2024

A Model for mapping crime hotspots using neural networks: a case of Nairobi.

Echessa, Rita Gloria
School of Computing and Engineering Sciences
Strathmore University

A Model for Mapping Crime Hotspots using Neural Networks: A Case of Nairobi

Echessa Rita Gloria

96393

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Information Technology (Msc. IT) at Strathmore University

School of Computing and Engineering Sciences, Strathmore University, Nairobi, Kenya.

April 2024

Declaration

I, Rita Gloria Echessa, hereby declare that the work presented here in is original work done by me and has not been published or submitted elsewhere for the requirement of a Master of Science in Information Technology. Any literature date or work done by others and cited within this thesis has been given due acknowledgement and listed in the reference section.

Signature…… Date…… 05/04/2024……...

Name Rita Gloria Echessa Registration Number: 96393

SUPERVISOR'S DECLARATION

This research proposal has been submitted for review with my approval as a university supervisor.

Signature…… Date…… 05/04/2024…… ……...

Name Dr Joseph Orero

School of Computing and Engineering Sciences,Strathmore University

Table of Contents

Declaration 2

Table of Contents 3

Dedication 8

Acknowledgement 9

Abbreviations/ Acronyms 10

Definition of Terms 11

Abstract 12

List of Figures 13

List of Equations 15

List of Tables 16

Chapter 1: Introduction 1

1.1 Background of the Study 1

1.2 Problem Statement 2

1.3 Research Objectives 3

1.3.1 Specific Objective 3

1.4 Research Questions 3

1.5 Justification of the Study 3

1.6 Scope and Limitation 4

Chapter 2: Literature Review 5

2.1 Introduction 5

2.2 Crime patterns in Kenya 5

2.3 Related Works 6

2.3.1 Twitter sentiment analysis tool for detecting crime hotspots: a case of Nairobi, Kenya. 6

2.3.2 Crime mapping using Artificial Intelligent Agents in Nairobi City County, Kenya 7

2.3.3 Machine Learning Approach to Crime Prediction and Identification of Hotspots 8

2.3.4 Spatio-Temporal CrimeHotSpot Detection and Prediction: A Systematic Literature Review 8

2.3.5 A Clustering Based HotspotIdentification Approach forCrime Prediction 9

2.4 Models 10

2.4.1 Space-Time Permutation Model 10

2.4.2 Geographic Information System 11

2.4.3 Push-Pins 13

2.5 Algorithms 13

2.5.1 Support Vector Machine (SVM) 13

2.5.2 Random Forest-Based Predictive Models 14

2.5.3 Neural Network (NN) 15

2.5.4 Naive Bayesian 17

2.5.5 Decision Tree 18

2.5.6 K-Means 20

2.6 Systems/Applications 21

2.7 Conceptual Framework 22
Chapter 3: Research Methodology 23
3.1 Introduction 23
3.2 Research Design and Philosophy 23
3.3 Population and Sampling 23
3.3.1 Population 23
3.3.2 Sampling 23
3.4 Data Collection Methods/ Analysis 24
3.4.1 Data Collection 24
3.4.2 Data Analysis 24
3.5 Research Quality and Reliability 24
3.5.1 Data Reliability. 26
3.6 System Development Methodology 27
3.6.1 Requirements Planning 27
3.6.2 User Design 27
3.6.3 Rapid Construction 27
3.6.4 Cutover 28
3.7 Utilization and Dissemination of Research Results. 28
3.8 Ethical Considerations/ Issues. 28
Chapter 4: System Analysis, Design and Architecture 29
4.1 Introduction 29
4.2 System Analysis 29

4.2.1 Requirement gathering 29
4.2.2 System Architecture 32
4.3 System Design 32
4.3.1 Use Case Diagram 32
4.3.2 Sequence Diagram 37
4.3.3 Context Diagram 37
4.3.4 Data Flow Diagram (Level One) 38
4.3.5 Wireframes of the system 39
Chapter 5: System Implementation and Testing 40
5.1 Introduction 40
5.2 Model Development 40
5.2.1 X Data Scrapping 40
5.2.2 Preprocessing 42
5.2.3 Identifying tweets 44
5.2.4 Location Identification 45
5.2.5 Neural Network training 45
5.2.6 Testing the model 47
5.2.7 Using the model to detect crime hotspots 49
Chapter 6: Discussions 51
6.1 Other classifiers experiments 51
6.1.1 Support Vector Machine Classifier (SVM) 51
6.1.2 Random Forest 52

6.1.3 Naive Bayesian 53

6.1.4 Decision Tree 55

6.1.5 K- Nearest Neighbor 56

6.1.6 Comparison results of the other classifiers 57

6.2 Discussions 57

Chapter 7: Conclusion and recommendations 59

7.1 Conclusion 59

7.2 Recommendations 59

7.3 Further work 60

References 62

Appendices 69

Appendix A: Ethical Approval 69

Dedication

This thesis proposal is dedicated to my mom, who has been a constant source of support and encouragement during the challenges in life. It is also dedicated to my dad for being here for me throughout every step and supporting me. I am truly thankful for having you both in my life and loving me unconditionally. Both of you have been my best cheerleaders. I also dedicate this thesis to my sisters, Lavender and Praise, for their continuous support throughout this journey and all my friends, who have been of great source.

Acknowledgement

I would like to express my sincere gratitude to all those who have contributed to the completion of this project. First and foremost, I am deeply thankful to my supervisor, Dr. Joseph Orero, for their invaluable guidance, support, and encouragement throughout this journey. Their expertise, patience, and constructive feedback have been instrumental in shaping the direction and quality of this project. I am also grateful to Strathmore University, for providing the necessary resources and facilities to conduct this research. The access to the library has greatly facilitated the smooth execution of the project. I would like to acknowledge the participants of this study, whose involvement and cooperation have been vital in gathering the data and insights necessary for this project. Lastly, I would like to express my heartfelt thanks to my family and friends for their unwavering support, understanding, and encouragement throughout this endeavor.

Thank you all for being part of this journey and for your invaluable contributions.

Abbreviations/ Acronyms

CNN	-	Convolutional Neural Network
COV	-	Coefficient of Variation
FN	-	False Negative
FP	-	False Positive
GIS	-	Geographic Information System
KDE	-	Kernel Density Estimation
MAE	-	Mean Absolute Error
MAPE	-	Mean Absolute Percentage Error
ME	-	Mean Error
NN	-	Neural Network
PWA	-	Progressive Web Application
RAD	-	Rapid Application Development
RMSE	-	Root Mean Square Error
SOM	-	Self-Organizing Maps
SVM	-	Support Vector Machine
TN	-	True Negative
TP		True Positive
VADAR	-	Valence Aware Dictionary for Sentiment Reasoning

Definition of Terms

Crime	An offence that violates the state's law and is disapproved by society can be punished by imprisonment or fine (CAP. 63, 2014).
Felony	An offense that is classified as a felony or, if not, a misdemeanor by law is punishable by death or by imprisonment for three years or more even without proof of a past conviction (CAP. 63, 2014).
Misdemeanor	A crime less serious than a felony (Definition of MISDEMEANOR, n.d.)
Subversive activities	Is when someone commits an offense when they carry out a subversive act, attempt to carry out a subversive act, prepare to carry out a subversive act, conspire to carry out a subversive act, or utter a subversive act (CAP. 63, 2014).

Abstract

Since the inception of the first modernized police agency, the primary objective of police organizations has been to prevent crime. Law enforcement, police, and crime reduction agencies commonly used hotspot mapping, an analytical technique, to visually determine the locations where a crime was most prevalent. This assisted in decision-making to determine the deployment of resources in target areas. This study aimed to investigate crime mapping techniques in crime analysis and suggest ways to enhance the implementation of crime mapping in Nairobi. Beginning with a historical analysis of GIS and crime mapping, the study then moved on to a consideration of the significance of geography in dealing with crime concerns. Neural Networks and K-Means machine learning models were used, and data was collected through quantitative and qualitative means in two phases. X was utilized in the first phase to collect information from the general public and important informants. The second phase involved collecting crime hotspot coordinates using a participative Geographic Information System. The study focused on utilizing social media data and machine learning techniques, particularly the KMEANS with NN (Neural Network) model, to identify and map crime hotspots in Nairobi. By analyzing crime-related tweets and categorizing them as either positive, negative or neutral using this NN (Neural Network) model then clustering them as either high risk or low risk using K-Means, the study achieved high accuracy, precision, recall, and F1-Score, suggesting the effectiveness of this approach for crime prediction and prevention.

Keywords: Hotspot mapping, X, Machine learning, crime, Neural Networks, K-Means

List of Figures

Figure 2.1:San Francisco dataset and K-Means clustering, evaluate classification methods 10

Figure 2.2:SVM 14

Figure 2.3:Random Forest Model 15

Figure 2.4:NN Supervised Learning Architecture Sample 17

Figure 2.5:Decision Tree 19

Figure 2.6:Conceptual Framework 22

Figure 3.1:RAD Methodology 27

Figure 4.1:System Architecture 32

Figure 4.2:Use case diagram 33

Figure 4.3:Sequence Diagram 37

Figure 4.4:Context Diagram 37

Figure 4.5:Level One Data Flow Diagram 38

Figure 4.6:Wireframes of the system 39

Figure 5.1:Code to scrap data 41

Figure 5.2:Code to scrap data 41

Figure 5.3:Raw scrapped data 42

Figure 5.4:Cleaning the data 43

Figure 5.5:Sample categorization 44

Figure 5.6:Classifying tweets 44

Figure 5.7:Vectorizing the tweets 46

Figure 5.8:NN training 46

Figure 5.9:ROC for K-means and NN 49

Figure 5.10:Search location page 50

Figure 6.1:ROC for SVM 52

Figure 6.2:ROC for Random Forest 53

Figure 6.3:ROC for Naïve Bayesian 54

Figure 6.4:ROC for Decision Tree 55

Figure 6.5:ROC for KNN 56

Figure 6.6:Performance of NN 58

List of Equations

Equation 2.1:Posteriori probability 18

Equation 2.2:Information Gain 19

Equation 2.3:Entropy 20

Equation 2.4:K-Means 20

Equation 3.1:Accuracy Ratio 26

Equation 3.2:Precision Ratio 26

Equation 3.3:Recall Ratio 26

Equation 5.1:Final score 45

Equation 5.2:Precision Formula 47

Equation 5.3:Calculated precision 47

Equation 5.4:Recall formula 48

Equation 5.5:Calculated recall 48

Equation 5.6:F-Measure Formula 48

Equation 5.7:Calculated F-measure 48

List of Tables

Table 2.1:Comparison of various systems 21

Table 4.1:Non- Functional Requirements 31

Table 4.2:Login Use Case 34

Table 4.3:Search Location Use Case 35

Table 5.1:Categorization of tweets 43

Table 5.2:Results of training 47

Table 6.1:SVM performance 51

Table 6.2:Random Forest performance 52

Table 6.3:Naïve Bayesian performance 53

Table 6.4 :Decision Tree performance 55

Table 6.5:KNN performance 56

Table 6.6:Performance comparison of other classifiers 57

Table 6.7:Performance of K-Means and NN 58

Chapter 1: Introduction

1.1 Background of the Study

Safety and security are the second human needs in Maslow's hierarchy. Aruma and Hanachor (2017) highlighted that individuals tend to seek environments that offer safety and security from unpredictable and hazardous situations such as communal crises, conflicts, wars, civil disturbances, riots, militancy, terrorism, kidnapping, armed robbery, and killings. These types of disruptions can significantly affect peaceful coexistence within communities, making it essential for individuals to prioritize safety and security. According to Durkheim's functionalistic theory, crime is inevitable and normal. However, in recent years, researchers and criminologists have emphasized the potential benefits of centralizing crime prevention efforts at crime scenes (Braga et al., 2019). Studies have shown that crime tends to concentrate in specific locations or 'hot spots,' which make up about half of all crimes. Braga et al. (2019) suggest that hot spot policing, a proactive and targeted policing strategy, can significantly reduce crime rates in high-crime areas. The authors conducted a systematic review of 25 studies on hot spot policing and found that it was consistently associated with crime reduction, particularly for violent and drug-related crimes. They also found that problem-oriented policing, which involves addressing the underlying causes of crime in hot spots, was more effective than traditional forms of policing.

The authors conclude that hot spot policing is a promising strategy for reducing crime, but it must be implemented in a thoughtful and evidence-based manner. This includes focusing on high- risk locations and offenders, using data to inform policing strategies, and engaging with community stakeholders to build trust and support. Overall, the findings suggest that hot spot policing can be an effective tool for improving public safety in high-crime areas. Crime pattern theory combines rationality and frequency activity theory to show the distribution of crime across different locations (Eck, Weisburd, & University, 2015). Mugging, kidnapping, car-jacking, bag-snatching in transport hubs, and armed robbery are regular occurrences in Nairobi, Mombasa, and other cities in Kenya.

Identifying where crime was committed is crucial in addressing the crime problem (Chainey, 2014). Crime does not occur randomly and is not evenly distributed across regions (Kedia, 2016).

Crime tends to focus on specific areas due to the interactions between victims and offenders and the availability of criminal opportunities (Campedelli, Aziani, & Favarin, 2021). These clusters of crime are known as hotspots (Shiode & Shiode, 2020). Hotspot mapping is a common analytical tool used by law enforcement, police, and crime reduction agencies to visually identify the most frequent regions for crime. Security technology has evolved significantly in the twenty-first century. This method aids in decision- making on resource deployment and targeting (Xiong et al., 2019). However, despite all the growth and advancement in cities worldwide, chronic insecurity, violence, and corruption remain a challenge, frequently linked to crime outside urban boundaries (Sarkar & Anup, 2017).

Despite the efforts of crime prevention units and security organizations, the volatility of crime remains a challenging issue in Kenya. Crime data from 2018 (Crime data, 2018) show that Nairobi has a higher crime rate than other Kenyan counties and cities. Street crimes involving multiple armed assailants and crimes of opportunity, such as pickpocketing and snatch-and-grab attacks in vehicles and crowded areas are a challenge in major Kenyan cities (Samoei, 2018). Technologies such as data mining systems, biometrics, and GPS tracking have been adopted to detect, investigate, and prevent crime in society (Samoei, 2018). The 21st-century policing models value accuracy and strive for real-time information collection for strategic planning, problem analysis, decision-making, threat detection, accountability, and many more functions (Samoei, 2018). However, the police often lack modern technological equipment and resources to respond to and monitor calls or emergencies (Samoei, 2018).

1.2 Problem Statement

The high population in cities in Kenya constrained resources and infrastructure, which negatively impacted citizens' security and privacy (Samoei, 2018). Crime affected economic stability, social welfare, and business operations in the affected areas (Braga et al., 2019). The ongoing increase in crime incidents in Nairobi County remained a challenge (Baraka & Murimi, 2021). Despite technological advancements, the police force in Nairobi continued to rely on pin maps for crime mapping as a crime prevention measure (Baraka & Murimi, 2021).

1.3 Research Objectives

The main aim of study is to develop a model for mapping crime hotspots using Neural Networks and K-Means in Nairobi County.

1.3.1 Specific Objective

i. To analyze crime patterns in Kenya.

ii. To review existing techniques used to map crime hotspots.

iii. To develop a model to map crime hotspots.

iv. To validate the model.

1.4 Research Questions

i. What are the crime patterns in Kenya?

ii. What are the existing techniques used to map crime hotspots?

iii. How to develop a model to map crime hotspots?

iv. How can the model be validated?

1.5 Justification of the Study

According to (United Nations, 2018), by 2018, 55% of people worldwide lived in cities, which was approximately 4.2 billion, and was expected to grow to 68% by 2050. Crime in urban areas was considered higher than in rural areas due to several factors that were added in urban areas compared to rural areas, such as homelessness, deprivation, poverty, unemployment, youth marginalization, poor urban planning, neglect of public spaces, and disintegration of culture (Onyeneke & Karam, 2022). Although social disadvantages did not directly cause crime, they did play a part in directing vulnerable citizens' criminogenic exposure and predisposition for crime (Onyeneke & Karam, 2022).

Crime had evolved proportionally to the rise of technology. Cybersecurity and terrorism are just

some of the crimes that have increased, with scarce resources to address them urgently (Baraka & Murimi, 2021). The correlation between the crime's occurrence and the crime's location has been reported and documented over the years (Mbani, Odera, & Kenduiywo, 2017). The efficiency of the police force and law enforcement officers increased greatly on hotspots (Mbani, Odera, & Kenduiywo, 2017). In the postmodern society, managing and controlling crime remained a major challenge as criminals continued to advance and evolve their techniques, making security management complex (Baraka & Murimi, 2021).

1.6 Scope and Limitation

The research assessed crime hotspots in Nairobi County. The study was restricted to crimes that occurred in Nairobi County. Even though the crime was pervasive throughout the nation, Nairobi was selected because of its high frequency of criminal behavior. The study only mapped and evaluated crimes against people and property. The report did not include white-collar offences like corruption and gender-based abuse. This was because acquiring such data might have been challenging.

Chapter 2: Literature Review

2.1 Introduction

One of the main responsibilities of any government worldwide is to protect its residents and their property because no other institution has the tools to deal with violence (Samoei, 2018). Despite their critical function, urban centers worldwide saw an increase in criminal activity (Samoei, 2018). Since the founding of cities itself, authorities and governments have been concerned about the higher crime rates in urban areas since crime was a serious danger to human security. Urban cities were bound to security and privacy challenges that affected citizens and the government (Samoei, 2018). Kenya had four major cities: Nairobi, Mombasa, Kisumu, and Nakuru. The greatest threats posed to these cities tended to be crime and road safety (Samoei, 2018).

In Nairobi, street crime that mostly involved armed attacks was a serious problem; pickpockets, snatch-and-grab attacks in crowded areas, idle traffic, and crimes of opportunity (Masiga, 2022) were other common crimes that occurred. Criminal activities could easily escalate to mob violence by crowds of street criminals (Skilling & Rogers, 2017). According to the Kenya National Police Service crime statistics reports (Crime Statistics, 2018), since 2014, Nairobi County has been leading in crime rate. The continuous pattern shift of criminal activities in urban areas, including Nairobi, defied the state and non-state efforts towards controlling and managing them (Masiga, 2022).

