University of
Salford
MANCHESTER

Real time crime prediction using social media

Fatai Jimoh

School of Science, Engineering and Environment

University of Salford, Salford, UK

Submitted in partial fulfilment of the requirements

of the degree of Doctor of Philosophy

October 2022

Table of Contents

Declaration of Authorship

I, Fatai Jimoh, declare that this thesis titled Real time crime prediction using social media and the work presented within are my own and no part of the work contained in this thesis has been given in support of any application for any other degree or qualification at the University of Salford or any other university or institution of learning.

I have maintained professional integrity during all aspects of my research degree and, I have followed the Institutional Code of Practice and the Regulations for Postgraduate Research Degrees.

Where I have consulted the published work of others, this is always clearly attributed. Where I have quoted from the work of others, the source is always given. This research received no external funding, and I declare no conflict of interest.

Acknowledgments

This opportunity has been granted to me by the Almighty Allah out of His boundless kindness, for which I am utterly, truly, and unreservedly grateful and appreciative. Without the assistance of my supervisors, family, and friends, it would not have been possible to complete my thesis. I would like to express my gratitude to Professor Mohamad Saraee, my primary advisor, for his exceptional skill as well as his support and guidance from the very beginning to the final draught of my thesis. Throughout my PhD journey, I have also been grateful to Dr. Azar Shahgholian for her additional supervision, inexhaustible assistance, and firm counsel. I would also like to convey my deepest appreciation to every participant in this initiative who volunteered their time for the benefit of others and without whom this endeavour would not have been possible. I am very grateful to my family for their unwavering love and support throughout my life, which has provided inspiration and motivation. I cannot begin to express the depth of my affection and gratitude for them, to whom I owe everything. My beautiful wife, Dr. Habeebah Ibraheem, and my daughters, Hamnah Jimoh, Fatimah Jimoh, and Maryam Jimoh, who encouraged me to persist and assisted me during the whole of my studentship, deserve my deepest gratitude. Without their assistance, empathy, comfort, and selflessness, this task could not have been accomplished. Finally, I would like to express my gratitude to the University of Salford for their kind assistance.

Abstract

There is no doubt that crime is on the increase and has a detrimental influence on a nation's economy despite several attempts of studies on crime prediction to minimise crime rates. Historically, data mining techniques for crime prediction models often rely on historical information and its mostly country specific. In fact, only a few of the earlier studies on crime prediction follow standard data mining procedure. Hence, considering the current worldwide crime trend in which criminals routinely publish their criminal intent on social media and ask others to see and/or engage in different crimes, an alternative, and more dynamic strategy is needed.

The goal of this research is to improve the performance of crime prediction models. Thus, this thesis explores the potential of using information on social media (Twitter) for crime prediction in combination with historical crime data. It also figures out, using data mining techniques, the most relevant feature engineering needed for United Kingdom dataset which could improve crime prediction model performance. Additionally, this study presents a function that could be used by every state in the United Kingdom for data cleansing, pre-processing and feature engineering. A shinny App was also use to display the tweets sentiment trends to prevent crime in near-real time.

Exploratory analysis is essential for revealing the necessary data pre-processing and feature engineering needed prior to feeding the data into the machine learning model for efficient result. Based on earlier documented studies available, this is the first research to do a full exploratory analysis of historical British crime statistics using stop and search historical dataset. Also, based on the findings from the exploratory study, an algorithm was created to clean the data, and prepare it for further analysis and model creation. This is an enormous success because it provides a perfect dataset for future research, particularly for non-experts to utilise in constructing models to forecast crime or conducting investigations in around 32 police districts of the United Kingdom.

Moreover, this study is the first study to present a complete collection of geo-spatial parameters for training a crime prediction model by combining demographic data from the same source in the United Kingdom with hourly sentiment polarity that was not

restricted to Twitter keyword search. Six unique base models that were frequently mentioned in the previous literature was selected and used to train stop-and-search historical crime dataset and evaluated on test data and finally validated with dataset from London and Kent crime datasets.

Two different datasets were created from twitter and historical data (historical crime data with twitter sentiment score and historical data without twitter sentiment score). Six of the most prevalent machine learning classifiers (Random Forest, Decision Tree, K-nearest model, support vector machine, neural network and naïve bayes) were trained and tested on these datasets. Additionally, hyperparameters of each of the six models developed were tweaked using random grid search. Voting classifiers and logistic regression stacked ensemble of different models were also trained and tested on the same datasets to enhance the individual model performance.

In addition, two combinations of stack ensembles of multiple models were constructed to enhance and choose the most suitable models for crime prediction, and based on their performance, the appropriate prediction model for the UK dataset would be selected. In terms of how the research may be interpreted, it differs from most earlier studies that employed Twitter data in that several methodologies were used to show how each attribute contributed to the construction of the model, and the findings were discussed and interpreted in the context of the study.

Further, a shiny app visualisation tool was designed to display the tweets' sentiment score, the text, the users' screen name, and the tweets' vicinity which allows the investigation of any criminal actions in near-real time. The evaluation of the models revealed that Random Forest, Decision Tree, and K nearest neighbour outperformed other models. However, decision trees and Random Forests perform better consistently when evaluated on test data.

CHAPTER 1 : INTRODUCTION

1.1 Introduction

There is no globally agreed definition for crime in modern criminal law, hence, definitions have emerged to meet specific objectives and benefits (Anderson, 2009, Cane and Conaghan, 2008). Literally, crime is defined as an illegitimate act prosecutable by the state or other authority. Crime is an unlawful act punishable by law, affecting not only individuals but the entire society. In other words, an action is said to be a crime if it is declared as such by the relevant and applicable law(Cane and Conaghan, 2008). As a result, the definition of crime varies between countries and is dependent on the country's judicial rulings. For example, in Saudi Arabia, an action is judged a crime from a religious perspective, whereas in most western countries, crimes are labelled based on common and international laws. There are actions that are generally judged as crimes and forbidden across all borders, for example, rape, robbery, and homicide (Mark, June, 2010). Crimes are difficult to solve in regions where frequent social issues are prevalent, and they have an influence on a region's personal happiness, financial success, and quality of life. As the population rises and the income imbalance in society widens, crime is expected to rise. Its rise can also be linked to the emergence of a variety of complex factors like unemployment, poverty, and violent ideologies.

Crime causes unrest in the community, and it is a prevalent issue in larger urban cities. It has far-reaching negative effects on the economy, social norms, and reputation of a country and it contributes significantly to the social and economic loss of any society. As urban areas expand at an unprecedented rate, residents need to maintain a heightened level of vigilance.

The overall crime rate in the United Kingdom rose dramatically in 2018, despite a few initiatives aimed at reducing it (O'Neill et al., 2022). Specific crime trends showed rising rates of murder, robbery, drug-related offences, and weapon possession. The found that violent and sexual crimes were the most common types of crime in the United Kingdom (Statistics, 2022). As an additional complication, various nations have used various strategies to combat crime.

Criminals today use technological advancements not only to commit crimes, but also to avoid detection. The internet's global reach has also made it a breeding ground for ever-more-intelligent criminals. Considering recent terrorist attacks and the use of technology to breach the most secure defence databases, innovative and effective approaches to preventing crime are in high demand. In addition, the explosion of social media has made it increasingly difficult for law enforcement and intelligence agencies to analyse massive amounts of criminal data in a fair and accurate manner(Canton, 2021).

Crime prediction could lower crime rates in high-crime regions, and it is obvious that crime reduction needs a diverse strategy (Aguirre et al., 2019). Although there are limitations evident in the prior studies, which might be due to a lack of understanding of data mining and the deployment of machine learning models to tackle criminal issues. It could also be because people did not know how to pre-process key crime datasets or could not investigate how social media could be used to find out crimes in near real time. This study, therefore, aims to identify and evaluate the most effective data mining methodologies and tactics for crime prediction. This chapter will provide a synopsis of the study by discussing the problem space, research issues, research aims, objectives, and questions, as well as the study's contributions and limitations.

1.2 Research Background

The rise of exploration of social media as an opportunity to perpetrate criminal acts or as a means of disseminating evil or influencing other innocent users to perpetrate evil acts has revealed the limitations of the so-called state-of-the-art prediction models. For instance, in a BBC online interview (BBC, 2020), a woman said, "Their website has nothing bad about it, but when people start reading their stories, they get brainwashed, which makes them commit crimes, and in most cases, violent crimes." Recent research studies are also deficient in real-time prediction or analysis, leading to most crimes being committed before the arrival of law enforcement officers. A good example is an attack on two mosques in New Zealand (NEWS, 2019). The offender not only reported his intention on Instagram, but he also posted a live video. However, the attack had been carried out before he was arrested. This might have happened because the area where the attack was carried out and/or the country was not a crime hotspot area. This was also the case with an incident that happened in the United Kingdom in November 2019, where a suspect invited people on Snapchat to attack JD stores in London (SUN, 2019). They met and looted JD stores and went away with substantial goods from the stores. This highlights the limitations of using a single variable in crime prediction. Cases like these could have been better handled if law enforcement had a real-time automatic detection model that used social media blogs (Medvedeva et al., 2016).

Although numerous crime-fighting tactics have been used in the past, the problems with previous research can be divided into two categories: data collection and prediction techniques. In terms of prediction techniques, hotspot techniques have been widely applied in the field of crime prediction. It is one of the most extensively used methodologies, in which crime density is forecasted using several methods such as Risk Terrain Modelling (RTM), Kernel Density Estimation (KDE), Hotspot mapping, Maximum Likelihood Theory, and Marked Point Process in Predictive Policing, to guide police patrol to some extent (Caplan et al., 2020, Dugato et al., 2020, Kim et al., 2021, Mingche et al., 2011, Mohler, 2014, Wang and Yuan, 2019). Predictive policing necessitates the competence to detect hidden patterns or implications. However, they are only designed to forecast either crime-dense locations or the next places of crime occurrence.

While most research focuses on identifying crime densely concentrated areas (also known as "hotspots"), others try to predict where crimes will occur. Some studies have focused on how to effectively incorporate social media as an alternative or additional source of information to boost the precision of crime prediction models considering the meteoric rise in social media usage. The use of machine learning algorithms in the prediction of crime hotspots has gained traction among researchers since 2019 even though most of these studies only compare the effectiveness of machine learning models to hotspot methods. However, research studies have shown that very little about the knowledge of how machine learning models work has been displayed in the previous work. Using social media as a secondary or tertiary source of data is rarely considered in these kinds of studies. Only a small number of current studies employ a data mining approach; the vast majority instead rely on statistical methods like ARIMA, SARIMA, and regression for crime prediction (Akhter et al., 2018, Nitta et al., 2019, Safat et al., 2021). In addition, integrating data from social media is rarely considered in their research because univariate time series are typically considered in their forecasting, making it challenging to combine data from multiple sources.

However, with the incorporation of machine learning models, crime hotspot and crime forecasting techniques have recently shown marked improvement. Even though the conventional methods of data mining are effectively implemented which has led to inappropriate result or some unreliable model performance. Overfitting of machine learning algorithms may also occur due to small sample size of each grid space, which is one of the most noticeable drawbacks with the use of machine learning in hotspot approaches because the partitioning into grids occurs largely before using machine learning. To determine the type of pre-processing needed by each prediction model, only a few of the recent studies that used feature engineering, feature selection, or class imbalance performed in-depth exploratory analysis of the data. Finally, most of the prior research into crime prediction models relied on data collected in the United States. Since each nation has its own system for recording criminal offences, this highlights the model's lack of generalizability.

Data mining has been applied in studies that incorporate different data sources (Elluri et al., 2019, Esquivel et al., 2019, Kadar et al., 2017) such as historical crime data, socio-economic indexes and demographic information, weather data, and census data into crime prediction. These data are sometimes used separately and in combination with other research projects. However, using an individual data source may not suffice in providing relevant and accurate predictions in some cases. So, caution must be exercised when extracting data from various sources to avoid duplication of data, which may be difficult to trace by the models. The limitation of this approach is that the results may be misleading. Therefore, this work ensures the quality of data by collecting data over the same duration of years and largely from the same source, except for Twitter, which is an external source but within the same time range.

Although some recent studies on crime prediction claim to have achieved high performance Nesa et al. (2022), their inability to predict drug use or drug-related crimes may be due to the approaches traditionally used in tackling crime prediction. Some of those crimes may not have been committed without drug intake or misuse. Since crime rates among teenagers and youth are higher than among adults and across genders, and since it is also said to be more popular among the black ethnicity groups than the white Pierce et al. (2017), adequate data exploration of the historical crime data is required in the Greater Manchester crime information to reveal the distribution of crimes across age and ethnicity groups, before model building. If not, this could result in inaccuracies in the model as the data would not be a good representation of the population.

Data collection methods are another difficulty faced by current research studies. One of the means of extracting information for crime prediction is through suspects' owned devices, like personal computers, laptops, notebooks, and smartphones, which serve as targets for further investigation (Bogomolov et al., 2014). These devices may hold important evidence relevant to the case under investigation and may also have valuable information about the social networks of the suspect, through which other criminals may be found. However, this raises ethical considerations. Also, in most cases, this source of information is useful only after the offence has been committed. It cannot serve as a precautionary tool for the public, especially when the criminal's intention is declared prior to the criminal event. Hence, this approach is not useful for the automatic detection or instant prediction of crime since the accessibility of the devices might require the user’s permission.

Crime data can also be generated from the official websites of legal information units. The problem with this is that some websites require official permission, and the available data from such units is insufficient for extensive research. Besides, most organisations do not update their websites in real-time, and this hinders the use of their data for automatic detection of any crime. While news websites are more regularly updated, events would have occurred before the update. So, none of the above are real-time data sources, and they are not enough to automatically find crimes.

Most studies on crime prediction utilise data from the United States. Few have utilised the United Kingdom data set for predicting crime. The few that utilised the UK dataset utilised the street-level dataset, which has an issue with the date because it is not the complete date. Due to data protection, only monthly information is provided on the website of the police department. This data set is deficient in quality even though it holds a vast quantity of information. Those that have used this dataset have had to extrapolate the remaining part of the dates, the reliability of the models built with this dataset would be inadequate.

Although other types of crime datasets (stop and search crime information) can be examined to obtain the demographic information of a specific offender, most researchers avoid this dataset because it has multiple categorical information as well as many missing values and data entry errors.

Also, most of the recent studies that incorporated social demographic factors into their prediction models either harvested the features from past census data or from databases where there is a substantial difference in collection dates when compared with historical crime data, and this could cause the models to be biased (Chandrasekar et al., 2015). It was also observed from recent research studies that crime prediction models could be location, data structure, time, and adequate pre-processing based. This is also noticed in the performance of decision trees, Nave Bayes, neural networks, and random forests in the prediction of crime occurrences in the works cited previously. This implies that some algorithms are better at predicting specific crime elements than others (El Bour et al., 2018, Hossain et al., 2020b, Llaha, 2020). Also, the application of data mining standard procedures has not been adequately applied in predicting crime occurrence. Although some of the past work achieved higher accuracy, the fact that the rate of crime is always increasing makes it clear that the previous models need improvement.

The recent success of online social networking sites and the continuous rise in crime rates have led researchers from a variety of fields, including law enforcement, social sciences, crime profiling, policing, and social media mining, to investigate social media as an alternative or as an added source of information to improve crime prediction. The problem with this, however, is that they are unable to fully utilise the potential of social media to explore the vast amounts of data available for real-time crime trend prediction. This is problematic because most of the research focuses on the use of specific keywords to extract information from social media or to extract characteristics from massive datasets. Most criminals' intentions would not be depicted negatively on social media, thereby posing a problem for crime prediction. The focus should not only be on data collection, but also on determining how these positive words can be investigated to determine if they have a positive meaning.

1.3 Research Problems

A review of the relevant literature revealed a number of important themes about predicting crime. Existing research has significantly contributed to our comprehension of the significance of social media in crime prediction. However, there are a number of significant research gaps that must be filled.

First, most studies from the UK and some other countries have only looked at keywords related to crime to get information from social media (Karmakar and Das, 2021, Kounadi et al., 2015, Tundis et al., 2019, Wang et al., 2019, Williams et al., 2019, Williams et al., 2017). This can make it difficult to obtain information beneficial for detecting offences in real-time. Given how important social media is for predicting crime, this is an important research gap.

Second, existing research has insufficiently utilised standard data mining to determine the best method for predicting crime or to solve crime prediction issues. This could be the result of insufficient data exploration, which would have led to the most appropriate data pre-processing and feature engineering required to create the most reliable crime prediction model (Castro et al., 2020, Esquivel et al., 2020, Gayathri et al., 2021). This gap is very important because crimes and illegal activities can hurt a country's economy and society as a whole. Clearly, addressing these research deficits is crucial.

Third, a few studies from the UK on crime prediction used unreliable data from a street-level crime prediction dataset on the UK police website to predict crime, and the few that did use reliable data did not consider demographic or social media effects on crime prediction (Kounadi et al., 2015, Toppireddy et al., 2018). This is another significant gap that must be filled, as the outcome of their research could be deceptive.

Fourth, some of the studies examined external factors that were not included in their database due to a lack of time or space in their models (Corso, 2015, Williams et al., 2017).

Fifth, regression analysis has been the only way to look at social media's ability to predict crime and find a correlation between the daily polarity of Twitter sentiment and the crime rate (Aghababaei and Makrehchi, 2018). In a previous study, it was suggested that combining the polarity of hourly sentiment extracted from Twitter post (tweets) with other demographic factors could help reduce crime.

By filling in these big gaps in research, my work would help advance knowledge and find practical uses in the field of crime prediction.

1.4 Research Aims:

Unlike previous studies, this research will consider the possible influence of both demographic factors as well as social media in the prediction of crime occurrences. Since Twitter is a popular social media platform where users have freedom of expression, it assumes that information from twitter and other social media website in addition to other socio-economic and demographic information would better reveal trends in crime prediction that previous research could not provide.

Given the paucity of research on the detailed use of data mining techniques and the analysis of the potential influence of social media information (Twitter) in crime prediction, the purpose of this study will be to identify and assess the best data mining approaches that would be suited for crime prediction using Greater Manchester as a case study for the United Kingdom.

1.5 Research Objectives:

The objective of this study is as follows:

- To conduct an extensive exploratory analysis of data (EDA), investigate, and reveal the relationship between different types of crime in the United Kingdom. Consequently, creating functions that can be applied to any crime dataset from the UK police stop-and-search dataset to perform data cleaning, pre-processing, and feature engineering as necessary.
- To train the six most prevalent machine learning base classifiers discovered in previous studies, as well as an ensemble of these base models. The performance of these models would be evaluated according to their accuracy, precision, recall,

and f1-score on the test data. Then, the performances of the models would be compared, and the most suitable model would be selected for crime prediction.

- A visualisation tool would be developed to reveal in near real-time the tweets, the screen name, and the approximate location of the person sending the tweets.

1.6 Research contributions

This study falls under the topic of crime prediction models, which evaluate the potential effect of merging historical crime data with hourly sentiment polarity extracted from social media information (Twitter) to predict crime in real time. The present study attempts to address multiple gaps and, in doing so, makes important contributions.

First, this study adds to the limited research that has been done on using keywords to get information from social media. This study is one of the first to look at the hourly sentiment polarity of all tweets in the UK without limiting the text collected to specific keywords. This is important for finding hidden information in the text, no matter how positive or negative it is.

Second, this study contributes to the existing literature on crime prediction by utilising standard data mining techniques to determine the optimal method for predicting crime in the United Kingdom. This is also one of the first studies to use standard data mining techniques to address crime-related issues. So, the best machine learning model for predicting crime in the United Kingdom was chosen after it was trained with demographic data from the same source as historical criminal records in the United Kingdom and hourly sentiment polarity that was not based on Twitter keywords.

Third, this study adds to the limited amount of research that has been done on using unreliable data from a street-level crime prediction dataset on the UK police website to predict crime. To make sure that researchers in the future have access to accurate and reliable crime records, an algorithm was made to get stop-and-search crime data from the UK police website, clean it up, and get it ready for analysis and model development. This is a huge accomplishment because it means that reliable and high-quality datasets will be available for future research, especially for non-experts to use when making models to predict crime or look into crimes in the United Kingdom.

Fourth, according to the literature review done for this study, no previous research has looked at the effects of demographic factors from the same source as this study, as well as the effects of hourly sentiment polarity, on predicting crime in the United Kingdom. This study demonstrated that demographic variables and hourly sentiment polarity extracted from social media (Twitter) can aid in the detection of offences in real-time or near real-time.

Fifth, existing research has focused on utilising Twitter content to predict either crime rate, location, and time, or trend analysis. This is one of the earliest studies to extract hourly sentiment polarity from Twitter in order to visualise the real-time crime trend and detect crime in real-time or near-real-time. This research has therefore effectively identified a temporary solution to the problem of real-time crime detection by developing a visualisation tool that displays sentiment polarity trends, tweets, and the approximate location of tweets in near real-time.

Hence, based on the CRISP-DM standard mining procedure, the study wants to find out how much Twitter information and demographic information about criminals help. This study would add to the new field of computational criminology by looking at past crime records through the lens of hourly Twitter sentiment polarity and criminal demographic factors. It would also help find crimes in real time.

1.7 The research structures

The study's context was provided in chapter one. The research problems, aims, and questions have been outlined, and the significance of this study has been justified. The study's weaknesses have also been examined.

The previous work on crime prediction from current research work will be examined in chapter two to identify research direction and research gaps within the larger framework of crime prediction.

In chapter three, a systematic analysis of the current literature will be conducted to identify critical skill development methodologies and tactics in the contest of crime prediction, particularly with the inclusion of social media data.

In chapter four, the theoretical framework will be presented. The adoption of a qualitative, inductive research approach will be justified, and the broader research design will be discussed, including the limitations.

A full discussion of the research results and analysis is offered in chapter five, and the conclusion of the research study and future research recommendations are appropriately covered in chapter six.

CHAPTER 2 : SYSTEMATIC LITERATURE REVIERW

2.1 Background and Related work

In this research, the feasibility of applying conventional data mining methods to the issue of predicting criminal behaviour is examined. To put a theoretical framework behind a proposed approach to using technology to solve crimes, we need one. Researchers in criminology and criminal psychology have studied crime extensively to identify patterns and provide explanations. At present, many theories exist to shed light on criminal behaviour, victim psychology, and the role of the surrounding environment in shaping all three. Several of these theories, and their applicability to crime prevention, are discussed in this study. Any strategy to reduce criminal activity needs to take these theories into account. The first step in crafting an approach to reducing crime is determining the causes of it. Start with a basic fact, like crime statistics, which can be broken down into more specific categories using spatial, temporal, and demographic information. This means that crime prediction models the researchers need to think about the best ways to incorporate additional data sources, such as social media data, to solve some crimes in real-time. Therefore, the purpose of this study is to develop a procedure that can incorporate all these elements into a unified answer.

This research was motivated by a desire to unify current concepts and methods in criminology, sociology, geography, and computer science. This study also aims to conduct a systematic literature review in crime predictive analytics, with a focus on the application of standard data mining procedures to the integration of historical crime data with social media information (Twitter), to evaluate the state of the art in terms of concepts and methods, given the unprecedented rate at which empirical studies are published. The study's research questions are listed below:

1. What are the different sorts of crime prediction for which data mining is useful?

2. What are the most frequent crime prediction methods?

3. What are the technical differences and similarities across crime prediction models?

4. How is predictive performance in crime prediction measured?

5. What are the most popular model validation procedures used in crime prediction?

6. What are the primary constraints and dependencies of crime prediction performance?

7. What are the primary sources of Twitter information for crime prediction?

8. What are the primary methods for integrating Twitter data with historical crime data?

9. What role may social media play in enhancing crime prediction models?

Prior to conducting the systematic literature review, a summary of the previous research on crime prediction and analytics ("related work") would be discussed. On this basis, it is believed that the issues with existing methods will be clearly understood. The systematic review was conducted in accordance with the steps outlined below:

Delivering the results ("Results") and analysing them in the form of a SWOT analysis ("Discussion"), and then discussing the methodology used to discover articles and ensure research quality ("Discussion") ("Methods"). In the final section, "Conclusion," the key findings of each study was summarised. It is hope that this research will shed light on future research topics and demonstrate the difficulties that may be encountered when attempting to predict crime in a particular area.

The remaining portions of this chapter are organised as follows: Related works is the section 2.2, Section 2.3 is a systematic literature review. The section concludes with Section 2.4, which discusses future projects.

2.2 Related work

This section is divided further into subsections based on various methods and data sources used in the existing literature. As a result, this section will investigate the following subheadings: Descriptive methods, Geographical Information system, Mathematical modeling methods, forecasting, regression, classification and others, social media, ensemble method, metric evaluation, and model comparison.

2.2.1 Descriptive

Badawy et al. (2018) developed analytics methods to analyse crimes by crime type and region, as well as heat maps for crime distribution, using Erie Police Department crime data. They conducted additional research in two high-crime areas. Their findings provided useful insights to decision-makers regarding crime prediction and prevention, and surveillance cameras have been installed in high-crime areas.

In addition, state governments across the United States that issued mandatory stay-at-home orders in response to the COVID-19 pandemic at the end of March 2020 realised that crime has decreased significantly since the year preceding the pandemic, 2019. However, there are a few indications that the decline in crime is due to a decline in minor offences, which are typically committed in peer groups. Simultaneously, serious crimes that are not typically committed by co-offenders (such as homicide and intimate partner violence) have remained stable or increased. As a result, the crime decline appears to be concealing a very troubling trend in which the number of homicides has remained unchanged, and the number of intimate partner assaults has increased. Because many offenders would likely commit less severe crimes in a world without a pandemic, we raise the possibility that mandatory lockdown orders have placed juvenile offenders in environments where intimate partner violence, serious battery, and homicide are prevalent (Abrams, 2021).

To determine whether the COVID-19 pandemic altered crime patterns in Mexico City, as measured by changes in public transportation passenger numbers. Following the discovery of the pandemic or the implementation of a national lockdown, Estvez-Soto (2021) revealed that most types of crime decreased significantly. In addition, the study revealed that a portion of the observed declines was attributable to a decline in the number of passengers using public transportation. However, they suggest that changes in mobility explain a portion of the observed decreases in crime, with significant variations by type of crime.

In the field of epidemiology, spatial-temporal Bayesian modelling, a regional statistics-based technique, is widely employed. Hu et al. (2018) employed Bayesian theory to develop a spatial-temporal Bayesian model tailored to urban crime to analyse its spatial-temporal patterns and identify any emerging trends. In addition, the associated covariates and their variations are investigated. From January to August of 2013, data on burglaries in Wuhan, China were analysed using the model. It was observed from their research that the burglary crime rate is strongly correlated with the average number of residents and the number of internet cafes in a community. Other socioeconomic factors associated with the crime rate include the number of internet cafes, hotels, shopping centres, the unemployment rate, and residential zones.

2.2.2 Geographical Information system

2.2.2.1 Kernel Density Estimation (KDE)

The convergence of public data and statistical modelling has enabled public safety officials to priorities the deployment of limited resources following predicted crime patterns. Current crime prediction methods are trained using various demographic factors and historical crime information. Due to the lack of observations at finer resolutions, scientists have favoured global models (such as those of entire cities) (e.g., ZIP codes). These global models and their underlying assumptions contradicted the evidence that the relationship between crime and demographic factors varies spatially. Al Boni and Gerber (2017a) created region-specific crime prediction models using hierarchical and multi-task statistical learning.

Multiple cutting-edge global models are outperformed in terms of accuracy of prediction by out-of-sample analysis of crime data from the real world. Caplan et al. (2020) investigated the optimal locations for allocating resources and focus. As a result, they examined how spatial vulnerabilities and exposures could be used to determine the most effective police target areas. Using the Theory of Risky Places, the results of these two methods were combined to predict crime more accurately. Risk terrain estimation (RTM) was used to measure crime exposures, while risk terrain modelling was utilised to determine crime vulnerabilities. The model was evaluated using a year's worth of data on street robberies in Brooklyn, New York.

Prediction accuracy index (PAI) was used to evaluate the performance of KDE, RTM, and their combined method over the course of one and three months. They discovered that the integrated method produced the most accurate forecasts on average and with the highest frequency. Their findings indicate that place-based policing and related policies can be improved with the aid of actionable intelligence generated by multiple crime analysis tools designed to measure various aspects of how crime changes over time and space Caplan et al. (2020). To help police practitioners and researchers make better use of KDE for targeting policing and crime prevention efforts, Chainey (2013) demonstrated that the implementation KDE can be improved by two key parameters: cell size and bandwidth size. As a result, experiments using different cell sizes and bandwidth values were conducted with data on residential burglaries and violent assaults as part of their research. According to their findings, cell size has a negligible effect on the ability of KDE crime hotspot maps to predict spatial crime patterns, while bandwidth size does.

The study of Hart and Zandbergen (2014) was backed by the findings of Chainey (2013) after examining the effects of user-defined parameter settings, including interpolation method, grid cell size, and bandwidth, on the predictive accuracy of crime hotspot maps generated from kernel density estimation (KDE). Across two types of interpersonal violence (such as aggravated assault and robbery) and two types of property crime, the impact of parameter setting variations on prospective KDE maps is investigated (e.g., commercial burglary and motor vehicle theft). They discovered that the interpolation method has a significant impact on predictive accuracy, whereas grid cell size has little to no impact and bandwidth has a moderate impact.

Flaxman et al. (2019) describe a generic spatiotemporal event forecasting method developed for the National Institute of Justice's (NIJ) Real-Time Crime Forecasting Challenge (NIJ (2017)). It was a spatiotemporal forecasting model that combined regularised supervised learning with scalable randomised Reproducing Kernel Hilbert Space (RKHS) methods for approximating Gaussian processes. While the smoothing kernels represent the two main approaches currently used in crime forecasting, KDE and self-exciting point process (SEPP), the RKHS component of the model can be interpreted as an interpolation to the well-known log-Gaussian Cox Process model.

They use the Poisson likelihood and highly efficient gradient-based optimization methods to discretize the spatiotemporal point pattern and learn a log-intensity function for inference. Cross-validation was used to learn model hyperparameters such as RKHS approximation quality, spatial and temporal kernel length scales, number of autoregressive lags and bandwidths for smoothing kernels, and cell shape, size, and rotation. For sparse events, the resulting predictions outperformed baseline KDE estimates and SEPP models.

Several relevant approaches in the literature assert that Kernel Density Estimation (KDE) can accurately predict crime and outperform other crime prediction methods. However, none of these strategies are effective due to the underlying assumption that police patrols are limited by road networks. Moreover, none of these approaches propose the continuous identification of crime hotspots. Junior et al. (2019) consequently presented two distinct algorithms.

1: Polygon Hotspots Approximated to Road Network (PHAR), a batch KDE algorithm that outputs crime hotspots approximated to the road network and helps allocate police patrols to prevent new crimes;

2: i- PHAR is an algorithm that incrementally revises the previous KDE calculation depending on newly reported crime incidents. This expedites the detection of crime hotspots because the KDE algorithm does not need to be recalculated for new data streams. In terms of the quality of their results, the performance of their PHAR and i-PHAR models was validated based on the state-of-the-art in the literature. In the experimental evaluation, Ristea et al. (2020) assessed crime prediction models by integrating tweets and extra variables as covariables of crime to historical crime data. During the same time frame, they analysed the spatial distribution of crime and its correlation with tweets. Using feature selection, the most vital properties were identified. Their findings revealed that Twitter data and a subset of violent tweets are beneficial for creating prediction models for the seven crime types analysed during home and away sporting events and nongame days, albeit to variable degrees.

2.2.2.2 RTM

Caplan et al. (2011) predicted shooting-related crimes using RTM. Using a range of contextual factors pertinent to the opportunity structure of shootings, RTM risk terrain maps estimate the risks of future shootings as they are dispersed throughout the landscape. On risk terrain maps across two six-month periods, the prediction accuracy of this approach was evaluated and compared to the forecasting capacity of retrospective hot spot maps. Risk terrains provide a statistically meaningful projection of future shootings across a wide range of cut sites and are far more accurate than retroactive hot spot mapping, according to their research.

The same method has also been adopted by Barnum et al. (2017) in their study to identify place features that increase the risk of robbery and their specific spatial influence in Chicago, Illinois, Newark, New Jersey, and Kansas City, Missouri. It was observed that the risk factors for robbery are similar but not identical across environments. However, some factors appear to be more susceptible to theft and have had a greater effect on their surrounding environment than others. The researchers concluded that the organisation of the environment influences the relationships and influences of individual place characteristics on crime.

Risk Terrain Modeling (RTM) was utilised by Dugato et al. (2020) to forecast Camorra homicides in an Italian city. They began by identifying and evaluating the underlying risk factors that can affect the probability of homicide. These factors were then used to predict the most likely locations for future events. In their study, they discovered that previous homicides, drug dealing, confiscated assets, and rivalries between groups could predict up to 85 percent of 2012 mafia homicides, identifying eleven percent of city areas at the highest risk, while socioeconomic conditions are not significantly associated with the risk of homicide. Even in a small area, the same risk factors can combine in different ways, resulting in areas with the same level of risk that require targeted solutions.

It was also used by Gerell (2018) to identify areas with high crime rates and high crime victimisation near bus stops and their result is compared to the same models with bus passenger count as the exposure variable. Their models consider the possible influence of environmental factors by fitting multi-level models with neighborhood-level predictors of concentrated disadvantage and collective efficacy. Certain types of facilities are risk factors for crime, but not for victimisation, according to the Gerell (2018). This provides new insights into how the flow of people affects forecasting, such as the fact that a school is a spatial risk factor for the crime despite not being associated with an increased risk per person. The findings also indicate that a neighborhood's level of collective efficacy is a stable and significant risk factor for both crime and victimisation, highlighting the potential for improved crime forecasting by combining different spatial and theoretical perspectives.

Ohyama and Amemiya (2018) chose RTM as a suitable crime prediction model because it is primarily based on environmental factors associated with crime and does not require previous crime data due to Japan's low crime rate in comparison to other developed nation like the US. Prior to 2014, they had applied (RTM) to cases of vehicle theft in Fukuoka, Japan, and evaluated the model's performance using the hit rate and predictive accuracy index. RTM outperformed models such as KDE, ProMap, and SEPP that use past crime events to predict future crime.

2.2.2.3 Near Repeat

The near-repetition phenomenon suggests that each victimisation may form a spatial and temporal contagion-like pattern. In recent years, in addition to the development of environmental criminology, the spatial and temporal distribution of crime has been studied extensively, and numerous crime patterns have been identified. These findings made substantial contributions to the advancement of problem-oriented policing policy, particularly crime forecasting and the deployment of police resources. However, this research, particularly spatial and temporal crime analysis, remains unavailable in China for various reasons. To investigate property crime patterns in the second largest city in China, Chen et al. (2013b) collected 5-month burglary recordings in Beijing and applied spatiotemporal methods. First, they conducted a descriptive analysis of crime patterns in space and time separately; their findings indicate that crimes clustered significantly in space and were sequentially correlated with history and time. Utilizing a disease contagion testing strategy, the spatial and temporal burglary risk was evaluated. Their findings revealed that a high-risk offender communicates for at least three weeks within 200m, exhibiting a pattern of repetition. This result is supported by the research conducted by Wang and Liu (2017).

Using the near-repeat method, a recent burglary dataset from southeastern China was utilised to analyse the risk levels around hot spots. Before and after hot spots, they proposed a temporal expanded near-repeat matrix for quantifying risk undulation. Their research showed that hot spots always form. They observed that high-risk space-time regions are always spatially and temporally variable, and regions near hot spots share this risk. In addition, they showed that crime risks manifest as a wave diffusion process around hot spots, and they provided a comprehensive analysis of criminal patterns, which not only advances prior findings but also provides valuable research results for crime prediction and prevention.

This is also supported by prior research demonstrating that when a crime occurs, the likelihood of crime in nearby areas increases (Chen et al., 2013b, Wang and Liu, 2017). Considering this, earlier grid-based studies of crime prediction combined all the cells surrounding the predicted location. In contrast to a geographic information system (GIS), the actual land is continuous as opposed to a collection of independent cells. Due to the variable nature of detailed method crime patterns, it is necessary for crime prediction to comprehend the spatial characteristics of the surrounding land. Therefore, Kim et al. (2021) utilised the Max-p region model (a spatial clustering technique) to classify cells with similar spatial characteristics, and compared its performance to that of the existing method, random forest (a tree-based machine learning model). According to their findings, spatial clustering increased the F1 score of the model by approximately 2%. Consequently, the physical environmental factors influenced by the specific method of crime vary. It was also demonstrated that the same person is likely to commit another crime near the scene of the initial offence. This means that repeat crimes are likely to occur in areas with similar spatial characteristics to the initial crime scene.

2.2.2.4 Hotspot Mapping

Using ArcGIS, three years of theft-related crime data were analysed. By applying theft location to a standard point in each district, their findings were deemed useful for predicting thefts. Baloian et al. (2017) presented a solution for crime prediction in large Chilean cities. Its innovative strategy included three autonomous software modules that make predictions based on distinct algorithms.

Crime Hotspots are regions with a higher crime rate than the national average. Yang et al. (2018a) presented a platform that predicts and visualises crime hotspots based on multiple data types. Their platform is said to continuously collect crime data, as well as urban and social media data. Using the extracted features, it then identifies crime hotspots and displays them on an interactive map.

Crime prediction contributes significantly to the improvement of public safety in cities (Zhao and Tang, 2017b). With the development of advanced telecommunications and intelligent transportation in urban areas, urban regions become increasingly interconnected and integrated. In-depth geographical and contextual inter-area spatial correlations between regions are difficult to capture. presented a Grid-embedding based Spatio-Temporal correlation (GeST) model, which consists of a module for crime graph prediction and a module for grid embedding. Crime prediction is essential for enhancing public safety and reducing crime's financial impact. It has also been observed sophisticated methods for collecting and integrating urban, mobile, and public service data at the granular level. This information improves our understanding of the dynamics of crime and has the potential to enhance crime prediction. Zhao and Tang (2017b) examined the use of temporal-spatial correlations in urban data for crime prediction. They established that temporal-spatial correlations in crime indeed exist, and they devised a logical method for modelling these associations inside the unified framework TCP for crime forecasting. They conducted experiments using real-world data and found that their suggested framework worked well.

2.2.3 Mathematical Modelling and Others

In a quest to discover criminal hangouts, taking into account the time and location of previous crimes and predicting the potential location of the next crime, one must consider the time and location of previous crimes. Ke and Jin (2014) used probability and statistical techniques to calculate the likelihood that each point would become a hangout for criminals. For the development of the geographic profile, the Distance Function Method and the Distribution Function Method were investigated over the course of three distinct phases. Using the distance function's output as a benchmark, the criminal's anchor point could be identified in stage one. At the second stage, the Multivariate Analysis Method was used to define the Euclidean and Manhattan Distances between the criminals' hangouts and the locations of the previous crime scenes. Based on the distributional characteristics of these distances, the corresponding distribution function was chosen to derive the concrete expression. Crime hotspots are regions where the crime rate is significantly higher than the national average. Multiple data types were utilised by Yang et al. (2018a) to predict and visualise crime hotspots. Apparently, their platform continuously collects crime data, along with urban and social media data. It then identifies crime hotspots using the extracted features and displays them on an interactive map. Crime prediction significantly contributes to the improvement of urban public safety.

With the development of advanced communications technologies and intelligent transportation in urban areas, urban regions have become more interconnected and integrated. It is difficult to capture in-depth geographical and contextual inter-area spatial correlations between regions. The Grid-embedding based Spatial-Temporal Correlation (GeST) model consists of modules for crime graph prediction and grid embedding. Predicting crime is essential for enhancing public safety and reducing the economic impact of crime. Researchers have observed sophisticated methods for collecting and integrating granular data on urban, mobile, and public services.

This data enhances our comprehension of the dynamics of criminal behaviour and has the potential to improve crime prediction. Zhao and Tang (2017b) investigated the use of temporal-spatial correlations in urban data for crime prediction factors, as well as the Analytic Hierarchy Process to compute different weights of social index factors from different regions of the world. Consequently, the weighted scores for each domain were determined. They concluded that calculating the probability distribution of where the next crime will occur based on both natural and social factors. This enables the system to determine where the next crime is most likely to occur.

