PREDICTION OF SOUTH AFRICAN CRIME RATE USING SUPERVISED MACHINE LEARNING TECHNIQUES

by

Mutobvu Ronewa

MINI-DISSERTATION

Submitted in fulfilment of the requirements for the degree of
MASTER OF SCIENCE
in

E-SCIENCE

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)
at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: DR T I MODIPA

2023

DEDICATION

I would first like to dedicate this thesis to four beloved people who have meant and still mean a great deal to me. Even though they are no longer among us, I still live by the memories of them. I want to dedicate this to my father, Fannie Mutobvu. Since you departed when I was still very young, I didn't have nearly enough time with you, yet you still inspire almost every element of my life. Thank you, "Bakali." Next is my maternal grandfather Elias Ramasala, who taught me the value of hard work and had an unending amount of love for me. I really appreciate you, Mushavhi "Ishe," and I won't ever forget you. May you find peace and pleasure in paradise! I'm dedicating this to my maternal uncle Ramasala Edward "Duki," who reared me, loved me, and filled a father figure gap before having his life tragically cut short in a taxi violence incident. Last but not least, I dedicate this to my late baby brother Wavhudi Mutobvu, who is no longer in our loving eyes and has left a vacuum in our hearts that will never be filled. Despite the shortness of your life, I will make sure that your memory is preserved for as long as I am alive. Beyond words, I miss each one of you.

DECLARATION OF AUTHORSHIP

I Mutobvu Ouma Ronewa declare that prediction of south African crime rate using supervised machine learning techniques hereby submitted to the university of Limpopo , for the degree of masters of science in e-science has not previously been submitted by me for a degree at this or any other university; that it is my work in design and in execution, and that all material contained herein has been duly acknowledged.

Mutobvu R O	14 Apr. 23
Surname, Initials (title)	Date

ACKNOWLEDGEMENT

First and foremost, I want to acknowledge my supervisor, Dr. T. I. Modipa, for his ongoing support and insightful guidance on this project. I also want to thank my mother, Anna Mutobvu, and my entire family for their love, support, and constant encouragement to work hard and care for my mental health. Additionally, I would like to express my gratitude to Mamorare Thantsha and Rofhiwa Ramasala for their support during the most trying period of my life. Without you two, this thesis would not have been a success. I will always be grateful to you guys for your unwavering encouragement, support, and tolerance during this journey. Finally, I want to thank The National e-Science Postgraduate Teaching and Training Platform (NEPTTP) for helping me out financially this year.

ABSTRACT

Kaggle crime statistics for South Africa were used to create machine learning categorization models. Although the techniques used in the experiments that came before this one differed, the dataset that was used was. The accuracy of other previous studies conducted on different datasets from this one and compared during the experiment stage were utilized to identify the three classification algorithms employed in this study. The study chose to use the random forest, K-nearest neighbor, and Naive Bayes classifier models. The Python-based algorithms were trained on a pre-processed crime dataset. Data preparation and processing, missing value analysis, exploratory analysis, and finally model construction and evaluation made up the analytical process. The best model should be chosen in accordance with the results. In both approaches, RF is outperforming the other models. According to the study's evaluation of both metrics and logloss, RF appears to be doing better.

TABLE OF CONTENTS

ABBREVIATIONS

ML ……..... Machine Learning

SA ……..... South Africa

RF ……..... Random Forest

KNN ……… K-Nearest Neighbor

SAPS……... South Africa police Service

NB………… Naive Bayes

RF………… Random Forest

DT………… Decision Tree

SVM…........Support Vector Machines

MR………… Multiple Regression

LR…………. Linear regression

LogR………. Logistic regression

Stats SA…...Statistics South Africa

SML………. Supervised Machine Learning

EDA………. Exploratory Data Analysis

TP………… True Positive

TN………… True Negative

FP………… False Positive

FN………… False Negative

NN………… Neural Network

CNN………. Convolutional Neural Network

LSTM……… Long Short-Term Memory

CSIR………. Council for Scientific and Industrial Research

IACP………. Association of Chiefs of police

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1. PROBLEM STATEMENT

South Africa (SA) is one of the still-developing countries in Africa; however, it is known for being of the country with high crime rate. Breetzke (2010) stated that Crime rates in SA after apartheid are high and increasing. SA distinct socio-political background and a system of inefficient social methods of control, which suggest significant degrees of social disorganization within some areas, are frequently stated as explanations for these high rates of crime. Other arguments emphasize the existence of disgruntled youngsters, deprivation, and the quick immigration of immigrants from neighbouring African nations into SA (Crush et al., 2013).

Crime rates across much of the nation have been growing since the mid-1980s (Schönteich and Louw, 2001). According to SAPS (2003) the numbers reported by the police after 1994 suggest that recorded crime in South Africa has grown by 30% over the preceding decade and stated the violent crime records shows a significant increase compared to any other crime category (by 41% compared to 28% for property crime). As of today, the crime index report in Africa shows that out of 20 crime hotspots countries in Africa, SA is at the top of the list of countries with the highest crime rate in Africa, with a crime index of 77.07% and a safety index of 22.93%.

Obagbuwa and Abidoye (2021) classified SA as one of the most dangerous, homicidal, and violent places across the globe and highlighted social violence and homicide as the two elements that place SA at the top of the list of crime ranks. However, since these numbers omitted criminal incidences in the apartheid-era they are commonly viewed as erroneous. Previous literature suggests that the crime rate has not been stable. The crime rate in South Africa sometimes significantly drops, and sometimes, the crime rate notably rose; for instance, the overall crime rate in 2002/03 peaked and years after presented a gradual decline of crime rate until it started to increase again in 2007/2008 (Bhorat *et al.*, 2017).

In 2021 SA police officials reported a high increase of 8.9% in murder cases between October and December (SAPS,2022). Despite the investigations around predicting the crime rate, there seem to be underlying issues with law enforcement officials predicting the crime rate accurately. The law enforcement officials release crime rate statistics annually; however, predicting future crime accurately seems to be still an issue due to human error and other underlying contributing factors. According to Schneider (2002) One can observe that predicting future crime is getting more and more important when comparing the amount of information available in earlier periods with that created throughout the 1990s.

The ability to produce precise, long-term projections of the character and extent of criminality is likely to show to be very challenging, notwithstanding ongoing improvements to the increasingly scientific prediction techniques. and the study also highlighted that examining previous predictions on impending crime rates shows issues with reliable predictions. Obagbuwa and Abidoye (2021) argue that the studies conducted using the SA dataset on crime rate predictions using a machine learning approach were not adequately considered; however, the traditional solving crime approach was mostly considered.

In the current context of crime increasing rapidly, traditional solving crime approaches are unable to deliver results, it is being slow paced and less efficient (Shah et al., 2021). There is a need to assist law enforcement officials in determining the possible changes that need to be implemented, ways to avoid crimes, and the factors that could be contributing drastically to the high crime rate in South Africa. According to shah et al (2021) states that every few years, police deploy new technologies like facial recognition software and sting rays. These facial recognition software and sting rays can fundamentally alter how the work is conducted for the better.

The importance of using machine learning and deep learning technologies in addressing crime in SA represent a big step forward in improving the efficient and effectiveness of law enforcement and crime prevention strategies. The international Association of Chiefs of police (IACP) (2016) has acknowledged the potential benefits of predictive policing, stating that it can enhance proactive strategies and tactical decision making in the fight against crime.

Similarly, the Council for scientific and Industrial Research (CSIR) (2018) has highlighted the importance of intelligent surveillance and response systems in improving the efficient and effectiveness of law enforcement. By leveraging the power of these technologies to analyse the data, detect criminal activities , and developing more objective sentencing guidelines, SA has the potential to make significant progress in reducing crime rates and creating a safer society (Maverick., 2019).

According to Forbes (2018) SAPS is exploring the use of predictive policing to anticipate and prevent crime using ML algorithms to analyse crime data and identify trends and patterns that can help police predict where and when crimes is likely to occur.

The goal of this project is to employ machine learning methods to accurately anticipate crime rates in South Africa because these methods are good at forecasting the future from historical data (Prithi et al., 2020). However, Vaidya et al. (2012) noted that despite the use of various modern technologies, the crime rate has not greatly lowered. While it is difficult to predict which perpetrators will actually conduct the crime, it is possible to make an educated guess as to where and how likely it is to happen.

According to a study on crime predicting conducted by Shah et al in 2021, combining ML with computer vision can considerably increase the overall effectiveness of police forces. By combining machine learning and computer vision with security equipment like surveillance cameras and spotting scopes, a machine will soon be able to accurately predict future crimes without the need for human intervention. A potential automation would be a system that can predict and foresee the locations of a city that are likely to be crime concentrations. Police officers can be warned and crimes can be avoided by boosting monitoring within the warning zone. The study examines supervised machine learning methods to identify the method that will help officials by producing the most accurate findings.

1.2. AIM

The study aims to predict crime rates using supervised machine learning techniques from the South African crime and population statistics datasets.

1.3. OBJECTIVES

The objectives of this study are to:

i. Perform data exploration analysis using crime and population statistics.

ii. Data pre-processing.

iii. Develop classification models using KNN, NB, and RF techniques.

iv. Evaluate the classification accuracies of the models.

1.4. RESEARCH QUESTIONS

i. Which are the suitable techniques to perform data exploration?
ii. How do we effectively pre-process the dataset?
iii. What are the practical approaches to developing machine learning-based models?
iv. How do we evaluate the classification accuracy of the models?

1.5. SIGNIFICANT OF THE STUDY

The study on the use of machine learning technique in addressing crime in South Africa is significant. Firstly, South Africa has one of the highest crime rates in the world, which poses a significant threat to public safety and security. Therefore, any efforts to develop more effective crime prevention and law enforcement strategies are of utmost importance.

Secondly, machine learning and deep learning technologies have shown great potential in addressing crime in various contexts, such as forecasting crime hotspots, identifying crime trends, and sending early alerts to law enforcement authorities. However, their application in the specific context of South Africa has not been extensively explored. Therefore, this study seeks to fill this gap in knowledge and provide valuable insights into the potential of these technologies in addressing the crime problem in South Africa.

Thirdly, the study can inform the development of more objective and evidence-based sentencing guidelines, which can help ensure fairness and consistency in the criminal justice system. This can ultimately lead to greater trust in the system and a more just society. Fourthly, the study can also have broader implications for the development of smart cities in South Africa, where machine learning and deep learning technologies can be used to evaluate big data for crime prevention and enhance public safety.

Finally, the study can contribute to the growing body of literature on the use of machine learning and deep learning technologies in addressing crime globally. This can help inform future research in this area and lead to the development of more effective crime prevention and law enforcement strategies worldwide.

1.6. DISSERTATION OUTLINE

This study is organized into five chapters, each of which is presented in the order shown below:

Chapter 1: Introduction of the study

The reason behind the study, goals, problem, and aim are all discussed in this chapter with clarity. The concepts and analysis of the SA crime stats in comparison to other nations are introduced here.

Chapter 2: Literature review

This chapter encompasses the detailed related work on the methodologies deployed, the nature of crime rates in SA over the years and how other studies used ML to attain the final findings.

Chapter 3: Methodology

This study's chapter illustrates the methods used for the investigation. The chapter describes how the dataset looks, the data exploratory, and the preparation of data before applying ML techniques. It outlines the classification models used, technologies and finally detailed models’ evaluations.

Chapter 4: Results and Discussion

The findings from carrying out the experiment are discussed in this chapter.

chapter 5: Conclusion

This chapter contains the summary of the dissertation and the final reasoning of the findings. The future recommendation and significance of the study is provided in this chapter

1.6 SUMMARY

This part provides an introduction explaining why the study's focus is on crime rates, how they have impacted South African communities, how they increase and decrease, and how SA is ranked globally and among other African nations. This section also describes how the study will be conducted to accomplish the eventual goal of using SML to predict crime rates.