2.2 Crime patterns in Kenya

Even though crime occurs almost everywhere in the world, it tends to cluster in certain areas of a city and neighborhoods that the police and residents are well acquainted with (Skilling & Rogers, 2017). Identifying where crime tended to cluster in space and time through timely mapping of crime scenes and precise spatial concentration detection was crucial for law enforcement's attempts to reduce crime (Baraka & Murimi, 2021). According to (Catlett et al., 2018), crime rates were constantly shifting and changing due to factors such as geographical location of an area, which could be categorized either as high risk or low risk, and crime trends that differed based on seasonal

patterns. According to Mburu (2015), densities of illegal acts were induced by specific stimuli such as drug trafficking, weather patterns, or seasons. In the context of the time of the offence, studies showed that crime spiked mostly in the evening,at night, on holidays and in isolated areas over the weekend (Mburu, 2015).

The examination of crime patterns in Kenya reveals a spectrum of prevalent offenses, as highlighted in the "2020 National Crime Mapping: Crime Patterns and Trends in Kenya" report. Among the reported crimes are stealing, burglary, possession of illicit brews, disorderly conduct, gender-based violence, stock theft, assault, narcotic possession, defilement, and robbery. Victims encompass women, children, youth, and men, with women and children being particularly vulnerable. Root causes of crime include unemployment, substance abuse, idleness, and poverty. Crime peaks during early evening hours, often involving weapons like pangas, machetes, knives, and swords. Consequences of crime extend to property loss, fatalities, heightened fear, increased poverty, and hindered economic growth. Existing preventive measures, such as community policing, patrols, arrests, timely reporting, lighting, security meetings, and civic education, vary in efficacy.

2.3 Related Works

Based on machine learning techniques, several detection and prediction models have been created to aid in mapping crime hotspots. These models assist in predicting the crime pattern in a location based on the interaction with various variables and are trained using the available crime data.

2.3.1 Twitter sentiment analysis tool for detecting crime hotspots: a case of Nairobi, Kenya.

Onyango (2021), utilized X (formerly known as twitter) in crime hotspot detection in Nairobi by subjecting unstructured data to statistical analysis. The study aimed to fashion a machine-learning tool capable of detecting crime hotbeds in Nairobi utilizing data from X. To accomplish this, the X API retrieved tweets based on location concerning violence in Nairobi. The obtained text data set underwent preprocessing to be classified as either positive, negative, or neutral before being divided into a training set and a test set. To train the model, Support Vector Machine (SVM) was utilized. It was found to perform with the highest accuracy using bigram features compared to other classifiers

include Naive Bayes, Random Forest, and K-nearest Neighbor. The model was applied to classify similar tweets retrieved using the X search API. However, the study's limitation was that tweets in languages other than English and Swahili were not considered, also video, emotions and image-based content were not factored.

2.3.2 Crime mapping using Artificial Intelligent Agents in Nairobi City County, Kenya

BO et al. (2017) introduced an agent-based approach for crime mapping in Nairobi City County, utilizing Artificial Intelligent (AI) Agents. The study primarily focused on three agents, namely the offender, target, and guardian. The research employed the q-learning algorithm, which is part of the reinforcement learning technique. The reinforcement learning approach is driven by underlying Markov chains, where the system randomly transitions to another state in discrete time steps. Q-learning is a model-free method that enables agents to learn how to make the best decisions possible by exposing them to the consequences of their actions. The agent seeks to perform an action when it is in a specific state as a result of the learning. The outcome of such an action is weighed against the immediate benefit it receives upon successfully completing an episode. In contrast, a penalty is imposed when an episode is not successfully completed. In this manner, the agents try all potential actions in each state, i.e., each iteration tries a conceivable action, learning which action-state pairs are the best overall.

It should be noted that the q-learning approach makes no explicit attempt to specify the action-state pairings taken by agents. On the contrary, the agents are free to pursue any actions they like, implying that the q-learning algorithm allows the agents to explore through arbitrary experimentation while keeping the best estimate of an action-state combination in a matrix. Agents are not restricted to specific actions; they are allowed to explore any action they would like. Despite q-learning allowing agents to pursue unlimited options, the q-learning approach fails to define the action-state pairs taken by the respective agents.

Research conducted by (BO et al., 2017) examined crime occurrence as a result of both socio-physical and human behavior. Input factors for crime to occur considered during the research included land use, population density, poverty rate, access to street lighting and access to transport network. Spearman rank correlation coefficient was used for comparison, and a positive rank of

0.4 was achieved. The gaps in the study were: the research focused on street robbery, and yet it is not the only crime; guardians were not subjected to learn from their mistakes; instead followed the normal routine, and the victims were restricted with no free movement.

2.3.3 Machine Learning Approach to Crime Prediction and Identification of Hotspots

Singh (2021) suggested using a machine learning method to forecast crime and identify hotspots. The project delves into the crime dataset for the city of Denver and attempts to identify hotspots and types of crime occurring in the city. The FP-growth algorithm helps to identify the location and time at which crime would be likeliest to occur in the city, which can help in police monitoring and faster response teams in the area. Using the classification algorithm, it is possible to guess the type of crime occurring, which allows the police and the government to take action and better deal with crime, according to Singh (2021).

The neural network model worked best for this purpose, but if the objective is to have reasonable probability over the different categories of crime, random forest is more effective (Singh, 2021).

The existing gap in this model was that it cannot be applied in a real-world scenario. There is clear scope for improvement in utilizing the data to make improved predictions which can be achieved through adding different types of data as variables, improve the association rule mining in such a way that trends in the crime hotspot mapping can also be observed, and another way the model can be improved is by using the tonality feature over a period of days rather than using just one day as a feature (Singh, 2021).

2.3.4 Spatio-Temporal CrimeHotSpot Detection and Prediction: A Systematic Literature Review

Spatial and temporal crime hotspot identification was suggested by Butt et al. (2020). The concept contends that improvements in geographic information systems create new opportunities for crime hotspot detection by incorporating geographical and temporal data into crime databases. Furthermore, it makes it possible for researchers to locate hotspots and swiftly analyze and visualize the corresponding change. The method of clustering and classification was shown to be more useful in locating crime hotspots. Recently, it has been reported and proven that the Random Forest and DBSCAN algorithms are effective and productive in comparison to modern techniques (Butt et

al., 2020).

Researchers are, however faced with the accuracy-related gaps created by these techniques. The hot spot crime detection method must have the scalability to allow the method to deal with demographic, sparsity and population characteristics, among other things; however, some drawbacks were identified during the study (Butt et al., 2020). Extensive research has been conducted on crime prediction strategies, including the use of various data mining and machine learning techniques. However, these methods have not demonstrated success in real-world applications. To use crime hotspot detection in the real world, they want to find a way to get over its current limitations (Butt et al., 2020).

2.3.5 A Clustering Based HotspotIdentification Approach forCrime Prediction

In their study, Hajela et al. (2020) propose a clustering-based approach for crime prediction. The model consists of three phases: identification of crime hotspots, preparation of the crime history dataset, and crime prediction. The researchers test their model using the San Francisco crime dataset, considering factors such as location and time of occurrence during the identification phase. To identify clusters, present in the dataset and detect hotspots in the area, the model adopts the K-Means clustering approach, which allows for flexibility in selecting the number of clusters needed.

K-Means was found to have the highest accuracy when compared to other ensemble learning algorithms such as Naive Bayes (NB), Naive Bayes with kernel estimator (NB -k), Decision Tree (REPTree), and classifier.

Approach	Accuracy (in %)					
	k=9	k=8	k=7	k=6	k=5	k=4
NB	97.72	97.75	97.765	97.787	97.825	97.835
NB -k	97.72	97.75	97.765	97.787	97.825	97.835
REPTree	99.775	99.775	99.775	99.775	99.775	99.775
Bagging (NB)	97.66	97.657	97.647	97.697	97.78	97.757
Bagging (NB -k)	97.66	97.657	97.647	97.697	97.78	97.757
Bagging (REPTree)	99.772	99.772	99.772	99.772	99.772	99.772
Voting (NB)	97.72	97.75	97.765	97.787	97.825	97.835
Voting (NB -k)	97.72	97.75	97.765	97.787	97.825	97.835
Voting (REPTree)	99.775	99.775	99.775	99.775	99.775	99.775
Voting (NB + REPTree)	99.775	99.77	99.775	99.78	99.777	99.78
Stacking (NB)	94.615	94.55	94.7	94.652	94.765	94.785
Stacking (REPTree)	99.8	99.8	99.8	99.8	99.8	99.8
Stacking (NB+REPTree, meta=NB)	97.205	97.142	97.275	97.23	97.277	97.287
Stacking(NB + REPTree,meta=REPTree)	99.8	99.795	99.8	99.8	99.8	99.8

Figure 2.1:San Francisco dataset and K-Means clustering, evaluate classification methods

(Mahmud et al., 2020)

It is concluded that K-means clustering improves crime prediction accuracy and suggested future models to incorporate other clustering approaches to improve the results further.

2.4 Models

2.4.1 Space-Time Permutation Model

The primary method used for preventing crime is the spatial-temporal method, which has historically been used in conjunction with Kernel Density Estimation (KDE) and Time Series methods. These methods, however, only consider space or time independently, while crime is influenced by multiple interdependent factors. To identify dense event patterns and statistically assess if they deviate from chance occurrences, space-time cluster identification is employed (Hardie & Wikström, 2021). Numerous study areas, including criminology, epidemiology, demography, regional science, and geography, depend on this method. There are four basic subcategories of space-time cluster detection: permanent spatial clustering, permanent spatial clustering within a time window, permanent space-time clustering, and permanent space-time interaction. When a region is predisposed to particular events, spatial clustering always occurs, and clustering persists over time.

When hotspots are only present at certain times of the day, they tend to cluster spatially within a given time frame. When numerous events take place rapidly in a crowded area, space-time clustering happens. The connection between space and time is sporadicand only manifests occasionally. The relationship between space and time becomes more complex when there is a space-time interaction (Deb & Tsay, 2019). This is when a concentrated cluster begins to diffuse into non-concentrated areas, and the spread may occur at different seasons of the year.

The spatial-temporal method is the everyday method used to prevent crime. Previously, spatial-temporal has been used alongside KDE and Time Series methods. The KDE method only considers space independent of the temporal variations, while the Time Series method only takes time independently into consideration even though crime is a result of several interdependent factors.

When a set of events are densely gathered, as well as whether or not the dense event pattern deviates from a chance occurrence, can be determined using the space-time cluster identification technique (Hardie & Wikström, 2021). This method is crucial in a number of research areas, including geography, criminology, epidemiology, and demography. An essential and fundamental component of a spatial-temporal phenomenon can be revealed via space-time cluster detection (Hardie & Wikström, 2021).

The space-time cluster identification model can be divided into four main categories: spatial clustering that occurs constantly, spatial clustering that occurs only during a certain time, space-time clustering, and space-time interaction (Liu et al., 2012).

Spatial clustering all the time occurs when a region is predisposed to certain occurrences and lacks space-time interaction because clustering happens constantly (Liu et al., 2012). Hotspot clustering happens at specified time periods when there is spatial clustering (Belhadi et al., 2020). Space-time clustering occurs when numerous events are anticipated to take place in a brief period of time in a crowded area, and the connection between time and space only briefly manifests itself at specific times (Belhadi et al., 2020).

2.4.2 Geographic Information System

Research conducted by (Dağlar & Argun, 2016) shows for a very long period, GIS was only

available manually. The 1871 Battle of Yorktown map was created by French geographer Louis Alexander Berthier and featured hinged overlays that showed military movements. While these manual information systems had their drawbacks, they were nonetheless valuable at the time. However, using them was challenging due to a few issues. In addition, the data was not always current, and there were no defined scales. Dağlar & Argun (2016) provide a good example of how early crime mapping implementations were used in their research. Police departments have long used pin maps to analyze geographic information patterns. These maps typically have pins placed on them to indicate criminality. These maps can offer officers a lot of useful information about the distribution and density of crime.

However, GIS can hold as much data as the available computer technology will allow. The data may be combined, added to, subtracted from, multiplied, divided, and many other similar processes using GIS. The police may be able to provide highly complex answers using GIS. Additionally, the technology integrates geographic data with crime data. Like any excellent technology or application, crime mapping and GIS have their share of issues and challenges. The drawbacks or contentious issues and issues can be listed as high implementation costs, resistance-police-culture considerations, challenges in determining the precise locations of crimes, accurate data considerations, invasion of privacy violations, detrimental effects on social and economic life, consequences of "spatial labeling," and some technical challenges (Dağlar & Argun, 2016).

The use of GIS and crime mapping technology might be challenging due to the high cost of GIS and some hostile police cultures. The cost of implementing high technology is constant. Dalar & Argun (2016), Insist that economic concerns are the primary considerations when selecting whether to use GIS or crime mapping. This can be a significant challenge for the development of this crucial technology, particularly for small agencies who cannot afford to use GIS and crime mapping. There are also some cultural and political considerations for some police agencies. These police cultures oppose using the technology due to resistance and a general lack of technical competence. They made the decision to rely more on personal expertise than on data based on statistics. Two further issues with GIS and crime mapping are pinpointing the precise site of crimes and obtaining accurate information about them. It is crucial to document the precise location of crimes (Dağlar & Argun, 2016).

2.4.3 Push-Pins

Point mapping is the method used most frequently to show geographic patterns of crime (Baraka & Murimi, 2019). The major reason point mapping is so popular is that it is a straightforward digital adaptation of the well-known and traditional technique of putting pins denoting criminal incidents onto a wall map. Baraka & Murimi (2019) say in a digital application, groupings of points satisfying certain requirements may be easily and rapidly picked, provided these individual geographic point objects are adequately ascribed with information, such as codes specifying the kind, date, and time of the infraction. The appropriate symbol for the displayed crime category can then be used to display these choices.

But it can be challenging to discern geographic trends and hotspots in crime point data, especially if the data sets are huge. Furthermore, there may be more than one crime spot when there only seems to be one. It is hard to locate these coincident sites on a point map (Baraka & Murimi, 2019).

Although they have their uses, such as mapping individual criminal acts, modest volumes of criminal activity, and recurrent areas, they can lose some of their effectiveness when trying to locate crime hotspots, especially when dealing with enormous amounts of data.

2.5 Algorithms

2.5.1 Support Vector Machine (SVM)

To successfully solve classification or regression challenges, the SVM-supervised machine learning model is commonly used, although it is most frequently used for classification. Every data point is represented by the SVM method as a point in an n-dimensional space, where n is the number of features that are accessible, and each feature value is associated with a particular position (Sunil, 2019).

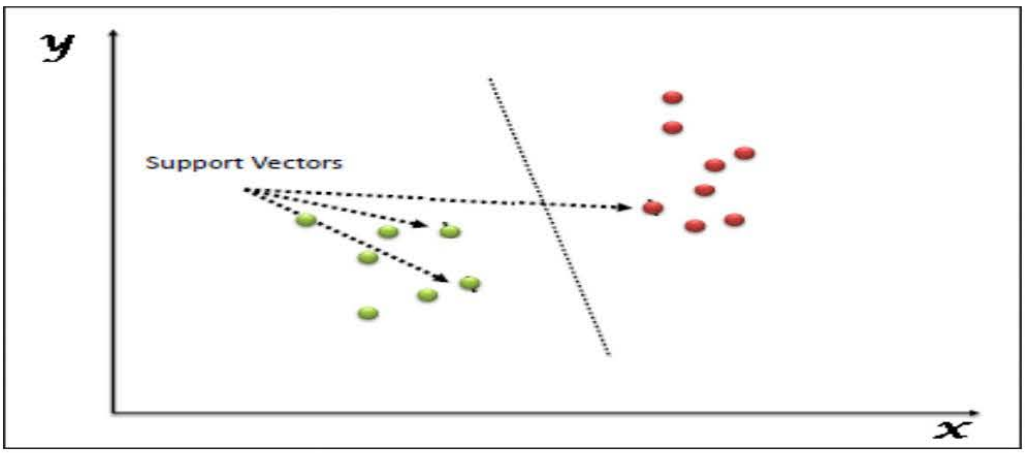


Figure 2.2:SVM

(Sunil, 2019)

To classify data using SVM, a hyper-plane that effectively separates the two classes is identified based on the coordinates of each observation, which form the support vectors. The SVM classifier is the frontier that best separates the two classes, either in the form of a hyper-plane or line.

2.5.2 Random Forest-Based Predictive Models

The random forest is a popular machine learning method that can handle classification and regression tasks by utilizing ensemble learning. By combining multiple decision trees, the algorithm can effectively address complex problems. The random forest algorithm trains a "forest" through bootstrap aggregation, also known as bagging, which increases the precision of machine learning systems (Mbaabu, 2020). Random forest algorithm generates the output informed by the decision tree predictions. The predictions are obtained by aggregating or averaging the outcomes from multiple trees; the number of trees is commensurate to the accuracy, i.e. more trees result in better accuracy. The random forest algorithm bypasses the shortcomings of the decision tree method, such as dataset overfitting and low accuracy (Mbaabu, 2020). Additionally, the random forest algorithm does not require any package parameters, such as those needed in sci-kit-learn.

