AFRICAN CENTER OF EXCELLENCE
IN DATA SCIENCE

Value Added Tax Fraud Detection using Naïve Bayes Data Mining Approach Case Study: Rwanda 2016-2019

By

CLAUDINE MUNEZERO

Registration Number: 219014391

A Dissertation Submitted in Partial Fulfilment of the Requirement for the Degree of Master of Data Science in Data Mining

University of Rwanda College of Business and Economic

Supervisor Name: Dr HAKIZIMANA JMV

September 2020

DECLARATION

I, Claudine MUNEZERO as a result of this dissertation, declare that the work presented in this dissertation titled "Value Added Tax Fraud Detection using Naïve Bayes data mining approach (Rwanda 2016-2019)", is my work and has never been presented in any institution of higher learning for any academic reward or qualification.

Signature  **Date:14/09/2020.....**

Student Name: Claudine MUNEZERO

Registration Number: 219014391

Our approval has submitted this thesis as a supervisor.

Signature  **Date:14/09/2020.....**

Supervisor Name: Dr. Jean Marie Vianney HAKIZIMANA

DEDICATION

This study is dedicated to my God. I also dedicate this work to my husband, Albert MURENZI. He has been a great source of motivation and inspiration all along with my courses over this very long and challenging journey. My children Jessy, Jovita and Jayden for your love, patience, and understanding during the endless evenings at the campus. My entire family and friends for their encouragement. In closing, this dissertation is dedicated to all those who interested in lifelong learning.

ACKNOWLEDGEMENTS

I am grateful to God for enabling me to pursue the dream of a master's degree in data science through the entire process from the course's study to dissertation write-up. This work would not have been possible without the support of the African Centre of Excellence in Data Science and technology of the University of Rwanda through critical scientific inputs and supervision. I recognise the support of my supervisor Dr JMV HAKIZIMANA for his guidance in the whole process of my dissertation writing by ensuring that it maintained the scientific quality and format required by the University of Rwanda. My heartfelt thanks also go to my beloved family, especially my husband, for the priceless moral and financial contributions towards my studies that they made. Also, to my colleagues and relatives who contributed in many ways to make sure that I could make it to the end. May Almighty God bless you all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CART: Classification and regression trees

CHAID: Chi-squared automatic interaction detector ()

CIT: Corporate Income Tax

CRISP-DM: Cross Industry Standard Process for Data Mining

CSV: Comma-Separated Values

DT: Decision Tree

EBM: Electronic Billing Machine

ETAX: Electronic Tax

FN: False Negative

FP: False positive

HS: classification models with harmony search

INTA: Iranian National Tax Administration

KDD: Knowledge Discovery from Data

KNN: K-Nearest Neighbour

LR: logistic regression

MLP: Multilayer Perceptron

NB: Naïve Bayes

OCA: Overall Classification accuracy

PIT: Personal Income Tax

RRA: Rwanda Revenue Authority

ROC: Receiver Operating Characteristic

SOM: Self-organising maps

SVM: Support Vector Machine

TP: True positive

TN: True Negative

VAT: Value Added Tax

ABSTRACT

Today's Tax fraud embraces various new means to commit fraud including declaring wrong information, underpaying tax due and carrying out financial businesses without considering legal frameworks. Like any other tax, Value-Added-Tax (VAT) is vulnerable to fraud which affects the growth of any country due to its numerous advantages and benefits. Recognising noncompliance for VAT's taxpayers is a weighty as well as challenging matter for Rwanda Revenue Authority (RRA), since there is a huge volume of VAT returns received daily and monthly that need complex techniques in order to discover new insights and analyse it effectively. Hence the need for a valuable intelligent tool to fight against fraud known as data mining to extract for patterns in massive volume of VAT data and automatically distinguish fraudulent patterns from legal ones. The main purpose of this present study is to analyse relationship between VAT's patterns, build and evaluate a data mining model for fraud detection on VAT historical data for RRA.The proposed solution used SQL queries to analyse patterns according to RRA business rules, and the model architecture is designed to reason using the classification techniques Naïve Bayes, Kneighbors and Decision Tree to classify the status for taxpayer VAT's compliance with two categories that are fraudulent and legitimate. Furthermore, the model performance is presented and compared. The classification results generated by our model on each technique are compared with respect to the performance measures such as accuracy, precision, recall, F1-Score and ROC curves. Generally, both algorithms showed a significant accuracy but the best performing being Naïve Bayes with 98% of accuracy. The developed data mining model is promising to effectively detect VAT fraud and therefore help to generate knowledge that can be used in the audit work performed by the RRA for feature decision making.

Keywords*: VAT Fraud, fraud detection, Data-mining, classification analysis, naïve Bayes, decision tree and kneighbors.*

CHAPTER I: GENERAL INTRODUCTION

1.0 Background

Tax revenue plays a critical part in the financial development of the nation. Tax avoidance as the illegitimate intentional actions by taxpayers taken in order to reduce their tax liability (Alm, J. 2012)**,** hamper revenue collection, which leads to inefficiency in government operations.

Value Added Tax as a consumption tax paid by every consumer on good or service to tax administration (F., Adegbie and Jayeoba,2016) has been introduced to a great majority of countries, in particular, more than 130 countries have implemented VAT, and there is a remarkable increase of tax revenues at different stages of economic growth reaching about 25% of the world's tax revenue (Harrison, Graham and Krelove,2005).VAT has allowed many countries to generate more revenues as it has a high-powered mechanism in the fact that citizen purchasing power increases based on economic growth. VAT is vulnerable to fraud like any other tax in many parts of the world because its characteristics offers opportunities for abuse. This study intent to analyse the exposure of the VAT fraud in Rwanda through noncompliance with a particular aim to provide what can be done as a preventive solution.

In Rwanda, VAT was introduced in 2001 with the hope to enhance domestic revenue mobilisation, and it became a key tax generator wherein 2013/14 was above the three-quarter of domestic revenues **(**Mascagni, Giulia, Monkam and Nell Christopher,2016**)**. In order to increase VAT compliance, the RRA introduced Electronic Billing Machine (EBM) in 2013, and it was mandatory to every business registered for VAT to install and use EBM for all VAT transactions whereby all data are stored on EBM digital card, and they are transmitted and stored to the RRA data repository in real-time. The recent VAT's evaluation shows that despite the increase for VAT revenues due to the use of EBM, some businesses are still underreporting their returns and sending incorrect transaction details (Mascagni G; Mukama, D. and Santoro,2019).VAT fraud has been a non-stop concern for RRA because the filing returns is made as self-assessment action; hence build enormous data that it increases in size and the ever-growing volume of tax-related records is the biggest issue as it includes huge bulk of data integrated from various data sources (Davia and Kastantin,2000) **.** Therefore, it is of great importance to illustrate how EBM together with other new technological developments known as data mining, can best be used to

detect VAT fraud while improving taxpayer compliance; here we explore how Rwanda Revenue Authority can argue that, and this can best be addressed when the use of EBM is combined with a well-built data rational models tools, which will lead RRA to gain insight information from their data-holdings together with a better understanding of their complex business structures, with the inclusion of new business models based on the digital economy to identify patterns of fraudulent tax schemes.

We have introduced the use of data mining techniques which is quickly becoming a popular artificial intelligence technique for better fraud detection such as Decision tree, K-nearest neighbour and Naïve Bayes so that RRA can benefit from their rich data with ability to quickly identify VAT fraudulent.

1.1 Statement of the problem

VAT is the most factor contributing a lot to government income for the growing economy in several countries. However, a big number of businesses habitually try to evade their VAT's properly and having an efficient way of detecting tax fraud is a challenging issue for tax administration in Rwanda like in other countries (Pomeranz and Dina,2013).
According to (Mascagni G; Mukama, D. and Santoro,2019), with the research done on Discrepancies in Taxpayers' VAT Declarations in Rwanda shown that the mismatch between sellers' and buyers' statements are still widespread in Rwanda, even with the great crack enforcement that the RRA has made on VAT together with the use of electronic billing machines (EBMs) and the implementation of validation control applied on VAT refund claims. Even though currently RRA has many systems in place to gather tax information and help in the investigation, it seems that their full potential remains unutilised in practice. There is a need for gaining insights from the available tax data in order to easily detect VAT fraud by using a power full data analysis tool to explore the information reported during VAT filing returns in the way to identify the existing threats to the proper operating of the VAT and taxpayers' compliance with this tax type.

It is upon this that the research would like to contribute on the improvement of knowledge discovery from the massive VAT tax data hold in RRA by using data mining approaches so that the future trends can be analysed and recognised effectively.

1.2 Objectives

1.2.1 General objective

The general objective of this study is to establish a VAT fraud detection model to enhance tax compliance in Rwanda.

1.2.2 Specific objectives

The specific objectives of this research are:

i. To explore patterns that relate to the relationship of VAT taxpayer status with other variables.
ii. To develop a fraud detection model appropriate to RRA.
iii. To assess the performance of the designed model to detect VAT fraud.

1.3 Research Questions

Based on the statement of the problem, the research questions are:

i. What are the main determinant factors (attributes) of fraud from the VAT data?
ii. By implementing a data mining model for VAT fraud detection will this improved audit and decision strategy?

1.4 Scope and Limitations

The expectation of this study is limited to identify fraud in the domain of tax administration in RRA with focus on VAT tax type using taxpayer's data for sales, importation and declaration in period of four year (2016-2019).Data mining will be used as a special tools to mine data specifically the classification method as the most dominant in fraud detection (Yue and & Chu,2007),whereby its three algorithms naïve Bayes (NB), decision tree (DT) and K-nearest kneighbors (KNN) will be assessed and rated upon the underlying VAT structured dataset.

CHAPTER II: LITERATURE REVIEW

2.0 Introduction

In this chapter we introduce the theoretical concepts of the DM methods applied in this work, interpolates key ideas related to this work and talk about the methodology and technologies applied in Data mining concerning fraud detection as a solution for this problem. After that, a review of related works in areas of application such as the banking Sector, Insurance, the Health, Education, Telecommunication and tax administration sector. Further, we will briefly tackle on academic studies using data mining in tax administration and describe the implemented data mining approaches which are Naïve Bayes, Decision tree and kneighbors. We finally end up with a literature summary and research gap.

2.1 Data Mining

(Azuaje and F. Witten ,2006) define “data mining as the mechanism of discovering patterns in raw data. The task must be automatic or semi-automatic. The perception discovered must be meaningful in, that they lead to some advantages, generally an economic advantage”.

According to (Jans, M., Lybaert, N., and Vanhoof, K. ,2007) "data mining is about explaining the past and predicting the future using data analysis, it is a multi-disciplinary field which combines statistics, machine learning, artificial intelligence and database technology" as it can be seen in Figure 2.1.

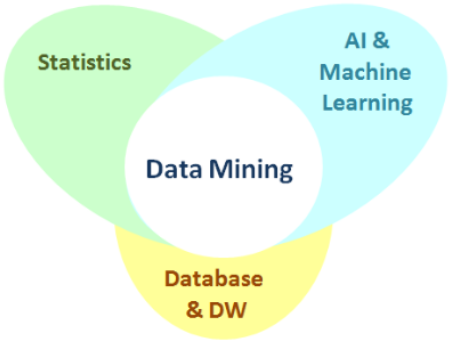

Figure 2.1: Disciplines for Data mining

According to (Deshpande, D., and Deshpande, S.,2017) Data mining is about choosing the right tools for the job and then using them skilfully to discover the information and issues in the data

where there is an identified problem to which an answer is needed or where it is suspected, or known that the answer is buried somewhere in the data but are not sure were exactly. In other words, (Gullo, F.,2015) we could describe data mining as the use of the sophisticated process to unearth uncertain, unspecified, and hidden patterns and connections in vast pre-accessible databases. It is a special technique of finding new facts and interconnection in the existing information that have not as yet been discovered by a specialist.

Nowadays, the improvement of device and technologies has amplified the frequency of electronic data creation. Technologies are readily available in handheld devices; according to (Bernus and Noran,2017), we are full of data but poor information. Even though the data are increasing at an alarming rate, processing this data and getting benefit out of the data is a challenging duty. Having those data only is just information/data overload, so to get the hidden pattern from the data, building organisation memory is needed. In order to get the most out of the available data, it is better to apply some processing techniques, called data mining so that the research has given rise to an approach to save and carrying out this valuable data for advanced decision making.

2.1.1 Data Mining Approaches

Data mining algorithms is divided into supervised and unsupervised techniques. "In supervised modelling, whether for the prediction of an event or a continuous numeric outcome, the availability of a training dataset with historical data is required. Models learn from past cases" (Tsiptsis, K. K., and Chorianopoulos, A.,2011). The supervised learning is used to construct a model to achieve tasks like classification, prediction, and it is called supervised because the classes are predefined before the exploration of the target data (Bramer, M.,2007). The supervised learning also called predictively is used in several domains like marketing, bank**,** insurance with the purpose to detect if any kind of fraud exists.

For unsupervised learning, there is no class label, and an example we have clustering used to group elements that share similar patterns. It is mostly applied to identify categories of houses based on their geographical location and values (Yongjian, F.,1997).

The supervised method is ranked as the dominant technique for detecting financial frauds in the public organisation (Ngai and Sun,2011). (Dhurjad and Rathod,2017) the different data mining methods are classification, prediction, time-series, association, clustering and summarization.