CHAPTER 2 LITERATURE REVIEW

2.1. INTRODUCTION

The crime studies are done worldwide and within South Africa; however, the techniques vary. Previous literatures done by Osisanwo *et al.*, 2017highlights the various utilisation of different approaches like traditional approach where the study is analysed based on socio-economy factors and the used of machine learning approach; machine learning either unsupervised or supervised learning.

The likes of Schneider (2002), Schonteich & Louw (2001), and Obagbuwa and Abidoye., (2021) used machine learning approach and Shaw (1997), Bhorat et al. (2017), and other literature included the traditional approach. This study followed the supervised machine learning approach to assess accuracy for crime prediction. The supervised machine learning approach was used in studies by Schneider (2002), Schonteich & Louw (2001), Shah et al (2021), and Obagbuwa and Abidoye (2021).

To achieve the goal, different machine learning approaches were applied. Schneider (2002), Schonteich & Louw (2001), and Shah et al (2021) used predictive classification while utilising combined a review of other studies dataset and data on the types of crimes that are most prevalent in the country, and the study used a range of data sources to compile their report, including crime statistics from the South African Police Service, surveys of victims of crime, and interviews with criminal justice professionals, respectively. Obagbuwa and Abidoye (2021) used a regression technique with crime data from the South African Police Service to analyse and predict crime trends in South Africa. In their machine learning technique, some of these investigations used comparative algorithms where the algorithms are being compared amongst each other to see which algorithm performs better.

Isafiade, Ndingindwayo, and Bagula (2021) used predictive policing model using deep learning techniques in a South African township. The study highlighted the need for predictive policing technology to be ethically sound, transparent, and accountable, and argue that deep learning algorithms can be used to develop more accurate predictive policing models. Isafiade, Ndingindwayo, and Bagula (2021) used historical crime data to train their predictive policing model. Specifically, they used data on reported crimes in a South African township, including information on the location, date, and time of each reported crime.

In the framework of worldwide comparative studies, with South Africa at the centre of this comparison, according to Schneider (2002), most of the crime prediction has come from the United States and Great Britain, and to a lesser degree, Australia. The following sections includes the literature on criminal behaviours, few factors contributing towards criminal behaviour, crime, and its history through all the eras in SA, the reasoning behind this study, the literature on the machine learning algorithms as well as the literature on the algorithms used in this study.

Some criminology studies such as McCafferty and Action (2003), Schonteich & Louw (2001) and many other   assessed the contributing factors towards the rise of crime rates in the apartheid and in democratic era as well as assessing when the crime rates spiked in these eras. According Schonteich & Louw (2001) apartheid and internal security regulations were included into South African law between the 1950s and the late 1980s. The state's security personnel committed crimes in their zeal to suppress the opposition to the government. Violent actions were frequently viewed and justified by those who committed them as a reasonable line of defence against political opponents and adversaries.

As a result, Schonteich & Louw (2001) highlighted that violence was increasingly employed in society to promote personal and political interests and recorded crime grew by roughly 35% in the 20 years leading up to 1993. While crime rates remained essentially stable in the 1970s, it surged considerably in the early 1990s. When there is political turmoil, instability, and violence, criminality tends to grow. Regular police activities are oriented toward suppressing violence during times of disruption, which leads to an increase in crime in the apartheid era.

Another study done by Shaw (1997) also highlighted that the growth in crime rates peaked and was at its highest in 1990, the year the political change began, lending support to this viewpoint. According to statistics, the total number of crimes that were reported between 1990 and 1994 increased significantly. Most crimes significantly rose over this period, including rape by 42%, assault with 18%, burglary with 20%, vehicle theft with 34%, and robbery with 40%.

According to Plessis and Louw (2005), The nation changed from an autocratic and authoritarian oligarchy to a constitutional democracy founded on human rights in 1994. The state's institutions and laws needed to be thoroughly reviewed in light of this fundamental change in the state's nature and operation. The new government faced a transitional and nation-building challenge at a time when crime rates and the general sense of fear were getting close to intolerable levels. Simply because the data didn't include criminal incidents in the apartheid-period "Bantustans," the apartheid era was ignored in the analysis.

While a study done on crime in South Africa by Schonteich et al. (2001) noted that while reported crime rates leveled between 1995 and 1996, crime has in fact been rising ever since. This is especially true given South Africa's high and rising levels of violent crimes. In 1999, the annual growth in the overall number of registered offenses was more consistent than it had ever been since 1994. During this time, violent crimes increased substantially more than the overall rate.

Other studies focused on investigating the association between crime and several features such as unemployment, Income level, education, and gender. According to Bhorat et al. (2017), studies worldwide show or indicate if there is association between crime and factors such as unemployment, poverty, and other factors in South Africa, even though it is classified as one of the countries with highest levels of unemployment, poverty, and inequality. After assessing other studies that outlined the factors that contribute towards crime rates, there is a clear indication that shows association between these factors with the rise of crime rates.

McCafferty and Action (2003) found that factors like low educational standards, alcohol abuse, a lack of social and vocational skills, and living conditions, and poor parenting abilities were to blame for the country's high crime rates on their study where they were comparing murder rates in South Africa in the past and present. The study done by Schonteich, and Louw (2001) placed it in a gender and crime perceptive, this study assessed which gender is most likely to fall victim and indicated that the likelihood of an ordinary male falling prey to crime is strongly impacted by his age, socioeconomic level, and geographic location. In females, only location and age were major determinants of crime risk.

Blackmore (2003) also stated that irrespective of their socioeconomic status, wealth, creed, colour, or culture, women and girls in South Africa appear to have become victims of physical, sexual, and psychological abuse. Those who were unemployed made up 49% of those who committed such crimes. Generally, there are two main ways to make money: through legal labor markets or through illegal operations. It is reasonable to expect that there is a positive relationship between the level of unemployment and the crime rate because the majority of research suggests that if the official unemployment rate is high, there is a certain incentive to turn to illicit activities to make a living.

2.2. THE NEED TO USE MACHINE LEARNING

Criminal justice and law enforcement experts have traditionally been in charge of conducting investigations. Computer data analysts help law enforcement personnel solve crimes more rapidly thanks to the increasing use of electronic systems to track crimes (Vaidya et al, 2012). It is debatable to what extent crime statistics can be accurately assessed (Burger et al, 2010). Since 1996, the South African Police Service (SAPS) has given data on the crime rate in the nation (Burger et al, 2010).

SA possesses some of the highest crime statistics reported in the world, hence the particular interest in predicting crime rates (Bhorat *et al.*, 2017). The rate of criminal activities experienced by individuals and households in the 2020/21 period dropped in SA due to strict measures implemented during the covid lockdown. However, the crime categories were either reducing or increasing. For instance, the rate of numerous types of violence against girls and women increased during that period (Stats SA, 2021). The overall contact with criminal activities increased by 2% between October and December 2021 (StatsSA, 2021). There is still a struggle to identify the contributing factors that cause the rate to increase. Crime contributes significantly to the country's economy declining. Hence it is vital to find solutions to reduce the crime rate (Obagbuwa and Abidoye., 2021).

A study done by Obagbuwa and Abidoye, (2021) stated that machine learning approaches may effectively uncover the Concealed patterns in crime data are valuable because they produce excellent visuals for crime prediction and, as a result, enhance South African efforts to avoid crime. Crime data analytics enable the officials to expedite the processes of solving crimes by extracting unknown, crucial information from raw data. McClendon and Meghanathan, (2015) stated that using machine learning tools may be a time-consuming and laborious procedure for law enforcement officials who have to sift through enormous amounts of data. Nonetheless, the protection and security of people are well worth the accuracy with which one may deduce and develop new knowledge on how to slow down crime.

2.3. REVIEW OF MACHINE LEARNING ALGORITHMS

Machine learning techniques may be unsupervised or supervised, and various research in the subject of investigating or forecasting crime rates have touched on both the approaches. Other similar studies employed the unsupervised technique, while others used the supervised strategy of dealing with the rise of crime rates, intending to minimise these rates utilising the electronic approach.

According to Vaidya et al. (2012), Computer data analysts assist law enforcement officers in solving crimes more quickly, and the usage of electronic systems to track crimes is on the rise. In addition to the current techniques, developing a data mining concept that improves the investigation of crimes. Numerous researchers have addressed situational crime prevention challenges and presented various classification, regression, and clustering approaches. This study presents a classification approach.

Classification algorithms are usually utilised in predictions depending on historical data (Zhang et al, 2020). This approach contains the potential to predict the label for classes, given that enough training instances are supplied. Several supervised algorithms are available, such as Naive Bayes, Decision Tree, K-Nearest Neighbor, support vector machines, Logistic regression, Random Forest, Linear regression, weighted voting, and Artificial Neural Networks. Even though there are many classification algorithms, previous studies usually select the algorithms that work for them, some are based on accuracy, and some go with what’s been recommended by studies done previously.

Some studies opt for one algorithm and others use multiple algorithms for comparison. The following study uses KNN, NB, and RF. These algorithms were selected based on the recommendation from previous studies and the accuracies they got while using these algorithms on crime dataset. A Study by Mahmud et al. (2021) employed linear regression to forecast the crime rate, attempting to examine specific algorithms and assess the crime rate in the United States. Mahmud et al. (2021) determined insufficient evidence to sustain model linearity. Obagbuwa and Abidoye (2021) additionally applied linear regression to assess the crime pattern and visuals and estimate future crime occurrence in SA.

The investigation that was carried out by Moeinizade and Hu (2020) used LogR, KNN, NB, DT, and SVM. However, the research revealed that among these algorithms' DT and LogR performances are better than the others, and KNN performs badly, notably for their case study. With the decision tree, the research merged numerous decision trees instead of just one. It was suggested that the ensembling procedures of bagging and boosting deal with lowering variation and bias, respectively, and that would assist with improved performance. With support vector machines, the research emphasised that this technique normally results in good accuracy, but the accuracy becomes impaired when the dataset is huge, this limitation is the same as with the Naive Bayes.

Vaidya et al. (2012) investigated crime rate prediction using an unsupervised learning approach utilising a clustering algorithm. The study emphasised the faster rise of crime in the world today and added that, For many other reasons, it is also challenging to predict when crimes will occur. The study stated that the clustering techniques are used to produce several clustering models to discover criminal trends. The study did not go too deeply into criminal justice for the sake of its modelling instead, it stuck to the primary categories of offenses. The crime clusters used are geospatially located on a map with precise geographical coordinates with the goal of producing the location where crime is most likely to occur. The clustering approach was aligned with the perspective of data mining.

The comparative approach of various algorithms is used to illustrate the strength and accuracy of each classification algorithm in term of performance efficiency. A comparison analysis will undoubtedly bring out the benefits and drawbacks of one strategy over the other.

2.4. MACHINE LEARNING TECHNIQUES FOR CRIME PREDICTION

Ahishakiye et al. (2017) used a different method and produced a crime detection prototype model using the DT algorithm, which projected 94.25% accuracy and concluded that the DT technique is excellent in forecasting future crime occurrence. However, another crime forecasting research done by Shah et al. (2021) focused on crime prediction and prevention, which demonstrated how unstable the decision tree is since the tiny changes in the dataset led to significant changes in the structure, and the projected accuracy gained was 83.95%.

Shah et al. (2021) contrasted DT with NB, KNN, SVM, RF, LR, and LogR. However, it selected RF as the best one for predicting crimes, obtaining an accuracy of 97%. The research also suggested KNN and NB as they both had 87.03% and 87% accuracy, respectively, when k = 10 for KNN. NB was reported to be more accurate than KNN, even if NB works only when the number of features is minimal.