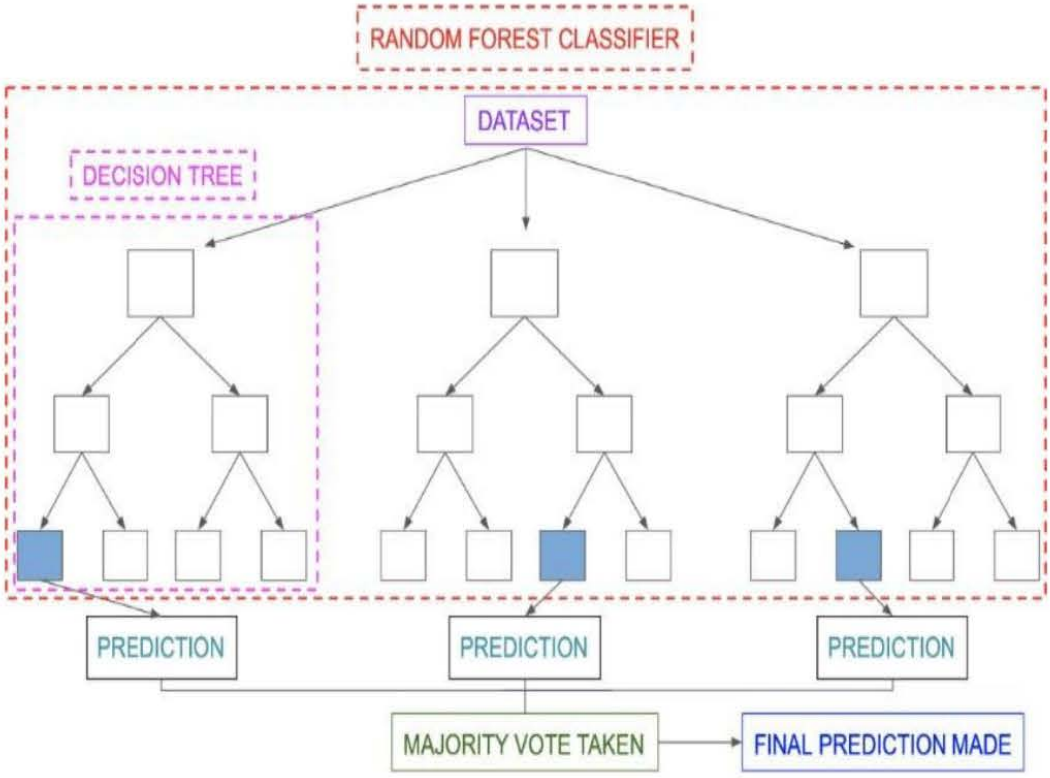


Figure 2.3:Random Forest Model

(Mbaabu, 2020)

Mbaabu (2020) argues that random forest regression is not the optimum method for extrapolating data. Unlike linear regression, which uses historical data to extrapolate values from the observed range. Random Forest also does not produce appropriate results when the amount of data is really low. In this case, an invariant space will be produced by the bootstrapped sample and the subset of features. Ineffective splits will follow from this, which will affect the outcome.

2.5.3 Neural Network (NN)

According to Walczak (2021), NNs are an effective machine- learning technique that can handle prediction and classification problems of any complexity since they learn from data. There are three main learning techniques used by neural networks: supervised learning reinforcement learning and unsupervised learning. Supervised learning involves training the neural network using labelled

data, where the historical data has known outcomes that the neural network tries to predict. On the other hand, unsupervised learning entails training the neural network without labelled data, which allows the network to discover patterns and relationships in the data on its own. For each type of NN learning, there are numerous training algorithms accessible. For supervised learning, backpropagation is the most used technique, although self-organizing maps (SOM) are frequently employed for unsupervised learning. Both methods require that the NN be given learning opportunities. Convolutional NN (CNN), for example, combines supervised and unsupervised learning in hybrid approaches.

A neural network, according to Walczak (2021), is made up of an input layer that identifies the training input variables and an output layer that displays the anticipated classification or prediction result. One or more hidden layers with weighted connections to the layer above them are present in both supervised and hybrid models. To learn, the neural network computes the difference between the predicted output and the actual value, back propagates the difference through the network, and modifies the connection weights to improve the correspondence between the prediction and the actual output.

To ensure that there is no implicit bias in the NN model and to simulate its potential application in real-world scenarios, NN models are validated or tested using data that was not used during the model training, as stated by Walczak (2021). The supplied data is typically used to construct a validation set, training set, and test set, each with its own collection of examples and known results. According to Walczak (2021), the training data should be changed to maintain balanced proportions of the classes to be predicted in order to avoid overfitting to more frequent classes and to guarantee that the NN learns all classes equally. To establish a balanced training dataset, this method, called class balancing, includes either under sampling the majority class or oversampling the minority class. This ensures that the NN does not become biased towards more frequent classes and that it is equally effective in predicting all classes. Figure 2.4 depicts a sample architecture of a supervised learning NN.

(Walczak, 2021)

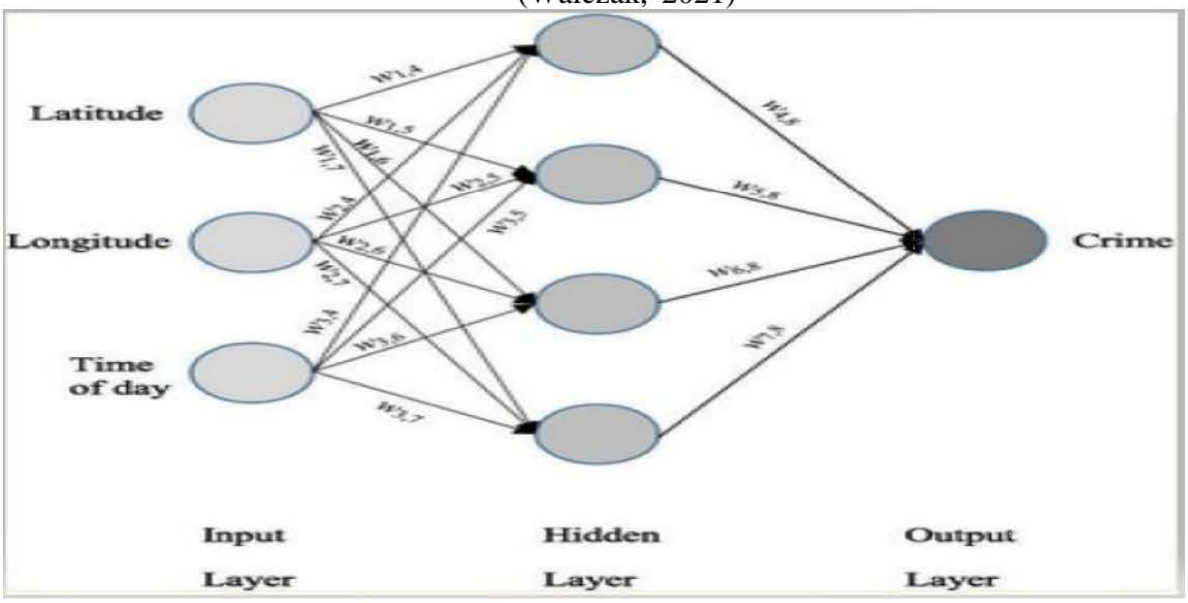


Figure 2.4:NN Supervised Learning Architecture Sample

2.5.4 Naive Bayesian

The Naive Bayes method is a classification technique that can estimate the likelihood of belonging to a particular class (Khan, Ali, Alharbi, & Farias, 2022). In the Naive Bayes approach, the

probability of a class is calculated independently, without considering the influence of other variables. This makes the Naive Bayes classifier highly scalable, as it only requires a linear number of parameters based on the number of variables in a given learning problem.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

X : data with unknown class
H : hypothesis of data X
$P(HIX)$: posteriori probability
$P(H)$: priori H
$P(XIH)$: probability X based on H hypothesis
$P(X)$: priori X

Equation 2.1:Posteriori probability

Simple Bayes and independent Bayes are two terms that are frequently used to describe naive Bayes models. Compared to other classifiers that use iterative approximations, they are particularly efficient since they can be trained using a closed-form expression that only requires linear time. With the help of equation 2.1, this classification is carried out.

2.5.5 Decision Tree

A decision tree is a structure used in the analysis of problem-solving and the mapping of potential solutions to the problem (Khan, Ali, Alharbi, & Farias, 2022). Because it is simple to understand, decision trees are also among the most often used categorization methods. Decision trees work well in situations where the results are discrete values (Jijo & Abdulazeez, 2021). The primary advantage of employing a decision tree is to analyze and characterize complex decision-making processes in order to make them simpler and easier to understand (Wibowo & Oesman, 2020).

The decision tree is a popular method for gathering data to support decision- making. The process starts with a root node, and the user makes decisions by following the branches to reach the leaf nodes. The final product of building a decision tree is a map of potential decision scenarios and their outcomes.

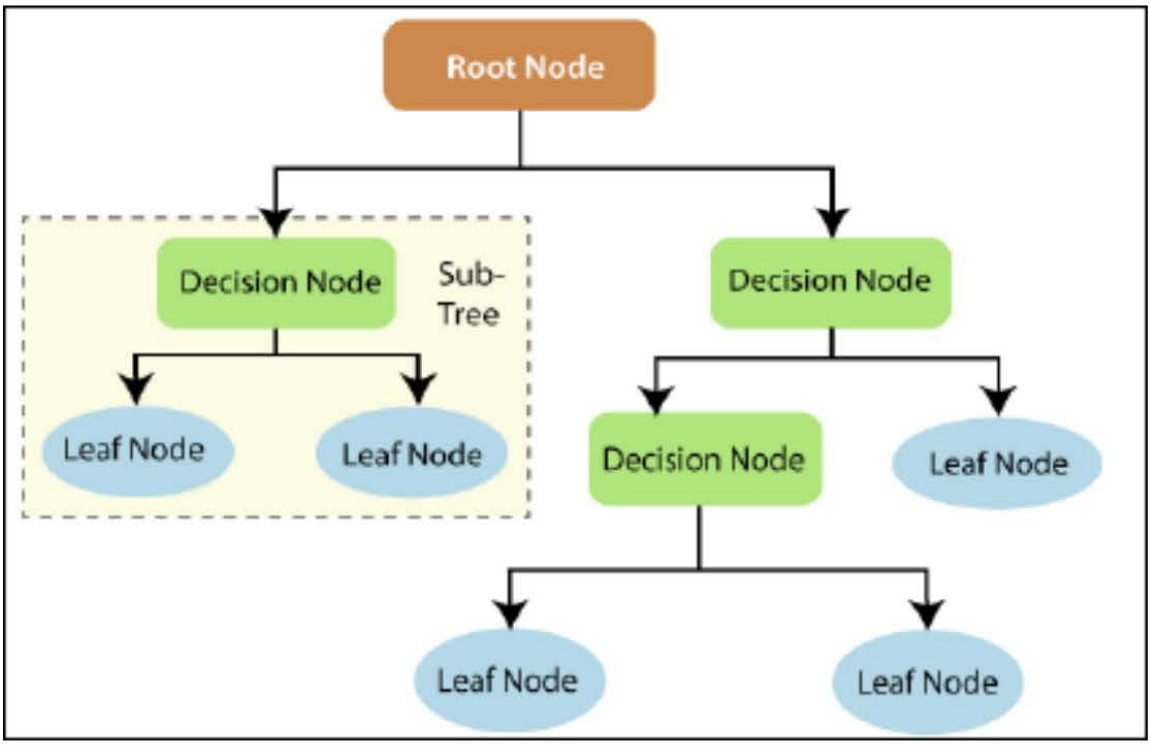


Figure 2.5:Decision Tree

(Jijo & Abdulazeez, 2021)

To construct a decision tree, the first step is to select the attributes for the root node based on their values. This selection is made using a formula that considers all the attributes available.

$$Gain(S,A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \text{Entropy}(S_i)$$

Equation 2.2:Information Gain

S : Case sets
A : Atribut
n : The total number of partition A
$|S_i|$: The total number of cases in partition to i
$|S|$: The total number of cases S

Equation 2.3 shows how to calculate entropy values:

$$Entropy(S) = \sum_{i=1}^{n} -p_i \times \log_2 p_i$$

Equation 2.3:Entropy

Pi: Proportion from Si to S

2.5.6 K-Means

K-means is a technique for grouping K clusters in a dataset. The dataset only contains one cluster for each data point. As noted by Imad Dabbura (2018), the initialization might considerably affect the outcomes, hence the clusters amount, K, must be specified beforehand. The cluster centroid and the data points' sum of squared distances from each other is minimized by the algorithm. The mean of all the cluster's data points is the cluster centroid. As a result, the cluster's variance decreases, enhancing the homogeneity or similarity of its data points. The process starts by randomly calculating K initial centroids, then assigning each data point to the closest centroid to cluster the data into K groups. Next, the centroid of each cluster is determined by taking the mean of all the data points assigned to it. According to Sinaga and Yang (2020), the algorithm repeats assigning data points to the closest centroid and recalculating centroids until the centroid positions stop changing or a predetermined number of iterations is reached. In the end product, the data are divided into K clusters, and each data point is assigned to the cluster with the centroid that is closest to it.

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

Equation 2.4:K-Means

(Sinaga & Yang, 2020)

Wik=1 if data point xi belongs to cluster k; else, wik=0. Furthermore, the cluster's centroid for xiis k.

2.6 Systems/Applications

Table 2.1:Comparison of various systems

Author	Title	Methods	Dataset
(Onyango, 2021)	Twitter sentiment analysis tool for detecting crime hotspots: a case of Nairobi, Kenya.	SVM Model Naïve Bayes Random Forest K-nearest Neighbor	Nairobi
(Singh, 2021)	Machine Learning Approach to Crime Prediction and Identification of Hotspots	The FP-growth algorithm Neural network model	Denver
(BO et al., 2017)	Crime mapping using Artificial Intelligent Agents in Nairobi City County, Kenya	Q-Learning Spearman's Correlation	Nairobi
(Butt et al., 2020)	Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review	Random Forest DBSCAN algorithms	New York
(Hajela et al., 2020)	A Clustering Based Hotspot Identification Approach for Crime Prediction	K-Means Naïve Bayes, Decision Tree	San Francisco

2.7 Conceptual Framework

Data will be collected from X and web UI open data sets available then stored in a database. The data then will go through the machine learning logic, where it will be trained and tested using neural network then clusters as either high risk or low risk area using K-means. The model is to map the type of crime, its frequency, where it happened and the time it occurred to a web interface dashboard. Figure 2.6 shows the conceptual framework.

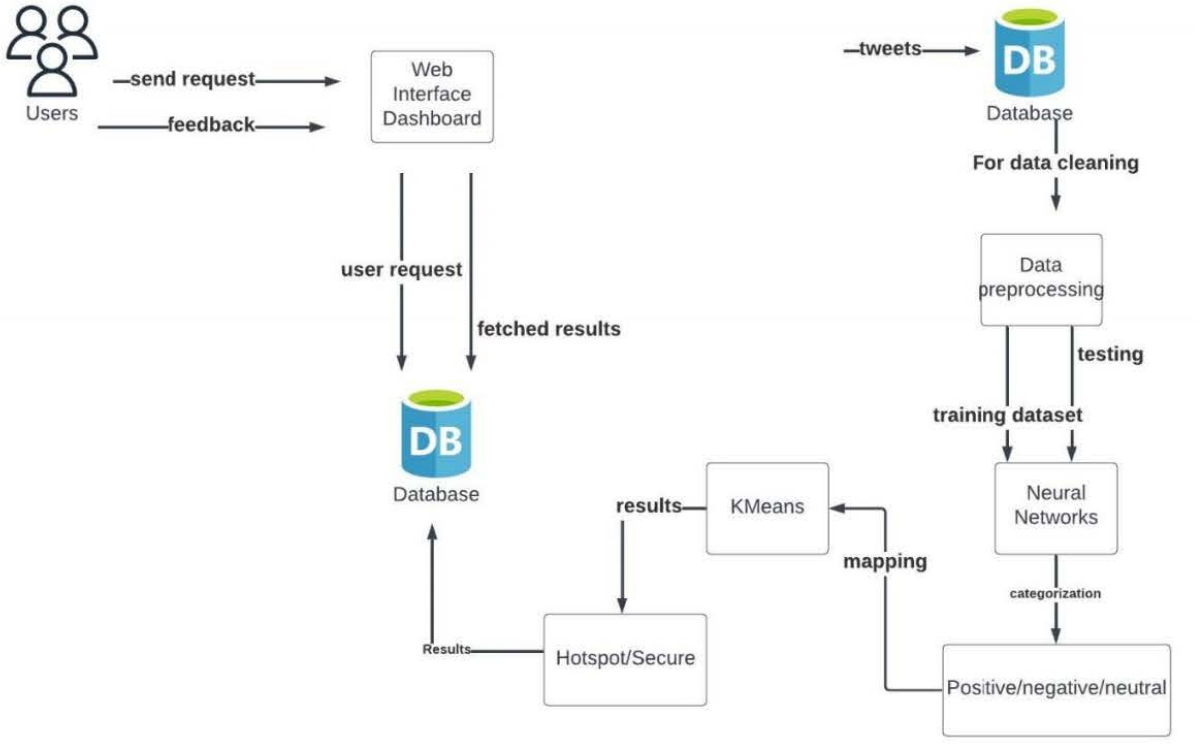


Figure 2.6:Conceptual Framework

Chapter 3: Research Methodology

3.1 Introduction

Research methodology specifically shows how a researcher will systematically design the study to ensure the reliability and validity of the results while addressing the research question and the objective (Jansen & Warren, 2020). Methodology works as a framework that is associated certain set of paradigmatic assumptions that might be adopted while conducting the research (Glatthorn & Joyner, 2005).

3.2 Research Design and Philosophy

The study embraced two research approaches, namely qualitative and quantitative methods, to identify the relationship between the features under observation and how they related to the probability of a crime occurring in Nairobi. We also assigned weights to the variables depending on their influence on the outcome in determining crime hotspots.

3.3 Population and Sampling

3.3.1 Population

The target population of this study was Nairobi residents. The study focused on the frequency of crime occurring in Nairobi. The source of the data set to be used was from open (public) data sets and X sentimental.

3.3.2 Sampling

Probability sampling was used in the study to reduce bias and guarantee that each population member had an equal chance of being picked, producing a sample that was representative of the entire community. The study also made use of stratified sampling, using 80% of the data obtained for model training, 10% for testing, and 10% for validation.

3.4 Data Collection Methods/ Analysis

3.4.1 Data Collection

The study utilized a secondary dataset by mining crime-related tweets from X which contained frequent crimes and areas where they occurred in relation to time. The dataset included Nairobi residents and crime encounters, which were important in creating, trial, and corroborating the model. The latest publicly available data that met the study requirements were used, given that crime data sets are available annually.

X data used for training and testing were mined from X. Corpora were built from crime - related tweets, and the Bag of Words technique was used for information retrieval to convert text into a format that the machine learning algorithm could process. The dataset was in comma - separated value (CSV) format, and stop words were removed to ensure uniformity and high-quality evaluation.