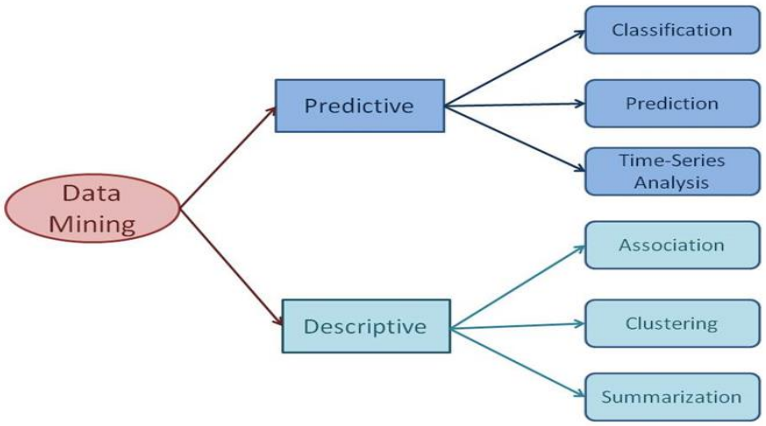

Figure 2.2: Data Mining Techniques

i) Classification

Classification is a process to identify the category of an object based on its properties. A classification model is created by exploring the interconnection between the target variable and features in training set with the aim to classify future element with a better understanding given data source mostly based on current or past assumptions (Han and Kamber,2011).

ii) Prediction

Prediction is a function to determine unavailable numerical data for a new event. Some dataset fields are used to predict the new event of other variables of interest, mostly based on the feature assumption (Williams, G.,2006).

iii) Time Series Analysis

Time series imply predicting numeric outcomes; it is a method to analyse trend analysis, series data to discover meaningful statistics and other features of the data. Time series deal with data of particular time intervals to predict future values based on previously discovered ones **(**Sugumaran and Gopal,2017**).**

iv) Association

Association is usually used to discover interesting relationships hidden in a big data set. It is a method to mine how objects are associated between them, and the discovered interconnection of the co-occurrence elements are expressed as association rule (Han and Kamber,2011)**.**

v) Clustering

Clustering is the way of building a cluster of abstract elements with the same similarities. The objects are clustered by dividing the set of data into interclass similarities with respect to follow the criteria defined on the attributes of elements and then assign them to their corresponding groups. Combined feature of the objects in a cluster is summarised to form the group description (Han and Kamber,2011)**.**

vi) Summarisation

summarisation is data mining concept to expedite KDD tasks by intelligently compacting the size of processed data in the form of tabulating to make them more understandable for data visualisation and automated reports (Ahmed, M.,2019).

2.1.2 Data mining process models

Data mining refers to a particular task among the entire process of knowledge discovery for extracting patterns from the database. The knowledge discovery in Database (KDD) seeks to identify what has deemed knowledge insights in the context of vast data repositories. It is described as an iterative process at each stage of integrating new data, transforming them in order to identify valid and potentially useful patterns by prioritising flexible and scalable databases as a primary source with the purpose to get more relevant results with a rapid response time to the data demands of modern businesses (Dzhurenko and Cherednichenko,2015). Knowledge discovery illustrates in Figure 2.3.

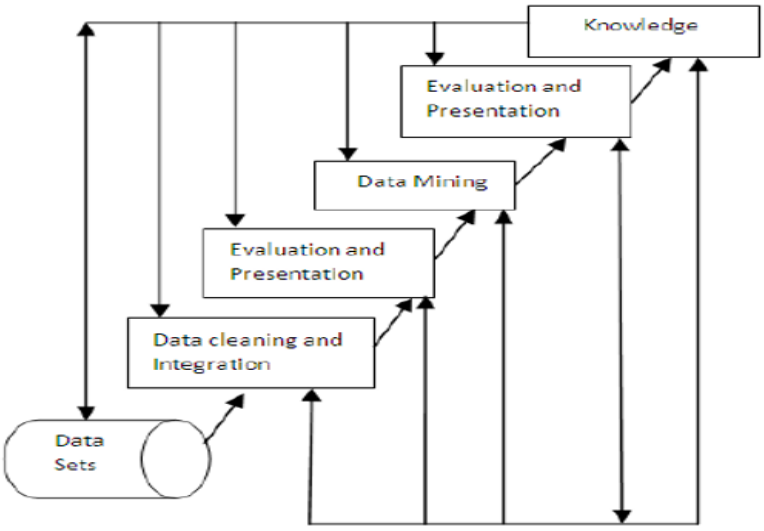

Figure 2.3: the process of knowledge discovery

According to (Padhy and Panigrahi,2012), the overall process for KDD is grouped into the following stages:

i) ***Learning the application domain****:* involves appropriate prior understanding and the intent of the application.

ii) ***Creating a target dataset:*** consist in determining a dataset on which discovery is to be carried out.

iii) ***Data cleaning and integration:*** defined as the process to remove noise and inconsistent data and handling missing data where required.

iv) ***Evaluation and presentation:*** data selection from the dataset by finding useful features relevant to the analysis task using dimensionality reduction and transform data into an appropriate form for a mining operation.

v) ***Data mining:*** apply the intelligent techniques able to discover insights of interest in a particular set and delivering potentially useful models.

vi) ***Interpretation*****:** includes explaining the discovered patterns by showing and translating the interestingness score of each pattern into terms understandable by users based on some interesting measures.

vii) ***Knowledge representation:*** represent and visualise discovered knowledge performance from data mining results and taking actions based on the gained knowledge.

2.1.2 Data Mining in business area

Data mining has been helpful in many fields with aim to find actionable strategies that can be used in a proper way to enhance efficiency. Some of the preliminary implementations were noticed for the retailer, more precisely in the form of groceries analysis (Olson & Delen,2008). Table 2.1 shows the broader application areas.

Table 2.1 Data mining application areas

Application area	Applications	Specifics
retailer	Affinity positioning, cross-selling	Position products effectively find more products for customers
Banks	customer relationship management	identify custom value, develop programs to maximise revenue
Credit card	Lift	
Insurance	Fraud detection	Identify claims meriting investigation
Telecommunications	Churn	Identify likely customer turnover
Human resource	Churn	Identify potential employee turnover.

2.2 Related works

Much publication talking about fraud detection (FD) with the use data mining methods has been reviewed. Various researchers have allocated a remarkable amount of interests in studying FD in some domains like banks, health, finance and tax.

2.2.1 Credit Card Fraud Detection

(Behera and Panigrahi,2015) have presented a novel algorithm to detect fraud using fuzzy c-means to find out the pattern's utilization of cardholders in relation with their past operation. The designed approach decreased the wrong classification score based on the feature's occurrence, but does not evaluate the time gap between transaction.

(Khandare,2016) discuss about different techniques implemented for credit card fraud. In application of Hidden Markov Model (HMM) three classes have been create (low, medium and high) based on their frequencies of transaction in terms of amount to classify customer's profile. The probability in term of money has been appointed to every client. The developed system was found much faster to detect fraud, but it does not show the experiment for the incoming transaction.

2.2.2 Tax Fraud Detection

As stated by (Carvalho and Souza,2016)**,** there are a big number of tax authorities which have implemented data mining methods to discover tax compliance of their taxpayers. However, despite being a matter of great interest, many constraints have been identified for inhouse projects. Given that taxpayer archive is labelled and should be covered by internal auditors; hence most of them are keeping pointed compliance risk as secrets.

(Boezio and Taboureau,2017) with his model on business tax fraud detection he proposed coloured network-based model to resolve the challenge of the usefulness of classic data mining-based tax evasion detection approaches but the developed method with the use of heterogeneous network information were ranked as time-consuming and tedious.

(Chen and S.,2016) have used hybrid algorithm to design a fraud detection for financial statements. The combination for regression trees (CART) with Chi-squared automatic interaction detector (CHAID) was adopted to identify the key variables. However, both had limitations as they were not able to compute the continuous numerical data.

(Halsteinslid,2019) tackling collinearity and imbalanced class using logistic regression for statistical fraud detection to model VAT fraud probabilities in the data set provided by the Norwegian Tax administration and logistic regression have been tested ridge and elastic net. Ridge uses the highest number of covariates since it does not perform variable selection. Elastic net struggled to find a uniquely best value which introduces much uncertainty, and the computational effort is much larger.

(Vasco, Rodríguez and de Madrid,2018) with the implementation and estimation for Personal Income Tax (PIT) model to identify and detect PIT evasion using Multilayer Perceptron (MLP) and one of the main problems of the designed approach was a large number of independent

variables which were multilinear and the significant imbalanced distribution of the target variable.

(Rahimikia and Ghazanfari,2017) have implemented hybrid intelligent using MLP, support vector machine (SVM), and logistic regression (LR) for the Iranian National Tax Administration (INTA) with a feature to detect corporate fraud. The proposed method was mainly interested in binary however; the probability outcome is the one most used practice.

(Pérez López, Delgado Rodríguez and de Lucas Santos,2019) implements neural networks intending to create taxpayer groups and calculating the probability for a single taxpayer to evade taxes. However, during the test, it was not possible to include all the available variables in the model, and the speed has been highlighted to be emphasised.

(González and Velásquez,2013) clustering algorithms like Self-organising maps (SOM) and neural gas have been applied to classify and identify in the universe of taxpayers the ones with false invoices. Self-organising map, together with artificial neural networks, have been used for clustering, segmentation with the main objective to create taxpayer's groups having similar behaviour, but the result has shown that:

i) The created groups were different from the objects of another group and clusters obtained were also mainly different because the neurons were required to maintain a fixed neighbourly correlation,
ii) Some patterns of behaviour have been identified without a specific relation with the use of false invoices.
iii) There is no significant variation from one group to another while exploring the variable associated with historical behaviour and irregularities.
iv) For neural gas, there is greater variability between the selected variables used to detect false invoices.

(Cút, S. ,2015) with the risk assessment of VAT attributes, the selected data mining techniques like decision trees, random forests and neural network algorithm MLP have been implemented, and respectively a classification rule to classify selected taxpayers accurately into the appropriate category (default/non-default). The best overall classification ability of 94.20% was achieved by the feedforward MLP neural network using the backpropagation algorithm of

errors. However, despite the MLP to be the most appropriate data mining tool, it has shown some limitations that may appear with:

i. Its practical implementation which can eventually result in highly biased estimates.
ii. Problematic randomization of the data sample selection due to the infrequent occurrence of default cases.

(Gayathri and Malathi,2013) summarize the existing challenges for the various huge dataset in fraud detection using data mining classifiers. The strengths and limitation of each technique are given in table 2.2.

Table 2.2: Strengths and limitations of data mining techniques

Technique	**Strengths**	**Limitations**
Artificial Neural Network	Ability to learn from the past and predict future activities	difficulty to confirm the structure with the high processing time poor explanation capability and difficult to set up and operate sensitivity to data format and high expense Need a high computational power unsuitable for real-time operations, re-training is required for new types of fraud several parameters (Raghavendra Patidar,2011) have to be set before any training can begin The topology placed a major role in a network

		performance but, there is a lack of methods exists to determine the optimal topology for a given problem due to its high complexity of large networks
Hidden Markov Model	Fast in detection	Highly expensive low accuracy not scalable to large size data sets
Support Vector Machine	deliver a unique solution since the optimality problem is convex	poor in process large dataset expensive has a low speed of detection Not easy to process the results due to the transformation of the input data The biggest limitation of SVM lies in the choice of the kernel (the best choice of the kernel for a given problem is still a research problem). A second limitation is a speed and size (mostly in training - for large training sets, it typically selects a small number of support vectors, thereby

		minimising the computational requirements during testing).
fuzzy logic	very fast in detection	it is expensive
Decision tree	high flexibility and good haleness	the requirement to check each condition one by one in fraud detection condition is a transaction the potential of overfitting Decision-tree learners can create over-complex trees that do not generalise well from the training data The reliable information in the decision tree depends on providing the required internal and external information properly Large changes can be made in a tree with even small changes incorporated in the input data. Variable change, without including the same information, or sequence alteration midway, can lead to major changes and might require redrawing the tree. the decision tree analysis is that the decisions contained

		in the decision tree are based on expectations, and thus these expectations lead to many errors in the decision tree it follows a natural way by tracing relationships between events, contingencies which arise from a decision may not be possible and thus it, in turn, leads to bad decisions
k-Nearest Neighbour	Very simple implementation	Assumption of class conditional independence usually does not hold. If the sample size increases significantly, it cannot be handled efficiently. High calculation complexity: To find out the k nearest kneighbour samples Dependency on the training set: The classifier is generated only with the training samples, and it does not use any additional data. No weight difference between samples: All the training samples are treated equally; there is no

		difference between the samples with a small number of data and a huge number of data
Linear Regression	Optimal result between linear independent and dependent variable	Sensitive to noise and limited to numeric values only
Logistic Regression	Easy to implement	Poor classification performance

(Uddin, Khan and Moni,2019) said that among the various methods that have been advanced to handle fraud one method is the support vector machines (SVM) but it has the main challenge to choose the kernel and building the model seems to be complex and entails time demanding calculation.

(Barros, De Carvalho and Freitas,2015) talked on a decision tree with the main disadvantage to present instability since small discrepancy in the data may result in the generation of an entirely different tree. Decision trees are also susceptible to the overfitting problem where the implemented classifier was fitting perfectly but many instances have been note considered in training process.

(Zhou and Kapoor,2011) “examine the effectiveness, scalability and speed of data mining methods like regression, decision trees, neural network and naïve Bayes” and the final result is shown in below Table 2.3.

Table 2.3: effectiveness, scalability and speed of data mining

Data mining technique	Effectiveness	Scalability	Speed
Naïve Bayes	Good	Excellent	Excellent
Decision tree	Excellent	Poor	Good
Neural network	Good	Excellent	Poor

2.3 Proposed RRA VAT fraud detection technique

2.3.1 Naive Bayes Classifier

“A naive Bayes classifier is a simple and powerful probabilistic classifier based on applying Bayes’ theorem with strong independence assumptions”. “The achievement goal is to forecast the class of test occurrence as accurately as possible. This kind of classifier is named naive because it is composed with two common assumptions: firstly, it assumes that the predictive instances are conditionally independent given the class and secondly, the values of numeric features are equally distributed within each class and used for both binary and multi-class classification issues” (Taheri and Mammadov, 2013).