To keep up with the rapid development caused by the social and economic transformations of urbanization and various threat poses by urbanisation. A crime risk prediction system is essential to urban crime prevention and control, as well as system improvement. Therefore, Han et al. (2020) presented a model of daily crime prediction by combining Long Short-Term Memory Network (LSTM) and Spatial-Temporal Graph Convolutional Network (ST-GCN) to automatically and effectively detect high-risk areas in a city in order to combat crime-related issues in urban communities. The model is supported by topological maps of urban communities, which primarily consist of two modules — the spatial-temporal feature extraction module and the temporal feature extraction module — to extract the factors of theft crimes collectively. They conducted an experimental analysis of Chicago's existing crime statistics and discovered that the integrated model accurately predicts the number of crimes over a sliding time range.

Criminals do not commit crimes when opportunity presents itself in a concentrated or familiar area as revealed by numerous criminological studies. Consequently, using crime pattern theory, it is possible to create a model capable of analysing past crime data and predicting future criminal activity. Mahmud et al. (2017) introduced CRIMECAST, a prediction and strategy direction software that utilises a probabilistic model and an artificial neural network to predict the likelihood of future criminal activity. Their software is a spatial crime analysis method that predominantly employs actual historical crime data. It forecasts criminal activity, outlines a master plan on a map, and raises a security flag. According to reports, their product outperformed other crime prediction methods.

In developing countries like Bangladesh, street crime is a major problem. In Dhaka, Bangladesh, one of the world's most populous megacities, despite the fact that this problem has been identified for a long time, there is no visible solution or action to combat or overcome these street crimes. Using historical Dhaka city street crime data, Parvez et al. (2016) proposed a novel spatiotemporal street crime prediction model that forecasts the likelihood of a crime occurring in a specific region at a specific time. When predicting a future crime, their algorithm took into account the spatial and temporal proximity of previous crimes. Their model was found to have a sensitivity of 79.24 percent and a specificity of 68.2 percent after being evaluated. In other words, their approach can accurately identify 79.24 percent of crime-ridden regions and 68.2 percent of crime-free areas. This indicates that their model cannot forecast around 20% of crime hotspots, which is a significant amount and would have a significant impact on the country's social and economic life.

In their study, Ristea et al. (2018) predicted Portland, Oregon hotspots for street crime. Using geographically weighted regression (GWR) models, the geocoded Twitter post predictors were determined. Two distinct characteristics were extracted from the Twitter data. The first is the population at risk of becoming a victim of street crime, and the second is tweets pertaining to criminal activity. These two variables were utilised by GWR to develop models depicting future hotspots for street crime. In only 1% of the study area, the predicted hotspots accounted for over 23% of future street crimes. This exceeded the capabilities of a baseline method.

As a major social issue during urban development, crime is closely tied to socioeconomic, geographical, and environmental factors. Traditional crime prediction models reveal the spatiotemporal dynamics of crime risks, but disregard the environmental context of crime hotspots. As a result, it is challenging to improve the spatial accuracy of crime prediction. Tang et al. (2019) presented the use of anisotropic diffusion in the traditional crime prediction model to incorporate environmental factors of the evaluated geographic area, with the goal of predicting crime occurrence on a finer scale in terms of spatiotemporal aspects and environmental similarity. The proposed method has a prediction accuracy of 28.8% on average, based on a variety of evaluation criteria. This represents a 77.5 percent improvement over conventional procedures. It is anticipated that their proposed techniques will provide robust policing support in the form of targeted hotspot policing and facilitate long-term community development.

2.2.4 Forecasting

Continuous urbanisation is causing substantial economic and social changes in cities, posing numerous challenges for city management. Given that the crime rate increases as the size of a city increase crime spikes are quickly becoming one of the most critical social issues in large cities. To combat the rise in crime, new technologies enable police departments to access increasing volumes of crime-related data, which can then be analysed to identify patterns and trends, resulting in the more effective deployment of police officers across the territory and crime prevention (Catlett et al., 2018).

Crime forecasting is essential for criminal justice decision-makers and crime prevention efforts (Vomfell et al., 2018b). Accurate crime forecasting can aid law enforcement in more effectively allocating resources, such as patrol routes and placements. Crime is unpredictable and disrupts social life. As the population of Bangladesh grows, so does the rate of crime, which is wreaking havoc on society in a variety of ways. As a result, analysing crime data to gain a better understanding of future crime trends has become critical. Machine learning and data mining techniques can be extremely useful in predicting future crime trends and patterns in this case (Awal et al., 2016a).

Awal et al. (2016a) investigated a linear regression algorithm for predicting future crime trends in Bangladesh in order to assist Bangladesh police and law enforcement agencies in predicting, preventing, or solving future crime in Bangladesh. Using actual crime data from the Bangladesh police website, the linear regression model was trained. Following the training of the model, crime forecasting in various regions of Bangladesh is conducted for dacoit, robbery, murder, women's and children's repression, kidnapping, burglary, and other crimes. Misyrlis et al. (2017) also investigated the predictive ability of traditional regression methods in Portland, Oregon. The study area was subdivided into cells of equal size, and the spatial autocorrelation of crime rates in adjacent cells was analysed. In addition to the cell's time series, they investigated the use of information from neighbouring cells in regression models to improve forecast accuracy. According to the results, the regression approach outperforms the moving window averaging method, especially as the predicted future horizon extends. However, it was reported that the addition of neighbourhood cells decreased the accuracy of the crime prediction model.

To enhance the capabilities of existing models and predict the likelihood of a new crime incident occurring at small spatiotemporal analysis units. Rummens et al. (2017) examined the viability of predictive analysis in an urban setting by examining an ensemble model consisting of logistic regression and a neural network on publicly available crime data that was spatially separated into 200m-by-200m grids. It was a bi-weekly forecast for 2014 based on crime data from the previous three years. Additionally, monthly forecasts were generated based on the time of day and night. Using the direct hit rate, precision, and prediction index, the performance of the forecast model was assessed. Their findings demonstrated that predictive analysis of crime data at the grid level can produce useful predictions. In general, monthly predictions that distinguish between day and night perform better than biweekly ones. This demonstrates that temporal resolution has a significant impact on the accuracy of predictions.

Catlett et al. (2018) utilised spatial analysis and auto-regressive models to identify high-risk urban crime regions and forecast crime trends in each region. Their spatial-temporal crime forecasting model is comprised of a set of crime-dense regions and a set of associated crime predictors, each of which is a predictive model for estimating the number of crimes that will occur in its respective region. The experimental evaluation using actual data from a substantial portion of Chicago demonstrates that the proposed method is effective at predicting crime in space and time over rolling time horizons. In addition, Ashby (2020) utilised a seasonal auto-regressive integrated moving average (SARIMA) model to forecast the expected number, which was then compared to the number of crimes committed during the pandemic to determine the trend of common crimes in the United States during COVID-19. There were no significant changes in the number or frequency of violent crimes committed in public or private residences. In some cities, residential burglaries have decreased while non-residential burglaries have changed little. In some cities, thefts of motor vehicles have decreased, while in others they have increased. Likewise, Butt et al. (2021) proposed a novel method to improve threat visualisation and identify and predict crime hotspot zones in the smart city environment to ensure safety and security in light of potential threats resulting from the significant increase in urban population over the past few decades. First, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) method was used to identify crime hotspots with a higher incidence risk. Second, spatial and temporal data were utilised with the Seasonal Auto-Regressive Integrated Moving Average (SARIMA) to predict the number of crimes that would occur in each high-crime area in the future.

2.2.5 Machine learning

2.2.5.1 Regression

Data analytics is becoming more important in addressing various societal challenges. Belesiotis et al. (2018) investigated how data from various sources can be used to gain insights and predict the spatial distribution of crime in cities. The estimation of their prediction performance revealed the relevance of a variety of data sources. In order to do this, they analysed not just the accuracy of predictions across all studied locations but also the effect of these predictions on the accuracy of identified crime hotspots. They also analysed the impact of individual features in identifying the most significant components. Finally, they studied how model performance and individual characteristics vary across several forms of criminal activity. Approximately 3,000 characteristics were used to generate the model from six independent datasets. According to their results, combining information from several sources significantly improves prediction ability.

Traditional crime prediction models based on census data are limited because they do not account for the complexities and dynamics of human activity. With the rise of ubiquitous computing comes the opportunity to improve such models with data, resulting in better proxies of human presence in cities. Therefore, Kadar and Pletikosa (2018) use large amounts of human mobility data to create a comprehensive set of features for crime prediction, informed by criminology and urban studies theories. They studied the predictive power of averaging and boosting ensemble machine learning methods in predicting yearly counts for various types of crimes in New York City at the census tract level. It was presented in their findings that spatial and spatio-temporal features derived from Foursquare venues and check-ins, subway rides, and taxi rides improved baseline models based on census and Point of interest (POI) data. On a geographical out-of-sample test set, their proposed techniques achieved approximately 65% absolute R2 and approximately 89% on a temporal out-of-sample test set. They demonstrated the significance of population in predicting a region's vulnerability to crime. They also demonstrated that the contribution of human dynamics features varied with crime type.

The crime rate in Bangladesh has risen dramatically in recent years. As a result, Biswas and Basak (2019b) found it critical to analyse and forecast crime so that authorities may readily decrease or prevent crime. In their case, machine learning is being investigated in order to identify crime trends and patterns in Bangladesh. They forecasted crime trends and patterns in Bangladesh by applying machine learning models such as linear regression, polynomial regression, and random forest regression using publicly available datasets acquired from the Bangladesh police's website. The forecasted results were compared to the actual results and the computed model assessment metrics for the various regression models used. It was discovered that on their dataset, polynomial and random forest regression outperformed linear regression in predicting crime trends and patterns.

Similarly in Brazil, relevant rates of violence have climbed in recent years. To tackle this issue and decrease public money expenditure and time spent by public authorities, intelligence and efficiency are essential and to also increase population safety. Da Silva et al. (2020) investigated the potentials of four machine learning models in their study to forecast crime location in the city of Fortaleza, Brazil. It was presented in their result that a base model might be useful in predicting crime and in terms of crime prediction, Decision Tree and Bagging Tree regressor models outperformed other models.

While the use of crime data has been broadly endorsed in the research, it is often limited to large urban areas and isolated databases that do not allow for spatial comparisons. Lamari et al. (2020) presented an efficient machine learning framework capable of predicting spatial crime occurrences at a relatively high resolution while using the U.S. Census Block Group level, without using past crime as a predictor. Their proposed framework was based on a thorough multidisciplinary literature review, which allowed the selection of 188 best-fit crime predictors from socioeconomic, demographic, spatial, and environmental data. Such information is made available on a regular basis for the entire United States. A comparative study of the application of different machine learning algorithms, including generalised linear models, deep learning, and ensemble learning, highlighted the gradient boosting model as the most accurate for the predictions of violent crimes, property crimes, motor vehicle thefts, vandalism, and total crime count. Evaluation of the gradient boosting model on real-world datasets of crimes reported in about eleven U.S. cities revealed that the proposed framework predicts property crimes and violent crimes with an accuracy of 73% and 77% respectively.

Crime is a complex social problem that is caused by a variety of factors, including population factors, political factors, economic factors, social factors, cultural factors, and environmental factors. However, how and to what extent the factors influence crime is mostly determined by industry experts based on their experience. It almost certainly introduces the risk of subjective experience, and it is difficult to quantify the effect of objective factors. Another issue is that the feature dimension is too large when choosing multi-factor indicators based on experience, resulting in inefficient operation due to doped insignificant factors.

Shi (2020) presented a bagging method which was based on the "good and different" principle. They employed heterogeneous base classifiers to create an integrated learning device that identified the impact of occurrence factors and then improved the efficiency and accuracy of crime prediction with fewer dimensions of factors. Their experimental results revealed that the Ensemble feature voting (EFV) bagging algorithm proposed has better generalisation ability and stability than other algorithms, as well as better prediction accuracy on the test sample. Also, it was said that their model didn't need any prior knowledge to manually set the selected feature subset dimension, which is helpful for analysing and predicting criminal data.

To estimate the impact of the COVID-19 pandemic on crime, data from 25 major US cities were compiled. There was an immediate decline in criminal incidents and arrests, particularly for drug crimes, theft, residential burglaries, and the most violent offenses. The decline appears to precede home detention orders, and arrests appear to follow a similar pattern, according to the reports. In most cities, there has been no decline in homicides and shootings, but there has been an increase in non-residential burglary and auto theft, suggesting that criminal activity has been redirected to areas with fewer people. In Pittsburgh, New York City, San Francisco, Philadelphia, Washington, D.C., and Chicago, the overall rate of crime decreased by at least 35 percent. However, most of the observed changes were not attributable to changes in crime reporting, according to evidence from police-initiated reports and geographic variation in crime change (Abrams, 2021).

Forradellas et al. (2021) suggested a crime prediction model based on Buenos Aires neighbourhoods. Using one of the data science techniques (Sample, Explore, Modify, Model, and Assess), and Buenos Aires crimes dataset reported between 2016 and 2019. They created a crime prediction model based on the city using k-means clustering to classify the areas and a neural network to predict crime. The mean absolute error (MAE) in the validation and training sets was 0.3317 and 0.4095, respectively.

Also, due ongoing research as a result of increase in the number of criminal records, to track the criminal nature and behaviour for better understanding and to secure the society from the criminal. Mishra et al. (2021) explored four types of machine learning models (the locally weighted linear regression learning (LWL) algorithm, linear regression, naive Bayes, and decision trees) to predict crime. It was observed LWL algorithm outperformed others with mean average error (MAE) of 3.35 and correlation coefficient of about 0.75.

South Africa has been identified as one of the world's most deadly, violent, and deadly nations. The rates of social violence and murder, however, have driven South Africa to the top of the crime rankings. According to Business Insider, South Africa is one of the world's 15 most hazardous nations. South Africa was the second most murderous country in 1995. However, following many years of decline, the crime rate has rebounded sharply in recent months. Foreign investors are no longer interested in maintaining or launching a company in South Africa as a result of the country's high levels of social violence and crime, and the economy is in decline.

The government of South Africa was searching for answers to the crime issue in order to enhance the country's image in terms of its high crime rate and to boost investor confidence. In South Africa, the machine learning method to data analysis in crime-related research has gotten little attention. Numerous crime-fighting organisations, notably the police, have huge datasets that may be used to anticipate or analyse criminal behaviour throughout the provinces of South Africa.

As a result, Obagbuwa and Abidoye (2021) set out to solve the problem by developing a linear regression model that could predict crime using South African crime data. Twenty-seven different crime categories were obtained from the popular data repository Kaggle and a machine learning technique was used to extract hidden information from crime data from South Africa's nine provinces. It was expected that appropriate South African authorities and security agencies will gain insight into crime trends and alleviate them in order to encourage foreign stakeholders to continue doing business.

Shukla et al. (2021) identified crime patterns by applying mathematical and statistical models on crime datasets from the state of North Carolina to forecast the likelihood of crime occurrences. They used univariate and bivariate exploratory analysis to find the most prevalent elements that contributed to criminal behaviour. Before testing the performance of the model on the test data using Mean Absolute Error (MAE), Median Squared Error (MSE), and Root Mean Square Error (RMSE), the Akaike Information Criteria (AIC) approach was applied to exclude unnecessary variables (RMSE). R-squared was 0.52 on the training test, while the mean absolute error (MAE) and root mean squared error (RMSE) for the test data were 0.008 and 0.011, respectively. There was a disparity between their assessment methodologies for the training and test datasets, which demonstrated the inconsistency of their model, despite their claims of research accomplishment. They concluded that crime predictability and criminology can be extremely beneficial in eradicating the threat to society.

Academics debate how social factors affect city events. The Ordinary Least Squares (OLS) linear regression model, the Random Forest (RF) regression model, the Artificial

Neural Network (ANN) model, and others have been used to study crime and social variables. These studies' prediction accuracy is poor, and past research has produced several contradicting findings. The non-Gaussian distributions and multicollinearity of urban social data, as well as the inaccuracy and insufficiency of processed data, contribute to these disputes (Wang et al., 2020a).

To address these gaps, they examined the influence of 18 urban variables across six categories on crime risk in China's prefecture-level cities by year. This included geography, the economy, education, housing, urbanisation, and population structure. They employed the big data algorithms Least Absolute Shrinkage and Selection Operator (LASSO) and Extremely randomised Trees to forecast crime risk and evaluate the effect of urban characteristics on crime (Extra-Trees). Their model accurately predicted crime risk with an accuracy of 83%, and the significance of urban indicators was rated. According to their study, the amount of land utilised for residential purposes, the number of mobile phone users, and the employed population are the three most influential elements in China's crime rate. In this age of big data, their study contributes to a better understanding of the impacts of urban indicators on crime in a socialist society and offers solutions for crime prediction and crime rate control with governments.

As public security informatization has progressed, crime prediction has become a crucial instrument for public security agencies to execute precise attacks on criminals and effective governance of the country. Crime prediction method based on feature modelling of combined historical crime data and textual data of case details to forecast the number of suspects. Deep Neural Networks (DNN) and machine learning techniques were examined for extracting features from various dimensions of case data, while Convolutional Neural Networks (CNN) were utilised to extract text features from case descriptions. The two kinds of characteristics were merged and fed into a fully linked layer and a SoftMax layer. The DNN-CNN model was 20% more precise than the DNN model, which used solely numerical data. The incorporation of textual data greatly enhances precision and accuracy (Zhang et al., 2020a).

2.2.5.2 Classification

The ability to accurately predict domestic violence (DV) recidivism may accelerate and improve risk assessment procedures for police and frontline providers. Accurate real-time crime projections contribute to a reduction in the crime rate (Berk et al., 2016). Using an algorithm that anticipates crime hotspots, the crime rate may be reduced. Therefore, Adesola et al. (2020b) analysed crime data given by the Obalende Statistics Department in Lagos, Nigeria, using classification techniques based on decision trees. In contrast to earlier research, their model included accuracy, recall, and F-measure. The efficacy of their prediction model differs across various types of criminal activity. It had an accuracy of around 76% for armed robbery, 77% for abduction, 75% for rape, 76% for serious assault, 72% for murder, and 75% for ritual murder. Also, the consequences of other assessment indicators were not considered.

To improve the performance and reaction of law enforcement authorities, it is necessary to identify the patterns of criminal activity in a certain area. Almaw and Kadam (2019) presented a crime prediction approach based on ensemble learning. They conducted statistical exploratory data analysis, created crime prediction models based on individual models (Naive Bayes, J48, and Random Tree), and then constructed an ensemble classifier based on all the baseline classifiers. It was discovered that random forest outperformed (1-ensemble model, 81.6%) and (3-ensemble model, 79%) respectively. Their exploratory research found that time, month, and season influence the prevalence of crime. According to the experiment, the most crimes occurred between 3:00pm and 6:59pm, and the fewest between 3:00am and 6:59am. July has the greatest monthly rate of criminal activity, while February has the lowest. In comparison to other seasons, summer has the greatest crime rate, while winter has the lowest. As summer, autumn, spring, and winter progress, the frequency of crime decreases.

To enhance the ability to forecast crime trends in order to detect crimes before they occur and allow authorities to take precautionary measures, Hajela et al. (2020) used crime timing, weather, location, and census characteristics such as yearly income and literacy rate to forecast crime. They deployed a spatiotemporal crime prediction method based on machine learning and 2-D hotspot analysis. The two-dimensional hotspot analysis method clusters data using the K-Means algorithm. According to their findings, hotspot analysis may enhance the accuracy of crime prediction compared to a model that simply employs cutting-edge approaches. Each sort of crime is influenced by time, weather, location, and census variables, such as yearly income and literacy rate.

Due to the many characteristics of crime data, it is difficult to extract useful information using standard or statistical data analysis methods. Improving this procedure would expedite law enforcement agencies' efforts to reduce the crime rate. Also, criminals might frequently be detected using crime statistics. Therefore, Khatun et al. (2020) applied decision tree, K-nearest neighbors(KNN), and random forest algorithms on crime data obtained from diverse places, such as Northern Ireland, Chicago, San Francisco, and the open data port. Their forecast was applied to crimes like as robbery, assault, and theft that occur often. According to their findings, decision tree classifiers outperformed others in predicting theft, robbery, car theft, and assault classes, whereas random forest outperformed decision tree classifiers in predicting arrest attribute. Although the performance of their decision tree classifier was superior, they favoured random forest owing to its reliability.

Kshatri et al. (2021) proposed an ensemble-stacking based crime prediction technique (SBCPM) based on SVM algorithms for determining the suitable predictions of crime by integrating learning-based methodologies in MATLAB. They used SVM approach to produce domain-specific setups in contrast to the machine learning models J48, SMO Nave byes bagging, and Random Forest. Their findings suggested that individual models were often ineffective. In some instances, the ensemble model outperforms the individual models with the greatest correlation coefficient and the lowest average and absolute errors. On the testing data, their suggested technique attained a classification accuracy of 99.5%, followed by random forest (97.2% accuracy) and decision tree (94.4% accuracy). However, the output of their model on the training data set was not disclosed so that the consistency of their model could be evaluated. Their model was shown to have a greater predictive impact than the Almaw and Kadam (2019) baselines, which focused primarily on violent crime datasets. In addition, the findings demonstrated that any actual data on crime is consistent with criminological ideas. And indicate that the prediction accuracy of the stacked ensemble model is superior than that of the individual classifier.

The number of complaints received by the Philippine National Police—Anti-Cybercrime Group (PNP-ACG) surged from double digits in 2013 to triple digits in 2017 as a consequence of the proliferation of online scams caused by the growth in online transactions and other internet activities. However, the scarcity of empirical research on cybercrime analytics in the Philippines shows that data mining is not being utilised to help cybercrime investigations, despite the importance of data mining being recognised in prior studies. Consequently, Palad et al. (2019) employed the Weka text mining approach to get insight by categorising a particular online scam dataset using an online scam unstructured dataset including 14,098 words that were mostly Filipino. Classification models were developed using a J48 Decision Tree, Nave Bayes, and Sequential Minimal Optimization (SMO). J48 has the highest accuracy and the lowest error rate among the three classifiers, followed by the Nave Bayes and SMO classifiers. In addition, the responses acquired during validation demonstrate that J48 is preferred over other classifiers since it is simple to learn and use while investigating cybercrime.

Identifying patterns among crime incidents has been the subject of numerous studies, and predicting crime occurrences with high accuracy is a significant challenge in criminology. Currently, the technique of statistical crime hotspot is used, along with empirical knowledge from law enforcement agencies. Patil et al. (2020) investigated the use of an artificial neural network model to forecast crime events in an area. Their model can predict where the majority of crime would occur. Their model is capable of predicting crime in a short period of time and over a small geographic area with an accuracy of 81%.

Using Twitter data and meteorological data, Sandagiri et al. (2020) created a crime prediction system. Using a Bidirectional Encoder Representations from Transformers (BERT) strategy, they accurately identified crime-related tweets with 92.8% precision. With a claimed 98% accuracy, the random forest classifier outperformed previous methods.

Sathyadevan (2014) evaluated the accuracy of categorisation and prediction using several test sets. The classification was performed using naive Bayes, and 90% accuracy was attained. Multiple news items were learned using this method. For testing purposes, test data was incorporated into the model, which now yields more accurate results. Their methodology utilises a location's characteristics, and the Apriori algorithm determines the location's prevalent patterns. On the basis of these regular patterns at each site, a model was developed. There was a need to uncover new elements that led to crime by filtering through crime data, but because they were only evaluating a restricted number of components, complete accuracy could not be attained. Instead of fixing particular features, it is necessary to uncover additional crime-attributes of locations to improve prediction performance. The prediction performance was limited to a few elements or predictors, which may be enhanced with the addition of more factors or predictors.

Safety and security are essential for enhancing the quality of life of inhabitants and ensuring the sustainable growth of cities. Therefore, Wang et al. (2020b) suggested a Deep Temporal Multi-Graph Convolutional Network model that combines graph generation and spatial-temporal components to describe the interdependencies between crime and several external elements. This graph construction would encapsulate the Euclidean and non-Euclidean interactions between areas into several graphs, representing their varied nature. The spatial-temporal component employs a graph convolutional network to recognise spatial patterns and an encoder-decoder temporal convolutional network to simultaneously characterise temporal characteristics. The experiment using a real-world crime dataset from Chicago demonstrated that the suggested DT-MGCN model is effective. This model is believed to be superior than the existing baseline for the state of the art.

In addition to being one of the most prevalent crimes, arson is also characterised by its low cost and significant damage. In addition to inflicting fatalities and physical damage, arson often has enormous societal repercussions and induces psychological terror among the general populace. Given that arson committed by a gang is more destructive, identifying gang involvement in arson cases has become an essential problem. Wang et al. (2021) presented a hybrid approach that combines ensemble learning and intelligent optimization methods to address this issue. They created the recursive feature elimination (RFE)-based feature selection approach in order to eliminate duplicate features. The performance of individual models was discovered to vary among feature selection procedures. In terms of accuracy, XGBoost surpassed others (83.7%), and in terms of recall (41.1%), it was a decision tree. XGBoost fared the best in terms of precision, followed by logistic regression and decision tree. Regarding f1-score, the decision tree outperformed others. After experimenting with numerous combinations of basic classifiers, they discovered the ideal combination. Lastly, they used the differential evolution (DE) technique to improve the parameters of the base classifier and the weight of the combination, so enhancing the model's prediction capability. Their proposed DE-ensemble model achieved an accuracy of 83.1%, recall of 75.5%, precision of 56.9%, and a f1-score of 55.5%. The proposed model was tested using the database of the United States National Fire Incident Reporting System (NFIRS). According to their findings, the recommended approach outperformed other current machine learning techniques.

The basic objective of Public Security Prevention and Control is prediction. By encoding the area-specific crime episodes, Zhang et al. (2016) categorised crime hot spots into various heat levels, therefore transforming the prediction of Hot Spots into a multiclass classification issue. As with rotational invariance, histogram-based statistical approaches were used to construct neighbourhood heat level characteristics. Ultimately, LDA (Linear Discriminant Analysis) was used for dimensionality reduction of mixed spatial-temporal features, and KNN was utilised for prediction. When crime data are collected on a "Weekly" basis, the new prediction model may perform optimally, according to their findings.

2.2.6 Classification and other Methods

With the advent of new tools for collecting and combining fine-grained crime-related information, it is now possible to advance crime prediction and get a better knowledge of crime dynamics. Due to the unequal distribution of data, however, it is difficult for a city to create a unified framework for all boroughs. Zhao and Tang (2017a) analysed spatial-temporal trends in urban data from one city borough before using transfer learning methods to enhance crime prediction in other boroughs. They validated the presence of spatial-temporal patterns in urban crime and retrieved crime-related characteristics from cross-domain data sets. Lastly, a framework for transfer learning was proposed to integrate these characteristics and construct spatial-temporal patterns for crime prediction.

Gang-related murders constitute a significant proportion of worldwide criminal activity, particularly in Latin American countries. As opposed to other types of homicide, they are often the result of territorial conflicts and are distinguished by area-specific risk factors. Due to underlying links between gangs and geographic locations, area-specific crime patterns, and a lack of spatially fine-grained predictive signals, gang-related murders have been largely neglected in current crime modelling and prediction research. Akhter et al. (2018) presented a novel context-aware multi-task multi-level learning framework for the acquisition of area-specific crime prediction models and potential gang operation territories. Their model learned more task-specific information at a finer level, and multi-task learning was used to analyse the large amount of knowledge from coarse-grained tasks. Support Vector Machine (SVM), Logistic Regression, regularised LASSO, and the baseline technique to monotonic multi-task (MMT) were applied to analyse 10,672 newspaper articles gathered between April 2015 and May 2016 from various news agencies, including El Colombiano, El Universal, RCN Radio, El Tiempo, El Confidencial, NTN24, and El Nuevo Herald. Their suggested model, MLMT, outperformed other techniques on average. In terms of violent intensity, MLMT beat the baseline technique by 5% to 10% in terms of accuracy, recall, and F1-score. MLMT outperforms the baseline in terms of murder count by 10.3% to 20.8% in accuracy, recall, F1-score, and AUC.

Due to human behaviour dependent nature of crime, it has been seen as random for millennia. Even today, it requires too many variables for earlier machine learning algorithms to reliably predict. Nitta et al. (2019) developed a forecasting model to anticipate future occurrences of crime and the sort of crime that may occur in a particular location. Latitude and longitude, as well as other geologically important characteristics, were explored. The LASSO approach was applied to different machine learning classifiers (Naive Bayes, SVM, KDE, and DNN) for prediction after feature selection. According to their results, multiclass naive Bayes models surpassed other models with an accuracy of 97.47 percent.

Yang et al. (2018a) presented a platform that predicts and visualises crime hotspots in New York City utilising many data sources and real-world data. Their algorithm continuously collects crime data as well as urban and social media data from the Internet, picks relevant aspects from the obtained data using statistical and linguistic analysis, and identifies crime hotspots by using the extracted features. The hotspots are then shown on an interactive map.

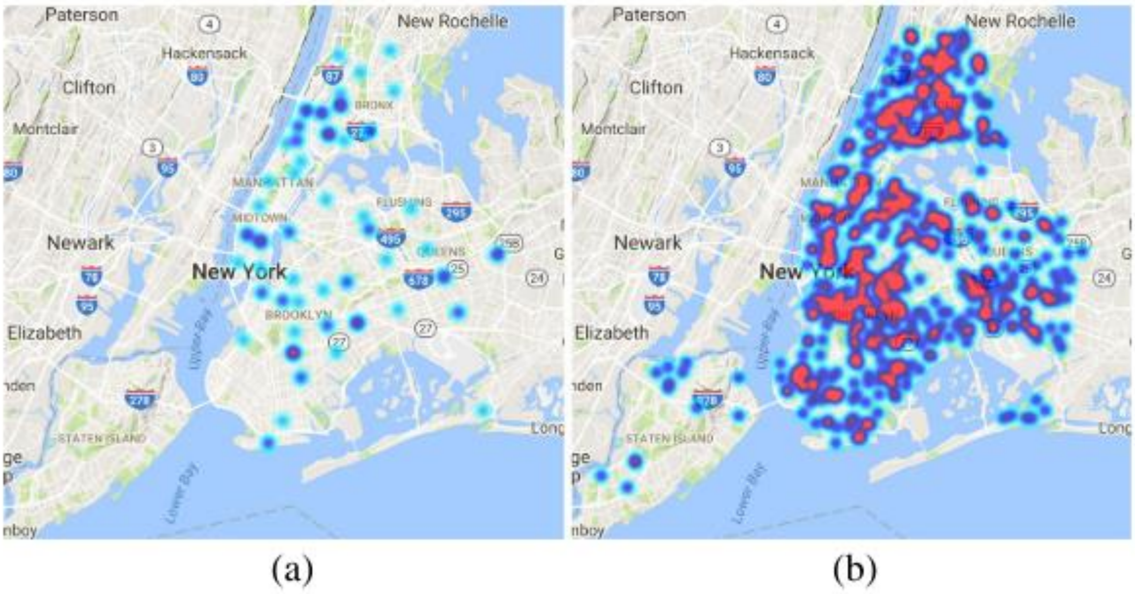


Figure 2-1: Crime hotspot map for two crime types: a Rape; b Grand Larceny of Motor Vehicle. (Source: (Yang et al., 2018a))

Yu et al. (2020) assessed the accuracy of prediction model studied the role of probable offenders' actions in crime prediction by including this information into their predictions and assessing the accuracy of their models. In order to anticipate the movements of future criminals, they examined the data obtained during ordinary police stop-and-question operations on the movements of prior offenders. In a Spatio-Temporal Cokriging model for crime prediction, it is argued that offender mobility data compensates for past crime data. Their models were used in the XT police district in ZG city, China, for weekly, biweekly, and quadruple weekly forecasts. Their outcomes are much better with the offender mobility data than without it. The weekly model improves the most, followed by the biweekly and quarterly models.

2.2.7 Sentiment Analysis and its Roles

Sentiment analysis is an extremely active field of research in natural language processing, which allows extracting the opinions from a set of documents (Behdenna et al., 2018). Sentiment analysis uses the natural language processing (NLP), text analysis and computational techniques to automate the extraction or classification of sentiment from sentiment reviews (Hussein, 2018). Analysis of these sentiments and opinions has spread across many fields such as Consumer information(Iqbal et al., 2022, Luo et al., 2022), COVID-19 cases and its vaccines (Alamoodi et al., 2022, Aljedaani et al., 2022), application on foreign languages such as Turkish and Spanish (Altinel et al., 2022, Arias et al., 2022), comments (Lin, 2022), and Social (Jain et al., 2022a, Jain et al., 2022b). Sentiment analysis becomes a hot area in decision making. Hundreds of thousands of users depend on online sentiment reviews. Over 50% of world population depended on social media (Iqbal et al., 2022, Luo et al., 2022). The main goal of analysing sentiment is to analyse the reviews and examine the scores of sentiments.

2.2.7.1 Different levels of sentiment analysis

In general, sentiment analysis has been investigated mainly at three levels (Behdenna et al., 2018, Devika et al., 2016).

•Document level analysis: At this stage, the aim is to determine the document's general viewpoint. At the document level, sentiment analysis implies that each document represents thoughts about a single entity. Example is in the clinical de-identification(Catelli et al., 2021),

•Sentence level analysis: At this level, the primary responsibility is to determine if each statement represented a good, negative, or neutral attitude. This level of analysis is closely connected to subjectivity classification, which divides objective sentences expressing factual information from subjective phrases expressing subjective thoughts and opinions. Document level and sentence level studies do not reveal what individuals liked and disliked. It is applied in both customer sentence review and medical settings (Denecke and Deng, 2015, Zhou et al., 2015).

•Aspect level analysis: This level does finer-grained analysis and necessitates the use of natural language processing. At this level, opinion is distinguished by polarity and a target of opinion. This can be applied to a product review or in the financial sector (Mai and Le, 2021, Yang et al., 2018c).

2.2.7.2 Methods of Sentiment Analysis

There are many methods to carry out sentiment analysis. Still many researches are going on to find out better alternatives due to its importance. Some of the methods are discussed below:

2.2.7.2.1 Machine learning Approach

Machine learning techniques function by first training an algorithm on a training data set, then applying it to the real data set. Machine learning approaches initially train the algorithm with specific inputs and known outputs, allowing it to operate with fresh new data later on(Devika et al., 2016).

2.2.7.2.2 Rule Based Approach

The rule-based method entails setting several criteria for collecting views, tokenizing each phrase in each document, and then confirming the existence of each token, or word. If the term is present and has a good connotation, a +1 rating was assigned. Each post begins with a score of zero and is regarded as positive. If the final polarity score was larger than zero, or if the overall score was less than zero, the value would be negative. The output of a rule-based method will be validated or questioned to see if it is accurate. If the input sentence includes a term not already in the database that might aid in the analysis of a movie review, such a word must be added to the database. This is supervised learning in which the system is taught to learn whenever it receives fresh information(Devika et al., 2016).

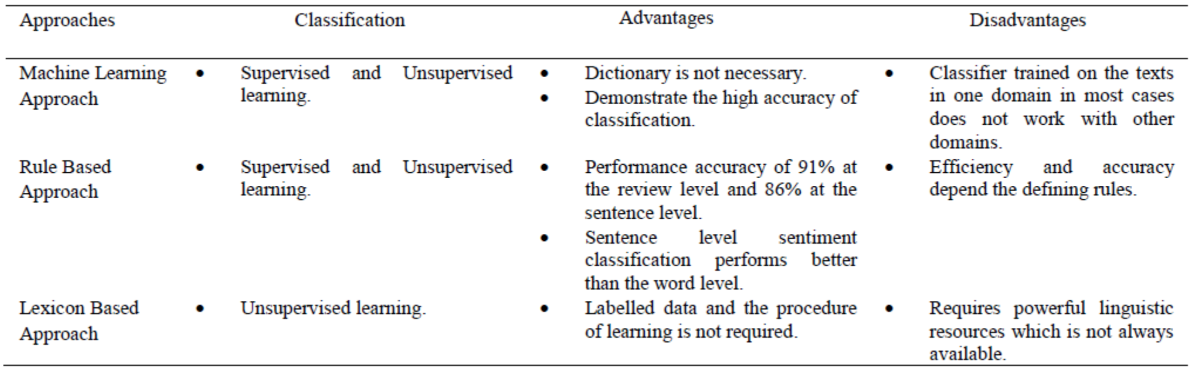
2.2.7.2.3 Lexical Based Approach

This approach is based on the premise that the overall polarity of a phrase or document is equal to the sum of the polarities of the constituent phrases or words. This technique is based on emotional studies for each domain's sentiment analysis dictionary. Next, each domain dictionary was supplemented with the highest-weighted evaluation words from the most relevant training collection, as determined by the relevance frequency approach. The word modifier adds or subtracts a specified proportion of weight to the next evaluation word. Word-negation modifies the weight of the subsequent evaluation word by a particular amount: for positive words, the weight decreases, and for negative words, the weight increases. The process for classifying the sentiment of a text was as follows. The first weights of all training texts are determined for the categorised text. Each sentence is positioned in a single-dimensional emotional space. The cross-validation approach was used to calculate the fraction of deletions. The average weights of training texts were then determined for each sentiment class. The categorised text was referred to the class that was closest to the emotional centre of the one-dimensional space (Devika et al., 2016).

2.2.7.3 Contrast and Integration

Table shows the three primary sentiment analysis methods compared and integrated. Sentiment analysis using different methods yields diverse findings. Each method has merits and downsides. Machine learning offers the highest performance, efficiency, and accuracy, and much of the work has been done in this technique.

Table 1: Comparison of Three Sentiment Analysis Methodologies (source: (Devika et al., 2016))



Approaches	Classification	Advantages	Disadvantages
Machine Learning Approach	• Supervised and Unsupervised learning.	• Dictionary is not necessary. • Demonstrate the high accuracy of classification.	• Classifier trained on the texts in one domain in most cases does not work with other domains.
Rule Based Approach	• Supervised and Unsupervised learning.	• Performance accuracy of 91% at the review level and 86% at the sentence level. • Sentence level sentiment classification performs better than the word level.	• Efficiency and accuracy depend the defining rules.
Lexicon Based Approach	• Unsupervised learning.	• Labelled data and the procedure of learning is not required.	• Requires powerful linguistic resources which is not always available.

This work employed a hybrid technique by merging the lexical approach with machine learning approaches. This would be accomplished by extracting the emotion polarity from the information collected on Twitter using the lexical technique, merging the extracted information with existing historical crime data, and constructing a crime prediction model.

2.2.7.4 Application of Sentiment Analysis

Using tweets from UK-based remote care professionals, we want to discover how COVID-19 impacted UK-based health and care discussions. Analysis of Twitter data may reveal users' opinions on COVID-19's use of telemedicine. User perspectives may evolve in response to fast improvements in care delivery, and studies of long-term remote care dialogues can be conducted using this strategy. Considering the importance of remote visits to health care delivery, it is important to continue studying patients' preferences despite a decline in public support for remote care since the onset of the epidemic. As stated by (Ainley et al., 2021). Further, many additional pandemics posed a threat to civilization around the end of 2019, when the globe was suffering a worldwide health crisis owing to the COVID-19 virus. Participation in social media is crucial in these situations because it helps health systems respond to emergencies by revealing public concerns, revealing signs of infection, and tracking the spread of the virus. Topic modelling and sentiment analysis were used to the Twitter data that was gathered via keyword searching. Most discussions focused on malaria, influenza, and TB, with the illnesses receiving centre stage. The most common tweet topics were HIV/AIDS, which may explain the prevalence of the feelings of fear, trust, and hate(Qin and Ronchieri, 2022).

Differences between the three waves of confinement in 2020 and 2021 were also identified, and hitherto unsuspected patterns were revealed, as a means of analysing the results of lockdown processes in Malaysia. As seen by the proliferation of trending topics, the initial lockdown captured the attention of the public. There were more compliments than complaints during each stoppage. Topic modelling revealed that during the first lockdown, people spoke mostly about staying home, quarantine, and lockdown, whereas during the second and third lockdowns, they talked more about the need for health measures and government actions. This enables governments to ascertain public sentiment and to address citizen concerns with preventive measures targeted at the most pressing problems. It was also stressed how important it is to provide words of encouragement during times of crisis, create digital solutions, and pass new laws to better equip people to deal with disasters. As stated by (Alamoodi et al., 2022).