3.4.2 Data Analysis

This study implemented both qualitative and quantitative research designs to establish relationships and patterns of crime with respect to space and time. Data cleaning was performed by eliminating missing values, checking, and changing the data to a format that could be analyzed to provide meaningful insights. Maps were used to visualize the crime zone areas and non-crime zone areas.

3.5 Research Quality and Reliability

The X and police crime datasets were divided into two sections, high-risk and low-risk areas, using neural networks and the k-means clustering algorithm. The labelled dataset was then split into a training dataset for the model to learn the testing dataset and patterns in the data to validate and assess the model's accuracy in detecting and mapping crime hotspots. Frameworks like Keras, TensorFlow and Scikit Learn were used. Keras is a modular neural network library for Python that emphasizes ease of use and flexibility. It offers a high level of modularity, allowing users to build and customize neural networks with minimal coding required. A wide range of modules are also included in Keras, such as modules for activation functions, layers, preprocessing, goal functions,

and the selection of optimization methods (Jiang & Shen, 2019). Activation function modules and optimization method selection modules, among others, incorporate the majority of the most cutting-edge and effective activation functions and optimization techniques currently available (Jiang & Shen, 2019). These modules make it simple to create network models and improve important parameters of neural networks.

TensorFlow is a popular open-source library that is widely used for numerical computing and machine learning tasks (Yegulalp, 2018). It entails numerous machine learning and deep learning methods and models and provides a user-friendly front-end API for building applications in Pythonor JavaScript. TensorFlow uses high-performance C++ to execute these programs. The library is capable of training deep neural networks to handle duties like natural processing of language and image recognition and PDE-based simulations. One of the key features of TensorFlow is its ability to use the same models used for training to provide production prediction at scale (Yegulalp, 2018). Dataflow graphs, where each node denotes a mathematical process and each edge denotes a multidimensional data array or tensor, are used to build TensorFlow applications. Numerous platforms, including local computers, cloud clusters, iOS and Android smartphones, CPUs, and GPUs, can run TensorFlow. TensorFlow 2.0 includes several changes aimed at improving efficiency and user-friendliness, including a new API for distributed training and support for TensorFlow Lite to enable deployment of models on a wider range of systems (Yegulalp, 2018).

Python's Scikit-learn package is employed in machine learning. More precisely, it's a collection of straightforward and effective tools for data analysis and mining (Ruczyński, 2021). The frameworkis based on a number of well-known Python tools, including matplotlib, scikit-learn, and numpy. The BSD license that this library is offered under is a significant advantage. You are free to choose whether to upstream your improvements under this license without any limitations on commercia l use (Ruczyński, 2021). This solution's accessibility and simplicity are its key benefits. Scikit - Learn, on the other hand, is not the ideal option for deep learning. (Ruczyński, 2021)

3.5.1 Data Reliability.

In the study, the reliability of the model was assessed by its ability to produce consistent outcomes for the specified set of inputs. Performance indicators such as accuracy, precision, and recall ratios were used to evaluate the model. Accuracy was calculated as the proportion of correctly identified observations to all observations using the following formula.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Equation 3.1:Accuracy Ratio

Where FP (False Positives) denoted positively predicted data that was incorrectly predicted, FN (False Negatives) denoted negatively predicted data that was incorrectly predicted, TP (True Positives) denoted positively observed data that was correctly predicted, and TN (True Negatives) denoted negatively observed data that was correctly predicted.

Precision measured the ratio of correctly identified observations into the positive class to all correctly categorized observations into the positive class as determined by the model. This equation below was employed to convey it:

$$Precision = \frac{TP}{TP+FP}$$

Equation 3.2:Precision Ratio

Recall was defined as the proportion of actual positive class observations that were correctly classified by the model. It was expressed as follows:

$$Recall = \frac{TP}{TP+FN}$$

Equation 3.3:Recall Ratio

3.6 System Development Methodology

During the systems development phase, the rapid application development (RAD) technique was used. The RAD development paradigm prioritized rapid prototyping and prompt feedback instead of extensive development and testing cycles (Murad et al., 2018).

The RAD methodology enabled quick software updates and several revisions without having to start from scratch each time. This made it possible to ensure that the result was higher quality-focused and met the needs of the end customers. The overview of the RAD methodology is shown in Figure 3.1.

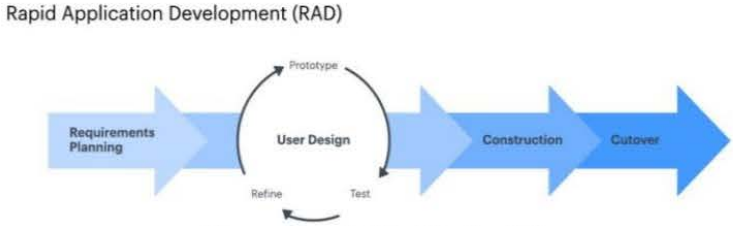


Figure 3.1:RAD Methodology

(Murad et al., 2018)

3.6.1 Requirements Planning

This stage involved the gathering of necessary data from existing literature and stakeholders of the crime hotspot mapping model. The scope and the goal of the system were explicitly defined.

3.6.2 User Design

During this stage, the proposed system's user interface will be designed. The system's potential users will design and test a number of prototypes to make sure there are no flaws.

3.6.3 Rapid Construction

The prototypes from the design phase were converted into functional models. This stage enabled the rapid development of the planned system. System testing was performed to ensure that everything was working as per the requirements. All of the above was achieved using Python programming language.

3.6.4 Cutover

The target users were introduced to the proposed system. The system underwent its last adjustments. To evaluate the performance of the suggested system, testing was carried out in the actual production environment. The model had a website for the users to interact with the model.

3.7 Utilization and Dissemination of Research Results.

The findings of this study should help citizens of Nairobi become aware of crime hotspot zones within Nairobi. The study also provides benefits to the police and security firms to know areas that needed more security resources based on the frequency and type of crimes that were common in a particular area.

3.8 Ethical Considerations/ Issues.

The researchers followed all the regulations, terms of use, privacy policies, and copyright laws that were applicable to the data used in the study. They obtained the secondary datasets from openly available sources and utilized both proprietary and open-source software. The proprietary software was obtained through freely available educational licensing. The study's research proposal was reviewed and approved by the Strathmore University Institutional Ethics Review Committee.

Chapter 4: System Analysis, Design and Architecture

4.1 Introduction

The application development involved an evaluation of multiple machine- learning techniques to determine the suitable framework. The analysis of the interaction between different components of the system was performed using use case and sequence diagrams. The design of the system specified both functional and non-functional requirements, outlining its overall structure and specific details.

4.2 System Analysis

System analysis involves determining the needs and specifications of a system, including its intended functionality. In the context of this study, the goal was to create an algorithm capable of identifying potential crime hotspots and designating specific areas as such. This analysis section provides details on both the functional and non-functional requirements that were taken into consideration during the development of the implemented solution.

4.2.1 Requirement gathering

To collect data for this study, the researcher obtained information from various sources to guide the investigation into the research questions. Relevant data on crime characteristics and frequency were retrieved from existing research literature conducted by researchers who have worked with similar data and implemented different tools. This information helped to define credible rules for the development of the algorithm.

The researcher also collected data from X and the Capital Hill Police Station to identify locations that have experienced crime in the past. This dataset was transformed into a geospatial dataset that could be loaded as a geodatabase in Geographic Information Systems (GIS).

Data analysis in this study involved cleaning, inspecting, and transforming the data into a geospatial format that could be analyzed to gain insights on the study objectives. Both qualitative and quantitative data analysis methods were utilized.

4.2.1.1 Functional Requirements

Functional requirements refer to the essential capabilities that a system must possess to perform its intended purpose. During the functional requirements gathering phase, the focus was on determining the specific operations and behaviors that the system should exhibit, including its response to user inputs and potential errors. The system was designed to cater to two main categories of users: regular users and administrators. The functional requirements identified for the system include:

i. The system must be launchable by the user to initiate its use.

ii. Users must be able to register an account on the system.

iii. Users must be able to log in to their accounts.

iv. Users should be able to check the safety status of a particular area by entering the relevant parameters into the system.

v. Users should be able to generate reports containing statistical data and visual elements like map images based on the analyzed information to determine if a crime event is likely to occur.

vi. Administrators should be able to manage datasets, including retrieving metadata from tweets, tweet content, and tweet coordinates, and retrain the model.

vii. The model must perform sentiment analysis on tweets and use coordinates to map the physical location of the crime, labelling the tweets as a secure area or a crime hotspot.

viii. Administrators should be able to manage system usage, scaling, and maintenance issues, including assisting users with system crashes and providing support as required.

ix. Users and administrators should be able to save their progress and exit the system after use.

4.2.1.2 Non-Functional Requirements

Non-functional requirements refer to the aspects of a system that relate to its design, quality, constraints, and external factors that may influence its performance. In this study, the following non-

functional requirements were identified:

Table 4.1:Non- Functional Requirements

Availability	The system should be accessible at all times for use by authorized users
Reliability	The system should consistently produce accurate and reliable results during data analysis and processing.
Accuracy	The system should have a minimal error rate as it will be providing sensitive output.
Usability	The system should be user-friendly, with a graphical user interface (GUI)that a layman user can easily navigate.
Maintainability	The system should be easy to maintain and fix in case of any errors or issues.
Scalability	The system should be able to accommodate a growing volume of crime reports and cover a wider geographical area

4.2.2 System Architecture

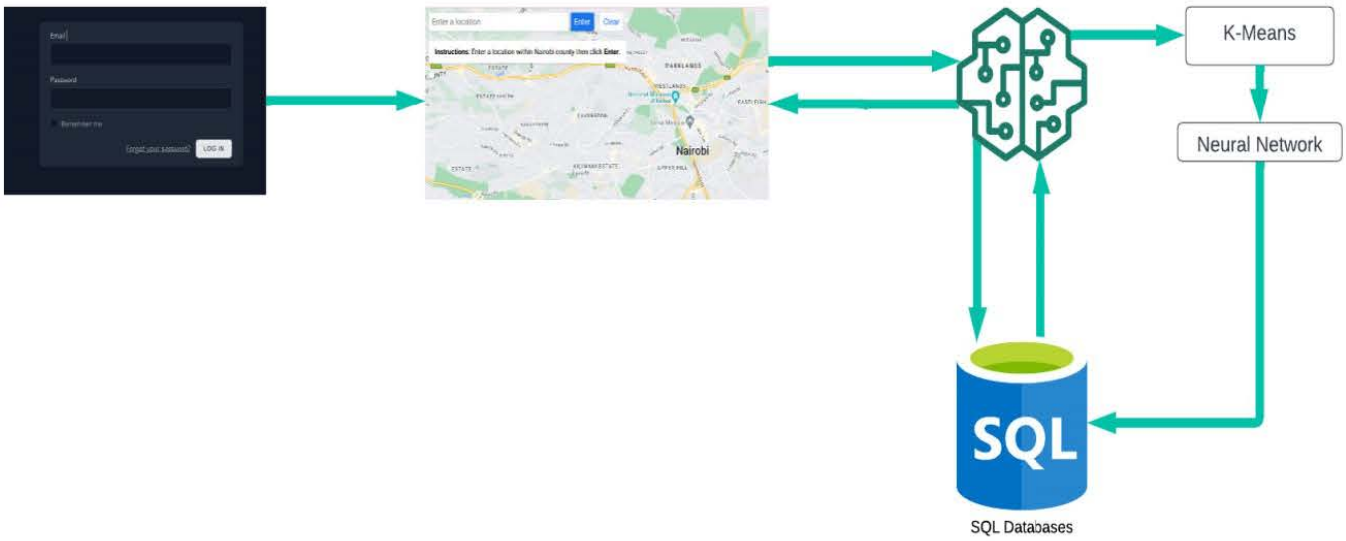


Figure 4.1:System Architecture

4.3 System Design

In the System Design phase, UML diagram models were created based on the analysis report to develop the system. The Structured System Design methodology was utilized in this study.

4.3.1 Use Case Diagram

The use case diagrams depict the interactions and relationships between various actors and the hot spot detection system. These diagrams illustrate scenarios in which external entities like users, organizations, or other systems use the system. Figure 4.2 illustrates such scenarios.

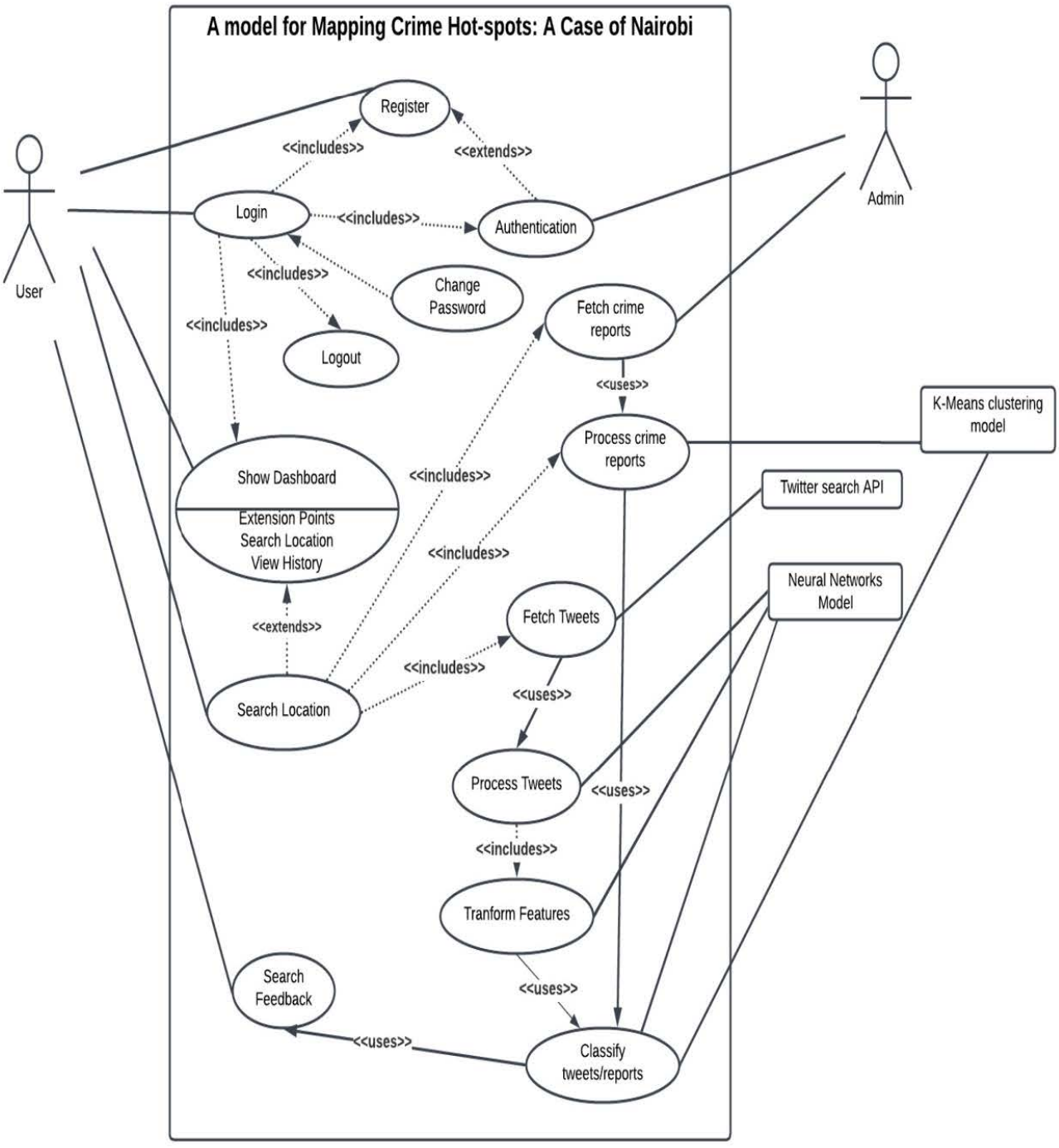


Figure 4.2:Use case diagram

4.3.1.1 Login Use Case

Table 4.2 below shows the login use case sequence

Table 4.2:Login Use Case

Use Case Name	Login, Register, Authentication, Change Password, Logout
Use case description	User logins if they have credentials, user creates a profile if they have none, user can logout after logging in
Actors	User
Pre-Condition	The user must have credentials to log in
Post-Condition	User can successfully login to the model
Main Success Scenario	
Flow of events	The user creates an account and logins to the model successfully. The user successfully changes the password. The user can view their dashboard after logging in

4.3.1.2 Search Location Use Case

Table 4.3 below shows the search location use case sequence

Table 4.3:Search Location Use Case

Use Case Name	Search Location, Search Feedback
Use case description	Registered users search for a location, and the model fetches data from tweets and crime reports uploaded by the admin and processes the data. The processed data is then used to classify an area as either hotspot or safe. The model then provides the user search feedback on whether the area searched is a crime hotspot or safe.
Actors	User, Admin, X search API, Neural Network, K-Means Clustering.
Pre-Condition	A user has to be registered. The area being searched has to be withinNairobi county.
Post-Condition	Tweets are successfully retrieved from X search API based on the specific location provided by the user. Tweets are classified validly as safe or crime hotspot areas.
Main Success Scenario	

Flow of events	The search is a location, and the searched location is mapped as either a crime hotspot area or a safe area.

