih et al., 2023). This case demonstrates that technological advancement must be matched by supportive policy frameworks and human capital development. NamRA should ensure that any AI or ML initiatives are accompanied by clear policies, legal frameworks, and investments in staff training. Attempting to implement advanced analytics without these prerequisites risks project failure and resource wastage.

The Australian Taxation Office (ATO) uses real-time analytics and behavioural nudges, such as pop-up alerts in the myTax system, which have proven effective in improving compliance and generating additional revenue (ATO, 2024). These interventions are relatively low-cost and scalable, relying on integrated data and taxpayer digital literacy. However, their effectiveness may be limited in environments with lower digital adoption or fragmented data systems. For NamRA, adopting behavioural insights such as personalised reminders or peer comparisons could enhance compliance, if data integration and digital engagement are sufficiently developed.to review their figures before submission.

During tax time in 2020, nearly 340,000 taxpayers (around 7.5% of myTax users) received a pop-up message through myTax suggesting they review a specific label. These prompts caused taxpayers to adjust their estimated tax payments to have a revenue impact of around $37 million.

According to the study by Kamara, (2021), tax administration is central to a country's development, and therefore, tax compliance, tax structure, and taxpayer services are key to an effective tax administration. He further continues by concluding that tax compliance is the most critical tax administration strategy that has a huge and significant impact on tax revenue collection. In relation to taxpayer services and revenue collection, he concluded that the emergence of technology and electronic taxation systems enhances effective communication between taxpayers and the revenue authority, hence increasing tax awareness, which ultimately has a positive impact on tax collection. This study focused on the effects of tax revenue collection in the Kenya Revenue Authority, tax

evasion, enforcement mechanisms, and financial penalties. The research relied on data collected through surveys using questionnaires.

Jørgensen, (2021) evaluated the idea of becoming a data-driven organisation, the reconfiguration of work situations in the Danish Customs and Tax Administration. In this study, he looks at front-line workers as one of the major impacts of the introduction of automation. He argues that, as much as machine learning algorithms automate their work, they still play a crucial role at the margins of infrastructure. By being available on email, telephone, and social media platforms, they ensure that taxpayers can deliver data to the self-help platforms of the tax administration.

2.5 Challenges and lessons learned by the Kenya Revenue Authority

The Kenya Revenue Authority (KRA) experience underscores the centrality of data quality and organisational culture in the transition to data-driven administration. Persistent data quality issues and resistance to change have hindered analytics initiatives, while leadership commitment and staff engagement have been identified as critical success factors. For NamRA, early investment in data cleaning, integration, and governance is essential. Leadership should actively champion data-driven decision-making and support staff through the transition. Establishing clear ethical guidelines for data privacy and AI use will further support sustainable analytics adoption.

2.6 Public sector application of ML

Beyond tax administration, ML has been applied in various public sector domains, including healthcare, education, and social services. In tax contexts, ML has been used to:

- Predict audit outcomes (OECD Independent External Evaluation Final Report, n.d.).
- Identify refund fraud (Achakzai & Juan, 2022).
- Segment taxpayers for targeted interventions (Del Carmen et al., 2022).

These applications highlight the growing relevance of ML in enhancing operational efficiency and policy effectiveness.

2.7 Digital Tax Administration Transformation

Digital Tax Administration Transformation may save a lot of money on compliance and administrative costs by leveraging technology. It makes it possible to collect taxes more effectively, enhances taxpayer services and transparency, and makes managing massive amounts of data easier (Stern et al., 2022). To identify and mitigate challenges of tax administration, this research explored the ways in which tax administrations could use technology to understand taxpayer behaviour and improve compliance. By using ML analytics, the tax administration can turn a vast amount of raw data into actionable insights for evidence-based decision-making.

Over the past decade, the role of tax administration has undergone a substantial transformation, with the pace of change accelerating in recent years. This evolution is largely driven by the emergence and adoption of new technologies, which have been integrated into various processes and functions within tax administration. Beyond digitisation, tax authorities are actively exploring innovative technological solutions to modernise their operations (Dina, 2021). This proactive approach enables them to seize new opportunities, enhance operational efficiency, expand their capabilities, and foster greater transparency and accountability, ultimately leading to increased revenue generation. The Organisation for Economic Co-operation and Development (OECD) Forum on Tax Administration has described the digital transformation path for tax administrations. This conceptualisation outlines the transformation process' beginning, middle, and aspirational endpoints. Below are the three phases:

i. **Tax Administration 1.0**
ii. **Tax Administration 2.0**
iii. **Tax Administration 3.0**

2.8 Research gap

While there is a growing body of work on ML in tax administration, few studies focus specifically on individual salaried taxpayers in developing economies. Most existing research is concentrated on corporate tax or VAT compliance, and often in high-income countries with mature data infrastructures. This study addressed that gap by applying ML to Namibia's ITAS data, offering insights into taxpayer behaviour in a developing context with unique data challenges. Here is why it is a game-changer:

2.7.1 Improved Compliance Monitoring:

Predictive models can identify patterns and behaviours associated with late or non-compliant taxpayers, enabling targeted interventions.

2.7.2 Resource Optimisation:

Instead of a one-size-fits-all approach, predictive analytics allow revenue authorities to allocate resources more effectively, focusing on high-risk cases and reducing unnecessary audits.

2.7.3 Fraud Detection:

Advanced machine learning models can flag anomalies in taxpayer behaviour or submissions that may indicate fraud, improve detection rates, and deter fraudulent activities.

2.7.4 Enhanced Decision-Making:

By analysing historical data, revenue authorities can forecast revenue collection trends and make more informed policy decisions to close gaps or improve processes.

2.7.5 Personalised Taxpayer Engagement:

Predictive insights can inform tailored communication strategies, ensuring taxpayers receive relevant guidance and support, improving their overall experience.

A predictive approach helps revenue authorities remain agile in addressing external changes (e.g., economic shifts, pandemics) by identifying emerging trends and planning responses effectively. Data governance encompasses the people, processes, and technologies required to effectively manage and protect an organisation's data assets. Its goal is to ensure that corporate data is understandable, accurate, complete, trustworthy, secure, and easily discoverable, thereby supporting informed decision-making and regulatory compliance (Business Application Research Centre, 2018). The governance framework focuses on the entire life cycle of the data, including its acquisition, utilisation, storage, and data ethics. The Data Governance Framework is designed to guide the effective management of data through clearly defined rules, regulations, and controls. It ensures the security, accountability, and integrity of data as it flows from internal systems, external platforms, and third-party sources, and as it is utilised for business operations and decision-making. By establishing a structured approach to data handling, the framework supports consistent, reliable, and compliant data practices across the organisation. In addition, a data-driven governance framework for revenue authorities includes improved revenue collection, increased compliance, and reduced tax fraud and evasion. A data-driven framework can help revenue authorities to make informed decisions, identify patterns and trends, and take proactive measures to ensure compliance.

By outlining the theoretical underpinnings and going over pertinent topics, this chapter establishes the study's background. The framework for the investigation is laid forth in this chapter. The chapter also describes how the variables and important ideas examined in the study relate to one another. To accomplish the research goals, the research methodology is finally summarised. The methodology will go over the data collection technique, data pre-processing, model selection, training, and assessment of the suggested models.

Recent advancements in ML have significantly transformed tax administration, particularly in enhancing taxpayer compliance and risk profiling. Contemporary research emphasises the use of supervised and unsupervised learning algorithms such as decision trees, random forests, and gradient boosting to predict non-compliant behaviour and optimise audit detection (Alm & Soled, 2016). These models leverage historical tax data, third-party information, and behavioural indicators to identify high-risk taxpayers with greater accuracy than traditional rule-based systems.

Moreover, the integration of explainable AI (XAI) techniques, including SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), has improved the interpretability of ML models, enabling tax authorities to justify automated decisions and enhance transparency (Lundberg & Lee, 2017). In the African context, countries such as South Africa and Kenya have begun piloting ML-driven tools for fraud detection, taxpayer segmentation, and service delivery optimisation, offering valuable insights for emerging economies like Namibia.

The money collected is used to support the government initiative, such as building roads, hospitals, schools, buying medicine and medical equipment, etc. For the 2022/2023 financial year, NamRA collected N$57 billion, surpassing the revised target of N$ 53,4 billion (Namibia Revenue Agency, 2023). This money is directly deposited into the state coffers to support government developmental projects. The timely and accurate filing of returns by individual taxpayers results in more money collected, which in turn will circulate into the economy through tax refunds.

2.9 Data-driven tax administration

Analytics and Machine Learning are the new game changers in today's tax administration as they help forecast and assist with evidence-based decision-making. According to the Intra-European Organisation of Tax Administration, they define advanced analytics as the practice of using statistical techniques to make predictions and draw inferences about cause and effect. (OECD, 2016a). Analytics has helped many organisations make real-time, informed decisions. In tax administration, analytics are used in areas such as Audit Case Selection, Debt Management, Fraud Detection, etc. The South African Revenue Service's (SARS) criminal and illicit economic activity intervention and risk-management programs used AI and data analytics to recover R89.3bn from 26 million taxpayers in the 2022-23 financial year, according to its commissioner, Edward Kieswetter(OECD, 2016). The Tax Administration has been utilising emerging technologies and deriving insights from data for its tax collection efforts.

The experiences of international tax authorities demonstrate that while advanced analytics and AI offer significant potential for improving tax administration, their success is contingent upon foundational investments in data quality, governance, and human capital. NamRA should adopt a phased approach, beginning with the establishment of robust data governance frameworks and the implementation of basic analytics to address immediate operational challenges. The integration of explainable AI tools, such as SHAP or LIME, will be critical for ensuring transparency and building stakeholder trust in predictive models. Ultimately, the most effective strategies will be those that are tailored to Namibia's specific context, resource constraints, and organisational maturity.