As an added bonus, we can now automatically analyse tweets from Arabic users to do sentiment analysis on COVID-19 vaccinations. It was chosen to conduct this poll over a longer time period in order to more truly portray the shift in public attitude on vaccines. I learned about the most common immunisations in the Arab world and the factors that make some people reluctant to be vaccinated. Furthermore, it was found that 38% of tweets were negative, compared to just 12% that were positive. It is remarkable that most tweeters show either indifference to or hostility against vaccination, showing they are not persuaded of the procedure's need. This research provides a summary of the most often voiced Twitter concerns and advises incorporating them into vaccine marketing initiatives. Several researchers (Aljedaani et al., 2022).

The development of ML-based tools for analysing Arabic tweets and identifying signs of sadness in that community. Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), AdaBoost, and Nave Bayes (NB) are only some of the machine learning classifiers and natural language processing techniques that went into building the model to determine user sentiment. With an accuracy of 82.39 percent, the RF classifier was shown to be superior than its rivals. This was shown to be the case (Musleh et al., 2022).

Many fields make use of sentiment analysis; one such area is the examination of online opinions regarding Saudi Arabian cruises (Al sari et al., 2022). This work is the first to examine the quality of these sentiment analyses (SAs) of first-hand accounts of Saudi cruises. The results of this research will aid in comprehending the perspectives of passengers and viewers by analysing their reactions expressed in social media postings on their first cruises to Saudi Arabia. Analyzing the number of tweets, favourite tweets, and retweets each day and by political party is only one way that this method of studying political communication has been investigated. This was shown to be the case by Costa et al. (2021).

For the leisure service business to successfully cater to its clientele, raise the bar on product quality, and strengthen its customer relationships, sentiment analysis of leisure consumption is crucial (Luo et al., 2022). Moreover, it has been put to use in a circumstance when it was crucial for political leaders and strategic decision-makers to grasp people's feelings about the crisis from a geographical perspective (Sufi and Khalil, 2022). Furthermore, it has been used to gauge public opinion of audiovisual media (Wang et al., 2022).

2.2.8 Social Media

Social media postings provide real-time data that has been used to predict social trends in the past. Prevention is superior than treatment. Preventing a crime is better than investigating why or how it occurred. In the same way that a vaccine is given to a child to prevent illness, it is crucial in today's society, with its high crime rate and horrendous crime incidents, to have a vaccination system that prevents crimes from happening. Vaccinating society against crime refers to a number of deterrent approaches, such as education, awareness-raising, boosting efficiency and proactive enforcement operations, and others.

Security is the most critical element that has been prioritised in the current climate. Despite the rapid growth and usage of digital devices, effective crime assessment in developing countries remains challenging. Crime monitoring technologies are essential for public health and law enforcement authorities to allocate appropriate resources and develop targeted responses. Twitter, for instance, has been shown to be an effective tool for monitoring and predicting public health events such as disease outbreaks. Social media may be a feasible method for conducting illegal surveillance (Wang et al., 2019).

Featherstone (2013) claimed that current technology can make communication more effective in terms of data collection, prediction, and recognising larger trends. An evaluation was conducted to discover if South African tweets were evaluated to determine what information was being shared among individuals and whether it may be a beneficial source of information for crime prevention. Also to provide direction for future research into automatic crime prediction, Wang and Gerber (2015) investigated the prediction of social media users' spatial trends by incorporating textual content into existing next-place prediction models, and they achieved a significant improvement in next-place prediction compared to several previously published research baselines. They also investigated if there was a correlation between their forecasts and crime in a major US metropolis.

In our culture, crime is a regular occurrence. Particularly of interest to the scientific community is the manner in which individuals communicate their worry about crimes. The quantity and diversity of communication channels have increased throughout time. Today, social media platforms like Twitter make it simple to voice one's ideas and worries about crime. In a quest to examine how the kind and proximity of a murder impacted public attention. Kounadi et al. (2015) examined tweets from 2012 that referenced killings that happened in London.

Particularly, the reliance of tweeting propensity on the closeness, in geography and time, of a criminal incidence and the number of persons who were worried about the incident was investigated. In addition, the criminal features of homicides were analysed by logistic regression. The findings indicate that the closeness of the estimated home locations of Twitter users to the sites of homicides influenced whether and how fast connected crime news gets disseminated. More than half of the tweets linked to homicides were made within the first week, and the vast majority are sent during the first month. Certain criminal features, such as the presence of a weapon, a young victim, a British victim, or a gang-committed murder, were predictors of the posting frequency of crime-related tweets (Kounadi et al., 2015).

Inspiring by two current techniques to crime prediction, Azeez and Aravindhar (2015a) proposed a visual analytics approach that offers decision-makers with a proactive and predictive environment to aid them in making successful resource allocation and deployment choices. Malik et al. (2014) stated that majority of crime prediction was based on past crime data in addition to other geographical and demographic data without taking into account the rich and fast developing social and digital media background surrounding instances of interest, despite their promise. Using latent Dirichlet allocation, topic recognition, and sentiment analysis, Malik et al. (2014) developed an approach based on semantic analysis and natural language processing of Twitter tweets. Both approaches, however, have inherent limits. Today's crimes are characterised by the incidence of recurrent crimes, crimes that occur as a consequence of another action, and crimes whose existence is anticipated by other information.

In support of a research study proposal (Corso, 2015) examined a possible correlation between social media, incident-based crime data, and public domain data. Creating a predictive crime-based artefact that integrates data mining, natural language processing, and graphical information system design was challenging. An effort was made for social media to observe the data flexibility, process control, and prediction capabilities of such an artefact. Observations were made of the data and their capabilities during the preparation of noisy social media data, government-based structured data, and obscurely obtained field data for use in a predictive GIS artefact. The approach for artefact architecture, data collection, and discussion of results were bundled into an exploratory study to meet the project's objectives. The findings demonstrate a high correlation between social media data and domain-specific datasets.

When statistical analysis is applied to unstructured data, social networking platforms offer the latent potential to uncover significant insights. A previous study has shown that GPS-tagged Twitter data permits the prediction of future crimes in major cities in the United States. However, current crime prediction models that use Twitter data are limited in their ability to describe criminal episodes owing to the lack of sentiment polarity and meteorological variables. The inclusion of sentiment analysis and weather forecasts into these models would provide substantial insight into the commission of crime (Chen et al., 2015b). Therefore, they attempted to predict the time and place when a certain sort of crime would occur. Their solution was based on sentiment analysis using lexicon-based methodologies and a comprehension of categorised meteorological data, as well as kernel density estimates based on past crime incidences and prediction using linear modelling. When it came to predicting future crime in each part of the city, their model was better than the benchmark model, which only used kernel density estimation to predict crime.

Almehmadi et al. (2017) used Twitter public data to forecast crime rates. The crime rate has risen in recent years. Although crime stoppers use a variety of methods to minimise crime rates, none of the prior approaches have focused on using the language used (offensive vs. non-offensive) in Tweets as a source of data to anticipate crime rates. Thus, they hypothesised that analysing the vocabulary used in tweets was a reliable method for predicting city-level crime rates. New York city-related tweets were gathered over three months. In addition, crime-related tweets from the two cities were collected to validate the accuracy of the prediction system. For crime prediction, a Support Vector Machine (SVM) classifier was used to analyse the gathered tweets. They discovered a correlation between tweets and crime rates in certain localities. Their data validated the hypothesis that the vocabulary used by Twitter users in diverse locations with varying crime rates varies. Label assignment acted as a bottleneck, revealing that text mining was not totally automated. The studies in this paper largely rely on a small number of Twitter fields. This kind of comprehensive data may miss some surprising trends. Putting the strategies in this research to better use would be to find out if crime rates change over time.

Williams et al. (2017) analysed the advantages and disadvantages of social media data for the research of crime and disorder. They claimed that disorder-related tweets correlated with genuine police crime statistics. Their findings demonstrated that naturally occurring social media data might be exploited as an additional source of knowledge on the crime issue.

Prior statistical crime prediction research has not analysed the micro-level mobility patterns of people in the studied region. Geotagged social media inherently indicates these patterns for many people; however, strategies for collecting such trends and incorporating them into a prediction strategy have yet to be researched. Al Boni and Gerber (2017c) investigated the use of spatiotemporally labelled Twitter tweets to infer micro-level movement patterns, and they created and evaluated a model informed by such patterns using real-world crime data. In comparison to a model with a baseline that does not incorporate micro-level movement patterns, their findings showed an improvement in performance for 15 of the 20 categories of crimes analysed.

Additionally, Cesur et al. (2017) created a unique approach for detecting possible criminals by analysing the content of Tweets in order to reveal criminals and make it simpler for security agencies to identify offenders. In their investigation, both deep learning and big data analytics were used. They used the multilinear perceptron (MLP) classification algorithm for deep learning on a 384-word dataset, taking into account the date, time, and location of transmitted Twitter shares as characteristics. It was stated that their algorithm effectively detected offenders around 71.61 percent of the time. In addition, they devised a method for identifying prospective criminals who utilise the most prominent social networking sites in the globe to perpetrate organised crime or cybercrime.

Vomfell et al. (2018b) evaluated the descriptive and predictive power of human activity patterns derived from Uber, Twitter, and Foursquare data. An evaluation of six months of crime data in New York City found that incorporating these data sources increases the forecast accuracy of property crimes by 19% compared to crime data in New York City found that incorporating these data sources increases the forecast accuracy of property crimes by 19% compared to simply using demographic data. When unique characteristics were integrated, this impact was enhanced, resulting in new insights into crime prediction.

Notably, and in line with the idea that society is not well organized, the unique traits did not help predict violent crimes.

Aghababaei and Makrehchi (2018) investigated if a social media setting may give socio-behavioural "signals" for a crime prediction issue. Although conventional crime prediction approaches depend on historical crime data and geographical information for the place of interest, they argued that Twitter's crowd sourced public data may include predictive characteristics that might signal changes in crime rates without requiring access to past crime statistics from particular places. They built a prediction model for crime trend forecasting, with the goal of using Twitter content to forecast crime rate directions in the future. Their approach made use of content, mood, and subjects as predictive markers to infer changes in crime indices. Since their work is sequential, the temporal topic identification approach was used to infer predicted themes across time. Instead of examining a vocabulary in bulk, their suggested topic identification algorithm constructs a dynamic vocabulary to identify new subjects. Using previous tweets, their approach was used to obtain data from Chicago to forecast crime trends. According to their findings, there exists a link between content-based characteristics retrieved from the material and crime patterns. Their work demonstrated the viability of a temporal topic detection model in finding the most predictive characteristics across time, as opposed to a static model without time consideration. In addition, the research highlighted the contribution of socioeconomic indices and temporal characteristics as auxiliary characteristics. In their research, they also show that content-based features are much better at making accurate predictions than additional features.

Curiel et al. (2020) gathered a huge number of tweets from the 18 most populous Spanish-speaking nations in Latin America over the course of 70 days in 2020. The tweets were categorised as crime-related or not, and further information was retrieved, such as the kind of incident and, if feasible, the city-level geolocation. It has been established that around 15 out of every 1000 tweets include language about crime or fear of crime. The frequency of crime-related tweets is then compared to the number of murders, the murder rate, and the degree of fear of crime as indicated by polls. According to their study, social media is not an effective tool to learn about crime patterns.

Due to the spread of COVID-19, an increasing number of individuals are using internet services. A rise in social media use has also been seen, leading to the belief that cyberbullying has also grown. Das et al. (2020) evaluated the likelihood of an increase in cyberbullying incidences as a result of the pandemic and high social media use in their research. To examine this trend, they gathered 454,046 public tweets on cyberbullying sent between January 1, 2020 and June 7, 2020. They analysed the presence of at least one statistically significant changepoint for the majority of these terms, the majority of which were centred near the end of March. Almost all of these changepoint locations and times were associated with COVID-19, supporting our initial hypothesis that cyberbullying is worsening based on Twitter chats.

Anyone interested in avoiding criminal behaviour or determining where crime enforcement is required would benefit from the ability to detect and anticipate crime patterns or follow criminal movement, especially for crimes where continual police is unfeasible, such as cable theft. Numerous South African neighbourhoods have developed volunteer community police units that communicate through SMS and two-way radios. Some have chosen websites and even Twitter as a method of communicating more rapidly and openly. The Crime Line (@CrimeLineZA) and the South African Police Service (@SAPoliceService) are two prominent Twitter users.

Although domain knowledge shows that hotels are hubs for criminal activity, hotel customers may be trustworthy. Kostakos et al. (2019) employed geographical clustering and sentiment feedback to map real crime occurrences into London hotel reviews to examine the viability of utilising consumer hotel evaluations to anticipate crime. Their early data reveal a negative correlation between hotel review sentiment and crime rate. Crime hotspots are more likely to be located near hotels with positive reviews than vice versa. A potential reason for such a puzzling finding is that the review data is not matched to particular crime categories, therefore the crime data mostly reflects police presence on the site. To demonstrate that hotel evaluations may be utilised to forecast crime, further research and domain expertise are required.

Today, terrorist networks and organised crime constitute an increasing threat to civilization. Emerging as a result of the increasing use of IT technology is the use of social media sites such as Twitter, Facebook, and YouTube to promote and support criminal activities, employ terrorists, and build alliances. In cyberspace, conventional approaches and mitigating tactics against cyberterrorism are ineffective. Tundis et al. (2019) developed software that automates the process of extraction and extension of social network analysis methods combined with well-known clustering techniques and association rules, to identify users with similar characteristics and ties to specific criminal acts, as well as the process of group leaders and their mediators. Their findings highlight the social network operations of both terrorist organisations.

Using natural language processing, separate research in Pakistan examined eight years of crime-related news archives to forecast the behaviour of criminal networks and derive usable information. Umair et al. (2020) conducted a hotspot-based geographical analysis and using two distinct machine learning models (KNN and Random Forest) to forecast criminal activity. KNN had the highest accuracy at 92%. Wang et al. (2019) analysed the association between drug-related tweets and crime statistics using Twitter data from May to December 2012 and crime statistics from 2012 to 2013. Their analysis revealed a high link between tweets from 2012 and county-level crime statistics from 2012 and 2013 for both 2012 and 2013. Their results indicate that social media data may be utilised to forecast future crimes.

Data mining applications have been used in the banking business for customer categorization and efficiency, credit scores and authorization, payment default prediction, publicity, and fraud detection. Lekha and Prakasam (2018) presented a comprehensive model of data mining methods and cybercrime kinds in banking apps. In addition, they present a comprehensive study of efficient and effective data mining strategies for cybercrime data analysis. They utilised novel data mining techniques such as K-means, influenced association classifier, and J48 prediction tree to investigate cybercrime data sets and classify the solvable issues. In influenced association classification for unsupervised learning clusters, the K-Means method is used. The initial centroids are selected using K-means, which enables the classifier to mine the data and build cybercrime predictions using a decision tree model. The combined knowledge of K-Means, Influenced Association Classifier, and J48 Prediction Tree will give an enhanced, integrated, and accurate cybercrime prediction model for the banking industry. Our law enforcement authorities must be well armed to fight and prevent cybercrime.

2.2.9 Metrics Evaluation

Many physical and sociological processes are represented as discrete events in time and space. These spatio-temporal point processes are often sparse, meaning that they cannot be aggregated and treated with conventional regression models. Models based on the point process framework may be employed instead for prediction purposes. Evaluating the predictive performance of these models posed a unique challenge, as the same sparseness prevents the use of popular measures such as the root mean squared error. Statistical likelihood was a valid alternative, but it did not measure absolute performance and was therefore difficult for practitioners and researchers to interpret. Adepeju et al. (2016) proposed a practical toolkit of evaluation metrics for spatio-temporal point process predictions. Their metrics were based around the concept of hotspots, which represent areas of high point density. In addition to measuring predictive accuracy, their evaluation toolkit considered broader aspects of predictive performance, including a characteristic of the spatial and temporal distributions of predicted hotspots and a comparison of the complementarity of different prediction methods. They showed that the application of these evaluation metrics using a case study of crime prediction, comparing four varied prediction methods using crime data from two different locations and multiple crime types. The relationship between predictive accuracy and spatio-temporal dispersion of predicted hotspots were revealed from their result.

Although, there were numerous hot spot mapping techniques that can be used in research and in practice for predicting future crime locations. Due to differences in the varying techniques, metrics were developed to compare the accuracy and precision of these techniques. (Drawve, 2016) explore predictive accuracy index (PAI) and recapture rate index (RRI) to evaluate six different hot spot techniques. Spatial and Temporal Analysis of Crime, Nearest Neighbor Hierarchical, Kernel Density Estimation (KDE), and Risk Terrain Modeling (RTM). As the results indicated, KDE was the most accurate while RTM was the most precise. Dugato (2013) showed similar results when metrically comparing RTM and KDE on a one-year prediction time period. Additionally, when examining just the accuracy, the current findings, with the inclusion of RTM, supported Chainey et al. (2008) results which found KDE to be the most accurate technique. The precision of RTM in the current study could be attributed to the integration of the built environment to expect where crime will occur, allowing for more consistent results.

Also, Rummens and Hardyns (2021)assessed the influence of changing grid resolution, temporal resolution, and historical time frame on prediction performance. To investigate, they analysed home burglary data from a large city in Belgium and predicted new crime events using a range of parameter values, comparing the resulting prediction performances. Given the potential prediction performance costs associated with prediction at a high spatio-temporal resolution, consideration should be given to balancing practical requirements with performance requirements.

2.2.10 Model comparison

The creation of police tactics and the execution of crime prevention and control rely heavily on crime forecasting. Machine learning is the prevalent approach to prediction right now. Nevertheless, little research has rigorously examined various machine learning techniques for crime prediction. Zhang et al. (2020b) gathered public property crime data from 2015 to 2018 from a region of a big coastal city in the southeast of China as research data to evaluate the forecasting potential of multiple machine learning algorithms. We demonstrated that the LSTM model outperformed KNN, random forest, support vector machine, naive Bayes, and convolutional neural networks using just historical crime data. When the built environment data of points of interest (POIs) and urban road network density are put into the LSTM model as covariates, they found that the model with built-environment covariates has a more accurate prediction effect than the original model, which is based only on historical crime data.

In an effort to evaluate and compare the prediction performance of three of the most prevalent predictive policing methods, Drawve et al. (2016) measured the possible differences in the accuracy and precision of two methodological mapping techniques as predictors of future gun crimes in Little Rock, AR: risk terrain modelling and nearest neighbour hierarchical (NNH), a traditional hot spot technique that uses past crime to predict where future crime is most likely to occur. Using data from the Little Rock Police Department, the Little Rock Treasury Department, and the 2000 census, the NNH hot spot and RTM gun crime prediction algorithms were evaluated. The RTM included measurements of crime generators and crime attractors, while the NNH hot spots were derived from 2008, gun crime data. Using their prediction accuracy index (PAI) and recapture rate index (RRI) values, the two measurements were compared. Six of the seven social and physical environmental factors in the RTM accurately identified future gun crime locations, whereas the NNH hot spots accurately predicted 7% of future gun crime. PAI and RRI values indicated that the RTM method was more precise than the NNH hot spot technique, but the NNH hot spot technique was more accurate than the RTM approach. Relying on a single spatial prediction method may pose accuracy and reliability issues. There may be a need for many methods to thoroughly examine the phenomena. RTM's accuracy may be inferior than that of other approaches. Due to the incorporation of the environmental background, RTM is more trustworthy than NNH hot spots.

The finding of Drawve et al. (2016) corroborated by the research of Rummens and Hardyns (2020). Using retrospective analysis of house burglary crime data from a Belgian city, they analysed the performance of three models: a near-repeat model, a supervised machine learning model, and a risk terrain model. Inclusion of hotspot analysis as a baseline. To account for seasonal changes, predictions are given for three separate months (January, May, and September 2017). Also investigated were the variations in geographical context (city core vs suburbs) and the number of anticipated risk areas. Accuracy, near-hit rate, precision, and F1-score are used to evaluate the performance of a prediction model. Their findings indicated that there were significant variances in the predictive performance of the model types among the investigated variants. In general, the ensemble model is the most consistent performer among all versions studied. Notable is also the fact that hotspot analysis is not demonstrably surpassed by other approaches. Each technique has its own set of benefits and drawbacks, and the ideal prediction performance is largely reliant on the unique geographical context. To provide a full picture, more comparative assessments of predictive policing systems in diverse scenarios are required. Future studies should also examine how combining multiple methodologies can improve crime prediction, they suggested.

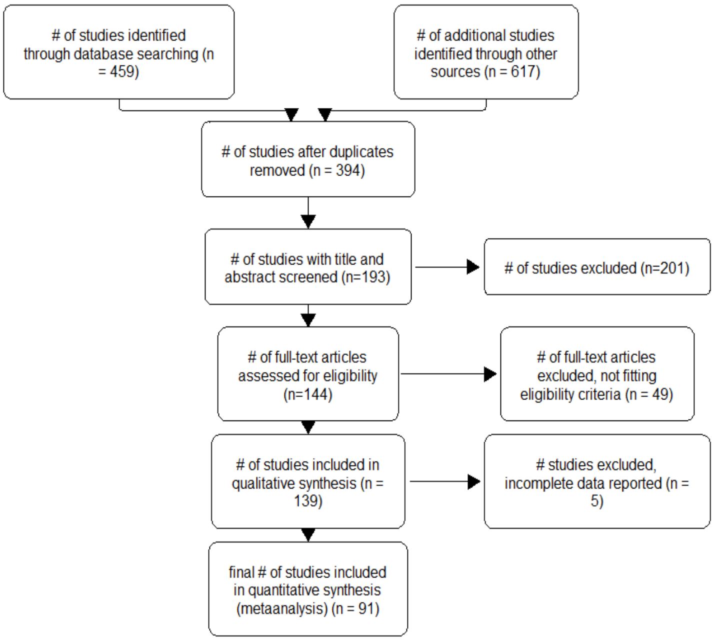
2.3 Systematic Literature Review

Based on the review of previous studies, this section performed a systematic review of only those studies that seemed to apply data mining techniques to predict crime.

2.3.1 Methods

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines (Page et al., 2021) guided the selection of this research. PRISMA recommends a checklist of 27 components for a systematic literature review's sections and content, as well as a four-phase flowchart for article selection. In order to meet the requirements of this research project, the PRISMA guideline was modified. The flowchart for paper selection consists of three stages: identification, screening, and eligibility.

- Identification: This stage comprises identifying information sources and a search strategy that yields a list of possible publications. The number of articles selected during the identification phase is based on the keywords' crime AND prediction, AND social AND media, AND data AND mining'. Screening means choosing articles from the first phase and getting rid of the ones that don't fit with the study's goal.
- Eligibility: The third phase comprises a more in-depth examination of publications and the selection of those that are relevant to the research themes., displays the number of papers at each level and their subsequent progression.



# of studies identified through database searching (n = 459)

# of additional studies identified through other sources (n = 617)

# of studies after duplicates removed (n = 394)

# of studies with title and abstract screened (n=193)

# of studies excluded (n=201)

# of full-text articles assessed for eligibility (n=144)

# of full-text articles excluded, not fitting eligibility criteria (n = 49)

# of studies included in qualitative synthesis (n = 139)

# studies excluded, incomplete data reported (n = 5)

final # of studies included in quantitative synthesis (metaanalysis) (n = 91)

Figure 2-2: The three phases of the study selection process: identification, screening, and eligibility

Selection: Selecting information sources for performing literature searches is the next step. Although there are several search engines and academic databases, the focus of the research was on scholarly and research databases, including areas where data mining methods for crime prediction and social media usage may be gathered.

This research studied an optimal search process that integrates several academic search databases, with searches done at the greatest attainable level of specificity. According to Bramer et al. (2017), if the research topic is more interdisciplinary, a bigger scientific database like Web of Science is likely to be beneficial. Nevertheless, according to this research, Scopus is a bigger database than the Web of Science. Scholarly opinion is divided regarding the usefulness of Google Scholar for multidisciplinary research. According to Sirotkin (2013), some search engines, particularly Google Scholar, have a predisposition to selectively reveal information by using algorithms that personalise content for users.

This phenomenon is known as the filter bubble effect. Haddaway et al. (2015) revealed that while looking for certain papers in Web of Science, the majority of the results were also found in GS. When similar search terms were used in Web of Science and GS (10-67 percent overlap), GS missed some crucial information in five out of six case studies, according to their results. Consequently, publications that are unavailable through Google Scholar or Scopus were collected through the library of the University of Salford. The identified papers from each database were imported straight into the free citation manager Endnote. Finally, duplicates were removed from each database, resulting in 394 papers for the screening step.

While the use of statistical and geostatistical analysis for crime forecasting has been studied for some time, in the past two decades there has been an increasing interest in the development of tools that use massive data sets to provide crime predictions (Perry, 2013). Thus, predictive analytics have been used into law enforcement strategies (Brayne, 2017). This is why, during the screening phase, articles older than 10-year-old were rejected. Second, duplicates were removed from the selected datasets. Thirdly, each article was reviewed to exclude those deemed "irrelevant." This determination was achieved by defining "relevant" articles based on the following three criteria. The first element is that a document must cover criminal acts that occurred inside defined geographical boundaries. Common examples of rejected papers include those dealing with the fear of crime (Taylor and Hale, 2017), offenders' characteristics (Browne et al., 2022), offender, victim characteristics (Vakhitova et al., 2021), geographical profiling (Goodwill et al., 2013), hot spot (Almanie et al., 2015, Yang et al., 2018b), and geographical information systems (GIS) (Kedia, 2016). The second condition for research to be considered "relevant" is that it uses a data mining technique that exhibit exploratory or cluster analysis. The third component is the incorporation of social media data. This indicates that there are set parameters for the analysis, such as the location of the research, whether or not to utilise social media data, and whether or not to use a data mining approach in the study area. The relevant aspects were identified using the following methodology: (a) read the title and screen graphics and/or maps; (b) if confused about relevance, read the abstract; and (c) if still uncertain about relevance, search the text for relevant terms (location, data mining, twitter).

The last step of the screening procedure included excluding irrelevant articles that the authors were unable to access due to subscription restrictions. The screening process returned 193 articles suitable for the third and final rounds. During this last step, the eligibility phase, the abstracts and main body of each of the 193 articles were examined and analysed (e.g., study area, data, methods, and results). The objective was to gather data that would comprise the paper's qualification requirements. These are classified into three groups: (a) procedures; (b) significance: relevance and purpose of the data item; and (c) research characteristics: data item study area, data source, data mining techniques, and evaluation metrics. Following that, each category and the corresponding data elements were examined. The kind of publishing is the first data point. Occasionally, conference papers are omitted from literature reviews because their quality is not evaluated in the same manner as International Scientific Indexing (ISI) publications. However, several conferences are recognised as very legitimate publishing outlets in certain fields, such as computer science. high number of papers written by professionals were noticed in computer or information science during the screening process; hence, conference papers were not rejected at this time. Seven novels or book sections were removed in total. The next two "relevance" criteria (i.e., relevance and purpose) address the conformity of the papers' content to the subject of this research. During this stage, the relevance of the document was double-checked. Several articles that seemed pertinent during the screening step (e.g., research on criminal occurrences, big data, and prediction) were found to be irrelevant when reading the paper's main body. Even though crime prediction was discussed in the abstract, the authors alluded to the fact that they were giving a framework or survey for future study. the data item "data sources" have been added to this research effort in order to integrate techniques for modelling and examining correlations between dependent and independent variables (e.g., crime predictors). There were 49 papers that were disqualified due to these criteria.

Lastly, four additional criteria related to the quality and consistency of the chosen articles were presented. The study area should not be limited to the United Kingdom or Greater Manchester to begin with. Cities are less susceptible to border effects than smaller administrative units within cities that share borders for example a district. Although the smaller the research location, the more likely it is that the findings would be applicable to the study's characteristics, there have been few studies on crime prediction using data mining methods published in the United Kingdom, and none using the Manchester dataset. Second, the period of the crime sample is not limited to one year; it may be shorter or longer in order to reduce selection bias and ensure transparency.

These two factors also contribute to the consistency of the study. Nonetheless, as shown in further detail in the Results section, there are substantial differences in the studies. The last two conditions are the restriction to empirical data analysis (e.g., elimination of proof-of-concept or purely methodological research) and the use of measurements to assess model performance (e.g., accuracy, precision). The last two criteria ensure that we exclusively review publications relevant to this study. There were 53 papers that were rejected based on factors related to their relevance to research. Figure 2-1, also depicts the number of articles selected between 2013 and 2021. The screening technique ultimately yielded 91 articles.

2.3.2 Study quality

Before each phase, the primary supervisor and co-supervisor of this research evaluated the articles that comprised this study. This research was further analysed and debated until all sides reached consensus on the following step. Multiple cross-checks were performed on the study's results to guarantee methodological consistency. Throughout the last step (eligibility), the research student evaluated the papers many times to verify that all eligible publications were included. Regarding the results subsections of the four study stages ("Study characteristics," "Overview of selected publications on application of ensemble machine learning for crime prediction," "Prediction of crime using social media (Twitter) and historical data," and "factors to consider when analysing prediction performance"). A three-step procedure was utilised to extract information that was structured as data items: extract, discuss, and analyse.

First, the authors manually extract the data components and their values from the articles by reading them (1-extract). The researcher and supervisors then reviewed and evaluated the data items and their values (2-discussion/consensus). If information is still unclear, it is compared against information that is readily available for clarification (3-analysis). This information was organised as a matrix, with rows representing papers and columns indicating different processing data (e.g., a data item is the year of publication). Table 2-1, give a list of prior studies that applied data mining techniques on historical crime data, while provides a list of studies that used Twitter data alone or in combination with other data sources. In the "Results" section, the characteristics of the items are explored.

Using the study scale, the risk of bias in each research was evaluated. As indicated throughout the eligibility procedure, spatial and temporal limitations were imposed to guarantee that we analyse medium-to-large-scale research and that the results are not biased by location or season. In addition, we did not discover any duplicate publications (i.e., two or more papers including identical samples and experiments) nor any study characteristics, such as unique and uncommon traits or research topics.

2.3.3 Results

2.3.3.1 Study characteristics

This investigation ultimately yielded 91 papers. About 79% of the selected articles examined just historical crime data, whereas about 19 of the total publications (21%) analysed. The proportion Twitter data to other sources of information was presented in the Figure 2-2 below. The distribution of selected publication between year 2012 and 2021 was shown in Figure 2-3.

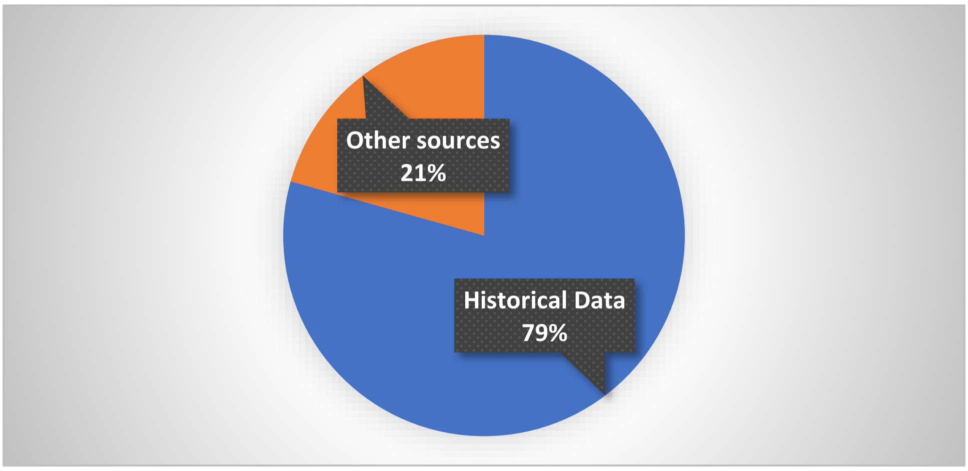


Figure 2-3: Graph Showing the percentage of articles that used historical crime information to other sources of information including Twitter extracted information

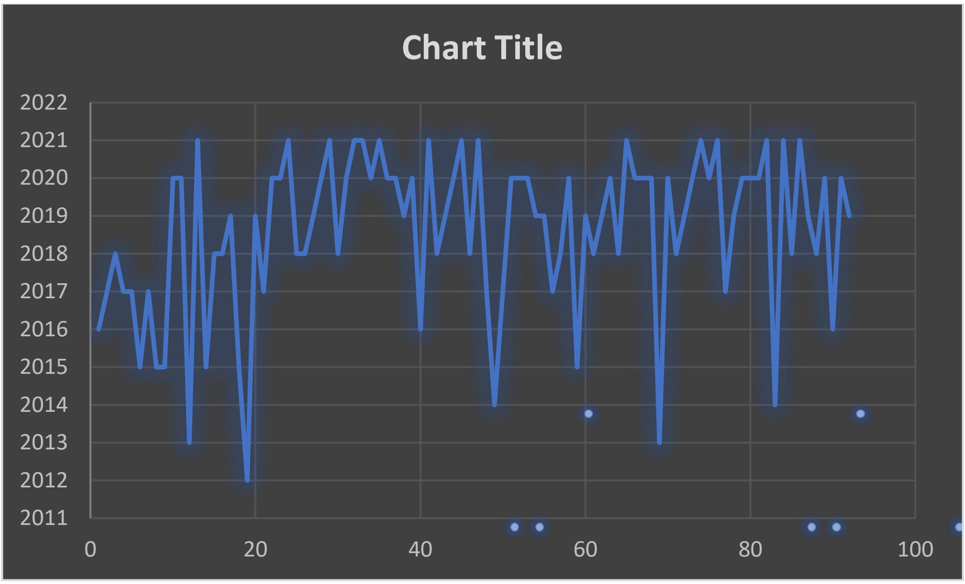


Figure 2-3: Graph showing the distribution of chosen publications from 2012 to 2021

2.3.3.2 Classification task

About 37 out of the 62 articles that used only historical crime data employed classification strategies. Fourteen used regression techniques, while the remaining eleven utilised forecasting methods. There were seventeen citations in different journals, twenty-one in conference papers, and just six in serial journals or proceedings that applied classification across different data sources. Eleven of them articles from the US were cited in a journal article using the US data for classification. Five of the eighteen were cited in conference papers and the remaining one in serial journal. the Five of the articles cited using the US data were from hem used information from San Francisco, two from Chicago and, two from a generic subset of the US dataset, and one from each of Baltimore, Boston, Denver, and Portland. The most current US dataset was compiled in 2018(El Bour et al., 2018, Elluri et al., 2019, Esquivel et al., 2020), while the oldest was compiled in 2001(El Bour et al., 2018). The first paper to reference the US dataset was published in 2013(Iqbal et al., 2013), and the most current were published in 2021(Na et al., 2021, Sharma et al., 2021, Wang et al., 2021). Accuracy (n = 8), precision (n = 4) and recall (n = 4) were the most often cited performance metrics in these articles. F-measured was mentioned three times, but AUC-ROC and log loss were mentioned just once.

Random forest (n = 10) was the most often cited machine learning model in the United States, followed by decision tree (n = 6), neural network (n = 5), nave bayes (n = 5), and k-nearest neighbors (n = 3). Using only historical data from the United States, it was determined that random forest classifiers outperformed other models in only two (n = 2) instances, whereas decision trees, gradient boosting, neural networks, and support vector machines each outperformed other models in just one (n = 1) instance. Additionally, it has been revealed that model performance changes under three distinct conditions.

In every paper that referenced US datasets, the sampling methodology was described. The majority (n=3) used the 70% to 30% splitting method, while two out of nine utilised 75% by 25%, two employed 10-fold cross validation, and one employed 90% to 10% and 60% to 40%. Only four of the studies that constructed classifiers on the US dataset included feature engineering. Two of the nine investigate the impact of class imbalance on model performance, while four of the nine apply feature selection to the variables prior to data partitioning (correlation, percentile, decision tree, and manual selection).

India (n=4) ranked second in this analysis in terms of the number of datasets cited. The most recent dataset used dates back to 2015, while the oldest dates back to 2001. The earliest article referencing these datasets appeared in 2019, while the most recent appeared in 2021. On the Indian datasets, the nave bayes classifier and the K-nearest neighbors' classifier were each employed three times (n=3). The random forest model was employed twice (n = 2), whereas the decision tree, bagging tree, adaboost, support vector machine, and stacking models were only employed once (n = 1). The 60 percent to 40 percent divide was used once in the essay, whereas the 80 percent to 20 percent split was utilised twice. Before creating their models, none of the three artists used feature selection or feature engineering. In addition, none of them considered the impact of class imbalance on the accuracy of crime prediction models. None of these three considers the possibility of solving a problem via exploratory analysis.

Brazil (n=3) ranked second to India (n=3) in terms of the number of datasets cited in this research. The most current dataset used dates back to 2018, while the oldest dates back to 2012. In 2020, these two articles were referenced. Other machine learning classifiers (Naive bayes, K-nearest neighbors, decision tree, bagging tree, adaboost, support vector machine, and stacked models) were only used once (n = 1) in the two papers. Only one of the articles discussed 10-fold cross validation, whereas the other divided the data set by year (2 year for training and 1 year for testing). The feature selection was performed manually. Before developing their models, none of the two articles explored feature engineering, class imbalance problems, or exploratory data analysis.

The number of Canadian studies listed was second only to those from Brazil (n=1). The most recent dataset used dates back to 2018, while the oldest dates back to 2003. Only of the papers published in 2018 were able to meet the criterial of this study. Only one instance (n = 1) of the logistic regression, K-nearest neighbors, boosted tree, support vector machine, and random forest models were employed. Only one of the papers discussed 5-fold cross-validation, and none of them addressed the splitting method. Despite the fact that both studies studied feature engineering, the effects of class imbalance and feature selection were not examined. In addition, none of the two studies include exploratory data analysis. Accuracy is the only criteria used to assess the performance of the models in these two publications.

Both Saudi Arabia and Canada have the same number of listed datasets. Previous coverage of the Saudi dataset occurred in 2019, while the most recent coverage emerged in 2020. The naive Bayes model was the most popular machine learning classifier, followed by random forest, decision tree, and deep learning algorithms. It appeared that Naive Bayes performed the best on the Saudi Arabia dataset. FAMD and PCA were used to choose features, and their effects on prediction models were evaluated. Each study used exploratory data analysis. However, neither the influence of feature engineering nor the class imbalance issue was investigated.

In this research, Nigerian and Tunisian datasets were cited once. In each of the studies (n = 1), the decision tree model was referenced once. While the performance of decisions on the Nigeria dataset is high, the performance of models on the Tunisia dataset differs between crime categories. In the two studies that mentioned datasets from both countries, decision tree, KNN, SVM, and random forest were the most often used performance metrics. In both articles, accuracy was the most often used metric.