Kim et al. (2018) focused their work on ML-based crime prediction. The crime data evaluated for prediction is from the previous 15 years in Canada (Vancouver). ML-based crime analysis covers models' prediction, identification of trends, data classification, data collection, and visualisation. The study employed classification algorithms that boosted decision trees and KNN to analyse the crime dataset, When utilizing ML techniques, it was possible to predict the crime with an accuracy of 39% to 44%. Although the accuracy of the prediction model was low, the authors came to the conclusion that by changing the algorithms and crime data for applications, the accuracy might be raised or improved.

Isafiade, Ndingindwayo, and Bagula (2021) employed historical crime data, such as the location, date, and time of reported offenses, to train their predictive police model using long short-term memory (LSTM) and convolutional neural network (CNN) models. The model was evaluated using precision, recall, and F1 score metrics, and the authors reported that it achieved a precision score of 0.83, recall score of 0.79, and F1 score of 0.81 in predicting crime hotspots. These results suggested that the model was able to accurately identify areas where crimes were likely to occur.

An experimental study done by Iqbal et al. (2013) on classification algorithms for crime prediction highlighted that the accuracy of the NB and DT algorithms used to assess crime stats was 70.81% and 83.95%, respectively. the decision reached was that the DT outperforms NB.

2.5. SELECTED CLASSIFICATION TECHNIQUES

This study applies algorithms; KNN, NB, and RF to predict crime rates in South Africa. The supervised machine learning approaches were picked based on their accuracy when forecasting crime rates indicated by previous studies on this issue. These algorithms behave differently according to every algorithm has its own benefit and downside in terms of accuracy, complexity, and training time and might offer various results from the same dataset. This section will also encompass the benefits and downsides of these algorithms

KNN is a non-parametric SML approach that is utilized for regression and classification as well as issue classification and prediction. A target variable is predicted using one or more independent factors. With this method, feature comparison is used to estimate the values of various data points, meaning that the value given to each data point will rely on how closely it resembles the points in the training set. KNN is one of the top 10 data mining algorithms, according to Wu et al. (2008), because of its simplicity, effectiveness, and implementation. Many real-world and practical classification problems in a variety of areas, including intelligent systems and experts, can be successfully solved using the KNN-based classification approach.

Pednekar et al. (2018) studied crime rate prediction using KNN, and the primary goal was to identify clustered crimes based on occurrence frequency throughout different years using KNN classification for crime prediction. The study suggested an approach of criminal analysis based on available information to derive crime trends. The approach was mainly focused on data mining as it is essential for crime analysis since an iterative process of extracting information concealed from enormous amounts of raw data. Furthermore, the approach included the frequency of occurring crimes predicted based on the geographical distribution of available data. The suggested approach can anticipate locations with high crime rates and forecast crime-prone areas.

The data used in a crime rate prediction study done by Pednekar et al. (2018) was in raw form and includes some erroneous and missing values. The suggested method of predicting crime-prone areas and places with a high crime rate introduces a novel framework for grouping and forecasting crimes based on data. Pednekar et al. (2018) stated that the KNN algorithm would be more accurate if the study examined a specific state/region. Also, another concern was that the algorithm would not forecast the time in which the crime occurred. KNN has the benefits of being robust to noisy training data, training quickly, and being simple to comprehend and use. Some of the limitations of KNN include memory constraints, sensitivity to the local normality of the model, and the fact that it is lazy supervised learning (Jadhav et al., 2016).

Vural and Gok (2017) conducted a study based on the NB classifier predicting criminal issues. The suggested approach was to utilise a practical model based on the NB classifier provided with unique techniques. The suggested technique may be applied in criminology with its 83% success rate for security personnel to aid explaining the situations. An effective model based on the NB classifier is provided with unique techniques utilised for the criminal prediction issue.

The experimental findings of Vural and Gok (2017) study demonstrated that the suggested model may be employed in criminology with its average 78.05% success rate to enable security personnel to discover the offender. A model based on the NB classifier was presented with unique techniques utilised for the criminal prediction issue. Incident-level crime data are created synthetically by the model itself, otherwise it is impossible to collect. The experimental findings demonstrated that the suggested model may be applied in criminology with its average 78.05% success rate.

Moeinizade and Hu (2020) implemented random forest alongside with other algorithms however the study concurs that RF in their original form, often do well with categorical data. And the experiment includes the two important parameters for this algorithm, the number of trees and the number of features chosen for each tree. These parameters are adjusted by running a grid search between several values. This study used log loss to evaluate the models and the results showed the RF performing better.

A method developed by Hossain et al., (2020) is presented that predicts crime by examining a dataset of past crimes in San Francisco and their trends. The decision tree and KNN ML algorithms are primarily used by the proposed system. The forecasting model’s accuracy was increased by using the random forest algorithm and adaptive boosting. In the study, which combined random under sampling with resampling, the RF method yielded an accuracy score of 99.16% and a logloss value of 0.17. After under sampling, the best accuracy score was attained. By utilising the same datasets and other cutting-edge machine learning approaches, Somayeh et al

(2013) found that the KNN computation accuracy rate is greater (89.50%) compared to other ML approaches.

Random Forest advantages includes operating well even when the data includes null/missing values, overcoming the issue of overfitting since output is based on majority vote or average, preservation variety since all the traits are not evaluated when building each decision tree, but not true in all circumstances, random forest is the most stable when the average responses supplied by many trees are chosen, and each decision tree formed is independent of the other. Some disadvantages of random forest include processing time is greater compared to other models owing to its intricacy and that random forest is exceedingly complicated in comparison to decision trees where choices may be reached by following the route of the tree.

2.6. SUMMARY

The previous studies suggest that SA has the highest crime rates. It is still difficult to determine the cause of the increase in crime rates in SA. This section includes the prior literature and discusses the need to use ML. It also discusses how ML may successfully help uncover hidden patterns or provide an accurate assessment of the crime statistics to avoid human error. To select the classification algorithms to utilize, the study considered prior accuracy and recommendations. It also offers details about the chosen algorithms based on previous findings. This includes both the benefits and drawbacks of each algorithm.

CHAPTER 3 METHODOLOGY

3.1 INTRODUCTION

This section briefly outlines the project workflow and how the study employs targets to forecast crime rates using a South African dataset. It includes the dataset used, data exploration, data pre-processing, model classification, and the evaluation of models. The study followed the supervised machine learning (SML) approach to predict crime rates. The term SML is one of the machine learning methods consisting of several Predictive models; predictive modelling entails creating models that do in prediction. SVM, DT, KNN, LR, MR, LogR, and RF are all examples of predictive models within supervised learning.

Osisanwo *et al.*, 2017 stated that SML is a branch of machine learning and artificial intelligence that looks for algorithms that use publicly provided examples to reason about general hypotheses that produce forecasts for the future cases; machine learning predicts the future using previous knowledge. The subsections in this section cover information about the datasets and the study's objectives, such as data exploration, data preparation, classification models, and model assessments.

3.2. DATASET

This study uses a public dataset of crime and population statistics from SA obtained from Kaggle. The dataset can be retrieved on Kaggle by following the Kaggle repository (Kaggle, 2020). The crime stats dataset contains 14 features: police stations, categories, provinces, and year intervals from 2005 to 2016, with an overall size of 30861 crimes and 27 categories. The population stats contain four features: provinces, density, population, and area. Table 3.1 present all types of crimes covered in this investigation.

Table *3*.1: All crime types in the crime statistics dataset.

NUMBER	CATEGORIES
1	all theft not mentioned elsewhere
2	malicious damage to property
3	theft out of or from motor vehicle
4	theft of motor vehicle and motorcycle
5	stock-theft
6	shoplifting
7	sexual offences as result of police action
8	sexual offences
9	robbery with aggravating circumstances
10	robbery at residential premises
11	robbery of cash in transit
12	robbery at non-residential premises
13	murder
14	illegal possession of firearms and ammunition
15	arson
16	drug-related crime
17	driving under the influence of alcohol or drugs
18	common robbery
19	common assault
20	commercial crime
21	carjacking
22	burglary at residential premises
23	burglary at non-residential premises
24	bank robbery
25	attempted murder
26	assault with the intent to inflict grievous
27	truck hijacking

To evaluate predictive accuracy objectively, one must divide the data. Typically, it suffices to arbitrarily split the data into three sets, namely, the training set, testing set, and validation, to prevent biased evaluation of model prediction performance and it can be used to prevent models from underfitting, overfitting. Train-Valid-Test split is a method to assess the performance of the ML models, which could be classification or regression. The study focuses on classification models. Below is a brief description of the train, validation, and test set.

A training dataset fits the parameters of classifiers for producing a fitted model that generalises well to new unknown data. Training datasets can be applied even in other models besides classifiers. For classification, the supervised ML algorithms look at the training data to learn the optimal combinations of variables that generate a good model. The evaluation of the trained model happens using new examples from the held-out data, validation, and test set to estimate the model's accuracy in classifying new data. This study uses 70% of the train set to design the classification model. The validation set is the subset of the dataset used to give an unbiased evaluation of a fitted model while performing the model hyperparameters tuning.

The study sampled 10% of the validation set to refine the models. Finally, the testing set is the subset of the primary dataset. The studies use the training set to provide an unbalanced evaluation of the fitted model on the trained dataset; unlike the validation set, the testing set evaluates the final model fit. The test set is independent of the train set. The study used 20% of the dataset to test the models.

3.3 DATA EXPLORATION ANALYSIS

Exploratory data analysis (EDA) is an essential phase in machine learning. It is a fundamental step of analysing the data to understand how it looks and uncovers patterns and areas to dig into more using visuals. This phase includes detecting outliers, missing values/ null values, the relationship between variables, descriptive statistics, and data type. This phase presents two diverse types of EDA, namely: quantitative or graphical.

Graphical and Quantitative analysis:

Quantitative EDA forms part of the general analysis approach, which detects nulls/missing values and duplicates. The Quantitative approach includes calculations of descriptive statistics, which consists of a measure of spread (standard deviation, variability, and variants), the measure of central tendency (mean, mode, median), and checking the existence of outliers in the shape of the distribution. Graphical EDA also forms part of the general analysis approach; however, it focuses on the visualisation side of EDA. Either of these two types of EDA can be univariate or multivariate/bivariate analysis

Univariate and Multivariate/Bivariate analysis

The univariate analysis only focuses on a single variable, whereas multivariate analysis explores the relationship between two or more variables, and when only comparing two variables, it is known as bivariate analysis. The EDA divides the following types into four categories: Univariate non-graphical, Multivariate non-graphical, Univariate graphical, and Multivariate graphical analysis. This study only outlines the functions of the EDA types individually. The study uses the multivariate approach since the datasets have more than one variable. This study used all general EDA types to explore crime and population datasets. This part encloses all the commands used and a summary drawn from the outcome of the analysis.

Univariate and Multivariate Quantitative analysis

The study assessed the null, missing values, and duplicates for each feature using the following commands: data.IsNull().sum() and data.duplicated.sum(), respectively. The outcome confirms that there are no nulls or missing values in the dataset and no duplicates**.** We looked at the descriptive statistics of each numerical feature by using the following command: data.describe() This command provides the mean, standard deviation, and each numerical feature's five-number summation. Minimum, Lower Quartile (Q1) = 25%, Median (Q2) = 50%, Upper Quartile (Q3) = 75%, and Maximum make up the five-number summary.

The standard deviation measures the degree of variance or dispersion among a set of data. While the mean interval summarizes data, the outlying of the data points might affect the mean. A low standard deviation suggests that the values are likely to be close to the set's mean, whereas a large standard deviation suggests that the values are dispersed out over a wider range. When the data are symmetrical or evenly distributed, the mean is useful..