3. Chapter Three: Research Methodology

3. Introduction

3.1 Methodology

This chapter presents the research methodology employed to develop predictive models for taxpayer compliance behaviour using machine learning techniques. The approach was organised into two main phases: data preparation and model development with validation. The study was rooted in data-driven methodologies, utilising the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework. This approach was customised to align with NamRA's operational context while incorporating international best practices in tax analytics. The data preparation phase addressed key challenges inherent in tax data, including data cleaning, feature engineering of compliance-related variables, and the handling of class imbalances. The modelling phase involved a comparative assessment of supervised learning algorithms, with emphasis on both predictive accuracy and interpretability within a tax administration context.

3.2 Research Philosophy

This study was guided by a positivist research philosophy. Positivism is rooted in the ontological belief that reality is objective, measurable, and limited to observable phenomena (Shannon-Baker, 2022). This approach was appropriate for the current research, as it relied on factual, secondary data without making unverifiable assumptions about individual tax return submission timelines within the ITAS system.

3.3 Research Paradigm

The study adopted a data-driven research paradigm, employing an exploratory approach to analyse and extract insights from taxpayer data using various machine learning and analytical techniques. Additionally, the research was informed by post-positivist principles, which acknowledge that complete objectivity is unattainable. Rather than seeking absolute truth, post-positivism aims to present the most accurate approximation of reality based on empirical evidence (Maksimović & Evtimov, 2023).

3.4 Research Strategy

The study strategy was hybrid, using analytical and predictive research models. Since the data was already available from NamRA (ITAS database), an analytical method was applied to conduct this study. The dataset was analysed using different Machine Learning models using Python, and models were created for prediction.

3.5 Research horizon

The research utilised a cross-sectional methodology, which entailed concurrently assessing the study taxpayer's exposure and outcome. This methodology was applied to the study because taxpayers' data was selected based on predetermined inclusion and exclusion criteria, containing taxpayer records from 2019 to 2023.

3.6 Research design

The study employed an inductive research design, which facilitated the development of theories or generalisations derived from specific observations and empirical data (John, 2024). The dataset was extracted from the ITAS database, and subsequent preprocessing and exploratory analysis were conducted using Python. This design enabled effective visualisation and interpretation of secondary data, allowing for hypothesis testing and informed conclusions.

This section also outlines the methodological flow diagram used in the study, detailing the transformation of pre-processed data and the predictive modelling process using various machine learning algorithms. The flow diagram, presented in Figure 8, illustrates the sequence of steps, including data collection, cleaning, and segmentation, as previously discussed.

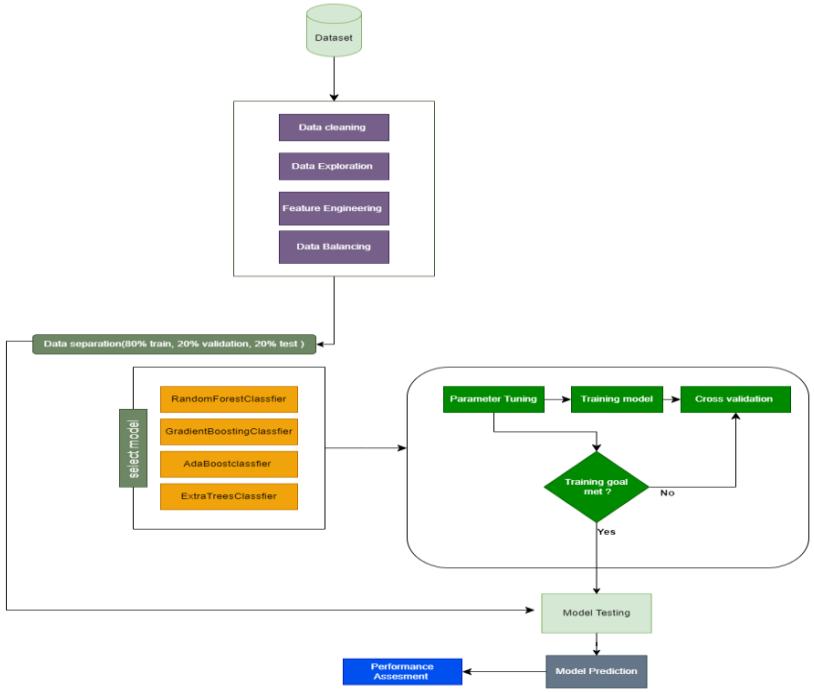

Figure 3-1 Research methodology adopted

Summary of the research methodology

This chapter outlined the research methodology used in the study, covering philosophical foundations, strategies, data preparation, and analytical techniques. It provided a robust framework for analysing taxpayer behaviour and predicting submission patterns while ensuring data integrity and ethical compliance. See the table below:

Table 3-1 Research tools used in the thesis

Research questions	**Tools**	**Outcome**
What is the present filing pattern of individual taxpayers at NamRA?	Python Random Forest Classifier Extra Trees Classifier Decision Tree Classifier Bagging classifier	The present filing pattern of individual taxpayers at NamRA indicates a significant portion of taxpayers are submitting their returns on time, while a notable number are still filing late**.**
How is the submission pattern of individual taxpayers? Do they submit late or on time?	Python	**Late Submissions:** 528,645 **On-Time Submissions:** 758,069
How is the accuracy of ML models in forecasting taxpayer filing patterns?	Python Random Forest Classifier Extra Trees Classifier Decision Tree Classifier Bagging classifier	Bagging classifier performed 3rd with the accuracy of 73%, with a precision of 78%.
Which ML model performs best/interpretable predictions?	Radom Forest	Random forest classifier performed best with an accuracy of 83%, with a precision of 83%
How can NamRA leverage emerging technologies such as ML and AI to enhance understanding of taxpayer behaviour (compliance) and data-driven decision making?	AI and ML Forecasting Clustering Prediction (Classification)	NamRA can leverage emerging technologies like Machine Learning (ML) and Artificial Intelligence (AI) to enhance understanding of taxpayer behaviour and improve data-driven decision-making.

3.7 Ethics and Confidentiality

The study took ethics and integrity as paramount. Before data collection, ethical clearance was obtained from Namibia University of Science and Technology. The clearance certificate was submitted to the Commissioner of NamRA, and the purpose of the research was explained. The data set provided by NamRA was treated with utmost confidentiality, and the study ensured complete anonymisation in all fields when it came to PII. Besides NUST, the research will only be shared with NamRA, which owns the data.

3.8 Data Collection

For the purposes of data visualisation and insight exploration, the researcher examined a sample of data that was extracted/collected from the ITAS database and supplied in CSV file format. The data used in this study was from the financial years of 2016 to 2023. Features related to taxpayer return submissions, including gross amount, tax year, taxpayer age, and submission status (on-time or late), were included in the databases. The information covered the periods prior to, during, and following the COVID-19 epidemic.

The process of obtaining, organising, and cleansing the raw data is known as data preparation. The datasets that include the individual salaried records were extracted from a database into an Excel sheet (CSV). One of the most crucial and difficult aspects is data (data cleaning). The dataset was extracted and pre-processed by cleaning and transforming it into a structured and analysable format, enabling the identification of patterns for future predictive analysis. Datasets, exploratory data analysis, and data segmentation are the subsections under "Data preparation" in this section. As part of feature engineering, a new feature called tax period was created based on the tax year. These features labelled all submissions into three periods of: Before COVID (tax years of 2016–18), During COVID (tax years of 2019–21) and After COVID (tax years of 2022–24). The goal of creating tax periods was to observe the impact the COVID pandemic had on taxpayer submissions. Based on the goal of reducing unneeded data, the researcher kept rows only where the tax year was greater than or equal to 16 (i.e., tax year of 2016 onward, representing the potential effect of COVID for this sample).

3.9 Data preparation

The main component of the data preparation process done on this dataset included the following:

3.9.1 Data Loading and Concatenation

All datasets from 2019 to 2022 from the ITAS database were loaded into memory and concatenated into a single atomic dataset. This step combined taxpayer submissions from all years for the purpose of analysis. The final combined dataset had 274,549 rows and 8 columns after filtering was performed to exclude unneeded data. The figure below shows the total descriptive statistics of the dataset:

Out[8]:

	taxpayer_age	number_of_days_late_submsn	taxyear	gross_amount
count	292684.000000	292686.000000	292686.000000	2.813990e+05
mean	44.386834	-644.473637	2021.256514	2.047036e+05
std	11.527418	1186.101273	3.268806	1.171453e+06
min	2.000000	-13289.000000	1987.000000	-2.586602e+04
25%	35.000000	-736.000000	2021.000000	5.376999e+04
50%	43.000000	-137.000000	2023.000000	1.390522e+05
75%	52.000000	-17.000000	2023.000000	2.856165e+05
max	126.000000	360.000000	2024.000000	6.059970e+08

Figure 3-2 Descriptive statistics of the dataset

3.9.2 Filtering for Tax Year

Based on the goal of reducing unneeded data, we kept rows only where the tax year was greater than or equal to 2016 (i.e., tax year 2016 onward, representing the potential effect of COVID for this sample). Furthermore, the researcher dropped the following attributes because they were of no value to the research: Taxpayer name, return receipt date, return due date, return submission year, and other columns. All remaining columns were analysed based on taxpayer age, marital status, and job status.