As stated by (Korb and Nicholson, 2010), the advantages for the Naive Bayes algorithm can be observed in three different ways. First of all, it is easy to be implemented due to its simplicity which requires fewer data to get a good result in many cases and with the use of probabilistic models which are able of giving overall judgement for uncertainty transactions. Secondly is his ability to experience through self-adaptive direct learning. Thirdly it can achieve high speed and high accuracy in a huge amount of data. “Naive Bayesian algorithm is considered as a group of collection of classification algorithm based on Bayes theorem where all the algorithm shares a common principle that every pair of features which is classified is not dependent on each other” (Berrar,2018).

Last but not least, the training using Naive Bayes takes a short computational time, and the model is easily developed. The estimation for parameter iteration is less complex with the capability of being robust, suitable for large dataset across the board and easily interpret the discovered knowledge.

2.3.2 Decision Tree

Decision trees are considered as simplest among data mining algorithms. Those algorithms are an entirely transparent technique of characterising observations that look like a sequence of if-then statements ordered into a tree (Segaran, 2007). Classification starts from topmost node called root node down the tree to the leaves based on the outputs. Leaves are the target classes of our dataset. Classification can be easily applied by answering question down the decision

tree. The logic applied for DT is to divide training data in two more than one set using the most significant predictors in order to provide different classes.

2.3.3 K-Nearest-Neighbor

The last implemented and tested data mining method is KNN. "The k-nearest-neighbor method is labor intensive when given large training sets, and It has since been widely used in the area of pattern recognition" (Al-Faiz and Miry,2010). "In pattern recognition, the k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning or lazy learning where the function is only approximated locally, and all computations are deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor." (Triguero and Herrera,2019).

2.4 Summary

Data mining plays a vital role in fraud detection due to its performance ability to compute a huge data set. Besides, data mining approaches is able to cope with the complexity of the patterns within a vast amount since they require to use low-cost computation and giving a high precision outcome as discovering potential fraud is a challenging and daunting task.

According to research literature about fraud related to tax administration in general, we have noticed that the tools used for detecting tax evasions and financial statement fraud are almost the same. Most of the related work focused on applying only one data mining technique such as classification or clustering to mine the fraud data based on historical data. Few studies used a hybrid algorithm which is based on cluster analysis and classification. From the review of the literature, we can say that in many studies the supervised classification methods such as DT, SVM, random forest and neural network were mostly used to predict fraud in different domains like insurance, banks, telecommunication companies and financial institutions and given that the performance assessment may differ as every problem is set up in different ways and varies with the use of different data sources the comparison between classifier algorithms are difficult to make.

To conclude, we mention that classification accuracy mainly depends on the data mining application domain, so influencing factors that can affect fraud should be considered to come up with high classification accuracy.

2.3 Gaps in knowledge

Given all proposed improvements as discussed from great works of other researchers, it is important to criticise the literature by finding gaps in developed models. The listed below gaps were highlighted as the missing links on learning as regards the investigation on the paper of previous authors in the field of data mining fraud detection.

i) Generally, have worse performance than ensemble methods
ii) Expensive and slow to predict new instances
iii) May overfit when provided with large numbers of features.
iv) May not handle irrelevant features well.

Another drawback is that many researchers have designed their classification model using balanced datasets with equal size of fraudulent and no fraudulent cases which is not reflecting the real situation in real life since in many cases the ratio of fraudulent cases is very small compared to no fraudulent ones. Some researchers have implemented data mining techniques to assess business operational continuity and thereby ignore performance evaluation.

This research aims to build a better model appropriate to RRA to help in discovering the hidden potential fraudulent taxpayers in the existing massive data for VAT so that the losses incurred by VAT evasion and the time implied for conducting professional audit can be reduced by designing a supervised classification model with the ability to cope with all the gaps mentioned above.

CHAPTER III: METHODOLOGY

3.0 Introduction

For the methodology part we are focusing on the ways through the guidance of which the specific objectives for this study can be streamlined and can be designed and be put effectively into action.

It is important to highlight that "the nature of methodology applied and used is dependent upon the kind of research that is conducted. The nature of methodological principles applied varies in terms from usage, whether it is quantitative or qualitative, and changes in methodology are incorporated accordingly "(Bryman,2008)**.** In our methodology, we are looking at the challenges that show problems in the present work.

(i) Can we select significant input patterns? We extract new fraud indicators from specific fields of the VAT dataset and show the relationship of these features.
(ii) Can we develop a fast model to deal with the large VAT data size for RRA? Different models are developed, and we illustrate their scalability both theoretically as well as experimentally.
(iii) How can we evaluate the performance of the designed fraud detection model? An evaluation methodology is proposed that provides reliable performance indications and guarantees that the proposed methods effectively detect fraud cases.

This study employs three different classification data mining approaches (Naïve Bayes, Decision tree and kneighbors) using VAT data (sales, importation and declaration) from 2016-2019 to uncover patterns and extract relationships among variables that are useful for detecting VAT fraud. In this study, a screening model will be developed based on usage patterns discovered from selected VAT dataset features. This model is utilised to select the cases that are classified as VAT fraud for further auditing checks hence enhance tax auditor's productivity in recovering VAT revenue losses.

3.1 Data Mining Process

The business logic of the present study is based on the well-known process called Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is defined as an independent

standard methodology used by data mining experts in every business to design a data mining project. It outlines each step every data mining work should pass. All stages have the same significance to contribute on a successful data analysis workflow, and none of them should not be ignored. (Wirth and Hipp,2000) The phases of CRISP-DM are:

i) **Business Understanding:** every data analysis technique should answer business questions to achieve business goals. In this phase, we first have to understand the requirement by finding what the business requirement is, secondary evaluate different resources and existing assumptions while considering other important factors and finally utilise data mining with a well-detailed plan and establish new data to achieve business objectives (Siraj and Abdoulha,2007).

ii) Data Understanding: The second phase is designed to evaluate the source of data together with its quality and characteristics. The initial data exploration provides insights by tackling the data mining questions using querying, reporting and visualisation. The result is a detailed understanding of the key data elements that will be used to build models. This phase is very critical and can be time-consuming when you have many data sources. Once the main issues are clear, it is time to understand the needed requirement that will be used to achieve the business goals highlighted in previous stage. The identified patterns are mapped to every target to identify the existing gaps and lack of useful details. "It is stated that there is a close link between business understanding and data understanding. The formulation of the data mining problem and the project plan requires at least some understanding of the available data" (Wirth and Hipp,2000).

iii) Data Preparation: The main tasks in data preparation are dataset construction, feature selection, data cleaning and transformation. The constructed dataset is then converting into a suitable format able to be mined and based on organisation's goals you can determine which type of algorithm to apply.

iv) Modelling: The modelling phase includes selecting and building data mining algorithms that extract the knowledge from the data. There are a variety of data mining techniques; each suitable for discovering a specific type of knowledge tax agency would use classification or regression models, for example, to discover the characteristics of more productive tax audits. Each technique requires specific types of data, which may require a return to the data preparation phase. The modelling phase produces a model or a set of

models containing the discovered knowledge in an appropriate format. In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Data preparation and modelling steps are closely linked because, in data preparation, the data issues are realised and give ideas to construct new data during the modelling process.

v) Evaluation: In this phase, we are focusing on evaluating the quality of the model and the result in the context of the business goals.

Data mining algorithms can uncover an unlimited number of patterns; many of these, however, may be meaningless. This phase helps determine which models are useful in terms of achieving the project's business objectives. At this stage in the project, you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

vi) **Deployment:** In the deployment phase, we have to present the result we gained through the data mining process for the decision-making process. Depending on the pertinence of the results, minor modifications may be required, or the necessity for a major reengineering of the whole model development will be recommended, and the proposed model is not generally the end of the project. In general, the knowledge gained will need to be organised and summarised in a way that the stakeholders can use it whenever they want. (Wirth and Hipp,2000**)** The whole CRISP-DM process is illustrated in the below Figure 3.1.

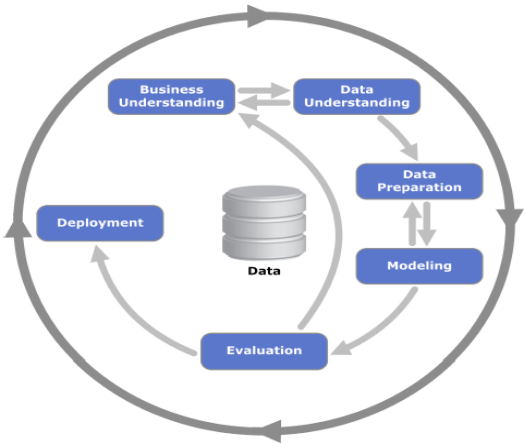

Figure 3.1: CRISP-DM Model process

3.2 Research Design

Research design is the overall workflow to link the conceptual research problems together with the suitable achievable empirical research (Lelissa and Kuhil,2018). Specifically, the research design demonstrates what will be the process flowchart of your work from the starting point to the end.

As a data mining technique, the materials and methods are based on a database repository and data mining technique. Our work is organised around the following stages: data acquisition, data cleaning, data integration, data pre-processing-processing, data mining modelling and evaluation of the result. The proposed design is in Figure 3.2 illustrated below.

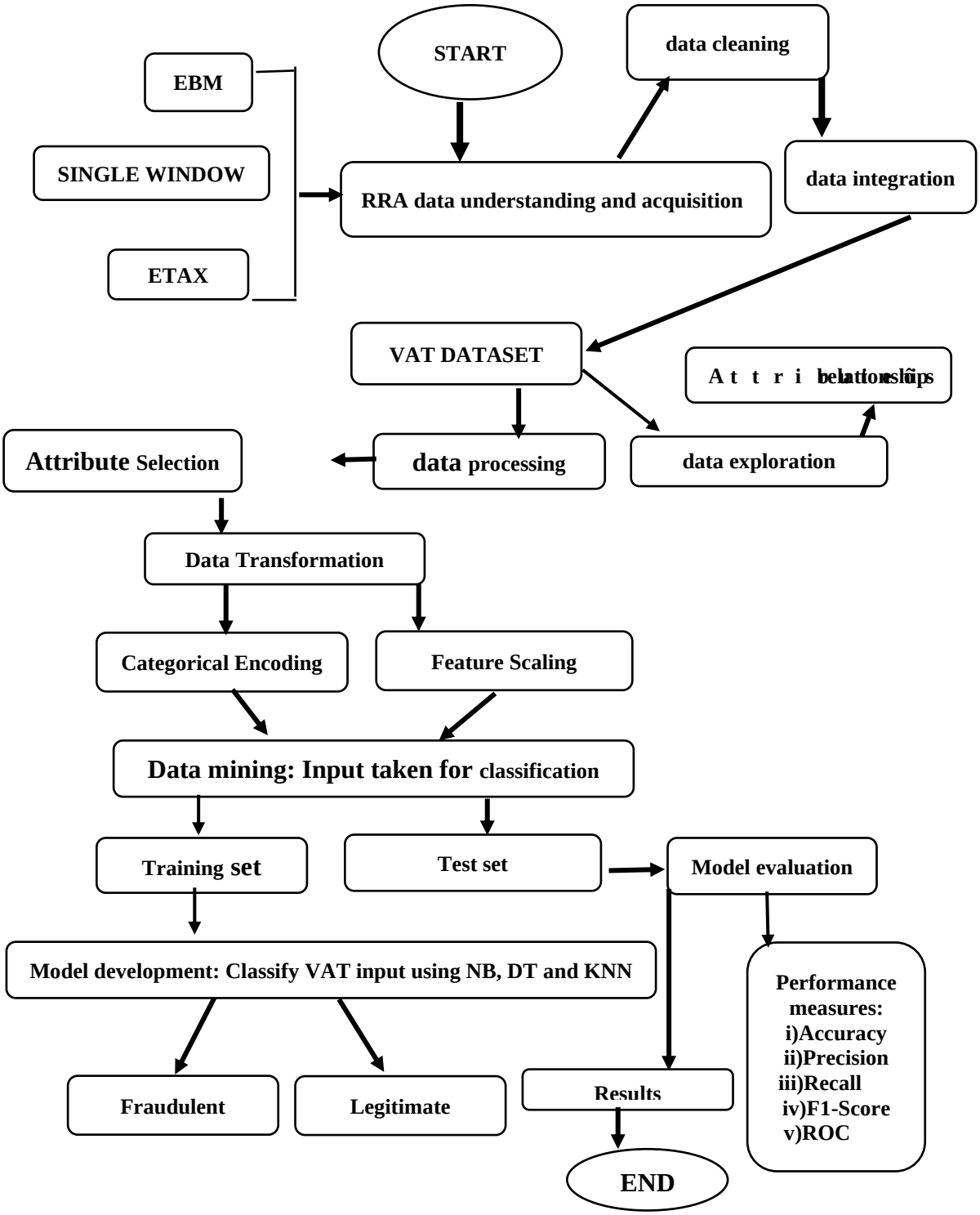

Figure 3.2: Proposed Methodology for VAT fraud detection model in RRA

In this study, the RRA VAT fraud detection methodology illustrated in Figure 3.2 has been implemented using taxpayer's historical VAT data which have been transformed into a suitable required format to our classifier algorithm with the help of data pre-processing-processing together with data exploration to first discovering the relationship between the independent variables and target variable and extracting useful features. VAT fraud detection model will use VAT data from the operational database to be uploaded into data mining software where data mining techniques will be applied.

3.2.1 Data Understanding and Preparation

The first phase of work was to analyse the database in order to extract important data to be mined. Data understanding and data acquisition tasks should be carried out carefully to come up with good output in data mining. The reason is that the models that will be built mainly depend on these tasks. In my study, I manage to use secondary data collected from RRA into three different databases for VAT tax type from 2016 to 2019. The three data sources are:

i)sales from EBM

ii)importations from SINGLE WINDOWS

ii)declaration from ETAX. The historical data is from.