4.3.2 Sequence Diagram

The sequence diagram of the model is shown in Figure 4.3 below

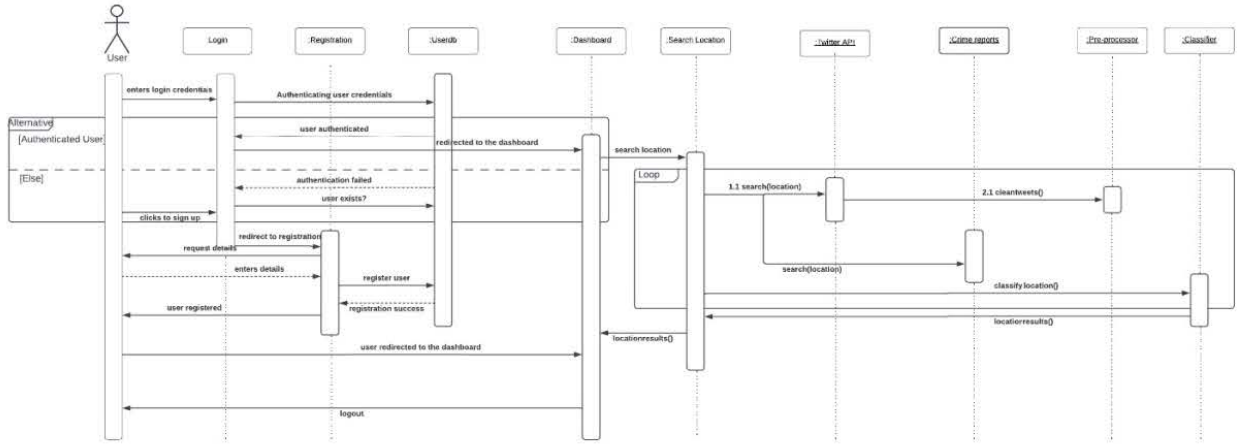


Figure 4.3:Sequence Diagram

4.3.3 Context Diagram

Figure 4.4 below shows the context diagram of the model.

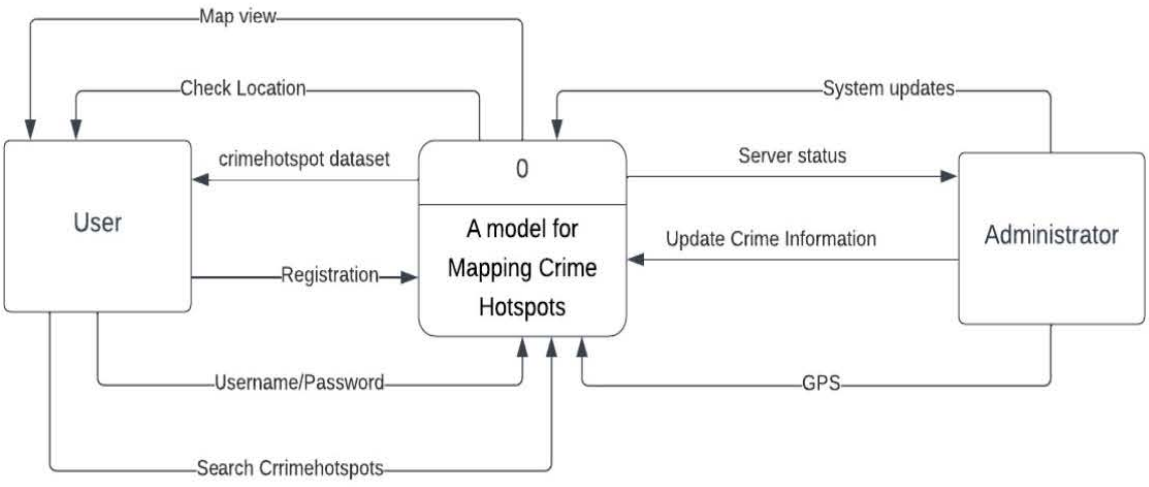


Figure 4.4:Context Diagram

4.3.4 Data Flow Diagram (Level One)

Figure 4.5 below shows the Level One: Data Flow Diagram

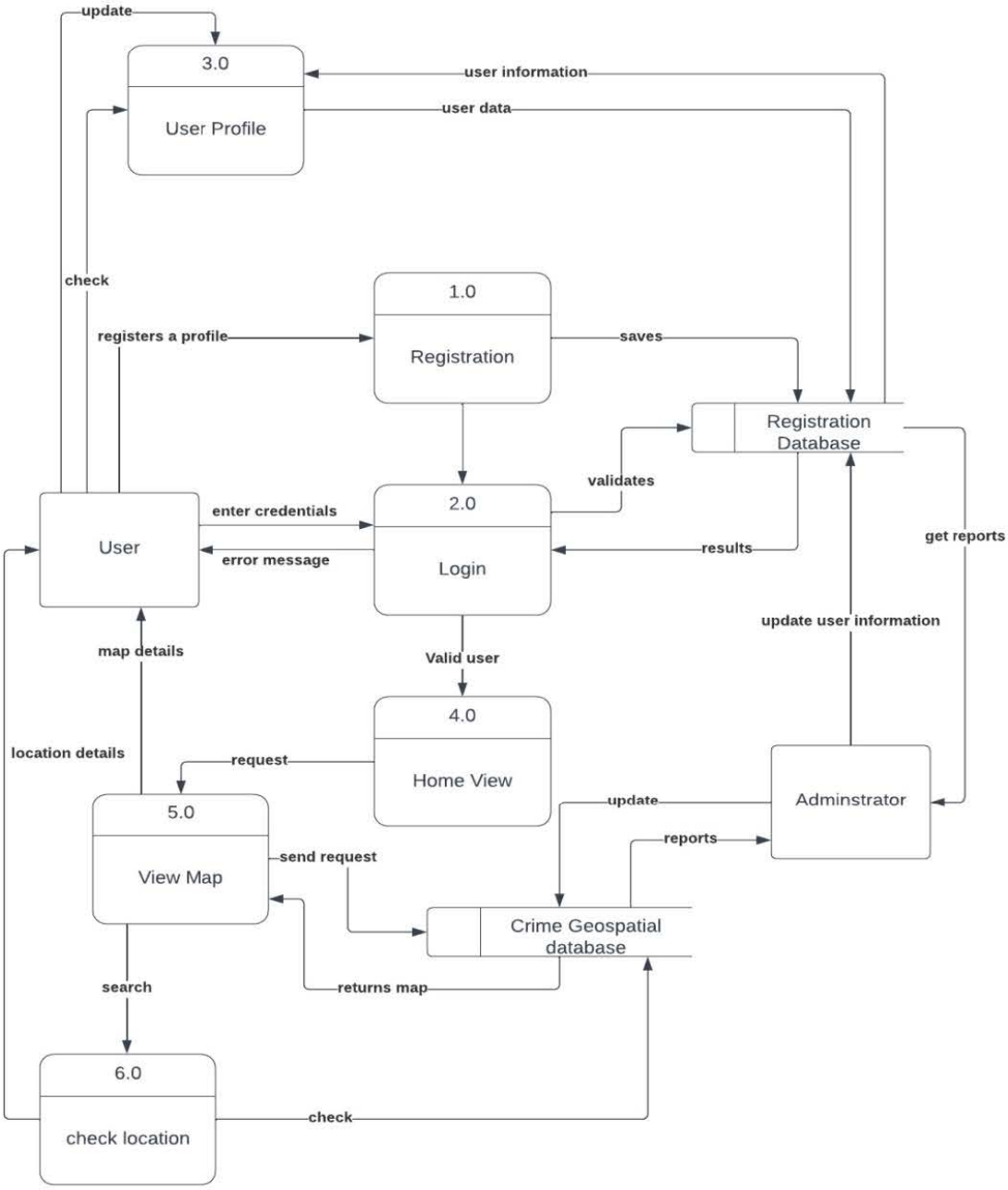


Figure 4.5:Level One Data Flow Diagram

4.3.5 Wireframes of the system

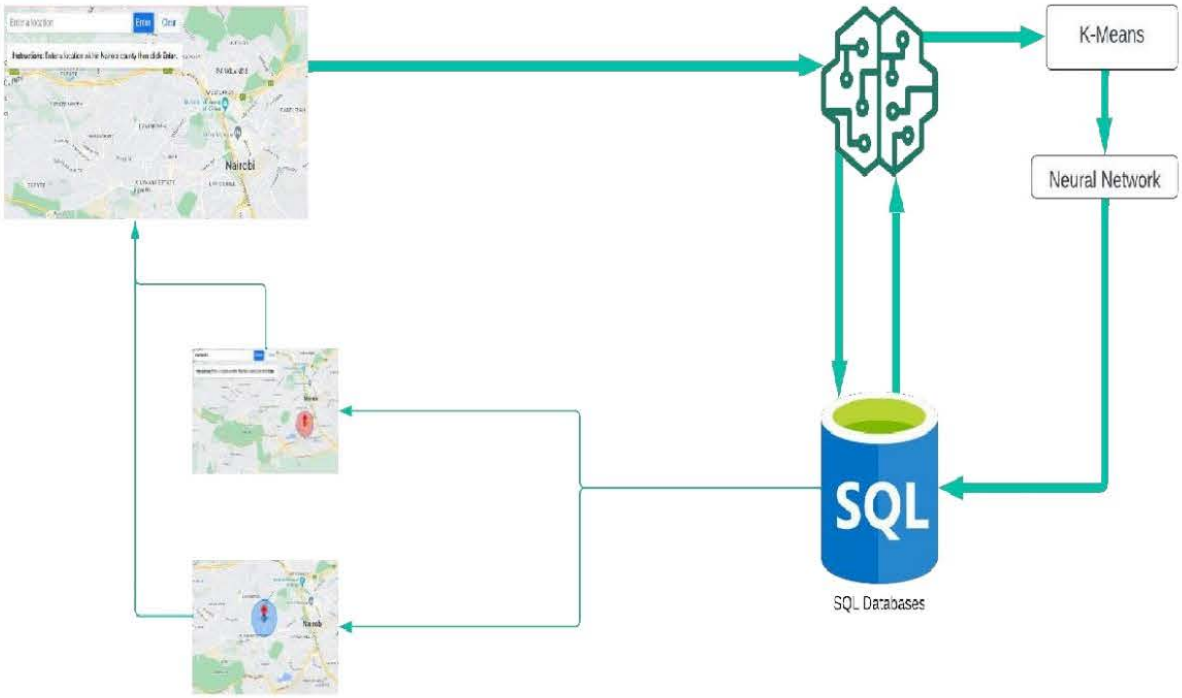


Figure 4.6:Wireframes of the system

Chapter 5: System Implementation and Testing

5.1 Introduction

The model is an amalgamate web solution with both frontend and backend modules. The model's front end provides a platform for the users to interact with the model. On the other hand, the backend is an internal linkage between the frontend requests raised by users and the data obtained from X and police reports to map the crime hotspot zones within Nairobi County.

5.2 Model Development

The model development process began with building the corpus for X API, designed to guide developers in analyzing ongoing conversations on X.

5.2.1 X Data Scrapping

The researcher used a sample of X users, such as @DCI_Kenya, @nrbcrimealert, @Kenyan_Report, @NPSC_KE, @PoliceKE, @NPSOfficial_K E, and @IPOA_KE, who had posted tweets about crime in Nairobi, to collect the tweets. The Python code snippet below was used to retrieve the tweets through the X API. The data was collected using the keywords #CrimeinNairobi, #NairobiCrime, and #CrimeNairobi.

```
import tweepy as tw
import pandas as pd

import tkinter as tk
from tkinter import messagebox
import tkinter.font as font


def scrapeData():
    consumer_key= '[redacted]'
    consumer_secret= '[redacted]'
    access_token= '[redacted]'
    access_token_secret= '[redacted]'

    path = direc.get()
    search_words = searchWord.get()
    no=n.get()

    auth = tw.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    api = tw.API(auth, wait_on_rate_limit=True)

    tweets = tw.Cursor(api.search_tweets,
                       q=search_words,
                       lang="en").items(no)
    users_locs = [[tweet.created_at, tweet.id, tweet.user.screen_name,tweet.text]
                  for tweet in tweets]
    twitterDf = pd.DataFrame(data=users_locs,
                             columns=['location', "id","user","Content"])
    path=path+"/TwitterCSV.csv"
    twitterDf.to_csv(path)
    messagebox.showinfo("Twitter Data Scraper","CSV file is saved successfully!")

Ln 10, Col 5  Spaces: 4  UTF-8  LF  Pyth
```

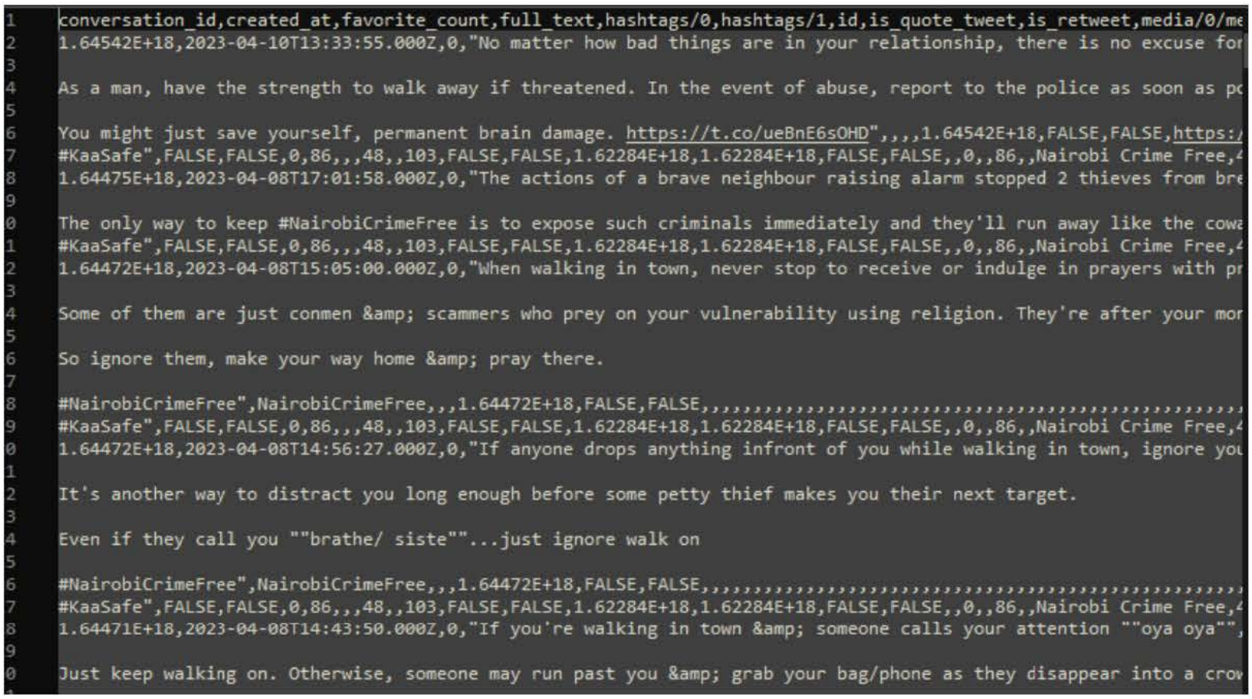
Figure 5.1:Code to scrap data

```
wn = tk.Tk()
wn.geometry("500x500")
wn.configure(bg='azure2')
wn.title("Twitter Data Scraper")
searchWord = tk.StringVar()
direc=tk.StringVar(wn)
n=tk.IntVar(wn)

headingFrame1 = tk.Frame(wn,bg="gray91",bd=5)
headingFrame1.place(relx=0.05,rely=0.1,relwidth=0.9,relheight=0.16)

headingLabel = tk.Label(headingFrame1, text=" Welcome to Twitter Data Srcaper", fg='grey19', font=
headingLabel.place(relx=0,rely=0, relwidth=1, relheight=1)


tk.Label(wn, text='Enter the word to be searched on twitter',bg='azure2', font=('Courier',10)).pla

tk.Entry(wn, textvariable=searchWord,  width=35,font=('calibre',10,'normal')).place(x=20,y=170)

tk.Label(wn, text='Please enter number of data values you require',bg='azure2', anchor="e").place(
tk.Entry(wn,textvariable=n, width=35, font=('calibre',10,'normal')).place(x=20,y=220)

#Getting the path of the folder
tk.Label(wn, text='Please enter the folder location where csv file is to be saved',bg='azure2', an
tk.Entry(wn,textvariable=direc, width=35, font=('calibre',10,'normal')).place(x=20,y=270)

ScrapeBtn = tk.Button(wn, text='Scrape', bg='honeydew2', fg='black', width=15,height=1,command=scr
ScrapeBtn['font'] = font.Font( size=12)
ScrapeBtn.place(x=15,y=350)
```

Figure 5.2:Code to scrap data

5.2.2 Preprocessing

The data collected from the X search API was busy and unstructured and could not be processed as it was by machine learning techniques. This made it mandatory to clean the data to a format that neural networks could process before proceeding to training the model. The corpus was built using the date and time of the tweets and retweets, and the geolocation of the tweet was also considered. Figure 5.3 shows a snip of the raw data fetched using X API.



```
conversation_id,created_at,favorite_count,full_text,hashtags/0,hashtags/1,id,is_quote_tweet,is_retweet,media/0/me
1.64542E+18,2023-04-10T13:33:55.000Z,0,"No matter how bad things are in your relationship, there is no excuse for

As a man, have the strength to walk away if threatened. In the event of abuse, report to the police as soon as po

You might just save yourself, permanent brain damage. https://t.co/ueBnE6sOHD",,,,1.64542E+18,FALSE,FALSE,https:/
#KaaSafe",FALSE,FALSE,0,86,,,48,,103,FALSE,FALSE,1.62284E+18,1.62284E+18,FALSE,FALSE,,0,,86,,Nairobi Crime Free,4
1.64475E+18,2023-04-08T17:01:58.000Z,0,"The actions of a brave neighbour raising alarm stopped 2 thieves from bre

The only way to keep #NairobiCrimeFree is to expose such criminals immediately and they'll run away like the cowa
#KaaSafe",FALSE,FALSE,0,86,,,48,,103,FALSE,FALSE,1.62284E+18,1.62284E+18,FALSE,FALSE,,0,,86,,Nairobi Crime Free,4
1.64472E+18,2023-04-08T15:05:00.000Z,0,"When walking in town, never stop to receive or indulge in prayers with pr

Some of them are just conmen &amp; scammers who prey on your vulnerability using religion. They're after your mon

So ignore them, make your way home &amp; pray there.

#NairobiCrimeFree",NairobiCrimeFree,,,1.64472E+18,FALSE,FALSE,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
#KaaSafe",FALSE,FALSE,0,86,,,48,,103,FALSE,FALSE,1.62284E+18,1.62284E+18,FALSE,FALSE,,0,,86,,Nairobi Crime Free,4
1.64472E+18,2023-04-08T14:56:27.000Z,0,"If anyone drops anything infront of you while walking in town, ignore you

It's another way to distract you long enough before some petty thief makes you their next target.

Even if they call you ""brathe/ siste""...just ignore walk on

#NairobiCrimeFree",NairobiCrimeFree,,,1.64472E+18,FALSE,FALSE,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
#KaaSafe",FALSE,FALSE,0,86,,,48,,103,FALSE,FALSE,1.62284E+18,1.62284E+18,FALSE,FALSE,,0,,86,,Nairobi Crime Free,4
1.64471E+18,2023-04-08T14:43:50.000Z,0,"If you're walking in town &amp; someone calls your attention ""oya oya"",

Just keep walking on. Otherwise, someone may run past you &amp; grab your bag/phone as they disappear into a crow
```