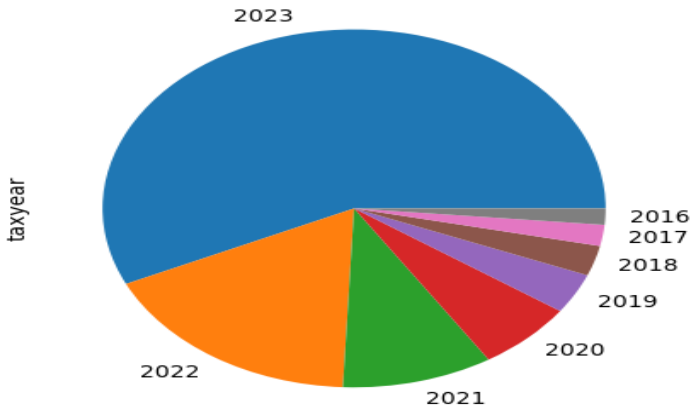

Figure 3-3 Distribution of tax years from 2016 to 2022

3.9.3 Feature Engineering

Following the selection of relevant variables, a new feature tax period, was engineered based on the tax year. This feature categorised all return submissions into three distinct periods: Before COVID (2016–2018), During COVID (2019–2021), and After COVID (2022–2024). The purpose of this transformation was to assess how the timing and impact of the COVID-19 pandemic influenced taxpayer filing behaviour. Once the tax period feature was created, the original tax year variable was removed from the dataset, as it had become redundant.

3.9.4 Handling of Missing Data

Missing data was common within the sample, and some kind of handling of missing values was important to explore how the COVID pandemic may have influenced the dataset. The handling of missing data was decided based on the categorical or numerical nature of each column in the dataset by using the median.

3.9.5 Numerical Columns

Filling missing data for numerical columns, such as gross amount and taxpayer age, was completed using the median value in the respective columns, to limit the potential impact of skewness in the data.

3.9.6 Data Preprocessing

After the dataset was cleaned by filling in missing values and dropping features that did not add value to the research, further processing was performed on columns in the dataset by:

i. Encoding Categorical Variables

The target variable such as the status on return, was an ordinal categorical variable with values like on time or late, requiring label encoding. Other categorical columns were one-hot encoded, wherever necessary, so that they could be used with machine learning algorithms.

ii. Recasting Some Columns

Several columns were mistakenly coded as type objects. For example, gross amount and taxpayer age columns were recognised as those with a float type.

iii. Standardising Numerical Features

The numerical attributes were standardised with StandardScaler to transform the data into a consistent form: a mean of 0 and a standard deviation of 1. This was necessary to prepare all the numerical predictors to perform at the same scale and produce optimal results.

3.9.7 Exploratory Data Analysis

This section highlights the importance of data preprocessing in machine learning-based prediction tasks. In the given dataset, the return year played a critical role in identifying whether observations were typical or anomalous. Data preprocessing is among the most essential and complex stages of any predictive modelling project, as it ensures the dataset is clean, consistent, and suitable for analysis. Key steps include assessing data quality, handling missing values and duplicates, and performing necessary transformations. Although the dataset was normalised, further inspection revealed that certain numerical variables contained missing values. These were addressed by imputing the median or mean, depending on the skewness of each column, to preserve data integrity and improve model accuracy.

3.9.8 Counting Entries by Return Status (target variable)

This analysis categorised the dataset based on the target variable, *status-on-return*, as illustrated in Figure 3-6, to count the number of records within each category. This facilitated a clearer understanding of how return submissions were distributed across different compliance statuses.

```
status_on_the_return     gender  marital_status
Submissions is late      Female  Divorced                          158
                                 Married in Community of Property  3157
                                 Married out Community of Property  839
                                 Others                           86863
                                 Single                           12480
                                 Widow/Widower                       76
                         Male    Divorced                           46
                                 Married in Community of Property  1726
                                 Married out Community of Property  239
                                 Others                           99518
                                 Single                           11688
                                 Widow/Widower                       21
Submissions is on time   Female  Divorced                          109
                                 Married in Community of Property  1301
                                 Married out Community of Property  620
                                 Others                           20938
                                 Single                            4014
                                 Widow/Widower                       37
                         Male    Divorced                           16
                                 Married in Community of Property   871
                                 Married out Community of Property  236
                                 Others                           25082
                                 Single                            4505
                                 Widow/Widower                        9
dtype: int64
```

Figure 3-4 Return status on target variables

Assessing data imbalance

Since the target variable (status-on-return) was imbalanced as shown in the in Figure 3-7 (more submissions were "late" than "on time"), random oversampling was employed to balance the sample size associated with the minority class, "late", to minimise the potential of bias when modelling as illustrated in figure 3-8.

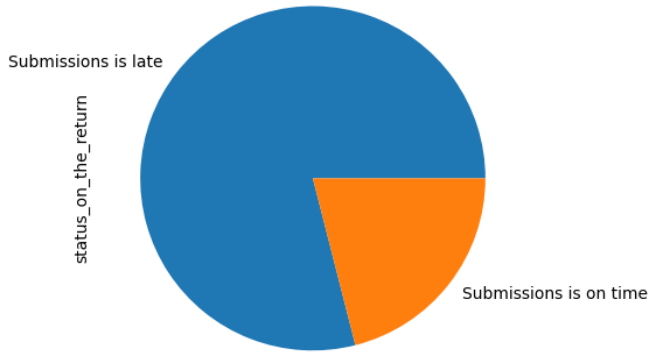

Figure 3-5 Dataset results before balancing the dataset

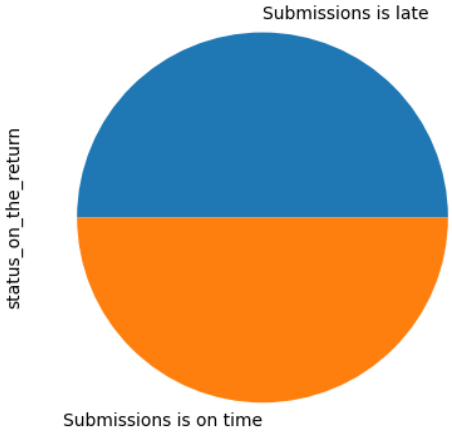

Figure 3-6 Balanced dataset after the random oversampling technique

3.9.9 Summary of the research methodology

This chapter outlined the research methodology used in the study, covering philosophical foundations, strategies, data preparation, and analytical techniques. It provided a robust framework for analysing taxpayer behaviour and predicting submission patterns while ensuring data integrity and ethical compliance. See table 3-1 below:

Table 3-1 Research tools used in the thesis

Research questions	**Tools**	**Outcome**
What is the present filing pattern of individual taxpayers at NamRA?	Python Random Forest Classifier Extra Trees Classifier Decision Tree Classifier Bagging classifier	Individual taxpayers at NamRA indicate a significant portion of taxpayers are submitting their returns on time, while a notable number are still filing late**.**
How is the submission pattern of individual taxpayers? do they submit late or on time?	Python	**Late Submissions:** 528,645 **On-Time Submissions:** 758,069
How is the accuracy of ML models in forecasting taxpayer filing patterns?	Python Random Forest Classifier Extra Trees Classifier Decision Tree Classifier Bagging classifier	Bagging classifier performed 3rd with the accuracy of 67%, with a precision of 69%
Which ML model performs best/interpretable predictions	Radom Forest	Random forest classifier performed best with an accuracy of 68.2%, with a precision of 68%
How can NamRA leverage emerging technologies such as ML and AI to enhance understanding of taxpayer behaviour (compliance) and data-driven decision making?	AI and ML Forecasting Clustering Prediction (Classification)	NamRA can leverage emerging technologies like ML and AI to enhance understanding of taxpayer behaviour and improve evidence-based decision-making.

4. Chapter Four: Implementation and Results

4. Introduction

This chapter presents the results of evaluating the performance of ten machine learning models tested on the same dataset, using an 80/20 train-test split. The primary objective was to identify the model that delivers the most accurate and reliable predictions for taxpayer return submissions. Model performance was assessed using standard classification metrics, accuracy, precision, recall, and F1-score, to quantify predictive effectiveness and support model comparison.

4.1 Model Selection

The study analysed individual tax return submissions using data extracted from the ITAS database, covering the period from the 2016 to the 2023 tax year. Python was used to perform the analysis and interpret the results. Several machine learning models were trained to identify the most suitable classifier for predicting taxpayer return submission behaviour. These included the Random Forest Classifier, Gradient Boosting Classifier, Decision Tree Classifier, K-Nearest Neighbours Classifier, Support Vector Machine (SVM), and Logistic Regression Classifier.

These models were selected based on their effectiveness in classification tasks and their ability to handle both categorical and numerical data. Random Forest was chosen for its robustness to noisy and incomplete data and its ability to highlight feature importance, which is valuable for understanding compliance drivers. Gradient Boosting was selected for its high predictive accuracy and its strength in handling imbalanced datasets, which is common in compliance modelling. SVM was included due to its effectiveness in high-dimensional spaces and its suitability for binary classification problems. While models like neural networks were considered, they were excluded due to their complexity, lower interpretability, and higher data requirements, which are less ideal in a public sector context where transparency and explainability are essential.

4.2 Objectives and goals of the models

The specific objective of the model was to conduct the data wrangling and training of a machine learning model to identify whether a taxpayer (individual salaried person) filed/submitted on time or late. A model should classify the submission status accurately, allowing for the identification of patterns and trends. The prediction task is a supervised learning task, specifically in the form of classification, and the target variable (status-on-return) is a binary variable (on time or late). The framework of assessing model performance consisted of metrics such as accuracy, precision, recall and F1 score.

4.3 Model Training and Validation

Data Partition: The dataset was divided into training (80%) and test sets (20%) to evaluate model performance on a new dataset.