Table 2-2 List of studies that applied classification method on historical crime information

Author	Year	Model types	Country of Research
Adesola et al. (2020a)	2020	SVM	USA
Adesola et al. (2020b)	2020	DT	Nigeria
Albahli et al. (2020)	2020	NB	Saudi Arabia
Almaw and Kadam (2019)	2019	NB, J48, RF	Denver, USA
Alsaqabi et al. (2019)	2019	NB, RF, DT, DL	Saudi Arabia
Baculo et al. (2017)	2017	BNET, NB, DT, RF, KDE	Philippines
Bappee et al. (2018)	2018	LR, SVM, RF	Canada
Castro et al. (2020)	2020	KNN, SVM, RF, XGB	Brazil
Chandrasekar et al. (2015)	2015	NB, RF, SVM, GB	San Francisco, USA
Das and Das (2019)	2019	DT, KNN, NB, RF, ADt,	India
El Bour et al. (2018)	2018	NB, DT, RF, NN, SVM	Chicago, USA
Elluri et al. (2019)	2019	MLP, DT, NN, LR, RF	NY, USA
Esquivel et al. (2020)	2020	CLSTM-NN	Baltimore, USA
Feng et al. (2018)	2018	NB, KNN, RF, XGB, Holt-winter	San Francisco, Chicago, Philadelphia
Gayathri et al. (2021)	2021	NB, RF	Kaggle data
Hajela et al. (2020)	2020	NB, DT	San Francisco
Hossain et al. (2020a)	2020	DT, KNN, RF	San Francisco, USA
Ippolito and Lozano (2020)	2020	NN, NB, DT, RF, LR, EL	Sao Paulo, Brazil
Iqbal et al. (2013)	2013	DT, NB	USA
Khatun et al. (2020)	2020	DT, KNN, RF	USA
Kim et al. (2018)	2018	KNN, Boosted DT	Vancouver
Kiran and Kaishveen (2019)	2019	NB, KNN	India
Kissi Ghalleb and Ben Amara (2020)	2020	DT, KNN, SVM, RF	Tunisia
Kshatri et al. (2021)	2021	SVM, j48, c4.5	India

Author	**Year**	**Model types**	**Country of Research**
Llaha (2020)	2020	DT, SLR, LR, MLP, NB, BN	
Na et al. (2021)	2021	LR, NN	USA
Nguyen et al. (2017)	2017	SVM, RF, GB, MLNN	
Palad et al. (2019)	2019	DT, NB, SMO	Philippines
Patil et al. (2020)	2020	ANN	Boston, USA
Sandagiri et al. (2020)	2020		
Sarma et al. (2021)	2021	KNN, RF, DT, LR, SVM	
Sathyadevan (2014)	2014	NB	
Sharma et al. (2021)	2021	RF, DT, with PCA	
Toppireddy et al. (2018)	2018	visualisation, KNN, NB	UK
Wang et al. (2021)	2021	Ensemble method	USA
Wijaya et al. (2019)	2019	NB, LSVM, RF KNN	Indonesia
Wijenayake et al. (2018)	2018	DT	Australia
Yao et al. (2020)	2020	RF	San Francisco, USA
Zhang et al. (2016)	2016	KNN	Nanchang, China
Deshmukh et al. (2020)	2020	RF	Mumbai, India
Lee et al. (2019)	2019	LR	South Korea

2.3.3.3 Regression Task

Fourteen of the sixty-two (n=14) articles that investigated just historical crime data used regression models. Seven (n=7) journal articles, three (n=3) conference papers, and two (n=2) serial magazines mentioned them. The earliest of these works was released in 2015(Cavadas et al., 2015), while the most recent appeared in 2021(Abrams, 2021, Forradellas et al., 2021, Mishra et al., 2021, Obagbuwa and Abidoye, 2021, Shukla et al., 2021). Five of them (n = 5) used US datasets, while three of the five were cited them in journal articles and one (n = 1) each was cited in a conference and serial paper. The earliest American dataset was from 2001(Alves et al., 2018), while the most recent was from 2019 (Forradellas et al., 2021, Kadar et al., 2018). Two of these articles deal with the forecast of violent crime, while the other two deal with the prediction of crime occurrence. Two of these studies utilised a data splitting ratio of 80% to 20% for the training and testing datasets, but only one applied five-fold cross validation. Only one of these studies made use of exploratory data analysis, whereas the other two applied both feature engineering and feature selection. Random Forest was the most cited publication among the four that mentioned the US dataset, followed by MARS, ERT, LGB, Dl, and Poisson regressors. Tree models outperformed other models in terms of performance. R-squared is the most often used model performance metric.

Two (n = 2) of the twelve studies that applied regression on historical crime data referred to Brazilian and American datasets, respectively, in a journal article and a conference paper. The dataset covers the years 2001 to 2019, with the most current data being collected in 2019. In addition, the earliest of these pieces was published in 2018 and the most recent in 2020. One of these studies examined murder prediction statistics, while the others examined the location of criminal activity. The two trials separated the training and test datasets 80 percent to 20 percent, respectively. Only one research used feature engineering and feature selection, but the others depended on data exploratory analysis. In none of the two studies is the impact of class imbalance investigated. In their investigation, they emphasised random forest, choosing tree, bagging tree, and extra tree. When numerous models from the same study were examined, random forest outperformed the others. These studies used precision, mean square error, and root mean square error as measures of performance.

Two (n = 2) of the twelve regression studies on historical crime data used Brazilian datasets as opposed to American datasets, and both were cited in a journal article and a conference poster. The dataset spans the years 2001 through 2019, with the most recent data being from 2019. In addition, the oldest of these works was published in 2018, while the most recent was released in 2020. One of these studies focused on murder prediction statistics, while the others focused on crime location. For the training and testing datasets, the two trials used a data splitting ratio of 80% to 20%. Only one study used feature engineering and selection, while the others relied on data exploration. In neither of the two research is the effect of class disparity assessed. Their research indicated a random forest, a selection tree, a bagging tree, and an extra tree. In a study comparing many models, random forests fared better than other models. In these experiments, accuracy, mean square error, and root mean square error were used as performance indicators.

Two of the twelve studies analysed historical Chinese crime data using regression methods. Both a journal article and a conference paper cited them. The original version of the dataset was developed in 2013, while the most recent version was generated in 2016. In 2020, the two probes were discussed. One of these studies focused on theft prediction statistics, while the others examined criminal risk. In both studies, bootstrapping and cross validation were performed. There was no use of feature engineering or feature selection in any of the two experiments that used China datasets. Furthermore, none of them used exploratory data analysis. In one of these experiments, attributes were selected using the bagging tree. Neither of the two studies also analyses the influence of class disparity in their study. Support vector machine regressors, additional tree models, and LASSO models were used in the research. Extra Tree, in contrast, outperformed other models. As performance measures, the two investigations utilised accuracy and modified R-squared.

Each of the datasets from Bangladesh, Argentina, the United Kingdom, and South Africa were subjected to regression analysis. On the UK data set, the split ratio for the training and the test dataset was 80% and 20% respectively, it was 94% and 6% for Bangladeshi. In one study, Argentina used feature engineering, but the United Kingdom and South Africa did not. In none of the investigations was feature selection used. In two studies

(Argentina and South Africa), exploratory data analysis was undertaken, but not in Bangladesh or the United Kingdom. Class differences were only the subject of one of the study initiatives. Bangladeshi data collection began in 2002 and finished in 2019. (Argentina). Crime factors discussed include criminal behaviour in the United Kingdom, crime patterns in Bangladesh, murders in Argentina, and future crimes (South Africa). Linear regression was applied three (3) times, whereas polynomial regression, random forest, neural network, and locally weighted linear regression linear regression (LWL) were each employed just once. R-squared, modified R-squared, MAE, and MSE were the metrics used to assess the performance of the model. Polynomial regression and LWL outscored the comparison's other models.

Table 2-3 List of studies that applied regression method on historical crime information

Author	Year	Model types	Country of research
Abrams (2021)	2021	regression	US
Belesiotis et al. (2018)	2018	RF, SV, Ridge	London, UK
Alves et al. (2018)	2018	RF	Brazil
Biswas and Basak (2019a)	2019	LR, Polynomial regression, RF	Bangladeshi
Da Silva et al. (2020)	2020	DT, BT, RF, ET	Fortaleza, Brazil
Forradellas et al. (2021)	2021	NN	Buenos Aires
Kadar and Pletikosa (2018)	2018	RF, GB, ERT	NYC, USA
Lamari et al. (2020)	2020	LGB, DL, PR	USA
Mishra et al. (2021)	2021	DT, LWL, NB, Linear R	Hampshire, UK
Obagbuwa and Abidoye (2021)	2021	Linear Reg	South Africa
Shukla et al. (2021)	2021		North Carolina, USA
Wang et al. (2020a)	2020	LASSO, ET	China
Zhang et al. (2020a)	2020	DL	
Williams et al. (2019)	2019	Neg.Binomial	London

2.3.3.4 Forecasting

Twelve of the 91 articles predicted different types of crime using machine learning and statistical techniques. The first was published in 2016, while the most current was published in 2021. Five of the eight publications were journal articles; two were conference papers; and the last publication was a series. The United States was referenced in four of the papers, while Spain, Poland, and China each cited one dataset. In the United States, the oldest dataset was from the 1960s, while the most recent was from Spain in

2020. The 90% to 10% data split was used twice. The ratio of 70% to 30% was used twice, although cross validation was performed just once. Deep learning and its variations were cited a total of six times. The support vector machine was mentioned three times. Two references were made to linear regression, random forest, logistic regression, decision tree, extreme gradient boosting, and the Gaussian Process.

Table 2-4 List of studies that applied forecasting method on historical crime information

Author	Year	Model types	Country of research
Ashby (2020)	2020	SARIMA	US
Awal et al. (2016b)	2016	linear regression	Bangladeshi
Butt et al. (2021)	2021	HDBSCAN and SARIMA	US
Catlett et al. (2018)	2018	auto-regressive	Chicago, US
Flaxman et al. (2019)	2019	RKHS	
Karmakar and Das (2020)	2020	Bayesian	
Shi et al. (2021)	2021	STL-FNN	
Vomfell et al. (2018b)	2018		NY, USA
Wang and Ma (2021)	2021	SVM, RF	China
Misyrlis et al. (2017)	2017	linear regression	Oregon, USA
Rahmani (2014)	2014	ARCGIS software	
Rummens et al. (2017)	2017	Ensemble (LR, NN)	Amsterdam, Netherlands

2.3.3.5 Historical data and Twitter Information combined.

In this section, there were 22 studies. Thirteen conference papers referenced the US dataset. Eight of them used a subset of US data, while one was cited from Chicago and another from Phoenix, both of which utilised Twitter data. The first US Twitter data in this area dates back to 2010, and it was published in a scholarly journal. The most current information is from 2015, according to a journal article. Journal articles comprised nineteen of the twenty-three pieces. While three of them referenced the United States dataset, one of them obtained data from New York City. The earliest reference to a journal paper was in 2012, while the most recent was in 2015.

Only one journal article was cited in the South African and United Kingdom databases. In addition, one research based on Tunisian data was cited in a conference paper. In 2012, the first of these categories was stated, and in 2014, the most recent. Twelve of the twenty-two articles used Twitter data exclusively. Four of the individuals were U.S. citizens. Two were referenced in conference papers, one in a journal article and one in a series. Using just Twitter data from the United Kingdom, South Africa, and Latin America, one was mentioned. The oldest reference dates back to 2012, while the most recent is from 2020. Ten of the twenty-two articles combined Twitter data with previous crime or meteorological statistics. The United States specified nine of these categories. One from Chicago, one from New York, and one from Phoenix. The UK dataset was cited by just one of the papers in this category. The most recent data collection in this category was from 2020, while the oldest was from 2010.

Table 2-5 List of studies that applied machine learning method on Twitter data

Authors	Date	Region	Citation Type	Model Types	Extraction Method	Task
Aghababaei and Makrehchi (2016)	2016	US	C	NA	active users	daily
(Aghababaei and Makrehchi, 2017)	2017	US	C	NA	active users	daily
(Aghababaei and Makrehchi, 2018)	2018	US	J	LinearSVC	active users	daily
(Al Boni and Gerber, 2017b)	2017	US	C	KDE	GPS tagged	daily
(Almehmadi et al., 2017)	2017	NA	C	SVM classifier	GPS tagged	daily
(Azeez and Aravindhar, 2015b)	2015	NA	C	Graph database	NA	
(Cesur et al., 2017)	2017	NA	J	MLP	NA	
(Chen et al., 2015b)	2015	Chicago USA	C	KDE, LR	GPS tagged	6-hour
(Corso, 2015)	2015	Phoenix, USA	C	NA	GPS tagged	
(Curiel et al., 2020)	2020	latin America	J	hotspot	GPS tagged Spanish speaking	daily
(Das et al., 2020)	2020	NA	J	trend analysis	keywords	
(Featherstone, 2013)	2013	South Africa	J	Exploration	no. of tweets	
(Karmakar and Das, 2021)	2021	undefined	C	forecasting	keywords	monthly
(Kounadi et al., 2015)	2015	UK, London	J	Logistic regression	keywords	
(Lekha and Prakasam, 2018)	2018	NA	C	Proposal	NA	
(Vomfell et al., 2018a)	2018	NY, USA	J	Forecasting	no. of tweets	weekly
(Tundis et al., 2019)	2019	Tunisia	C	RF	keywords	
(Wang and Gerber, 2015)	2015	USA	C	regression	geotagged tweets	
(Wang et al., 2012)	2012	USA	S	LDA	all tweets	
(Wang et al., 2019)	2019	USA	J	regression	keywords	
(Williams et al., 2017)	2017	UK, London	J	regression	keywords	yearly
(Vishwamitra et al., 2020)	2020	USA	C	BERT	keywords	

Two of the twenty-two combined historical crime data and weather data with Twitter data, while one used just weather data and Twitter data. In terms of the extraction procedure, a total of six methods have been identified. Three of the twenty-two research projects in this field made use of the number of active users. Three were derived using data collected from all publicly available tweets. Six of the twenty-two relied on geotagged tweets and six relied on keywords to get data from Twitter. Two people from the United States collected all the tweets, and both of them were featured at a conference and in a journal article. The second was published in a journal and originated in South Africa. Four Americans used GPS-tagging. One was described using Chicago data, while the other was described with Phoenix data. In a conference paper, they were all cited. A journal article cited the only individual from Latin America. In their inquiry, six of the twenty-three studies applied daily analysis, while one employed annual, monthly, six-hourly, and hourly analysis. One of the two pieces from the United Kingdom had a yearly analysis. In addition, the bulk of the published works used binary classification. In conference papers and scholarly journals, three researchers on binary classification were referenced three times.

2018 had a rise in the number of relevant articles, while 2019 saw a sharp fall. 2020 marked its zenith, followed by a significant decline in 2021, likely due to COVID-19. In this field, the number of publications pertaining to this issue changes throughout time. The greatest number of relevant cited papers that studied Twitter information peaked in 2017 and then sank drastically until 2020, when it was roughly 80% of what it had been in 2017, and then reduced to around 25% of what it had been in 2020 in 2021.

Fifty-three percent of the papers cited used previous crime data with or without weather data, while about 47 percent utilised Twitter data with or without historical crime data. N = 31 of the selected papers were analysed purely using classification and regression. Twenty-four (77%) used solely classification procedures, whereas seven (23%) employed only regression techniques. Twenty-two (71%) of them relied only on crime figures from the past, disregarding all other information sources. The USA dataset was used by eight (26%) of the 22 respondents. Three (38%) of the eight datasets originated from San Francisco, while one (n = 1) each originated from Denver, Baltimore, Portland, and Boston. Five (n = 5) of the American researchers were referenced in a journal article, two (n = 2) in a conference paper, and one (n = 1) in a series of journal publications. The oldest reference was made in 2013, while the most current reference was made in 2020. The oldest dataset included in these studies was the San Francisco dataset from 2003, while the most recent was the Baltimore dataset from 2018.

The random forest model (n = 6) is the most often cited machine learning model in the United States, followed by the decision tree algorithm (n = 4) and the nave bayes, k-nearest neighbors, and neural network combined algorithms (n = 3). Support vector machine and gradient boosting were discussed on two separate occasions, however severe gradient boosting was only mentioned once. Using just historical datasets from the United States, it was observed that random forest outperformed other models in only two (n = 2) domains, whilst choice, neural networks, and gradient boosting each performed better in only one (n = 1) domain. In addition, it has been shown that model performance varies, for instance, between random forest and decision in one inquiry and between random forest, decision tree, and support vector machine in another.

India (n = 3) ranked second (n = 3) in terms of the quantity of datasets cited in this research. K-nearest neighbors and random forest were used twice (n = 2), but decision tree, nave bayes, adaboost, support vector machine, and stacked models were employed just once (n = 1). n = 2 references were made to the datasets from Brazil, Canada, and Saudi Arabia. On the Brazil dataset, random forest was employed twice (n = 2), as opposed to k-nearest neighbors (KNN), support vector machine (SVM), extreme gradient boosting (XGBoost), logistic regression (LR), decision tree (DT), and nave bayes (NB), which were each used once. Their efficacy varied across tree models based on time intervals (KNN and SVM). In one (n = 1) investigation, random forest outperformed KNN, and in another (n = 2), gradient boosting beat KNN. The naive Bayes model outperformed random forest, decision tree, and deep learning models in studies conducted in Saudi Arabia. This research cited datasets from Nigeria and Tunisia once each. The decision tree model was cited twice in these publications (n = 2). The performance of judges on the Nigeria dataset is exceptional, but the performance of judges on the Tunisia dataset varies by kind of crime.

In the 22 (n = 22) research investigations that only used historical datasets and classification models, accuracy (n = 14) is the most frequently utilised performance metric, followed by precision (n = 2). There was a single mention of recall, F-scores, and specificity.

2.3.4 Analysis and Discussion

As indicated previously, chosen publications incorporate the authors' suggested machine learning baseline models, statistical models, or ensemble models. Evaluation measures for such models are well-known in criminology and past crime prediction studies. However, aside from predicting, few writers emphasise the relevance of integrating or applying several evaluation measures.

It is challenging to compare the assessment conclusions of the 67 publications due to a range of factors, including the varying number of variables, study areas, and data mining approaches. It may be required to investigate their similarities. Nonetheless, the principal prediction result, such as crime counts, binary classification (crime is absent or present), or multi-class classification, has a substantial impact on the choice of an evaluation measure (predicting different crime types). For the classification task, the most frequently used evaluation metrics are accuracy (n = 18), precision (n = 8) and recall (n = 7), whereas for the regression task, the most frequently used evaluation metrics are Mean Squared Error (MSE, n = 2), Root Mean Squared Error (RMSE, n = 2), R-squared (n = 3) and modified R-squared (n = 1). The use of the top three evaluation metrics (accuracy, precision, and recall) for assessing model performance was investigated. While the majority of studies investigates other assessment metrics, only a minority considers the influence of higher or lower values of other metrics. And just one experiment investigated specificity.

2.3.5 Algorithms and validation strategies

The following algorithms have attracted the most attention: (i) Random Forest, (ii) Decision Tree, (iii) Nave Bayes, and (iv) deep learning and (v) KNN. Random Forest classifiers outperformed other models in around five cases, followed by naive bayes. Nonetheless, for naïve bayes, the same datasets from Saudi Arabia and the same author were employed. Furthermore, there are several validation approaches. Approximately 37% of the researchers (n = 25) consider dividing the data into training and testing groups. The bulk of these studies are made up of 80% training and 20% evaluation sets. Adesola et al. (2020b) use the decision tree method in conjunction with stratified cross validation and the default splitting technique. Shi (2020) employed the Bootstrapping sampling strategy on the support vector machine regressor in a regression task. Classification metrics are evaluation metrics that are designed specifically for classification activities (e.g., accuracy, precision and recall). Regression metrics are a collection of error measurements used in regression analysis (e.g., MSE, RMSE, MAE).

Random forest algorithms are preferred for crime prediction (n = 14). Interestingly, model performance varies according to different research, regardless of machine learning tasks, and it is not geography dependent (Castro et al., 2020, Elluri et al., 2019, Khatun et al., 2020, Kissi Ghalleb and Ben Amara, 2020, Nguyen et al., 2017). Nguyen et al. (2017) also indicated in their research findings that support vector machine is not appropriate for some of their tasks. This contradicts the findings of Nitta et al. (2019), in which support vector outperformed LFSNBC in terms of prediction. Some research has also challenged linear models due to their poor performance (Kadar and Pletikosa, 2018). Although Lamari et al. (2020) anticipated a linear connection between crime and its variables, their analysis violates this assumption. This confirms that criminal behaviour cannot be considered to be linear, and hence a simple linear equation cannot be used to forecast crime. Only a few research evaluated the impact of feature engineering while developing their models, with some models, such as logistic regression and neural networks, performing better with sufficient feature engineering. Most research did not include the impact of class imbalance while developing their models. And not one of them mentioned the difference in model performance between training and testing data.

Although Wang et al. (2019) identified a substantial correlation between social media and crimes in the United States, Curiel et al. (2020) discovered that less than 1% of social media is connected to crime or the fear of crime. This might be due to the fact that the bulk of the research that used Twitter used an additional source of information retrieved tweets using keywords. Using merely keywords, while useful for text categorization, is insufficient in criminology and can be deceptive (Aghababaei and Makrehchi, 2018), because criminals are sophisticated enough not to divulge all of their activity on social media (Shi et al., 2021). The two studies done in the UK retrieved their twitter data using keywords, and both employed regression analysis (Kounadi et al., 2015, Williams et al., 2017), and even the one that used just historical crime data (Mishra et al., 2021) applied regression on data from Hampshire in the UK. It is worth noting that none of the UK research utilised the original dataset's demographic information, which might be attributed to the dataset's unorganised nature.

None of them really predicted crime using typical data science methods. Furthermore, tree-based models were not used, and while random forest performed well in certain experiments, the fact that its performance varied from that of other models indicates that more robust machine learning methods should be explored. (Kshatri et al., 2021) used stacked ensembles of different machine learning models to predict crime on an Indian dataset.

It was also reported that the majority of studies (Elluri et al., 2019, Sandagiri et al., 2020, Williams et al., 2017) that combined information from multiple sources were merging unrelated information, such as those that integrated census data with crime data, where the census took place before the crime was committed and there is no evidence that the offenders were still living in the area or were still alive. Although a decent result may be obtained, the outcome may be deceptive. Only one of the sixty-seven research included in the study used stacked ensembles of heterogeneous models, and it was never used to predict crime using social media data. Discussion In this section, we perform a SWOT analysis of the most significant findings.

2.3.6 Strengths

The merging of social media information taken from Twitter into standard prediction algorithms is one of the most powerful aspects of current research endeavours. This method is seen in the work of (Aghababaei and Makrehchi, 2018, Al Boni and Gerber, 2017b, Chen et al., 2015a, Corso, 2015, Williams et al., 2017). Even if the correlation of social media (Twitter) was minor, it was discovered that if correctly mined, some vital hidden trend may be disclosed, which will be highly valuable for security personnel and police departments. Furthermore, the linked work area demonstrates the scientific community's interest in the incorporation and influence of social media in prediction. This enthusiasm is bolstered by the tendency of incorporating weather, demography, and socioeconomic aspects into the modelling process. Surveillance plots (Fig. 4) offer a more complete view of the accuracy of the various machine learning predictions in terms of performance evaluation.

2.3.7 Weaknesses

In general, key experimentation details are not usually provided and are often ambiguous. In the majority of instances, feature selection and feature-engineering dependent procedures are inadequately stated and not used. Few research conducted in-depth exploratory data analysis, and the number of studies that explored class imbalance issues was negligible. The bulk of research that investigated Twitter relied only on keywords, particularly those from the United Kingdom, although other studies gathered tweets from active Twitter users. This vulnerability allows criminals to modify prediction algorithms and deceive security personnel. Those studies that used the UK dataset did not use the demographics included in the UK dataset; rather, they combined the crime dataset with census data from different years and places, as has been the case in the majority of previous studies. In addition, only a small number of studies have investigated the impact of different sampling methods, including stratified sampling, bootstrap sampling, and stratified geographic sampling.

It is difficult to repeat research or compare its results to those of a prospective future study due to the aforementioned issues. Nevertheless, in all modelling approaches, a rigorous data mining process, such as CRIPS-DM, must always be followed. In addition, the Random Forest is the most popular machine learning technique, which may not be optimal for all datasets, as shown by past studies. Until now, only one research has addressed this problem, and unfortunately, this study did not use additional data science techniques, such as feature engineering, class imbalance issues resolution, and data exploratory analysis. Most authors assess performance using standard metrics. A "universal" benchmark statistic, such as accuracy, may not be enough to reflect the performance of models created expressly for sensitive subjects, such as crime. Therefore, we recommend developing a stacked ensemble of a heterogeneous model based on accuracy, precision, and recall, which would combine Twitter data with historical crime data. Modelling concerns such as overfitting, multicollinearity, sample bias, and a lack of data are hardly, if ever, addressed.

2.3.8 Opportunities

Crime prediction research is growing in popularity. From the pool of 67 selected articles, seventeen, twelve, and four were published in 2019, 2010, and 2021, respectively, compared to one or two each year between 2013 and 2017. This illustrates the growing interest in this sort of research from specialists in other disciplines. The most studied kinds of crime are robbery and theft. Exploratory data analysis on detailed data may reveal the most common sorts of crime in the United Kingdom, as well as the relationship between each type of crime. Due to the fact that the majority of crime prediction research use US datasets, which vary in reporting format and structure from UK datasets, it may be possible in the future to rigorously analyse if there is a pattern of consistency in the pattern of crime in the UK. In addition, with the exception of a single research, stacked ensemble methods have not been used in crime prediction. This may be due to the authors' different academic backgrounds, including computer science and criminology, and their unfamiliarity with data science techniques and methodologies. By combining criminal history information with social media data, this creates an opportunity for researchers to examine and compare less frequently used data mining methodologies with baseline machine learning models and ensembles of various machine learning algorithms. Another thing that may be done is to assess the predictive accuracy of the algorithms using both training and testing data.

In this review, although methodological trends were discussed, but a meaningful comparison of regional methods was not possible. Certain methods were not initially compared to a control approach. Other studies explored the same process with various settings. Even if there were papers with a comparable set of characteristics, comparing them would be flawed owing to variations in sample data, study sites, sampling dates, etc. Future empirical study should compare the constantly rising number of algorithms.

The selected publications were divided into three data mining categories: classification, regression, and machine learning-based forecasting. The majority of proposed solutions include machine learning algorithms, with MLP and RF being the most prevalent, and AR models being the most often employed baseline methods. Instead of continuing to compare old or simpler techniques that have repeatedly underperformed, compare freshly generated or newly developed algorithms against the most often presented algorithms.

2.3.9 Threats

The most prevalent predicting approach is the binary classification (n = 21), whereas the classification (n = 20) is the most prevalent predictive method. It is essential to know the cause of this trend. Is regression analysis less predictive or less useful? We recognise the importance of both prediction tasks and the continual increase in the development of classification algorithms and features to be included into classification work. For effective and efficient crime control, it is also essential to investigate the potential advantages of utilising social media to anticipate crime or a stacked ensemble categorization to predict crime categories.

2.3.10 Summary of the findings

This research focuses on "crime prediction using crime history data and social media data," which is an inference approach for both temporal and geographical crime. A comprehensive literature review utilising the reporting guidelines "PRISMA" (Liberati et al., 2009) was conducted in order to comprehend and assess the state of the art in empirical research on crime with a focus on crime and its different applications, especially with the integration of social medial information. A variety of research questions, including the conventional strategy for data mining in crime prediction, the applied techniques, predictive performance, and model validation processes, among others. It was discovered that the following types of inferences: (1) violent crime (the majority of articles), (2) crime rate, (3) criminal category, (4) murder, and (5) theft and robbery were investigated. In terms of data mining methods, authors often offered traditional machine learning techniques and, less frequently, deep learning approaches. It was observed from the systematic review that the performance of most of the crime prediction models was inconsistent and that the performance varied sometimes according to the dataset and the pre-processing applied. Various performance evaluation metrics were used, with prediction accuracy, precision, and recall ranking as the top three. In the end, the train-test split was the most common way to test a model, but cross-validation and bootstrapping were also used.

In addition, crucial features of research investigations were presented ambiguously or not at all. Regarding the last point, we suggest offering the following information: study area, size, sampling period, months, type, sample, feature engineering, feature selection, class imbalance problem data, exploratory analysis, and temporal unit (total of 10 items). Also needed are a proposed method, a best-suggested method, a baseline method, evaluation metrics, and a validation method.

It was also discovered that only a few studies investigated the use of Twitter information, and those that did either extracted the information with specific keywords, which could limit access to some hidden information, or did not provide an hourly analysis.

CHAPTER 3 : MODEL DEFINITION

3.1 Introduction

This chapter of the research introduces the reader to the models that will be investigated in more detail in subsequent sections, with a focus on their ubiquity and usefulness in data mining and their potential for use in criminal-prediction applications. They have been selected because they were the most common machine learning models found in the previous studies on crime prediction using a data mining approach. Because of this, the next section will give a short summary of the pros and cons of each model.

3.2 Machine Learning Classifiers

The following machine learning models will be briefly discussed: Decision Tree (DT), Random Tree (RF), Gradient Boosting Tree (GB), XGBoost, k-nearest neighbors (KNN), Support Vector Classifier (SVC), Neural Network (NN), and Naive Bayes classifier (NB).

3.2.1 Decision Tree

A non-parametric supervised learning technique, decision trees may be used for classification and regression. Decision tree learning employs a greedy search to partition a tree optimally. After initial splitting, the method iterates top-down until all or most records fall into specified categories. Complexity of decision tree affects reliability of data point grouping. Smaller trees make it easier to get single-class leaf nodes. As a tree grows, it gets difficult to maintain its purity, leading to too little information in a subtree. Data fragmentation causes overfitting. Decision trees favour smaller trees because of Occam's Razor, which asserts that "entities should not be multiplied beyond necessity." The simplest explanation is frequently the best; thus, decision trees should not add complexity until essential. Pruning removes unimportant forks to avoid overfitting. Cross-validation evaluates the model's fit. Using a random forest approach to combine numerous decision trees into an ensemble improves classification accuracy, particularly when the trees are independent(Jadhav and Channe, 2016).

3.2.1.1 Advantages of decision tree classifier

- Easy to interpret
- Handles both categorical and continuous data well.
- Works well on a large dataset.
- Not sensitive to outliers.
- Non-parametric in nature.
- Normalization or scaling of data is not needed.
- Handling missing values: No considerable impact of missing values.
- Easy visualization
- Automatic Feature selection: Irrelevant features won’t affect decision trees.

3.2.1.2 Limitation of decision tree

- These are prone to overfitting.
- It can be quite large, thus making pruning necessary.
- It cannot guarantee optimal trees.
- It gives low prediction accuracy for a dataset as compared to other machine learning algorithms.
- Calculations can become complex when there are many class variables.
- High Variance (Model is going to change quickly with a change in training data) (Gupta et al., 2017).

3.2.2 Random Forest

As a predictive modelling and machine learning approach, Random Forest (RF) is a collection of supervised learning algorithms that may be used for both classification and regression. The optimum output is determined by averaging the predictions of many decision trees and selecting the mode (most often occurring value) of the classes or the mean forecast. The dataset is divided into a training set and a test set for RF to function. After that, choose several training samples at random. The next step is to use the decision tree to find the optimal split of each sample into two daughters. Finally, after voting on all of the predictions, the outcome with the most votes are chosen.

Random Forest's primary hyperparameters are adjusted to either improve the model's prediction ability or boost its speed. Increasing the number of trees improves performance and provides more reliable forecasts, but at the cost of more processing time. In addition, the algorithm's speed may be enhanced by using a maximum number of features in conjunction with a minimum number of leaves that are required to divide internal nodes. As soon as the training phase is complete, the model may be employed on a dataset that was not used during training. This method enables predictions to be estimated and then compared against norms.(Tarsha Kurdi et al., 2021)

3.2.2.1 Advantages of Random Forest classifier

- Robust to outliers.
- Works well with non-linear data.
- Lower risk of overfitting.
- Runs efficiently on a large dataset.
- Better accuracy than other classification algorithms.

3.2.2.2 Limitations of Random Forest classifier

- Random forests are found to be biased while dealing with categorical variables.
- Slow Training.
- Not suitable for linear methods with a lot of sparse features (Speiser et al., 2019).

3.2.3 Gradient Boosting

Gradient boosting merges ineffective trees. Repeated training helps trees perform somewhat better than chance. Similar to boosting, gradient boosting classifies data (Chen et al., 2020). Gradient boosting overfits if the iterative approach is not regularised. When the pseudo-residuals are zero after the following iteration, the method finishes (such as quadratic loss). AGB is controlled by many regularisation factors. Shrink gradient decent steps to regularise gradient boosting. Imposing model complexity limits improves regularisation. This strategy limits a decision tree's depth or the number of occurrences required to divide a node. In gradient boosting, the default values for these parameters restrict tree expressiveness. Randomly selecting the base learners may enhance gradient boosting's generalisation. A random sampling approach is one example. Final gradient boosting settings tested:

• The rate of shrinkage (learning rate) vs the rate of learning.

• The deepest tree (max depth): random forest trees. Random sample sizes are calculated using fractional subsample sizes. The random forest replaces previous instances. Maximum features used to determine optimum split is like a random forest. Min samples split indicates the minimum number of samples required to divide a tree node (Adler and Painsky, 2022).

3.2.4 XGBoost

XGBoost is a decision tree ensemble with gradient boosting. Similar to gradient boosting, XGBoost adds a loss function to the goal function. Since XGBoost uses decision trees as its base classifier, the loss function modifies tree complexity. This loss function may be used to pre-prune decision trees. Higher values provide simpler trees. Parameter specifies minimum loss reduction gain needed to divide an internal node. Shrinkage reduces the additive expansion step size in XGBoost. Finally, tree depth may be used to limit complexity. Reducing tree complexity speeds up model training and saves data. XGBoost uses randomness to delay training and speed outcomes. Random subsamples for training individual trees and column subsampling at the tree and node levels provide unpredictability to XGBoost. It employs strategies that speed up decision tree training yet do not affect ensemble accuracy. It also optimises the time-consuming process of finding the optimal split in decision tree building methods. To find the ideal split for each node, a linear scan of all sorted characteristics must be done. XGBoost employs a compressed column-based structure with pre-sorted data to avoid sorting in every node. This decreases how often each attribute must be sorted. This columnar data format allows simultaneous partitioning for all important attributes. XGBoost uses an approach based on data percentiles, in which just a sample of candidate splits is reviewed and their benefit is computed using aggregated statistics. This is analogous to data subsampling at CART nodes. XGBoost uses a sparsity-aware technique to delete missing values from split candidate loss gain computations. In this study, XGBoost parameters were changed. • Shrinkage-to-learning-rate ratio (learning rate). Maximum tree depth equals minimum loss reduction (gamma). Maximum tree depth (max depth) and column sample level.

• Subsample rate: data are obtained without duplicating previous ones (Wang and Ni, 2019).

3.2.4.1 Advantages of XGBOOST classifier

- Less feature engineering required (No need for scaling, normalizing data, can also handle missing values well)
- Feature importance can be found out (it output importance of each feature, and can be used for feature selection)
- Fast to interpret
- Outliers have minimal impact.
- Handles large-sized datasets well.
- Good Execution speed
- Good model performance (wins most of the Kaggle competitions)
- Less prone to overfitting

3.2.4.2 Limitations of XGBOOT Classifier

- Difficult interpretation, visualization tough
- Overfitting is possible if parameters are not tuned properly.
- Harder to tune as there are too many hyperparameters (Wu et al., 2021).

3.2.5 K-nearest neighbors (KNN)

KNN is an instance-based learning method used in pattern recognition to classify objects based on their closest training examples in the feature space. A unit is assigned to the class with the highest frequency among its k closest neighbors (Figure 2), where k is a positive integer higher than zero (Hmeidi et al., 2008). The KNN method determines the classification of a new test feature vector based on the classes of its k-nearest neighbors.

KNN is immune to noisy training data and is ideal for massive training data. For this procedure, the value of the parameter number of nearest neighbors (K) and the kind of distance to be used must be given. Due to the need to calculate the distance between each test point and every training sample, the computation time may be lengthy. In addition, when the number of variables rises, calculation time gets much slower (Statnikov et al., 2008). There is no need to build a model, tweak a lot of parameters, or make additional assumptions. KNN is a simple, versatile, and easy-to-implement supervised classifier that can be employed for classification, regression, and search problems. The system believes that similar items are found nearby. In other words, similar things are clustered together, as the proverb "birds of a feather flock together" suggests. This premise must be true for the KNN approach to be beneficial (Porwal et al., 2004).

Due to the fact that it grows much slower as the number of inputs increases, KNN is inappropriate for applications requiring rapid prediction. In addition, faster algorithms may provide more accurate classification and regression outcomes. KNN may still be advantageous for solving issues whose solutions depend on finding related things, provided there are sufficient computer resources to analyse the data and provide predictions rapidly. (Halvani et al., 2013).

To discover the best K for a dataset, the KNN algorithm is run numerous times with different values of K. The K that makes the fewest mistakes while keeping the model's ability to make accurate predictions based on new information is chosen.(Chen et al., 2013a). Depending on the situation, one approach may be selected for determining distance. Straight-line distance, or Euclidean distance, is a frequent alternative (Chan and Paelinckx, 2008).

As K nears 1, projections become less stable. Inversely, as K grows, projections become more stable due to majority voting/averaging and are more likely to be right. Errors eventually increase. K has been exceeded at this moment. K is commonly an odd number to offer a tiebreaker when a majority vote is needed, such as when picking the mode in a classification problem (Vincenzi et al., 2011). KNN is useful for classification, regression, and search. It helps solve problems that require locating similar objects.

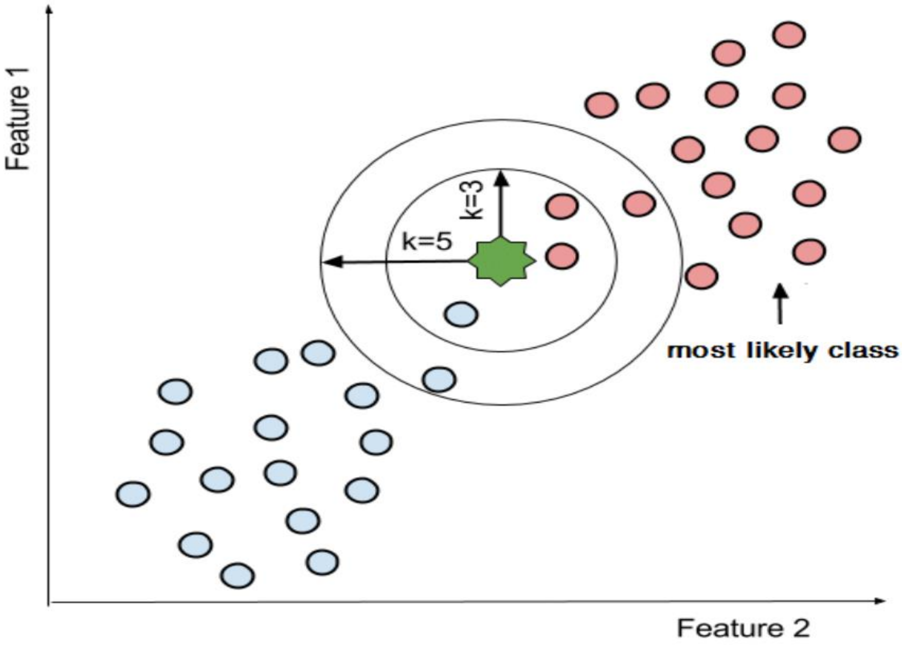


Figure 3-1: Two-class KNN classification is a common scenario.

Referenced from: https://www.researchgate.net/figure/A-typical-example-of-a-KNN-classification-for-a-two-class-problem-ie-the-pink-and_fig2_322358139

3.2.5.1 Advantages of K-nearest neighbors Classifier

- Simple to understand and implement
- No assumption about data (In the case of linear regression dependent variable and independent variables are assumed to be linearly related, and for Naïve Bayes features are assumed to be independent of each other., but KNN makes no assumptions about data)
- Constantly evolving model: When it is exposed to new data, it changes to accommodate the new data points.
- Multi-class problems can also be solved.
- One Hyper Parameter: KNN might take some time while selecting the first hyperparameter but after that rest of the parameters are aligned to it (Singh et al., 2017).

3.2.5.2 Limitations K-nearest neighbors Classifier

- Slow for large datasets.
- Does not work very well on datasets with a large number of features.
- Scaling of data absolute must.
- Does not work well on Imbalanced data. So before using KNN either under sample majority class or oversample minority class and have a balanced dataset.
- Sensitive to outliers.
- Cannot deal well with missing values

3.2.6 Support vector Machine (SVM)

SVMs are supervised learning models with related learning algorithms that analyse data for classification and regression analysis. Cortes and Vapnik's (1995) SVM method attempt to locate the best hyperplane in n-dimensional classification space with the greatest margin between classes (Figure 5).

The SVM approach is often stated to provide better results than other classifiers (Abedi et al., 2012), while it has been suggested that the major reason to use an SVM instead is because the issue may not be linearly separable (Sluiter and Pebesma, 2010). In such situation, an SVM with a non-linear kernel, such as the Radial Basis Function (RBF), might be appropriate. Another reason to employ SVMs is if you are working in a high-dimensional space. SVMs, for example, have been shown to perform better for text classification, however training takes a long time (Boateng et al., 2020).

Although the SVM classifier is optimised for binary classification, it is not suggested when there are many training examples since it is a kernel-based supervised learning approach that divides the data into three or more classes. A kernel function mapping process is used to make the training set more similar to a linearly separable data set. The goal of mapping is to add more dimensions to the data set, and it accomplishes this goal effectively via the use of a kernel function. Linear, RBF, quadratic, Multilayer Perceptron kernel, and polynomial kernels are some of the most frequent types of kernel functions.

Results for linearly separable data sets are optimal for the linear kernel function, whereas results for non-linear data sets are optimal for the RBF kernel function. Time spent training the SVM using the linear kernel function is much shorter than that required using the RBF kernel function. (Boateng et al., 2020) say that overfitting is also less likely with the linear kernel function than with the RBF kernel function.