Figure 5.3:Raw scrapped data

The corpus had to be cleaned to remove duplicates, irrelevant data, and incorrect data and handle all the missing data from the corpus. The study did not need phrases like external URLs, hashtags, favorite count and retweets. The sentences had to be converted to lowercase with the .lower() method in Python. Removal of special characters and stop-words that have not much information about the text being used given the project focused on the English language. A code snippet used for text cleaning is shown in Figure 5.4 below.

```
# Text-preprocessing

# Missing Values
num_missing_desc = data.isnull().sum()[2]    # No. of values with msising descriptions
print('Number of missing values: ' + str(num_missing_desc))
data = data.dropna()

TAG_CLEANING_RE = "@\S+"
# Remove @tags
X['text'] = X['text'].map(lambda x: re.sub(TAG_CLEANING_RE, ' ', x))

# Smart lowercase
X['text'] = X['text'].map(lambda x: x.lower())

# Remove numbers
X['text'] = X['text'].map(lambda x: re.sub(r'\d+', ' ', x))

# Remove links
TEXT_CLEANING_RE = "https?:\S+|http?:\S|[^A-Za-z0-9]+"
X['text'] = X['text'].map(lambda x: re.sub(TEXT_CLEANING_RE, ' ', x))

# Remove Punctuation
X['text']  = X['text'].map(lambda x: x.translate(x.maketrans('', '', string.punctuation)))

# Remove white spaces
X['text'] = X['text'].map(lambda x: x.strip())

# Tokenize into words
X['text'] = X['text'].map(lambda x: word_tokenize(x))

# Remove non alphabetic tokens
X['text'] = X['text'].map(lambda x: [word for word in x if word.isalpha()])
```

Figure 5.4:Cleaning the data

The clean tweets were categorized into three groups. Negative tweets, Positive tweets and neutral tweets. The table below illustrates how the tweets were categorized

Table 5.1:Categorization of tweets

Tweet	Category
Safe zone	Positive
Crime hotspot	Negative
Neither crime hotspot nor safe zone	Neutral

The tweets have been classified as a 1 for crime hot spots, 0 for secure sites and -1 for neutral locations to carry out machine learning through neural networks. Figure 5.5 below illustrates the labelling

5.2.3 Identifying tweets

0	janet mbugua is one of the most easily recognisable of television personali
-1	death by dangerous driving fail to stop at rtc fail to report rtc disqualified dr
1	reuben ndolo ex-mp arrested for allegedly threatening to kill
-1	this young soul was killed by known people his phone and other valuables
0	policemen should be our friends not our enemies
-1	the situation seems under control but we ll wait the report of investigator

Figure 5.5:Sample categorization

Tweets associated with crime in Nairobi County were fetched, and the huge dataset was collected; not all tweets could be classified as positive or negative, as some of them were unrelated to crime despite being posted on crime reporting X handles. VADAR, a sentiment analysis tool, was used because it does not require prior training or punctuation elimination since it understands even slang (Elbagir & Yang, 2019). Figure 5.6 shows how VADAR was used for sentiment analysis of the tweets.

```
def sentiment_scores(sentence):

    # Create a SentimentIntensityAnalyzer object.
    sid_obj = SentimentIntensityAnalyzer()

    # polarity_scores method of SentimentIntensityAnalyzer
    # object gives a sentiment dictionary.
    # which contains pos, neg, neu, and compound scores.

    for i in range(len(sentence)):

        sentiment_dict = sid_obj.polarity_scores(sentence[i])

    print("Sentence Overall Rated As", end = " ")

    # decide sentiment as positive, negative and neutral
    if sentiment_dict['compound'] >= 0.05 :
        print("Positive")

    elif sentiment_dict['compound'] <= - 0.05 :
        print("Negative")

    else :
        print("Neutral")

# Driver code
    if __name__ == "__main__" :

        sentiment_scores(sentence[i])
```

Figure 5.6:Classifying tweets

5.2.4 Location Identification

Identifying the location was achieved by dividing the sum of all positive tweets by the sum of all negative tweets of a particular place. Based on the score, the place is categorized as safe or crime

$$Final\ Score = \frac{positive\ tweets}{negative\ tweets}$$

Equation 5.1:Final score

5.2.5 Neural Network training

The neural network is a popular machine learning model due to its powerful computational capacity and ability to handle large datasets (Mahmood et al., 2020). It comprises three layers: the input layer, the hidden layer, and the output layer, and the parameters are called weights. To learn, the neural network computes the difference between the predicted output and the actual value during training, propagates this difference back through the network, and adjusts the connection weights to improve the alignment between the predicted and actual output values (Mahmood et al., 2020).

The dataset acquired from X had to go through preprocessing for clean-up and then was analyzed to determine positive tweets from negative tweets to be used to train the model. Python programming language was used based on its wide ability to accommodate several library importations. Pandas is a software library used for data analysis and machine learning. It is designed to work with structured data and is built on top of another library called Numpy. Numpy provides support for handling multi-dimensional arrays and matrices in Python. Pandas reduce performing repetitive tasks that would be time-consuming, such as data cleaning, filling, normalization, merging, joining, statistical analysis and visualization.

Another Python package is the Scikit-learn which is used to implement machine learning. Scikit - learn provides functionality for dimensionality reduction, feature selection, feature extraction, inbuilt datasets and ensemble techniques. The Scikit-learn library, particularly the CountVectorizer module, was used in this study to convert the training data into a matrix that includes token counts. Scikit-learn was also used to handle vectorization, transformation, and classifier specification in a single pipeline.

```
from sklearn.cluster import KMeans
from sklearn.neural_network import MLPClassifier
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from keras.models import Sequential
from keras.layers import Dense

# Load and split the dataset
X = df_balance['Tweet_lemmatized']
y = df_balance['sentiments_array']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

X_train_strings = [' '.join(lemmas) for lemmas in X_train]
X_test_strings = [' '.join(lemmas) for lemmas in X_test]

# Convert text data to numerical data
vectorizer = CountVectorizer()
X_train_vec = vectorizer.fit_transform(X_train_strings)
X_test_vec = vectorizer.transform(X_test_strings)
```

Figure 5.7:Vectorizing the tweets

```
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.feature_extraction.text import CountVectorizer


# Split the corpus and labels into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(df_balance.Tweet_lemmatized, df_balance.sentiments_array, test_size=0.2, random_state=42)
X_train.shape , X_test.shape, y_train.shape, y_test.shape


# Join the lemmas in each list into a string
X_train_strings = [' '.join(lemmas) for lemmas in X_train]

# Vectorize the training data
vectorizer = CountVectorizer()
X_train_vec = vectorizer.fit_transform(X_train_strings)
# Vectorize the training data
# vectorizer = CountVectorizer()
# X_train_vec = vectorizer.fit_transform(X_train)

# Train a neural network classifier
clf = MLPClassifier(hidden_layer_sizes=(100,), max_iter=1000, random_state=42)
clf.fit(X_train_vec, y_train)

X_test_strings = [' '.join(lemmas) for lemmas in X_test]

# Vectorize the testing data
X_test_vec = vectorizer.transform(X_test_strings)

# Test the classifier
accuracy = clf.score(X_test_vec, y_test)
print(f"Accuracy: {accuracy:.4f}")
```

Figure 5.8:NN training

5.2.6 Testing the model

The model was trained using 80% of the labelled tweets dataset, which was fed into the model to help it learn and make accurate predictions. To evaluate the accuracy of the model, a confusion matrix was utilized. The results of the confusion matrix are presented in Table 5.2.

Table 5.2:Results of training

Predicted Values		Actual Values	
		Positive	Negative
	Positive	43	5
	Negative	15	49

5.2.6.1 Precision

This refers to the proportion of true positives to the total number of true and false positives and negatives in a classification task, known as the precision score.

$$Precision = \frac{\sum true\ positives}{\sum(true\ positives + true\ negatives)}$$

Equation 5.2:Precision Formula

The precision in this instance will be:

$$Precision = \frac{49}{49+5} = 0.9074$$

Equation 5.3:Calculated precision

Python has a function called precision_score() that is used to calculate the precision

5.2.6.2 Recall

To calculate this metric, the number of true positive instances is summed up and then divided by the sum of true positive and false negative instances.

$$Recall = \frac{\sum true\ positives}{\sum (true\ positives + false\ negatives)}$$

Equation 5.4:Recall formula

$$Recall = \frac{49}{49+15} = 0.765$$

Equation 5.5:Calculated recall

Recall can be calculated in Python using the recall_score() sci-kit-learn function.

5.2.6.3 F-Measure

The F-measure is a metric that considers both precision and recall to provide a comprehensive evaluation of a model's performance. It can be calculated using equations 5.6 and 5.7, as shown below.

$$F\ Measure = \frac{(2 * presision * recall)}{(precision + recall)}$$

Equation 5.6:F-Measure Formula

$$F\ Measure = \frac{(2 * 0.9074 * 0.765)}{(0.9074 + 0.765)} = 0.8305$$

Equation 5.7:Calculated F-measure

To calculate the F-measure score using Python, the Scikit-learn library's f1_score() function was utilized.

5.2.6.4 ROC Curve for the model

Figure 5.9 shows how well the classifier is performing by displaying a Receiver Operating Characteristic (ROC) curve for both the neural networks and k-means clustering.

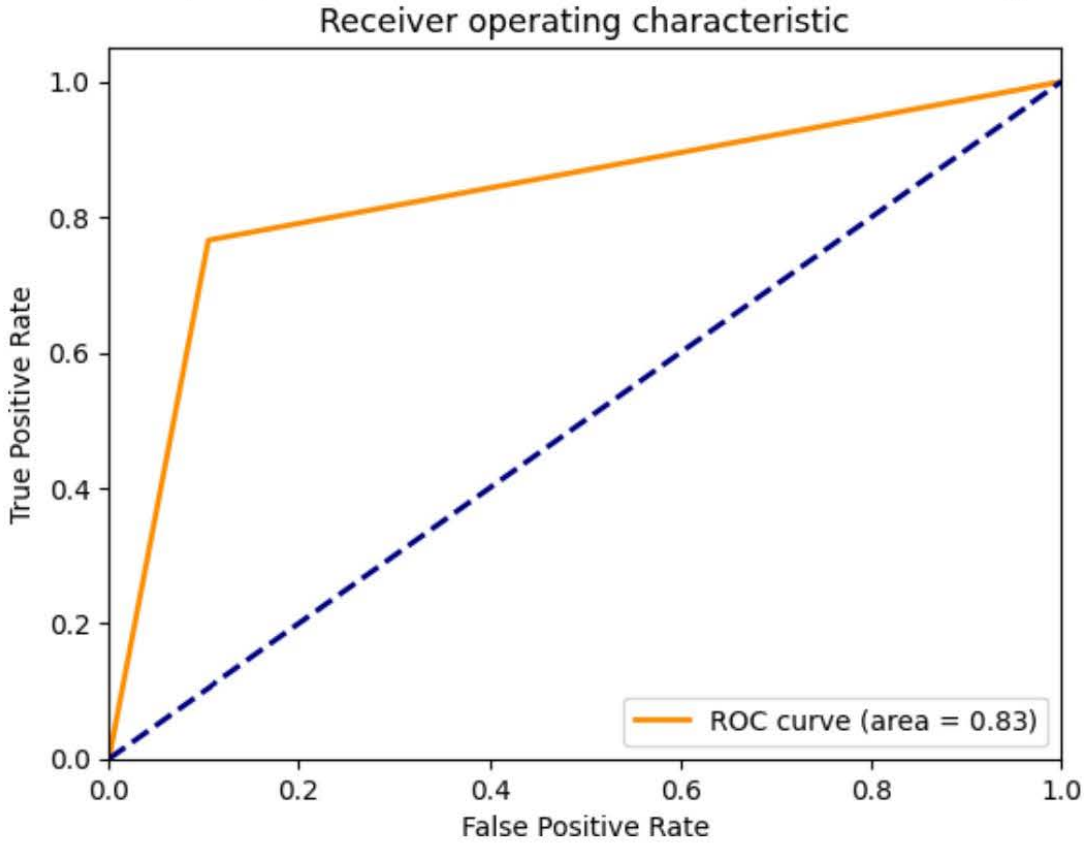


Figure 5.9:ROC for K-means and NN

5.2.7 Using the model to detect crime hotspots

The crime hotspot detection system requires users to choose a location to detect a crime hotspot. Then, the system used the X Search API to extract tweets from the selected location, and these tweets were passed through the built model for hotspot detection.

Figure 5.10 displays the interface design of the crime hotspot detection system. It included a home screen that prompted the user to choose a location for the model to analyze.

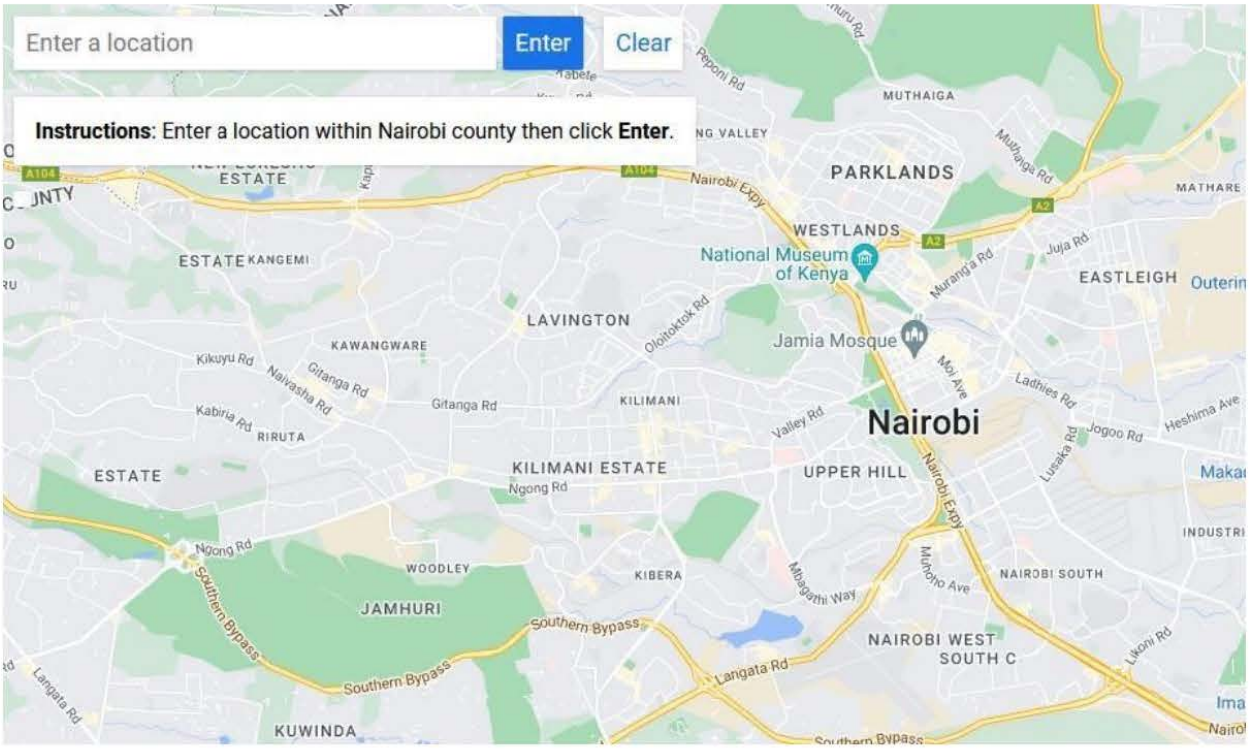


Figure 5.10:Search location page

First, the model will receive a specific location as input and use it to generate a search phrase for the X search API. The API will retrieve tweets related to the given location, and then the tweets will be preprocessed. The preprocessing will involve organizing the tweets into a table structure that includes information such as username, date and time, retweet count, text, and geolocation. After that, the tweets will be cleaned up to remove unwanted phrases like hashtags, URLs, and retweets. Furthermore, to preprocess the data, all tweets will be changed to lowercase, and any punctuation marks will be eliminated. Then, the pre-trained model will categorize the tweets as having negative, positive, or neutral sentiment. Based on the quantity of positive or negative tweets, the location will be designated as either a safe area or a crime hotspot.

Chapter 6: Discussions

The study aimed to create a crime hotspot mapping model for Nairobi County using K-Means and Neural Networks with unigram, bigram, and trigram. The model's performance was evaluated using a subset of the dataset. Several experiments used different classifiers to validate the method for detecting crime hotspots in Nairobi via X. The findings revealed that the model's accuracy was 88.39% using trigrams.

6.1 Other classifiers experiments

These tests aimed to evaluate and compare the performance of the K-Means and Neural Networks model against other widely-used machine learning algorithms, such as Support Vector Machine, Naive Bayes, Random Forest, Decision Tree, Naive Bayesian, and K-Nearest Neighbor.

6.1.1 Support Vector Machine Classifier (SVM)

The dataset was tested using SVM, and table 6.1 below shows the accuracy level achieved using unigrams, bigrams and trigrams.

Table 6.1:SVM performance

No of grams	Accuracy	Precision	Recall	F1-Score
Unigram	0.7946	0.8868	0.7344	0.8034
Bigram	0.8571	0.9000	0.8438	0.8710
Trigram	0.8839	0.8923	0.9062	0.8992
average	0.8452	0.8930	0.8281	0.8579

The performance of the Support Vector Machine (SVM) classifier is demonstrated in Figure 6.1, which shows a ROC curve.