Model Fitting: The previously mentioned models were trained on the pre-processed dataset, predicting each target (status-on-return) from the rest of the features used as input variables.

Model Validation: After training the respective models, the researcher adopted measures of accuracy, precision, recall, F1 score and ROC-AUC score to quantify how well the models predicted taxpayer submissions.

Model Cross-Validation: To ensure that overfitting would not occur, the researcher implemented cross-validation that involved taking training set and dividing it into a smaller number of subsets, training the model on each subset of data iteratively.

4.4 Results

Eight variables were explored using Python to analyse and interpret results, namely: gender, taxpayer age, marital status, status-on-the-return, number-of-days-late-submission, taxpayer-registration-office, tax year and gross amount.

After determining the top-performing models, hyperparameter tuning was initiated to maximise classification performance. The process entails:

i. **Initial Model Selection:**

The study employed a uniform dataset (individual tax filings) across all models to enable direct performance comparisons. SVM, Random Forest, Decision Tree, and Gradient Boosting classifiers

were prioritised due to their established robustness in supervised classification, yielding accuracies of 62–67% with optimal bias-variance equilibrium.

Out[44]:

	Model	Accuracy	Precision	Recall	F1 Score	CV Score (mean)
4	SVM	0.679647	0.684353	0.679647	0.677528	0.685590
1	Decision Tree	0.678526	0.682688	0.678526	0.676626	0.682563
2	Random Forest	0.677928	0.681975	0.677928	0.676069	0.683198
3	Gradient Boosting	0.679273	0.693009	0.679273	0.673364	0.681666
0	Logistic Regression	0.674490	0.691032	0.674490	0.667173	0.677181
5	K-Nearest Neighbors	0.625906	0.625906	0.625906	0.625904	0.638985

Figure 4-1 Models' performance on predicting submission patterns

An additional methodological consideration involved the target variable (status-on-return). As illustrated in Figure 4.2, the initial dataset exhibited significant class imbalance. Following the balancing procedures, Figure 4.3 demonstrates the equitable distribution achieved across target classes, which was essential for robust model training.

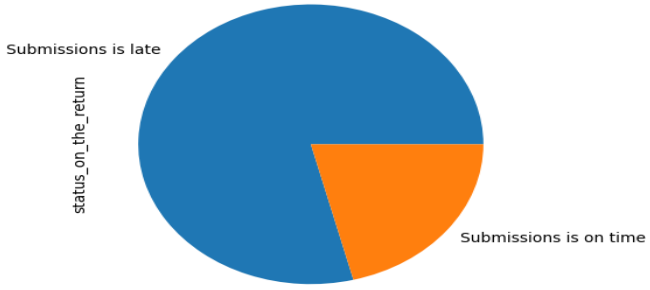

Figure 4-2 Unbalanced dataset

To balance the dataset, the class RandomOverSampler was used to generate new samples for the minority class to balance the dataset. The instance created was used to resample the data, and the fit resample method was applied to oversample the dataset; it took the x feature and the target

variable y as inputs. The method returned resampled feature x and target labels y, to ensure classes in y were balanced as per Figure 4-3.

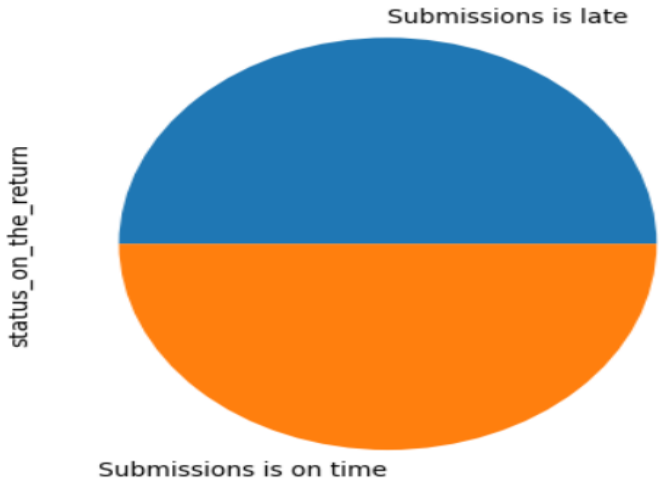

Figure 4-3 balanced data

Analysis of the submission pattern showed the submission behaviour of individuals based on the number of days they submitted their returns. A total of 120949 (approximately 78.3%) taxpayers submitted their returns late, while 33446 (21.7%) taxpayers filed on time, as indicated in Figure 4-4 below. These insights can help NamRA understand the compliance behaviour of taxpayers and identify areas where interventions might be needed to improve on-time submission rates.

```
In [103]: df["status_on_the_return"].value_counts()

Out[103]: Submissions is late       120949
          Submissions is on time     33446
          Name: status_on_the_return, dtype: int64
```

Figure 4-4 Prediction of submission pattern behaviour

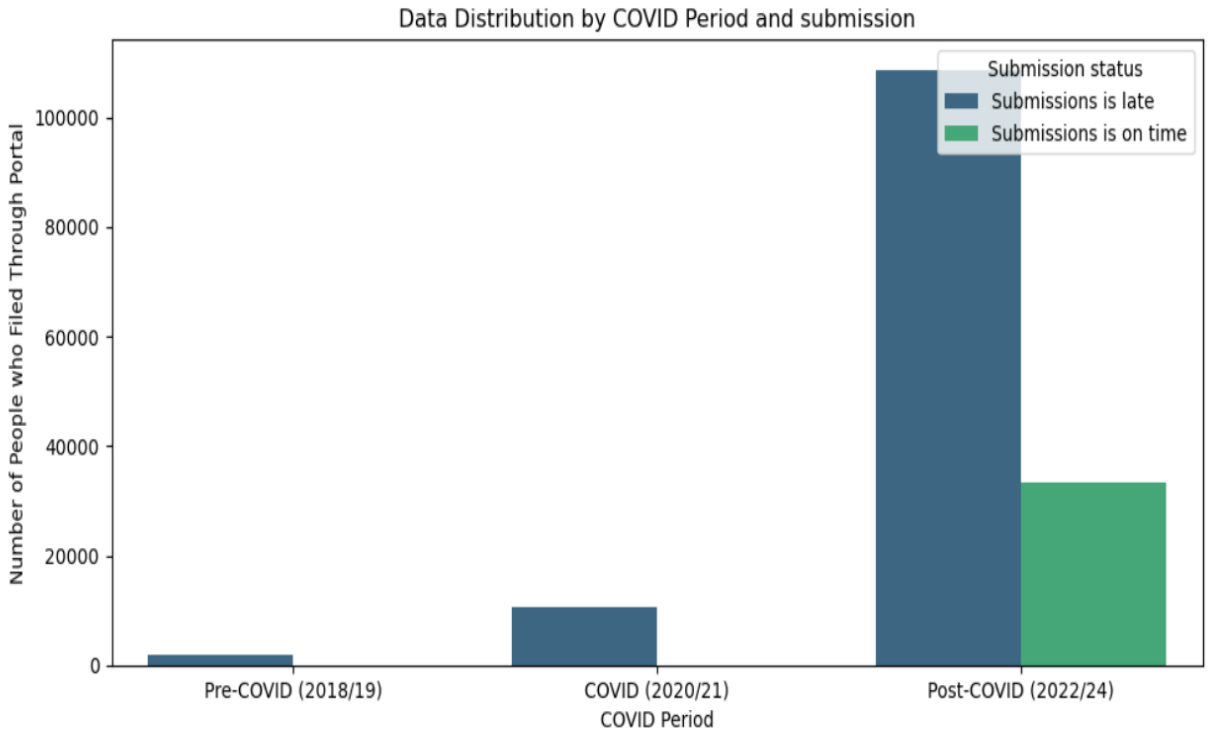

Figure 4-5 pre- and post-COVID analysis of return submission patterns

Figure 4-5 illustrates the number of people who filed through a portal during three distinct periods: pre-COVID (2018/19), COVID (2020/21), and post-COVID (2022/24). Each period is divided into two categories based on submission status: "Submissions are late" (blue bars) and "Submissions are on time" (green bars).

Here is a breakdown of the data:

- **Pre-COVID (2018/19)**:
    - Late submissions: approximately 5,000
    - On-time submissions: approximately 20,000
- **COVID (2020/21)**:
    - Late submissions: approximately 80,000
    - On-time submissions: approximately 10,000

- **Post-COVID (2022/24)**:
    - Late submissions: approximately 90,000
    - On-time submissions: approximately 40,000

The chart shows a significant increase in late submissions during the COVID period, which continued into the post-COVID period, although there is also a notable rise in on-time submissions post-COVID.

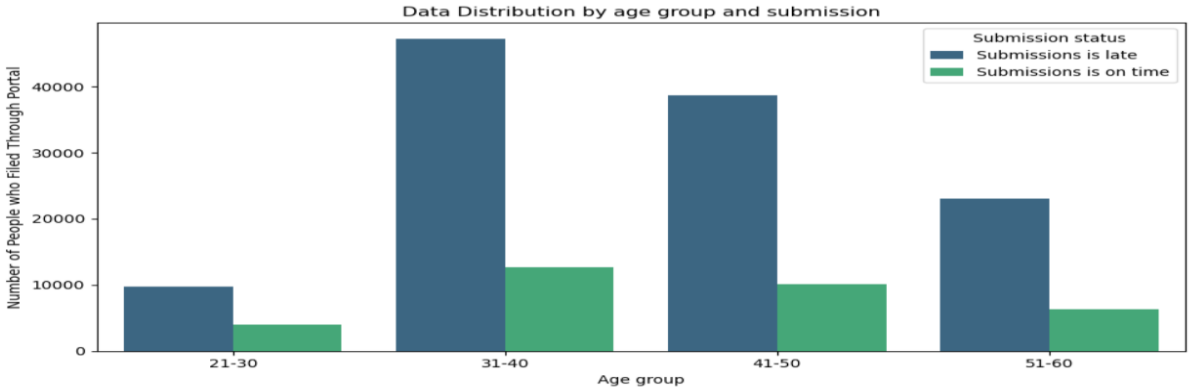

Figure 4-6 Return submission analysis by different age groups

Figure 4-6 shows the number of people who filed through a portal, categorised by age group and submission status. The x-axis represents different age groups: 21-30, 31-40, 41-50, and 51-60 years. The y-axis represents the number of people, ranging from 0 to 10,000.