3.2.2 Data Cleaning

Data cleaning aims to improve data quality by identifying any dirty data and then replacing, modifying or deleting them to prevent data corruptions from happening again. We have to manage our data by making them free of irrelevances which may hinder the knowledge discovery process and at the end and provide inaccurate results.

3.2.3 Data Integration Construction

The other important step is the integration for attributes from different data sources and the fact of generating other useful variables from the existing ones. Extracting significant patterns that mean the relationships in the dataset are likely to help increase the likelihood of the knowledge discovery process yield accuracy result.

3.2.4 Data Preprocessing

When it comes to developing a data mining model, data pre-processing becomes the first step to start the logical process. (Akaranga and Makau,2016) Generally, real-world data contains outliers with lack of specific features trends and with gaps to be incomplete, inconsistent and

inaccurate. This inaccuracy is where data pre-processing plays a big role to clean, format, and transform the raw data, thereby preparing it ready-to-go for knowledge discovery.

i) Attribute Selection

Feature selection is an important step to come up with an accurate predictive model. Reducing irrelevant and redundant inputs improve the prediction models by focusing only on the most meaningful features. Feature extraction helps to avoid overfitting so that the developed model can predict not only the test dataset but also other instances that are new to the trained model.

ii) Data Transformation

Reshaping business data can ensure a well-structured dataset which is vital to gain precise analysis, and look for valuable patterns that will ultimately reinforce data-driven resolutions.
In our methodology, we have used two types of data transformation:

ii.1 Categorical Encoding: most of machine learning techniques are not working directly with categorical data as they lack any intrinsic mathematical connections, so it is required first to do some modification on the data to be mined before feed it to a machine learning model.

ii.2 Feature Scaling: is a technique to standardise the independent features present in the data in a fixed range" (Thara, PremaSudha and Xiong,2019). If the standardisation is not done, then the data mining algorithm tends to weigh greater values since the range may vary widely, so there is a necessity to put our data on the same scale.

3.2.5 Data Exploration

In data exploration we will show the relationship of VAT taxpayer status with other variables to have a clear idea on how the independent variables are contributing to dependent variable.

3.2.6 classification engine development

Classification engine is the process of providing the processed data to the candidate classification algorithm and identifying the model that shows better performance.
There are several tasks involved in the phase. Some of the tasks include splitting data into training and validation sets, selection of modelling technique, building model, evaluate the model and selection of the best model.

The classification engine development is the main targeted part of our study know as data mining and below in figure 3.4 we show the main stages for our data mining engine process.

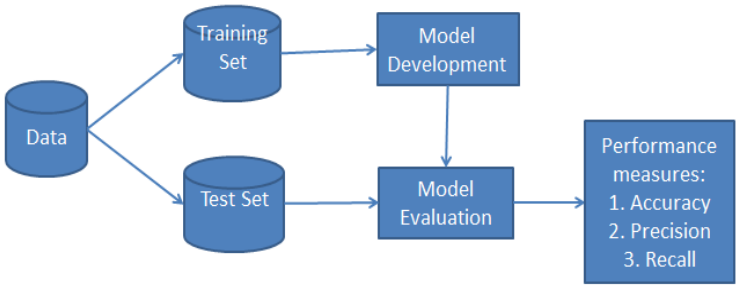

Figure 3.3: Engine Development Process

To Build VAT fraud detection model will use classification as supervised data mining learning which has aptness to identify and characterise each taxpayer into a category, it may belong to (fraudulent or legitimate. An experimental plan should first be set to guide the training, testing and evaluation process of the model. It is advisable to split the whole dataset into training and test sets with a large proportion of training data because the performance of your machine learning algorithm may decrease or increase based on the training data volume. In this study, we have used 80% of our data set to train our model and 20% for the testing process.

i)Model Selection

we have selected Naïve Bayes as our main classifier, and input corresponding to both training and testing phases will be initially processed. Naïve Bayes discover future insights based on prior experience. (Ahmed and Chung,2014) Using Bayesian probability terminology, the equation can be described as:

Posterior = (Prior * likelihood) / evidence, from the calculation formula as bellow:

P (X|Y) = (P (X)) * P (Y|X) / P (Y)

Where:

P (X): Hypothesis X's prior probability

P (Y): Training sample Y's prior probability

P (Y/X): Probability of Y given X(Likelihood)

P (X/Y): Probability of X given Y (Evidence)

The equation and formula described above they can be explained as such:

P (X): Hypothesis X's prior probability is the certain belief you have on a certain domain or situation.

P (Y/X): Probability of Y given X(Likelihood); this is representing the arguments that are supporting your prior belief after doing experiment, it is called also likelihood because it shows the degree of your experiment to be true or false.

P (Y): Training sample Y's prior probability is the prior predictor; means the set of arguments you have used to support your prior result.

P (X/Y): Probability of X given Y (Evidence) is your updated belief after considering the real effect of your class prior given to your predictor prior in your experiment.it is the posterior probability for the prior one; hence the formula **P (X|Y) = (P (X) * P (Y|X)) / P (Y)** is same as **Posterior = (Prior * likelihood) / evidence s**ince **P (X|Y) is Posterior, P (X) is the Prior.**

P (Y|X) is the likelihood and P (Y) is the evidence.

ii)Model training and prediction

the model training is the process of feeding to the selected data mining algorithm a large amount of relevant data and then teach it a certain thought-action process for learning good values from the labelled variables. Prediction refers to the result from the trained model when discovering the likelihood of a particular target variable in previously unseen data (Amazon, M. L.,017). For training our model, we will compare different classification algorithm like decision tree, kneighbors and naïve Bayes to find the best accuracy result.

3.2.7 Model evaluation

Model evaluation is an integral process of the model design aims to find the best model with generalisation accuracy on unseen data. The evaluation stage is where the data not seen by the trained model comes into play to give a picture of how the model will perform in real life.

The performance measures to evaluate our classification model are:

i)confusion matrix: is a table layout used to visualise and describe the performance of a classification algorithm on a set of unseen data for which the true values are known. "it contains values of true positive (correct classification) and false positives (incorrect classifications), and it gives you a better idea of what your classification model is getting right and what types of errors it is making" (Brownlee,2016).

Table 3.1: Confusion Matrix for evaluation

Confusion metric (standard metrics)		Predicted connection label	
		Not suspected	Fraud suspected
	Fraudulent	True Negative (TN)	False-positive (FP)
	Legitimate	False Negative (FN)	True positive (TP)

The key elements for a confusion matrix are explained as follows:

i) True Positive (TP): contains the number of fraudulent transactions that are correctly identified.

ii) True Negative (TN): shows the total number of non-fraud transactions that are correctly classified.

iii) False Positive (FP): highlight the total number of transactions with the wrong classification means suspect to be fraudulent, but in fact, they are not.

iv) False Negative (FN): is the total number of with improper classified records as legitimate, but in fact, they are fraudulent ones.

ii)Accuracy: the "Overall Classification accuracy (OCA) is the essential measure of the performance of a classifier. It measures how many observations, both positive and negative, were correctly classified. From the confusion matrix, we can say that accuracy is the percentage of correctly classified instances over the total number of instances in the total test dataset, namely the situation TP and TN "**(**Ahammed, Taheri and Ott,2010). The accuracy formula is the result explained below:

Accuracy = **[(TP+TN) *100%]/ (TP+TN+FP+FN)**

iii) Precision: Precision is the total number of class segments with correct classification over the total instances identified as class members by telling us when the model predicts positive, how far is it correct. The precision is obtained using the formula below:

Precision= (TP*100%) / (TP+FP).

iv)Recall: Recall is also called "True Positive Rate (TPR)". It used to measure the number of records with a correct classification relative to the total number of positive examples (Davis, and& Goadrich,2006)

Recall= (TP*100%)/ (TP+FN)

v) F1 score: F1 score is a combination of precision together with recall into one measure by calculating the balanced mean between those two and then the two on the positive class (Nicholson,2019)**.**

F1=2 * (Precision * recall)/(precision+recall)

vi) Receiver Operating Characteristics (ROC): ROC is a graph that allows visualising the interchange between TPR and FPR. "the higher TPR and the lower FPR is for each threshold, the better and so classifiers that have curves that are more top-left-side are better (Ekelund,2012).

3.2.8 Software Materials

Python, as a powerful and a friendly programming language, will be used in data exploration and analysis with the use of its built-in Libraries such as pandas, sci-kit learn and matplotlib etc. The following libraries will be used in our data analysis and for building the proposed model.

i) NumPy: NumPy is dynamic python's library most useful for the multidimensional array, and it helps to store large amounts of data.
ii) Pandas: it is known in full as Python Data Analysis Library, and it provides functions used to organise data, enhance code legibility, provide speed during data processing and allows Excel and comma-separated values (CSV) files to be accessed, be read and converted into a structured table format called DataFrames.
iii) Scikit-Learn: Scikit-learn is one among the most notable python's libraries which are helpful for data manipulation. It is mainly applied for feature selection, cross-validation and construction of confusion matrix during data modelling.
iv) Matplotlib: Matplotlib allows to create 2D data visualisation and to plot a variety of graphs such as bar, histogram, scatter and line graphs for a better result demonstration

3.2.9 Ethical Considerations

The goal of ethics in the study is to ensure that there are no harm adverse weights from the research activities (Akaranga and Makau,2016). It was agreed to keep the given data as confidential and use them only for academic purpose.

CHAPTER IV: DATA ANALYSIS AND RESULTS

This chapter presents the model development and experimentation results. It describes the data mining techniques used to help RRA to detect VAT fraud by first exploring patterns that exist between VAT data variables with the category status of taxpayer's compliance then develop a fraud detection model for VAT within three phases: (i) dataset pre-processing-processing, (ii) engine development to classify the compliance status for VAT's taxpayers, (iii) test and evaluate the designed model.

4.1 Business Understanding and Data Acquisition

The data we have used is a combination of three different data sources (EBM sales, VAT imports, VAT returns). The data attributes found in each data repository are the following with a detailed description, as seen in **Table 4.1.**

Table 4.1: List of attributes for VAT dataset.

EBM Sales

Attribute Names	Description of Attribute	Data type
TIN	seller identification number	Text
TAXTYPEDESC	tax description	Text
TAX PERIOD	Tax period	Number
INVOICE DATE	date of invoice	Date
ASSETS NO	transaction Number	Number
BUYERTIN	buyer identification number	Text
GOOD NATURE	item name	Text
INVOICENO	EBM invoice number	Text
VATTAX	VAT to be paid	Number
TOTALSALESAMOUNT	total sales	Number
EXEMPTSALESAMOUNT	Exempted sales	Number
ZERORATE	zero rate	Number
EXPORTAMOUNT	export amount	Number
VATSALE	vat on sales	Number

VAT Imports

Attribute Names	Description of Attribute	Data type
TIN	seller identification number	Text
TAXCENTRE	Taxpayer location	Text
TAXPERIODYEAR	Year of import	Number
TAXPERIODMONTH	Month and year for import	Number
CUSTOMSVALUE	transaction Number	Number
CUSTOMSVATPAID	buyer identification number	Number
CUSTOMSDECLARATIONNUMBER	item name	Text
CUSTOMSSTATION	Station name	Text
CUSTOMSDECLARATIONDATE	Import date	Date

VAT Declaration

Attribute Names	Description of Attribute	Data type
ANONYMIZEDTIN	seller identification number	Text
ENTGROUP	Enterprise group	Text
ENTDESC	Enterprise description	Text
BUSINESSDESC	Business description	Text
MAINENTACTVTYDESC	Activity description	Text
TAXCENTERTYPE	Business category	Text
TAXCENTER	Tax Centre	Text
TAXPERIOD	Declaration period	Number
ASSESSMENTNUMBER	Declaration number	Number
ASSMNTTYPESHORTDESC	Transaction description	Test
SUBMITDATE	Declaration date	Date
DUEDATE	Due date	Date
TAXTYPE	Taxtype description	Text
TOTALVALUEOFSUPPLIES	Total value	Number
EXEMPTEDSALESAMOUNT	Exempted sales	Number
ZERORATEDSALES	Zero rate sales	Number
EXPORTAMOUNT	Export amount	Number
TOTALNONTAXABLE	Nontaxable amount	Number
TAXABLESALESSUBJECTTOVAT	Taxable sales	Number
VATONTAXABLESALES	Vat on taxable sales	Number
VATREVERSECHARGE	Vat reverse charge	Number
VATPAYABLE	Vat payable	Number
VATPAIDIMPORT	Vat paid on import	Number
VATPAIDLOCALPURCHASE	Vat on local purchase	Number
TOTALVATPAIDONINPUT	Vat paid on input	Number
VATREVERSECHARGEDEDUCTIBLE	Vat deductible	Number

PREVIOUSMONTHCREDIT	Monthly credit	Number
VATWITHHOLDINGRETAINED	Vat retained	Number
VATDUECREDITPAYABLE	Vat payable on credit	Number
OTHERCREDITS	Other credit	Number
VATREFUNDPAID	Vat paid on the refund	Number
RUFUNDINTERESTADJUSTMENT	Refund interests	Number
VATREFUNDCLAIM	Vat refund claim	Number
VATDUE	Vat due	Number
VATRECEIVEDRRA	Vat paid	Number

4.2 Data cleaning

In this task, we have checked all the missing values and filling them for resolving inconsistencies. Most machine learning models require all features to be complete; therefore, missing values must be dealt with. After doing datasets integration we have to identify missing values and their provenance, and for the VAT dataset we have a missing value for:

i) TURNOVER, DECLAREDIMPORT and VATDUE due to lack of declaration

ii) TOTAL SALES due to lack of using EBM

iii) CUSTOMSVATPAID for those who have no importation

we have treated the existing missing data by Replace the NAN value with mean.