Univariate and Multivariate Graphical analysis

This section includes data exploration visuals that showcase the crime report from 2005-2016 and the relationship among the features. The findings in figure 3.1 shows the overall crimes reported in SA from 2005 to 2016 clustered by provinces.

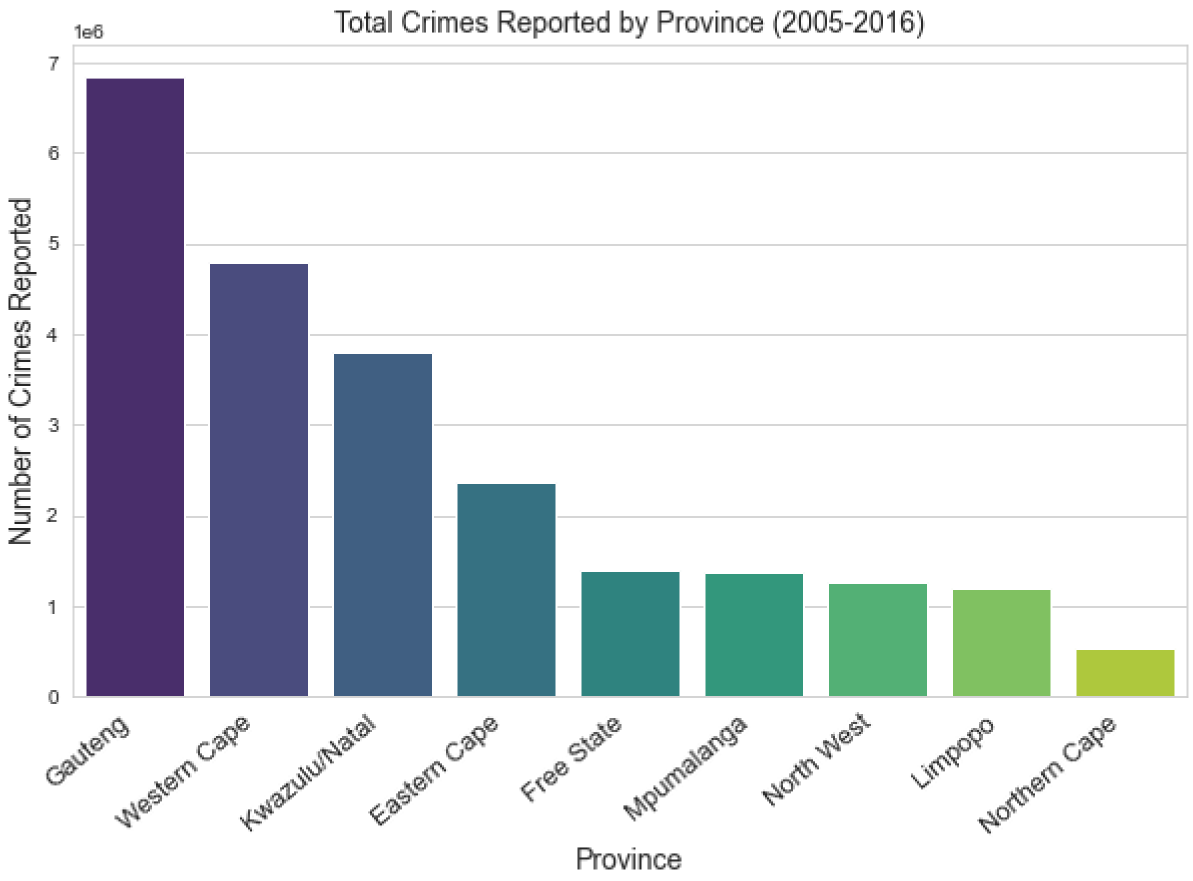

Figure 3.1: Total crime reported per province.

Figure 3.1; shows how overall crime has changed in each province between 2005 and 2016; Gauteng Province has the greatest percentage of offenses. Gauteng, the Western Cape, and KwaZulu-Natal are the provinces that exhibit a propensity for crime. Moreover, the three provinces with the fewest incidences are the Northern Cape, Limpopo, and Northwest.

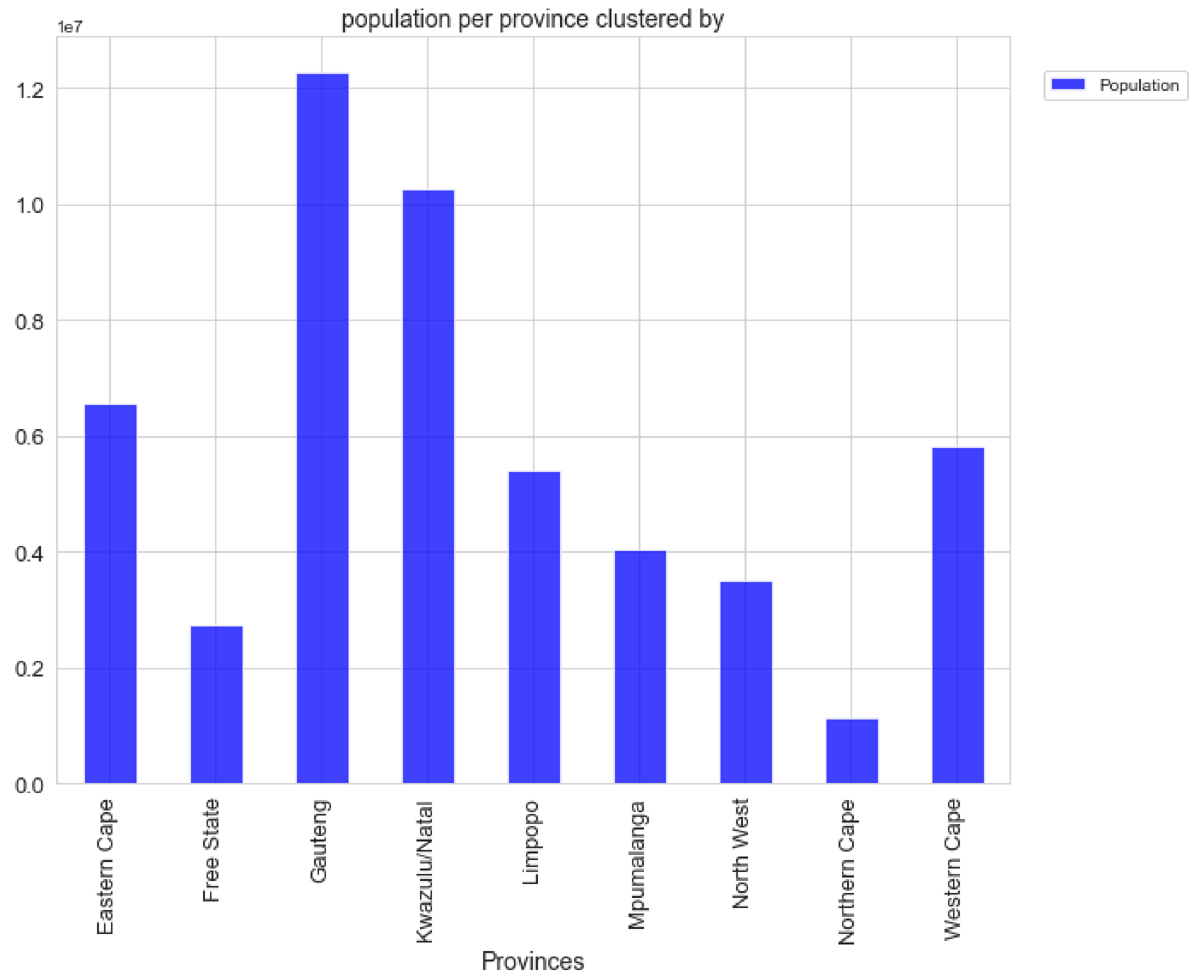

Figure 3.2: Population clustered by province.

Figure 3.2; displays the population per province. Gauteng has the highest population compared to other provinces, followed by KwaZulu Natal. The provinces with the least number of provinces are Northern Cape and Free State, respectively. According to figure 3.2, the province with the highest crime reported is Gauteng, and Gauteng has the highest population. The following highlight could mean that the population influences criminal activities in the provinces; however, KwaZulu Natal shows the third highest number of crimes but the second highest population.

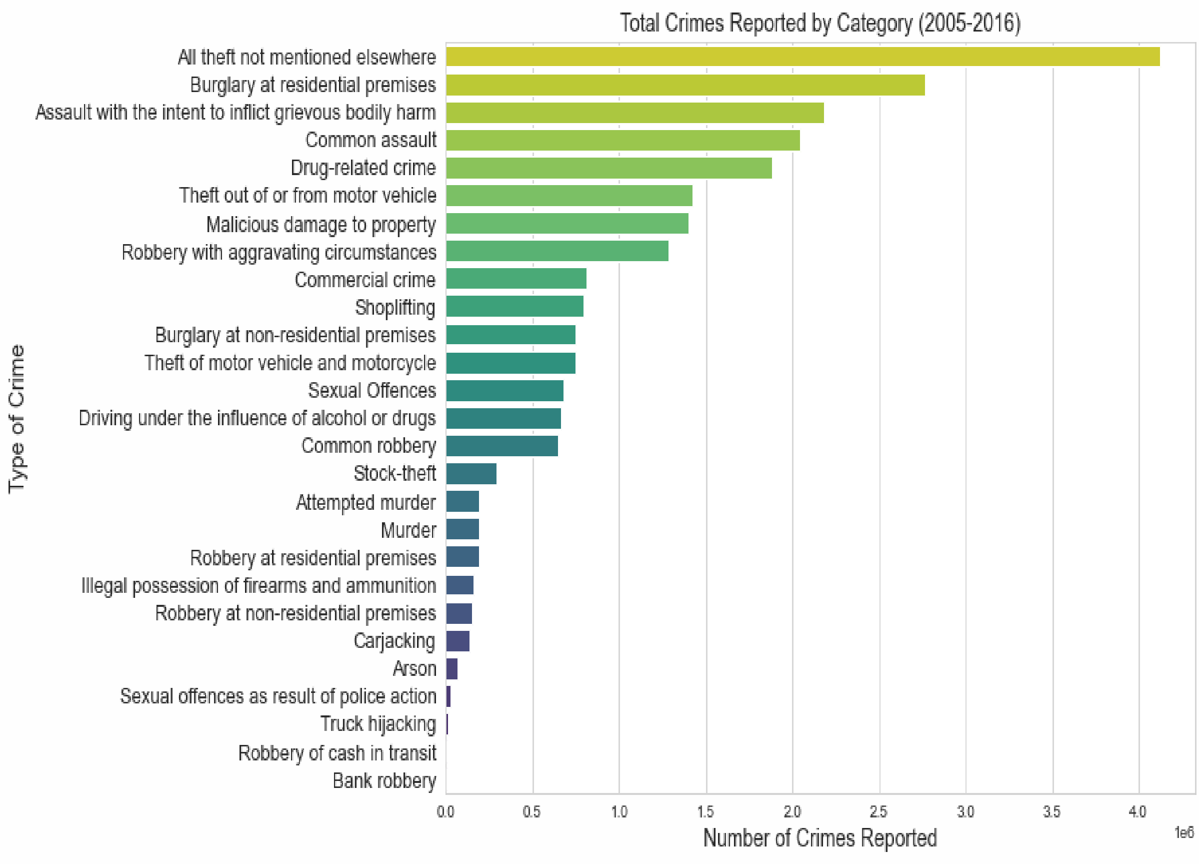

Figure 3.3: Number of crimes reported per category.

Figure 3.3; depicts the 27 crime categories and the total crime committed. The most crimes reported are burglaries at residential properties, assaults with the purpose to do great bodily injury, and all theft not specified above, whereas the least crimes reported are robberies of currency in transit, bank robberies, and truck hijackings. The most crimes reported are burglaries at residential properties, assaults with the purpose to do great bodily injury, and all theft not specified above, whereas the least crimes reported are robberies of currency in transit, bank robberies, and truck hijackings.

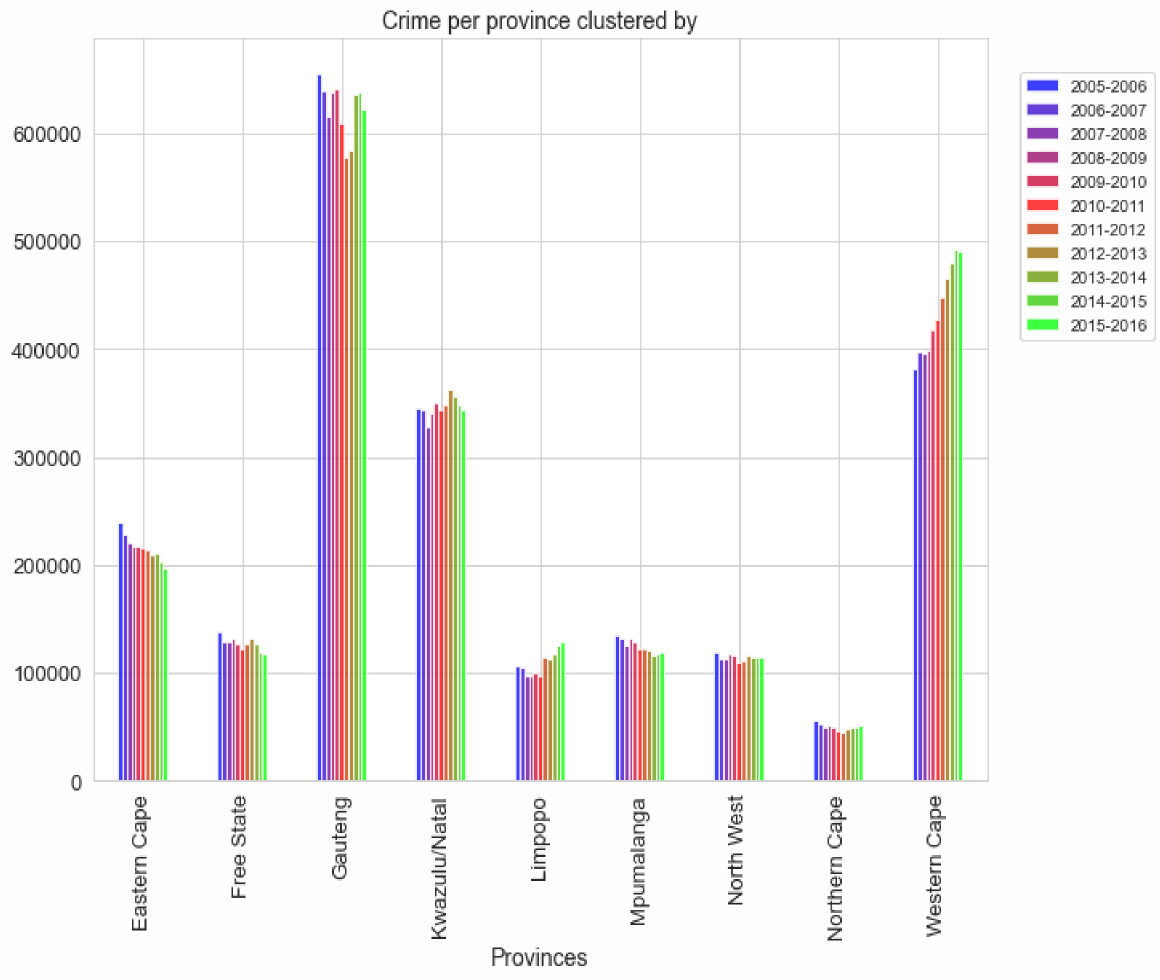

Figure 3.4: Crime per province clustered by year intervals.

Figure 3.4; The highest number of crimes were reported in Gauteng; the number of crimes decreased from 2006 until a point around 2016, after which it was relatively constant. This shows that 2016, despite both years having a significantly lower crime rate (in total) than the years before, is no better than 2015. The dataset collected reveals an inconsistent increase and decline in crimes reported over the previous 11 years. The second province with the most significant number of crimes recorded is the western cape, with the number of crimes growing each year for the previous 11 years.

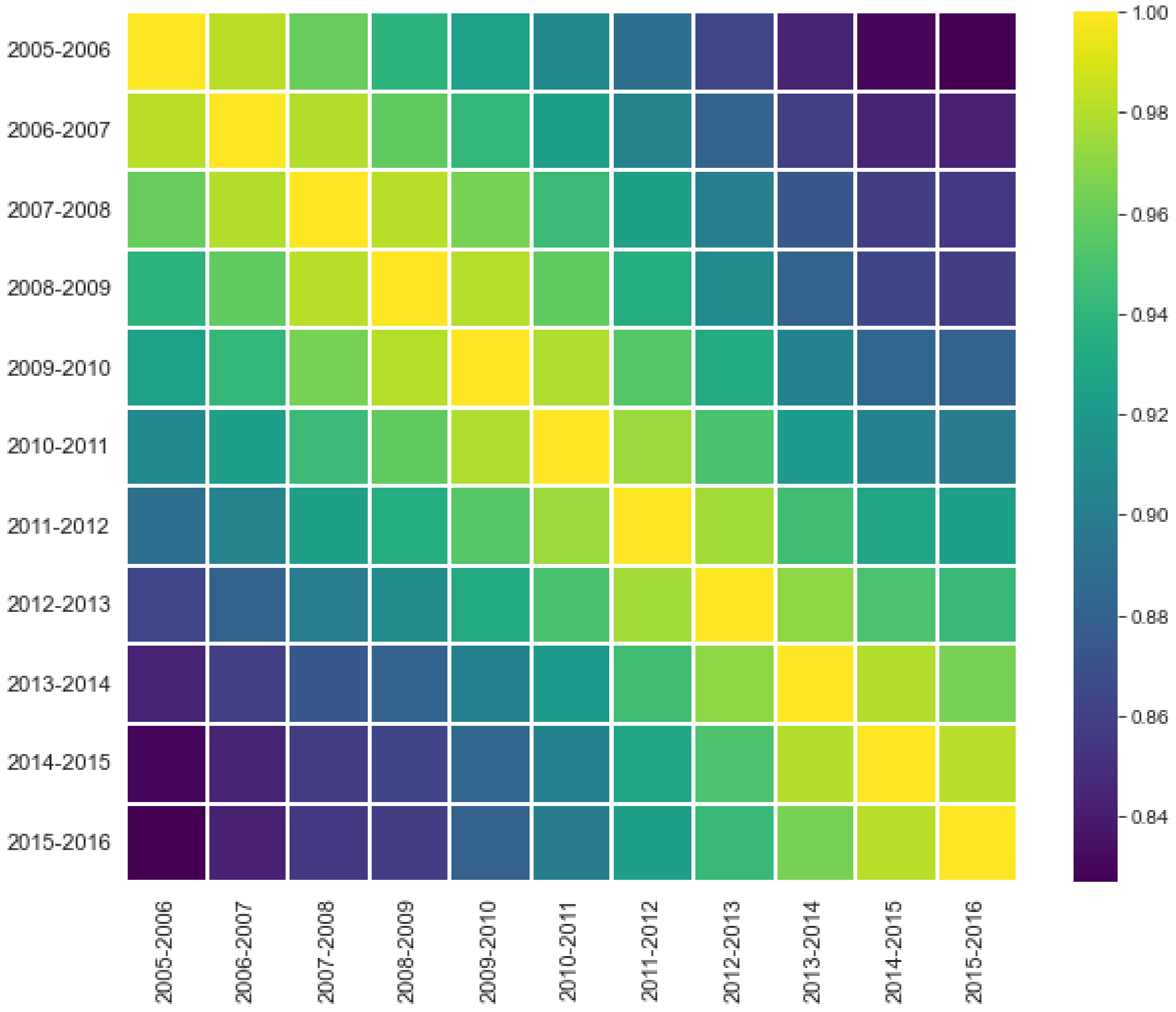

Figure 3.5: Features association heatmap.

From the above Correlation Heatmap, there is a strong positive correlation between the features, and as shown in figure 3.5, the correlation score is between positive 0.8 to 1, indicating a strong relationship between the features. The strongest correlation exists between the year intervals before and after the year period in question, with green denoting a score between 0.94 and 0.98.

3.4. DATA PRE-PROCESSING

Data pre-processing is one of the essential phases in data mining. This method is vital for cleaning, formatting, and sampling the data as the data collected could be incomplete or contain null values. Having data containing null or missing values can significantly affect the model's accuracy; hence, cleaning of data and making it suitable for a machine learning model is necessary as it increases the accuracy and efficiency of a ML model.

3.4.1 Data Cleaning

Data must be cleaned before it can be used. Everything from deleting duplicate entries to missing outlier data is part of this tedious procedure. It is simpler to extract value from datasets throughout the manipulation of data, modelling, and algorithm learning phases the more precise the cleaning step is.

In this section, the raw dataset becomes a clean dataset. All categorical features were converted into numerical features to avoid inconsistency in the models. This study does not contain missing items, duplicates, and outliers. Initially, the dataset contained 14 features; the police station feature was removed from the crime stats dataset using the drop function because it was of no use in this study. In the population stats, we removed the density and area features.

After converting the raw data to a clean dataset, the study manipulated the data because it was difficult to determine the labels with the features provided. When dealing with supervised machine learning labels are one that the data must have so that the algorithm does not result in underfitting. The study decided to utilise the categories as the labels however when all categories are used the models underfits due to that reason the study decided to categorise the crimes using different approaches. More details on the approaches are discussed in the section that follows.

3.4.2 Categorising crime

Initially, the study used 27 categories; however, since it was mentioned that the study will use a supervised ML technique. The study considered two approaches, namely: approach I and approach II. using Category as the target variable. All the categories were converted to lowercase for consistency purposes and to make sure that the integrity of the dataset is not compromised by allocating wrong crime types into wrong categories. These approaches are constructed motivated by the study done on the analysis of through ML by Kim et al (2018).

Approach I: This approach comprises of the relevant crime types recorded in the dataset supplied. The crime types are bundled to build up three fundamental crime categories, such as violent crimes, robberies, and property crimes. In this technique, the study simply focuses on those three areas. The study constructed a dictionary with new crime types grouping the categories based on the type of the criminal activity. All criminal activities that happened on the premises were given the label property crime, all robbery crimes were given the label robbery, and all violent activities were labelled violent crimes. The study selected the studies that are the most contributing towards the crime rate increment and the once that are least contributing.

Table 3.2: Crime types that are grouped into three categories: Property crimes, Robbery and Violent crime.

CATEGORIES	CRIME TYPE
Property Crime	Burglary at non-residential premises. Burglary at residential premises. Theft of motor vehicle and motorcycle. Theft out of or from motor vehicle. Stock-theft
Robbery	Common robbery. Robbery at residential premises. Robbery at non-residential premises. Carjacking. Bank robbery. Robbery of cash in transit. Truck hijacking.
Violent Crime	Sexual offences as the result of police action Sexual offences Assault with the intent to inflict grievous Common assault Attempted murder Murder

Approach II: In this approach, the categories were grouped into two groups, one consisting of violent crimes and all other crimes not included in violent crime types encompassed in one category labelled other crimes. No crime type was left out in this approach. This approach was chosen because of an examination that looked at the model's behaviour with fewer labels than when using more than two labels. Everything that was done was conducted in a manner identical to how Approach 1 was carried out. The crime types are distributed as follows:

Table 3.3: Crime types that are grouped into two categories: violent crime and all other crimes not included in the violent crimes.

CATEGORIES	**CRIME TYPE**
Other Crimes	malicious damage to property shoplifting sexual offences as the result of police action sexual offences murder illegal possession of firearms and ammunition arson drug-related crime driving under the influence of alcohol or drugs common assault commercial crime carjacking attempted murder to assault with the intent to inflict grievous *truck hijacking*
Violent Crime	robbery with aggravating circumstances robbery of cash in transit Burglary at non-residential premises. Burglary at residential premises. Theft of motor vehicle and motorcycle. Theft out of or from motor vehicle. Stock-theft all theft not mentioned elsewhere robbery at residential premises robbery of cash in transit bank robbery common robbery robbery at non-residential premises

After all the approaches were implemented, the target variable which is categories was converted to numerical using Label Encoder. The label encoding in Python may be accomplished using the SKlearn library. Sklearn provides a very efficient way for transforming the levels of categorical characteristics into numerical values. Where n is the number of unique labels, Label Encoder encodes labels with a number in the range of 0 and n classes.

The features were scaled using standardisation method. By using feature scaling, one ensures that all features are around the same size, giving them all the same weight and making them simpler for most ML algorithms to comprehend. The data was transformed into a helpful format and made ready to feed into the model. The splitting of datasets into test, train and validation sets was done after the transformation of the dataset.

3.4.3 Technologies and software used for pre-processing.

This study utilised the python 3.6 package to carry out the experiment because it is now the most used and very easy to navigate and understand programming language. Because of its simplicity, Python is presently the programming language with the fastest growth globally due to its rapid learning curve and supply of high-quality packages for data science and ML (Vallat, 2018). In data pre-processing Importing a couple of the crucial libraries needed is the first step. A library is a set of callable and useable modules. Here are software modules in Python that are used for pre-processing data as well as the data Exploration. The significant modules used includes:

NumPy, is often used to build or apply complex mathematical computations for ML. It is helpful when working with multidimensional arrays. Python data structure and analysis tools with excellent performance and ease of use are offered by the open-source package known as Pandas. Working with relational and tagged data is simple and straightforward because of its architecture.

Matplotlib is a Python visualisation package for 2D plots and arrays. It is constructed using a NumPy array and intended to operate with a larger SciPy stack. When there is a lot of data accessible, visualisation of the datasets is useful. In matplotlib, you can create line, bar, scatter, histogram, and other types of plots. Seaborn: Python also provides a visualisation library. It offers a sophisticated user interface for creating visually appealing and useful statistics graphs.

The SKlearn package provides various common utility methods and transformer classes to transform raw feature vectors into a format that is more suited for the following estimators. For learning algorithms, standardizing the data set is generally useful. In the event that there are any outliers in the collection, transformers or robust scalers are advised. In Assessing the influence of various scalers on data including outliers, the operations of a number of scalers, normalizers, and transformers are highlighted.

3.5. CLASSIFICATIONS MODELS

The KNN, NB, and RF classifiers implementations are discussed in this section. All the parameters used are discussed here.

3.5.1 K-Nearest Neighbors:

The selected algorithm KNN simply produces a prediction of a class of a target variable which is the categorised data based on number of closest neighbors. Every instance in the training dataset will be measured in relation to the occurrence one wishes to classify, and then your occurrence will be classified according to the overall class of a *k* closest instances. The prediction is made using the validation and train set first before testing the model using the test set to check how the final model performs. This algorithm utilises the distance formula to find the input and training data points that are the closest to one another.

The algorithm converts the data points into measured values since it evaluates the distance between the measured values of the data points. It determines the distance between each data point and the test data before estimating the likelihood that the points will match the test data. In developing KNN the study declared a function called predict () with these inputs: the same k value, train set, validation set, and test set, implemented on both approaches. The KNN algorithm parameter used for this study are n_neighbors, p, metric, and weights, these parameters were selected based on how they improve the model. The assessment was done to evaluate them and see which parameter is improving the score and use only those instead of using all the parameters under this algorithm. The *k* value corresponds with the n_neighbors' parameter.

n_neighbors: Its default value is 5. n_neighbors is the most basic setting for KNN algorithms, it establishes the minimum n_neighbors that must be inspected when classifying an object. The KNN algorithm's k parameter specifies how many n neighbors will be considered in order to categorize a single query point. Given that different values could result in overfitting or underfitting, defining k may require some careful balance. Bigger values of k may result in strong bias and lower variance, whereas reduced values of k may have high variation but low bias. To avoid ties in classification, it is advised to use an odd value for k, and cross-validation techniques help discover the best k for the set.

To select the value, the study tested the values to see which works better or which *k* is giving better accuracy. The training and validation sets were used to test the k-values and assess which k-value works well with the algorithms used. To test different k values, we use the little portion of the train set that has been selected as the validation dataset. After determining which value of k performs best on the validation set, we use that value as the final configuration of our algorithm to minimize the prediction error. We use k equitable to 1, k equitable to 2, and k equitable to 3 so on until the value that works better is found to predict the label for each instance in the validation set.

Weights: (default: “uniform “) Another crucial option, weights, specifies how weight should be divided amongst adjacent values. there are different weights under this algorithm and are as follows: “uniform”: This quantity allows weights to be spread evenly across all n_neighbors elements. “distance”: This quantity allows weights to be spread depending on actual distance (inverse correlated). Nearby neighbours have a larger weight in the algorithm. finally, "callable": You may also create a function and assign it to this argument.

Weights would be tailored depending on the collection one is given. p has a default of 2 and it helps choose the metric to utilise or apply in the method. p parameter denotes the power for Minkowski. Minkowski distance (l *p*) may be used for any *p. Manhattan distance (l1) is represented by p = 1, and Euclidean distance (p = 2) (l2).* In the experiment, the n_neighbors value that worked better compared to other values is *k* = 11, the *p* used is *p* = 2, and the metric is Euclidean. Other parameters are set as defaults, because when tried out the model score was not changing at all hence the are set to default.

3.5.2 Naive Bayes:

The study separates the columns into dependent and independent variables and assumes no feature is dependent, and each feature will be given the same importance. Different NB classifiers differ in the assumptions they make regarding probability.

NB classifiers are a collection of supervised learning algorithms on the Bayes' theorem. The Bayes' theorem may be presented as seen in equation 3.1

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \tag{3.1}$$

Here,

$P(A|B)$ represent the conditional likelihood that A will happen if B is true, $P(B|A)$ represent the likelihood that B will happen if A is true, and $P(A)$ and $P(B)$ represent the likelihood that A and B will happen independently.

In developing NB, the study declared a function called predict () with these inputs: train set, validation set, and a test set of the categories, and made use of hyperparameters to optimise the model. These are the most regular parameters used with various NB Algorithms; priors: When priors are given (in an array), the prior class chances will not be changed in response to the dataset. var smoothing: (1e-9 by default) In order to smooth out the variance, the given float value will be utilised to determine each feature's greatest variations and apply those variances to the stable calculation variance. In both approaches, the study set the parameter the var smoothing to 200.

3.5.3 Random Forest:

In RF, from a set of k records in a data collection, n records are randomly chosen. For each sample, a unique decision tree is formed, and the output from each decision tree is assessed for classification using a majority vote or a mean. The study randomly selects the samples from the training dataset and structures the DT for each sample to get prediction outcomes as of the decision tree. This study will utilise relative feature importance to select the features of the classifiers that contribute the most.

Now that the data preparation is complete, like other algorithms implemented, random forest needs to be initiated and trained using a training set. In developing random forest, the study declared a function called predict () with these inputs: train set, validation set, and test set of the categories. In developing the random forest, the decision tree is built in the process, and it appears to have some similarities in parameters, but random Forests have their unique factors that might be crucial as the forest is larger and more complicated than the decision tree parameter.

The parameters that are used while building the RF Classifier model can significantly affect how machine learning is implemented. This study used n estimators: (default 100), this option represents the quantity of trees in the forest. This is perhaps the most distinctive optimization parameter of a random forest method and uses max depth: (default None) Another crucial option, max depth represents allowable depth of individual decision trees. The estimators were set to 10000, the estimators were selected based on how it was affecting the model when the number of estimators is increased or decreased. The same goes for Max depth however the depth size used is 800.

3.6. Evaluation of the classification models

After developing the classification, this study evaluates how good the predictions made by the models are. This study uses performance metrics to help improve the models built. Both algorithms' outcomes are assessed based on the following performance metrics which are stated below: Accuracy which outlines the proportion of successfully categorised instances by the classifiers. Precision refers to the fraction of data accurately identified using classification algorithms. Recall is an important performance indicator that relates to the proportion of material that is significant to the class and is properly categorised.

Accuracy

Accuracy is the most in-built execution measure, proportioning predicted observations to the total observations. High accuracy implies that the model is the best; however, that only works if the dataset is symmetric. In the case of having an imbalanced class, this metric is not a good one to use. Accuracy is measured as the sum of true positives(TP) and negatives (TN) divided by the sum of true positives (TP), true negatives (TN), false positives, and false negatives.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \qquad (3.2)$$

Precision

Precision can be used as an individual machine learning metric, forming part of the F1 score. Unlike accuracy, precision considers class imbalance. It counts the correct percentage within everything predicted as a positive. Precision is determined by dividing the overall amount of true positives and false positives by the total amount of true positives across all classes.

$$Precision = TP / (TP + FP) \qquad (3.3)$$

Recall

Recall Is a metric that counts the number of correct optimistic predictions made from all predictions that may have been made and are positive. It also forms part of the F1 score and deals well with the class imbalance. The recall is determined by dividing the number of true positives by the total amount of true positives and false negatives.

$$Recall = TP / (TP + FN) \qquad (3.4)$$

F1- score

The F1 score is a brief way to determine whether a classifier is good at finding elements of a class or at finding shortcuts. It is less spontaneous because it combines precision and recall into a single metric; if both are high, F1 will be high; if both are low, F1 will be low; and if one is high and the other is low, F1 will be low. Between accuracy and F1, accuracy is easier to understand than F1; however, F1 is more valuable than accuracy, especially when there are unbalanced classes (Prithi et al., 2020). F1 score is measured from recall and precision.

$$F1\,Score\ =\ 2 * (Recall\ *\ Precision)\ /\ (Recall\ +\ Precision) \qquad (3.5)$$

3.7 SUMMARY

The research and analytical techniques that were used to create the models are summarized in this section**.** The EDA was performed to draw insights into the data and check the relationship between the variables. The study then performed data cleansing in the pre-processing stage ensuring that all the gaps within the data are closed so that the models can perform accurately. This section includes details about the classification algorithms selected based on the previous studies' recommendations. The study used the scoring method to evaluate the performance. The score provides information about the algorithms' mean accuracy for the provided data. The performance was assessed on the train set first, then the validation set, and then the test set.

CHAPTER 4 RESULTS AND DISCUSSION

Here, the study discusses the classification results and evaluation of models when using different sets. It encompasses the results of the RF, KNN, and NB algorithms. The crime reported from 2005 to 2016 as the independent variables and the categories as the target or dependent variables with 3 labels namely property crime, Robbery, and violent crime for the approach I and 2 labels namely other crimes and violent crimes in approach II whose values are modelled to predict the crime reported from 2005 to 2016. As mentioned in the previous section, this study used two approaches, and the results of these approaches are included in this section.

Approach I

Table 4.1: The counts of approach I categories.

CATEGORIES	***COUNTS***
Property crime	4572
Robbery	8001
Violent crime	6858

Table 4.1 shows that the total count of **Property crimes**, **Robbery** and **Violent crime** are 4572, 8001, 6858, respectively. Even though the definitions of robbery and property crimes are largely similar, and the only significant differences relate to the application of force and terror, it is important to note that the two crimes differ in terms of where they are more likely to occur: Most robberies happen outside of a home, most frequently on the street, while the majority of property crimes occur at residential locations. This shows that whereas robbery crime may be more of a snap decision, property crime is more likely to be planned.

Approach II

Table 4.2: The counts of approach II categories.

CATEGORIES	***COUNTS***
Violent crime	13716
Other crimes	17145

Table 4.2 shows that all violent crime counts to 13716 and other crime type not mentioned in violent category sum to 17145. The violent crime encompasses all the crimes that are violent, and the other crimes includes the crimes such as theft and hijacking and other crimes mentioned in table 3.3.

4.1 COMPARISON AMONG K-NEAREST NEIGHBOR, NAIVE BAYES, AND RANDOM FOREST

4.1.1 Classification Algorithms on training, and validation set

Table 4.3: Results of Accuracy and Log loss for Classification

Classification Algorithms		**Accuracy (%)**		**Log loss**	
		Approach I	**Approach II**	**Approach I**	**Approach II**
Random Forest	Train set	94.81	93.21	0.31	0.26
	Val set	68.65	61.00	0.80	0.72
	Test set	50.65	61.49	1.15	0.68

K-Nearest Neighbor	Train set	64.11	66.37	0.79	0.61
	Val set	64.30	59.31	1.47	0.74
	Test set	47.41	56.96	2.84	0.91
Naïve Bayes	Train set	48.15	55.79	1.05	2.38
	Val set	48.18	54.25	1.04	2.41
	Test set	33.17	54.36	1.33	2.67

K-Nearest Neighbor

KNN algorithm was applied using training and validation set in both approaches. Table 4.3 highlights that for approach 1, accuracy and log loss for the training set are 64.11% and 0.789, respectively, while for approach 2, accuracy and log loss are 66,37% and 0.613, respectively. The accuracy and log loss for KNN using Approach I on the validation set are 64.30% and 1.47, respectively, while the accuracy and log loss for Approach II are 59.31 and 0.74, respectively.

Random Forest

The estimators were set to 10000, and the estimators were selected based on how it was affecting the model when the number of estimators increased or decreased same goes for Max depth however the depth size used is 800. The training set was subjected to the RF algorithm. Table 4.3 shows that for approach 1, the training set's accuracy and log loss are 94.81% and 0.309, respectively, while for approach 2, they are 93.21% and 0.262. The accuracy and log loss for RF using Approach I on the validation set are 68.65% and 0.80, respectively, while the accuracy and log loss for Approach II is 61.00% and 0.72, respectively.

Naïve Bayes

In all methods, the study set the parameter the var smoothing to 200. NB algorithm was applied to the training set. Table 4.3 shows that the accuracy and log loss for the training set for approach 1 and approach 2 is 48.15% and 1.045, respectively, and 55.79% and 2.380, respectively. The accuracy and log loss for RF using Approach I on the validation set are 48.18% and 1.04, respectively, while the accuracy and log loss for Approach II is 54.25% and 2.41, respectively.

4.1.2 Classification Algorithms on the testing set.

The main objective of Shah et al. (2021) was to identify the machine learning (ML)-based techniques that were most effective at forecasting crime rates and to assess how well they applied to the Chicago crime dataset. When the classifications were assessed, accuracy scores for RF, KNN, and NB were all higher than 80%; KNN scored an accuracy of 87.03 on a K=10 test run on the testing set, NB scored an accuracy of 87%, and RF scored the maximum accuracy of 97%. This shows that the RF categorization is the best one for forecasting crime rates. The RF performs better than the other categories, according to the experiment conducted for this study, however, accuracy is poor for both methods.

The models were underfitting because the study first evaluated the model with default parameters before analysing extra values. Table 4.3 has a complete listing of all the outcomes from the test set. For methods, I and II, KNN's accuracy on the testing set is 47.41% and 56.96%, respectively. On the testing set, Approach I's accuracy for RF is 50.65%, whereas Approach II's accuracy is 61.49%. The testing set's accuracy ratings for methods I and II for NB are 33.17% and 54.36%, respectively.

However, a study by Kim et al. (2018) on a dataset from Canada, which contained statistics features similar to those for crime in South Africa, only used KNN against DT, and the accuracy obtained was 39% and 44%, respectively. This suggested that the accuracy was low and that it may be due to the dataset or the categorical labelling. According to Kim et al. (2018), adding hyperparameters or changing the models could help them get better. The study tried altering the models to make them better, but the models didn't become any better at all.

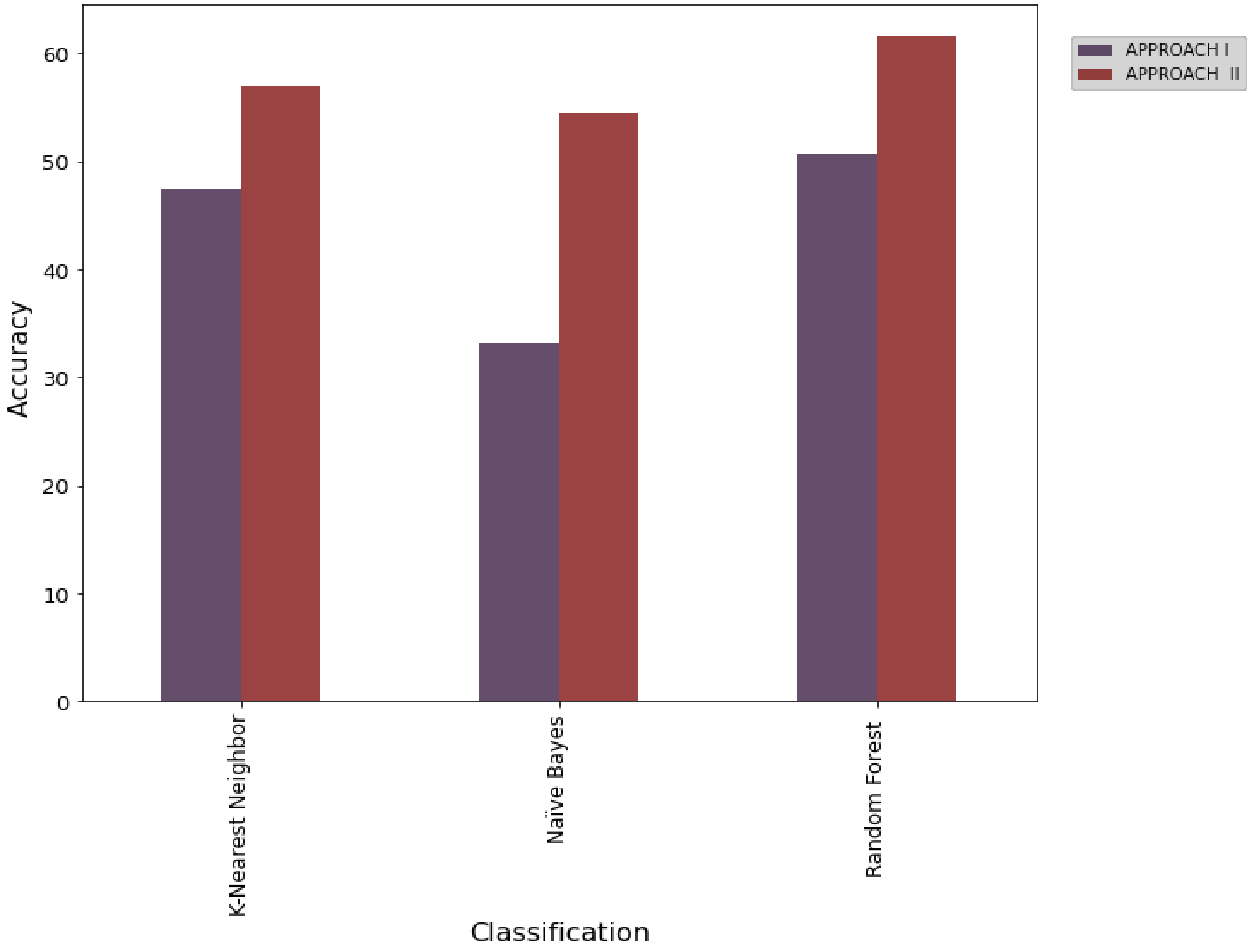

Figure 4.1: Comparison between approach I and II on the test set

As can be observed, the RF is operating more accurately on Approach II compared to Approach I, while using an accuracy set. The study tested the model on default parameters first before evaluating additional values, and the models were underfitting. All the results performed using the test set are recorded in table 4.3. The accuracy and log loss for KNN on the testing set for approaches I and II are 47.41% and 2.84, respectively, and 56.96% and 0.91, respectively. The accuracy and log loss for RF using Approach I on the testing set are 50.65% and 1.15, respectively, whereas the accuracy and log loss for Approach II is 61.49% and 0.68, respectively. The accuracy and log loss for NB evaluated on the testing set for approaches I and II are 33.17% and 1.33, respectively, and 54.36 % and 2.67, respectively.

4.2 MODEL EVALUATION ASSESSING CLASSIFICATION METRICS: PRECISION, RECALL, AND F1-SCORE.

The classification metrics for our models are printed in the study using the classification report () function. The report was run on testing set for evaluation.

Approach I

Table 4.4: Classification metrics results for the approach I

	F1 Score	**Precision**	**Recall**
Random Forest	0.48	48.70	50.65
KNN	0.44	45.30	44.82
NB	0.24	42.83	33.17

Table 4.4: Approach I classification report reveals that RF has a higher precision of 48.70% in predicting crime rates than KNN and NB, which have precisions of 45.30% and 42.83%, respectively. KNN and NB correctly reported 44.82% and 33.17% of the reported crimes, respectively. RF accurately reported 50.65% of the recorded crimes. The f-scores for both models are below 1, which clearly shows that KNN, NB, and RF are not very good at forecasting crime rates.

Approach II

Table 4.5: Classification metrics results for approach II

	F1 Score		**Precision**		**Recall**
	Violent crime	Other crimes	Violent crimes	Other crimes	Overall
Random Forest	0.55	0.66	60	63	51
KNN	0.44	0.68	57	60	53
NB	0.15	0.69	55	54	8

Model I: KNN

For recorded crimes, precision shows that only 57% of violent crimes and 60% of other crimes match the model's prediction. Recall that only 35% of the violent crimes recorded were accurately predicted by the model out of the total reported incidents. F1 score: The KNN model performs poorly in predicting crime rates, with the f1 score for violent crimes being 0.44 and the f1 score for other crimes being 0.68, neither of which is closer to 1.

Model I: RF

Precision: Only 60% of reported crimes are violent crimes, while 63% are other crimes, despite the model's prediction that this proportion would be higher. Recall: Of all reported violent crimes, the model only accurately predicted this outcome for 51% of the recorded offenses with an F1 score for violent crimes of 0.55 and an f1 score for all crimes of 0.66. Since these scores are not closer to 1, the RF model performs poorly in predicting crime rates.

Model I: NB

Precision: Only 55% of reported crimes are violent crimes, compared to 54% of reported crimes that the model projected would be violent crimes. Remember that the model only accurately predicted this outcome for 8% of the reported violent offenses out of all reported crimes. F1 score: Because the f1 scores for violent crimes and other crimes are 0.15 and 0.69, respectively, and are not closer to 1, the NB model performs terribly when forecasting crime rates.

4.3 DISCUSSION

The log loss values for three classification algorithms was performed on the SA crime dataset. It can be noticed that the two techniques (RF and KNN) are capable of handling categorical data and are doing better than NB. The RF technique has outperformed the other two algorithms that were employed in the study in both approach I and II with accuracy of 50.65% and 61.49%, respectively. The study also found that KNN technique had a higher performance in approach I and II of 47.41% and 56.96%, respectively compared to NB technique with performance scores of 33.17% and 54.36 % in Approach I and II, respectively. The contribution of our suggested method resides in the process of model selection and hyperparameters.

After complete model integration on the test set and based on the log losses comparing the approaches, it is observed that approach II performs better than the approach I. When comparing the outcomes of these models, on both approaches based on accuracies RF outperforms KNN and NB classifier. This is evident after evaluating all the results, which show that RF, KNN and NB perform better in approach II than in approach I, however the accuracies for these algorithms are low on this dataset. The accuracies are less than 65% in both approaches. The overall results on the training, validating, and testing sets, RF appears to be working okay, and this view is supported by the log loss, which is low in comparison to others.

As mentioned in previous section Obagbuwa and Abidoye (2021) used LR techniques to analyse and predict crime trends in SA, achieving high accuracy in their predictions. The findings of this study highlight the potential of ML techniques to provide valuable insights into crime trends, which can help law enforcement agencies. The accuracy score of this study was not stated however, the study does highlight that the accuracy is high, meaning the comparison between LR, RF,NB and KNN is not possible.

Overall evaluation of Obagbuwa and Abidoye (2021) study suggest that the LR model is an effective method for projecting SA's crime rates on the same dataset. Further evaluation of the three models (KNN, RF, and NB) for predicting crime rates based on recall, precision, and F1 score metrics for approach II. The precision score shows that the models' ability to forecast the percentage of violent crimes or other crimes in reported instances is insufficient. The recall score indicates that the models are not able to correctly identify a large proportion of actual violent crimes among the reported incidents. And the F1 score, which considers both precision and recall, indicates that the models are low overall.

These results suggest that the models are not sufficiently capturing the complexity and variability of the data, and further analysis and experimentation may be needed to improve the models' performance. It is possible that the models are not including all relevant features that could improve their predictive power. Alternatively, the models may be using features that are not sufficiently informative. Another possible issue could be related to the model selection. It is possible that other models, such as neural networks (NN) or SVM, may perform better on this dataset. Additionally, the hyperparameters of the models may not be optimized correctly, leading to suboptimal performance. Therefore, further experimentation with different models and hyperparameter settings may be necessary to improve the models' performance.

The low performance of the KNN, RF, and NB models in predicting crime rates based on the given evaluation metrics suggests that further analysis and experimentation are needed to improve their accuracy. This may include exploring different feature sets and feature engineering techniques, experimenting with different models and hyperparameter settings, and validating the models on different datasets and scenarios.

4.4 SUMMARY

It is crucial to remember that the quality and amount of the available data have a significant impact on how well machine learning models predict crime. Additionally, the specific features used in the models can have a significant impact on the accuracy, and the models need to be regularly updated as new data becomes available. The study evaluated two techniques, I and II, using categories as the target variable. The findings collected reveal that approach II performs better than approach I, although in both ways, the RF performs better than other classification models (KNN and NB). The classification report, accuracy, and logloss were utilized in the study to support the conclusion that the chosen classifiers accuracies are low on the SA dataset and that more research on RF is still required to determine whether this model will improve since it is the only one that seems to be doing well compared to KNN and NB.

CHAPTER 5 CONCLUSION

5.1 INTRODUCTION

This chapter outlines the key research findings, addresses the objectives, and aims, describes the most important contributions made by the study, evaluates any limits, and finally presents Future work recommendations for the future research purpose.

5.2 CONCLUSION

The first goal of this investigation was attained In Chapter 3, the study gained an insight in the Exploratory Data Analysis phase. Data pre-processing was successfully implemented and the classifications and their evaluation. Data visualisation was developed to identify intriguing statistics that aided in analysing the dataset. The final overall findings were done on a test set, and the outcomes of these models based on accuracy reveal that RF outperforms the KNN and NB classifiers. Random Forest appears to be working better compared with other algorithms, although this method is not doing extremely well as emphasized in previous studies, and it could be improved with a thorough assessment.

5.3 CONTRIBUTION OF THE STUDY

Crime prediction, which determines the crimes that are most likely to occur through data and statistical analysis, is one of the methods employed by law enforcement. Crime creates anxiety in the minds of SA citizens, hinders their equitable participation in the nation's advancement, and robs them of the freedom and rights that the constitution guarantees. The results of the study will be advantageous to both the general public and the businesses impacted by the crime rate. The study's findings will assist law enforcement in anticipating the crime rate by using the most accurate model, which is RF, for doing so. This will enable law enforcement to successfully detect, prevent, lessen, and address crimes occurring in SA. Despite the accuracy found in testing, the visualization used in this study does give a clear indicator of the areas that need better protection and planning to lower the rates. It is possible that the accuracy can be increased with further, and depth implementation of these algorithms.

5.4 FUTURE WORK AND RECOMMENDATION

Future studies could evaluate why the accuracies are this low and what are most contributing factors towards the behaviour of these algorithm using SA dataset. Applications can modify the algorithm and the data to improve prediction accuracy. Despite having low prediction accuracy, this model offers a basic framework for additional investigations. More depth evaluation on these algorithms might provide an exceptional accuracy that could be relied on. The use of different algorithms might help in figuring the model that works efficiently in this dataset.

BIBLIOGRAPHY

Bagula, A. & Ndingindwayo, B., (2021). Predictive policing using deep learning: A community policing practical case study. In Towards new e-Infrastructure and e-Services for Developing Countries: 12th EAI International Conference, AFRICOMM 2020, Ebène City, Mauritius, December 2-4, 2020, Proceedings 12 (pp. 269-286). Springer International Publishing.

Bhorat, H., Thornton, A., & Van der Zee, K. (2017). Socio-economic determinants of crime in South Africa. *An empirical assessment*, (March), pp.1-44.

Blackmore, F. L. E. (2003). A panel data analysis of crime in South Africa. *South African Journal of Economic and Management Sciences*, *6*(3), pp.439-458.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014, November). Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction*, pp.427-434.

Breetzke, G. D. (2010). A socio-structural analysis of crime in the city of Tshwane, South Africa. *South African Journal of Science*, *106*(11), pp.1-7.

Burger, J., Gould, C., & Newham, G. (2010). The state of crime in South Africa. *SA Crime Quarterly*, *34*, pp.1-12.

Council for Scientific and Industrial Research (2018). Intelligent surveillance and response systems: A new era in public safety. Retrieved March 30, 2023 from https://www.csir.co.za/sites/default/files/Documents/Intelligent%20Surveillance%20and%20Response%20Systems%20White%20Paper.pdf

Crime Stats Sa. (2022). Crime Stats SA. Retrieved May 18, 2022, from https://www.crimestatssa.com/.

Crush, J., Ramachandran, S., & Pendleton, W. (2013). Soft targets: Xenophobia, public violence and changing attitudes to migrants in South Africa after May 2008.

Forbes (2018). SAPS to explore predictive policing using machine learning. Retrieved March 30, 2023 from https://www.forbes.com/sites/tobyshapshak/2018/11/08/saps-to-explore-predictive-policing-using-machine-learning/?sh=26f3a10a4a2d

Hossain, S., Abtahee, A., Kashem, I., Hoque, M. M., & Sarker, I. H. (2020, March). Crime prediction using spatio-temporal data. In *International Conference on Computing Science, Communication and Security*, pp. 277-289. Springer, Singapore.

International Association of Chiefs of Police (2016). Predictive policing: A smart policing tool for the 21st century. Retrieved March 30, 2023 from https://www.theiacp.org/resources/predictive-policing-a-smart-policing-tool-for-the-21st-century

Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., & Khanahmadliravi, N. (2013). An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*, *6*(3), pp.4219-4225.

Ivan, N., Ahishakiye, E., Omulo, E. O., & Taremwa, D. (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm. *International Journal of Computer and Information Technology*, *6*(3), pp.188-195.

Kaggle, (2020). *Crime Statistics for South Africa*. Kaggle. Retrieved May 11, 2022, from https://www.kaggle.com/datasets/slwessels/crime-statistics-for-south-africa.

Kim, S., Joshi, P., Kalsi, P. S., & Taheri, P. (2018, November). Crime analysis through machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp.415-420. IEEE.

Mahmud, S., Nuha, M., & Sattar, A. (2021). Crime Rate Prediction Using Machine Learning and Data Mining. *Soft Computing Techniques and Applications*, pp.59–69. 10.1007/978-981-15-7394-1_5

Maverick, R. (2019). How machine learning is helping South Africa fight crime. Retrieved from https://www.itweb.co.za/content/GxwQDMrNMV7qX9l8

McCafferty, R., & Action, U. C. (2003). Murder in South Africa: a comparison of past and present. *United Christian Action*, pp.1-24.

McCluskey, A., & Lalkhen, A. G. (2007). Statistics II: Central tendency and spread of data. *Continuing Education in Anaesthesia, Critical Care and Pain*, *7*(4), pp.127-130.

Moeinizade, S., & Hu, G. (2020). Predicting Metropolitan Crime Rates Using Machine Learning Techniques. *INFORMS International Conference on Service Science*, pp.77-86.

Obagbuwa, I. C., & Abidoye, A. P. (2021). South Africa Crime Visualisation, Trends Analysis, and Prediction Using Machine Learning Linear Regression Technique. *Applied Computational Intelligence and Soft Computing*, pp.1–14. 10.1155/2021/5537902.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), pp.128-138.

Pednekar, V., Mahale, T., Gadhve, P., & Gore, A. (2018). Crime rate prediction using KNN. *International Journal on Recent and Innovation Trends in Computing and Communication*, *6*(1), pp.124-127.

Plessis, A., & Louw, A. (2005). Crime and Crime Prevention in South Africa: 10 Years After. *Canadian Journal of Criminology and Criminal Justice*, *47(2)*, pp.427-446. DOI: 10.3138/cjccj.47.2.427.

Prithi, S., Aravindan, S., & Anusuya, E. A.K. (2020). Gui Based Prediction of Crime Rate Using Machine. pp.221–229.

SAPS. (2021). *Crime statistics 2021/2022*. South African Police Service. Retrieved 26/03/,2022, https://www.saps.gov.za/newsroom/msspeechdetail.php?nid=38194

Schneider, S. (2002). Predicting Crime: The Review of Research. *Department of Justice Canada*, pp.1-37.

Schonteich, M., & Louw, A. (2001). Crime in South Africa: A country and cities profile. *Institute for Security Studies Papers*, *2001*(49).

Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art, 4(1)*, pp.1-14. DOI: 10.1186/s42492-021-00075-z.

Shojaee, S., Mustapha, A., Sidi, F., & Jabar, M. A. (2013). A study on classification learning algorithms to predict crime status. *International Journal of Digital Content Technology and its Applications*, *7*(9), p.361.

Sun, C. C., Yao, C., Li, X., & Lee, K. (2014). Detecting Crime Types Using Classification Algorithms. *J. Digit. Inf. Management*, *12*(5), pp.321-327.

Tseloni, A., Mailley, J., Farrell, G., & Tilley, N. (2010). Exploring the international decline in crime rates. *European Journal of Criminology*, *7*(5), pp.375-394.

Vallat, R. (2018). Pingouin: statistics in Python. *J. Open-Source Software.*, *3*(31), p.1026.

Vural, M. S., & Gök, M. (2017). Criminal prediction using Naive Bayes theory. *Neural Computing and Applications*, *28*(9), pp.2581-2592.

Xu, H., Zhou, J., G. Asteris, P., Jahed Armaghani, D., & Tahir, M. M. (2019). Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate. *Applied sciences*, *9*(18), p.3715.

Yiu, T. (2019). Understanding random forest. *Towards data science*, 1-11. Retrieved May 04, 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

Zhang, X., Liu, L., Xiao, L., & Ji, J. (2020). Comparison of machine learning algorithms for predicting crime hotspots. *IEEE Access*, *8*, pp.181302-181310.