- **Age 21-30**:
    - Late submissions: approximately 8,000
    - On-time submissions: approximately 3,000
- **Age 31-40**:
    - Late submissions: approximately 10,000
    - On-time submissions: approximately 4,000

- **Age 41-50**:
    - Late submissions: approximately 9,000
    - On-time submissions: approximately 3,000
- **Age 51-60**:
    - Late submissions: approximately 7,000
    - On-time submissions: approximately 2,000

The chart indicates that late submissions are consistently higher across all age groups, with the 31-40 years age group having the highest number of both late and on-time submissions.

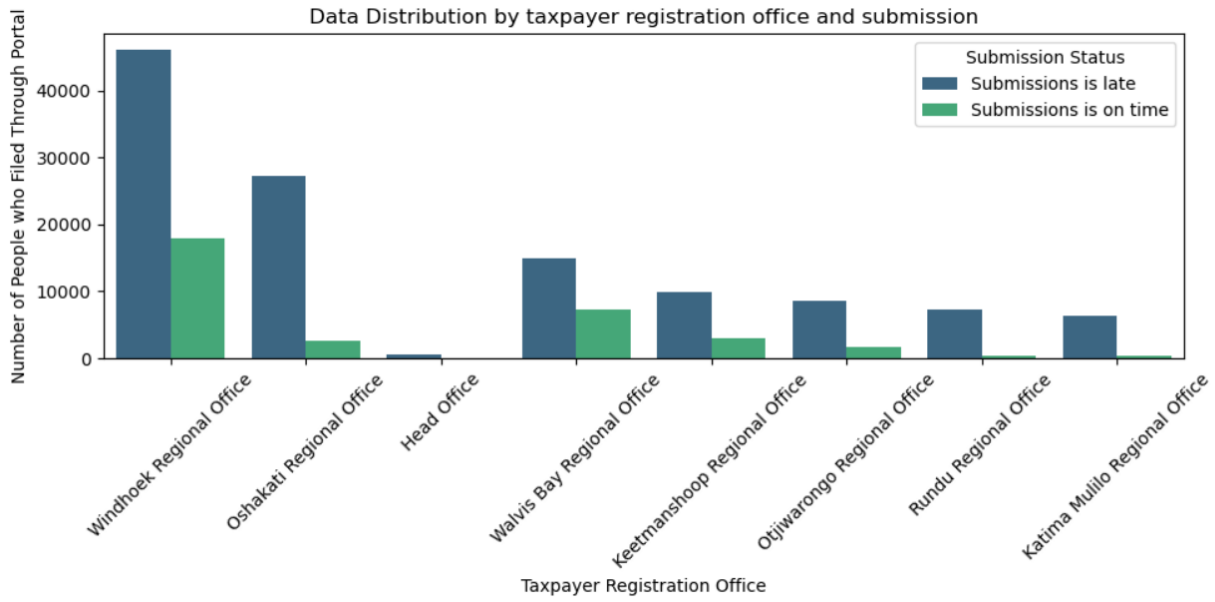

Figure 4-7 Number of online taxpayer filers by region (demographic)

Figure 4-7 illustrates the number of people who filed through various taxpayer registration offices, categorised by submission status (late or on-time). The number of people who filed through the portal is displayed on the y-axis, while the x-axis reflects the various taxpayer registration offices.

- **Windhoek Regional Office**:
    - Late submissions: approximately 9,000
    - On-time submissions: approximately 4,000

- **Oshakati Regional Office**:
    - Late submissions: approximately 7,000
    - On-time submissions: approximately 3,000
- **Head Office**:
    - Late submissions: approximately 8,000
    - On-time submissions: approximately 5,000
- **Walvis Bay Regional Office**:
    - Late submissions: approximately 6,000
    - On-time submissions: approximately 2,000
- **Keetmanshoop Regional Office**:
    - Late submissions: approximately 5,000
    - On-time submissions: approximately 1,000
- **Otjiwarongo Regional Office**:
    - Late submissions: approximately 4,000
    - On-time submissions: approximately 2,000
- **Rundu Regional Office**:
    - Late submissions: approximately 3,000
    - On-time submissions: approximately 1,000
- **Katima Mulilo Regional Office**:
    - Late submissions: approximately 2,000
    - On-time submissions: approximately 1,000

The chart indicates that late submissions are consistently higher across all offices, with the Windhoek Regional Office having the highest number of late and on-time submissions.

Figure 4-8 below visualises the correlation between the taxpayer age, tax year, and gross amount heatmap. The results show no strong correlations between the variables, with the highest correlation being a weak positive correlation between taxpayer age and gross amount.

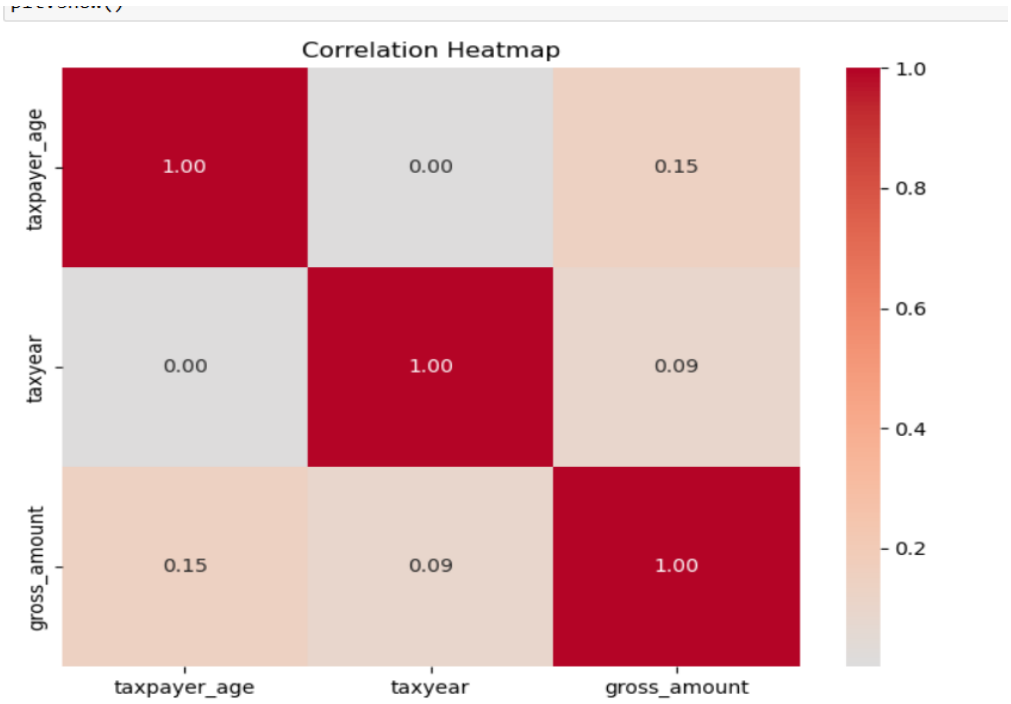

Figure 4-8 Correlation heat map

ii. **Hyperparameter Tuning**

Hyperparameter optimisation was performed for all models using 5-fold cross-validated Grid Search and Randomised Search (from the scikit-learn library v1.2) to maximise predictive accuracy. The search space for each algorithm was tailored to its structure: for instance, the Random Forest's space included n-estimators [50, 500] and max-depth [None, 50], while Gradient Boosting models focused on learning-rate [0.01, 0.2] alongside architectural parameters. This process significantly improved model performance, with final cross-validated accuracy scores ranging from 64% to 68%. The top-performing model was a tuned Random Forest classifier, which achieved the highest accuracy (68%). We therefore recommend the optimised Random Forest model for final deployment, as it demonstrated the efficacy of the hyperparameter tuning process.

Out[49]:

	Model	Test Accuracy	Test Precision	Test Recall	Test F1
3	Gradient Boosting	0.681142	0.688379	0.681142	0.677976
4	SVM	0.679647	0.684353	0.679647	0.677528
2	Random Forest	0.680395	0.686997	0.680395	0.677477
1	Decision Tree	0.678825	0.683079	0.678825	0.676889
0	Logistic Regression	0.672173	0.682883	0.672173	0.667205
5	K-Nearest Neighbors	0.647956	0.648190	0.647956	0.647797

Figure 4-9 Model performance after hyperparameter tuning

iii. Explainable AI (XAI)

The SHAP and LIME results provide critical insights into the determinants of taxpayer compliance, even in the context of a modestly performing model. The consistent importance of *income group* and *COVID-19 period* underscores the role of financial capacity and macroeconomic disruptions in shaping compliance. These findings are consistent with prior research linking income stability and external shocks to tax compliance behaviour (Alm & Torgler, 2011).

Administrative variables, particularly the *taxpayer-registration-office*, highlight structural differences that may be linked to variations in service delivery, enforcement practices, or taxpayer education. Similarly, the role of *marital status* suggests that demographic characteristics may indirectly influence compliance, perhaps through differences in household responsibilities or perceptions of tax obligations.

The dual application of SHAP and LIME strengthens the interpretability of the model. SHAP offers a global perspective on the most important features across the taxpayer population, whereas LIME provides individualised explanations as per Figure 4-10 and Figure 4-11 that can enhance communication with taxpayers and compliance officers. Together, these methods enhance trust in machine learning models and support actionable policy recommendations.

Although the predictive accuracy was limited, the interpretability benefits demonstrate the value of XAI in guiding compliance strategies. For instance, tax administrations could use these insights to design targeted outreach for high-income taxpayers or implement office-level interventions

where late submissions are more prevalent. Future work should focus on improving predictive accuracy by incorporating richer behavioural, transactional, and historical data, while also assessing fairness to ensure that specific demographic groups are not disproportionately flagged.

.

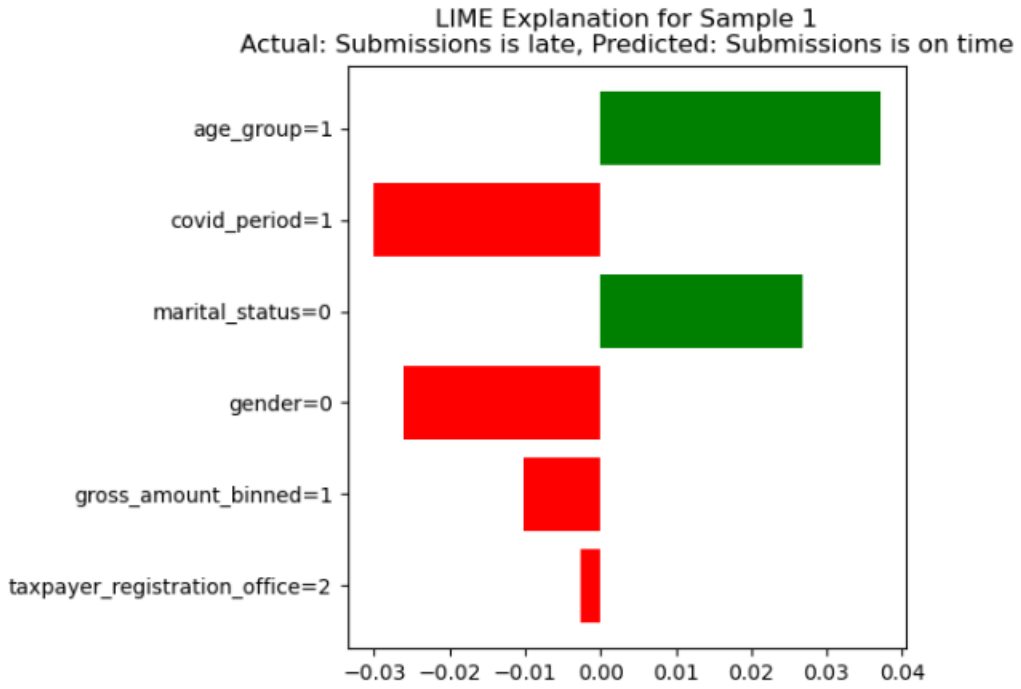

Figure 4-10 Individual model decisions (Lime)

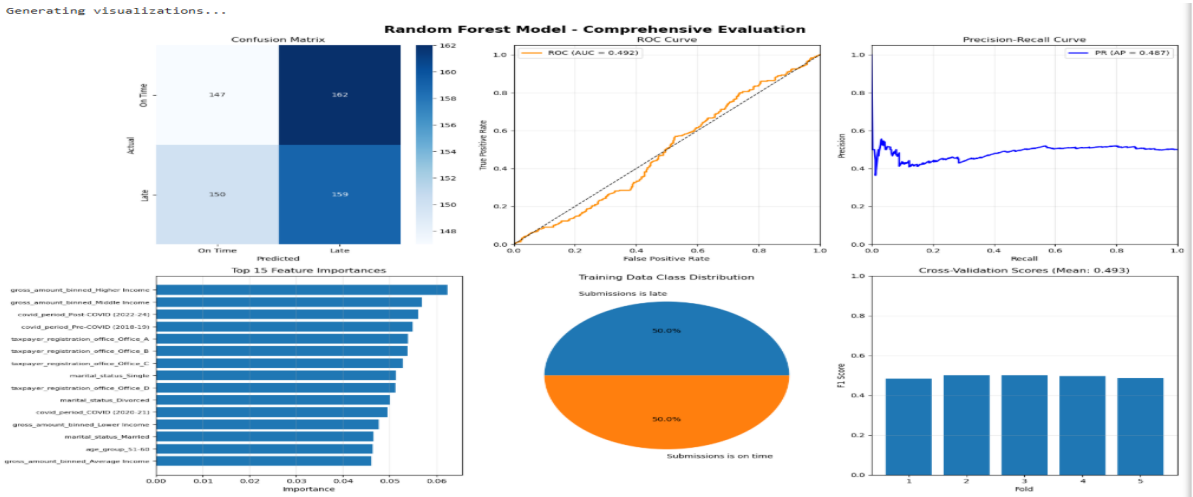

Figure 4-11 ROC AUC evaluation

iv. **Final Model Selection**

Comparative analysis revealed SVM, Gradient and Random Forest as the optimal algorithms (accuracy: 68.1% ±), outperforming alternative models during cross-validation. The tuned SVM, Gradient and Random Forest (F1-score: 0.67) development was subsequently validated against independent test data.

4.5 Conclusion

This chapter explored taxpayer compliance behaviour using machine learning and explainable AI. Late submissions were consistently higher across all registration offices, with Windhoek leading in both late and on-time filings. While correlation analysis showed weak relationships among age, tax year, and gross amount, machine learning models especially the tuned Random Forest achieved moderate accuracy (68%).

SHAP and LIME revealed that income group, COVID-19 period, and registration office were key drivers of compliance. These insights, though based on modest model performance, offer valuable guidance for targeted interventions and policy design. Future improvements should focus on richer data and fairness assessments to enhance predictive power and equity.

5. Chapter Five: Conclusion and Future Work

5. Introduction

This chapter provides an overview of the main conclusions and findings as well as an answer to the study's main objective and goals. The chapter emphasises the significance of the findings for the whole dataset by going into detail about the theoretical and practical implications of the findings. Moreover, future study subjects are supplied, emphasising areas of concern about further research. These results indicate the need for better research methods when dealing with open-ended questions. The last phrase of this chapter reiterates the importance of the results and how they might impact further study in the field of Data Science.

5.1 Assumptions

The hypothesis was that the NamRA data is complete regarding accuracy and quality. The study did not consider the patented data or real-time data sources that could have additional, deeper insights into tax-related matters.

5.2 Summary of findings

This study used historical data and machine learning algorithms based on historical filing patterns. The following ML models were utilised:

- Random Forest Classifier
- Gradient Boosting Classifier
- SVM
- Logistic Regression
- K-Nearest Neighbours
- Decision Trees

The ML models were used to explore insights into an individual salaried person's return. Among these, the SVM stood out with higher performance in metrics like accuracy, precision, recall, and F1-score after parameter tuning and cross-validation. The models were built by splitting historical data into 80% Training and 20% Testing. It is important to note that both pre-and post-COVID-2019 in both the training and testing datasets were covered. The dataset incorporated pre- and post-COVID-2019 occurrences to enrich the training data, which improved the model's predictive capabilities.

Key variables:

- **Status-on-the-return** indicates whether the return was on time or late.
- **Date submitted** was essential in assessing compliance with the annual deadline of June 30.

Based on the key variables above, the thesis adopted the Machine Learning models, SVM, Random Forest Classifier, Decision Tree Classifier, Logistic Regression, K-Nearest Neighbours, and Gradient Boosting.

The thesis was guided by the primary objective of conducting data wrangling and training a machine learning (ML) model to classify taxpayer return submissions as either "on time" or "late." Based on the performance evaluation of various ML models, the research demonstrated confidence in the effectiveness of data exploration through ML techniques. Out of the 6 models trained, the **Random Forest Classifier**, **Extra Trees Classifier**, **Decision Tree Classifier**, and **Bagging Classifier** emerged as the best-performing models.

In addition to the main objective, the thesis was supported by the following sub-objectives:

1. **To explore data analysis of the individual taxpayer returns to gain a better understanding of the dataset:**
   The thesis recommends the adoption of the three best-performing models for data insights and exploration. These models can help identify taxpayer behaviour patterns, enabling proactive measures and improved understanding of trends. This, in turn, can foster voluntary compliance. Individual salaried taxpayers represent the largest segment of tax filers. Simplifying the filing process and understanding their filing patterns can enhance voluntary compliance, increasing revenue collection. Higher revenue supports the national budget, addressing socioeconomic needs and stabilising economic growth.
2. **To identify patterns and trends in the submission behaviour of individual taxpayers:**

   Both past research and this study have demonstrated that ML and accurate data are powerful tools for driving reforms. These capabilities enable tax administrations, including NamRA, to achieve their strategic visions by leveraging analytics to generate actionable insights and identify trends.

3. **To enhance prediction accuracy through enhancement/feature engineering:**
   Techniques such as Principal Component Analysis (PCA), handling missing values, and encoding categorical variables were used.
4. **To assess the performance of ML models in predicting individual taxpayer submission patterns:**
   Different ML models were evaluated using Python to see how they performed on the dataset.
5. **To analyse and recommend modern ways that NamRA can leverage from using emerging technologies (ML/AI):**

   The different ML models were tested with the same dataset, and the Random Forest Classifier performed the superlative in this research, combining high accuracy with the ability to provide interpretable results through feature importance analysis. This made it an ideal choice for both predictive performance and generating actionable insights.