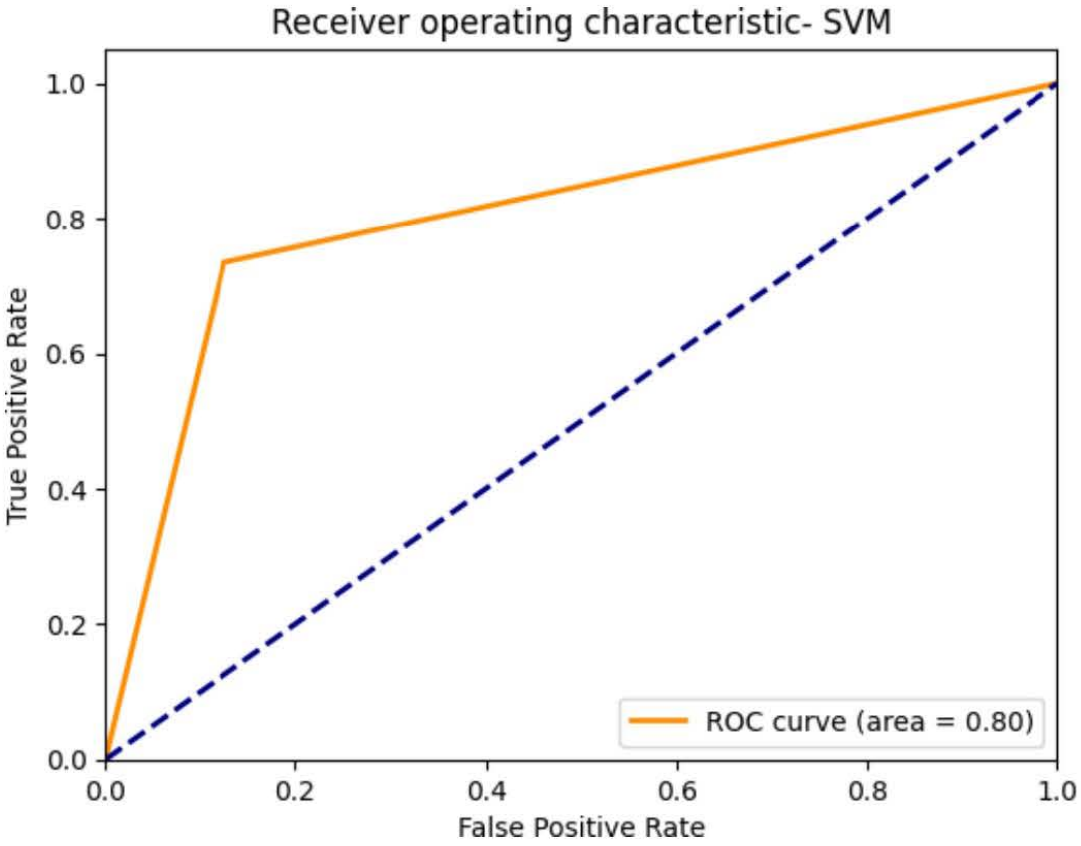


Figure 6.1:ROC for SVM

6.1.2 Random Forest

Table 6.2 shows the performance of the random forest classifier with the same dataset.

Table 6.2:Random Forest performance

No of grams	Accuracy	Precision	Recall	F1-Score
Unigram	0.8750	0.9032	0.8750	0.8889
Bigram	0.8929	0.8824	0.9375	0.9091
Trigram	0.9018	0.8955	0.9375	0.8378
average	0.8899	0.7746	0.9584	0.9160

Figure 6.2 below displays the performance of the Random Forest Classifier through a Receiver Operating Characteristic (ROC) curve.

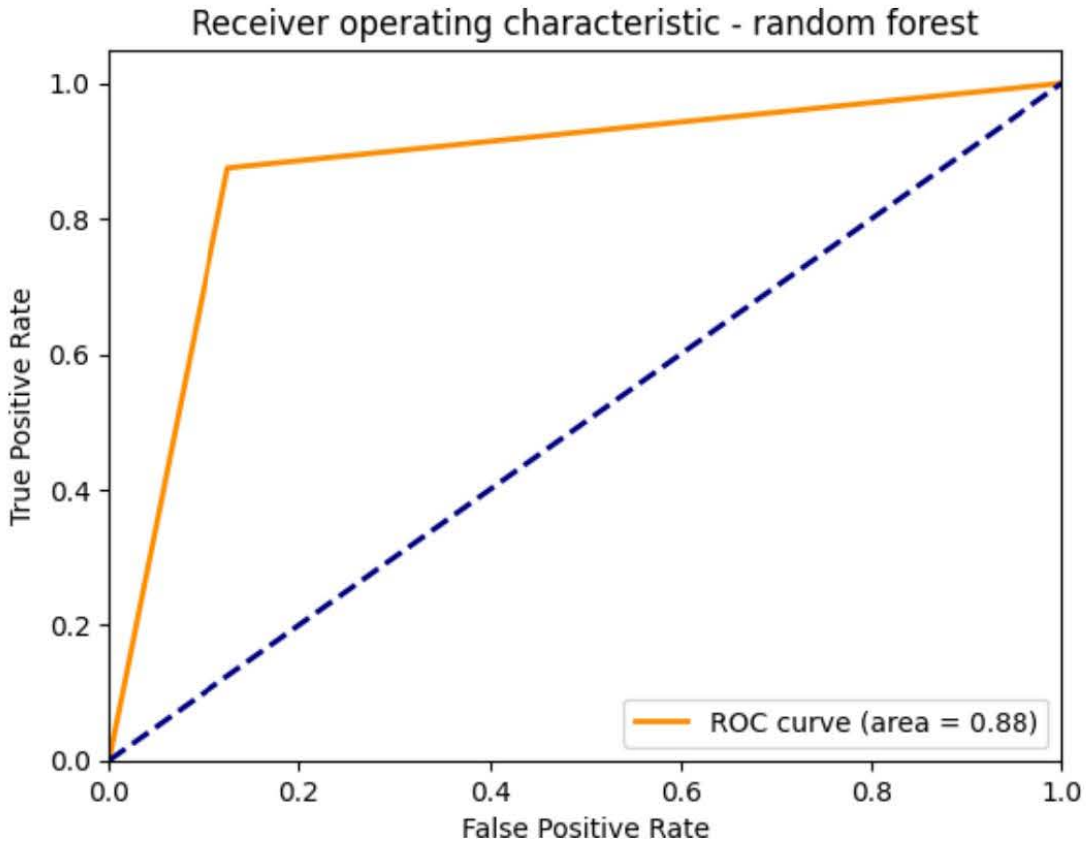


Figure 6.2:ROC for Random Forest

6.1.3 Naive Bayesian

Table 6.3 below shows the performance of the naïve Bayesian classifier on the same dataset.

Table 6.3:Naïve Bayesian performance

No of grams	Accuracy	Precision	Recall	F1-Score
Unigram	0.8125	0.8525	0.8125	0.8320
Bigram	0.8036	0.8750	0.7656	0.8167

Trigram	0.8036	0.8750	0.7656	0.8167
Average	0.8066	0.8675	0.7812	0.8218

A ROC curve was plotted to depict the performance of the Naïve Bayesian Classifier, and the resulting chart is presented in Figure 6.3 below. The curve shows how well the Naïve Bayesian classifier performed.

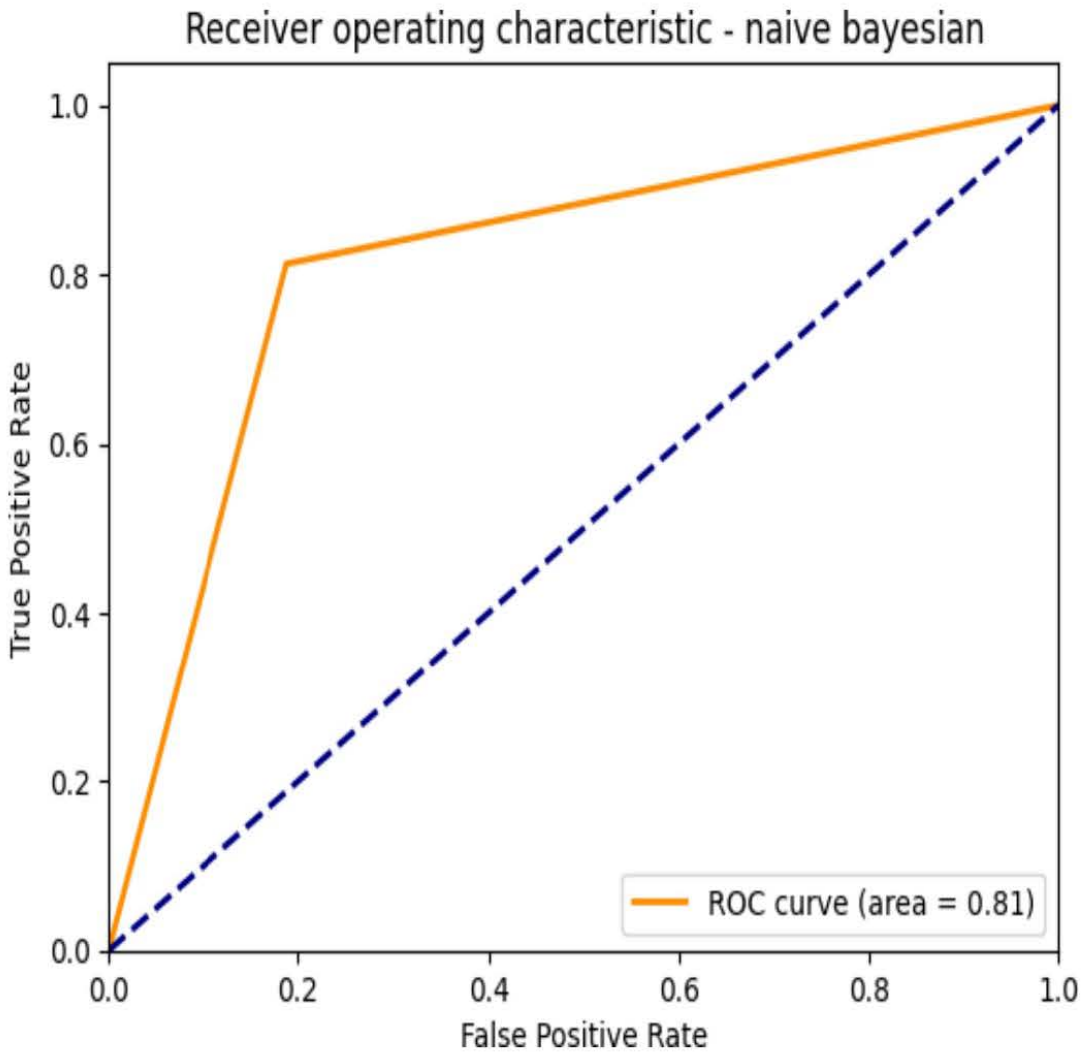


Figure 6.3:ROC for Naïve Bayesian

6.1.4 Decision Tree

Table 6.4 below shows the performance of the decision tree algorithm classifier on the same dataset.

Table 6.4 :Decision Tree performance

No of grams	Accuracy	Precision	Recall	F1-Score
Unigram	0.7143	0.8636	0.5938	0.7037
Bigram	0.6964	0.8409	0.5781	0.6852
Trigram	0.7232	0.8511	0.6250	0.8378
average	0.7113	0.7746	0.9584	0.7207

A graphical representation of the performance of the Decision Tree Classifier was generated using the Receiver Operating Characteristic (ROC) curve. The resulting figure is displayed below as Figure 6.4.

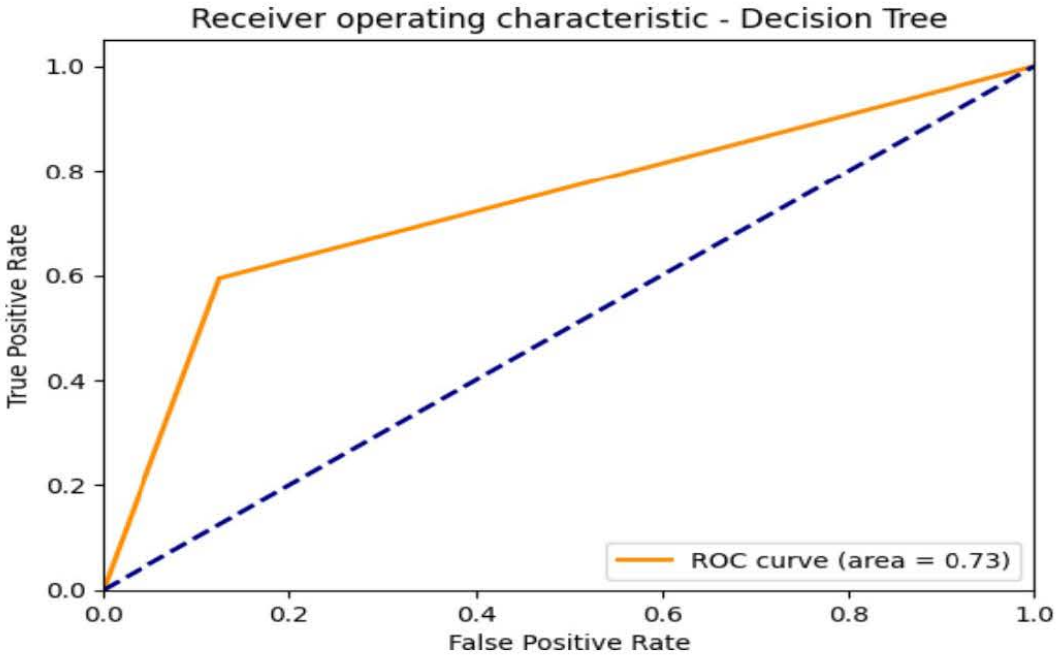


Figure 6.4:ROC for Decision Tree

6.1.5 K- Nearest Neighbor

The dataset used was tested using K-Nearest neighbor when k was 2, and the performance was as shown in Table 6.5 below

Table 6.5:KNN performance

No of grams	Accuracy	Precision	Recall	F1-Score
Unigram	0.8393	0.8108	0.9375	0.8696
Bigram	0.8214	0.7750	0.9688	0.8611
Trigram	0.7857	0.7381	0.9688	0.8378
average	0.8155	0.7746	0.9584	0.8562

A graph depicting the performance of the K-Nearest Neighbor (KNN) classifier was created using a Receiver Operating Characteristic (ROC) curve. Figure 6.5, displayed below, illustrates the performance of the KNN classifier.

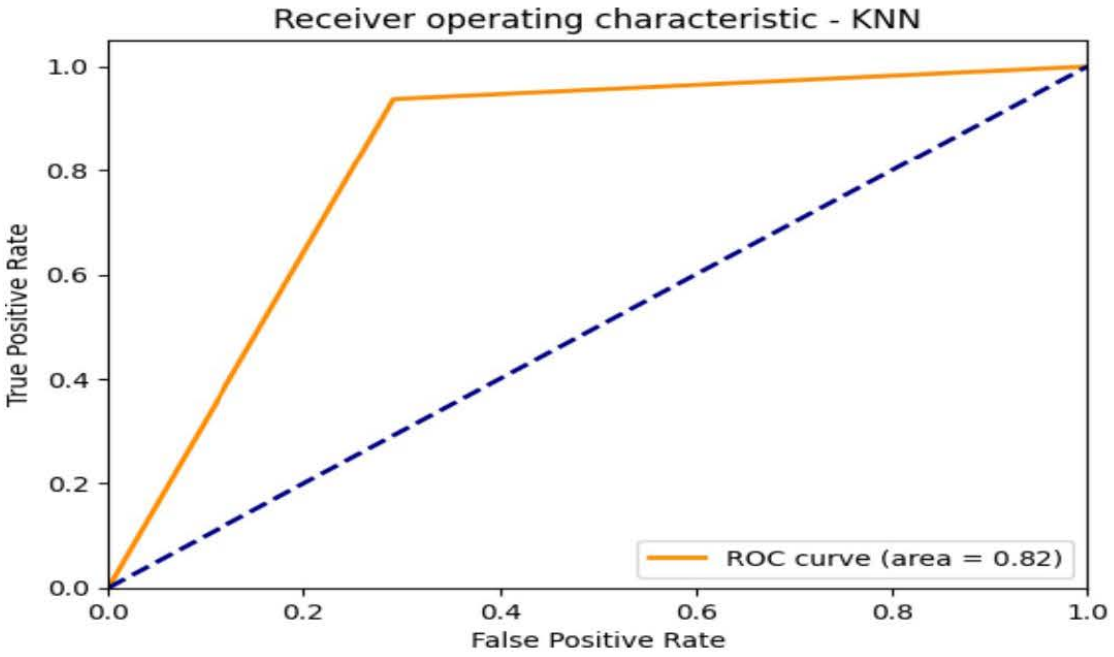


Figure 6.5:ROC for KNN

6.1.6 Comparison results of the other classifiers

According to Table 6.6, the random forest classifier scored the highest in accuracy, 0.8899. The SVM model's Precision and Recall scores, however, were higher than those of the Random Forest model, coming in at 0.8930 and 0.8281, respectively. The Naive Bayesian model outperformed the SVM model in terms of precision while having a lower accuracy score of 0.8066. With the lowest accuracy score of 0.7113 and the lowest F1-Score of 0.7207 among all models, the Decision Tree model performed poorly. Both the Random Forest and SVM models fared well overall, with the Random Forest model scoring higher in accuracy and the SVM model scoring higher in precision and recall.

Table 6.6:Performance comparison of other classifiers

	Accuracy	Precision	Recall	F1-Score
SVM	0.8452	0.8930	0.8281	0.8579
Random Forest	0.8899	0.7746	0.9584	0.9160
Naïve Bayesian	0.8066	0.8675	0.7812	0.8218
Decision Tree	0.7113	0.7746	0.9584	0.7207
KNN	0.8155	0.7746	0.9584	0.8562
NN	0.8214	0.9074	0.7656	0.8305

6.2 Discussions

According to the developed model, there was an increase in complexity from unigram to trigram, and the performance also improved. Both bigram and trigram outperformed the unigram in all metrics. Among them, the trigram had the highest accuracy of 0.88839. The obtained results are shown in Table 6.7.