4.3 Data Integration

For being able to process the three different data sources, we have to integrate them to one single source. The initial data for EBM sales and importation are stored in real-time taxpayer's transaction by for VAT returns taxpayer and tax period group the data, and our historical data is in 4 years from 2016-2019.

Table 4.2: EBM real-time sales for a given taxpayer from EBM database

TIN	TAXPERIOD	INVOICENO	AMOUNT
901901128	2019 / 11	SDC007022805/1	4600847
901901128	2019 / 11	SDC007022805/3	9277118

Table 4.3: real-time importation for a given taxpayer from SINGLE WINDOW database

TIN	TAXPERIOD	CUSTOMNUMBER	CUSTOMSVATPAID
901901128	2019 / 11	C16220	490574
901901128	2019 / 11	C17266	499619
901901128	2019 / 11	C17460	491559

Table 4.4: VAT return transaction for a given taxpayer from ETAX database

ANONYMIZEDTIN	TAXPERIOD	TAXTYPE	TAXPAYERCATEGORY	TURNOVER	DECLAREDIMPORT
901901128	2019 / 11	VAT QUARTERLY	SMALL	13877965	1481752

We have grouped EBM and importation data by tin and tax period:

Table 4.5: EBM sales grouped by tin and tax period

TIN	TAXPERIOD	TOTALSALES
901901128	2019 / 11	13877965

Table 4.6: Importation grouped by tin and tax period

TIN	TAXPERIOD	CUSTOMSVATPAID
901901128	2019 / 11	1481752

After having the same format for our three sources of data, we have integrated them based on their common foreign key (TIN, TAX PERIOD), and we are still keeping constraint integrity.

Table 4.7: Integration for EBM sales, importation and VAT returns

TIN	TAXPERIOD	TAXTYPE	TURNOVER	TOTALSALES	DECLAREDIMPORT	CUSTOMSVATPAID
901901128	2019 / 11	VAT QUARTERLY	13877965	13877965	1481752	1481752

Our combined data source has **502,802** rows for four tax period **2016-2019**.

After understanding the provided RRA VAT datasets and investigate the similar data mining work published about fraud detection and given that Using Relational Database Management System (RDBMS) technology is limited as their SQL queries are slow, works as a database constraint without the ability to discover new facts, we, therefore, have devoted a significant striving in developing a classification model for VAT fraud detection able to mine patterns from VAT historical data and then flag a given transaction with a scientific basis for the intelligent as fraudulent or legitimate.

4.3 Data patterns and Relationship

4.3.1 Patterns extraction

our dataset contains 506,982 instances, and 19 columns contain numerical and categorical data. For **categorical data type**, we have **ten features**: (taxpayercategory, district, province, declaration category, declaration, EBM, importation, declarationebm, declaration import and fraud). Each VAT transaction has a class label that indicates whether the transaction is fraud or not. For **numerical data type**, we have **nine features**: (tin, tpyear, tpmonth, taxperiod, turnover, totalsales, declaredimport, customsvatpaid and vatdue).

Based on our main objective to detect VAT fraud from the historical VAT data, we have to dig deep and find hidden insight information that is relating to taxpayer compliance classification. Hence, we need to have a predefined hypothesis and based on them; we can know that a given transaction is fraudulent or not.

i) Taxpayer id identified by TIN (Taxpayer Identification Number) and to know if a given taxpayer has declared for a given period her/his turnover should be different to null.

ii) To know if a given taxpayer is using EBM TOTALSALES should be different from null.

iii) To know if a given taxpayer has imported some goods CUSTOMSVATPAID should be different from null.

iv) Where TOTALSALES > TURNOVER means the taxpayer has declared more than he has bought

v) Moreover, where DECLAREDIMPORT > CUSTOMSVATPAID means the taxpayer has declared more that she/he has imported.

With the use of those highlighted discovered information, we have extracted new patterns as described into the bellow summarised table 4.8**.**

Table 4.8: VAT Patterns

Generated Feature	Hypothesis	Feature Value
DECLARATION	TURNOVER is null with TOTALSALES >=0	**N**
	TURNOVER is null, and TOTALSALES is null and CUSTOMSVATPAID >=0	**N**
	TURNOVER>=0	**Y**
EBM	TOTALSALES >= 0	**Y**
	TOTALSALES is null	**N**
IMPORTATION	CUSTOMSVATPAID>=0	**Y**
	CUSTOMSVATPAID is null	**N**
DECLARATIONEMB	TOTALSALES > TURNOVER	**EBMSUP**
	TOTALSALES=TURNOVER	**EBMEQUAL**
	TOTALSALES<TURNOVER	**EBMINF**
	TOTAL SALES is null	**EBMZERO**
DECLARATIONIMPORT	DECLAREDIMPORT>CUSTOMSVATPAID	**IMPORTSUP**
	DECLAREDIMPORT=CUSTOMSVATPAID	**IMPORTEQUAL**
	DECLAREDIMPORT<CUSTOMVATPAID	**IMPORTINF**
	CUSTOMSVATPAID is null	**IMPORTZERO**
FRAUD	DECLARATION='Y', EBM='Y', IMPORT='N', DECLARATIONEBM='EBMSUP', DECLARATIONIMPORT='IMPORTZERO'	**Y**
	DECLARATION='Y' and EBM='Y' and IMPORT='Y' and DECLARATIONEBM='EBMEQUAL' and DECLARATIONIMPORT='IMPORTSUP'	**Y**
	DECLARATION='Y' and EBM='Y' and IMPORT='Y' and DECLARATIONEBM='EBMINF' and DECLARATIONIMPORT='IMPORTSUP'	**Y**
	DECLARATION='Y' and EBM='Y' and IMPORT='Y' and	**Y**

	DECLARATIONEBM='EBMSUP' and DECLARATIONIMPORT='IMPORTEQUAL'	
	DECLARATION='Y' and EBM='Y' and IMPORT='Y' and DECLARATIONEBM='EBMSUP' and DECLARATIONIMPORT='IMPORTINF	**Y**
	DECLARATION='Y' and EBM='Y' and IMPORT='Y' and DECLARATIONEBM='EBMSUP' and DECLARATIONIMPORT='IMPORTSUP	**Y**
	DECLARATION='Y' and EBM='N' and IMPORT='Y' and DECLARATIONEBM='EBMZERO’ and DECLARATIONIMPORT='IMPORTSUP	**Y**
	DECLARATION='N' and EBM='Y' and IMPORT='N' and DECLARATIONEBM='EBMSUP' and DECLARATIONIMPORT='IMPORTZERO	**Y**
	DECLARATION='N' and EBM='Y' and IMPORT='Y' and DECLARATIONEBM='EBMSUP' and DECLARATIONIMPORT='IMPORTSUP	**Y**
	DECLARATION='N' and EBM='N' and IMPORT='Y' and DECLARATIONEBM='EBMZERO' and DECLARATIONIMPORT='IMPORTSUP	**Y**
	DECLARATION='Y' and EBM='Y' and IMPORT='Y' and DECLARATIONEBM='EBMEQUAL' and DECLARATIONIMPORT='IMPORTEQUAL'	**N**
	DECLARATION='Y' and EBM='Y' and IMPORT='N' and DECLARATIONEBM='EBMEQUAL' and DECLARATIONIMPORT='IMPORTZERO	**N**
	DECLARATION='Y' and EBM='N' and IMPORT='Y' and DECLARATIONEBM='EBMZERO’ and DECLARATIONIMPORT='IMPORTEQUAL	**N**
	DECLARATION='Y' and EBM='N' and IMPORT='N' and DECLARATIONEBM='EBMZERO’ and DECLARATIONIMPORT='IMPORTZERO	**N**
	DECLARATION='Y' and EBM='Y' and IMPORT='N' and	**N**

	DECLARATIONEBM='EBMINF' and DECLARATIONIMPORT='IMPORTZERO	
	DECLARATION='Y' and EBM='Y' and IMPORT='Y' and DECLARATIONEBM='EBMINF' and DECLARATIONIMPORT='IMPORTINF	**N**
	DECLARATION='Y' and EBM='Y' and IMPORT='Y' and DECLARATIONEBM='EBMEQUAL' and DECLARATIONIMPORT='IMPORTINF	**N**
	DECLARATION='Y' and EBM='Y' and IMPORT='Y' and DECLARATIONEBM='EBMINF' and DECLARATIONIMPORT='IMPORTEQUAL	**N**
	DECLARATION='Y' and EBM='N' and IMPORT='Y' and DECLARATIONEBM='EBMZERO' and DECLARATIONIMPORT='IMPORTINF	**N**

Data mining is a knowledge discovery and with this exercising (known as feature engineering) of bringing out information from data gives us additional meaningful information to our dataset (declaration, EBM, importation, declarationebm, declaration import and fraud features). From TAXCENTER feature We have also extract geographical information district and province so that we can have a complete meaningful dataset.

4.3.2 Patterns relationship with Taxpayer status

By using Exploratory Data Analysis as a tool of storytelling, it helps us to understand our data how they are related to each other (independent variables to target variable). Our target variable is FRAUD, and the remaining are the independents' ones (TIN, TAXPAYERCATEGORY, DISTRICT, PROVINCE, DECLARTIONCATEGORY, TPYEAR, TPMONTH, TAXPERIOD, DECLARATION, EBM, IMPORTATION, TURNOVER, TOTALSALES, DECLAREDIMPORT, CUSTOMSVATPAID, DECLARARIONEBM, DECLARATIONIMPORT, VATDUE).

i) Fraud with TIN

COUNT(TIN)	CLASSIFICATION
28887	LG
13311	FR

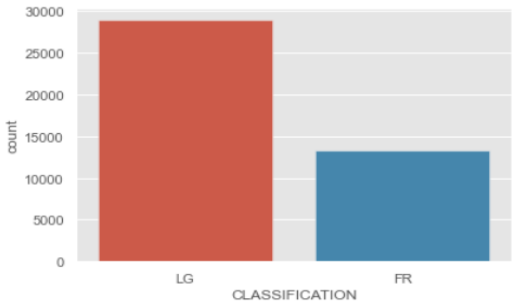

Figure 4.1: Fraud vs TIN

In our dataset, we have different TIN (number of taxpayers) of **29,129** within **28,887** are classed Legitimate (LG) and **13,311** are Fraudulent (FR).

ii) Fraud with Taxpayer Category

TAXPAYERCATEGORY	FRAUDULENT	LEGITIMATE
LARGE	10653.0	29325.5
MEDIUM	5448.0	11730.0
SMALL	41164.5	155170.0

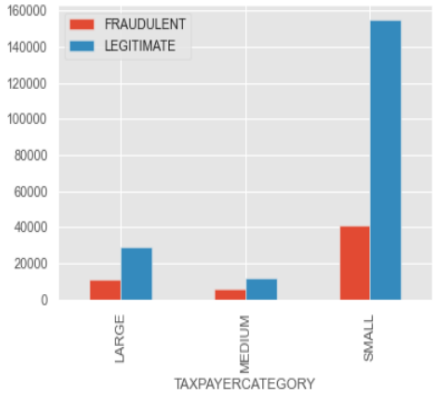

Figure 4.2: Fraud vs Taxpayer Category

For **taxpayercategory** we have three categories Large, Medium and Small. For Large category in a total of **39,979** of large's transactions, **10,653** are Fraudulent, and **29,326** are Legitimate. For Medium category in a total of **17,178** medium's transaction **5,448** is Fraudulent and **11,730** are Legitimate. For Small category with a total of **196,334** transactions, **41,165** are Fraudulent, and **155,170** are Legitimate. Given the statistics above we can say that more fraudulent cases are in a small category, and as the number of category's members increase the fraudulent cases also increase.

iii) Fraud with Tax period

TPYEAR	FRAUDULENT	LEGITIMATE
2016	11460.0	95617.0
2017	41782.0	79980.0
2018	14272.0	123085.0
2019	47017.0	93769.0

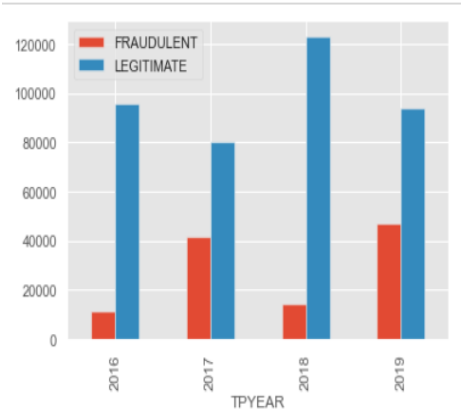

Figure 4.3: Fraud vs Tax Period

For tax period by a year, we find that the fraudulent cases have increased for the year 2019 with a total number of **47,017** cases and 2017 with total number of **41,782** cases, and this shows that fraudulent should be controlled; otherwise, it will increase more in the feature.

Maybe in 2017, some cases have been identified, and in 2018 some strategies were implemented which have reduced the fraudulent cases, and 2019 taxpayers changed their strategies also.

iv) Fraud with District

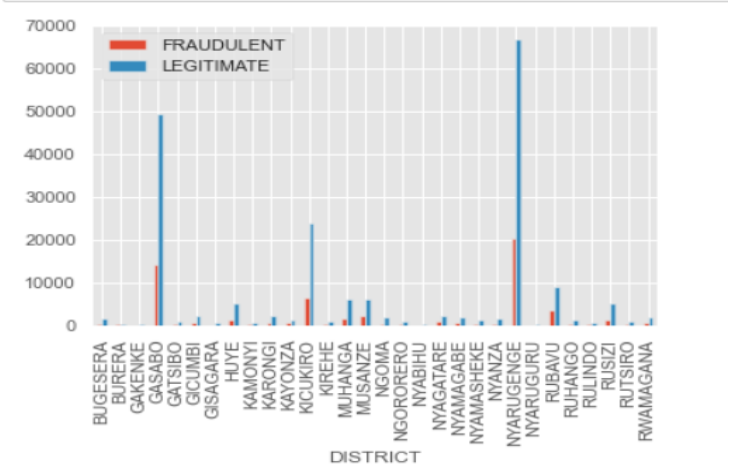

Figure 4.4: Fraud vs District

With **district** independent variable, we find out that RRA has to emphasise the control in Nyarugenge and Gasabo even though the number of legitimate is more.

v) Fraud with Province

PROVINCE	FRAUDULENT	LEGITIMATE
EAST	5993.0	21382.0
KIGALI CITY	81581.0	279581.0
NORTH	6199.0	18493.0
SOUTH	8342.0	33697.0
WEST	12416.0	39298.0

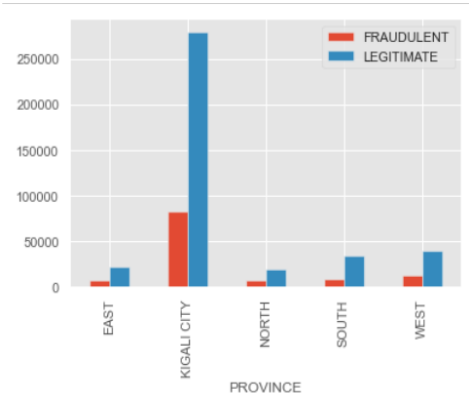

Figure 4.5: Fraud vs Province

For **province**, fraud is high in Kigali City with total number of 81581 cases, and we find an approximatively equal number of frauds in EAST (5993 cases) and NORTH (6199 cases).

vi) Fraud with Importation

IMPORTDECLARATION	LEGITIMATE	FRAUDULENT
IMPORTEQUAL	21043	17169
IMPORTINF	239	363
IMPORTSUP	0	531
IMPORTZERO	371169	96468

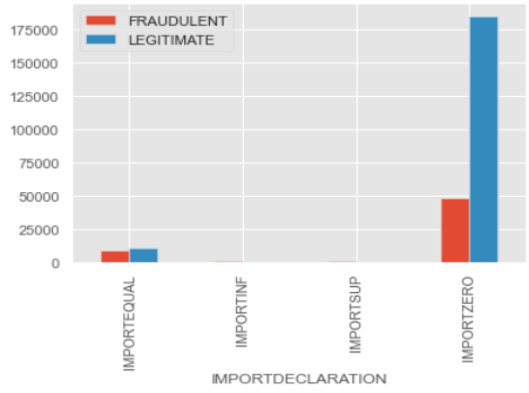

Figure 4.6: Fraud vs Importation

For importation variable we find that a higher number of fraudulent is noticed on the category of taxpayers who is not doing importation (**96,468** cases), those who declared a high volume of importation are definitively categorised in fraud (**512** cases) and for the ones who declare real volume of their imports we see that fraud (**17,169** cases) and legitimate (**21,043** cases) are approximately equal means even they declare importation correctly they are cheating in another side may be EBM sales.

vii) Fraud with EBM

EBMDECLARED	LEGITIMATE	FRAUDULENT
EBMZERO	233877	25
EBMINF	71555	142
EBMEQUAL	87019	91
EBMSUP	0	114273

Figure 4.7: Fraud vs EBM

For EBM variable, we have four groups, and a big number of fraudulent is **114,273** for those who declare a big volume of sales compared to their turnover.

4.4 Model development

4.4.1 Feature Selection

Data mining is based on a simple rule if you put garbage in, you will only get garbage as output. (Danubianu,2015) When you have a very large number of features it is more important to select those features that are important which will enable the learning algorithm to train faster while reducing the complexity, overfitting of a model, improve the accuracy with easy interpretation.

After a deep analysis of our dataset, together with the main objective of our research, the selected features process for our VAT fraud detection model has been done as follow:
FRAUD is the target variable, then TIN, TAXPERIOD, DECLARATION, EBM, IMPORTATION, DECLARATIONEBM and DECLARATIONIMPORT are the independent's variables.

The remaining independents variable we have to drop them from our dataset because some are duplicate for others like tpyear and tpmonth for them we keep tax period. Others are not contributing to our model building like taxpayercategory, declaration category, district, the province we have used them to understand the distribution for fraudulent transaction in different aspects. For turnover, totalsales, declaredimport, customsvatpaid from them, we have extracted meaningful attributes which will help us to illustrate the hypothesis of our model.

Table 4.9: VAT dataset after feature selection

	TIN	TAXPERIOD	DECLARATION	EBM	IMPORTATION	DECLARATIONEBM	DECLARATIONIMPORT	FRAUD
0	908638472	20165	Y	N	N	EBMZERO	IMPORTZERO	LG
1	908638472	20168	Y	N	N	EBMZERO	IMPORTZERO	LG
2	908638472	201611	Y	N	N	EBMZERO	IMPORTZERO	LG
3	908638422	20162	Y	Y	N	EBMEQUAL	IMPORTZERO	LG
4	908638422	20165	Y	Y	N	EBMEQUAL	IMPORTZERO	LG

4.3.2 Data Transformation

We have transformed some of our data in appropriate forms suitable for mining process as because most of data mining algorithms require the dataset's variables to have the same range of scale and be converted into numerical values for being able to work better.

i)Categorical Encoding

The majority features for our dataset are categorical ones, and we cannot mine them directly. We need to encode the Categorical Variable by replacing their values with numerical ones able to be interpreted in the machine learning equations.

Table 4.10: VAT dataset after categorical Encoding

TIN	TAXPERIOD	DECLARATION	EBM	IMPORTATION	DECLARATIONEBM	DECLARATIONIMPORT	FRAUD
908638472	20165	1	0	0	0	0	1
908638472	20168	1	0	0	0	0	1
908638472	201611	1	0	0	0	0	1
908638422	20162	1	1	0	1	0	1
908638422	20165	1	1	0	1	0	1

ii)Feature Scaling

Feature scaling enables to compare independent variables on common grounds after giving them the same limited ranges.

Table 4.11: VAT dataset before scaling

TIN	TAXPERIOD	DECLARATION	EBM
908638472	20165	1	0
908638472	20168	1	0
908638472	201611	1	0
908638422	20162	1	1

As illustrated in Table 4.11, you can see that the tin and tax period attributes do not have the same scale. If you compute tin and taxperiod values, the value for taxperiod will be dominated, which will also lead to inaccuracy result as variables with different range do not contribute equally to the learning process.

4.3.3 classification engine development

The fraud feature for our dataset is composed with two groups (**LG: Legitimate, FR: Fraudulent**) and from the total of **506,982** records, **392,451** are legitimate transaction, and **114,531** are fraudulent which means that legitimates returns are about **77.4%** and for Fraudulent is **22.6%** this shows that our target variable is imbalanced.

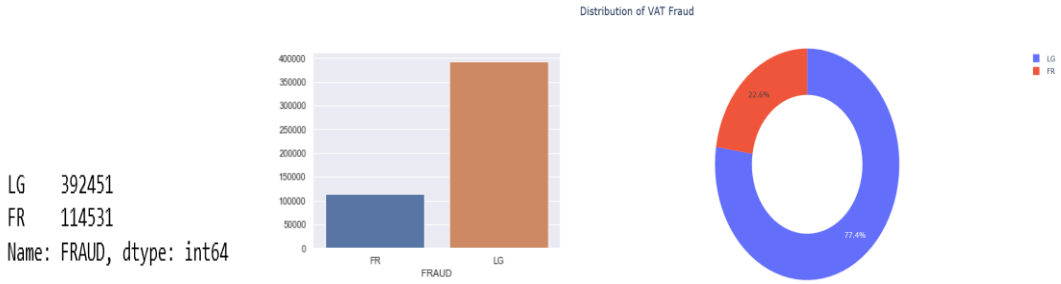

Figure 4.8: Imbalanced Fraud distribution

Imbalanced data refers to a state where, for a given dataset, total observations are not equally assigned to all defined classes. Most classifier learnings fail to cope with imbalanced set due to its sensitivity to have different proportions of the class's observation. As a consequence, the trained model tends to favor the class with the majority class, which will give misleading accuracy (Apostu,2020). Before classifying our data, we first have to deal with this imbalance dataset to avoid misclassification of our data. We have used over-sampling methods (Santacruz,2018) which classifies all instances as the majority class by randomly replicating the

number of instances in minority class to be increased while representing them with a higher figure with aims to have a balanced dataset.

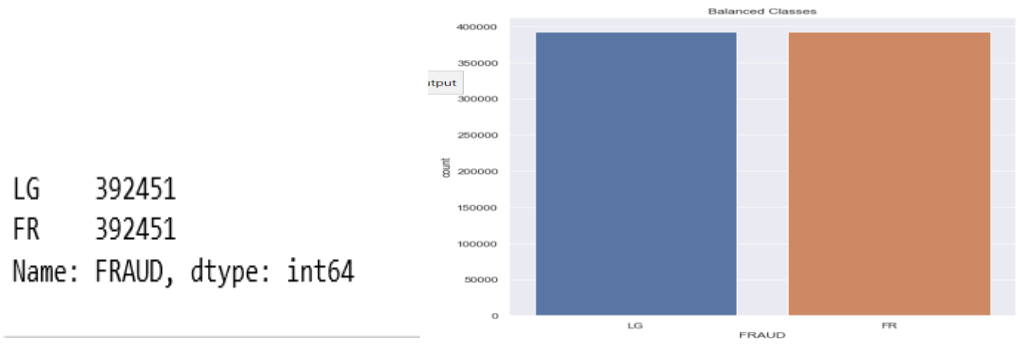

Figure 4.9: Oversampling VAT dataset for a Balanced Fraud distribution

The above bar plot shows that by applying the oversampling method, we have a balanced dataset.

i)Selecting of Models/Technique

according to our research objectives, we have developed a supervised classification model using different techniques first with naïve Bayes together with decision tree and Kneighbors.

For Naïve Bayes, we have used mixed Bayes due to our numerical and categorical data. We have combined Gaussian for numerical data together with multinomial Bayesian for categorical data.

ii) Organising Data into Sets

The ML model development requires separate sets (training and testing). (Mahani and Ali,2019) The whole lifecycle is data-driven because the model output using test data is linked with the trained dataset. We have allocated 80% of VAT dataset for training, and 20% were used for testing the model. To split data for training and testing, we have used the equation below:

"X_train, X_test, y_train, y_test = train_test_split (X, y, test_size = 0.2, random_state = 0) where:

X_train is the training part of the matrix of features.

X_test is the test part of the matrix of features.

y_train is the training part of the dependent variable that is associated with X_train here.

y_test is the test part of the dependent variable that is associated with X_train here".

iii)Model Creation

Model creation starts by initiating our selected model techniques, and by using the initiated model, we train the part for our training dataset to run smoothly and make it ready for prediction. With the use of our trained data, we can then make predictions using test data and get the accuracy of the model. “Making predictions involves calculating the probability that a given data instance belongs to each class, then selecting the class with the largest probability as the prediction” (Yadwadkar, Gonzalez and Hellerstein,2018).

iv)Measuring the accuracy

the confusion matrix has been used to calculate the correct number of predictions with the corresponding accuracy, as shown in Table 4.12.

Table 4.12: Accuracy measurement

Algorithm training/testing data size	**80%-20%**
Naïve Bayes	98%
Decision Tree	94%
Kneighbors	94%

Experimental result analysis shows that the test datasets are tested using the trained model are giving an excellent response. Accuracy table distinguishes between naive Bayes, decision tree and Kneighbors algorithm.

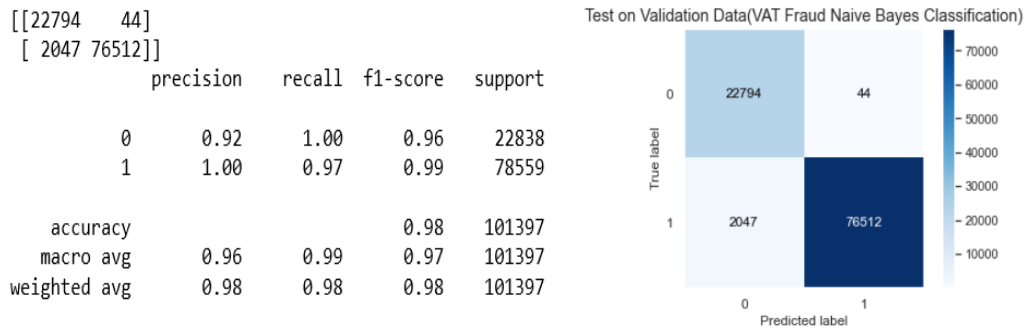

Figure 4.10: Naïve Bayes classification report and confusion matrix

With naïve Bayes, we can achieve accuracy for 98%.the classifier made a total of **101,397** predictions and out of those **101,397** cases naïve Bayes classifier predicted legitimate cases 76,512 times and fraudulent cases 22,794 times. The given result rank naïve Bayes as the best classifier in our experiment as the true predicted values are approximatively the same from our initial prior result **77%** for legitimate cases and **22%** for Fraudulent ones**.**

```
[[77061  1512]
 [ 8153 70255]]
              precision    recall  f1-score   support

         0.0       0.90      0.98      0.94     78573
         1.0       0.98      0.90      0.94     78408

    accuracy                           0.94    156981
   macro avg       0.94      0.94      0.94    156981
weighted avg       0.94      0.94      0.94    156981
```

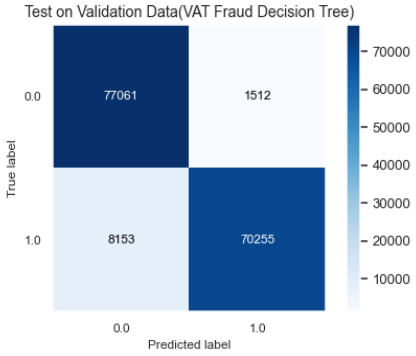

Figure 4.11: Decision tree classification report and confusion matrix

```
[[77061  1512]
 [ 8153 70255]]
              precision    recall  f1-score   support

         0.0       0.90      0.98      0.94     78573
         1.0       0.98      0.90      0.94     78408

    accuracy                           0.94    156981
   macro avg       0.94      0.94      0.94    156981
weighted avg       0.94      0.94      0.94    156981
```

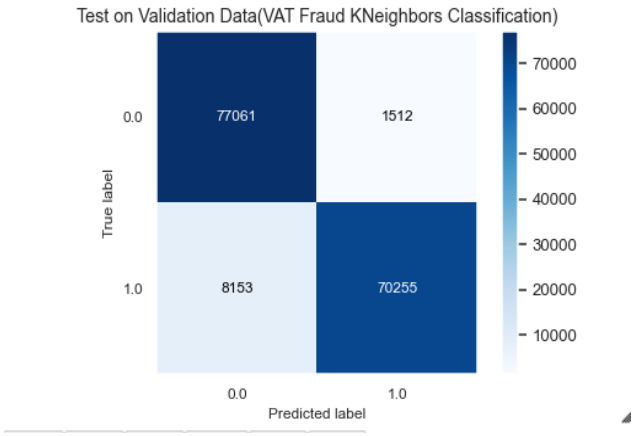

Figure 4.12: Kneighbors classification report and confusion matrix

With decision tree and Kneighbors the accuracy is the same with 94% of score.

For decision tree and Kneighbors the classifier made a total of **156,981** predictions and out of those **156,982** cases the both classifiers predicted **fraudulent** cases **77,061** times and **legitimate** cases 7**0,255** times. As seen this a wrong assumption because based on real situation we cannot have more fraudulent than legitimate cases.

4.3.4 Model Evaluation

To evaluate the developed VAT fraud detection model, we have used statistics metrics (accuracy, precision, recall and F1 score) together with receiver operating characteristic (ROC) so that we can have a real picture of how our model will work in the future.

i) Overall Performance Statistics

the accuracy, precision, recall and F1 Score result for VAT fraud detection model is described in Table 4.13.

Table 4.13: Overall performance

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.979378	0.999425	0.973943	0.986520
Decision Tree	0.938432	0.978932	0.896018	0.935642
KNeighbors	0.938432	0.978932	0.896018	0.935642

Comparing the performance statistics of all the model developed, as seen in the table above we see that naïve Bayes has the highest F1 Score (0.986520), precision (0.999425) and accuracy (0.979378). The Decision tree and kneighbors have the same scores. Almost all the models have an accuracy greater than 0.90.

ii) Receiver Operating Characteristic (ROC)

the ROC curves for our model is shown in Figure 4.14

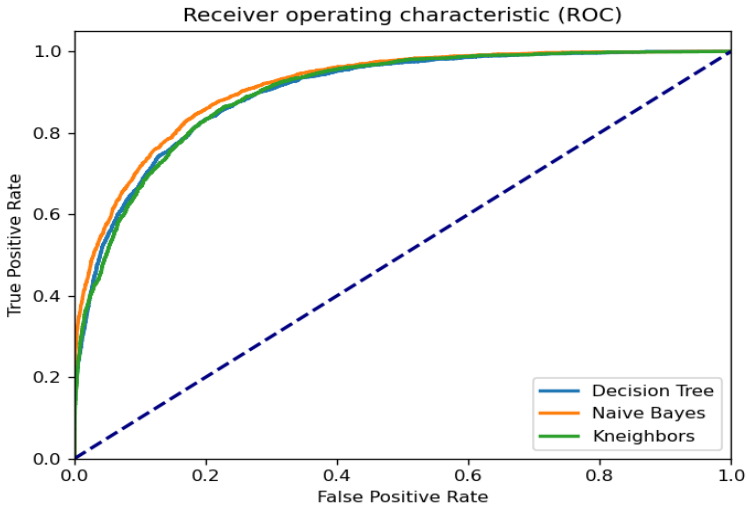

Figure 4.13: Receiver Operating Characteristic

From The plot above of the receiver operating characteristic curve for our three models; Naïve Bayes, Decision tree and Neighbors model, we can see that for Naïve Bayes, the ROC curve indicates the highest lift and is closest to the top left corner of the plot means one as TPR and 0 as FPR. The Naïve Bayes model's curve separates itself from the ROC curves of the other two models, which overlap with each other.

The Naïve Bayes classifier has been chosen as my preferred approach, to not only it has the highest accuracy, but also the highest precision and F-measure together with the correct assumption for which the true value is known via confusion matrix report. (Keogh,2006) The advantages of using Naïve Bayes over other models are that they are very simple to implement with the facility to handles both continuous and discrete data very fast with a highly scalable prediction. The below Figure 4.15 displays the histogram for our three classifier's accuracy.

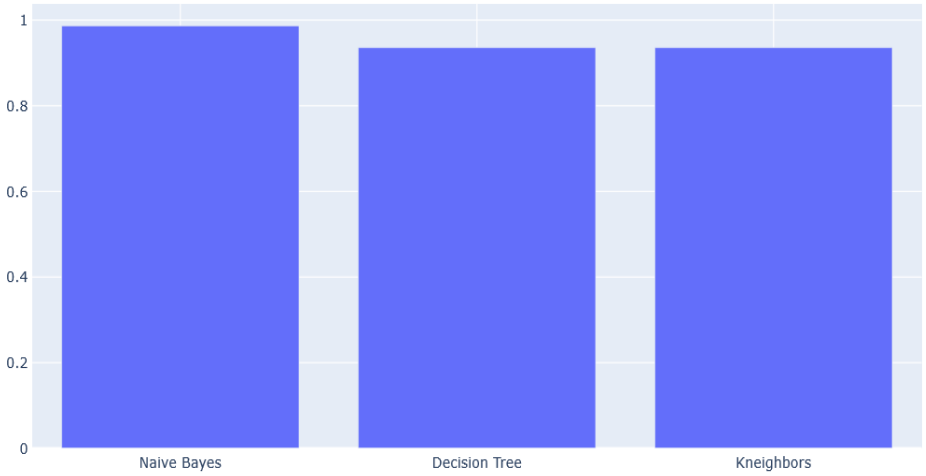

Figure 4.14: Histogram for our model evaluation

4.3.5 Summary

We have tried to deal with the VAT imbalance dataset by using the oversampling method then we parse and process the data with several things to clean it up:

i) We shuffled the data to ensure that there is no bias and the rows are not in a certain order
ii) We dropped any rows where the class is undefined.
iii) We encoded the nominal labels.
iv) We used a standard scaler in order to scale the data to a standard normal distribution.

v) Moreover, finally initiate, train, predict and evaluate our model by generating the classification report in which various statistics are displayed for helping us to judge our model and also a confusion matrix which give us a clear picture of the model accuracy and how our model is fitted.

We obtained fairly well accuracy in all the three classifier algorithms, and more attention needs to be given on the fraudulent data, which is much less than the correct one and which is overseen.

CHAPTER V: CONCLUSION AND RECOMMENDATION

5.1 Conclusion

Rwanda finds its resources in taxation like other countries. The responsibility of the RRA as a tax administration is to increase the compliance of the people in order to guarantee the revenues of public policies. Better use of tax audits saves the cost of resources while increasing the tax collection performance. As tax fraud becomes more sophisticated with significant growth of data, it becomes more difficult to discover fraudulent patterns from the bulk of data. Data mining experts may not abolish fraud but surely reduce it with the help of data mining to uncover hidden insight and deliver useful knowledge.

The developed approach for VAT fraud detection models to improve the risk-based selection has shown promise based on early results. The researcher used the secondary data and tried to meet the objective of the study by strictly following the six steps of data mining on collected data set and the result shows the business area is still at risk (22% of fraudulent in our study) and need for data mining tools and techniques to discover the hidden insights from existing data.

Our study has tested the suitability of various classifier algorithms (Naïve Bayes, Decision Tree and KNN) with data set arrangement and compared the results. All the three classifiers have more than 90% of accuracy. Moreover, the proposed Naïve Bayes with 98% of accuracy has an additional significant factor as a less time were required to train our model and given that the smaller patterns databases are used for fraud detection rather than the vast transaction data source.

The created model of compliance risk will give a room to taxpayers to whether to decide or not to comply with RRA with an increase of specificity. We also believe that the study and its implementation are useful for researchers from neighbouring fields.

5.2 Further research

i) In this work, we have only used VAT taxpayer's data, in the feature, it will be better to work with a whole vast amount of data by integrating more tax types and including more taxpayer's details.

ii) Personal Income Tax, together with Corporate Income Tax data, could be interesting while combined with VAT data for explanatory variables and will be used in future applications to the better performance and improvement of the model.

iii) Thus, the researcher believes that the application of other data mining techniques (rather than classification) with different algorithms in the new tool is potential research area to improve performance and compliance of taxpayer in the institution.

iv) It does not mean that are occurring only in E-filing, it also noticed in other disciplines within authority by auditors, staffs and experts. These can also be taken as another area for further research.

5.3 Recommendations

In the belief of the researcher, findings of the study will help the organisation, to work on the implementation of data mining techniques for the achievement of the targeted goals. The highlighted recommendations linked to our findings are forwarded:

i) Based on our study with 22% of VAT frauds shows that it should not be neglected and more risks have been noticed location wise in Kigali and taxpayer category wise more fraudulent are into small category.

ii) Taking account of the confidentiality of RRA taxpayer's data, some important details have been kept as secrets, and anonymized dataset have been provided. However, the usefulness of data mining application is not influenced thereof. As all attributes of VAT data have their respective influence, it seems impossible to implement comprehensive mining and summarization; RRA should try to avail all information in order to get the improved outcome.

iii) This research was conducted for academic purpose with constraint limit of time. To deploy the model into the institution some modification is required to come up with more comprehensive models, and it is recommended that RRA can conduct experimental tests with the inclusion of many datasets by using extensive training and testing datasets.

iv) RRA should continue in this type of study in the further to get better improvement fraud prediction and for other issues. Nowadays, fraud is complex and assorted.

REFERENCES

Alm, J. (2012). "Measuring, explaining and controlling tax evasion: lessons from theory, experiments and field studies" *Int Tax Public Finance 19,54-77(2012).* *https://doi.org/10.1007/s10797-011-9171-2*.

Adegbie and Jayeoba (2016). "Assessment of Value Tax on the Growth and Development Economy: Imperative for Reform" *Accounting and Finance Research, Vol.5, No.4,2016, SSRN: https://ssrn.com/abstract=3024642*.

Harrison, Graham and Krelove (2005). "Vat Refunds: A Review of Country Experience" *IMF Working Paper No.05/218*.

Mascagni, Giulia, Monkam and Nell Christopher (2016). "Unlocking the Potential of Administrative Data in Africa: Tax Compliance and Progressivity in Rwanda" *ICTD Working Paper 56, SSRN: https://ssrn.com/abstract=3120309*.

Mascagni, Giulia, Mukama and Santoro (2019). "An Analysis of Discrepancies in Taxpayers VAT Declarations in Rwanda" *ICTD Working Paper 92, Brighton, DS*.

Davia, H.R., Coggins, P.C, Wideman, J.C., & Kastantin, J.T. (2000). "Accountant's guide to fraud detection and control.

Mascagni, Giulia, Monkam and Nell Christopher (2016). "Unlocking the Potential of Administrative Data in Africa: Tax Compliance and Progressivity in Rwanda" *ICTD Working Paper 56, SSRN: https://ssrn.com/abstract=3120309*.

Promeranz, Dina (2013). "No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax" *Working Paper No.13-057*.

Yue, D., Wu, X., Wang, Y., Li, Y., & Chu, C.H (2007). "A review of data mining-based financial fraud detection research" *International Conference on Wireless Communications, Networking and Mobile Computing, pp.5519-5522*.IEEE.

Azuaje, F. Witten IH, Frank E (2006). "Data Mining: Practical Machine Learning Tools and Techniques" *2nd editon, BioMed Eng OnLine 5,51, https://doi.org/10.1186/1475-925X-5-51*.

Jans, M., Lybaert, N., & Vanhoof, K. (2007). "Data mining for fraud detection: Toward an improvement on internal control systems" *In European Accounting Association-Annual Congress (Vol.30)*.

Deshpande, D., & Deshpande, S (2007). "Analysis of Online User Behavior Detection Methodologies and its Evaluation" *International Journal of Compute Application,975,8887*.

Gullo, F. (2015). “From patterns in data to knowledge discovery: What data mining can do” *Physics Procedia,62,18-22*.

Bernus, P., & Noran, O. (2017). “Data rich-but information poor” *In Working Conference on Virtual Enterprises(pp.206-214)*.

Tsiptsis, K.K, & Chorianopoulos, A (2011). “Data mining techniques in CRM: inside customer segmentation.

Bramer, M (2007). “Principles of data mining” (Vol.180), London Springer.

Yongjian, F. (1997). “Data mining: tasks, techniques and applications” *IEEE Potentials,16(4),18-20*.

Ngai, E.W., Hu, Y., Wong, Y.H., Chen, Y., & Sun, X. (2011). “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature” *Decision support systems,50(3),559-569.*

Dhurjad, P.S., Marothu, V.K., & Rathod, R (2017). “Post-acquisition data mining techniques for LC-MS/MS-acquired data in drug metabolite identification” *Bioanalysis,9(16),1265-1278*.

Han, J., Kamber, M., & Pei, J. (2011). “Data mining concepts and techniques third edition” *The Morgan Kaufmann Series in Data Management Systems,5(4),83-124*.

Williams, G.(2006). “Data Mining Desktop Survival Guide” *http://www.togaware.com/datamining/survivor. Usage2.html*

Sugumaran, V., Geetha, T.V., Manjula, D., & Gopal, H. (2017). “Guest editorial: Computational intelligence and applications” *Information Systems Frontiers,19(5),969-974*.

Ahmed, M. (2019). “Data summarization Knowledge” *Information Systems,58(2),249-273*.

Dzhurenko, T., Myakshylo, O, & Cherednichenko, G. (2015). “Analysis of Text Mining methods in Web search” *Ukrainian Food Journal,4(3),508-519.*

Mascagni, Giulia, Monkam and Nell Christopher (2016). “Unlocking the Potential of Administrative Data in Africa: Tax Compliance and Progressivity in Rwanda” *ICTD Working Paper 56, SSRN: https://ssrn.com/abstract=3120309*.

Padhy, N., Mishra, D., & Panigrahi, R. (2012). “The survey of data mining application and feature scope” *arXiv preprint arXiv:1211.5723*.

Olson, D.L., & Delen, D. (2008). “Advanced data mining techniques” *Springer Science & Business Media*.

Khandare, N.B. (2016). “Credit card fraud detection using hidden markov model” *International Journal,1(4)*.

Behera, T.K., & Panigrahi, S. (2015). “Credit card fraud detection: A hybrid approach using fuzzy clustering & neural network” *Second International Conference on Advances in Computing and Communication Engineering (pp.494-499). IEEE*.

Da Silva, L.S., Rigitano, H., Carvalho, R.N., & Souza, J.C.F. (2016). “Bayesian Networks on Income Tax Audit Selection-A Case Study of Brazilian Tax Administration” *In BMA@ UAI (pp.14-20)*.

Boezio, B., Audouze, K., Ducrot, P. & Taboureau, O. (2017). “Network-based approaches in pharmacology” *Molecular informatics,36(10),1700048*.

Chen, S. (2016). “Detection of fraudulent financial statements using the hybrid data mining approach” *SpringerPlus,5(1),1-16*

Halsteinslid, E.L. (2019). “Addressing collinearity and class imbalance in logistic regression for statistical fraud detection” *Master’s thesis*.

Vasco, C.G., Rodriguez, M.J.D., de Lucas Santos, S., & de Madrid, U.A. (2018). “Characterization and detection of potential fraud taxpayers in Personal Income Tax using data mining techniques”.

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

Pérez López, C., Delgado Rodríguez, M. J., & de Lucas Santos, S. (2019) “Tax fraud detection through neural networks: an application using a sample of personal income taxpayers” *Future Internet,11(4),86*.

González, P. C., & Velásquez, J. D. (2013). “Characterization and detection of taxpayers with false invoices using data mining techniques” *Expert Systems with Applications,40(5),1427-1436*.

Cút, S. (2015).). “Risk assessment of VAT entities using selected data mining models”.

Gayathri, R., & Malathi, A. (2013). “Investigation of data mining techniques in fraud detection: credit card” *International Journal of Computer Applications,82(9)*.

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). “Comparing different supervised machine learning algorithms for disease prediction” *BMC medical informatics and decision making, 19(1), 1-16*.

Barros, R. C., De Carvalho, A. C., & Freitas, A. A. (2015). “Automatic design of decision-tree induction algorithms. Springer”

Zhou, W., & Kapoor, G. (2011). “Detecting evolutionary financial statement fraud. Decision” *Support Systems,50(3),570-575*.

Zhang, H., & Li, D. (2007) “Learning the naive Bayes classifier with optimization models” *IEEE International Conference on Granular Computing (GRC 2007) (pp.708-708).*

Taheri, S., & Mammadov, M. (2013). “Naïve Bayes text classifier” *International Journal of Applied Mathematics and Computer Science, 23(4).*

Korb, K. B., & Nicholson, A. E. (2010). “Bayesian artificial intelligence” *CRC press.*

Berrar, D. (2018). “Bayes’ theorem and naive Bayes classifier” *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands, 403-412*.

Segaran, T. (2007). “Programming collective intelligence: building smart web 2.0 applications” *O'Reilly Media, Inc*.

Al-Faiz, M. Z., Ali, A. A., & Miry, A. H. (2010). “A k-nearest neighbor-based algorithm for human arm movements recognition using EMG signals” *1st International Conference on Energy, Power and Control (EPC-IQ) (pp. 159-167). IEEE.*

Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). “Transforming big data into smart data: An insight on the use of the k-nearest neighbors’ algorithm to obtain quality data” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(2), e1289.*

Bryman, A. (2008). “Of methods and methodology Qualitative Research in Organizations and Management” *' An International Journal*.

Wirth, R., & Hipp, J. (2000). “CRISP-DM: Towards a standard process model for data mining” *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1). London, UK: Springer-Verlag.*

Siraj, F., & Abdoulha, M. A. (2007). “Mining enrolment data using predictive and descriptive approaches” *Knowledge-Oriented Applications in Data Mining, 53-72.*

Lelissa, T. B., & Kuhil, A. M. (2018). “The structure conduct performance model and competing hypothesis” *A review of literature. Structure, 9(1).*

Akaranga, S. I., & Makau, B. K. (2016). “Ethical Considerations and their Applications to Research: a Case of the University of Nairobi” *Journal of educational policy and entrepreneurial research, 3(12), 1-9.*

Thara, D. K., PremaSudha, B. G., & Xiong, F. (2019). “Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques” *Pattern Recognition Letters, 128, 544-550.*

Ahmed, I., Guan, D., & Chung, T. C. (2014). “Sms classification based on naive bayes classifier and a priori algorithm frequent itemset” *International Journal of machine Learning and computing, 4(2), 183.*

Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., & Cheung, D. (2009). "Naive bayes classification of uncertain data" *Ninth IEEE International Conference on Data Mining (pp. 944-949).*

Amazon, M. L. (2017).). "Amazon machine learning.

Brownlee, J. What is confusion matrix in machine learning (2016). "Machine Learning Mastery" *https://machinelearning mastery. com/confusion-matrix-machine-learning*

Ahammed, F., Taheri, J., Zomaya, A. Y., & Ott, M. (2010). "Using distance measurements to improve the accuracy of location coordinates in gps-equipped vanets" *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services (pp. 149-161).*

Mishra, A. (2019).). "Amazon Machine Learning" *Mach Learn AWS Cloud, 317-51.*

Davis, J., & Goadrich, M. (2006). "The relationship between Precision-Recall and ROC curves". *In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).*

Nicholson, C. (2019). "Evaluation metrics for machine learning accuracy, precision, recall, and F1 defined.

Ekelund, S. (2012). "Roc Curves—What are they and how are they used?" *Point of care, 11(1), 16-21.*

Danubianu, M. (2015). "Step by step data preprocessing for data mining. A case study" *In Proc. Of the International Conference on Information Technologies (InfoTech-2015) (pp. 117-124).*

Apostu, S. (2020). "Using machine learning algorithms to detect frauds in telephone networks" *The Annals of "Dunarea de Jos "University of Galati. Fascicle III, Electrotechnics, Electronics, Automatic Control, Informatics, 43(3), 16-20.*

Santacruz, A. (2018). "Why it is important to work with balanced classification dataset.

Garcia, R., Sreekanti, V., Yadwadkar, N., Crankshaw, D., Gonzalez, J. E., & Hellerstein, J. M. (2018). "Context: The missing piece in the machine learning lifecycle" *In KDD CMI Workshop (Vol. 114).*

Keogh, E. (2006). "Naive bayes classifier" pp. *16-21.*

Mahani, A., & Ali, A. R. B. (2019 "Classification problem in imbalanced datasets" *Recent Trends in Computational Intelligence, 1-23.*

PLAGIARISM REPORT

Submission date: 12-Sep-2020 02:33PM (UTC+0300)
Submission ID: 1385241937
File name: inal_draft_-Master_s_Dissertation_MUNEZERO_ClaudineVersion2.docx (1.1M)
Word count: 14062
Character count: 78271

munezero final dissertation

ORIGINALITY REPORT

11%	8%	5%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	research.ijcaonline.org Internet Source	1%
2	docplayer.net Internet Source	1%
3	library.iugaza.edu.ps Internet Source	<1%
4	www.jatit.org Internet Source	<1%
5	www.wikicoursenote.com Internet Source	<1%
6	Submitted to University of Stirling Student Paper	<1%
7	Jellis Vanhoeyveld, David Martens, Bruno Peeters. "Value-added tax fraud detection with scalable anomaly detection techniques", Applied Soft Computing, 2020 Publication	<1%
8	"Proceedings of International Ethical Hacking Conference 2018", Springer Science and	<1%

Business Media LLC, 2019
Publication

9	umexpert.um.edu.my Internet Source	<1%
10	eprints.uthm.edu.my Internet Source	<1%
11	repository.out.ac.tz Internet Source	<1%
12	Submitted to CVC Nigeria Consortium Student Paper	<1%
13	hdl.handle.net Internet Source	<1%
14	ethesis.nitrkl.ac.in Internet Source	<1%
15	Submitted to Yonsei University Student Paper	<1%
16	Submitted to Cardiff University Student Paper	<1%
17	Submitted to University of Strathclyde Student Paper	<1%
18	ccsenet.org Internet Source	<1%

19	"Data Mining and Big Data", Springer Science and Business Media LLC, 2016 Publication	<1%
20	usir.salford.ac.uk Internet Source	<1%
21	Submitted to Universiti Teknologi Malaysia Student Paper	<1%
22	Submitted to University of Southampton Student Paper	<1%
23	experfy.com Internet Source	<1%
24	dspace.unipampa.edu.br:8080 Internet Source	<1%
25	c4i.gmu.edu Internet Source	<1%
26	Eghbal Rahimikia, Shapour Mohammadi, Teymur Rahmani, Mehdi Ghazanfari. "Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran", International Journal of Accounting Information Systems, 2017 Publication	<1%
27	Gonzalo Mariscal, Óscar Marbán, Covadonga Fernández. "A survey of data mining and	<1%

knowledge discovery process models and methodologies", The Knowledge Engineering Review, 2010

Publication

28	Submitted to London School of Marketing Student Paper	<1%
29	Submitted to University of Northumbria at Newcastle Student Paper	<1%
30	Submitted to Maastricht School of Management Student Paper	<1%
31	Submitted to National College of Ireland Student Paper	<1%
32	Submitted to Institute of Technology Blanchardstown Student Paper	<1%
33	educationalresearchtechniques.com Internet Source	<1%
34	Submitted to BRAC University Student Paper	<1%
35	es.scribd.com Internet Source	<1%
36	repository.unib.ac.id Internet Source	<1%
37	Submitted to Bournemouth University Student Paper	<1%
38	dzone.com Internet Source	<1%

39	Submitted to University College London Student Paper	<1%
40	Charalambos Themistocleous, Marie Eckerström, Dimitrios Kokkinakis. "Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks", Frontiers in Neurology, 2018 Publication	<1%
41	Submitted to Deakin University Student Paper	<1%
42	Tsiptsis. "An Overview of Data Mining Techniques", Data Mining Techniques in CRM, 01/15/2010 Publication	<1%
43	Charikleia Chatzaki, Matthew Pediaditis, George Vavoulas, Manolis Tsiknakis. "Chapter 7 Human Daily Activity and Fall Recognition Using a Smartphone's Acceleration Sensor", Springer Science and Business Media LLC, 2017 Publication	<1%
44	www.cactusindustrial.com Internet Source	<1%
45	S. Brindha, K. Prabha, S. Sukumaran. "A survey on classification techniques for text mining", 2016 3rd International Conference on Advanced Computing and Communication Systems	<1%

(ICACCS), 2016
Publication

46	www.ijsret.org Internet Source	<1%
47	neptune.ai Internet Source	<1%
48	www.eee.metu.edu.tr Internet Source	<1%
49	repository.up.ac.za Internet Source	<1%
50	www.researchsquare.com Internet Source	<1%
51	Submitted to Universiti Teknologi MARA Student Paper	<1%
52	Submitted to Kingston University Student Paper	<1%
53	worldwidescience.org Internet Source	<1%
54	Submitted to South Bank University Student Paper	<1%
55	pdfs.semanticscholar.org Internet Source	<1%
56	id.123dok.com Internet Source	<1%

57	archive.org Internet Source	<1%
58	aut.researchgateway.ac.nz Internet Source	<1%
59	Hisham ElMoaqet, Mohammad Eid, Martin Glos, Mutaz Ryalat, Thomas Penzel. "Deep Recurrent Neural Networks for Automatic Detection of Sleep Apnea from Single Channel Respiration Signals", Sensors, 2020 Publication	<1%
60	ir.jkuat.ac.ke Internet Source	<1%
61	caia.swin.edu.au Internet Source	<1%
62	dspace.pacuniversity.ac.ke:8080 Internet Source	<1%
63	uniassignment.com Internet Source	<1%
64	studentsrepo.um.edu.my Internet Source	<1%
65	news.mak.ac.ug Internet Source	<1%
66	"Intelligent Computing, Networking, and Informatics", Springer Science and Business	<1%

Media LLC, 2014
Publication

67	Submitted to International Islamic University Chittagong Student Paper	<1%
68	www.dtic.mil Internet Source	<1%
69	"Advanced Data Mining Techniques", Springer Science and Business Media LLC, 2008 Publication	<1%
70	ir.msu.ac.zw:8080 Internet Source	<1%
71	Submitted to University of West London Student Paper	<1%
72	Cakir, A.. "Data mining approach for supply unbalance detection in induction motor", Expert Systems With Applications, 200911 Publication	<1%
73	Mohd Jawad Ur Rehman Khan, Anjali Awasthi. "Machine learning model development for predicting road transport GHG emissions in Canada", WSB Journal of Business and Finance, 2019 Publication	<1%
74	"Trends in Applied Knowledge-Based Systems and Data Science", Springer Science and	<1%

On