The research questions that guided the thesis are as follows:

a. **What is the present filing pattern of individual taxpayers at NamRA?**

Analysis of current filing patterns among individual taxpayers at NamRA revealed that while a portion of taxpayers submitted their returns on time, a substantial majority continued to file late. According to the available data, approximately 120,949 taxpayers submitted their returns late, compared to 33,446 who filed on time. This equates to 78.3% late submissions and 21.7% on-time submissions, respectively. These figures underscore the need for targeted interventions to enhance compliance and promote timely filing behaviour.

b. **How can ML models be used to achieve NamRA's objectives? Objectives such as improved voluntary compliance and improved data management and analytical capability can be attained using emerging technologies such as ML and AI at NamRA.**

ML models may help NamRA increase voluntary compliance by predicting taxpayer behaviour using supervised learning (e.g., XGBoost, Random Forest) to identify tax payers likely to delay submissions based on historical filing patterns, demographics, and economic activity. The models were able to predict with an F1 score of 68% after hyper-tuning.

c. **Which ML model achieves the highest accuracy/F1-score while maintaining interpretability?? In this thesis, ML were used to forecast taxpayer filing behaviour based on the submission status.**

The Random Forest Classifier exhibited the strongest performance in this study, achieving superior predictive accuracy in classifying tax payer return submissions as either "on time" or "late." While Random Forest models are not inherently interpretable, the use of feature importance analysis provides some transparency by highlighting which variables most significantly influence classification outcomes, such as previous filing history, business sector, or return complexity. By identifying these influential features, the model can offer policymakers evidence-based insights into the determinants of taxpayer compliance behaviour, supporting more informed decision-making beyond the model's predictive outputs. Furthermore, the other models evaluated in this research also demonstrated competitive accuracy and robust predictive capabilities, viz:

i. SVM
ii. Decision Tree Classifier
iii. Gradient Boosting.

d. **How can NamRA leverage emerging technologies such as ML and AI to enhance understanding of taxpayer behaviour (compliance) and data-driven decision making?** NamRA can leverage emerging technologies like Machine Learning (ML) and Artificial Intelligence (AI) to enhance understanding of taxpayer behaviour and improve data-driven decision-making in several ways:

Predictive Analytics for Compliance

- **Fraud Detection:** Unsupervised ML algorithms can analyse historical data to identify patterns indicative of fraudulent behaviour. By predicting potential non-compliance, NamRA can proactively address issues before they escalate.
- **Risk Scoring:** AI can assign risk scores to taxpayers based on their filing history and other relevant data, helping prioritise audits and investigations.

Enhanced Data Management

- **Data Integration:** AI can consolidate data from various sources, ensuring a comprehensive view of taxpayer information. This integration helps in making more informed decisions.
- **Data Quality Improvement:** ML algorithms can detect and correct errors in data, ensuring higher accuracy and reliability.

Advanced Analytics for Decision-Making

- **Scenario Analysis:** ML models can simulate various scenarios to predict the impact of policy changes, helping NamRA make data-driven decisions.
- **Trend Analysis:** AI can analyse trends in taxpayer behaviour, providing insights into compliance patterns and helping to develop targeted interventions.

NamRA can increase compliance rates, gain a better understanding of taxpayer behaviour, and make data-driven, well-informed decisions by utilising the above tools. Advanced technologies like machine learning, generative AI, and data analytics empower tax professionals to generate precise, real-time calculations, filings, and reports. These tools also enhance transparency, helping organisations to maintain compliance with tax regulations.

5.3 Limitations of the study

The prediction is limited to historical data from 2019; it does not cover the year before because, before that, there were no e-filing services, and the services were still under the Inland Revenue (Ministry of Finance). Another limitation was that we tried to run 30 machine learning models in just a few seconds to give us a grasp of how the models would perform on the dataset. However, we could not achieve our goal due to the computational power. The lazy predictor got stuck at 59% while training the models.

```
from lazypredict.Supervised import LazyClassifier


clf = LazyClassifier(verbose=0,ignore_warnings=True, custom_metric=None)
models,predictions = clf.fit(x_train, x_val, y_train, y_val )

print(models) |
```

```
59%|                    | 17/29 [09:07<03:32, 17.70s/it]
```

Figure 5-1: Lazy Predict

5.4 Contribution

Based on the research aim and objective of this thesis, the thesis contributes immensely to the body of knowledge through:

- **Improvement in tax administration practices:** By applying machine learning to predict whether taxpayer return submissions are on time or late, the research introduced innovative methods for automating and enhancing tax compliance monitoring. This can contribute to more efficient management of tax systems by providing insights into submission patterns, potentially leading to improved decision-making and resource allocation.
- **Theoretical Contribution to Predictive Modelling:** The study advanced academic knowledge by demonstrating the applicability of various machine learning classifiers, such as Random Forest, Gradient Boosting, SVM, and Logistic Regression, to behavioural prediction in a public sector context. It contributed to the literature by comparing model performance on real-world administrative data, highlighting the trade-offs between interpretability, accuracy, and data quality. Furthermore, the study provided empirical evidence on how predictive modelling can be used to forecast compliance behaviour in environments with limited labelled data, high dimensionality, and imbalanced classes. These insights can inform future research in public policy analytics, behavioural modelling, and applied machine learning.
- **Identification of patterns and trends:** Through identifying patterns and trends in the submission status, the research added value to the broader field of data analytics in tax systems. This can help policymakers and tax authorities better understand the factors contributing to late submissions and design targeted interventions.
- **Exploration of Predictive Modelling in Tax Systems:** The study's focus on using supervised learning classification models to assess the timeliness of submissions filled a gap in the application of machine learning within the context of tax administration. The contribution lies in offering a practical framework for predicting taxpayer behaviour, which has not been widely explored in this domain.

- **Methodological contribution:** This study contributed to the existing literature on machine learning in public sector management by providing practical insights into data preparation, feature selection, and model development, specifically applied to a real-world challenge in tax administration. With the booming of data governance and management, there is a shift from the traditional way of analysing data.

In summary, the research aimed to bridge the gap between machine learning techniques and practical applications in tax compliance, potentially leading to more efficient tax systems and better policy interventions.

5.5 Recommendations for future research

To enhance model estimation accuracy, future studies may consider utilising a longer historical period of taxpayer records, preferably exceeding five years. The current study explored six supervised learning models; however, future research could expand this scope by investigating unsupervised learning techniques, such as clustering algorithms and Isolation Forest, to identify anomalous transactions or inconsistencies between declared income and observed spending patterns.

This study primarily relied on two variables status-on-return and date as other attributes were outside the scope of the current research objectives. Future investigations should assess the predictive value of additional variables within the same dataset to uncover deeper behavioural patterns.

Moreover, enriching the modelling framework with external data sources, such as those from the Business and Intellectual Property Authority (BIPA), Social Security Commission (SSC), Ministry of Home Affairs, Immigration and Safety, Banks and digital tax platforms, could significantly improve the training dataset and enhance forecast accuracy. Incorporating variables related to economic activity, institutional factors, and socio-economic indicators may further strengthen model robustness and support evidence-based tax policy development in Namibia.

To further improve predictive performance, future research could explore advanced techniques such as:

- **Deep learning models** (e.g., neural networks) for capturing complex, non-linear relationships in taxpayer behaviour.
- **Ensemble stacking**, which combines multiple models to leverage their individual strengths and improve overall accuracy.
- **Integration of behavioural datasets**, including taxpayer sentiment analysis from social media platforms like X (formerly Twitter) and Facebook, to gain insights into perceptions, service satisfaction, and institutional trust. These approaches may also support fraud detection through machine learning-based behavioural analysis.

5.6 Concluding remarks

This study illustrated the transformative potential of machine learning (ML) in enhancing tax administration by accurately classifying individual taxpayer return submissions as either on time or late. Through rigorous data wrangling, feature engineering, and the implementation of supervised ML models, the research demonstrated how predictive analytics can be effectively applied to modernise tax compliance monitoring systems. By employing robust evaluation metrics such as accuracy, precision, recall, and F1-score, the study offers a replicable framework for assessing model performance in real-world tax environments.

The findings carry important implications for tax authorities aiming to improve operational efficiency, reduce manual processing burdens, and adopt more targeted, evidence-based compliance strategies. The ability of ML models to uncover behavioural patterns and submission trends enables proactive interventions, thereby shifting tax administration from reactive enforcement to strategic, data-informed governance. Such capabilities not only support improved taxpayer segmentation and risk profiling but also inform the design of more equitable and responsive policies.

Furthermore, the study makes a substantive contribution to the growing body of research on the application of artificial intelligence and data science in public sector management, particularly in fiscal and revenue administration domains. It sets a methodological and conceptual foundation for

future studies seeking to explore the integration of emerging technologies into tax systems, including real-time analytics, anomaly detection, and behavioural prediction.

Ultimately, the research reinforces the case for embracing intelligent, data-driven solutions in building resilient, adaptive, and service-oriented tax institutions. By doing so, tax authorities like NamRA can enhance compliance outcomes, optimise resource allocation, and align more closely with national development goals and digital transformation agendas.

6. References

Achakzai, M. A. K., & Juan, P. (2022). Using machine learning Meta-Classifiers to detect financial frauds. *Finance Research Letters*, *48*, 102915. https://doi.org/10.1016/J.FRL.2022.102915

Alm, J., & Soled, J. A. (2016). W(h)ither the Tax Gap? *Working Papers*. https://ideas.repec.org/p/tul/wpaper/1618.html

Alm, J., & Torgler, B. (2011). Do Ethics Matter? Tax Compliance and Morality. *Journal of Business Ethics*, *101*(4), 635–651. https://doi.org/10.1007/S10551-011-0761-9

ATAF Communication. (2022, March 4). *ATAF vows to build data-intelligent Tax administrations in Africa*. ATAF Communication. https://www.ataftax.org/ataf-vows-to-build-data-intelligent-tax-administrations-in-africa

ATO. (2024, August 12). *How we use data and analytics | Australian Taxation Office*. https://www.ato.gov.au/About-ATO/Commitments-and-reporting/Information-and-privacy/How-we-use-data-and-analytics/

Bassey, E., Mulligan, E., & Ojo, A. (2022). A conceptual framework for digital tax administration - A systematic review. *Government Information Quarterly*, *39*(4), 101754. https://doi.org/10.1016/J.GIQ.2022.101754

Business Application Research Centre. (2018, February 28). *Data Governance: Definition, Challenges & Best Practices [Interactive]*. https://bi-survey.com/data-governance

Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P. J., Ma, Q., & Zhang, Y. (2020). Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. *Computers in Biology and Medicine*, *123*. https://doi.org/10.1016/j.compbiomed.2020.103899

Del Carmen, G., Espinal Hernandez, E. E., & De Gouvea Scot De Arruda, T. (2022). *Targeting in Tax Compliance Interventions: Experimental Evidence from Honduras*. https://doi.org/10.1596/1813-9450-9967

Gichohi, B. W. (2020). Leveraging on big data and advanced technologies to enhance domestic revenue mobilization. *Statistical Journal of the IAOS*, *36*(S1), S111–S119. https://doi.org/10.3233/SJI-200706

Hayek, A. F., & Noordin, N. A. (2024). Tax Data Analytics. *Springer Briefs in Applied Sciences and Technology*, *Part F3223*, 23–31. https://doi.org/10.1007/978-3-031-63326-3_4

International Monetary Fund. (2024). Tax Administration: Essential Analytics for Compliance Risk Management. In *Technical Notes and Manuals* (Vol. 2024, Issue 001). International Monetary Fund. https://doi.org/10.5089/9798400260063.005.A001

Jørgensen, B. (2021). *BECOMING DATA-DRIVEN?* [PHD]. IT University of Copenhagen.

jvanzyl. (2019). *Launch of a tax incentive programme in encouraging e-filing through ITAS Tax First Alert*.

Kamara, R. G. (2021). *THE EFFECT OF TAX ADMINISTRATION ON TAX REVENUE COLLECTION IN KENYA REVENUE AUTHORITY*.

Liuhong, C. (2022). *Tax Collection and Administration Application and Legal Issues Based on Big Data Analysis*. https://doi.org/10.1155/2022/6578964

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *2017-December*, 4766–4775. https://arxiv.org/pdf/1705.07874

Maksimović, J., & Evtimov, J. (2023). Positivism and post-positivism as the basis of quantitative research in pedagogy. *Research in Pedagogy*, *13*(1), 208–218. https://doi.org/10.5937/istrped2301208m

Milner, C., & Berg, B. (n.d.). *Tax Analytics Artificial Intelligence and Machine Learning-Level 5*.

Mok, M. S. (2021). *Data Quality Management BRITACOM Seminar 3 ~ Tax-related Data Governance and Application ~ Inland Revenue Authority of Singapore (IRAS)*.

Murorunkwere. (2022). Fraud Detection Using Neural Networks: A Case Study of Income Tax. In *Future Internet* (Vol. 14, Issue 6, p. 168). MDPI. https://doi.org/10.3390/fi14060168

Mwafongwe, J. (2020). *NAMIBIA ESAAG 27TH ANNUAL INTERNATIONAL CONFERENCE Appropriate Resource Management to Increase Revenue Streams Presented at the 27 th Annual Conference for East and Southern African Association of Accountants General*.

Namibia Revenue Agency. (2023). Communique Wagon Lazarus Amukeshe Happy 2 nd Anniversary to NamRA Serving with passion How much of a relief is this? How it works. In *NamRA Communique Wagon* (Vol. 10). https://www.namra.org/newsletter

OECD. (2016a). *Advanced Analytics for Better Tax Administration: Putting Data to Work*. https://doi.org/10.1787/9789264256453-en

OECD. (2016b). *Tax revenue | Tax | OECD iLibrary*. OECDiLibrary. https://www.oecd-ilibrary.org/taxation/tax-revenue/indicator/english_d98b8cf5-en

OECD Independent External Evaluation Final Report. (n.d.). Retrieved September 29, 2025, from www.iodparc.com

Paul-Emeka George, E., Idemudia, C., & Bolatito Ige, A. (2024). *Predictive analytics for financial compliance: Machine learning concepts for fraudulent transaction identification*. https://doi.org/10.53022/oarjms.2024.8.1.0041

PricewaterhouseCoopers. (2018). *The Data Intelligent Tax Administration Meeting the challenges of Big Tax Data and Analytics 2*.

Puyeipawa, N. (2021, April 11). Namra opened in Namibia - Google Search. *The Namibian*. https://www.namibian.com.na

Saragih, A. H., Reyhani, Q., Setyowati, M. S., & Hendrawan, A. (2023). The potential of an artificial intelligence (AI) application for the tax administration system's modernization: the case of Indonesia. *Artificial Intelligence and Law*, *31*(3), 491–514. https://doi.org/10.1007/s10506-022-09321-y

Shannon-Baker, P. (2022). Philosophical underpinnings of mixed methods research in education. In *International Encyclopedia of Education (Fourth Edition),* (pp. 380–389). https://digitalcommons.georgiasouthern.edu/. https://doi.org/10.1016/B978-0-12-818630-5.11037-1

Stern, R. (Consultant of A. D. B., Sanger, C., & Asian Development Bank. (2022). *Launching a digital tax administration transformation : what you need to know.* 55.

Zhou, F., Zhu, J., Qi, Y., Yang, J., & An, Y. (2021). Multi-dimensional corporate social responsibilities and stock price crash risk: Evidence from China. *International Review of Financial Analysis*, *78*. https://doi.org/10.1016/J.IRFA.2021.101928

7. Appendix A: NUST Ethical Clearance Letter

P.O.BOX 30937

Windhoek

Student no:222074299

07/03/2023

The Secretary

NUST Ethic Clearance Committee

Windhoek

Namibia

Dear Sir/Madam,

RE: The omission of the questionnaire and Informed Consent Form

My name is Sifani Sifani, Student number: 222074299 and I am pursuing the Master of Data Science degree program at the Namibia University of Science and Technology; my research topic is Towards a data-driven governance framework for a Revenue Authority. This letter is to inform the committee that my research does not require a questionnaire and Informed Consent form and the following reasons support why I am not submitting those documents: My research is based on archived data available at Namibia Revenue Agency, Secondly, the study does not require an informed consent form as I am not interacting with users or participants to ensure informed consent.

Thank you for your consideration.

Yours Sincerely,

Sifani

8. Appendix B: NamRA data collection acceptance letter

Namibia Revenue Agency
PO Box 569
Windhoek, Namibia

NamRA
Namibia Revenue Agency

Enquiries: Fanuel Uugwanga
Email: fanuel.uugwanga@namra.org.na
Telephone:061 209 2196

02 March 2023

Mr. Sifani Sifani
P. O. Box 30937
Windhoek

Dear Mr. Sifani,

REQUEST FOR PERMISSION TO CONDUCT RESEARCH

The subject matter refers.

This letter serves to confirm receipt of your letter, wherein you requested for consent to conduct research on the following topic *"Towards a data driven governance framework for a Revenue Agency"*.

We wish to inform you that your request has been approved, on condition that the survey results will strictly be used for academic purposes, shared with NamRA, and should not be disclosed or shared with the public without expressed written consent by NamRA.

The focal person will be our Manager: IT Governance and Security, Mr. Mbaunguraije Tjikuzu, (email: mbaunguraije.tjikuzu@namra.org.na)

Trust that the above is in order and good luck with your research and studies.

Yours Sincerely,

FANUEL UUGWANGA
CHIEF: HUMAN CAPITAL & STRATEGY

cc: Ndapewa Hedipinge & Mbaunguraije Tjikuzu

NAMIBIA REVENUE AGENCY
NamRA
REPUBLIC OF NAMIBIA

9. Appendix C: Dataset Dictionary

1. TIN = Tax Identification Number
2. Taxpayer name = Taxpayer name
3. Taxpayer category = Taxpayer category, e.g., individual
4. Taxpayer type = Type of tax, is it a farmer or an individual
5. Office = NamRA office/centre where the application was submitted
6. Tax type = Tax type is VAT or Tax
7. Return status = The status of the tax is assessed, posted, or rejected
8. Tax year = The year the tax was filed
9. Return period = Period of the return
10. Source = Where the return was submitted (manual/portal)
11. Received Date = The date the return was submitted for filing
12. Due Date = closing date or due date of filing

RVS means a return is filed by a representative or bookkeeper filing on behalf of the taxpayer.

PST means the return was posted and assessed.

TP means the return filed in person at the NamRA office.

PT means the person submitted online on the portal (ITAS).

POST means submitted via NamPost.

REP means return filled by a representative.

REPPT means a return filed by a representative on the portal.

TA means Tax Authority (NamRA on behalf of the taxpayer).

EML means sent via email.

TRU means Trust.

GOV means Government.

BUS means Business.

ITX means Income tax.

ETX means Employee's tax.

VIA means Value Added Tax on the import account.

VAT means Value Added Tax.

WTS Withholding tax on services.

STD means Stamp Duty.

TFD means Transfer duty.