The regularisation parameter C, also called the box constraint, and the kernel parameter, sometimes called the scaling factor, are critical to the SVM classifier's performance. They constitute what is called the hyperplane parameter. While in its training phase, SVM creates a model, maps the decision boundary for each class, and defines the hyperplane that divides them. It is possible to improve classification accuracy by expanding the hyperplane margin, hence increasing the distance between classes. It has been shown that SVMs may be utilised to efficiently conduct non-linear classification as well.

Text and hypertext categorization, image detection, verification, and recognition, speech recognition, bankruptcy prediction, remote sensing image analysis, time series forecasting, information and image retrieval, information security, biology (including protein classification in bioinformatics), and chemistry are just some of the many areas where SVMs have been successfully applied (Boateng et al., 2020).

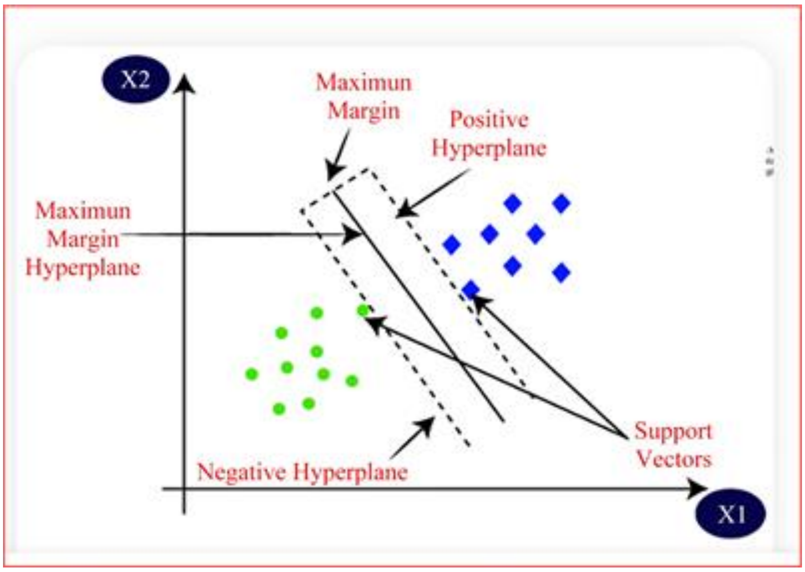


Figure 3-2: A 2-dimensional depiction of the Support Vector Machine (SVM) technique.

Referenced from: https://www.scirp.org/journal/paperinformation.aspx?paperid=104256#return33

3.2.6.1 Advantages of Support Vector Classifier

- Performs well in Higher dimension
- Best algorithm when classes are separable
- Outliers have less impact.
- SVM is suited for extreme case binary classification.

3.2.6.2 Limitations of Support Vector Classifier

- For a larger dataset, it requires a large amount of time to process.
- Does not perform well in case of overlapped classes.
- That will allow for sufficient generalization performance.
- Selecting the appropriate kernel function can be tricky (Anguita et al., 2010).

3.2.7 Naïve Bayes

It is a model of machine learning based on probabilities, and it's used to tell one type of object from another by analysing their shared and unique characteristics. Bayes' theorem forms the backbone of the classifier.

$$P(A/B) = \frac{P(A/B)P(A)}{P(A)} \qquad 5$$

The likelihood of event A given event B can be calculated using Bayes' theorem. A is a hypothesis, and B is the supporting evidence. All of the predictors/features are assumed to be unrelated to one another. That is, having one trait does not negate the usefulness of the other and that is why it is labelled naïve (Singh et al., 2017).

3.2.8 Types of Naive Bayes Classifier

3.2.8.1 Multinomial Naive Bayes:

Its primary application is in solving the problem of categorising documents. The word count in the document serves as the classifier's features/predictors (Abbas et al., 2019).

3.2.8.2 Bernoulli Naive Bayes:

A naive Bayes approach with Boolean predictors; it's like multinomial naive Bayes, but simpler. Class prediction parameters, such as whether a given word appears in the text, can only have the values yes or no (Singh et al., 2020).

3.2.8.3 Gaussian Naive Bayes:

When the predictors have a continuous value rather than being discrete, the values are assumed to be drawn at random from a normal distribution (Jahromi and Taheri, 2017).

3.2.8.4 Advantages of Naïve Bayes Classifier

- It is very fast and can be used in real-time.
- It is scalable with large datasets
- It is insensitive to irrelevant features.
- Multiclass prediction is effectively done in Naive Bayes
- Good performance with high dimensional data

3.2.8.5 Limitations of Naïve Bayes Classifier

- Independence of features does not hold
- Probability outputs from predicted probability are not to be taken too seriously.
- Training data should represent the population well

3.2.9 Artificial neural networks

Neural networks are a model developed by scientists to process information in a way that is similar to how biological nervous systems, like the brain, process information. One of the most notable aspects of this new paradigm is the new data processing architecture. It consists of numerous neurons (processing elements) that are highly interconnected and work together to solve problems. Learning occurs in neural networks in much the same way that it occurs in humans. In other words, neural networks are capable of acquiring knowledge, generalising to new situations, and making decisions based on that knowledge.

Most neural networks have an input layer, a hidden layer, and an output layer. This class of network is formally known as a multilayer perceptron (MLP). The "input" level connects to the "hidden" level, which connects to the "output" level. During training, a network is exposed to a vast amount of information, all of which is reflected in the activity of its input layers. The collective behaviour of the hidden units is determined by analysing the input units' actions and the connection weights between them. The behaviour of the visible units is determined by the actions of the hidden units and the relative importance of the hidden and output units. Figure 1 shows the input, output, and a single hidden layer.

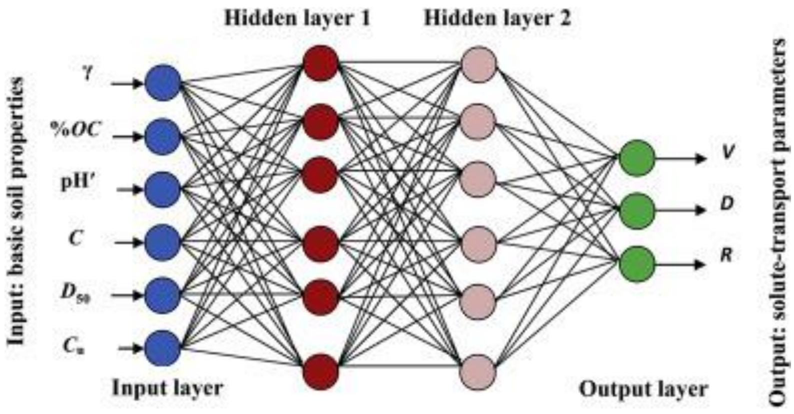


Figure 3-3: Figure Outlining the Structure of an Artificial Neural Network.

Referenced from: https://www.sciencedirect.com/science/article/abs/pii/S0269749119312631

Multilayer networks can find the discrepancy between each predicted output and its corresponding target output by first calculating the expected outputs for each training example. The network then modifies the error so that the next time the training sample is presented, it produces fewer mistakes. This allows the algorithm to learn what features of the inputs are most important for achieving mastery of the goal function (Zhang et al., 2021).

3.2.9.1 Advantages of Neural Network Classifier

- Neural networks are flexible and can be used for both regression and classification problems. Any data which can be made numeric can be used in the model, as a neural network is a mathematical model with approximation functions.
- Neural networks are good to model with nonlinear data with a large number of inputs; for example, images. It is reliable in an approach to tasks involving many features. It works by splitting the problem of classification into a layered network of simpler elements.
- Once trained, the predictions are pretty fast.
- Neural networks can be trained with any number of inputs and layers.
- Neural networks work best with more data points.

3.2.9.2 Limitations of Neural Network Classifier

- Because of their complex nature, neural networks make it hard to deduce the relative importance of each external factor in determining the outcomes of experiments.
- The training process takes a long time and a lot of computational resources when using conventional central processing units.
- Training data is crucial for neural networks. This causes issues with over-fitting and generalisation. The mode may be adjusted based on the information it has learned from its training.

3.3 Evaluation Metrics

Accuracy: the proportion of the total number of correct predictions.

Accuracy = $\frac{\boldsymbol{Number\ of\ Correct\ predictions}}{\boldsymbol{Total\ number\ of\ predictions\ made}}$ 6

True Positive Rate (Sensitivity): True Positive Rate is defined as *TP/ (FN+TP)*. True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, concerning all positive data points.

True Positive Rate = $\frac{\boldsymbol{True\ Positive}}{\boldsymbol{False\ \ Negative\ +False\ Poitive}}$ 7

True Negative Rate (Specificity): True Negative Rate is defined as *TN / (FP+TN)*. False Positive Rate corresponds to the proportion of negative data points that are correctly considered as negative, concerning all negative data points.

True Negative = $\frac{\boldsymbol{True\ Negative}}{\boldsymbol{True\ Negative\ +False\ Positive}}$ 8

False Positive Rate: False Positive Rate is defined as *FP / (FP+TN)*. False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered positive, concerning all negative data points.

False Positive Rate = $\frac{\boldsymbol{False\ Positive}}{\boldsymbol{True\ Negative\ +False\ Positive}}$ 9

Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Postive} + \textit{False Positive}} \qquad 10$$

Recall: It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Postive} + \textit{False Negatives}} \qquad 11$$

3.4 Model Performance Improvement

3.4.1 Hyperparameter Tuning

Hyperparameters may be likened to an algorithm's settings that can be altered to achieve optimal performance. As opposed to the slope and intercept in linear regression, which are learned during training, the data scientist specifies the hyperparameters before training. In the process of model tweaking, machine learning transitions from a science to an engineering approach based on trial and error. This is owing to the fact that it is often difficult to anticipate in advance which hyperparameters will provide the best results.

Since hyperparameter tuning is more empirical than theoretical, the best way to determine the optimal settings is to experiment with a wide range of values and compare the results across several models. However, if models are evaluated only on the training set, overfitting may arise in machine learning.

It is conceivable that training-data-specific model optimization cannot generalise to test-set data that is newer than the training data. When a model performs very well on the training set but badly on the test set, this is known as overfitting. This makes the model incapable of meeting new difficulties.

3.4.2 Cross Validation (CV)

Since k-fold cross-validation (CV) is the most popular approach for illustrating CV, we will utilise it here. In machine learning problems, the dataset is often separated into a training set and a testing set. In K-Fold CV, however, the training set is partitioned into a hierarchy of subsets, or folds. The model is then fitted K times iteratively, using data learned on K-1 folds and tested on the Kth fold at each iteration. Consider the process for fitting models with K set to 5. For this first cycle, it will be taught on the first four folds and our progress will be evaluated with the fifth. This is repeated on the first, second, third, and fifth folds of the second try, and the results are reviewed on the fourth. The procedure is repeated three times to assess performance. As the last stage of training, the model's validation metrics are determined by averaging the results from all the folds. To fine-tune hyperparameters, the complete K-Fold CV operation is done several times with different model settings. Following an evaluation of each model, we would choose the most promising alternative. After training the best model on the complete training set, we test it on the testing set. The training data would be divided into K folds and trained and evaluated K times, one for each conceivable hyperparameter combination. Using a 5-fold CV and 10 distinct hyperparameter configurations yields 50 training iterations.

CHAPTER 4 : RESEARCH METHODOLOGY

4.1 Introduction

Data mining methodologies were developed to provide a more complete picture of the knowledge discovery process rather than applying only statistical or machine learning techniques. Their purpose is to establish a general workflow that begins with the identification of pertinent questions and continues with the focused processing of raw, frequently unstructured data in order to discover new information. Even though model construction is a crucial aspect of any knowledge discovery process, selecting and deploying models takes only a fraction of the total time involved. In contrast, data collection, manual cleaning, and preparation can require a considerable amount of time.

Most data science research has focused on the technical capabilities required for data science while ignoring the difficulty of managing data science projects, and only few researchers have analysed the various techniques used by data science teams (Saltz and Hotz, 2020). Consequently, this research employed cross-industry standard procedure for data mining (CRISP-DM) to investigate the possibility of improving crime prediction through exploration of the social media data (Twitter) by merging Twitter sentiment polarity with historical crime records. The CRISP-DM is a framework for transforming business challenges into data mining tasks and executing data mining projects independent of the application domain and technology employed (Wirth and Hipp, 2000). It is an industry-wide implementation of the Knowledge Discovery (KD) methodology outlined in (Brachman and Anand, 1994).

Figure 4-1, illustrates the relationships between the six stages of the CRISP-DM process model. Every data mining project begins with a description of the project's objectives, which is covered in the "Business Understanding" phase. Utilizing predictive analytics to improve the runtime and efficiency of machines is a typical corporate objective in crime prediction research. This objective is then translated into a specific data mining problem, such as defining the appropriate machine learning technique and quantifying the variables' contribution to the model. During the "Data Understanding" phase, hypotheses for hidden information relating the data to the project objective are generated based on experience and qualifying assumptions. In the case of crime prediction, it would be appropriate to retrain the model using various sampling techniques in order to identify the optimal hyperparameter that would produce the most successful models using all of the primary features or variables available in the dataset. As depicted in the figure, here is a quick overview of the various steps of the standard method for data science that employs CRISP-DM.

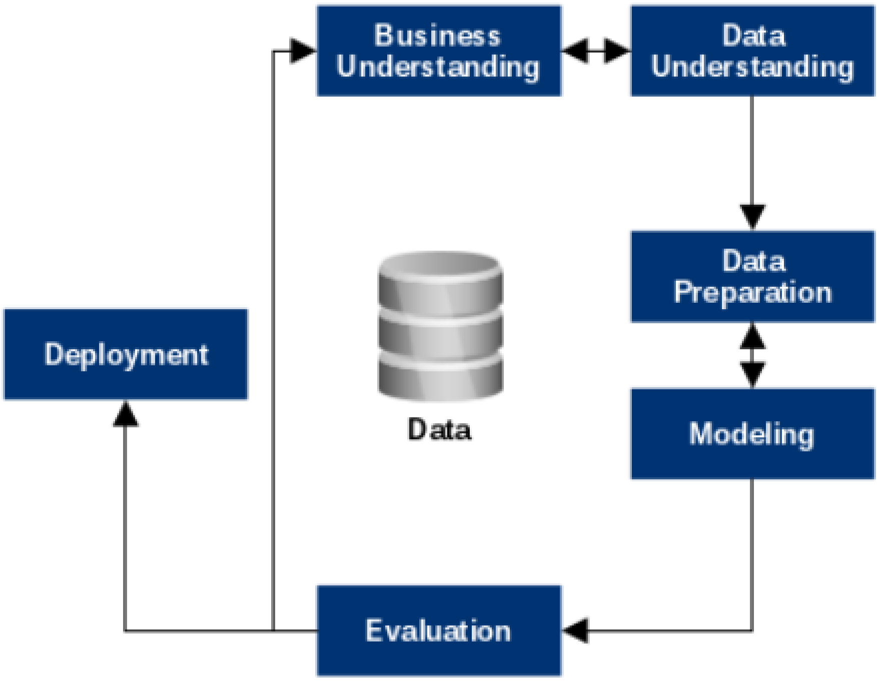


Figure 4-1: The CRISP-DM process. (Source: https://towardsdatascience.com/data-science-career-reflection-based-on-crisp-dm-process-model-aedd8542b019)

Business Understanding: Existing research and literature should be evaluated to determine the available and necessary resources. Identifying the data mining objective is one of the most crucial aspects of this step. The type of data mining (such as classification) and the data mining key performance indicators must be stated first (like accuracy, precision, or recall). A required project plan should be created.

Data understanding: The steps involved in this process include accumulating information from various resources, processing it to extract useful information, checking for errors, and ensuring its quality. This phase of the research process includes not only a thorough description of the data but also an exploratory analysis of the statistical patterns and relationships within the data used in the study.

Data preparation: Establishing inclusion and exclusion criteria should precede data selection. Poor data quality can be remedied by cleaning data. Derived characteristics must be constructed depending on the employed model (specified in the first phase). For each of these processes, there are available and model-dependent alternative methods.

Modelling: This process includes selecting a modelling strategy, developing a test case, and developing the model. All techniques for data mining are relevant. In general, the business challenge and the facts impact the decision. What is more important is how to justify the choice. To generate the model, some parameters must be set. Examining the model against evaluation criteria and selecting the best ones for assessment is appropriate.

Evaluation: During the evaluation phase, the outcomes are reviewed against the established business objectives. Consequently, the results must be analysed and additional steps must be planned. Another argument is that the entire procedure should be scrutinised. At this point in the project, one or more models with high-quality data analysis characteristics have been constructed. Before proceeding with the final deployment of the model, the model(s) are thoroughly evaluated, and the actions used to develop the model are examined to ensure that it meets the business objectives. A primary purpose is to assess if there are any significant business issues that have not been adequately addressed. At the conclusion of this step, a decision should be made about the application of the data mining results.

Deployment: Generally, the building of the model is not the conclusion of the project. Typically, the acquired knowledge must be structured and presented so that the consumer may utilise it. The deployment phase might be as easy as generating a report or as sophisticated as implementing a repeatable data mining process, depending on the requirements. In many instances, the user, and not the data analyst, will do the deployment processes. In any event, it is essential to understand in advance what steps must be taken to really utilise the produced models. The deployment step is broadly described in the user guide. The deployment stage includes deployment planning, monitoring, and maintenance. Therefore, the rest of this chapter will follow the phases recommended in the CRISP-DM framework.

4.2 Data understanding

This section gives a comprehensive explanation of the collected datasets and a thorough exploratory study of the historical crime dataset.

4.3 Study Area

The UK has been chosen for this research because it is not widely analysed in the field of crime prediction and the few studies that employed the UK dataset in their dataset have not used the appropriate dataset and none of the previous studies have utilised the Greater Manchester dataset, despite its population density.

Greater Manchester in the Northwest England was selected as the area of study. As one of the ten largest cities in England, it possesses many of the metropolitan topographical features and amenities that one would anticipate from a typical metropolis. This includes a bustling retail and entertainment sector in the city centre, an abundance of economic and commercial services, a mainline train station, a metro system, and around four notable colleges. The district also encompasses northern rural regions. The population of the research area is approximately 2,770,436 as of 2022.

4.4 Research Design

The Figure 4-2, below illustrates the steps involve to accomplish our research aims and objectives.

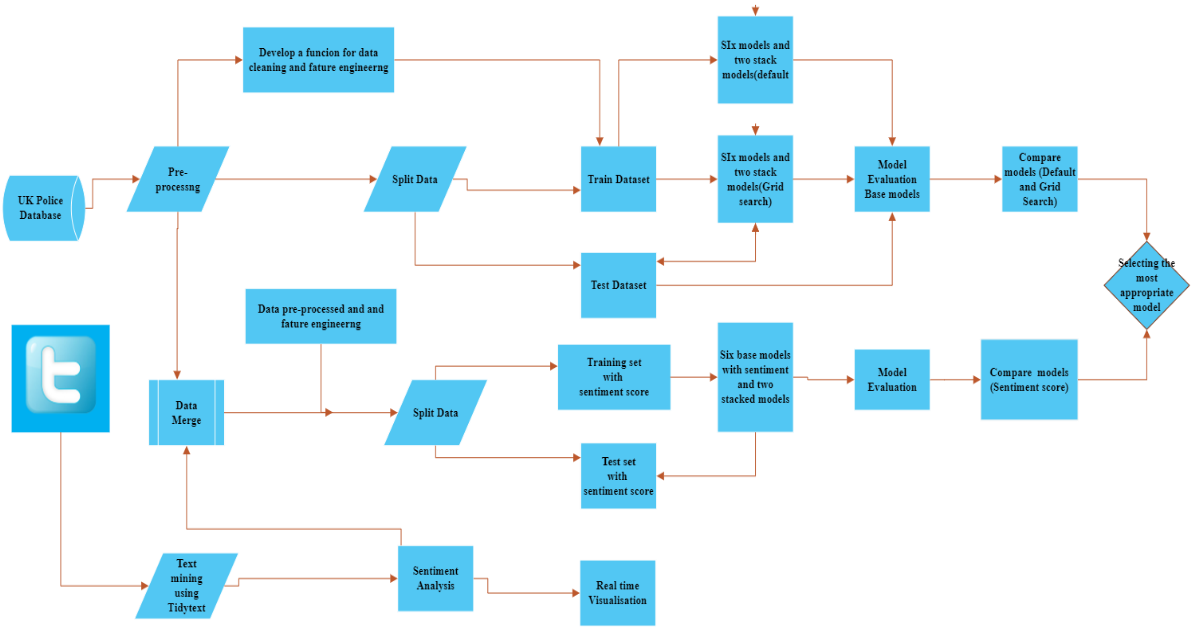

Develop a funcion for data cleaning and fature engineerng
UK Police Database
Pre-processng
Split Data
Train Dataset
SIx models and two stack models(default
SIx models and two stack models(Grid search)
Model Evaluation Base models
Compare models (Default and Grid Search)
Test Dataset
Selecting the most appropriate model
Data pre-processed and and fature engineerng
Data Merge
Split Data
Training set with sentiment score
Six base models with sentiment and two stacked models
Model Evaluation
Compare models (Sentiment score)
Test set with sentiment score
Text mining using Tidytext
Sentiment Analysis
Real time Visualisation

Figure 4-2: Research Process Framework

4.4.1 Data collection and Description

Two distinct data sources were utilised for this investigation. The first and most important dataset utilised in this study is the stop and search historical crime data taken from UK police website (DATA.POLICE.UK), while the second dataset was obtained from Twitter.

4.4.2 Historical crime records description

The official crime records issued by the UK Police from 2016 to 2019 were used as the source of crime occurrences for this study. This consists of 25,764 records of criminal occurrences organised into 10 distinct categories. This data was utilised for this work since it offers relevant information, such as the demographics information of offenders. This will also increase the credibility of the work, as the information is not gathered from external sources. With Spatial-temporal and demographic data acquired from criminals, this project tried to anticipate several types of criminal activity. Also, for this research sentiment polarity generated from Twitter data (tweets) was incorporated to demonstrate how social media, without duplication of information, might aid in the identification of different types of crime. Finally, each incident's category is stated. Table (4.1) displays examples of the list of offenses considered and the number of occurrences for each category from 2015 to 2019.

Table 4-1 provides a quick statistical summary of the dataset's properties. The "TYPE" variable across all of our samples reveals whether the search was conducted on a human or a vehicle. The "Date" element provides precise timestamps for the criminal occurrence. The "longitude" and "latitude" features describe the crime scene in relation to a certain reference point or place of interest (e.g., shopping area, supermarket, parking lot). The attribute "Part of the policing operation" is recorded as false or true, although the column "Policing operation" is always empty. The Gender feature shows whether the offender is male or female, whereas empty rows are supposed to represent those who did not choose to reveal their gender. The "Age range" attribute anonymizes the offender's age so that it is not disclosed to the general public. A similar situation is found for the "Self-defined ethnicity" attribute, but the "Officer-defined ethnicity" attribute makes it easier to assign a certain ethnicity to the offender. The "Object of search" can be compared to a sort of crime, but this comparison can only be proven by the "Object of the search-related outcome." The type of crime is one of the six categories used by the frequency police in the United Kingdom data utilised for this study. The information is disclosed at the time of offense, with each row representing a single event.

The crime statistics were then categorised into four categories: Offensive weapons (Anything to threaten or harm another person, Articles for use in criminal damage, Offensive weapons), drugs (Controlled drugs, Psychoactive substances), theft (Article used for theft, Stolen goods), and others (Firearms, Fireworks, goods on which duty has not been paid, Seals or hunting equipment). These crime types were selected because they are regularly analysed by police and crime reduction practitioners; hence, the implications of the research would be accessible and more easily applicable to policing and crime reduction practice. Table 4-2 enumerated, for each kind of crime, the number of occurrences and proportion for each year. Tables 4-3, 4-4, 4-5 exhibit a portion of the variables from the historical dataset used for this inquiry when the data is not cleaned, when the data is cleaned, and when the data is merged with the sentiment polarity of tweets, respectively.

Table 4-1: Brief statistic of the historical dataset

Variable type: character							
S/N	**Features**	**Missing**	**Complete rate**	**min**	**max**	**empty**	**n_unique**
1	Type	0	1	13	25	0	3
2	Gender	8859	0.79	4	6	0	2
3	Age range	12030	0.714	5	8	0	6
4	Self-defined ethnicity	1672	0.96	13	84	0	36
5	Officer-defined ethnicity	4660	0.889	5	5	0	4
6	Legislation	1710	0.959	30	55	0	11
7	Object of search	109	0.997	8	42	0	12
8	Outcome	504	0.988	6	39	0	14
9	Outcome linked to object of search	31753	0.246	4	5	0	4
10	Removal of more than just outer clothing	1710	0.959	4	5	0	4

Table 4-2: Brief Description of the target variables (crime categories)

S/N	Outcome	Numbers	Proportion
1	Anything to threaten or harm anyone	299	0.0525
2	Article for use in theft	779	0.137
3	Articles for use in criminal damage	84	0.0148
4	Controlled drugs	2902	0.51
5	Firearms	69	0.0121
6	Fireworks	9	0.00158
7	Offensive weapons	697	0.122
8	Psychoactive substances	2	0.000351
9	Seals or hunting equipment	2	0.000351
10	Stolen goods	833	0.146

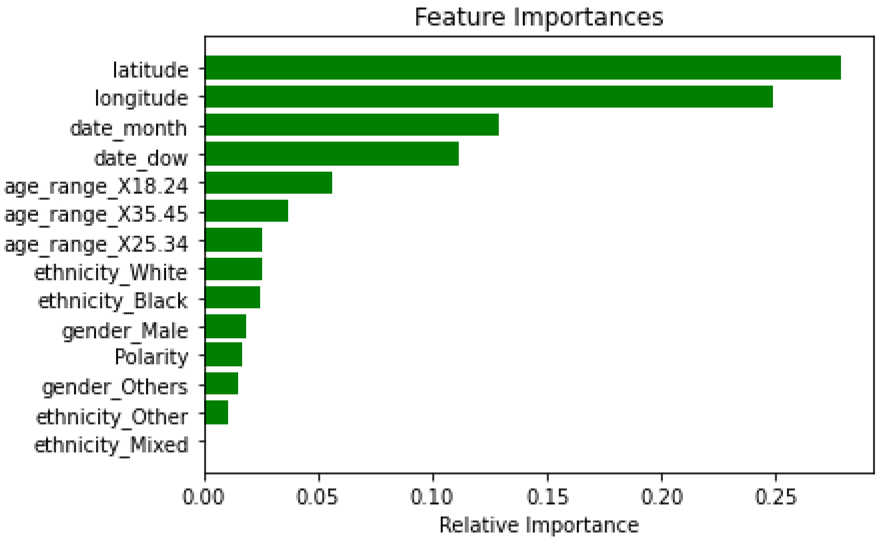


Figure 4-3: Ranking of Important features

Table 4-3 samples of Selected features from historical raw crime date

S/N	date	latitude	longitude	gender	Age range	self	Officer defined ethnicity	Object of search
1	10/09/2015 08:30	53.62467	-2.16278	Male	Oct-17	White	White	Anything to threaten or harm anyone
2	10/09/2015 08:30	53.62467	-2.16278	Male	Oct-17	Black	Black	Anything to threaten or harm anyone
3	10/09/2015 08:35	53.62467	-2.16278	Male	Oct-17	White	White	
4	11/09/2015 09:15	53.45487	-2.09396	Male	over 34	White	White	
5	15/09/2015 19:45	53.55871	-2.32392	Male	18-24	White	White	
6	17/09/2015 03:00	53.57997	-2.37873	Male	over 34	White	White	Anything to threaten or harm anyone

Table 4-4 samples of Selected features from historical cleaned crime data

S/N	date	latitude	longitude	gender	Age range	ethnicity	Object of search
1	10/09/2015 08:30	53.62467	-2.16278	Male	10 -17	White	Offensive weapons
2	10/09/2015 08:30	53.62467	-2.16278	Male	10 -17	Black	Offensive weapons
3	10/09/2015 08:35	53.62467	-2.16278	Male	10 -17	White	Offensive weapons
4	11/09/2015 09:15	53.45487	-2.09396	Male	35-45	White	drugs
5	15/09/2015 19:45	53.55871	-2.32392	Male	18-24	White	drugs
6	17/09/2015 03:00	53.57997	-2.37873	Male	35-45	White	Offensive weapons

Table 4-5: Sample of combined historical crime data with sentiment polarity

S/N	date	latitude	longitude	gender	age_range	ethnicity	object_of_search	Polarity
1	10/09/2015 08:30	53.6	-2.16	Male	Oct-17	White	Offensive weapons	1
2	10/09/2015 08:30	53.6	-2.16	Male	Oct-17	White	Offensive weapons	1
3	10/09/2015 08:30	53.6	-2.16	Male	Oct-17	White	Offensive weapons	2
4	10/09/2015 08:30	53.6	-2.16	Male	Oct-17	White	Offensive weapons	2
5	10/09/2015 08:30	53.6	-2.16	Male	Oct-17	Black	Offensive weapons	1
6	10/09/2015 08:30	53.6	-2.16	Male	Oct-17	Black	Offensive weapons	1

4.4.3 Twitter information description

Twitter was selected as a supplementary data source, and tweets sent through Manchester between 2016 and 2019 were collected from Twitter. The gathered tweets were anonymized such that only the date and text were kept for analysis. The Twitter data was subjected to sentiment analysis to determine the sentiment score. The tweet sentiment score generated was 440433, which is a record of the average hourly sentiment score of tweets within the same period. The sentiment score was used in two different ways. In one way, it was added to the historical crime data to develop a crime prediction model, and in another way, it was used to build a near-real-time shiny app to detect criminal activities on Twitter.

4.4.4 Data Exploratory Analysis

This section provides a short statistical exploration of the datasets to help in the supply of appropriate data processing, feature engineering, and solutions to class imbalance issues, as well as the answering of essential questions that may lead to a solution for crime prediction. This section also addresses the following questions: How has the crime rate varied over the last several weeks? How has the crime rate varied among various ethnic groups? And how has the crime rate varied across various age groups? The results sections included the responses to these questions.

4.4.5 Data Pre-processing

This section intends to prepare both historical crime data and Twitter data for additional research and model development. Prior to merging, the historical crime data was cleaned only for variable names, missing data imputation, target grouping, variable transformation, and feature engineering. Text mining was performed on Twitter tweets for data cleaning, and sentiment analysis was applied to the tweets to extract their sentiment score. Finally, the sentiment polarity associated with and the corresponding date column in the tweets were combined with historical crime data. One dataset would contain the polarity of Twitter sentiment, whereas the other would not. As a result, these datasets were used to train six different classifiers (Random Forests, Decision Trees, K-nearest neighbors, Support Vector Machine, Neural Network, and Nave Bayes).

4.4.5.1 Variable names cleaning, Missing Data Imputation

Gender: The missing rows in the gender column may be due to missing values or individuals who do not desire to disclose their gender. Thus, lacking information, this column is labelled "undefined."

Ethnicity: Due to the connection between the self-defined and officer-defined ethnicity columns, both self-defined and officer-defined ethnicity columns were updated to provide an ethnicity column with no missing rows.

Age range: The rows containing Oct-17 have been replaced with 10-17. It was presumed to be an Excel worksheet-related imputation mistake. The ages less than ten were classified under 10-17 since their numbers were inconsequential, and the missing rows were eliminated because if they were imputed using any technique and used to form a model, the model would be biased toward the new age group.

Longitude and Latitude: The missing rows in the location were removed to avoid misleading the model's evaluation. Because regardless of the type of imputation employed, a new location would be established, meaning that the place of the crime would be moved, for instance from Salford to Trafford.

Object of search: "Anything to threaten or hurt another person", "Articles for use in criminal damage", and "offensive weapon" were renamed violent crime, "Articles for use in theft" and "Stolen goods" were renamed theft, and "Controlled drugs" was renamed Drugs.

4.4.5.2 Twitter data Cleaning and Pre-processing

The date column of the Twitter data was converted to datetime format and the irregular characters were eliminated. The date and tweet columns were chosen to derive the sentiment score from the tweets. Each Twitter user was then matched to the tweets and sentiment polarity of each tweet. The range of polarity is from 5 to 0. In the result section, a sample of tweets with polarity and a word cloud that shows the positive and negative tweets would be shown.

4.4.6 Merging Datasets

The year, month, day, and hourly features were extracted from the date column of the two datasets so that hourly analysis of Twitter data and historical data could be performed, and the two datasets could be joined. After grouping the two datasets by year, month, day, and hour, the historical crime data and the sentiment data were combined for further analysis.

4.4.7 Feature Engineering

The target variable has a class imbalance, as indicated by a quick analysis of the dataset. Although there are numerous strategies (undersampling, oversampling, etc.) for addressing class imbalance concerns (undersampling, oversampling, and many more), the up-sampling method is employed in this research since it has been shown to have high performance with varied numbers of predictors and also because of the size of our dataset (Hossain et al., 2020, Wijenayake et al., 2018). This location was also chosen owing to the amount of historical crime data included in this inquiry. This is predicted to increase the number of the minority class to the same level as the majority class. Moreover, the categorical variables were converted into dummy variables in order to normalise the predictors and ensure that all variables were on the same scale as the model's inputs.

4.5 Model Building

This section addresses the selection of modelling approaches, the development of the test and training dataset, parameter settings, and model fittings. The optimal data splitting method for this crime prediction model was determined by comparing a variety of approaches. Eighty percent of the dataset was utilised for training, whilst the remaining twenty percent was used for testing the models' performance. Using random grid search cross validation with 5-fold cross validation, the hyperparameters were fine-tuned.

4.5.1 Building Model

Six distinct classifiers (Random Forest, XGBoost, neural network, support vector machine, k-nearest neighbors, and decision tree) were independently trained with and without hyper-parameter adjustment, using the two different datasets created earlier (one with sentiment score and one without sentiment score) Then, the separate models for models with and without sentiment scores were stacked using voting classier and logistic regression.

4.6 Evaluation

The performance of individual models with and without grid search and the stacked ensemble models for the two different datasets was evaluated using the accuracy, precision, recall, and f1-score metrics. The model with the highest performance was chosen as the final model and used to create a criminal prediction classifier pipeline ready for deployment.

4.7 Sentiment Trend Visualisation

The first and second goals of this study were met by combining historical crime records with and without emotion polarity determined from Twitter data. The final purpose of this project is to develop a visualisation tool capable of capturing users' tweets in real or near-real time. The visualisation was not limited to keyword data collection alone; it may be adjusted to stream Twitter information, but because to Twitter's limitation on the number of tweets that can be collected and to be able to gather sufficient tweets for the purpose of analysis, only tweet searches were done.

Thus, the gathered tweets were cleaned using an appropriate text mining technique before being subjected to sentiment analysis. The tweets were used to generate three separate charts. Sentiment polarity graph: this is a graph that displays Twitter user ids in relation to sentiment polarity. All tweets, regardless of polarity, were gathered for the objective of this study and to be able to uncover the hidden message underneath certain positive tweets. As a result, the graph includes the median sentiment score as well as a 95% confidence interval below and above the median sentiment score. Also, each point on the graph is intended to display the screen name and tweets. This is done so that any officer in authority may hover the mouse over any point in the graph and show the information that has been posted on Twitter. It is now up to criminology specialists to determine whether the message is harmful or not based on their experience with or knowledge with the nature of the tweets and the screen name.

The second graph depicted the tweets' approximate location. The final graph is a word cloud illustrating the frequency of bad and positive tweets, which would make it simpler for police officers or those in charge to correlate the screen name and tweets with a prospective offender or offender's location. The word cloud only displays aggregate data, not individual data.

Finally, the three graphs were integrated to form a gleaming app. This beautiful app would show the trend in sentiment polarity, the 95% lower- and upper-class bounds, and the median. In addition, each point on the chart might disclose the user's screen name and tweets. The last two graphs would show the approximate location as well as a word cloud representing the frequency of the aggregate negative and positive tweets.

The shiny app would allow for subject searches, the quantity of tweets to be returned, nation and city searches, and an estimated location. To minimise overpopulation, the maximum number of tweets set for this app is 2000, and the location is Manchester, United Kingdom, but it is flexible enough to accept other places, and the maximum number of tweets may be altered if required.

CHAPTER 5 : RESULT AND DISCUSSION

5.1 Introduction

The visualisation of the sentiment polarity trend with higher and lower confidence intervals was also developed for real-time crime detection on the Shiny app, using Twitter as a social media platform, where the upper confidence indicated the highest positive and the lower confidence indicated the lowest negative, at which a flag should be raised. The graph would indicate the tweets, the sentiment, the user's screen name, and the approximate location of the person who submitted the message.

5.2 Result of the analysis performed with only historical crime data

This section describes datasets that lack sentiment polarity data. The results revealed that males gender committed more crimes than females (85.4% vs 6.5%). Fifty-eight percent of documented crimes were committed by those aged 18 to 24. About 28.8% of offenders were between 25 and 34 years old. Those aged 35 to 45 perpetrated 9% of all offences, while those aged 10 to 17 were responsible for 6% (Figure 5-1 and 5-2).

Typically, the Friday-to-Saturday surge was when the crime rate was at its highest. However, most other sorts of crime occurred on Saturday (27%), followed by Monday (20%) and finally Sunday (17%). In addition, the prevalence of drug-related drug offences was almost same between Friday and Saturday (17%), but higher on Sunday (12%) than Monday (13%). Wednesday through Friday, however, the incidence of drug-related offences rises. On Tuesday, Wednesday, and Thursday, theft-related offences steadily increased. Additionally, there were minor increase in the rate of weapon use from Tuesday to Saturday (Figure 5-3. A). As shown in Figure 5-3.C, people of white ethnicity background were responsible for around 55% of drug crimes, 62% of crimes involving offensive weapons, over 70% of theft crimes, and over 80% of other crimes. The use of offensive weapons was the most prevalent crime among blacks, followed by drug offences, theft, and other sorts of criminal behaviour. In the Asian community, drug crime was the most prevalent offence. Other sorts of crime were more prevalent among Asians than crimes involving firearms and theft.

Except for other forms of crime, the frequency of crimes committed by people of different origins was comparable, and those of mixed origins committed these crimes rarely, if at all. People in the age range of 18–24 were the most criminalised, followed by those of 25–34. However, the distribution of crimes fluctuated daily. On Sundays, Mondays, and Wednesdays, 18-to 24-year-olds committed the most crimes. On Tuesday and Thursday, those aged 10 to 17 committed the bulk of the crimes. On Friday, the most crimes were perpetrated by people aged 35 to 45, followed by those aged 25 to 34. Moreover, Figure 5-3. D indicated that people of the male gender were about 50% more likely to commit crimes than individuals of other genders, and women were less likely to commit crimes than individuals of other genders. The majority of Sunday's offences involving offensive weapons, illegal substances, and theft were likely perpetrated by individuals aged 18 to 24 and people aged 10 to 17 (Figure 5-4). Also, it can be inferred that, with the exception of theft and obnoxious use on Saturday, the majority of drug-related and other offences were committed by those aged 25 to 34. For the age range 10-17, offensive weapon use and other forms of crime happened most often on Tuesdays and Thursdays, whereas theft and offensive weapon use occurred most frequently on Fridays for the age group 34-45.

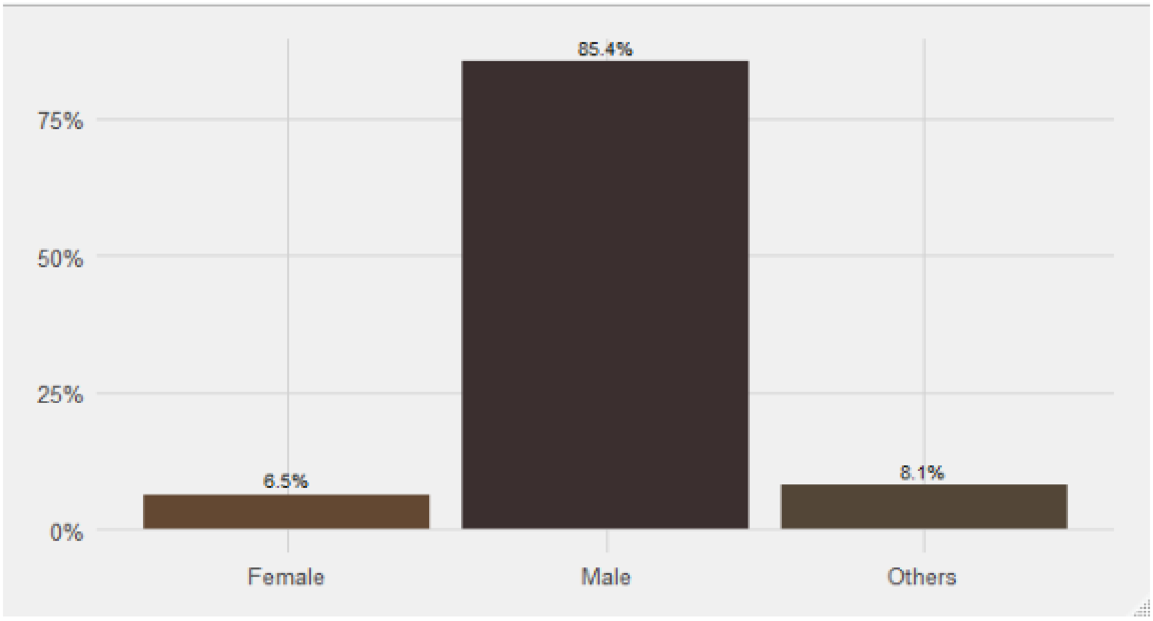


Figure 5-1: Crime percentage by gender

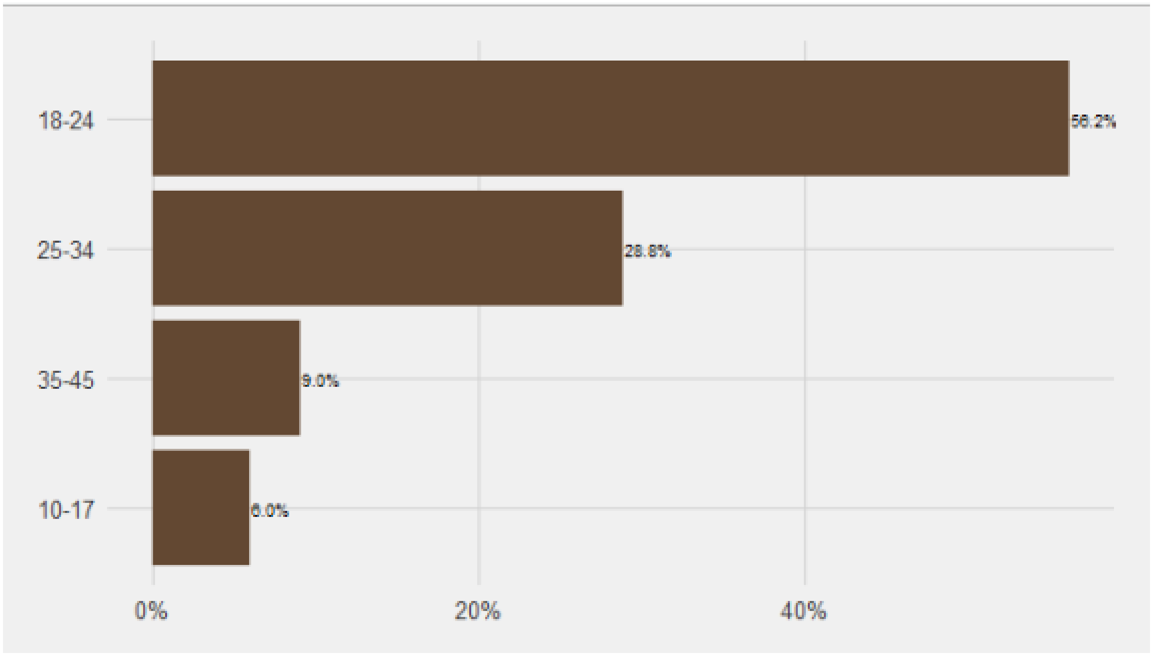


Figure 5-2: Crime percentage by age-range

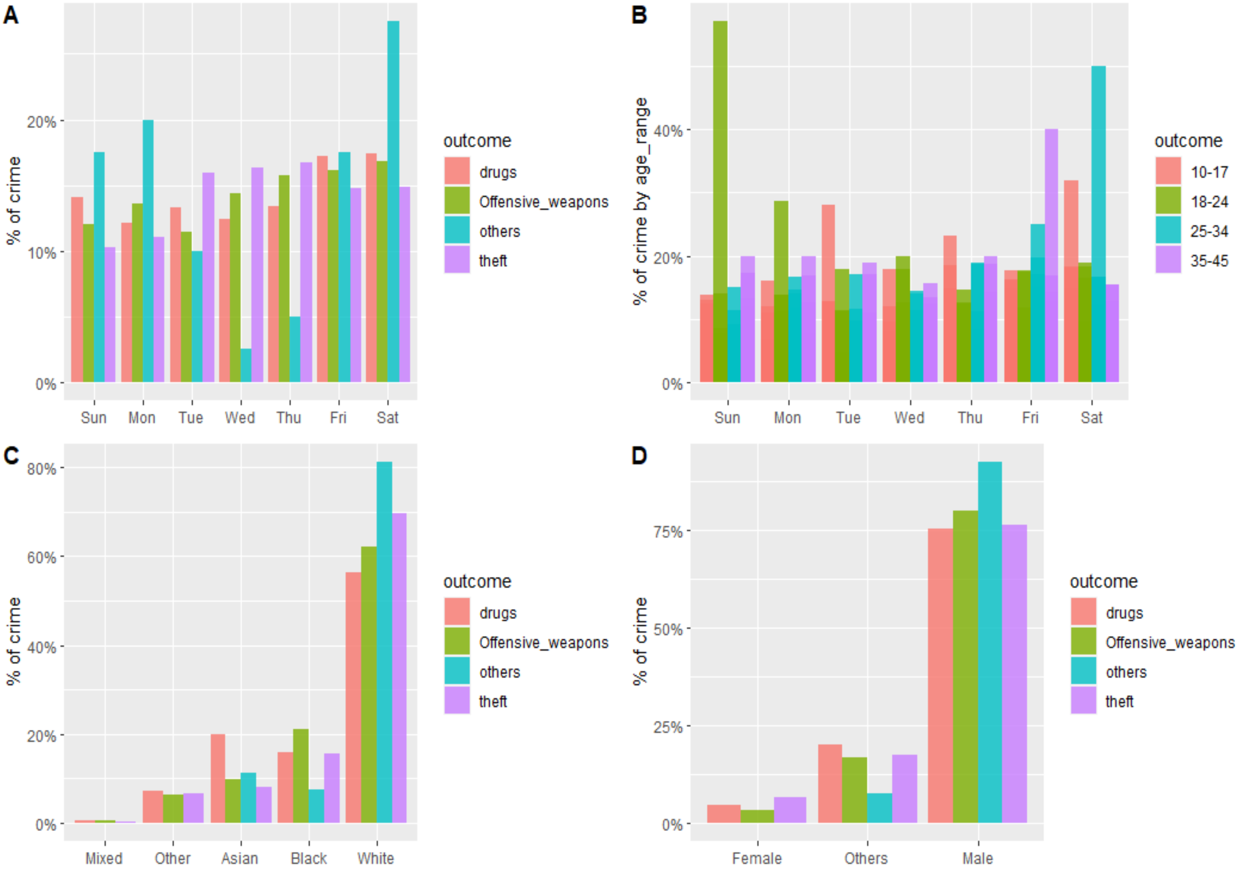


Figure 5-3: graph showing relationship between crime and various demographic factors

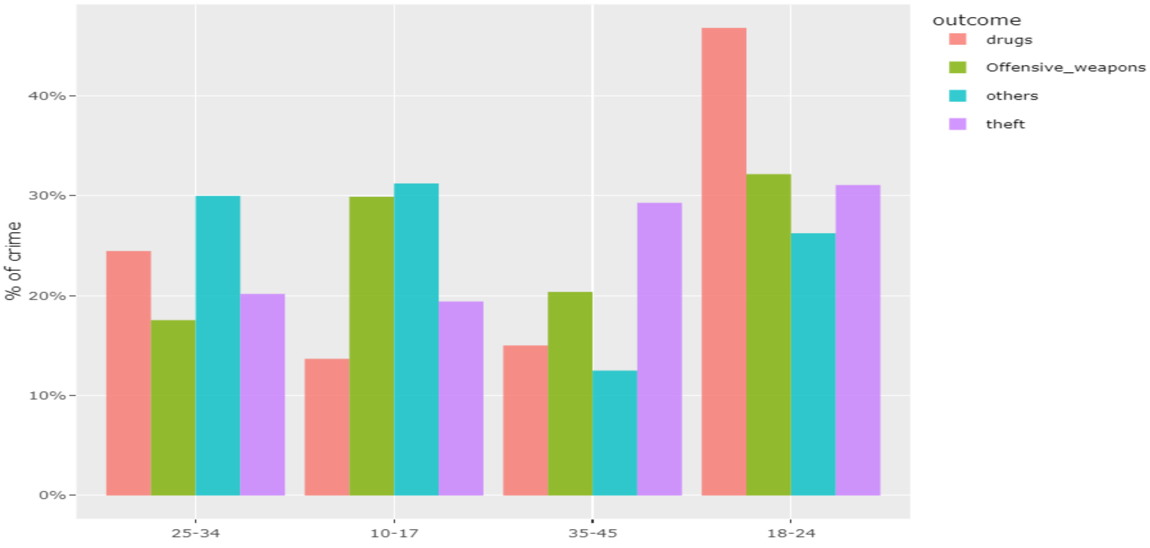


Figure 5-4: crime distribution with age range

5.2.1 Machine learning classifiers result built with only historical crime data

The result of the classification models without incorporation of sentiment polarity were presented in Table 5-1. It was demonstrated that the decision tree classifier and random forest classifier outperformed other models in terms of accuracy, precision, recall, and f1-score with a score of 100%. When the two classifiers were evaluated on the test data sets, the London dataset and Kent dataset, both decision and random forest got a perfect score across all assessment measures. On the training set, XGBoost achieved an accuracy, precision, recall, and f1-score of around 99%. On the test dataset, it scored 97% precision, 96% accuracy, 96% recall, and 96% f1-score. On the London dataset, it obtained 99% across all assessment parameters; however, on the Kent dataset, it achieved 82% accuracy, 84% recall, 84% precision, and an f1-score of 80%.

KNN achieved 81% accuracy, recall, and f1-score, and 82% precision on the training dataset. Its performance on test data was about 82% accurate, 83% precision, and 82% in terms of recall and f1-score. Except for f1-score, whose performance in London and Kent is 72% and 71%, respectively. KNN achieved an accuracy and recall of 73% and 72% for both London and Kent datasets, respectively.

The neural network's performance was much worse than the preceding three models, with an accuracy and recall of around 56%, a precision of 53%, and an f1-score of 52%. When evaluated on the test data, its performance increased, with an accuracy and recall of 64%, precision, and an f1-score of roughly 61%. Its performance on the Kent dataset was somewhat better than its performance on the London dataset, with an accuracy and recall of 67%, a precision of 62%, and an f1-score of 56%.

Accuracy and recall for the support vector classifier were 54% and 60%, respectively. It has an F1 score of 45%. On the test data, the performance of the support vector machine was 59% accuracy and recall, 47% precision, and 52% f1-score. It achieved 56% accuracy and recall, 32% precision, and 40% f1-score on the London dataset. On the Kent data set, it achieved 66% accuracy and recall, 43% recall, and an f1-score of 52%.

The performance of Nave Bayes on the UK dataset was dismal, with an accuracy and recall of around 6%, 7% f1-score and 34% precision on the training dataset. On the test dataset, its accuracy and recall further reduced by 2%. Its precision was reduced by about 50% and tits value of f1-score was reduced by about 4%. Likewise, its performance on the Kent and London and Kent was inadequate.

Table 5-1 Weighted Average performance for the models without sentiment polarity

Models	Data Split	Accuracy	Precision	Recall	f1-score
Decision Tree	training set	1	1	1	1
	Test set	1	1	1	1
	London	1	1	1	1
	Kent	1	1	1	1
KNN	training set	0.81	0.82	0.81	0.81
	Test set	0.82	0.83	0.82	0.82
	London	0.73	0.72	0.73	0.72
	Kent	0.73	0.72	0.73	0.71
Random Forest	Training set	1	1	1	1
	Test set	1	1	1	1
	London	1	1	1	1
	Kent	1	1	1	1
XGBoost	training set	0.99	0.99	0.99	0.99
	Test set	0.96	0.97	0.96	0.96

Models	Data Split	Accuracy	Precision	Recall	f1-score
	London	0.99	0.99	0.99	0.99
	Kent	0.82	0.84	0.82	0.8
Neural Network	training set	0.56	0.53	0.56	0.52
	Test set	0.64	0.61	0.64	0.61
	London	0.58	0.66	0.58	0.55
	Kent	0.67	0.62	0.67	0.56
Support vector Machine	training set	0.54	0.6	0.54	0.45
	Test set	0.59	0.47	0.59	0.52
	London	0.56	0.32	0.56	0.4
	Kent	0.66	0.43	0.66	0.52
Naïve Bayes	training set	0.06	0.34	0.06	0.07
	Test set	0.04	0.17	0.04	0.03
	London	0.12	0.59	0.12	0.2
	Kent	0.04	0.5	0.04	0.03

Using random grid search methods, model hyper-parameters were modified in order to enhance the performance of each unique model. Table 5-2 displayed the result of the model with grid search hyper-parameter tuned. The performance of decision tree, KNN, and XGBoost classifiers were found to be 100% on both training and test datasets across all assessment criteria (accuracy, precision, recall and f1-score). Although random forest performance on training set in terms of accuracy, precision, recall and f1-score was 100%, however it performed dropped about 96% for accuracy, recall, and f1-score, and 95% for precision. On the training dataset, the neural network performed with around 69% accuracy, precision, and recall and a 68% f1-score. On the assessment dataset, it obtained a perfect score across all evaluation measures. Even after grid search, Nave Bayes performance was quite poor, with an accuracy and recall of around 7% on training data and 48% on the test dataset.

As shown in Tables 5-3 and 5-4, a complete evaluation of each model's performance in predicting specific crimes revealed that the accuracy of individual models in predicting specific forms of crime varied. For example, on the training dataset, decision trees and random forests achieved 99% precision in predicting offensive weapons and 100% precision for other types of crimes, while the performance of KNN was 74% in terms of precision on offensive weapon prediction, 86% precise for drug-related crimes, 68% precision for other types of crimes, and 77% precision for theft-related crimes. This proved that some models are superior to others in predicting specific forms of criminal activity. When the models were evaluated on the test data, the decision tree and random forest achieved 100% accuracy, precision, recall, and f1-score across the crime categories. XGBoost performed better in predicting theft and other related crimes than it did on offensive weapon and drug-related crimes. KNN achieved 100% in predicting other related crimes. Its performance dropped by about 10 to 14% in precision in predicting offensive weapons and drug-related crimes. Although it achieved above 50% accuracy, it was very poor in predicting other related crimes with zero precision. Also, the support vector classifier was not good at predicting offensive weapons and other related crimes (zero precision). Likewise, nave Bayes achieved zero precision in predicting theft-related crimes.

As a result, a stack of different models was built to capitalise on the capabilities of certain base models; the result of the stack ensemble model is shown in Table 5-5. The stack ensemble model using the voting classifier technique achieved 0.968908 accuracy on training data and 0.964286 accuracy on the test dataset. The accuracy of the stacking method of logistic regression was 100% on the training data set and almost 100% on the test data set.

Table 5-2 Weighted Average performance for the random grid search models without sentiment polarity

Models	Data Split	Accuracy	Precision	Recall	f1-score
Decision Tree	training set	1	1	1	1
	Test set	1	1	1	1
KNN	training set	1	1	1	1
	Test set	1	1	1	1
Random Forest	Training set	1	1	1	1
	Test set	0.96	0.95	0.96	0.96
XGBoost	training set	1	1	1	1
	Test set	1	1	1	1
Neural Network	training set	0.69	0.69	0.69	0.68
	Test set	1	1	1	1
Naïve Bayes	training set	0.07	0.36	0.07	0.09
	Test set	0.48	0.67	0.48	0.44

Table 5-3 The result of the models trained on the training dataset portion of the dataset without sentiment polarity for specific crime type

Models	Crime categories	Accuracy	precision	recall	f1-score
Decision Tree	Offensive weapons	1	0.99	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
KNN	Offensive weapons	0.81	0.74	0.8	0.77
	drugs		0.86	0.84	0.85
	others		0.68	0.62	0.65
	theft		0.79	0.79	0.79
Random Forest	Offensive weapons	1	0.99	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
XGBoost	Offensive weapons	0.99	0.99	0.97	0.98
	drugs		0.98	1	0.99
	others		1	1	1
	theft		1	0.98	0.99
Neural Network	Offensive weapons	0.65	0.53	0.39	0.45
	drugs		0.69	0.84	0.76
	others		0.67	0.05	0.09
	theft		0.63	0.51	0.56
SVM	Offensive weapons	0.54	1	0	0
	drugs		0.55	0.89	0.68
	others		0	0	0
	theft		0.46	0.28	0.35
Naive Bayes	Offensive weapons	0.06	0	0	0
	drugs		0.55	0.08	0.13
	others		0.02	1	0.03
	theft		0.19	0.01	0.01

Table 5-4 The result of the models tested on the test dataset portion of the dataset without sentiment polarity for specific crime type

Models	Crime categories	Accuracy	precision	recall	f1-score
Decision Tree	Offensive weapons	1	1	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
KNN	Offensive weapons	0.82	0.86	0.6	0.71
	drugs		0.89	0.86	0.88
	others		1	1	1
	theft		0.7	0.88	0.78
Random Forest	Offensive weapons	1	1	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
XGBoost	Offensive weapons	0.96	0.91	1	0.95

Models	Crime categories	Accuracy	precision	recall	f1-score
	drugs		0.97	0.97	0.97
	others		1	1	1
	theft		1	0.94	0.97
Neural Network	Offensive weapons	0.64	0.5	0.2	0.29
	drugs		0.68	0.86	0.76
	others		0	0	0
	theft		0.6	0.56	0.58
SVM	Offensive weapons	0.59	0	0	0
	drugs		0.64	0.86	0.74
	others		0	0	0
	theft		0.47	0.5	0.48
Naive Bayes	Offensive weapons	0.04	0	0	0
	drugs		0.33	0.03	0.06
	others		0.02	1	0.04
	theft		0	0	0

Table 5-5 stack ensemble models result

Models	Data Split	accuracy
Voting Classifier	Training set	0.968908
	Test set	0.964286
Logistic regression	Training set	1
	Test set	0.998

5.2.2 Model Explanation for the model without sentiment score

It is important to note that after longitude, people with white ethnic origin contributed the most to the forecast of the crime prediction model, followed by those aged 25 to 34. the people with other ethnic backgrounds and latitude contribute equally to the crime prediction. People of age group 18-24 contributed twice less than days of the weeks but about twice more than other genders. People of black ethnicity group contribution to the prediction is not noticeable (Fig 5-6).

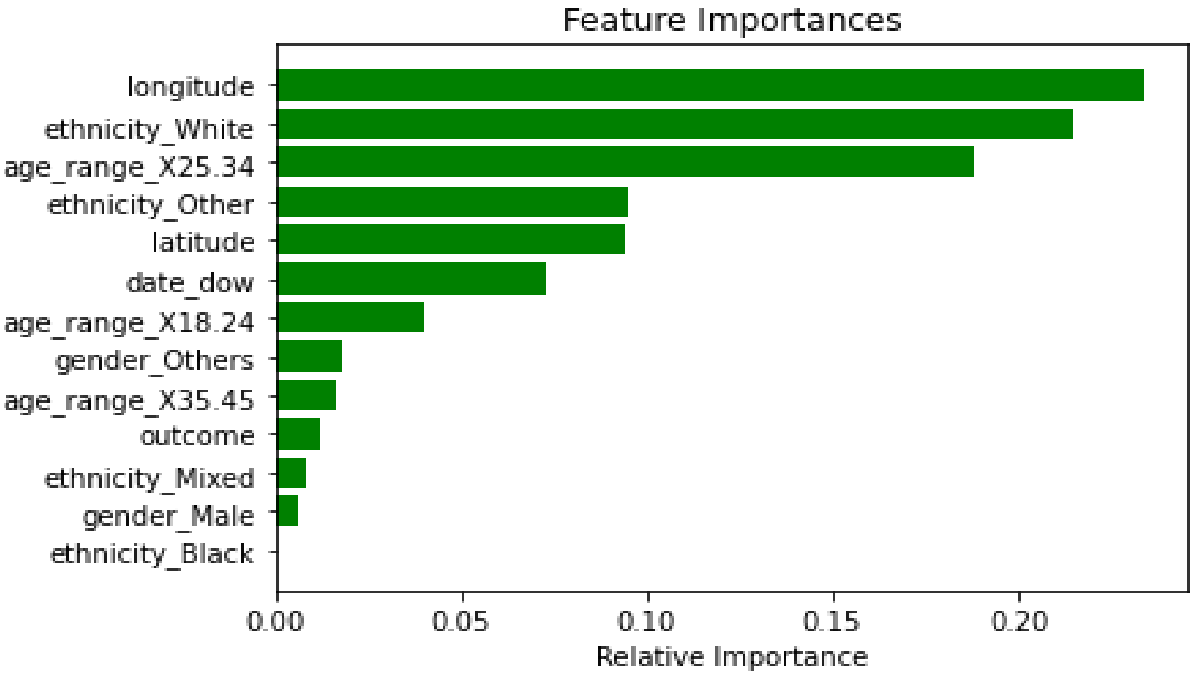


Figure 5-5: Dataset without sentiment score: Feature contribution to the Random Forest crime prediction

5.2.3 Result of the analysis performed on the combination of historical crime data with Twitter sentiment score

This section displays the outcomes of the exploratory analysis of the tweet data, as well as the outcomes of the models constructed using a mix of the historical crime data and Twitter sentiment score.

According to Table 5-6, decision tree, random forest, and XGBoost performed very well on our dataset, closely followed by KNN, with around 1% less prediction accuracy on the training dataset and approximately 2% less accuracy on the test set. On the training dataset and test set, the accuracy performance of neural networks and support vector machines was slightly over average (50%). However, the neural network was below average in terms of accuracy and f1-score, with 49% and 40%, respectively, while the support vector machine has 26% precision and 34% f1-score on both the training and test sets. Except for accuracy, where nave bayes has a performance average of almost 50%, its performance using all other assessment measures is approximately 35%.

Table 5-8 demonstrated that decision trees, random forests, and XGBoost are all capable of accurately predicting a wide variety of crimes. KNN is the most accurate at predicting drug-related crimes and other sorts of crimes, while it is around 99% accurate at predicting crimes involving offensive weapons and theft. The neural network achieved an accuracy of 86% for other sorts of crime, 53% for drug-related crimes, and 51% for theft-related crimes. With an accuracy of around 51%, a precision of approximately 51%, and a f1-score of 67%, support vector classier is only capable of predicting drug-related crimes. Even though the accuracy of the naive bayes classifier was lower than that of the support vector classifier, the naive bayes classifier performed better than the support vector classifier in predicting all forms of crime.

With the exception of XGBoost's accuracy on offensive weapons, the performance of decision tree, random forest, and XGBoost is close to 100 percent across all crime categories and evaluation metrics. There were no notable changes in the performance of neural networks, support vector machines, or naïve bayes classifiers. The KNN performance decreased by around 1% relative to the training set.

However, for models with customised hyper-parameters, as shown in Table 7, KNN performed best across all assessment measures, followed by XGBoost with a performance of about 99 percent. The performance of the random forest classifier decreased by around 20%, whereas the performance of the decision tree classifier decreased by approximately 30%. The performance of the neural network did not alter. However, the performance of the naive Bayes classifier has improved by around 16%. As demonstrated in Table 5-9, there were no discernible differences in the performance of the various crime prediction models.

The KNN classifiers beat other classifiers with a performance of 100% across all evaluation measures, followed by XGBoost, random forest, and decision tree in that order. However, decision trees are around 2% more accurate than random forests at predicting drug-related crimes, whereas random forests are approximately 5% more accurate than XGBoost at predicting offensive and theft-related crimes, respectively. The performance of the support vector classifier is around 16% better than the performance of the naive bayes classifier, while it is only excellent at predicting drug-related crimes and is extremely bad at predicting other types of crimes. Similar to models with default parameter values, with the exception of neural networks and nave bayes, there are no significant differences in the model's performance for distinct forms of crime. To take use of the prediction potential of individual models, a voting classifier and a logistic regression classifier were used to form a stack ensemble of individual models. Both stack ensemble classifiers performed well on both the training and test datasets, achieving a perfect score across all evaluation measures.

Table 5-6 The results of the individual models developed with historical dataset combined with sentiment score

Model	Data Split	Accuracy	Precision	Recall	f1-score
Decision Tree	training set	1	1	1	1
	Test set	1	1	1	1
KNN	training set	0.99	0.99	0.99	0.99
	Test set	0.98	0.98	0.98	0.98
Random Forest	Training set	1	1	1	1
	Test set	1	1	1	1
XGBoost	training set	1	1	1	1
	Test set	1	1	1	1
Neural Network	training set	0.52	0.49	0.52	0.4
	Test set	0.51	0.53	0.51	0.41
SVM	training set	0.51	0.26	0.51	0.34
	Test set	0.51	0.26	0.51	0.34
Naïve Bayes	training set	0.35	0.5	0.35	0.37
	Test set	0.35	0.5	0.35	0.37

Table 5-7 The results of the individual random grid search models developed with historical dataset combined with sentiment score

Model	Data Split	Accuracy	Precision	Recall	f1-score
Decision Tree	training set	0.72	0.71	0.72	0.71
	Test set	0.71	0.71	0.71	0.71

Model	Data Split	Accuracy	Precision	Recall	f1-score
KNN	training set	1	1	1	1
	Test set	1	1	1	1
Random Forest	Training set	0.8	0.84	0.8	0.79
	Test set	0.8	0.84	0.8	0.78
XGBoost	training set	0.9	0.9	0.9	0.9
	Test set	0.9	0.9	0.9	0.9
Neural Network	training set	0.52	0.44	0.52	0.42
	Test set	0.51	0.46	0.51	0.34
Naïve Bayes	training set	0.51	0.26	0.51	0.34
	Test set	0.51	0.26	0.51	0.34

Table 5-8 The result of the models trained on the training dataset portion of the dataset with sentiment polarity for specific crime type

Models	Crime categories	Accuracy	precision	recall	f1-score
Decision Tree	Offensive weapons	1	1	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
KNN	Offensive weapons	0.99	0.99	1	0.99
	drugs		1	1	1
	others		1	1	1
	theft		0.99	0.99	0.99
Random Forest	Offensive weapons	1	1	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
XGBoost	Offensive weapons	1	1	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
Neural Network	Offensive weapons	0.52	0.31	0.1	0.15
	drugs		0.53	0.95	0.68
	others		0.86	0.11	0.19
	theft		0.51	0.03	0.06
Support Vector Machine	Offensive weapons	0.51	0	0	0
	drugs		0.51	1	0.67
	others		0	0	0
	theft		0	0	0

Naive Bayes	Offensive weapons	0.35	0.43	0.01	0.02
	drugs		0.62	0.44	0.52
	others		0.06	0.82	0.12
	theft		0.38	0.37	0.38

Table 5-9 The result of the models tested on the test dataset portion of the dataset with sentiment polarity for specific crime type

Models	**Crime categories**	**Accuracy**	**precision**	**recall**	**f1-score**
Decision Tree	Offensive weapons	1	1	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
KNN	Offensive weapons	0.98	0.97	0.98	0.98
	drugs		0.99	0.99	0.99
	others		0.98	0.98	0.98
	theft		0.98	0.98	0.98
Random Forest	Offensive weapons	1	1	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
XGBoost	Offensive weapons	1	0.99	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
Neural Network	Offensive weapons	0.51	0.28	0.23	0.25
	drugs		0.55	0.89	0.68
	others		0.86	0.13	0.22
	theft		0.66	0.01	0.02
Support Vector Machine	Offensive weapons	0.51	0	0	0
	drugs		0.51	1	0.67
	others		0	0	0
	theft		0	0	0

Naive Bayes	Offensive weapons	0.35	0.4	0.01	0.01
	drugs		0.62	0.44	0.51
	others		0.06	0.82	0.11
	theft		0.41	0.39	0.4

Table 5-10 The result of the random grid search models trained on the training dataset portion of the dataset with sentiment polarity for specific crime types

Models	Crime categories	Accuracy	precision	recall	f1-score
Decision Tree	Offensive weapons	0.72	0.67	0.48	0.56
	drugs		0.75	0.86	0.8
	others		0.96	0.56	0.71
	theft		0.66	0.64	0.65
KNN	Offensive weapons	1	1	1	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
Random Forest	Offensive weapons	0.8	0.97	0.54	0.69
	drugs		0.73	0.99	0.84
	others		1	0.49	0.66
	theft		0.93	0.65	0.77
XGBoost	Offensive weapons	0.9	0.93	0.8	0.86
	drugs		0.88	0.96	0.92
	others		1	0.97	0.99
	theft		0.91	0.84	0.87
Neural Network	Offensive weapons	0.52	0.27	0.05	0.09
	drugs		0.54	0.91	0.68
	others		0.33	0.12	0.17
	theft		0.41	0.15	0.22
Naive Bayes	Offensive weapons	0.51	0	0	0
	drugs		0.51	1	0.67
	others		0	0	0
	theft		0	0	0

Table 5-11 The result of the random grid search models tested on the test dataset portion of the dataset with sentiment polarity for specific crime types

Models	Crime categories	Accuracy	precision	recall	f1-score
Decision Tree	Offensive weapons	0.71	0.67	0.48	0.56
	drugs		0.74	0.86	0.8
	others		0.97	0.57	0.72
	theft		0.66	0.63	0.64
KNN	Offensive weapons	1	1	0.99	1
	drugs		1	1	1
	others		1	1	1
	theft		1	1	1
Random Forest	Offensive weapons	0.8	0.98	0.53	0.69
	drugs		0.73	0.99	0.84
	others		1	0.48	0.65
	theft		0.93	0.65	0.77
XGBoost	Offensive weapons	0.9	0.92	0.79	0.85
	drugs		0.88	0.96	0.92
	others		1	0.96	0.98
	theft		0.91	0.85	0.88
Neural Network	Offensive weapons	0.51	1	0	0
	drugs		0.51	1	0.67
	others		0	0	0
	theft		0	0	0
Naive Bayes	Offensive weapons	0.51	0	0	0
	drugs		0.51	1	0.67
	others		0	0	0
	theft		0	0	0

Table 5-12 The weighted average result of the stack ensemble models for the training and test dataset of the dataset with sentiment polarity

Models	Data split	Accuracy	precision	recall	f1-score
Voting Classifier	Training set	1	1	1	1
	Test set	1	1	1	1
Logistic Regression	Training set	1	1	1	1
	Test set	1	1	1	1

5.3 Model Explanation for the model with sentiment score

The relevance of a feature is a helpful metric for estimating the extent to which it contributes to the desired value. Estimates of feature significance may be obtained automatically from a trained prediction model when employing ensembles of decision tree techniques such as gradient boosting, XGBoost, and Random Forest. Feature importance is a metric used to quantify the significance of a feature's contribution to the model's boosted decision trees. The more often a characteristic is relied on to make a decision, the more weight it should be given. For this reason, a trained tree model can offer the order of feature relevance for a given task. Features with higher scores are more likely to have an impact on the predicting outcomes than those with lower scores (Song et al., 2020).

According to their relative weight, the variables relating to the attributes are ranked from most to least significant. The horizontal position indicates whether the value has a positive or negative effect on crime prediction. When the value of a variable is high (represented by red) or low (represented by blue), the accompanying colour reflects that value's range. There is a positive correlation between high levels of median income and high levels of median house value.

According to the ordering of feature contributions given in Figure 5-10 by random forest classifier, longitude was the most influential factor, affecting the predicted crime types by an average of 0.15, followed by the age group of 18–24 with an average change of roughly 0.25. Both the days of the week and the date of the month had almost an equal role in the prediction, followed by the age group 35-45, then the age group 24-34, and persons of black race. In the crime prediction output, white ethnicity is more significant than male gender, whereas male gender is more significant than other genders. Polarity followed people with other gender in the rankings, followed by those of a different ethnicity. The significance of individuals with mixed racial ancestry was hardly noted.

The Figure 5-11 also revealed that all of the characteristics contributed to the prediction of drug-related crimes, theft, and crimes using offensive weapons. However, the contributions of persons of mixed race were limited to the drug usage and were modest. The association between each of the characteristics and the crime prediction was shown in Figure 5-12. It was discovered that while the contribution of sentiment polarity was significant, it had an indirect effect on the crime prediction. And sentiment polarity was deemed more significant than the model's other three features (people of another ethnicity and mixed ethnicity).

When the decision tree's explanatory model was completed, date of the months, was noted to be most important features in the model, followed by latitude, days of the week, longitude, age group 18-24 and age group 35-45. The sentiment polarity was at the top of the last six features in terms of its significant contribution to the crime prediction. The location of sentiment polarity shifted upward, leaving around five features (Figure 5-13) and still exhibiting an indirect relationship with the target variables (Figure 5-14).

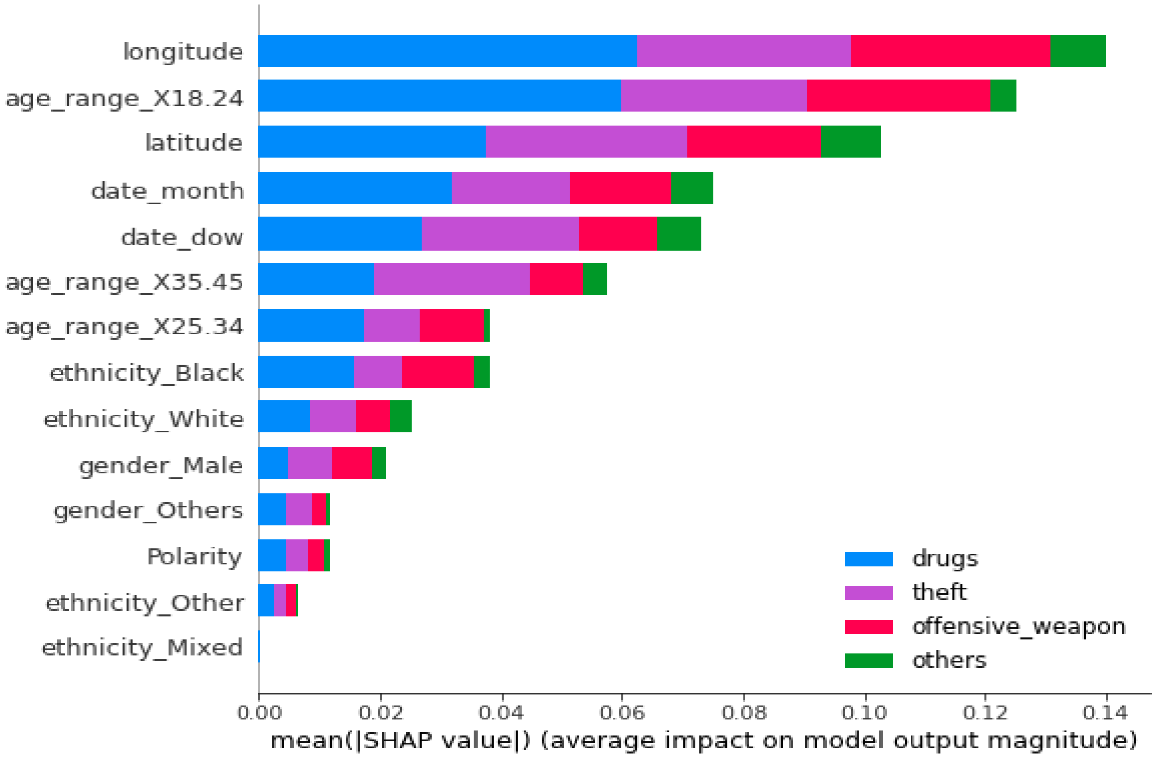


Figure 5-6: Dataset with sentiment score: Feature contribution to the Random Forest crime prediction and relationship with crime categories

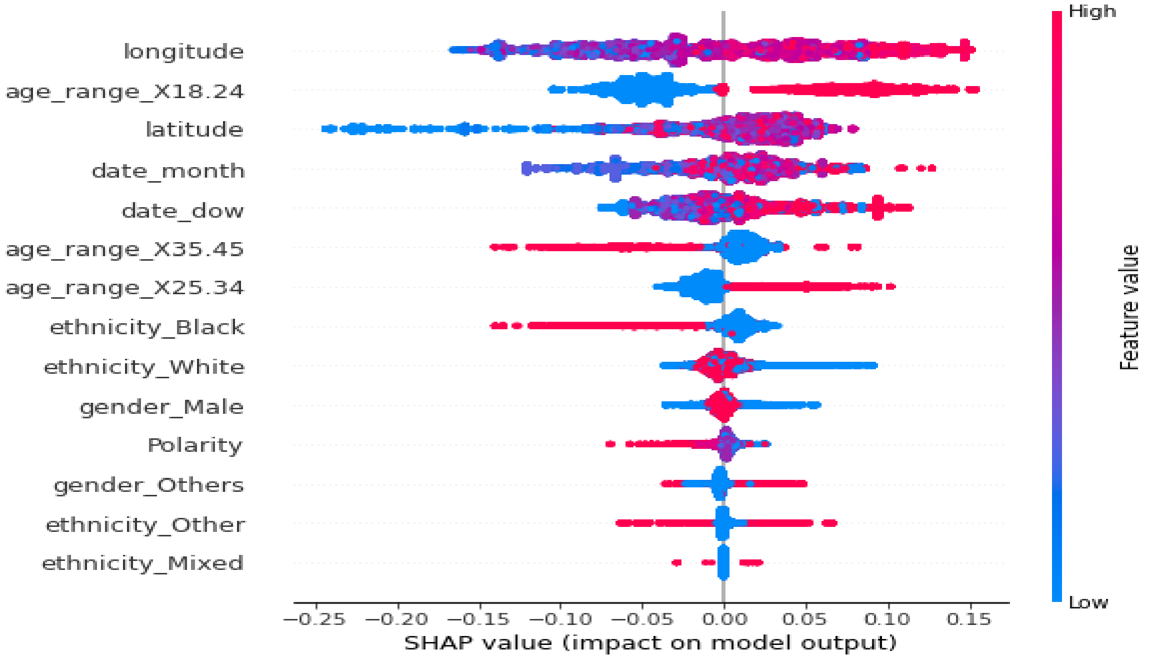


Figure 5-7: Feature dependency plot showing the correlation and importance of the variables for random forest crime prediction model

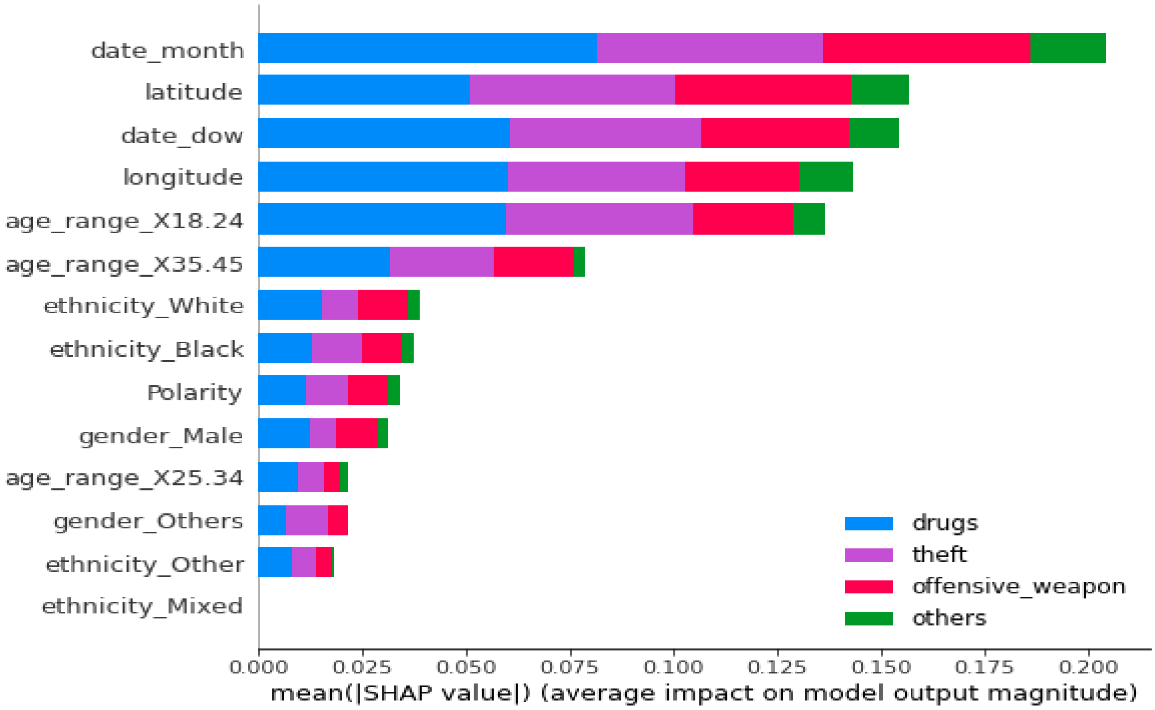


Figure 5-8: Dataset with sentiment score: Feature contribution to the Decision Tree crime prediction and relationship with crime categories

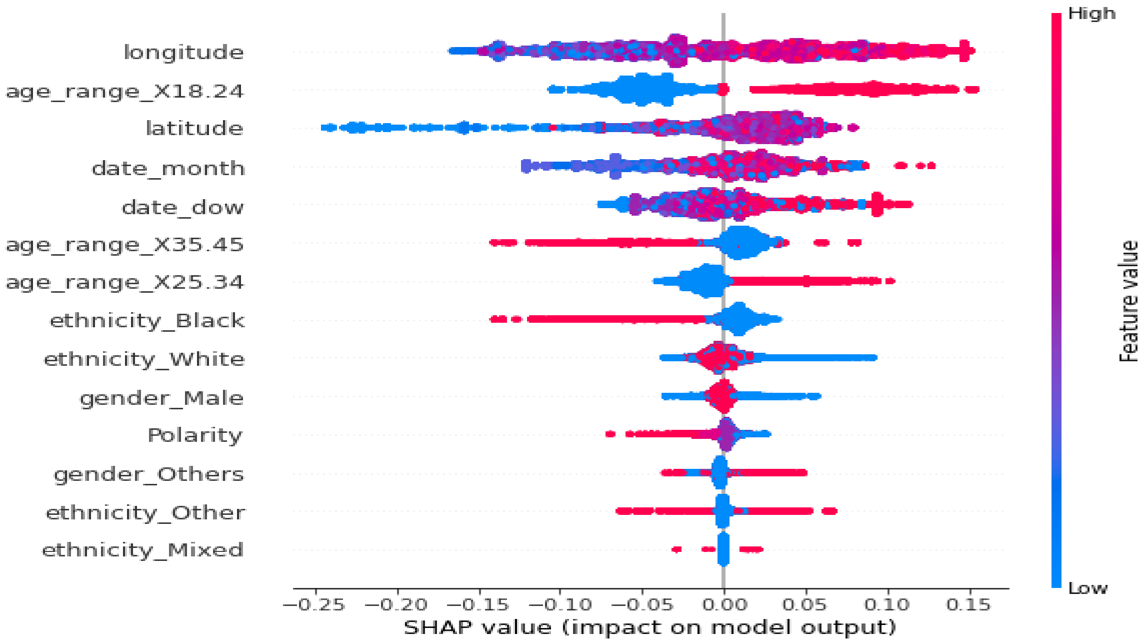


Figure 5-9: Feature dependency plot for decision tree crime prediction model

5.3.1 Rapid Twitter Information Search

A brief analysis of Fig 5-7 revealed the spread of tweets from 2017 to 2020. According to the graph, the number of tweets peaked in May and November of 2017, respectively. In 2018, the highest numbers of crime were recorded in May, while in 2019 and 2020, it occurred in December. Up until mid-July, when it began to steadily decline, the peak was seen in almost every month. The sentiment polarity distribution of Twitter users is seen in Fig 5-8. All tweets with a negative or good emotion may be seen alongside the tweet and the user, and in Figure 5-9 the tweets' positive and negative terms are shown as a word cloud. The size of each word determines its frequency.

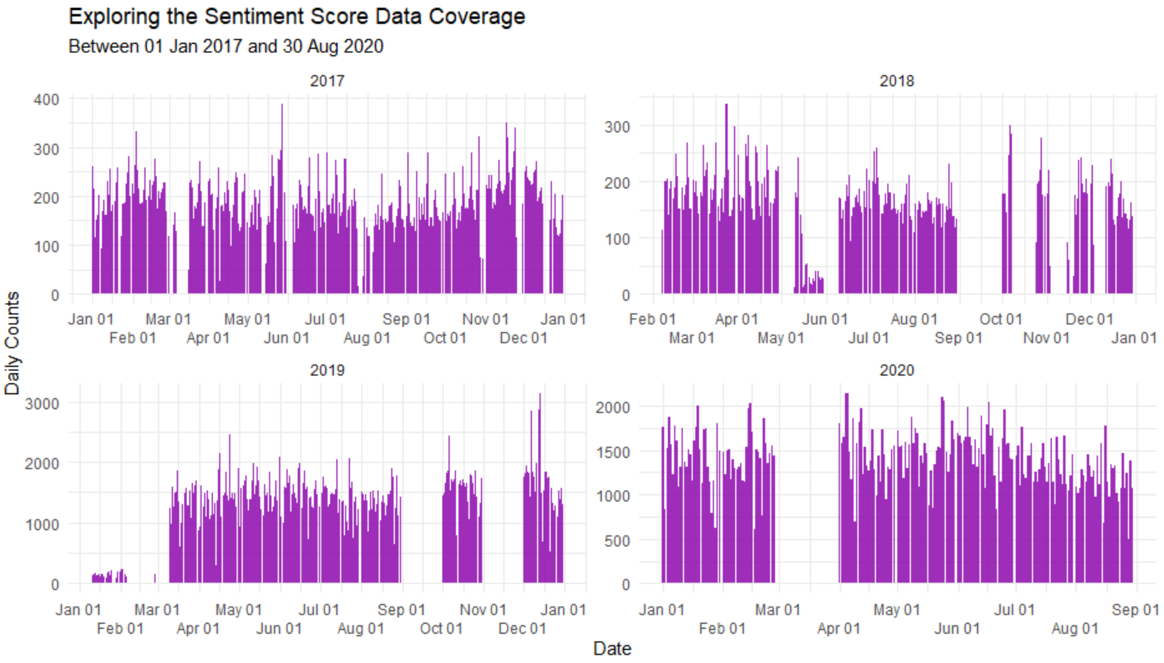


Figure 5-10: Time series plot of tweets information extracted from Twitter

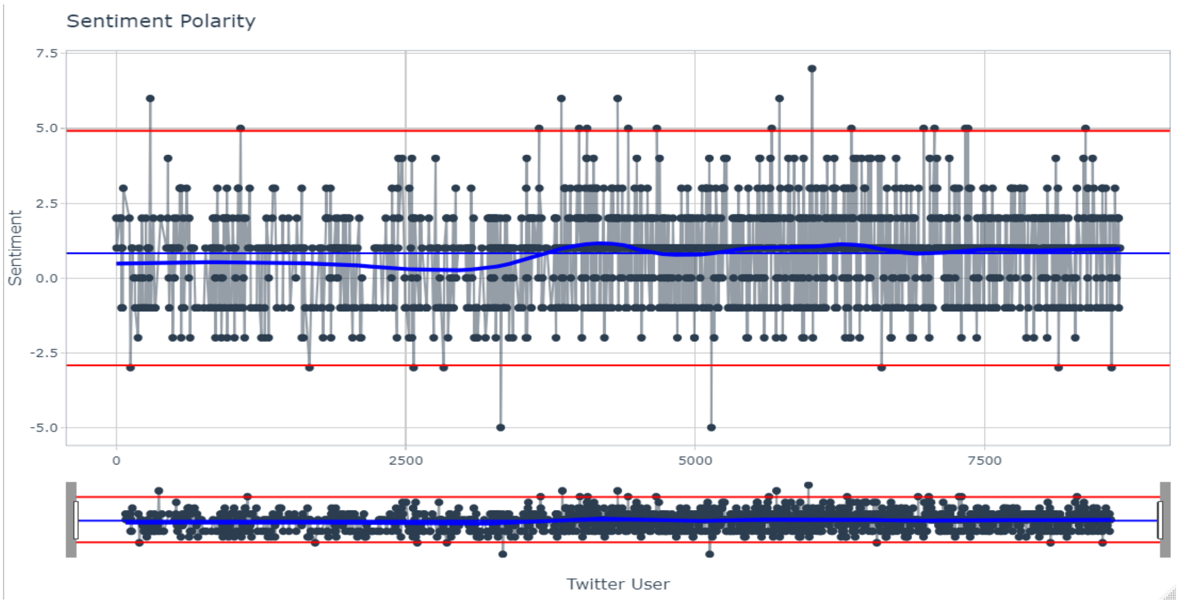


Figure 5-11: Sentiment Polarity plot against userid. with the upper- and lower-class boundary line.

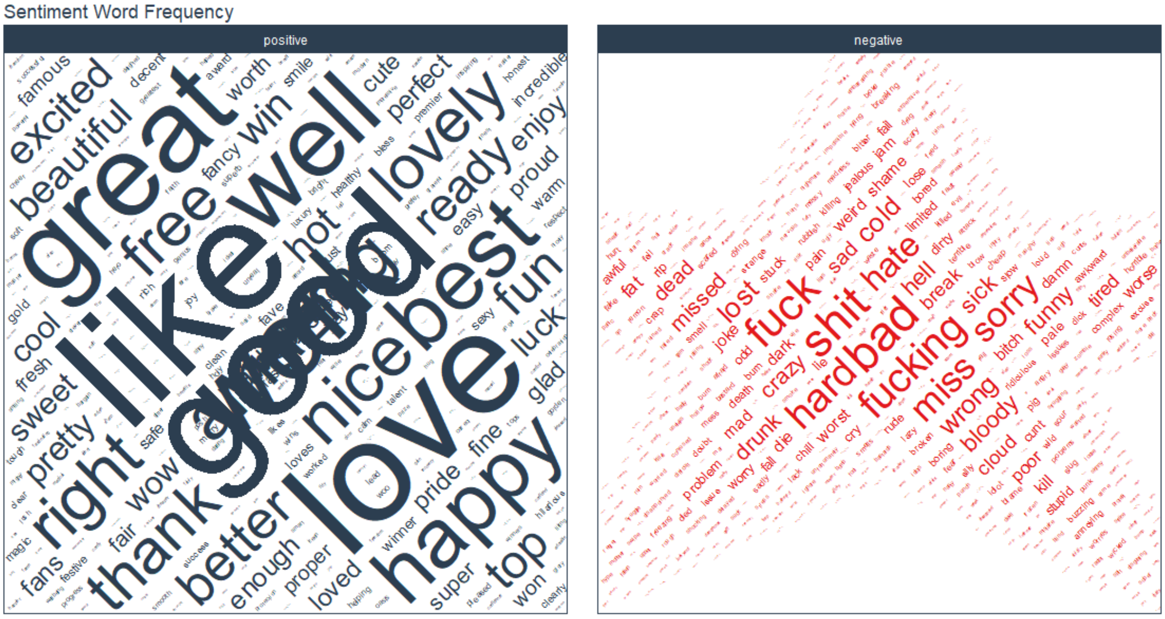
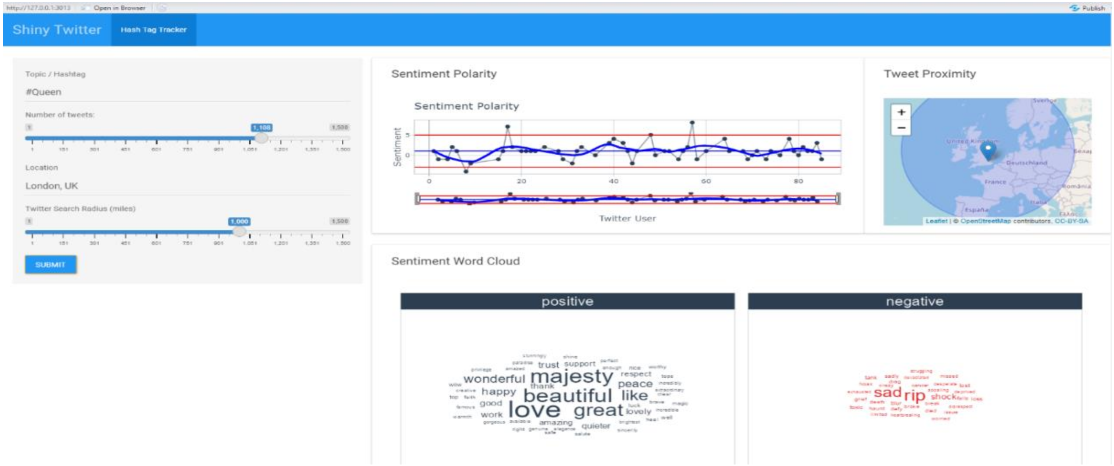


Figure 5-12: word cloud displaying the positive words in blue colour and negative words in red colour

Figure 5-5, depicted a picture of a Shiny App designed to evaluate the sentiment of tweets for social media crime detection in near real-time. The application reveals the sentiment score, and all tweets that fall within the upper- and lower-class are given ultimate attention. The software displays both positive and negative tweets in addition to the crime's vicinity. It allows for flexible location changes and keyword searches.



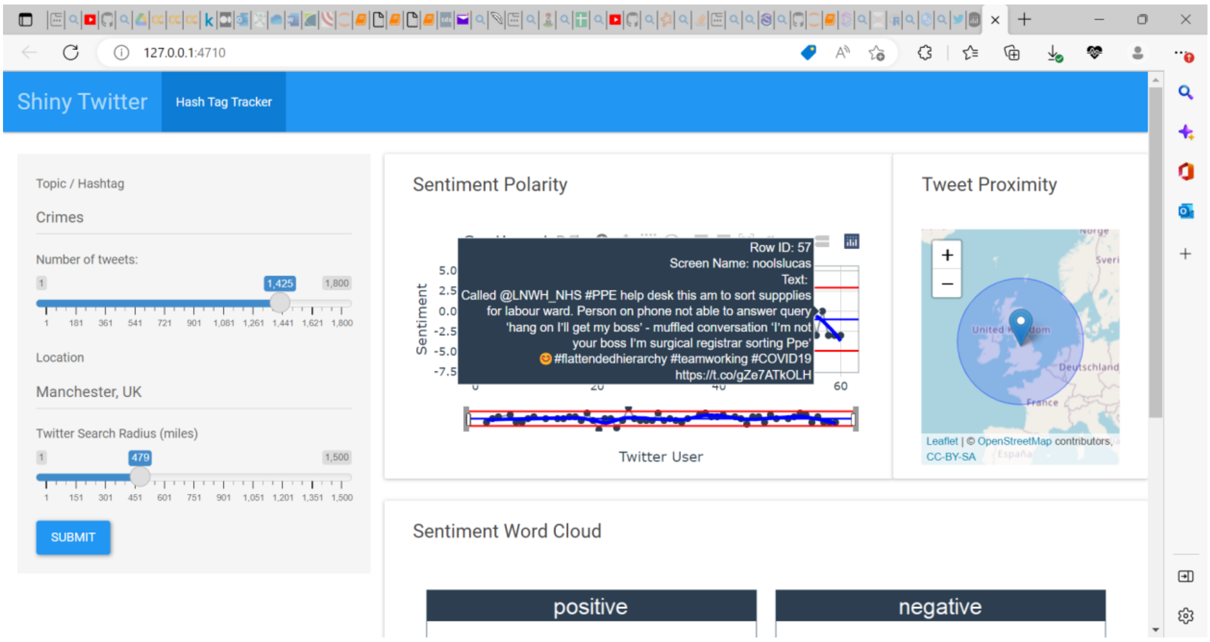
Topic / Hashtag: #Queen

Number of tweets: 1,100

Location: London, UK

Twitter Search Radius (miles): 1,000

Figure 5-13: Shiny App: Displaying Twitter sentiment analysis with the upper- and lower-class boundary, crime proximity and positive and negative tweet



127.0.0.1:4710

Shiny Twitter

Hash Tag Tracker

Topic / Hashtag

Crimes

Number of tweets: 1,425

Location

Manchester, UK

Twitter Search Radius (miles) 479

SUBMIT

Sentiment Polarity

Row ID: 57
Screen Name: noolslucas
Text:
Called @LNWH_NHS #PPE help desk this am to sort supplies for labour ward. Person on phone not able to answer query 'hang on I'll get my boss' - muffled conversation 'I'm not your boss I'm surgical registrar sorting Ppe' 😡 #flattendedhierarchy #teamworking #COVID19 https://t.co/gZe7ATkOLH

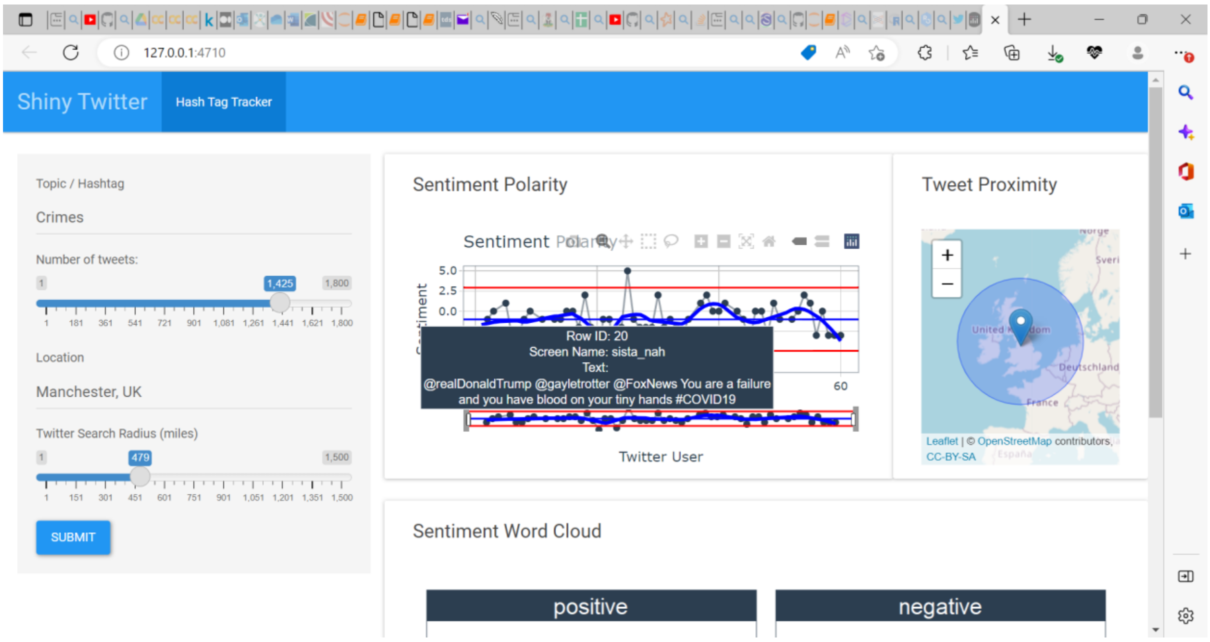
Tweet Proximity

Sentiment Word Cloud

positive

negative

Figure 5-14: Shiny App: Displaying Twitter sentiment analysis with the upper- and lower-class boundary, crime proximity and positive and negative tweet, displaying the Upper-class boundary sentiment polarity, tweets, the screen name and the approximate



127.0.0.1:4710

Shiny Twitter

Hash Tag Tracker

Topic / Hashtag

Crimes

Number of tweets: 1,425

Location

Manchester, UK

Twitter Search Radius (miles) 479

SUBMIT

Sentiment Polarity

Row ID: 20
Screen Name: sista_nah
Text:
@realDonaldTrump @gayletrotter @FoxNews You are a failure and you have blood on your tiny hands #COVID19

Tweet Proximity

Sentiment Word Cloud

positive

negative

Figure 5-15: Shiny App: Displaying Twitter sentiment analysis with the upper- and lower-class boundary, crime proximity and positive and negative tweet, displaying the lower-class boundary sentiment polarity, tweets, the screen name and the approximate

5.4 Result discussion

This study has offered insight into the performance differences between six distinct machine learning models applied to a UK crime prediction dataset using machine learning techniques. It also highlighted the potential of leveraging Twitter sentiment polarity to detect crimes in near real time. In addition, it demonstrated how the typical approach for data mining may successfully enhance crime prediction: Due to the limitations of the present study, the findings should be regarded with care. This section offers an analysis of the study findings. The design's shortcomings and possible ramifications, as well as their implications for the interpretation of the data, were examined. The chapter concludes with a number of suggestions for further investigation.

It is a well-known saying among data scientist that, data cleaning and pre-processing consumes about 75% to 80% of their task. This might be one of the reasons why previous researchers have been avoiding the stop and search data set from the UK. Therefore, a function has been created and tested on about 32 police districts in the UK, to enhance future research.

On the two datasets evaluated, the default parameter values for decision tree, random forest, and XGBoost classifiers produced predictions that were approximately 100% correct. The performance of the random forest classifier on crime prediction backed by prior research that utilised the algorithm on crime prediction, even though our result was superior. (Almaw and Kadam, 2019; Baculo et al., 2017; Hossain et al., 2020; Ippolito and Lozano, 2020). Although KNN performance was greater with random grid search which was validated by the work of Castro et al., 2020 and Das and Das, 2019, the performance differed for various kinds of crime. However, the KNN results achieved in this work were superior than those obtained by Feng et al., 2018, Kim et al., 2018, Kiran and Kaishveen, 2019) in their research. The findings of the decision tree classifier achieved in this study were also better than those acquired in prior studies (Adesola et al., 2020b; Iqbal et al., 2013). This study also shown that naive bayes was not suitable for crime prediction on the UK dataset and was particularly ineffective for predicting crime in Greater Manchester, even though the previous researchers who used naive bayes in their study received a better result (Alsaqabi et al., 2019, Kiran and Kaishveen, 2019). The poor performance of the support vector machine had been documented in the previous literature. Even though support vector machine performance on training data was excellent, it performed poorly on testing data (Adesola et al., 2020a, Castro et al., 2020, El Bour et al., 2018, Kissi Ghalleb and Ben Amara, 2020, Kshatri et al., 2021, Nguyen et al., 2017, Wijaya et al., 2019). It took extremely longer time to adjust the support vector machine's hyper-parameters. Therefore, it was only tested with the default settings. The average performance of the neural network across the two datasets was 50%. Even though earlier research shown high accuracy when applied to their dataset (El Bour et al., 2018; Esquivel et al., 2020), hyper-parameter adjustment did not improve their results.

It might be possible that the decline in performance of decision trees, random forests, and XGBoost was due to the selection of inappropriate hyper-parameters or to the selection of inappropriate values for the hyper-parameters. Since the researchers' performance on the datasets was flawless across all categories of crime, any of the three models may be applied to the UK dataset. The performance of stacked ensemble models on the dataset devoid of sentiment polarity was inconsistent. While they performed exceptionally well on both training and testing data for the dataset containing sentiment polarity, it would be preferable to trade off the stacked ensemble models due to the time required to train them, especially since other individual models could provide the same performance in less time. In addition, the decision tree, random forest, and XGBoost are the most acceptable base models for crime prediction using the UK stop and search dataset based on the consistency of their predictions on both training and testing data. Although a decision tree requires the least time to construct, a random forest classifier leverages on averaging the best selection of decision base model and requires little time to be trained, making it the best option for the UK dataset.

In addition to location, the 18–24 age group also plays a significant effect in crime prediction. The date of the months and days of the weeks were followed by the age ranges 35-45 and 25-34. Sentiment polarity, other ethnicity, and mixed ethnicity were the last three features that had significant contribution to the crime prediction. The features that contributed most to the decision tree model were the day of the month, latitude, days of the week, longitude, the age range of 18-24, and finally the age range of 35-45. However, among the final six factors included into crime prediction models, polarity performed the best. In addition, for the decision tree model, latitude has a low but negative connection with the prediction of crime; age group 35–45; and black ethnicity and polarity have a strong but negative association with the prediction of crime. This might be seen as having a substantial but indirect effect on crime prediction. The link between crime and longitude, the 18-24 and 25-34 age groups, and the other genders was robust and favourable. This demonstrates the substantial and direct effect that these characteristics have on crime. In addition, the white ethnic group and men have a strong and direct relationship with crime. Several factors could have contributed to the crimes committed especially by the people withing the age group 18-24. One of them might be that they grew up with those crimes from young age. Other factors might be the effect of their local areas or their quality of the schools they attended. Even, though, schools were not reported in the historical crime records, it is possible that majority of the crimes were being committed by the people from high school especially those from low ratings. It might also be that they were being used by the older age group to perpetrate evils.

This finding that the sentiment polarity has a strong indirect association with crime demonstrates that both positive and negative polarities conceal a hidden reality. Therefore, the positivism of the tweets should not be concluded as positive or negative until further investigation has been carried on the tweet. This demonstrates that the offender may still be identified regardless of how he or she communicates, particularly with the use of the developed visualisation tool.

CHAPTER 6 : CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

In this chapter the conclusions derived from the findings of this study on this research will consider the possible influence of both demographic factors as well as social media in the prediction of crime occurrences. The conclusions were based on the aims, research problems and results of this study. The implications of these findings and the resultant recommendations will also be explained. Recommendations were based on the conclusions and purpose of the study.

6.2 An overview of the Research

This study adopted the standard data mining procedure (CRISP-DM) to achieved the set objectives to address the research problems. an exploratory, descriptive and contextual qualitative study of the historical crime information containing the locations, time, demographic factors of the offenders and the crime categories were purposively selected as participants. Also, twitter data used were not limited to specific keywords to be able to reveal hidden pattern in the user’s communication. The sentiment polarity extracted from the twitter was finally merged with the historical data for further analysis and crime prediction.

The research methods and approach for this study were adopted as a result of gaps found in the previous work on crime prediction as well as the outcome of a comprehensive systematic review conducted. The findings and recommendations described below are centred on the research problems, the objectives and the results from the data analysis. The research problems were wrap up around the following research problems:

- The acquisition of additional data from unrelated external sources, the extraction of crime-specific keywords from Twitter, and the inappropriate implementation of data mining best practises.

- Inadequate planning while selecting the best method to predicting crime in the United Kingdom, which might be because of insufficient data exploration, would lead to the feature engineering required to develop the most effective machine learning models.
- To uncover a link between the daily Twitter sentiment polarity and the crime rate, earlier research that studied the potential of social media to reveal crime nearly exclusively used regression analysis approaches.

The following objectives were established to address the highlighted issues:

- Detailed exploratory analysis that revealed the possible connection between different features used for crime prediction, suggested the most appropriate pre-processing and transformation and engineering require prior to building the crime prediction model was obtained.
- The best suitable machine learning model for both datasets with and without sentiment from Twitter was developed, assessed on test data, and subsequently verified using crime data from London and Kent. Comparison of the two datasets to determine the impact of Twitter data on crime prediction. In addition, a stacked ensemble of crime prediction models was developed using separate machine learning techniques, and the optimal method of stacking the models using Voting classifier and Logistic regression was determined.

6.3 An overview of the result

Four main topics emerged from the data. The findings were discussed according to the four topics that emerged from the data:

6.4 The topics and their significance

Topic 1: It is a well-known saying among data scientist that, data cleaning and data pre-processing consumes about 75% to 80% of their task. Data pre-processing is a crucial stage in machine learning since it improves the data's integrity, consistency, algorithm readability, and quality. It also facilitates machine learning algorithms' ability to read, use, and analyse the data (Huang et al., 2015). This might be one of the reasons why previous researchers that have carried out studies on crime prediction (Mishra et al., 2021, Williams et al., 2019) did not use the stop and search data set from the UK. Unlike previous studies that employed dataset from the UK, a function capable of cleaning the data, pre-processing it, while considering the required feature engineering has been created to facilitate future study, particularly on UK crime data. This function has been verified and shown to perform well in 32 UK police districts. By applying our procedure for cleaning data to the stop-and-search crime data from the past, an analysis-ready dataset would be made.

Topic 2: Based on the exploratory analysis of historical crime records and the information extracted from Twitter information, most crimes were committed by men of mostly white ethnicity, and that drug-related offences were dominated among them. Drug-related crimes accounted for over 10% of all crimes committed between Sunday and Thursday, and 15% of all crimes committed between Friday and Saturday. This study revealed that most crimes were perpetrated by individuals between the ages of 18 and 24, who were mostly engaged in drug-related offences; the influence of drugs was also evident in the rise of other crimes. It may be argued that as individuals age, the quantity of crimes drops but the severity of crimes rises. The number of people between the ages of 10 and 17 is another noteworthy discovery that should be investigated by those in charge. These pupils are in their sixth year of elementary and secondary education, yet they have been engaged in a range of criminal activities, including drug trafficking. It is possible that some adults used them to sell their drugs, and by the time they reached adulthood, drug marketing had become a part of their lives in drug dealing and other types of crime. This may be one reason the population at that age continues to grow. So, it's likely that drug use among young people is to blame for the rise in crime.

Topics 3: The problem of random selection of models to be applied for crime prediction was addressed by selecting the most dominant machine learning algorithm found in the earlier literature. The six models were trained using both default and randomly generated hyper-parameters for both datasets with and without sentiment polarity. Also, both the voting classifier and the logistic regression classifier were used to make a stacked ensemble of individual models, and their performance was compared. Thus, it could be concluded from the result that although stack ensemble of individual models improves the performance of crime prediction models, individual models such as decision trees, random forests, and XGBoost can achieve the same result without the need to train many models, which would take a significant amount of time. Also, it can be concluded that the most appropriate models for the UK dataset are decision tree, random forest and XGBoost, but decision tree is the most consistence of the three models.

Topics 4: To be able to account for Twitter's contribution to crime, a sentiment analysis of all tweets, not just those containing specified keywords, was conducted. It may be inferred that integrating hourly sentiment polarity with other demographic factors may enhance the effectiveness of various crime prediction models. Based on the contribution of various characteristics to the performance of the model, it is possible to deduce that sentiment polarity has a large but indirect influence on crime. People in charge of security should study both negative and positive tweets, and should not be side tracked by the positive language used by certain criminals to do evil crimes.

6.5 Summary

The study confirmed that people aged 18–24 are involved in the most criminal events such as drugs, weapon usage, and theft, especially on Sundays and Saturdays. This is supported by the recommendation of (Aghababaei and Makrehchi, 2018) Therefore, people in authority should not restrict the solution to only police arrest, thereby overburdening the enforcement officers.

Not only did the random forest and decision tree do well on the UK dataset, but their results were also the same when they were tested on other data. The outcome of Hossain et al. (2020b) quest to predict crime using spatial-temporal information, in which Random Forest achieved an accuracy of about 99%, supported the performance of Random Forest in this research. The results of the decision tree in this study also back up what Adesola et al. (2020b) found about predicting violent crime in Nigeria. Likewise, the studies conducted on the analysis of crime information using an ensemble approach (Almaw and Kadam, 2019, Bappee et al., 2018) confirm the effectiveness of random forests in crime prediction. The performance of Random Forest might be due to its ensemble nature and the fact that it trained a number of decision tree algorithms to make its final prediction.

XGBoost's performance improved when its hyper-parameters were tuned to select the optimal parameters. Although it takes several times to tune the hyperparameters of the XGBoost, the performance of the algorithm in this study is better than that obtained by Castro et al. (2020) in their crime prediction on heterogeneous data. Thus, Random Forest and Decision Tree are suggested for use on the Stop and Search UK Crime dataset for crime prediction. However, Naive Bayes and neural network models have been reported to perform better than Random Forest in some other studies (Alsaqabi et al., 2019, El Bour et al., 2018). Even though it is not clearly stated in their report that standard data mining procedures were adopted, variation in the different models' performances is evidence that there is inconsistency in the performance of different algorithms across different datasets from different countries.

The result obtained from this research by using Random Forest is about 10% better than the one with the best accuracy found in the study conducted by Hossain et al. (2020b) on crime prediction while exploring spatial-temporal data. This improvement could be a result of the application of standard data mining procedures in this research.

Even though there is not a big difference between how well random forest and decision tree work for datasets with and without sentiment polarity, random forest is better because it uses several decision trees to come up with a result. XGBoost improved by about 8% when applied to a dataset with polarity to obtain an accuracy of 100%. This showed the contribution of sentiment polarity extracted from Twitter. The contribution of sentiment polarity extracted from social media has been discussed extensively by some of the previous researchers, except that most of them, including those that used the UK dataset, collected keyword-specific information from their sources (Aghababaei and Makrehchi, 2018, Kounadi et al., 2015, Williams et al., 2019), which although it would improve the model's performance, did not provide the real contribution of the sentiment polarity or be directly used for crime prediction. Unlike other studies, the contribution of individual variables was evaluated, and the result showed that the sentiment polarity of all tweets extracted from the Twitter social media platform has a strong but indirect relationship with the crime prediction. This supports the assumption that some criminals would not register their intentions in a negative way on the social media platform. Therefore, not even the most positive comment on any social media platform should be ignored at any point in time. They should be given reasonable consideration by the experts in the field of crime detection and prediction.

Additionally, one should not be misled by the negative sentiment of tweets because few criminals would express negative intentions, as proven by the indirect but significant impact of sentiment polarity on crime prediction based on their evaluations using decision tree and random forest models. Therefore, the Shiny App has been developed to ease the interpretation of the sentiment trend.

The sentiment trend app is determined by the 95% confidence interval below and above the median sentiment polarity. Unlike previous studies that only found the correlation between sentiment score and crime events (Aghababaei and Makrehchi, 2018), or those that evaluated the trend analysis based on the number of tweets collected (Das et al., 2020), or those that explored whether the volume of tweets from their location would be sufficient for crime prediction (Featherstone, 2013), the chart displayed in the app developed in this study would allow users to analyse individual tweets. It is so adaptable that multiple selections can be made for items such as location, number of tweets, and keywords. The chart displays the sentiment trends, the user's screen name, the tweets, their approximate location, and whether the tweet is positive or negative. If the tweets are suspicious, further investigation could be carried out based on the location provided in the app.

The function developed has been shown to be successful in cleaning, pre-processing and applied feature engineering on about 32 UK police districts.

6.6 Future Research

Base on the result obtained from this research, the following recommendations are made

- To answer the question of why crime is so prevalent among 18- to 24-year-olds, further research is necessary.
- Examine the link between crime, school quality, and local deprivation
- Explain the influence of schools and the quality of living neighbourhood (deprivation) on juvenile and young adult criminal behaviour (18-25).

- In addition, academics should examine how schools contribute to crime in the community.
- To prevent or reduce juvenile drug trafficking, it is vital to examine the criminal ties between adults and those aged 10 to 17 years.
- Even though, according to our data, gender did not play a role in feature selection, future research should investigate female genders to prevent parents from using their children as drug or criminal trade agents.

6.7 Research Limitations

This study analysed Twitter data as an extra source of information for crime prediction. However, limited access to historical crime information publicly available on the UK police website restricted the use of this research to up to 2019. Information from 2020 was not made available for public usage. Therefore, the Twitter information collected was filtered to be up to 2019. Those beyond 2019 were not included in the modelling, thereby reducing the sample size of the data.

On the UK police website, only rough information about the variables was given. This was assumed to affect whatever outcome derived from this research.

The biggest problem with this study is that it is hard to get a large enough number of tweets. This is because Twitter makes free information available every day.

Keywords were used to develop the app, which would affect the quality of the information displayed by it. And this is because streaming the most recent information on Twitter's free platform can only return about 1% of the total tweets at a time. Therefore, premium services would be required if this app were to be used in production.

This study's findings are limited to Twitter data, however other social media websites should also be included for a more precise forecast. Additionally, just a tiny portion of Twitter's data was examined in this research. This suggests that more Twitter data might enhance the generalizability of this study's results. The sentiment analysis was done using a publicly available vocabulary. It might be investigated how a lexicon of slang spoken by criminals could aid in the prediction of crimes.

6.8 Conclusion

This study explored the possible impact of Twitter data on crime prediction by combining historical crime data with sentiment polarity derived from Twitter analysis. The exploratory analysis of the data revealed that persons of varied ages (18–24) and genders contribute to crime prediction. The right prediction models for forecasting crime on the UK dataset were also selected. Moreover, the significance of Twitter data was emphasised.

It may be inferred that multitasking is the answer to reducing crime. Because there is no one-way answer, both the government and the general population must be engaged. If possible, social etiquette should be included into the school curriculum and the government should check children's morality from elementary school through secondary school. The government should update and completely finance secondary schools to offer the essential aid for educational and moral standards.

We think that this study has helped us learn more about how to forecast crimes using social media data, and its findings and recommendations might serve as a foundation for future research initiatives and the continuance of crime prediction research.

Reference List

1. ABBAS, M., MEMON, K. A., JAMALI, A. A., MEMON, S. & AHMED, A. 2019. Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS Int. J. Comput. Sci. Netw. Secur,* 19**,** 62.

2. ABRAMS, D. S. 2021. COVID and crime: An early empirical look. *J Public Econ*, 194**,** 104344.

3. ADEPEJU, M., ROSSER, G. & CHENG, T. 2016. Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study. *International Journal of Geographical Information Science,* 30**,** 2133-2154.

4. ADESOLA, F., AZETA, A., ONI, A., AZETA, A. E. & ONWODI, G. 2020a. Violent crime hot-spots prediction using support vector machine algorithm. *Journal of Theoretical and Applied Information Technology,* 98**,** 3187-3196.

5. ADESOLA, F., AZETA, A., ONI, A., CHIDOZIE, F. & AZETA, V. 2020b. Predicting Violent Crime Occurrence: An Evaluation of Decision Tree Model. *International Journal of Engineering Research and Technology*, 13**,** 1258-1265.

6. ADLER, A. I. & PAINSKY, A. 2022. Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. *Entropy,* 24**,** 687.

7. AGHABABAEI, S. & MAKREHCHI, M. Mining social media content for crime prediction. 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2016. IEEE, 526-531.

8. AGHABABAEI, S. & MAKREHCHI, M. Mining Social Media Content for Crime Prediction. 2017. 526-531.

9. AGHABABAEI, S. & MAKREHCHI, M. 2018. Mining Twitter data for crime trend prediction. *Intelligent Data Analysis,* 22**,** 117-141.

10. AGUIRRE, K., BADRAN, E. & MUGGAH, R. 2019. Designing and deploying predictive tools. *FUTURE CRIME:.* Igarape Institute.

11. AINLEY, E., WITWICKI, C., TALLETT, A. & GRAHAM, C. 2021. Using twitter comments to understand people's experiences of uk health care during the COVID-19 Pandemic: Thematic and sentiment analysis. *Journal of Medical Internet Research,* 23.

12. AKHTER, N., ZHAO, L., ARIAS, D., RANGWALA, H. & RAMAKRISHNAN, N. 2018. Forecasting gang homicides with multi-level multi-task learning. Springer Verlag.

13. AL BONI, M. & GERBER, M. S. Area-specific crime prediction models. 2017a. Institute of Electrical and Electronics Engineers Inc., 671-676.

14. AL BONI, M. & GERBER, M. S. Predicting crime with routine activity patterns inferred from social media. 2017b. 1233-1238.

15. AL BONI, M. & GERBER, M. S. Predicting crime with routine activity patterns inferred from social media. 2017c. Institute of Electrical and Electronics Engineers Inc., 1233-1238.

16. AL SARI, B., ALKHALDI, R., ALSAFFAR, D., ALKHALDI, T., ALMAYMUNI, H., ALNAIM, N., ALGHAMDI, N. & OLATUNJI, S. O. 2022. Sentiment analysis for cruises in Saudi Arabia on social media platforms using machine learning algorithms. *Journal of Big Data,* 9.

17. ALAMOODI, A. H., BAKER, M. R., ALBAHRI, O. S., ZAIDAN, B. B., ZAIDAN, A. A., WONG, W. K., GARFAN, S., ALBAHRI, A. S., ALONSO, M. A., JASIM, A. N. & BAQER, M. J. 2022. Public Sentiment Analysis and Topic Modeling Regarding COVID-19's Three Waves of Total Lockdown: A Case Study on Movement Control Order in Malaysia. *KSII Transactions on Internet and Information Systems,* 16**,** 2169-2190.

18. ALBAHLI, S., ALSAQABI, A., ALDHUBAYI, F., RAUF, H. T., ARIF, M. & MOHAMMED, M. A. 2020. Predicting the type of crime: Intelligence gathering and crime analysis. *Computers, Materials and Continua,* 66**,** 2317-2341.

19. ALJEDAANI, W., ABUHAIMED, I., RUSTAM, F., MKAOUER, M. W., OUNI, A. & JENHANI, I. 2022. Automatically detecting and understanding the perception of COVID-19 vaccination: a middle east case study. *Social Network Analysis and Mining,* 12.

20. ALMANIE, T., MIRZA, R. & LOR, E. 2015. Crime prediction based on crime types and using spatial and temporal criminal hotspots. *arXiv preprint arXiv:1508.02050*.

21. ALMAW, A. & KADAM, K. Crime Data Analysis and Prediction Using Ensemble Learning. 2019. Institute of Electrical and Electronics Engineers Inc., 1918-1923.

22. ALMEHMADI, A., JOUDAKI, Z. & JALALI, R. Language Usage on Twitter Predicts Crime Rates. 2017. Association for Computing Machinery, 307-310.

23. ALSAQABI, A., ALDHUBAYI, F. & ALBAHLI, S. Using machine learning for prediction of factors affecting crimes in Saudi Arabia. 2019. ICST, 57-62.

24. ALTINEL, A. B., BUZLU, K. & IPEK, K. Performance Analysis of Different Sentiment Polarity Dictionaries on Turkish Sentiment Detection. 2022. Institute of Electrical and Electronics Engineers Inc.

25. ALVES, L. G., RIBEIRO, H. V. & RODRIGUES, F. A. 2018. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications,* 505**,** 435-443.

26. ANDERSON, R. G. 2009. THE NEW OXFORD COMPANION TO LAW. Ed by P Cane and J Conaghan Oxford: Oxford University Press (www. oup. co. uk), 2008. lxxxiii+ 1306pp. ISBN 9780199290543.£ 39.95. Edinburgh University Press 22 George Square, Edinburgh EH8 9LF UK.

27. ANGUITA, D., GHIO, A., GRECO, N., ONETO, L. & RIDELLA, S. Model selection for support vector machines: Advantages and disadvantages of the machine learning theory. The 2010 international joint conference on neural networks (IJCNN), 2010. IEEE, 1-8.

28. ARIAS, F., GUERRA-ADAMES, A., ZAMBRANO, M., QUINTERO-GUERRA, E. & TEJEDOR-FLORES, N. 2022. Analyzing Spanish-Language Public Sentiment in the Context of a Pandemic and Social Unrest: The Panama Case. *International Journal of Environmental Research and Public Health,* 19.

29. ASHBY, M. P. J. 2020. Initial evidence on the relationship between the coronavirus pandemic and crime in the United States. *Crime Sci,* 9**,** 6.

30. AWAL, M. A., RABBI, J., HOSSAIN, S. I. & HASHEM, M. M. A. Using linear regression to forecast future trends in crime of Bangladesh. 2016a. Institute of Electrical and Electronics Engineers Inc., 333-338.

31. AWAL, M. A., RABBI, J., HOSSAIN, S. I. & HASHEM, M. M. A. Using linear regression to forecast future trends in crime of Bangladesh. 2016b. 333-338.

32. AZEEZ, J. & ARAVINDHAR, D. J. Hybrid approach to crime prediction using deep learning. 2015a. Institute of Electrical and Electronics Engineers Inc., 1701-1710.

33. AZEEZ, J. & ARAVINDHAR, D. J. Hybrid approach to crime prediction using deep learning. 2015b. 1701-1710.

34. BACULO, M. J. C., MARZAN, C. S., DE DIOS BULOS, R. & RUIZ, C. Geospatial-temporal analysis andclassification of criminal data in Manila. 2017. 6-11.

35. BADAWY, M., CIRELLI, J., SETYONO, H. & AQLAN, F. Analysis and visualization of city crimes. 2018. 1136-1145.

36. BALOIAN, N., BASSALETTI, E., FERNÁNDEZ, M., FIGUEROA, O., FUENTES, P., MANASEVICH, R., ORCHARD, M., PEÑAFIEL, S., PINO, J. A. & VERGARA, M. Crime prediction using patterns and context. 2017. 2-9.

37. BAPPEE, F. K., SOARES JNIOR, A. & MATWIN, S. 2018. Predicting crime using spatial features. Springer Verlag.

38. BARNUM, J. D., CAPLAN, J. M., KENNEDY, L. W. & PIZA, E. L. 2017. The crime kaleidoscope: A cross-jurisdictional analysis of place features and crime in three urban environments. *Applied Geography,* 79**,** 203-211.

39. BBC 2020. Far-right online: 'I got them back on social media'.

40. BEHDENNA, S., BARIGOU, F. & BELALEM, G. 2018. Document level sentiment analysis: a survey. *EAI Endorsed Transactions on Context-aware Systems and Applications,* 4**,** e2-e2.

41. BELESIOTIS, A., PAPADAKIS, G. & SKOUTAS, D. 2018. Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Transactions on Spatial Algorithms and Systems,* 3.

42. BERK, R. A., SORENSON, S. B. & BARNES, G. 2016. Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of empirical legal studies,* 13**,** 94-115.

43. BISWAS, A. A. & BASAK, S. Forecasting the Trends and Patterns of Crime in Bangladesh using Machine Learning Model. 2019a. 114-118.

44. BISWAS, A. A. & BASAK, S. Forecasting the Trends and Patterns of Crime in Bangladesh using Machine Learning Model. 2019b. Institute of Electrical and Electronics Engineers Inc., 114-118.

45. BOGOMOLOV, A., LEPRI, B., STAIANO, J., OLIVER, N., PIANESI, F. & PENTLAND, A. Once upon a crime: Towards crime prediction from demographics and mobile data. 2014. Association for Computing Machinery, Inc, 427-434.

46. BRACHMAN, R. J. & ANAND, T. The Process of Knowledge Discovery in Databases: A First Sketch. KDD workshop, 1994. 1-12.

47. BRAMER, W. M., RETHLEFSEN, M. L., KLEIJNEN, J. & FRANCO, O. H. 2017. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Systematic reviews,* 6**,** 1-12.

48. BRAYNE, S. 2017. Big data surveillance: The case of policing. *American sociological review,* 82**,** 977-1008.

49. BROWNE, K., GREEN, K., JARENO-RIPOLL, S. & PADDOCK, E. 2022. Knife crime offender characteristics and interventions–A systematic review. *Aggression and Violent Behavior***,** 101774.

50. BUTT, U. M., LETCHMUNAN, S., HASSAN, F. H., ALI, M., BAQIR, A., KOH, T. W. & SHERAZI, H. H. R. 2021. Spatio-Temporal Crime Predictions by Leveraging Artificial Intelligence for Citizens Security in Smart Cities. *IEEE Access,* 9**,** 47516-47529.

51. CANE, P. & CONAGHAN, J. 2008. The new Oxford companion to law.

52. CANTON, H. 2021. United Nations Office on Drugs and Crime—UNODC. *The Europa Directory of International Organizations 2021.* Routledge.

53. CAPLAN, J. M., KENNEDY, L. W. & MILLER, J. 2011. Risk Terrain Modeling: Brokering Criminological Theory and GIS Methods for Crime Forecasting. *Justice Quarterly,* 28**,** 360-381.

54. CAPLAN, J. M., KENNEDY, L. W., PIZA, E. L. & BARNUM, J. D. 2020. Using Vulnerability and Exposure to Improve Robbery Prediction and Target Area Selection. *Applied Spatial Analysis and Policy,* 13**,** 113-136.

55. CASTRO, R. M., RODRIGUES, M. W. & BRANDO, W. C. Predicting crime by exploiting supervised learning on heterogeneous data. 2020. SciTePress, 524-531.

56. CATELLI, R., CASOLA, V., DE PIETRO, G., FUJITA, H. & ESPOSITO, M. 2021. Combining contextualized word representation and sub-document level analysis through Bi-LSTM+ CRF architecture for clinical de-identification. *Knowledge-Based Systems,* 213**,** 106649.

57. CATLETT, C., CESARIO, E., TALIA, D. & VINCI, A. A data-driven approach for spatio-Temporal crime predictions in smart cities. 2018. 17-24.

58. CAVADAS, B., BRANCO, P. & PEREIRA, S. 2015. Crime prediction using regression and resources optimization. Springer Verlag.

59. CESUR, R., CEYHAN, E. B., KERMEN, A. & SAIROLU 2017. Determination of potential criminals in social network. *Gazi University Journal of Science,* 30**,** 121-131.

60. CHAINEY, S. 2013. Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *BSGLg,* 60**,** 7-19.

61. CHAINEY, S., TOMPSON, L. & UHLIG, S. 2008. The utility of hotspot mapping for predicting spatial patterns of crime. *Security journal,* 21**,** 4-28.

62. CHAN, J. C.-W. & PAELINCKX, D. 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment,* 112**,** 2999-3011.

63. CHANDRASEKAR, A., RAJ, A. S. & KUMAR, P. 2015. Crime prediction and classification in San Francisco City. *URL http://cs229. stanford. edu/proj2015/228 {_} report. pdf.*

64. CHEN, H.-L., HUANG, C.-C., YU, X.-G., XU, X., SUN, X., WANG, G. & WANG, S.-J. 2013a. An efficient diagnosis system for detection of Parkinson’s disease using fuzzy k-nearest neighbor approach. *Expert systems with applications,* 40**,** 263-271.

65. CHEN, P., YUAN, H. & LI, D. 2013b. Space-time analysis of burglary in Beijing. *Security Journal,* 26**,** 1-15.

66. CHEN, R.-C., CARAKA, R. E., ARNITA, N. E. G., POMALINGO, S., RACHMAN, A., TOHARUDIN, T., TAI, S.-K. & PARDAMEAN, B. 2020. An end to end of scalable tree boosting system. *Sylwan,* 165**,** 1-11.

67. CHEN, X., CHO, Y. & JANG, S. Y. Crime prediction using Twitter sentiment and weather. 2015 Systems and Information Engineering Design Symposium, 2015a. IEEE, 63-68.

68. CHEN, X., CHO, Y. & JANG, S. Y. Crime prediction using Twitter sentiment and weather. 2015b. Institute of Electrical and Electronics Engineers Inc., 63-68.

69. CORSO, A. J. 2015. Toward predictive crime analysis via social media, big data, and gis spatial correlation. *iConference 2015 Proceedings.*

70. COSTA, C., APARICIO, M. & APARICIO, J. Sentiment Analysis of Portuguese Political Parties Communication. 2021. Association for Computing Machinery, Inc, 63-69.

71. CURIEL, R. P., CRESCI, S., MUNTEAN, C. I. & BISHOP, S. R. 2020. Crime and its fear in social media. *Palgrave Communications,* 6**,** 1-12.

72. DA SILVA, A. R. C., DE PAULA JUNIOR, I. C., DA SILVA, T. L. C., DE MACEDO, J. A. F. & SILVA, W. C. P. Prediction of crime location in a brazilian city using regression techniques. 2020. 331-336.

73. DAS, P. & DAS, A. K. 2019. Application of Classification Techniques for Prediction and Analysis of Crime in India.

74. DAS, S., KIM, A. & KARMAKAR, S. 2020. Change-point analysis of cyberbullying-related twitter discussions during covid-19. *arXiv preprint arXiv:2008.13613*.

75. DATA.POLICE.UK.

76. DENECKE, K. & DENG, Y. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine,* 64**,** 17-27.

77. DESHMUKH, A., BANKA, S., DCRUZ, S. B., SHAIKH, S. & TRIPATHY, A. K. Safety App: Crime Prediction Using GIS. 2020. 120-124.

78. DEVIKA, M., SUNITHA, C. & GANESH, A. 2016. Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science,* 87**,** 44-49.

79. DRAWVE, G. 2016. A Metric Comparison of Predictive Hot Spot Techniques and RTM. *Justice Quarterly,* 33**,** 369-397.

80. DRAWVE, G., MOAK, S. C. & BERTHELOT, E. R. 2016. Predictability of gun crimes: a comparison of hot spot and risk terrain modelling techniques. *Policing and Society,* 26**,** 312-331.

81. DUGATO, M. 2013. Assessing the validity of risk terrain modeling in a European city: Preventing robberies in the city of Milan.

82. DUGATO, M., CALDERONI, F. & BERLUSCONI, G. 2020. Forecasting Organized Crime Homicides: Risk Terrain Modeling of Camorra Violence in Naples, Italy. *Journal of Interpersonal Violence,* 35**,** 4013-4039.

83. EL BOUR, H. A., OUNACER, S., ELGHOMARI, Y., JIHAL, H. & AZZOUAZI, M. 2018. A crime prediction model based on spatial and temporal data. *Periodicals of Engineering and Natural Sciences,* 6**,** 360-364.

84. ELLURI, L., MANDALAPU, V. & ROY, N. Developing machine learning based predictive models for smart policing. 2019. 198-204.

85. ESQUIVEL, N., NICOLIS, O., PERALTA, B. & MATEU, J. 2020. Spatio-Temporal Prediction of Baltimore Crime Events Using CLSTM Neural Networks. *IEEE Access,* 8**,** 209101-209112.

86. ESQUIVEL, N., PERALTA, B. & NICOLIS, O. Crime Level Prediction using Stacked Maps with Deep Convolutional Autoencoder. 2019. Institute of Electrical and Electronics Engineers Inc.

87. ESTVEZ-SOTO, P. R. 2021. Crime and COVID-19: effect of changes in routine activities in Mexico City. *Crime Science,* 10.

88. FEATHERSTONE, C. The relevance of social media as it applies in South Africa to crime prediction. 2013.

89. FENG, M., ZHENG, J., HAN, Y., REN, J. & LIU, Q. 2018. Big Data Analytics and Mining for Crime Data Analysis, Visualization and Prediction.

90. FLAXMAN, S., CHIRICO, M., PEREIRA, P. A. U. & LOEFFLER, C. 2019. Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ "real-time crime forecasting challenge". *Annals of Applied Statistics,* 13**,** 2564-2585.

91. FORRADELLAS, R. F. R., ALONSO, S. L. N., RODRIGUEZ, M. L. & JORGE-VAZQUEZ, J. 2021. Applied machine learning in social sciences: Neural networks and crime prediction. *Social Sciences,* 10**,** 1-20.

92. GAYATHRI, Y., SRI LALITHA, Y., ADITYA NAG, M. V. & ALTHAF HUSSAIN BASHA, S. 2021. Data-Driven Prediction Model for Crime Patterns.

93. GERELL, M. 2018. Bus Stops and Violence, Are Risky Places Really Risky? *European Journal on Criminal Policy and Research,* 24**,** 351-371.

94. GOODWILL, A. M., VAN DER KEMP, J. J. & WINTER, J.-M. 2013. Applied geographical profiling. *Encyclopedia of criminology and criminal justice***,** 86-99.

95. GUPTA, B., RAWAT, A., JAIN, A., ARORA, A. & DHAMI, N. 2017. Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications,* 163**,** 15-19.

96. HADDAWAY, N. R., COLLINS, A. M., COUGHLIN, D. & KIRK, S. 2015. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PloS one,* 10**,** e0138237.

97. HAJELA, G., CHAWLA, M. & RASOOL, A. A Clustering Based Hotspot Identification Approach for Crime Prediction. 2020. 1462-1470.

98. HALVANI, O., STEINEBACH, M. & ZIMMERMANN, R. 2013. Authorship verification via k-nearest neighbor estimation. *Notebook PAN at CLEF*.

99. HAN, X., HU, X., WU, H., SHEN, B. & WU, J. 2020. Risk Prediction of Theft Crimes in Urban Communities: An Integrated Model of LSTM and ST-GCN. *IEEE Access,* 8**,** 217222-217230.

100. HART, T. & ZANDBERGEN, P. 2014. Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting. *Policing,* 37**,** 305-323.

101. HMEIDI, I., HAWASHIN, B. & EL-QAWASMEH, E. 2008. Performance of KNN and SVM classifiers on full word Arabic articles. *Advanced Engineering Informatics,* 22**,** 106-111.

102. HOSSAIN, S., ABTAHEE, A., KASHEM, I., HOQUE, M. M. & SARKER, I. H. Crime prediction using spatio-temporal data. International Conference on Computing Science, Communication and Security, 2020a. Springer, 277-289.

103. HOSSAIN, S., ABTAHEE, A., KASHEM, I., HOQUE, M. M. & SARKER, I. H. 2020b. Crime prediction using spatio-temporal data. Springer.

104. HU, T., ZHU, X., DUAN, L. & GUO, W. 2018. Urban crime prediction based on spatiotemporal Bayesian model. *PLoS ONE,* 13.

105. HUANG, J., LI, Y.-F. & XIE, M. 2015. An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology,* 67**,** 108-127.

106. HUSSEIN, D. M. E.-D. M. 2018. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences,* 30**,** 330-338.

107. IPPOLITO, A. & LOZANO, A. C. G. Tax crime prediction with machine learning: A case study in the municipality of São Paulo. 2020. 452-459.

108. IQBAL, A., AMIN, R., IQBAL, J., ALROOBAEA, R., BINMAHFOUDH, A. & HUSSAIN, M. 2022. Sentiment Analysis of Consumer Reviews Using Deep Learning. *Sustainability (Switzerland),* 14.

109. IQBAL, R., MURAD, M. A. A., MUSTAPHA, A., PANAHY, P. H. S. & KHANAHMADLIRAVI, N. 2013. An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology,* 6**,** 4219-4225.

110. JADHAV, S. D. & CHANNE, H. 2016. Comparative study of KNN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR),* 5**,** 1842-1845.

111. JAHROMI, A. H. & TAHERI, M. A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. 2017 Artificial intelligence and signal processing conference (AISP), 2017. IEEE, 209-212.

112. JAIN, M., BHALLA, G., JAIN, A. & SHARMA, S. 2022a. Automatic keyword extraction for localized tweets using fuzzy graph connectivity measures. *Multimedia Tools and Applications,* 81**,** 42931-42956.

113. JAIN, P. K., QUAMER, W., SARAVANAN, V. & PAMULA, R. 2022b. Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis. *Journal of Ambient Intelligence and Humanized Computing*.

114. JUNIOR, F. C. F. N., DA SILVA, T. L. C., DE QUEIROZ NETO, J. F., DE MACDO, J. A. F. & PORCINO, W. C. A novel approach to approximate crime hotspots to the road network. 2019. Association for Computing Machinery, Inc, 53-61.

115. KADAR, C. & PLETIKOSA, I. 2018. Mining large-scale human mobility data for long-term crime prediction. *EPJ Data Science*, 7.

116. KADAR, C., ROSÉS BRÜNGGER, R. & PLETIKOSA, I. 2017. Measuring ambient population from location-based social networks to describe urban crime.

117. KADAR, C., ROULY, C., ROSÉS, R. & GERRITSEN, C. Agent-based simulation of offender mobility: Integrating activity nodes from location-based social networks. 2018. 804-812.

118. KARMAKAR, S. & DAS, S. Evaluating the Impact of COVID-19 on Cyberbullying through Bayesian Trend Analysis. 2020.

119. KARMAKAR, S. & DAS, S. Understanding the rise of twitter-based cyberbullying due to COVID-19 through comprehensive statistical evaluation. 2021. IEEE Computer Society, 2521-2531.

120. KE, Z. & JIN, Z. 2014. Research of crime prediction technology based on mathematical model. *Open Cybernetics and Systemics Journal,* 8**,** 860-868.

121. KEDIA, P. 2016. Crime mapping and analysis using GIS. *International Institute of Information Technology,* 1**,** 1-15.

122. KHATUN, R., AYON, S. I., HOSSAIN, R. & ALAM, J. 2020. Data mining technique to analyse and predict crime using crime categories and arrest records. *Indonesian Journal of Electrical Engineering and Computer Science,* 22**,** 444-452.

123. KIM, D., JUNG, S. & JEONG, Y. 2021. Theft prediction model based on spatial clustering to reflect spatial characteristics of adjacent lands. *Sustainability (Switzerland),* 13.

124. KIM, S., JOSHI, P., KALSI, P. S. & TAHERI, P. Crime analysis through machine learning. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018. IEEE, 415-420.

125. KIRAN, J. & KAISHVEEN, K. Prediction analysis of crime in India using a hybrid clustering approach. 2019. Institute of Electrical and Electronics Engineers Inc., 520-523.

126. KISSI GHALLEB, A. E. & BEN AMARA, N. E. Terrorist Act Prediction Based on Machine Learning: Case Study of Tunisia. 2020. Institute of Electrical and Electronics Engineers Inc., 398-403.

127. KOSTAKOS, P., ROBROO, S., LIN, B. & OUSSALAH, M. Crime prediction using hotel reviews? , 2019. Institute of Electrical and Electronics Engineers Inc., 134-137.

128. KOUNADI, O., LAMPOLTSHAMMER, T. J., GROFF, E., SITKO, I. & LEITNER, M. 2015. Exploring Twitter to analyze the public's reaction patterns to recently reported homicides in London. *PloS one,* 10**,** e0121848.

129. KSHATRI, S. S., SINGH, D., NARAIN, B., BHATIA, S., QUASIM, M. T. & SINHA, G. R. 2021. An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach. *IEEE Access,* 9**,** 67488-67500.

130. LAMARI, Y., FRESKURA, B., ABDESSAMAD, A., EICHBERG, S. & DE BONVILLER, S. 2020. Predicting spatial crime occurrences through an efficient ensemble-learning model. *ISPRS International Journal of Geo-Information,* 9.

131. LEE, I., JUNG, S., LEE, J. & MACDONALD, E. 2019. Street crime prediction model based on the physical characteristics of a streetscape: Analysis of streets in low-rise housing areas in South Korea. *Environment and Planning B: Urban Analytics and City Science,* 46**,** 862-879.

132. LEKHA, K. C. & PRAKASAM, S. Data mining techniques in detecting and predicting cyber crimes in banking sector. 2018. 1639-1643.

133. LIBERATI, A., ALTMAN, D. G., TETZLAFF, J., MULROW, C., GØTZSCHE, P. C., IOANNIDIS, J. P., CLARKE, M., DEVEREAUX, P. J., KLEIJNEN, J. & MOHER, D. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology,* 62**,** e1-e34.

134. LIN, R. Comment Texts Sentiment Analysis Based on Improved Bi-LSTM and Naive Bayes. 2022. Institute of Electrical and Electronics Engineers Inc., 407-412.

135. LLAHA, O. Crime analysis and prediction using machine learning. 2020. 496-501.

136. LUO, J., QIU, S., PAN, X., YANG, K. & TIAN, Y. 2022. Exploration of Spa Leisure Consumption Sentiment towards Different Holidays and Different Cities through Online Reviews: Implications for Customer Segmentation. *Sustainability (Switzerland),* 14.

137. MAHMUD, N., ZINNAH, K. I., RAHMAN, Y. A. & AHMED, N. CRIMECAST: A crime prediction and strategy direction service. 2017. Institute of Electrical and Electronics Engineers Inc., 414-418.

138. MAI, L. & LE, B. 2021. Joint sentence and aspect-level sentiment analysis of product comments. *Annals of Operations research,* 300**,** 493-513.

139. MALIK, A., MACIEJEWSKI, R., TOWERS, S., MCCULLOUGH, S. & EBERT, D. S. 2014. Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement. *IEEE transactions on visualization and computer graphics,* 20**,** 1863-1872.

140. MARK, E. June, 2010. *What is crime?* [Online]. BBC. Available: https://www.bbc.co.uk/blogs/thereporters/markeaston/2010/06/what_is_crime.html [Accessed 12/10/2022 2022].

141. MEDVEDEVA, M., AGBOZO, E. & NAVIVAYKO, D. 2016. Automatic detection of abuse on social media. *16th International Multidisciplinary Scientific Geoconference (SGEM 2016)-Albena, Bulgaria,* 30.

142. MINGCHE, S., HANYU, L., YIMING, Q. & XIAOHANG, Z. 2011. Crime location prediction based on the maximum-likelihood theory. NA.

143. MISHRA, S., SAHOO, S., RANJAN, P. & PANDA, A. R. 2021. Machine Learning Approach in Crime Records Evaluation.

144. MISYRLIS, M., CHEUNG, C. M., SRIVASTAVA, A., KANNAN, R. & PRASANNA, V. Spatio-temporal modeling of criminal activity. 2017. 3-8.

145. MOHLER, G. 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting,* 30**,** 491-497.

146. MUSLEH, D. A., ALKHALES, T. A., ALMAKKI, R. A., ALNAJIM, S. E., ALMARSHAD, S. K., ALHASANIAH, R. S., ALJAMEEL, S. S. & ALMUQHIM, A. A. 2022. Twitter arabic sentiment analysis to detect depression using machine learning. *Computers, Materials and Continua,* 71**,** 3463-3477.

147. NA, C., OH, G., SONG, J. & PARK, H. 2021. Do machine learning methods outperform traditional statistical models in crime prediction? A comparison between logistic regression and neural networks. *Korean Journal of Policy Studies,* 36**,** 1-13.

148. NESA, M., SHAHA, T. R. & YOON, Y. 2022. Prediction of juvenile crime in Bangladesh due to drug addiction using machine learning and explainable AI techniques. *Journal of Computational Social Science***,** 1-21.

149. NEWS, B. 2019. *Christchurch shootings: 49 dead in New Zealand mosque attacks* [Online]. BBC NEWS. Available: https://www.bbc.co.uk/news/world-asia-47578798 [Accessed 12/10/2022 2022].

150. NGUYEN, T. T., HATUA, A. & SUNG, A. H. 2017. Building a learning machine classifier with inadequate data for crime prediction. *Journal of Advances in Information Technology Vol,* 8.

151. NITTA, G. R., RAO, B. Y., SRAVANI, T., RAMAKRISHIAH, N. & BALAANAND, M. 2019. LASSO-based feature selection and naïve Bayes classifier for crime prediction and its type. *Service Oriented Computing and Applications,* 13**,** 187-197.

152. O'NEILL, M., DE MAILLARD, J. & VAN STEDEN, R. 2022. The enforcement turn in plural policing? A comparative analysis of public police auxiliaries in England & Wales, France and The Netherlands. *European Journal of Criminology***,** 14773708211070203.

153. OBAGBUWA, I. C. & ABIDOYE, A. P. 2021. South Africa Crime Visualization, Trends Analysis, and Prediction Using Machine Learning Linear Regression Technique. *Applied Computational Intelligence and Soft Computing,* 2021.

154. OHYAMA, T. & AMEMIYA, M. 2018. Applying Crime Prediction Techniques to Japan: A Comparison Between Risk Terrain Modeling and Other Methods. *European journal on criminal policy and research,* 24**,** 469-487.

155. PAGE, M. J., MCKENZIE, J. E., BOSSUYT, P. M., BOUTRON, I., HOFFMANN, T. C., MULROW, C. D., SHAMSEER, L., TETZLAFF, J. M., AKL, E. A. & BRENNAN, S. E. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic reviews,* 10**,** 1-11.

156. PALAD, E. B. B., TANGKEKO, M. S., MAGPANTAY, L. A. K. & SIPIN, G. L. Document Classification of Filipino Online Scam Incident Text using Data Mining Techniques. Proceedings - 2019 19th International Symposium on Communications and Information Technologies, ISCIT 2019, 2019. 232-237.

157. PARVEZ, M. R., MOSHARRAF, T. & ALI, M. E. A novel approach to identify spatio-temporal crime pattern in Dhaka city. 2016.

158. PATIL, A. P., NAWAL, D. J. & JAIN, D. 2020. Crime Prediction Application Using Artificial Intelligence. *Proceedings of ICETIT 2019.* Springer.

159. PERRY, W. L. 2013. *Predictive policing: The role of crime forecasting in law enforcement operations*, Rand Corporation.

160. PIERCE, M., HAYHURST, K., BIRD, S. M., HICKMAN, M., SEDDON, T., DUNN, G. & MILLAR, T. 2017. Insights into the link between drug use and criminality: Lifetime offending of criminally-active opiate users. *Drug and alcohol dependence*, 179**,** 309-316.

161. PORWAL, A., CARRANZA, E. & HALE, M. 2004. A hybrid neuro-fuzzy model for mineral potential mapping. *Mathematical geology,* 36**,** 803-826.

162. QIN, Z. & RONCHIERI, E. 2022. Exploring Pandemics Events on Twitter by Using Sentiment Analysis and Topic Modelling. *Applied Sciences (Switzerland),* 12.

163. RAHMANI, M. H. 2014. Predicting crimes using time series model and ARCGIS software. *Biosciences Biotechnology Research Asia,* 11**,** 1841-1847.

164. RISTEA, A., AL BONI, M., RESCH, B., GERBER, M. S. & LEITNER, M. 2020. Spatial crime distribution and prediction for sporting events using social media. *International Journal of Geographical Information Science,* 34**,** 1708-1739.

165. RISTEA, A., KOUNADI, O. & LEITNER, M. Geosocial media data as predictors in a GWR application to forecast crime hotspots. 2018.

166. RUMMENS, A. & HARDYNS, W. 2020. Comparison of near-Repeat, Machine Learning and Risk Terrain Modeling for Making Spatiotemporal Predictions of Crime. *Applied Spatial Analysis and Policy,* 13**,** 1035-1053.

167. RUMMENS, A. & HARDYNS, W. 2021. The effect of spatiotemporal resolution on predictive policing model performance. *International Journal of Forecasting,* 37**,** 125-133.

168. RUMMENS, A., HARDYNS, W. & PAUWELS, L. 2017. The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. *Applied Geography,* 86**,** 255-261.

169. SAFAT, W., ASGHAR, S. & GILLANI, S. A. 2021. Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. *IEEE Access,* 9**,** 70080-70094.

170. SALTZ, J. S. & HOTZ, N. Identifying the most common frameworks data science teams use to structure and coordinate their projects. 2020 IEEE International Conference on Big Data (Big Data), 2020. IEEE, 2038-2042.

171. SANDAGIRI, S. P. C. W., KUMARA, B. T. G. S. & KUHANESWARAN, B. ANN Based Crime Detection and Prediction using Twitter Posts and Weather Data. 2020.

172. SARMA, D., MITTRA, T., BAWM, R. M., SARWAR, T., LIMA, F. F. & HOSSAIN, S. 2021. Comparative Analysis of Machine Learning Algorithms for Phishing Website Detection. Springer Science and Business Media Deutschland GmbH.

173. SATHYADEVAN, S. Crime analysis and prediction using data mining. 2014 First International Conference on Networks & Soft Computing (ICNSC2014), 2014. IEEE, 406-412.

174. SHARMA, H. K., CHOUDHURY, T. & KANDWAL, A. 2021. Machine learning based analytical approach for geographical analysis and prediction of Boston City crime using geospatial dataset. *GeoJournal*.

175. SHI, S., CHEN, P., ZENG, Z. & HU, X. 2021. STL-FNN: An Intelligent Prediction Model of Daily Theft Level.

176. SHI, T. A Method of Predicting Crime of Theft Based on Bagging Ensemble Feature Selection. 2020. 140-143.

177. SHUKLA, A., KATAL, A., RAGHUVANSHI, S. & SHARMA, S. Criminal Combat: Crime Analysis and Prediction Using Machine Learning. 2021.

178. SINGH, A., HALGAMUGE, M. N. & LAKSHMIGANTHAN, R. 2017. Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms. *International Journal of Advanced Computer Science and Applications,* 8.

179. SINGH, M., BHATT, M. W., BEDI, H. S. & MISHRA, U. 2020. Performance of bernoulli's naive bayes classifier in the detection of fake news. *Materials Today: Proceedings*.

180. SIROTKIN, P. 2013. On search engine evaluation metrics. *arXiv preprint arXiv:1302.2318*.

181. SONG, K., YAN, F., DING, T., GAO, L. & LU, S. 2020. A steel property optimization model based on the XGBoost algorithm and improved PSO. *Computational Materials Science,* 174**,** 109472.

182. SPEISER, J. L., MILLER, M. E., TOOZE, J. & IP, E. 2019. A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications,* 134**,** 93-101.

183. STATISTICS, O. F. N. 2022. *Crime in England and Wales: year ending March 2022* [Online]. Office for National Statistics. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingmarch2022 [Accessed 12/10/2022 12/10/2022].

184. STATNIKOV, A., WANG, L. & ALIFERIS, C. F. 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics,* 9**,** 1-10.

185. SUFI, F. K. & KHALIL, I. 2022. Automated Disaster Monitoring From Social Media Posts Using AI-Based Location Intelligence and Sentiment Analysis. *IEEE Transactions on Computational Social Systems*.

186. SUN, T. 2019. *LAWLESS BRITAIN Shocking moment hooded gang of teens ransack JD Sports in Halloween 'shoplifting spree'* [Online]. THE SUN. Available: https://www.thesun.co.uk/news/10276108/gang-ransack-jd-sports-halloween/ [Accessed 12/10/2022 2022].

187. TANG, Y., ZHU, X., GUO, W., WU, L. & FAN, Y. 2019. Anisotropic diffusion for improved crime prediction in urban China. *ISPRS International Journal of Geo-Information,* 8.

188. TARSHA KURDI, F., AMAKHCHAN, W. & GHARINEIAT, Z. 2021. Random forest machine learning technique for automatic vegetation detection and modelling in LiDAR data. *International Journal of Environmental Sciences and Natural Resources,* 28.

189. TAYLOR, R. B. & HALE, M. 2017. Criminology: Testing alternative models of fear of crime. *The Fear of Crime.* Routledge.

190. TOPPIREDDY, H. K. R., SAINI, B. & MAHAJAN, G. Crime Prediction & Monitoring Framework Based on Spatial Analysis. 2018. Elsevier B.V., 696-705.

191. TUNDIS, A., JAIN, A., BHATIA, G. & MUHLHAUSER, M. Similarity analysis of criminals on social networks: an example on twitter. 2019. Institute of Electrical and Electronics Engineers Inc.

192. UMAIR, A., SARFRAZ, M. S., AHMAD, M., HABIB, U., ULLAH, M. H. & MAZZARA, M. 2020. Spatiotemporal analysis of web news archives for crime prediction. *Applied Sciences (Switzerland),* 10**,** 1-16.

193. VAKHITOVA, Z. I., ALSTON-KNOX, C. L., REEVES, E. & MAWBY, R. I. 2021. Explaining victim impact from cyber abuse: An exploratory mixed methods analysis. *Deviant Behavior***,** 1-20.

194. VINCENZI, S., ZUCCHETTA, M., FRANZOI, P., PELLIZZATO, M., PRANOVI, F., DE LEO, G. A. & TORRICELLI, P. 2011. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. *Ecological Modelling,* 222**,** 1471-1478.

195. VISHWAMITRA, N., HU, R. R., LUO, F., CHENG, L., COSTELLO, M. & YANG, Y. On Analyzing COVID-19-related Hate Speech Using BERT Attention. 2020. Institute of Electrical and Electronics Engineers Inc., 669-676.

196. VOMFELL, L., HÄRDLE, W. K. & LESSMANN, S. 2018a. Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems,* 113**,** 73-85.

197. VOMFELL, L., HRDLE, W. K. & LESSMANN, S. 2018b. Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems,* 113**,** 73-85.

198. WANG, H. & MA, S. 2021. Preventing crimes against public health with artificial intelligence and machine learning capabilities. *Socio-Economic Planning Sciences*.

199. WANG, J., HU, J., SHEN, S., ZHUANG, J. & NI, S. 2020a. Crime risk analysis through big data algorithm with urban metrics. *Physica A: Statistical Mechanics and its Applications,* 545.

200. WANG, M. & GERBER, M. S. Using twitter for next-place prediction, with an application to crime prediction. 2015. Institute of Electrical and Electronics Engineers Inc., 941-948.

201. WANG, N., ZHAO, S., CUI, S. & FAN, W. 2021. A hybrid ensemble learning method for the identification of gang-related arson cases. *Knowledge-Based Systems,* 218.

202. WANG, S., LI, M., YU, B., BAO, S. & CHEN, Y. 2022. Investigating the Impacting Factors on the Public's Attitudes towards Autonomous Vehicles Using Sentiment Analysis from Social Media Data. *Sustainability (Switzerland),* 14.

203. WANG, S. & YUAN, K. Spatiotemporal Analysis and Prediction of Crime Events in Atlanta Using Deep Learning. 2019. Institute of Electrical and Electronics Engineers Inc., 346-350.

204. WANG, X., GERBER, M. S. & BROWN, D. E. 2012. Automatic crime prediction using events extracted from twitter posts. NA.

205. WANG, Y., GE, L., LI, S. & CHANG, F. 2020b. Deep Temporal Multi-Graph Convolutional Network for Crime Prediction. Springer Science and Business Media Deutschland GmbH.

206. WANG, Y., YU, W., LIU, S. & YOUNG, S. D. 2019. The Relationship Between Social Media Data and Crime Rates in the United States. *Social Media + Society,* 5.

207. WANG, Z. & LIU, X. 2017. Analysis of burglary hot spots and near-repeat victimization in a large Chinese city. *ISPRS International Journal of Geo-Information,* 6.

208. WIJAYA, S. S., ANUGERAH AYU, M. & MANTORO, T. Providing Real-time Crime Statistics in Indonesia Using Data Mining Approach. 2019. Institute of Electrical and Electronics Engineers Inc.

209. WIJENAYAKE, S., GRAHAM, T. & CHRISTEN, P. 2018. A Decision Tree Approach to Predicting Recidivism in Domestic Violence.

210. WILLIAMS, M. L., BURNAP, P., JAVED, A., LIU, H. & OZALP, S. 2019. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology*.

211. WILLIAMS, M. L., BURNAP, P. & SLOAN, L. 2017. Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *British Journal of Criminology,* 57**,** 320-340.

212. WIRTH, R. & HIPP, J. CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000. Manchester, 29-39.

213. WU, J., LI, Y. & MA, Y. Comparison of XGBoost and the Neural Network model on the class-balanced datasets. 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC), 2021. IEEE, 457-461.

214. YANG, D., HEANEY, T., TONON, A., WANG, L. & CUDR-MAUROUX, P. 2018a. CrimeTelescope: crime hotspot prediction based on urban and social media data fusion. *World Wide Web,* 21**,** 1323-1347.

215. YANG, D., HEANEY, T., TONON, A., WANG, L. & CUDRÉ-MAUROUX, P. 2018b. CrimeTelescope: crime hotspot prediction based on urban and social media data fusion. *World Wide Web,* 21**,** 1323-1347.

216. YANG, S., ROSENFELD, J. & MAKUTONIN, J. 2018c. Financial aspect-based sentiment analysis using deep representations. *arXiv preprint arXiv:1808.07931*.

217. YAO, S., WEI, M., YAN, L., WANG, C., DONG, X., LIU, F. & XIONG, Y. Prediction of crime hotspots based on spatial factors of random forest. 2020. 811-815.

218. YU, H., LIU, L., YANG, B. & LAN, M. 2020. Crime Prediction with Historical Crime and Movement Data of Potential Offenders Using a Spatio-Temporal Cokriging Method. *ISPRS International Journal of Geo-Information,* 9.

219. ZHANG, C., CAI, M., ZHAO, X., CAO, L. & WANG, D. 2020a. Prediction model of suspect number based on deep learning.

220. ZHANG, M., LI, H., PAN, S., LYU, J., LING, S. & SU, S. 2021. Convolutional neural networks-based lung nodule classification: A surrogate-assisted evolutionary algorithm for hyperparameter optimization. *IEEE Transactions on Evolutionary Computation,* 25**,** 869-882.

221. ZHANG, Q., YUAN, P., ZHOU, Q. & YANG, Z. Mixed spatial-temporal characteristics based crime hot spots prediction. 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2016. IEEE, 97-101.

222. ZHANG, X., LIU, L., XIAO, L. & JI, J. 2020b. Comparison of machine learning algorithms for predicting crime hotspots. *IEEE Access,* 8**,** 181302-181310.

223. ZHAO, X. & TANG, J. Exploring transfer learning for crime prediction. 2017a. IEEE Computer Society, 1158-1159.

224. ZHAO, X. & TANG, J. Modeling temporal-spatial correlations for crime prediction. 2017b. 497-506.

225. ZHOU, F., JIANXIN JIAO, R. & LINSEY, J. S. 2015. Latent customer needs elicitation by use case analogical reasoning from sentiment analysis of online product reviews. *Journal of Mechanical Design,* 137**,** 071401.