Table 6.7:Performance of K-Means and NN

No of grams	Accuracy	Precision	Recall	F1-Score
Unigram	0.8214	0.9074	0.7656	0.8305
Bigram	0.8661	0.9153	0.8438	0.8780
Trigram	0.8839	0.9180	0.8750	0.8960

Figure 6.6 below is a graphical illustration of the performance of K-Means and NN

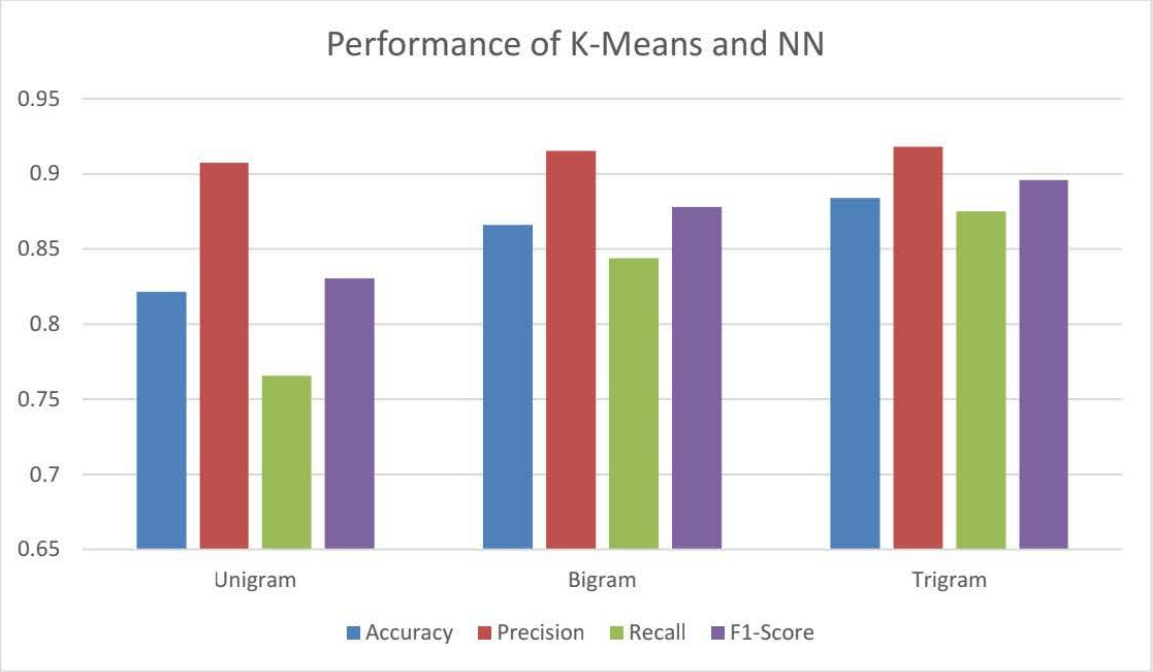


Figure 6.6:Performance of NN

The experiments' findings suggested that the optimal approach to constructing a model for identifying crime hotspots on X was through a combination of Neural Network and K-means machine learning algorithms, utilizing trigram features weighted.

Chapter 7: Conclusion and recommendations

7.1 Conclusion

The crime mapping model successfully illustrated how social media data and machine learning algorithms may be used to locate and map crime hotspots. Highly accurate crime activity forecasts were made using X data and cutting-edge machine learning algorithms, showing that social media sites like X can be useful data sources for law enforcement agencies and crime prevention organizations. Other businesses that rely on real-time data analysis, such traffic and environmental monitoring, could use the tools and techniques used in this study more widely.

The study concentrated on gathering tweets regarding crime in Nairobi using the X API and preprocessing the tweets to categorize them as good, negative, or neutral. The researchers ran a number of experiments to assess the efficacy of their strategy, contrasting it with other well-known machine learning classifiers as SVM, KNN, Naive Bayes, Decision Tree, and Random Forest.

The Random Forest model achieved the highest accuracy score of 0.8899, followed by the KMEANS with the NN model with an accuracy score of 0.8839. The SVM and KNN models also achieved reasonable accuracy scores of 0.8452 and 0.8155, respectively, while the Decision Tree model had the lowest accuracy score of 0.7113. The KMEANS with NN model performed best in precision with a score of 0.9180, followed by SVM with a score of 0.8930, while the Random Forest model had a lower precision score of 0.7746. For recall, the Random Forest and KMEANS with NN models achieved the highest scores of 0.9584, followed by SVM and Naive Bayesian models with scores of 0.8281 and 0.7812, respectively. Finally, the F1-Score was highest for the Random Forest model with a score of 0.9160, followed by the KMEANS with the NN model witha score of 0.8960. The results suggest that the Random Forest and KMEANS with NN models are the most effective for this problem. They achieve high accuracy and F1-Scores while having reasonable precision and recall metrics.

7.2 Recommendations

Based on the conclusions of the crime mapping project, it was recommended that law enforcement agencies and crime prevention organizations consider using machine learning algorithms and social

media data to identify and map crime hotspots. The findings suggested that X was a valuable real-time crime analysis and prevention data source.

The project's methodology and techniques could also be adapted and refined for broader applications in other industries that rely on real-time data analysis. For instance, the models and algorithms could be used for traffic and environmental monitoring.

Regarding the machine learning classifiers used in the project, the Random Forest and KMEANS with NN models demonstrated the most effective performance for this particular problem, achieving high accuracy, precision, recall, and F1-Scores. Therefore, it was recommended that these models be considered for future crime mapping and prevention initiatives. The researcher acknowledged that superior outcomes could have been achieved if a larger dataset had been utilized.

7.3 Further work

Several areas for further research and development were identified based on the results and limitations of our crime mapping project.

Firstly, expanding the dataset to include more diverse data sources could improve the accuracy and generalizability of the models. Additional sources, such as CCTV footage or police reports, could provide a more comprehensive understanding of crime patterns and aid in identifying hotspots.

Secondly, incorporating temporal data could enhance the models by allowing for more accurate predictions and monitoring of crime activity as crime patterns and trends change over time. Thirdly, developing a user-friendly interface for law enforcement agencies and crime prevention organizations could improve the practical application of our project by integrating the models and algorithms into existing software platforms. Finally, expanding the project to other cities or regions could provide further insight into the effectiveness and generalizability of the models. Adapting the models to new locations and evaluating their performance could help to refine and improve them.

Overall, there is significant potential for further research and development in crime mapping using

machine learning algorithms and social media data. By addressing limitations and exploring new avenues, we can continue to improve the accuracy and practical application of these methods for the benefit of society.

References

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics, 9(8), 1295. https://doi.org/10.3390/electronics9081295

Alam, M. J. (2016). Home Automation and Security System. *Daffodil International University*.

Alduraywish, M. A. (2021). Juvenile Delinquency and Differential Association Theory. *Advances in Applied Sociology, 11*(8), 341–349. doi:10.4236/aasoci.2021.118031

Aruma, D. E., & Hanachor, D. M. (2017, December). ABRAHAM MASLOW'S HIERARCHY OF NEEDS AND ASSESSMENT OF NEEDS IN COMMUNITY DEVELOPMENT. *International Journal of Development and Economic Sustainability, 5*(7), 15 - 27.

Baraka, G. E., & Murimi, S. K. (2021). Why geographic information systems in spatiotemporal crime analysis? Attitude of Kenyan police officers. *Police Practice and Research, 22*(5), 1453-1468. doi:10.1080/15614263.2019.1689131

Baraka, G., & Murimi, S. (2019). Stuck in the past with push-pins on paper maps: Challenges of transition from manual to computerized crime mapping and analysis in Kenya. *International Journal of Police Science & Management, 21*, 36-47. doi:10.1177/1461355719832620

Belhadi, A., Djenour, Y., Nørvåg, K., Ramampiaro, H., Masseglia, F., & Lin, J. C.-W. (2020). Space–time series clustering: Algorithms, taxonomy, and case study on urban smart cities. *Engineering Applications of Artificial Intelligence, 95*, 103857. doi:https://doi.org/10.1016/j.engappai.2020.103857

Braga, A. A., Turchan, B., Papachristos, A. V., & Hureau, D. M. (2019). Hot spots policing of small geographic areas effects on crime. *Campbell Systematic Reviews* , 1.

Campedelli, G. M., Aziani, A., & Favarin, S. (2021). Exploring the Immediate Effects of COVID-19 Containment Policies on Crime: an Empirical Analysis of the Short-Term Aftermath in Los Angeles. *American Journal of Criminal Justice, 46*(5), 704-727. doi:10.1007/s12103-020-09578-6

CAP.63.(2014).Retrieved from Kenyalaw.org: http://www.kenyalaw.org/lex/actview.xql?actid=CAP.%2063

Catlett, C., Cesario, E., Talia, D., & Vinci, A. (2018). A Data-Driven Approach for Spatio-Temporal Crime Predictions in Smart Cities. *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, 17-24. doi:10.1109/smartcomp.2018.00069

Chainey, S. (2014). Crime Mapping. In G. Bruinsma, & D. Weisburd, *Encyclopedia of Criminology and Criminal Justice* (pp. 699--709). New York, NY: Springer New York. doi:10.1007/978-1-4614-5690-2_317

Crime Statistics . (2018). Retrieved from Nationalpolice.go.ke: https://www.nationalpolice.go.ke/crime-statistics.html

Dağlar, M., & Argun, U. (2016). Crime Mapping and Geographical Information Systems in Crime Analysis. *Journal of Human Sciences, 13*(1), 2208. doi:10.14687/ijhs.v13i1.3736

Deb, S., & Tsay, R. S. (2019). SPATIO-TEMPORAL MODELS WITH SPACE-TIME INTERACTION AND THEIR APPLICATIONS TO AIR POLLUTION DATA. *Statistica Sinica, 29*(3), 1181--1207. Retrieved from https://www.jstor.org/stable/26705998

Definition of MISDEMEANOR. (n.d.). Retrieved from www.merriam-webster.com: https://www.merriam-webster.com/dictionary/misdemeanor

Eck, J. E., Weisburd, D., & University, H. (2015). CRIME PLACES IN CRIME THEORY. *Crime and Place: Crime Prevention Studies, 4*, 1-33.

Eck, J. E., Weisburd, D., & University, H. (2015). CRIME PLACES IN CRIME THEORY. *Crime and Place: Crime Prevention Studies, 4*, 1-33.

Hardie, B., & Wikström, P.-O. (2021). Space-Time Budget methodology: facilitating social ecology of crime. doi:10.1002/9781119111931.ch25

Jang, S. J., & Agnew, R. (2015). Strain Theories and Crime. *International Encyclopedia of the Social & Behavioral Sciences*, 495-500. doi:10.1016/B978-0-08-097086-8.45088-9

Jiang, C., Liu, L., Qin, X., Zhou, S., & Liu, K. (2021). Discovering Spatial-Temporal Indication of Crime Association (STICA). *ISPRS International Journal of Geo-Information, 10*(2), 67. doi:10.3390/ijgi10020067

Jiang, Z., & Shen, G. (2019). Prediction of House Price Based on The Back Propagation Neural Network in The Keras Deep Learning Framework. *2019 6th International Conference on Systems and Informatics (ICSAI)*, 1408-1412. doi:10.1109/ICSAI48974.2019.9010071

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends, 2*(1), 20 – 28. doi:10.38094/jastt20165

Kedia, P. (2016). Crime mapping and analysis using GIS. *International Institute of Information Technology*, 1-15.

Khan, M., Ali, A., Alharbi, Y., & Farias, G. (2022). Predicting and Preventing Crime: A Crime Prediction Model Using San Francisco Crime Data by Classification Techniques. *Complexity*, 1-17. doi:10.1155/2022/4830411

Kitteringham, G., & Fennelly, L. J. (2020). Chapter 19 - Environmental crime control. In L. J. Fennelly, *Handbook of Loss Prevention and Crime Prevention (Sixth Edition)* (Sixth Edition ed., pp. 207-222). Butterworth-Heinemann. doi:https://doi.org/10.1016/B978-0-12-817273-5.00019-3

Li, Y., Zhao, Q., & Zhong, C. (2022). GIS and urban data science. *Annals of GIS, 28*(2), 89-92. doi:10.1080/19475683.2022.2070969

Lin, Y.-L., Yen, M.-F., & Yu, L.-C. (2018). Grid-Based Crime Prediction Using Geographical Features. *ISPRS International Journal of Geo-Information, 7*(8), 298. doi:10.3390/ijgi7080298

Liu, Q., Deng, M., Shi, Y., & Wang, J. (2012). A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Computers & Geosciences, 46*, 296-309. doi:https://doi.org/10.1016/j.cageo.2011.12.017

Masiga, C. (2022). TRENDS OF CRIMINAL ACTIVITIES ON SAFETY OF PERSONS AND PROPERTY IN NAIROBI CITY COUNTY, KENYA 2011-2021. 1-29. Retrieved from http://ir-library.ku.ac.ke/handle/123456789/24135

Mbaabu, O. (2020, 12 11). *Introduction to Random Forest in Machine Learning*. Retrieved from Engineering Education (EngEd) Program | Section: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

Mbani, B. O., Kenduiywo, B., & Odera, P. A. (2017). Crime mapping using Artificial Intelligent Agents in Nairobi City County, Kenya. *Geoinformatics & Geostatistics: An Overview*. doi:10.4172/2327-4581.1000171

Mbani, B. O., Odera, P. A., & Kenduiywo, B. (2017, 11). Crime mapping using Artificial Intelligent Agents in Nairobi City County, Kenya. *Geoinformatics & Geostatistics: An Overview, 5*. doi:10.4172/2327-4581.1000171

Ogeto, F. (2018). crime mapping as a tool in crime analysis for crime management. *International Journal of Phytoremediation*.

Onyango, K. O. (2021). Twitter sentiment analysis tool for detecting crime hotspots: a case of Nairobi, Kenya. *Thesis, Strathmore University*. Retrieved from http://hdl.handle.net/11071/12937

Onyeneke, C. C., & Karam, A. H. (2022, 6 24). An Exploratory Study of Crime: Examining Lived Experiences of Crime through Socioeconomic, Demographic, and Physical Characteristics. *Urban Science, 6*(3), 43. doi:10.3390/urbansci6030043

Ruczyński, Ł. (2021, 2 22). *ML Frameworks Compared: Scikit-Learn, Tensorflow, PyTorch and More [Updated]*. Retrieved from www.netguru.com: https://www.netguru.com/blog/top-machine-learning-frameworks-compared

Samoei, P. C. (2018). Role of Information Communication and Technology in Enhancing Security in Urban Areas in Kenya: A Literature Based Review. *Journal of Information & Technology, 2*(1), 17-27.

Sarkar, R., & Anup, S. (2017). *Economics of Sustainable Development.* Business Expert Press.

Sevri, M., Karacan, H., & mAkcayol, M. A. (2017). Crime Analysis Based on Association Rules Using Apriori Algorithm. *International Journal of Information and Electronics Engineering, 7*(3), 99-102. doi:10.18178/ijiee.2017.7.3.669

Shiode, S., & Shiode, N. (2020). A network-based scan statistic for detecting the exact location and extent of hotspots along urban streets. *Computers, Environment and Urban Systems, 83*, 101500. doi:https://doi.org/10.1016/j.compenvurbsys.2020.101500

Singh, A. (2021). *Machine Learning Approach to Crime Prediction and Identification of Hotspots*. Retrieved from linkedin: https://www.linkedin.com/pulse/machine-learning-approach-crime-prediction-hotspots-amshumann-singh/

Skilling, L., & Rogers, C. (2017). Crime prevention and coping mechanisms in neighbourhoods: insights from Kibera, Nairobi. *Crime Prevention and Community Safety, 19*(2), 103. doi:10.1057/s41300-017-0017-4

Sleeuwen, S. v., Ruiter, S., & Steenbeek, W. (2021). Right place, right time? Making crime pattern theory time-specific. *Crime Science, 10*. doi:10.1186/s40163-021-00139-8

Sunil. (2019, 3 11). *Understanding Support Vector Machine algorithm from examples (along with code)*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Thotakura, D. .. (n.d.). Crime: A Conceptual Understanding. *Indian Journal of Applied Research, 4*, 196 - 198. doi:10.15373/2249555X/MAR2014/58

United Nations. (2018, 5 16). Retrieved from United Nations Department of Economic and Social Affairs: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

Venkatesh, A., S, H., Inamdar, J. S., & S, L. (2019). Detection and Analysis of Crime Patterns Using Apriori Algorithm. *International Research Journal of Engineering and Technology, 6*(4), 4119- 4123.

Wibowo, A. H., & Oesman, T. I. (2020). The comparative analysis on the accuracy of k-NN, Naive Bayes, and Decision Tree Algorithms in predicting crimes and criminal actions in Sleman Regency. *Journal of Physics: Conference Series, 1450*, 012076. doi:10.1088/1742-6596/1450/1/012076

Xiong, C., Srivastava, A., Kannan, R., Damle, O., Prasanna, V., & Southers, E. (2019). On Predicting Crime with Heterogeneous Spatial Patterns: Methods and Evaluation. *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 43-51). Association for Computing Machinery. doi:10.1145/3347146.3359374

Yegulalp, S. (2018, 6 6). *What is TensorFlow? The machine learning library explained*. Retrieved from InfoWorld: https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html

Zhao, X., & Tang, J. (2018). Crime in Urban Areas: A Data Mining Perspective. *SIGKDD Explor. Newsl., 20*(1), 1-12. doi:10.1145/3229329.3229331

Zhou, G., Lin, J., & Zheng, W. (2012). A Web-based Geographical Information System for Crime Mapping and Decision Support. 147 - 150. doi:10.1109/ICCPS.2012.6384228

2020 NATIONAL CRIME MAPPING: CRIME PATTERNS AND TRENDS IN KENYA. (2020). In *Crime Research Centre* (ISBN 978-9914-9955-3-4). https://www.crimeresearch.go.ke/2020-national-crime-mapping-crime-patterns-and-trends-in-kenya/

Appendices

Appendix A: Ethical Approval

Strathmore UNIVERSITY

24th March 2023

Ms Echessa Rita,
rita.echessa@strathmore.edu

Dear Ms Echessa,

RE: A Model for Mapping Crime Hot-spots: A Case of Nairobi

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC1651/23.** The approval period is from **24th March 2023 to 23rd March 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-ISERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-ISERC

Cc: Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
24-Mar-2023
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA