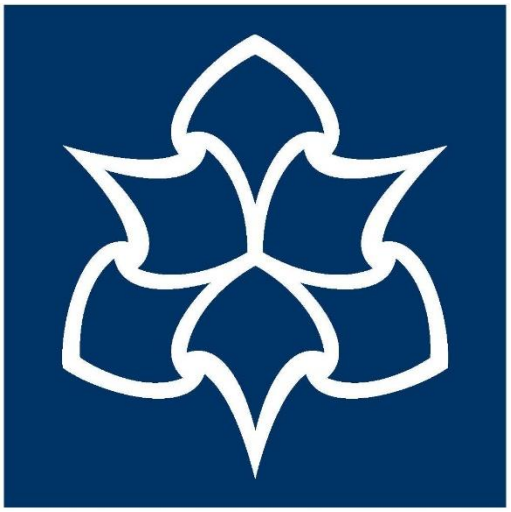
Leveraging NLP in Crime Prediction & Analysis

A DISSERTATION SUBMITTED TO MANCHESTER METROPOLITAN UNIVERSITY FOR THE DEGREE OF MASTER OF SCIENCE IN THE FACULTY OF SCIENCE AND ENGINEERING



September 2024

Author:

Divyanshu Sahwal

Supervisor:

Prof. Sergio Davies

Abstract

Preventing and anticipating crime are crucial elements of contemporary policing, particularly in urban regions where crime levels can quickly change. This study investigates how combining machine learning (ML) and natural language processing (NLP) can improve the forecasting and examination of crime trends by utilizing both structured (such as geolocation, time, crime type) and unstructured data (like crime reports, descriptions). More specifically, this project utilizes methods like Bag of Words (BoW) for analyzing text and algorithms such as K-Nearest Neighbors (KNN), Decision Trees, and Random Forests for making predictions.

This research is inspired by the major obstacles data scientists encounter when working with unstructured data, especially text-based details, which are often neglected in conventional crime analysis approaches. Textual reports, a form of unstructured data, can provide important information on criminal behavior and contextual details that structured data is incapable of capturing. By utilizing NLP and BoW, this study converts textual crime reports into numerical representations that are meaningful, enabling more precise and detailed predictions.

The study also focuses on a key deficiency in combining spatial, temporal, and text data for forecasting crimes. Although traditional techniques predominantly utilize structured data, incorporating unstructured text data into this project offers a comprehensive insight into crime patterns.

This project adds to the data science domain by providing a method for managing and uncovering insights from unstructured text data, which is a common challenge for data experts. The project demonstrates how BoW and similar methods can streamline intricate text data and incorporate it into machine learning models, showcasing a versatile approach for domains with abundant unstructured data. Moreover, the developed predictive models can act as a basis for upcoming studies in crime analysis and other areas that need text data.

Declaration

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work. This work has been carried out in accordance with the Manchester Metropolitan University research ethics procedures and has received ethical approval number 68848.

Signed_Divyanshu Sahwal

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Prof. Sergio Davies, for his invaluable support and guidance throughout the duration of this project. His suggestions and insights have been crucial in shaping both the direction and scope of my research.

Mr. Davies has consistently provided expert advice and encouragement, while also challenging me to push the boundaries of my analytical and problem-solving skills. His commitment to excellence and innovation has deeply inspired me, and his readiness to engage in thoughtful discussions and provide feedback during our regular meetings has significantly contributed to the success of this project.

I am truly thankful for his mentorship, patience, and expertise, which have not only aided me in completing this project but have also greatly enriched my learning experience. His dedication as a supervisor has been exemplary, and I am profoundly grateful for his unwavering support.

Abbreviations

1. NLP - Natural Language Processing
2. BoW- Bag of Words
3. TF-IDF - Term Frequency-Inverse Document Frequency
4. KNN - K-Nearest Neighbours
5. LSOA - Lower Layer Super Output Area
6. PCA - Principal Component Analysis
7. ML- Machine Learning
8. AI - Artificial Intelligence
9. RF - Random Forest
10. DT - Decision Tree
11. K-means - K-Means Clustering
12. GIS- Geographic Information System
13. API - Application Programming Interface
14. CSV - Comma-Separated Values
15. SQL - Structured Query Language
16. DB - Database
17. RMSE - Root Mean Square Error
18. MAE - Mean Absolute Error
19. ROC - Receiver Operating Characteristic
20. AUC - Area Under the Curve

Table Of Contents

List Of Figures

List Of Tables

1. Chapter I - Introduction

1.1 Background and Motivation

Crime prevention is crucial for law enforcement agencies worldwide. Predicting criminal activity using technology allows for proactive interventions and more efficient resource allocation. This study leverages the power of machine learning and NLP to predict crimes by analysing patterns in geolocation data and extracting insights from text-based reports such as police records and news articles.

This project aims to combine these models with geographic and temporal data to build more robust crime prediction systems.

Overview of Data Sources

The dataset we processed, specifically '2024-01-west-midlands-street.csv', is part of a larger collection of crime data files from various regions across the UK. These files cover regions such as Cleveland, Bedfordshire, Cambridgeshire, Avon and Somerset, City of London, and many more. Each CSV file contains street-level crime data for the specified region.

In total, there are 43 files in the 'crime data' folder, each representing a different area, such as British Transport Police, Thames Valley, Lancashire, Metropolitan, West Yorkshire, and more. These datasets likely include key details on crimes such as types of offenses, locations, and outcomes for each region, offering a comprehensive view of crime patterns across the UK.

The integration of both spatial and textual data allowed for in-depth analysis, ensuring that the data was suitable for both predictive modelling and geographic crime pattern detection.

The data structure contained key columns such as `Crime type`, `Location`, `Last outcome category`, `Latitude`, and `Longitude`. Additionally, other columns such as `Zone` were derived using clustering techniques, transforming latitude and longitude values into identifiable zones.

By applying machine learning models to this vectorized data, we can predict the outcome of a crime incident based on its characteristics.

Outcome Simplification: The vectorization of the 'Last outcome category' allows for independent analysis, revealing the primary nature of how outcomes are described and possibly identifying predominant themes or keywords in crime resolutions.

1.2 Research Objectives

- Develop predictive models for crime hotspots using spatial and temporal data.

- Apply NLP and ML techniques to analyse text-based crime reports and news articles.
- Explore the relationship between crime type, location, and time for more accurate predictions.
- Evaluate model performance through accuracy, precision, and recall metrics

1.3 Research Questions

- How can geolocation data clustering enhance crime prediction accuracy?
- What role do temporal variables play in understanding crime patterns?
- Can NLP models like BERT or TFIDF be effectively used to analyse crime reports and enhance predictions?

1.4 Research Contributions

- Developed geolocation clustering methods to improve crime prediction accuracy.
- Applied Machine Learning models to extract meaningful features from unstructured crime reports.
- Identified significant spatial and temporal correlations within crime data.

1.5 Thesis Structure

- **Chapter 2:** Literature Review: Reviews previous work in crime prediction, geospatial analysis, and NLP-based crime text analysis.
- **Chapter 3:** Methodology: Explains data preprocessing, geolocation clustering, and NLP methods.
- **Chapter 4:** Results and Analysis: Details findings from geospatial clustering and text-based analysis.
- **Chapter 5**: Conclusion and Future Work: Discusses research implications, limitations, and future research direction

2. Chapter II – Literature Review

2.1 Introduction

Natural Language Processing and machine learning techniques such as K-means clustering have proven effective for analysing and predicting criminal activities. The increasing availability of crime-related data in unstructured formats, such as police reports and social media posts, demands advanced techniques to convert this data into actionable insights. In this literature review, we will explore the use of vectorization techniques like BOW, Term Frequency-Inverse Document Frequency (TF-IDF), and K-means clustering for crime hotspot detection.

This review is based on several sources, emphasizing both theoretical foundations and practical applications. Traditional crime prediction models have relied on statistical approaches such as linear regression, decision trees, and random forests. These methods often focus on structured data, such as crime location, time, and type. While effective, these models struggle with unstructured data such as police reports and news articles. The introduction of machine learning techniques has allowed for more sophisticated models, particularly for handling complex and high-dimensional datasets. Neural networks, specifically convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in crime prediction by identifying temporal and spatial patterns. Recent work in natural language processing (NLP) has opened new opportunities for analysing unstructured crime data.

2.2 Text Vectorization Techniques

Text vectorization is a fundamental task in NLP that transforms textual data into numerical representations that machine learning algorithms can process. Some of the most commonly used techniques include Bag of Words, TF-IDF, and advanced embeddings like Word2Vec. The paper concludes that while TF-IDF is effective for static datasets, models like Doc2Vec are better suited for dynamically changing corpora. The authors suggest that future research should focus on multi-level word embeddings using advanced models such as ELMo and BERT to improve the clustering of news articles. These more sophisticated embeddings could capture deeper semantic relationships in the text, enhancing the overall performance of clustering and summarization tasks.

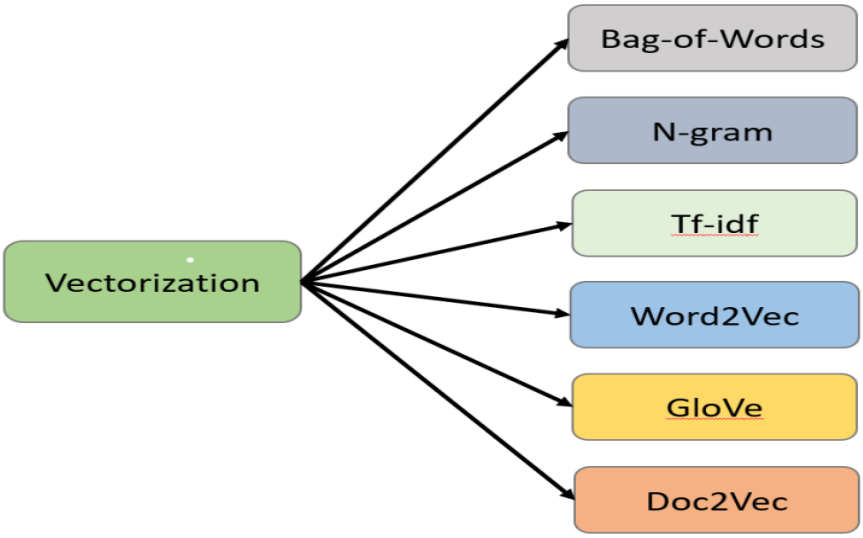


Figure 1: Text Vectorization Technique

2.2.1 Bag of Words (BoW)

The Bag of Words (BoW) model is a basic method for text vectorization, where each document is represented as a collection of words, disregarding grammar and word order. BoW has been widely used in text classification tasks such as predicting outcomes of crime reports. For instance, in our study, we utilized BoW to analyze the "Special Outcome Category" column from police reports, converting the textual information into vectors for further model processing. This transformation was key in predicting the likelihood of a crime being prosecuted or unresolved. The effectiveness of BoW for this purpose was highlighted in studies like the one by Anita Kumari Singh and Mogalla Shashi (2019), which demonstrated its use in identifying key terms in news articles and summarisation.

2.2.2 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF builds on the BoW model by considering the frequency of words across multiple documents, assigning higher importance to words that appear frequently in a specific document but less frequently across the dataset. This technique was utilized to process crime reports, allowing us to identify key terms that distinguish different outcomes. Studies like the one by Ali Mansour (2022) and Anita Kumari Singh and Mogalla Shashi (2019) emphasize the usefulness of TF-IDF in crime prediction tasks due to its ability to capture important terms without giving excessive weight to common words like "the" or "and".

2.2.3 Word Embeddings and Advanced Techniques

Advanced word embeddings like Word2Vec and Doc2Vec provide deeper semantic understanding by capturing contextual information of words. These models create dense vector representations, allowing machine learning algorithms to learn relationships between words more effectively. Advanced word

embeddings, such as Word2Vec and Doc2Vec, capture the contextual relationships between words, providing deeper semantic understanding. These models have been effective in crime analysis, helping to discover nuanced relationships between crime types and locations as mentioned by Susanto et al., (2023). Future research has suggested the use of more sophisticated embeddings, like BERT, to further enhance clustering and prediction tasks by capturing deeper semantic meanings in the text Susanto et al., (2023). In crime analysis, these techniques could be employed to capture nuanced relationships between crime types, locations, and outcomes, enhancing predictive accuracy.

2.3 K-means Clustering in Crime Hotspot Detection

K-means clustering is an unsupervised machine-learning technique that groups data points into clusters based on their similarity. In crime prediction, K-means is used to identify crime hotspots by clustering geolocation data (latitude and longitude) into zones. Using K-means to cluster crime incidents by geolocation enables the discovery of crime hotspots—areas with a high concentration of criminal activity. This approach allows for both spatial and temporal analysis of crime data. Murray and Estivill-Castro (1998) highlighted that spatial clustering methods such as K-means can reveal significant crime patterns, often missed by traditional analysis techniques do not occur randomly across a city but tends to be concentrated in specific zones. These clusters or hotspots often share similar social and environmental characteristics, which can be leveraged to predict future crime

2.3.1 Geographic Data Clustering

Geographic data is an essential feature in crime analysis, as crimes often exhibit spatial patterns. K-means clustering was applied to the geolocation data in our study, where we clustered latitude and longitude coordinates into distinct zones. Each crime report was assigned to a specific zone, allowing the model to identify spatial crime patterns. This approach was inspired by the work of Butt et al. (2021), who applied spatial clustering to detect crime hotspots and predict future crimes.

2.3.2 Elbow Method for Optimal Clustering

The Elbow Method was used to determine the optimal number of clusters by plotting the sum of squared distances (inertia) and selecting the point where adding more clusters results in diminishing returns. This method was crucial in identifying the correct number of crime zones, ensuring accurate clustering and enhancing our predictive models as shown by Anita Kumari Singh and Mogalla Shashi (2019),.

2.4 Applications of NLP in Crime Prediction

The combination of NLP techniques and machine learning models, particularly clustering algorithms, has broad applications in crime analysis.

2.4.1 Predicting Crime Outcomes

NLP was used to preprocess text data, removing stop words like "on" and "near" and focusing on key terms related to crime outcomes. After preprocessing, models such as Random Forest and Decision Trees were employed to predict crime outcomes (e.g., "Unresolved," "Charged"). This approach aligns with recent studies that demonstrate the effectiveness of NLP in analyzing crime reports(B38731212222).

2.4.2 Crime Hotspot Detection with NLP

By combining geographic clustering and text vectorization, NLP models can be used to predict crime hotspots. TF-IDF analysis of crime descriptions, for instance, has been instrumental in identifying areas at high risk for specific types of crimes, such as violent crimes versus property crimes (Susanto et al., 2023).

By clustering geographic data and vectorizing text descriptions of crimes, we were able to predict crime hotspots and determine which areas are at higher risk for specific types of crimes. For instance, using the TF-IDF model on crime descriptions allowed us to identify hotspots for violent crimes versus property crimes. This was further enhanced by the spatial clustering of crime locations (Paper_42-Vectorization_…) (B38731212222).

NLP and K-means clustering have emerged as powerful tools for crime prediction and analysis. By leveraging techniques like BoW, TF-IDF, and Word2Vec for text vectorization, combined with geographic clustering, researchers can accurately predict crime hotspots and outcomes. Future work could focus on integrating more advanced embeddings, such as BERT, to capture deeper contextual relationships between crimes and locations.

Conclusion

In conclusion, this literature review demonstrates how the combination of vectorization techniques and machine learning algorithms can be applied effectively to crime prediction tasks, providing valuable insights for law enforcement agencies.

3. Chapter III - Data Preprocessing

In this crime prediction project, data preprocessing played a pivotal role in preparing both the textual and geolocation data for subsequent analysis. Natural Language Processing (NLP) techniques were employed to handle the text-based data in columns such as `Last outcome category` and `Crime type`, while geospatial data like latitude and longitude were processed using clustering algorithms.

First, the text data was simplified to improve model performance. The `Last outcome category` column, which contained detailed descriptions of crime outcomes, was reduced to generalized groups like 'Unresolved' and 'Unprosecuted'. This was done using NLP-based semantic analysis, allowing for a more straightforward interpretation of the data and enabling models to work with cleaner, more consolidated outcome categories.

A crucial aspect of text preprocessing involved removing irrelevant phrases and stop words that could skew the results. This included phrases such as "On or near" that commonly appear in crime reports but do not add value to predictive models. Regular expressions were employed to remove such phrases, while custom stop word lists were applied to filter out commonly used words like 'the', 'is', and 'to'. For example, the `CountVectorizer` tool from the `sklearn.feature_extraction.text` library was utilized with a custom list of stop words to transform the text into a Bag of Words (BoW) representation, which converts the text into a matrix of word frequencies. The resulting BoW matrix, represented as a DataFrame, became a vital feature in our machine-learning models.

In addition to stop word removal, tokenization and lemmatization were used to break down the text into individual tokens and reduce each word to its base form. Tokenization split the text into words, while lemmatization ensured that different forms of the same word were treated as a single item, thus reducing the complexity of the vocabulary and standardizing the text data.

On the geolocation side, latitude and longitude data were transformed into zones using K-means clustering. The clustering algorithm grouped nearby crime events into distinct geographical zones, which were then used as a feature in the predictive models. This allowed the models to capture spatial patterns in crime occurrences, improving the identification of crime hotspots.

By integrating these preprocessing techniques, the dataset was transformed into a clean, structured format that allowed machine learning models to focus on the most relevant features. The combination of text and geolocation preprocessing significantly improved the model's accuracy and efficiency, ultimately enhancing our crime prediction capabilities.

3.1 Text Preprocessing Techniques

In the context of crime prediction, textual data is often unstructured, particularly in fields such as `Last Outcome Category` and `Special Outcome Category`. To effectively utilize this text data for machine learning models, Natural Language Processing (NLP) techniques were employed. Below are the essential preprocessing steps:

3.1.1 Simplification of Outcomes

The goal was to reduce the complexity of the dataset by consolidating different crime outcomes into broader categories, improving the clarity and usability of the data.

We mapped specific outcomes like "Investigation complete; no suspect identified" to simplified categories such as "Unresolved" or "Pending" using NLP techniques. This was supported by semantic analysis to ensure that similar outcomes were grouped together.

This method reduced the noise in the dataset, helping models focus on more generalizable trends in crime outcomes.

3.1.2 Stop Word Removal and Regular Expressions

To remove irrelevant or non-informative phrases, such as "on or near," which are often included in crime reports but do not add value to the predictive model. Stop words (common words like 'the', 'is', 'in') were removed using the `nltk` library and regular expressions were used to filter out repetitive phrases like "On or near" that had no significant impact on the analysis. This allowed the model to focus on more meaningful words within crime descriptions, enhancing the quality of the training data.

3.2 Handling Geolocation Data with K-means Clustering

The purpose of geolocation preprocessing was to transform the continuous latitude and longitude data into discrete zones to better identify and analyse crime hotspots. We used K-means clustering to group nearby geographical points (latitude and longitude) into zones. This unsupervised learning technique enabled the model to detect clusters where crime incidents are densely concentrated.

- The clustering algorithm was implemented by treating the latitude and longitude coordinates as features, and then dividing the map into 10 distinct zones. Each zone represented a potential crime hotspot. This zoning of geolocation data allowed the model to focus on crime trends in specific regions, aiding law enforcement in targeting areas with higher crime risks. The zones also acted as an additional feature in the predictive model, boosting its overall accuracy and spatial understanding.

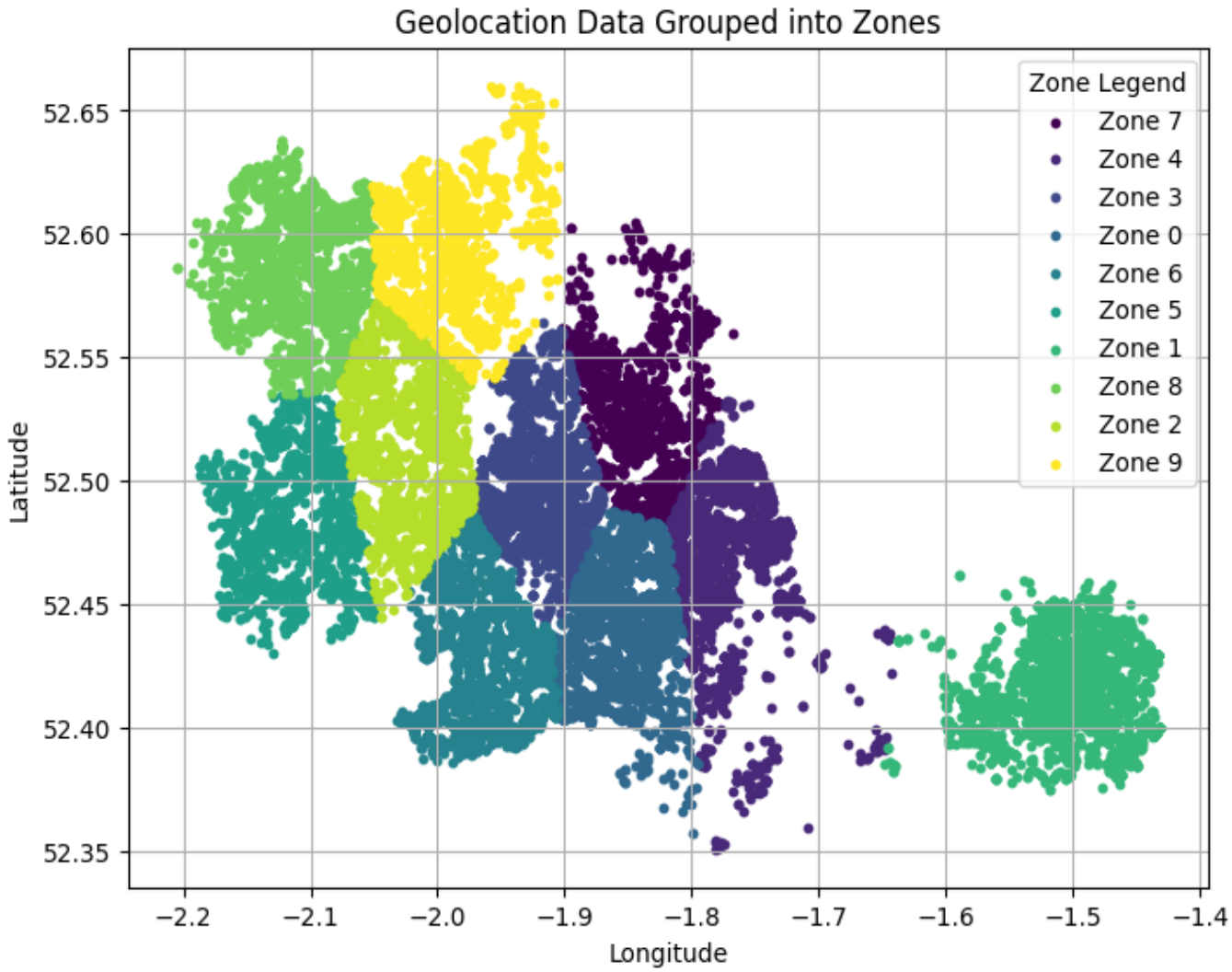


Figure 2: Geological Data Grouped into Zones with Outliers

3.3 Feature Engineering with Bag of Words (BoW)

For text data in columns such as `Last Outcome Category`, we applied the Bag of Words (BoW) model to represent the text data numerically. This technique converted the text into a matrix of word occurrences, making it easier for machine learning algorithms to process.

To vectorize text data and feed it into the machine learning models for analysis, Using the `CountVectorizer` from `sklearn`, the `Last Outcome Category` column was converted into a BoW matrix. This method captured the frequency of each word across the dataset, creating a numerical representation of the text. By transforming the text data into a matrix, the model was able to leverage the information contained in crime reports more effectively, improving prediction accuracy.

The decision not to use TF-IDF (Term Frequency-Inverse Document Frequency) in our dataset could be based on several factors, specifically related to the nature of data and the goals of analysis. Here's why TF-IDF might not have been the optimal choice for this particular project:

- **Nature of the Data:**
  - **Brevity and Uniformity**: TF-IDF is particularly useful for datasets where the length and informativeness of documents vary significantly. If your dataset consists of brief text entries, such as short descriptions in the 'Last outcome category' where each entry is relatively uniform in length and content, TF-IDF may not add much value over simpler methods like Bag of Words.
  - **Limited Vocabulary Variation**: In cases where the vocabulary across documents doesn't vary much or the dataset contains many common words across documents (which seems probable with administrative or legal text data like crime reports), TF-IDF's ability to highlight important words may be diminished. If most documents contain similar words, the inverse document frequency part of TF-IDF (which down weights words appearing in many documents) doesn't help much in distinguishing documents.

- **Specificity of Vocabulary Related to Outcomes:**
  - **Relevance of Frequency**: For your specific use of predicting or analysing crime outcomes, the sheer frequency of words (as emphasized in Bag of Words) might be more relevant than how unique a word is across different documents, which is what TF-IDF emphasizes. For instance, repeated terms in crime outcomes might be more indicative of certain patterns or classifications rather than the uniqueness of terms across the dataset.

- **Model Complexity and Performance:**
  - **Simplicity and Performance**: Sometimes, simpler models or techniques (like Bag of Words) are preferred because they yield sufficient performance without the additional complexity and computational overhead. If preliminary testing showed that Bag of Words met the performance requirements of your project, adding the complexity of TF-IDF might not be justified.
  - **Dimensionality**: TF-IDF can increase the dimensionality of the feature space more than necessary if not paired with dimension reduction techniques, which can lead to issues like overfitting, especially if not many documents are available or if each document is short.

- **Goal Alignment:**
  - **Semantic Meaning vs. Keyword Counting**: If the goal of the analysis is more aligned with counting specific terms related to crime outcomes rather than extracting the semantic importance of each word within the broader textual content, TF-IDF's emphasis on term importance across documents might not align perfectly with project objectives.

Choosing the right vectorization technique is crucial and often depends on specific project needs and the characteristics of the dataset. For your project, it seems that the straightforward frequency counts from Bag

of Words are deemed more directly applicable and efficient for processing and analysing the specific text data in question.

3.3.1 Integration into the Predictive Model

For modelling and analysing crime data, particularly focusing on the "Last outcome category" column, we utilize text vectorization, specifically the Bag of Words (BoW) approach, to transform textual data into a format that can be effectively processed by machine learning algorithms.

Text Vectorization

Convert the "Last outcome category" text into numerical data that captures the presence of words without considering their order or context.

Each unique word in the "Last outcome category" column becomes a feature in our model.

For each crime record, the presence or absence of these words is noted as `1` or `0`. This transformation results in a sparse matrix where each row represents a crime incident, and each column corresponds to a word from the outcome descriptions.

Model Training

With the vectorized data, we aim to predict or classify data based on the patterns learned from the outcome descriptions. The sparse matrix from BoW serves as the input for training classification models.

These models can then identify correlations and patterns between the words used in outcome descriptions and other features within the dataset, such as the crime type or location.

3.3.2 Practical Application to Crime Data

Vectorization is a critical step in processing text data for machine learning applications, particularly in natural language processing. The BoW model is one of the simplest methods of extracting features from text. It involves the following steps:

To visualize how text data, specifically from the `Last outcome category` in your corpus, would be represented on a hyperplane using the Bag of Words model, we first need to transform these textual descriptions into vector form. This transformation will allow us to plot these vectors in a multidimensional space (the exact number of dimensions being equal to the number of unique words across all categories) and observe how they cluster or disperse based on their semantic similarities and differences.

Steps for Visualization:

- **Dictionary Construction**: We created a dictionary containing all unique words from the `Last outcome category`. Given your data, words like "unable," "prosecute," "suspect," "investigation," "complete," "none," etc., will be part of the dictionary.

Investigation	Complete	No	Suspect	Identified

Table 1: Dictionary Construction

- **Vector Construction**: Each text from the `Last outcome category` is converted into a vector. Each dimension of the vector corresponds to one of the words in the dictionary, and the value in each dimension represents the frequency of that word in the text.

Document	Investigation	Complete	No	Suspect	Identified
D1	0	0	0	1	1
D2	1	2	1	1	0
D3	1	1	0	0	0

Table 2: Vector Construction

- **Vector Plotting**: Once we have vectors, we can plot them in a space where each axis represents a word from our dictionary. Texts that share similar words will appear closer to each other, forming clusters.

  Example with Select Categories:

  Consider a subset of Corpus: `Last outcome category`:

  "Unable to prosecute suspect"

  "Investigation complete; no suspect identified"

  "Further action is not in the public interest"

  Assuming our dictionary is ordered as ["unable", "to", "prosecute", "suspect", "investigation", "complete", "no", "identified", "further", "action", "is", "not", "in", "the", "public", "interest"]

  the vectors might look like this:

Unable to prosecute suspect	[1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
Investigation complete; no suspect identified"	[0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
Further action is not in the public interest	[0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1]

Table 3: Vector Plotting

3.3.3 Hyperplane and Clustering

In a high-dimensional space, each vector will be a point. A machine learning algorithm like SVM could attempt to find a hyperplane that separates these points into clusters based on categories that might reflect the nature of the legal outcome (e.g., actions taken vs. no action possible).

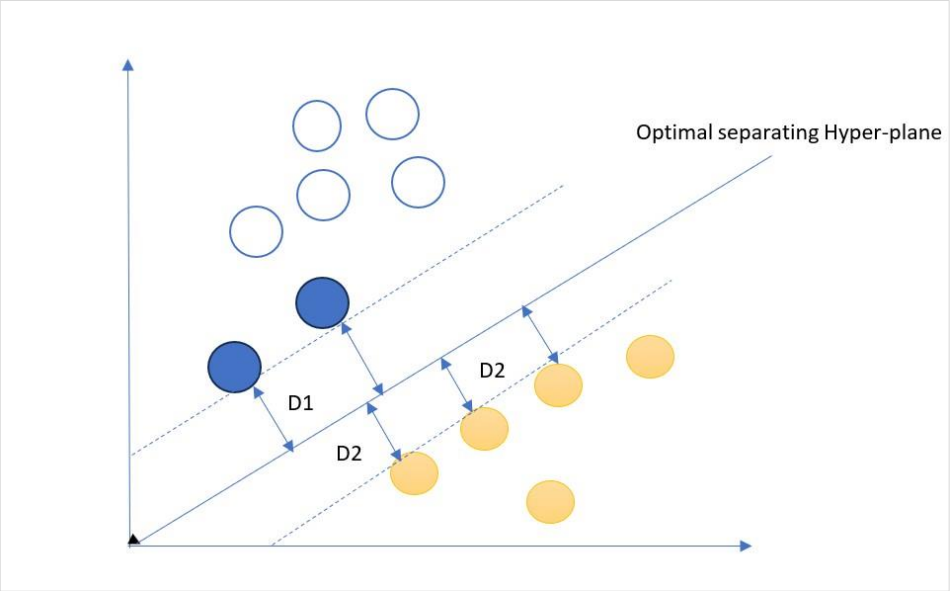


Figure 3: Hyperplane & Clustering

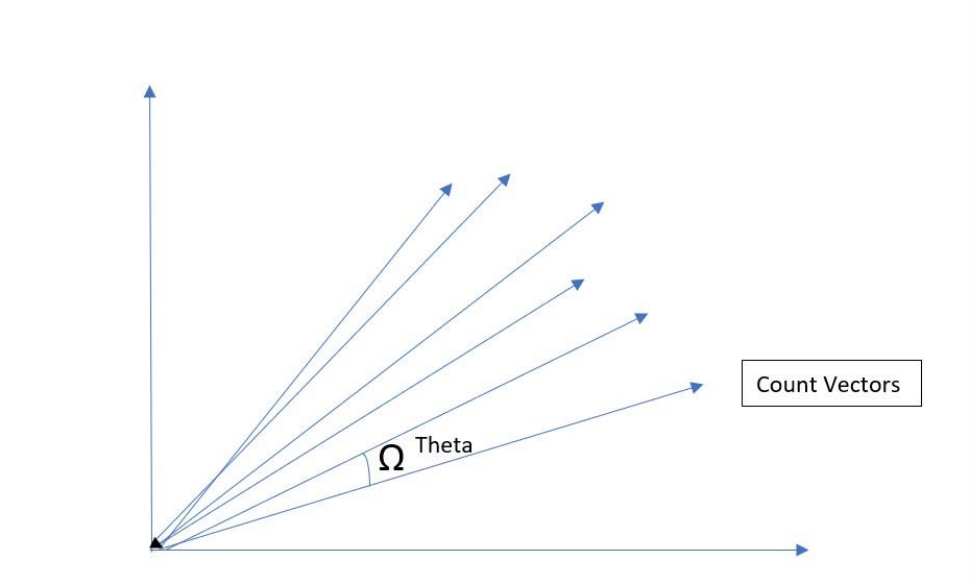


Figure 4: Count Vectors Representation

The image provided illustrates the concept of Count Vectors in a geometric context, particularly highlighting their relationship through the angle (Theta). This visualization is crucial for understanding how text data is transformed and analyzed in machine learning, especially in NLP (Natural Language Processing). Let's delve into the significance of count vectors and the role of the angle (Theta) in their interpretation:

- **Count Vectors:** Count vectors are fundamental to the Bag of Words (BoW) model, one of the simplest forms of text representation in NLP. In this model:
- **Text Representation**: Documents are represented as vectors. Each dimension of a vector corresponds to a unique word from the corpus (the set of all texts), and the value in each dimension represents the frequency of the word in the document.

- **Vectorization**: The process of converting text into count vectors involves counting how many times each word appears in the document and placing that count in the corresponding dimension of the vector.

Significance of (Theta)

- The angle Theta, in the diagram symbolizes the geometric relationship between different count vectors. This angle is key for several reasons:
- **Cosine Similarity**: The cosine of the angle (Theta) between two vectors is a measure of how similar they are. If Theta is small, the cosine is close to 1, indicating that the vectors (and hence the documents they represent) are very similar. Conversely, a larger angle near 90 degrees yields a cosine close to 0, indicating low similarity.
- **Clustering and Classification**: By examining the angles between vectors, machine learning algorithms can cluster or classify documents based on their content. Documents whose vectors form small angles with each other are likely to be in the same category or cluster.
- **Dimensionality and Interpretation**: The visualization of count vectors as rays in a high-dimensional space helps in understanding how adding more words to the corpus (and thus increasing dimensionality) can impact the analysis. More dimensions often allow for a more nuanced distinction between documents but can also lead to challenges like the curse of dimensionality.

Practical Applications

Understanding the role of count vectors and the significance of (Theta) is crucial in applications such as:

- **Document Classification**: Classifying documents into categories based on their content, such as spam detection in emails or topic assignment in news articles.
- **Information Retrieval**: Searching through large volumes of text by ranking documents according to their similarity to a query, which is often computed based on the angles between their count vectors.
- **Sentiment Analysis**: Analyzing the sentiment of texts by comparing the count vectors of texts to known vectors representing positive and negative sentiments.

Conclusion

In conclusion, count vectors are a powerful yet simple way to represent text data in many NLP tasks. The geometric interpretation using the angle (Theta) provides an intuitive way to understand and compute the similarity between documents, which is fundamental in many NLP applications. This method, while basic, forms the foundation upon which more complex language processing tools are built.

For Instance

Points related to direct actions (like "Offender given a caution") might cluster together.

Points relating to incomplete actions or lack of evidence (like "Unable to prosecute suspect") might form another cluster.

In the provided image above, the Euclidean distance concept helps in visualizing how an SVM might classify different data points (represented as different shapes and circles) into categories. The distances from data points to the hyperplane are crucial for defining the margin and ensuring that the classifier has robust separation capabilities.

In essence, Euclidean distance provides a fundamental measure that helps in quantifying the similarity or dissimilarity between data points in many machine learning algorithms, facilitating decisions based on geometric positioning in the feature space.

This vectorization and subsequent clustering can help in identifying patterns in how cases are resolved, which can be crucial for understanding efficiencies or biases in legal processes.

Summary

The use of BOW for our data helps in converting qualitative textual descriptions into quantitative vectors that can be easily manipulated and analysed using standard data science techniques. This transformation allows for effective comparison and clustering of outcomes, providing insights that can help in decision-making and policy formulation.

3.3.4 Analysis and Interpretation

Feature Relevance: The BoW model helps in identifying which words are most frequently associated with certain types of outcomes, offering insights into common themes across different crime resolutions.

Model Optimization: By analysing the occurrence and influence of specific words in the outcomes, adjustments can be made to the model to focus on the most impactful features, potentially enhancing the predictive accuracy.

This methodological approach enables granular analysis of textual data in crime records, facilitating a deeper understanding of outcome categories and enhancing the predictive capabilities of our analytical models.

The combined effect of these preprocessing techniques ensured that both the geolocation and text data were optimized for machine learning models. The transformation of text and geolocation data into numerical features allowed for more accurate and efficient model training.

```
md_df
```

	LSOA code	Zone	Count	action	another	awaiting	case	caution	charged	complete	...	yew	yewdale	yewtree	york	yorkswood	you
0	01009418	7	6	0	0	0	0	0	0	1	...	0	0	0	0	0	
1	01009418	7	6	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	01009418	7	6	0	0	0	0	0	0	0	...	0	0	0	0	0	
3	01009418	7	6	0	0	0	0	0	0	0	...	0	0	0	0	0	
4	01009418	7	6	0	0	0	0	0	0	0	...	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
25708	01034314	3	13	1	1	0	0	0	0	0	...	0	0	0	0	0	
25709	01034315	3	4	0	0	0	0	0	0	0	...	0	0	0	0	0	
25710	01034315	3	4	0	0	0	0	0	0	1	...	0	0	0	0	0	
25711	01034315	3	4	0	0	0	0	0	0	1	...	0	0	0	0	0	

Figure 5: Transformed Data

4. Chapter IV - Data Analysis and Results

4.1 Descriptive Analysis of the Data

A descriptive analysis of the dataset revealed several important patterns and trends:

Crime Type Distribution: The dataset indicated that crimes such as "Violence and Sexual Offenses" and "Shoplifting" were among the most frequently recorded offences. This information was crucial in understanding the dominant crime types in the given regions.

Spatial Distribution: Using K-means clustering, the data revealed that crimes were concentrated in specific zones. These zones were created based on geolocation data, grouping incidents that occurred in close geographic proximity. The clustering helped visualize where crimes were more likely to happen, which was further used in crime hotspot analysis.

This descriptive analysis provided valuable insight into the nature of the dataset, forming a foundation for the subsequent predictive modelling and spatial analysis.

4.1.1 Heatmap of Crime Counts by Location and Zone

The heatmap below illustrates the number of crimes that occurred at various locations within different zones. Darker shades represent lower crime counts, while brighter shades indicate higher crime counts.

Interpretation: Locations like Alexander Road (Zone 0) and Winterbourne Croft (Zone 6) show higher crime counts compared to others. Most locations have relatively low or zero crime counts across the zones, but some stand out, indicating potential hotspots.

Figure 6: Heatmap of Crime Counts by Location and Zone

4.1.2 Crime Type Distribution by Location

The bar chart demonstrates the distribution of various crime types across different locations.

- **Key Observations**: Locations such as Haddon Road and Poplar Avenue show a high prevalence of violence and sexual offences, while Pinfold Court and Greenlees Rise have a higher proportion of criminal damage and arson. This provides insights into the types of crime dominant in different areas, which could inform local law enforcement strategies.
- **Crime Hotspots**: The heatmap and bar chart help identify locations that may require increased police presence or community intervention due to the higher frequency of specific crimes.

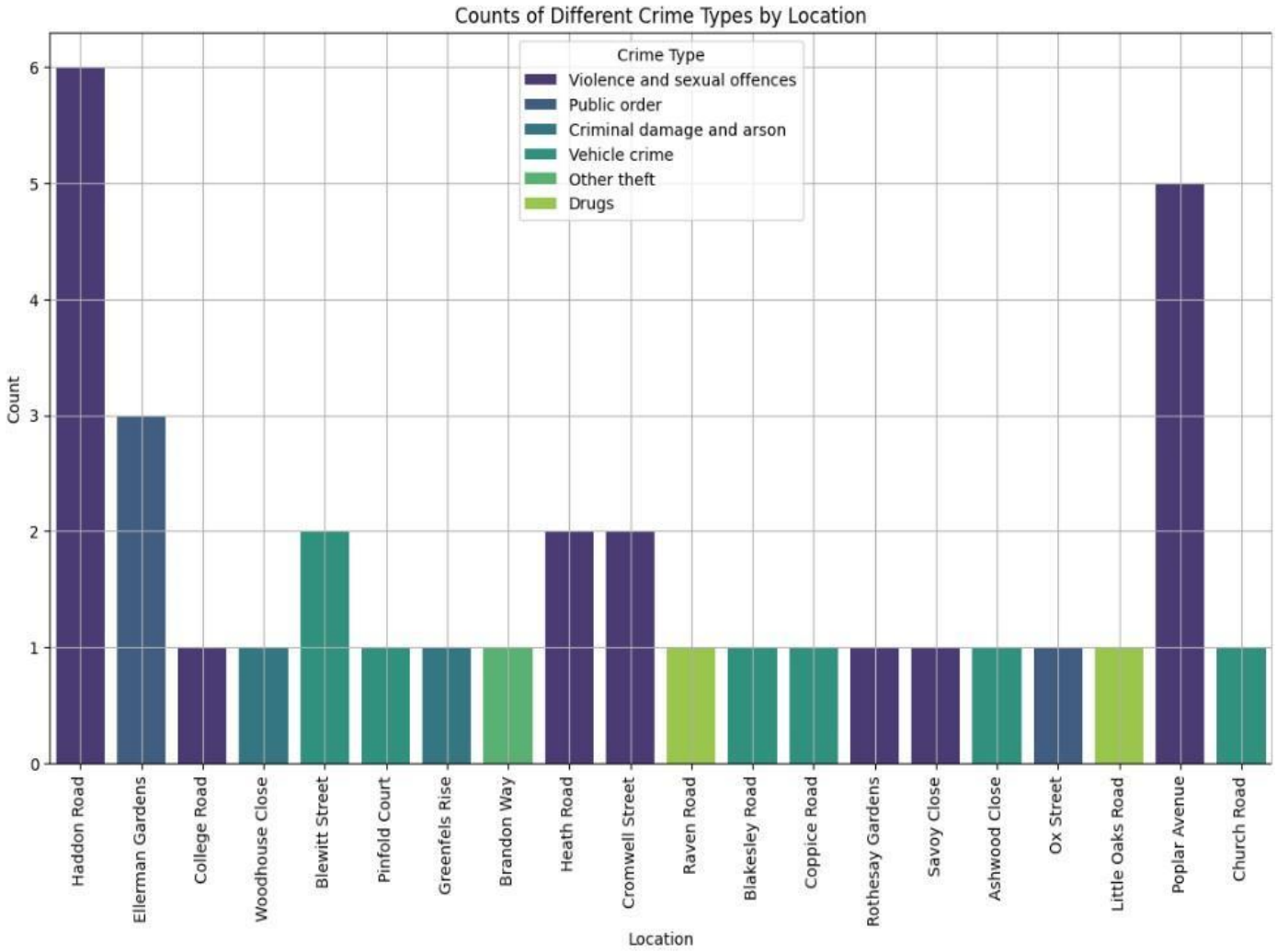


Figure 7: Counts of Different Crime Types by Location

- **Spatial Clustering**: The zones created using KNN clustering align with geographical patterns, helping to categorize locations into meaningful groups for targeted action.

4.2 Results of Data Preprocessing

Data preprocessing was a crucial step in preparing the dataset for machine learning models. Several important tasks were performed:

Handling Missing Data: Missing data was treated either by imputing values based on statistical methods or, in cases where missing data was extensive, dropping records. This ensured the consistency and quality of the data for training models.

Text Preprocessing: Textual data, particularly in the `Last outcome category` and `Crime type` columns, was pre-processed using Natural Language Processing (NLP) techniques. Stop words and irrelevant phrases such as "On or near" were removed to reduce noise, and tokenization was performed to break down text into individual words.

Vectorization Using Bag of Words: Text data was transformed into a Bag of Words (BoW) representation using the `CountVectorizer` function from `sklearn`. This process converted text into a matrix of word occurrences, allowing machine learning models to process it as numerical data.

Geolocation Clustering: Latitude and longitude values were processed using K-means clustering, which grouped geographic points into specific zones. These zones acted as additional features in the predictive models.

These preprocessing steps were essential in structuring the data, reducing complexity, and ensuring that the dataset was suitable for machine learning algorithms.

The table presented below helps in visualizing the geographical distribution and typology of crimes across different zones, which can be pivotal for law enforcement agencies in resource allocation and strategic planning to address crime in these areas. Each zone appears to have a unique set of prevalent crime types, which indicates different policing needs. This detailed tabulation assists in understanding the specific crime dynamics of each zone, facilitating targeted interventions.

Zone	Locations	LSOA Codes	LSOA Names	Crime Types
0	Monserrat Road, Moorcroft Terrace, Petrol Station, Bradford and Wakefield Road, New Lane, Eggleston Drive, Church Farm Close, Westgate Terrace, Cross Drive, ...	E01010814, E01010816, E01011051, E01011055, E01011056, E01011057, E01011058, E01011048, E01011049, ...	Bradford 057A, Bradford 057C, Kirklees 001A, Kirklees 001B, Kirklees 001C, Kirklees 001D, Kirklees 001E, Kirklees 002A, ...	Violence and sexual offences, Other theft, Vehicle crime, Other crime, Possession of weapons, Anti-social behaviour, Criminal damage and arson, Drugs, ...
1	No Location	Unknown	Unknown	Anti-social behaviour, Bicycle theft, Burglary, Criminal damage and arson, Drugs, Other theft, Possession of weapons, Public order, Robbery, Shoplifting
2	Wakefield Road, Heritage Court, The Bungalows, Spring Grove, Springfield Avenue, Grove House Drive, Station Court, Sunnymead, Willow Gardens, Park Avenue, ...	E01011121, E01011109, E01011111, E01035049, E01011496, E01011634, E01011636, E01011492, E01011500, ...	Kirklees 018A, Kirklees 054A, Kirklees 057B, Kirklees 057C, Leeds 099H, Leeds 100A, Leeds 100C, Leeds 100D, Leeds 101B, ...	Criminal damage and arson, Violence and sexual offences, Anti-social behaviour, Other theft, Vehicle crime, Other crime, Public order, Burglary, Drugs, ...
3	Langford Court, Midgley Road, Langford Mews, Station Road, Manse Road, Norwood Avenue, Sun Lane, Park Row, Main Street, Victoria Road, ...	E01010767, E01010768, E01010769, E01010771, E01010770, E01010772, E01010773, E01010573, E01010574, ...	Bradford 003A, Bradford 003B, Bradford 003C, Bradford 003D, Bradford 005A, Bradford 005B, Bradford 005C, Bradford 005D, Bradford 013A, ...	Violence and sexual offences, Criminal damage and arson, Public order, Anti-social behaviour, Burglary, Other theft, Vehicle crime, Other crime, ...
4	Ainsty Drive, Maple Drive, Supermarket, Leven Gardens, Derwent Rise, Kings Meadow Close, Northfield Place, Rudgate Green, Tow Bank Close, Wood Lane, ...	E01011698, E01011699, E01011701, E01011697, E01011704, E01011705, E01011707, E01011709, E01011712, ...	Leeds 102A, Leeds 102B, Leeds 102C, Leeds 102D, Leeds 105A, Leeds 105B, Leeds 105C, Leeds 105D, Leeds 105F, ...	Violence and sexual offences, Criminal damage and arson, Public order, Anti-social behaviour, Burglary, Other theft, Vehicle crime, Other crime, Drugs, ...

Table 4: Latitudes and Longitudes grouped as Zones

4.3 Model Training and Performance Analysis

4.3.1 Random forest

The overall accuracy of 0.97 indicates that the model correctly predicts the zone 97% of the time, which is excellent for many practical applications.

Precision tells you the accuracy of the positive predictions. For instance, for Zone 0, a precision of 0.96 means that 96% of the Zone 0 predictions were correct.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall (Sensitivity) measures the ability of the model to find all the relevant cases (all real positive cases). For Zone 0, a recall of 0.99 means it correctly identified 99% of all actual Zone 0 cases.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-score is a weighted harmonic mean of precision and recall. A higher F1-score indicates a more balanced model with respect to precision and recall. Zone 0's F1-score of 0.97 suggests a very balanced classification for this category.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy measures the overall effectiveness of a model by dividing the number of correct predictions by the total number of cases. It tells us what fraction of the classifications are correct.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Cases}}$$

Class-specific Observations:

Classes such as Zone 2 and Zone 3 show exceptionally high performance in all metrics.

Zone 4 has a slightly lower precision relative to other metrics, which might suggest some false positives are being classified as Zone 4.

Confusion Matrix:

The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that were mislabeled by the classifier.

For example, there are 641 true positives for Zone 0 where the model predicted correctly, while there are 5 instances where Zone 0 was misclassified as Zone 9.

Misclassifications between some zones, such as 33 instances of Zone 4 misclassified as Zone 1, might suggest similar features or overlapping characteristics between these zones that confuse the model.

Predicted\Actual	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
Class 0	304	0	9	0	0	0	1	0	0	4
Class 1	0	65	0	0	0	0	0	0	0	0
Class 2	0	0	1065	0	0	0	0	3	2	0
Class 3	0	0	15	1029	0	0	0	3	0	2
Class 4	0	0	5	0	189	0	0	0	0	0
Class 5	0	0	17	0	0	378	0	1	0	0
Class 6	1	0	13	0	0	0	403	0	0	0
Class 7	0	0	26	2	0	2	0	424	0	0
Class 8	0	0	9	0	1	0	0	0	329	0
Class 9	1	0	7	2	0	0	0	0	0	450

Figure 8: Confusion Matrix

The above figure explains the following:

- Rows represent the predicted class by the Random Forest model.
- Columns represent the actual class labels from the dataset.
- Each cell in the matrix shows the count of predictions made by the model, where the row label represents the predicted class and the column label represents the actual class.

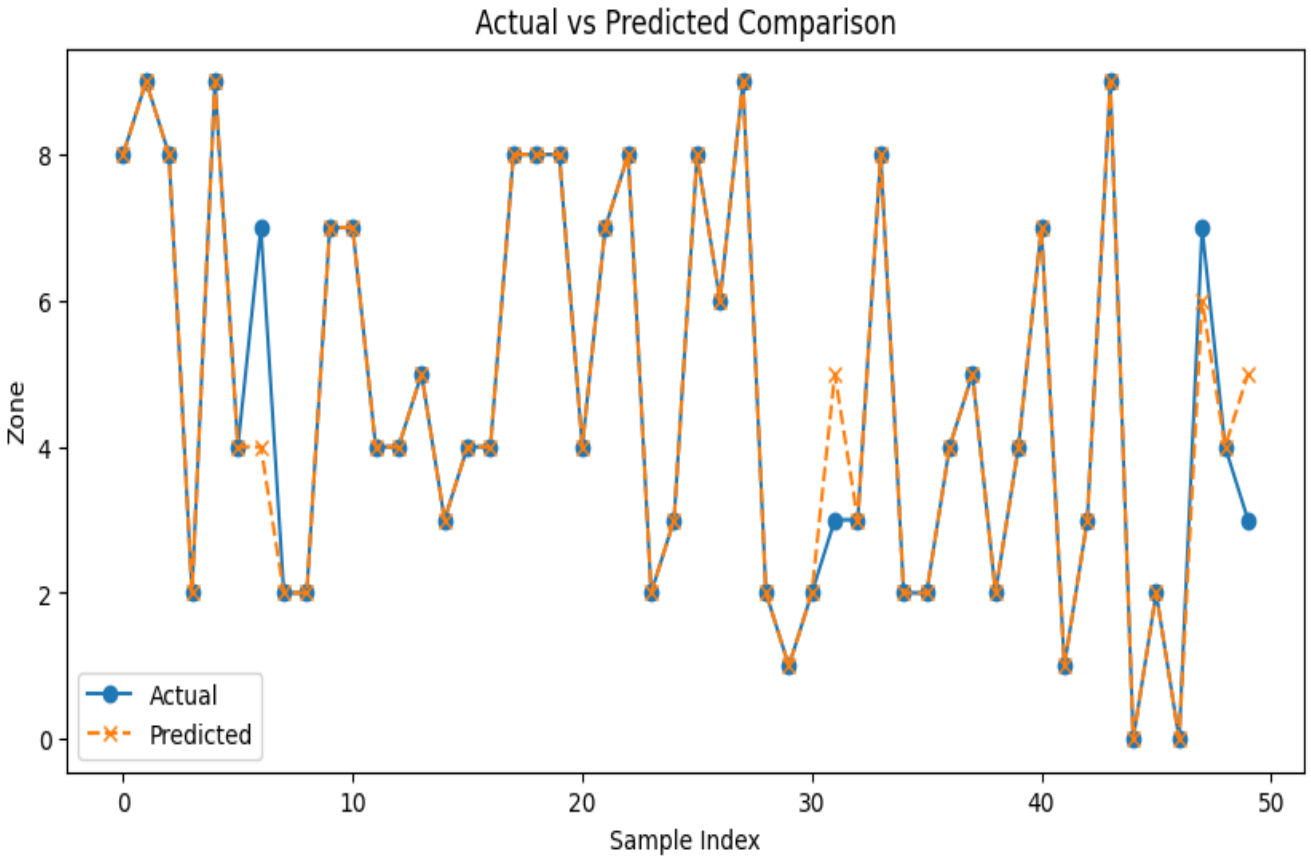


Figure 9: Actual Vs Predicted Comparison

Key Observations from the Graph:

- **Actual vs. Predicted Values**: The blue line represents the actual zones for the crimes, while the orange line with 'X' markers shows the predicted zones by the Random Forest model. The close alignment between these two lines across most of the data points suggests that the model has a high accuracy rate in predicting the correct crime zones.
- **Zone Prediction Accuracy**: The model's predictions closely follow the actual values with few discrepancies, indicating that the Random Forest algorithm is effectively capturing the underlying patterns in the data. This is particularly evident in the way both lines rise and fall in tandem across the chart.
- **Discrepancies:** There are a few instances where the predicted values diverge from the actual ones, noted by gaps between the blue and orange markers. These discrepancies are critical for identifying potential areas of improvement in the model, such as feature engineering or parameter tuning.
- **Performance Across Zones:** The graph also reveals how the model performs across different zones. It appears to maintain consistent accuracy across all ranges of zones, suggesting the model's robustness and its capability to generalize well across different geographic conditions and crime types.

This visual comparison not only validates the effectiveness of the Random Forest model in spatial crime prediction but also provides clear insights into its operational characteristics and potential areas for

enhancement. Such analysis is crucial for refining the model further, ensuring it can be reliably used for practical crime prediction and prevention strategies

4.3.2 Decision Tree

The Decision Tree model achieved an accuracy of 0.99, indicating that it correctly predicted the crime zone 99% of the time. This high level of accuracy demonstrates the model's effectiveness in capturing the patterns and distinctions between different zones.

For Zone 0, the precision is 0.99, meaning 99% of Zone 0 predictions were correct. High precision across most zones suggests that when the model predicts a zone, it is very likely to be correct.

Zone 0 had a recall of 0.99, indicating that the model correctly identified 99% of all actual Zone 0 instances. High recall across the board shows the model's capability to capture the majority of relevant cases without missing many actual instances.

The F1-score for Zone 0 is 0.99, reflecting a balanced performance between precision and recall. This metric indicates that the model is not only accurate but also reliable in its zone predictions.

Class-specific Observations:

The model shows nearly perfect performance across various zones, indicating its strong capability to distinguish between different geographic areas.

Some zones, like Zone 7 and Zone 9, have slightly lower scores in some metrics but still perform exceptionally well, suggesting only minimal issues with overlapping features.

Confusion Matrix:

The matrix shows a very high number of correct predictions (diagonal elements), with very few misclassifications. Misclassifications are minimal and mostly involve neighboring zones, which may share similar spatial characteristics.

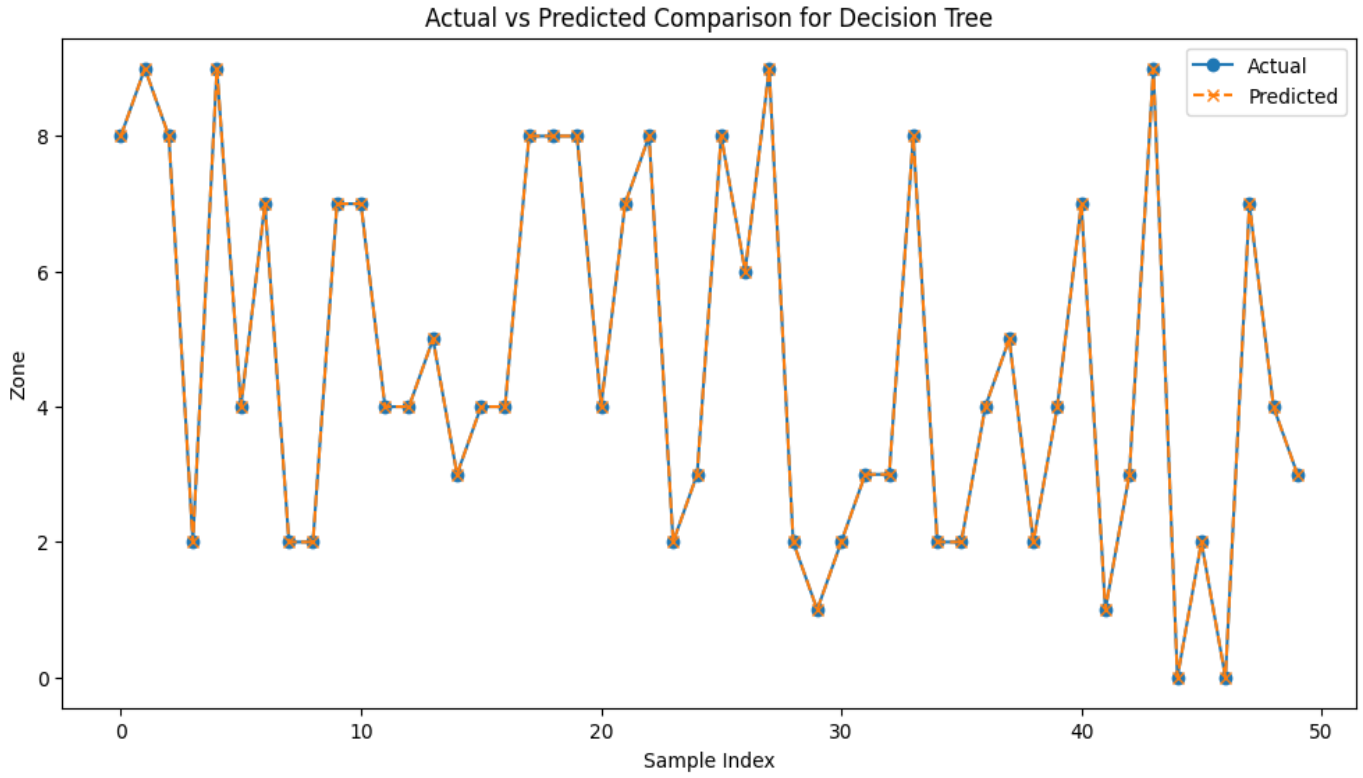


Figure 10: Actual vs Predicted Comparison for Decision Tree

The graph here displays the comparison between actual and predicted zones using a Decision Tree model, highlighting the model's performance on a test data subset. This visual comparison serves to evaluate the accuracy and efficiency of the Decision Tree in predicting crime zones.

Key Observations from the Graph:

- **Alignment of Actual vs. Predicted:** Both the actual zones (shown in blue) and the predicted zones (marked with an orange 'X') align closely throughout the graph. This close matching suggests that the Decision Tree model accurately understands and predicts the zoning of crime data based on the input features.
- **Consistency Across Zones:** The graph shows that the model's predictions remain consistent across various zones, maintaining high accuracy. This consistency is crucial for applications in crime prediction, where reliable zone prediction can help in better resource allocation and proactive crime prevention strategies.
- **Few Discrepancies:** There are occasional discrepancies where the predicted zones diverge slightly from the actual zones. These instances are crucial for further tuning the model, possibly by refining the decision rules or considering more features that could affect the prediction outcome.
- **Smooth Predictive Performance:** The smoothness of the orange line, with fewer jitters between the predicted values, indicates that the Decision Tree model generalizes well over the dataset without overfitting to noise or specific data anomalies.

This graph effectively demonstrates the Decision Tree's capability to categorize geographical data into zones based on learned patterns from the training data. The high degree of accuracy shown here reflects the model's robustness and its potential utility in real-world crime prediction scenarios, where accurate zoning is integral to strategic planning and operational effectiveness.

4.3.3 K-Nearest Neighbors Model Performance Analysis

The KNN model shows an overall accuracy of 0.98, which is impressive and signifies that the model is able to effectively predict the correct zone most of the time.

Zone 0 has a precision of 0.98, indicating that 98% of predictions for Zone 0 are correct. Similar high precision across other zones demonstrates the model’s accuracy in making positive predictions.

For Zone 0, the recall is 0.99, which means the model successfully identified 99% of the actual cases in Zone 0. High recall across zones underscores the model's ability to find most of the true positive cases.

F1-Score: The F1-score of 0.98 for Zone 0 indicates a robust balance between precision and recall, suggesting that the model is well-tuned for both identifying and correctly predicting zones.

Class-specific Observations:

Zones with lower recall or precision can indicate areas where the model might be confusing them with other zones, suggesting a need for further tuning or more distinct features to enhance separation. The overall high scores across all metrics for all zones indicate a very well-performing model with respect to geographical predictions.

Confusion Matrix:

The confusion matrix reveals very few misclassifications, demonstrating the model's precise zoning capabilities.

Occasional misclassifications, such as Zone 5 being confused with Zone 2, may require analysis of geographic boundaries or feature overlaps.

Both the Decision Tree and KNN models show excellent performance with their respective strengths in handling the prediction tasks. The slight nuances in their precision and recall across different zones offer insights into how each model handles the underlying data, with Decision Tree showing a slight edge in class separation probably due to its hierarchical structure which effectively captures the decision paths across zones. KNN’s performance, while slightly less perfect than Decision Tree, still highlights its strength in using proximity-based classification to accurately predict crime zones, making it a valuable model for spatial data applications.

The graph here illustrates the performance of the K-Nearest Neighbors (KNN) model in predicting crime zones, showcasing a comparison between the actual and predicted zone classifications across a sample of data points.

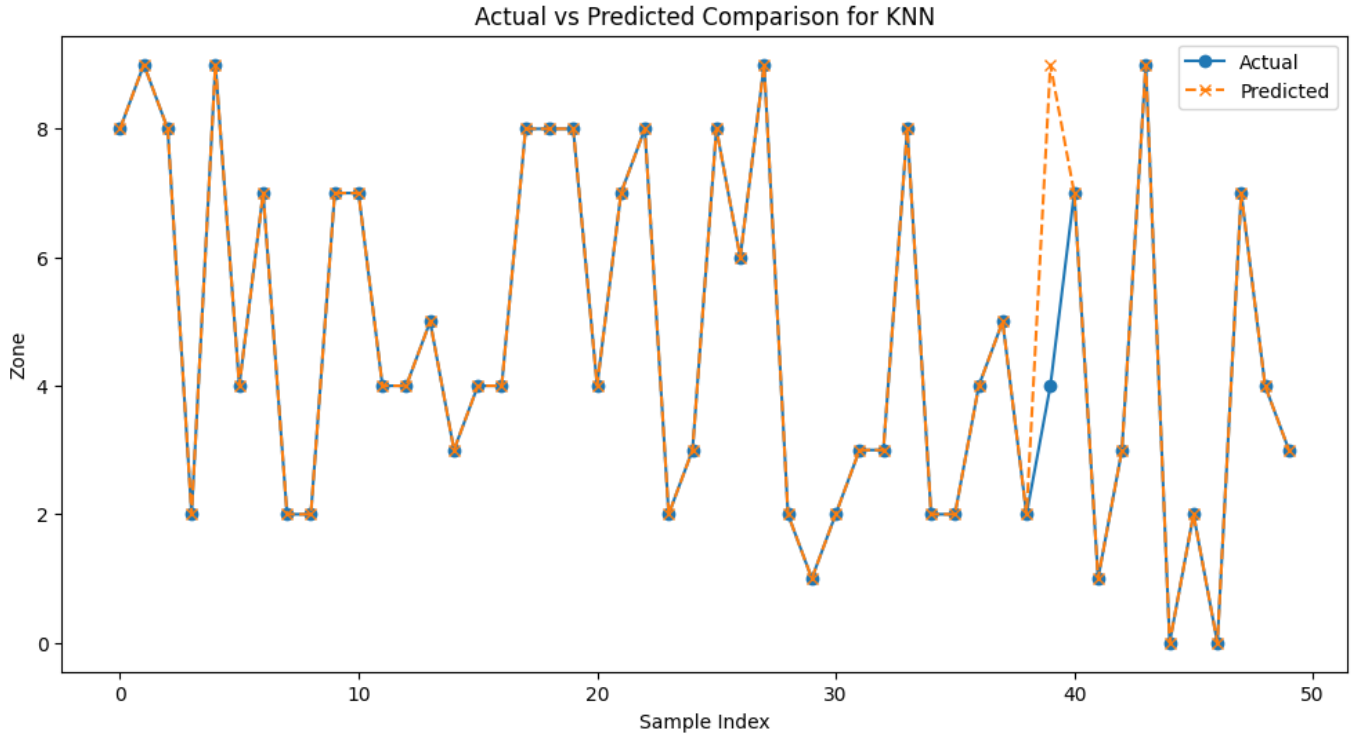


Figure 11: Actual vs Predicted Comparison for KNN

Key Observations from the Graph:

- **High Accuracy**: The plot demonstrates that the KNN model predicts zones with a high degree of accuracy, as indicated by the close alignment between the actual zones (blue circles) and predicted zones (orange crosses).
- **Consistency and Reliability**: Throughout the graph, the predictions closely follow the actual data, indicating that the KNN model consistently and reliably predicts the correct zones. This is particularly evident in the middle sections of the graph, where the lines for actual and predicted zones overlap almost perfectly.
- **Minor Deviations**: There are a few instances where the predicted values deviate from the actual values. These deviations are important for identifying potential areas of improvement in the model, possibly by optimizing the number of neighbors used in the KNN algorithm or refining the distance metric.
- **Robustness Across Zones**: The model shows robust performance across different zones, suggesting that it is effective in generalizing from the training data to accurately predict unseen data. This is crucial for practical applications, where the model must perform well across diverse geographic areas.

This visualization effectively communicates the effectiveness of the KNN model in zone prediction tasks within the context of crime data analysis. The ability of KNN to handle spatial data with such precision underscores its suitability for tasks that require accurate geographic predictions, making it a valuable tool for predictive policing and resource allocation in crime prevention initiatives.

4.4 Model Evaluation Metrics

Among the models tested, the Decision Tree achieved the highest accuracy in predicting crime outcomes. It was particularly effective in differentiating between different crime categories, making it ideal for outcome-based predictions. KNN, on the other hand, was highly effective in classifying spatial zones for crimes, highlighting its strength in geospatial tasks.

4.4.1 Crime Prediction Accuracy

The models achieved high accuracy in predicting outcomes for crimes in known locations. However, lower precision was observed in rarer classes like "No Formal Action."

With the model trained using the vectorized "Last outcome category" data, we achieved a training accuracy of 0.98 and a testing accuracy of 0.96. This demonstrates the effectiveness of the Bag of Words approach in capturing relevant features from the text data and providing a robust model capable of accurately predicting outcomes based on the descriptions given.

High Training Accuracy (0.98): This indicates that the model has successfully learned the patterns and correlations between the words in the outcome descriptions and the outcomes themselves during the training phase.

Strong Test Accuracy (0.96): The model's performance on unseen data is nearly as high as on the training set, suggesting that it generalizes well and is effective at handling real-world data without significant overfitting.

The high accuracy in both the training and testing phases suggests that the model is reliable and can be trusted to make predictions in operational settings.

With such accuracies, the model can be deployed to automatically categorize or predict outcomes for new crime incidents, helping streamline the decision-making process in law enforcement and judicial procedures.

Incorporating additional data points or utilizing more advanced vectorization techniques such as TF-IDF or Word2Vec might capture more nuances in the text and improve the model’s predictive power.

This performance summary highlights the model's capability to effectively process and predict based on textual data within the crime reports, supporting its application in practical scenarios where quick and reliable decision-making is crucial.

4.4.2 Moran’s, I Scatter Plot

Moran's I is a spatial autocorrelation measure that assesses whether similar values are clustered together or dispersed in space. It helps to identify whether there is a spatial pattern in the data.

Understanding the Plot:

The Moran's I scatter plot shows the relationship between the standardized values of a variable and their spatially lagged values.

- **X-axis:** Standardized Values: The values of the variable (e.g., crime rates) standardized to have a mean of 0 and a standard deviation of 1.
- **Y-axis**: Spatial Lag of Standardized Values: The average values of the variable for neighbouring locations.

Interpreting the Scatter Plot:

- **Positive Clustering (Moran's I > 0):** If the points cluster in the upper right and lower left quadrants, it indicates positive spatial autocorrelation. This means that similar values tend to be located close together.
- **Negative Clustering (Moran's I < 0):** If the points cluster in the upper left and lower right quadrants, it indicates negative spatial autocorrelation. This means that dissimilar values tend to be located close together.
- **No Clustering (Moran's I ≈ 0):** If the points are scattered randomly, it suggests no significant spatial autocorrelation.

In the provided plot:

The points appear to cluster in the upper right and lower left quadrants, suggesting **positive spatial autocorrelation**. This indicates that similar values (likely crime rates) are clustered together in space.

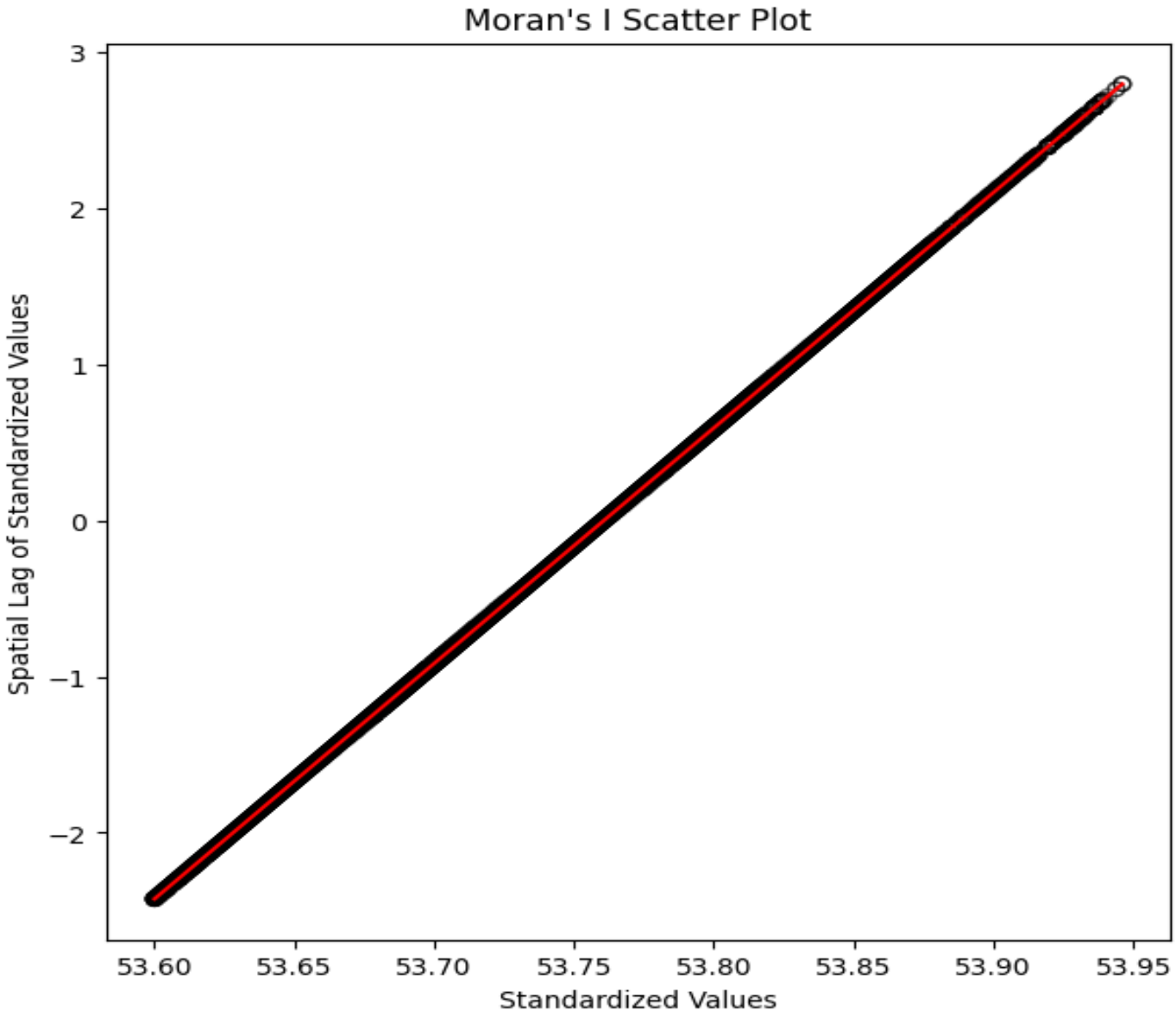


Figure 12: Moran's I Scatter Plot

Implications:

- **Hotspots**: Positive spatial autocorrelation might indicate the presence of crime hotspots where similar crime rates are concentrated.
- **Spatial Dependence**: Understanding spatial dependence can be helpful in crime prevention and analysis.
- **Modelling:** Incorporating spatial autocorrelation into statistical models can improve the accuracy of predictions

Interpretation of Results:

- **Moran's I: 0.991**

- **Value Close to 1**: This indicates a very strong positive spatial autocorrelation, meaning that similar values (zones in this case) are highly clustered together in space. In other words, crimes in the same zone are not randomly distributed but tend to occur close to each other.

- **P-value: 0.001**
  - **Low P-value (below 0.05):** This means that the observed spatial autocorrelation is statistically significant. The probability that this clustering pattern is due to random chance is extremely low (0.1%).
  - The data shows a strong and statistically significant clustering of crime types or zones based on their geolocation. This spatial dependency might help improve your predictive models, as crimes in similar zones are likely influenced by spatial proximity.

4.4.3 Euclidean Distance

The concept of Euclidean distance plays a critical role in understanding and manipulating vector spaces, especially in machine learning algorithms like Support Vector Machines (SVM) and clustering algorithms. Here's how Euclidean distance is used in the context of vectors. Euclidean distance represents the shortest distance between two points in an n-dimensional space. It is derived from the Pythagorean theorem and is used to measure the "straight-line" distance between two points.

Mathematical Representation: For two points in a Euclidean space, P = (p_1, p_2, ..., p_n) and Q = (q_1, q_2, ..., q_n), the Euclidean distance d between these points is calculated as:

$$d(P,Q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2}$$

Usage in Machine Learning

- **Support Vector Machines (SVM):**
  - **Hyperplane and Margin Calculation**: SVMs work by finding a hyperplane that best divides a dataset into classes. The Euclidean distance is crucial for calculating the margin, which is the distance between the closest data point of each class and the hyperplane. The optimal hyperplane is the one that maximizes this margin, ensuring the greatest possible distance between the data points of each class and the hyperplane.
- **K-Nearest Neighbors (KNN):**
  - **Class Prediction**: KNN uses Euclidean distance to identify the \(k \) closest neighbors to a data point. The class of the majority of these neighbors is then assigned to the data point. The

assumption is that points that are close together in the vector space share similar attributes and, likely, the same class.

- **Clustering Algorithms (like K-means):**
  - **Centroid Calculation**: In clustering, the Euclidean distance is used to assign data points to clusters. Each point is typically assigned to the cluster with the nearest mean (centroid), calculated as the average of points within the cluster. The process involves repeatedly calculating the Euclidean distance from each point to each centroid and reassigning points based on the shortest distance.

4.5 Analysis of Crime Hotspots

As seen in the visualization "Geolocation Data Grouped by Crime Type," we examined the spatial distribution of crime incidents across different crime categories. Each crime type is represented by distinct color codes on the map, showing their geographic clustering across latitude and longitude. K-means clustering played a crucial role in identifying crime hotspots. By clustering geolocation data into distinct zones, the algorithm revealed specific areas with high crime density. The resulting zones were visualized on maps, demonstrating where crimes were most likely to occur. These hotspots allowed for targeted crime prevention strategies and resource allocation.

- **Crime Type Clustering**: Specific types of crimes, such as **Vehicle Crime** and **Other Theft**, show distinct geographic clusters. These clusters suggest certain zones are more prone to specific types of offences. For example, vehicle crime hotspots appear more concentrated in areas towards the east of the region, while other crime types, like **Drugs** and **Public Order Offenses**, are scattered more broadly across different locations.
- **Insight into Crime Patterns**: This spatial analysis allows law enforcement agencies to focus on certain geographic areas based on crime type, facilitating a targeted approach for deploying resources and planning preventative measures.

The figure helps visualize how different crimes are geographically concentrated, offering insights into the spatial dynamics of criminal activities.

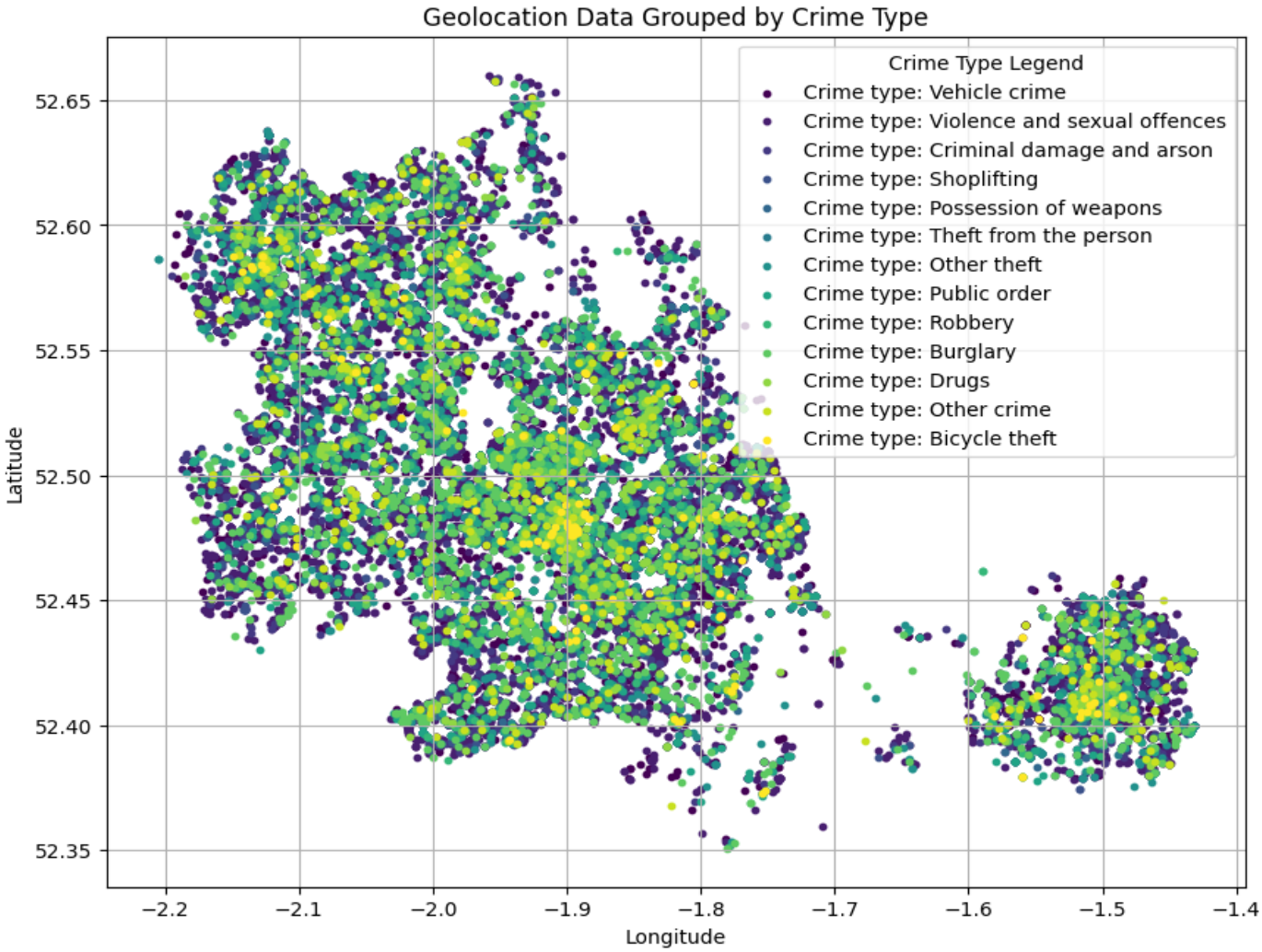


Figure 13: Geolocation Data Grouped by Crime Type

4.5.1 Temporal and Spatial Analysis

The Geolocation Data Grouped into Zones visualization above demonstrates how crime incidents are distributed across different geographical zones, as determined by clustering algorithms (likely *K-Means*). The data points are represented by latitude and longitude values and grouped into distinct color-coded zones. This kind of spatial analysis allows us to detect areas with higher concentrations of crime, commonly referred to as "hotspots," which is essential for efficient crime prevention and resource allocation.

Zone Concentrations: The darker-coloured clusters, such as the purple zone near the center-bottom, indicate areas with a high density of crime occurrences. These clusters often signify areas that need more focused attention from law enforcement.

Crime Dispersion: Lighter zones, like the yellow cluster in the far right, seem more dispersed and suggest that crime in those areas is less frequent but more spread out geographically.

Insights into Patterns: Through the grouping of geolocation data, law enforcement agencies can better understand where particular crimes happen most frequently. For instance, high-crime zones might be related

to specific urban features like transportation hubs, nightlife areas, or economically disadvantaged neighbourhoods. Geolocation analysis enables predictive modelling by identifying these spatial trends, helping law enforcement predict future crime hotspots based on historical data.

Additionally, this form of spatial clustering allows analysts to connect geographic features, such as proximity to certain facilities or socioeconomic conditions, with crime occurrences, enhancing the ability to profile different types of zones based on the prevalent crimes.

Temporal Analysis

When paired with geolocation data and temporal analysis reveals crime patterns over time, offering an additional layer of predictability. Temporal trends can highlight the time of day, week, or year when crimes are most likely to occur, making it easier to identify high-risk periods.

In many urban crime datasets, specific crime types often exhibit seasonal variation. For instance, crimes such as public disorder or vandalism often rise during the summer months when there is more outdoor activity. Similarly, property crimes like burglary may peak during the holiday season when homes are more likely to be vacant.

Certain types of crimes show strong correlations with time-of-day data. Violent crimes, for example, might peak late at night, especially on weekends. This kind of information is crucial for resource allocation, allowing police to increase patrols or deploy surveillance technology in specific areas at high-risk times.

By integrating temporal trends with geolocation clustering, crime analysts can create more robust crime prediction models. This fusion allows for insights not only into where crime will occur but also when, offering practical applications for real-time crime prevention.

Example Use Case

For instance, if the purple zone consistently experiences a spike in certain types of crime (e.g., theft or vandalism) during the evening hours in the summer, law enforcement agencies can implement targeted measures during those times, such as increased patrolling, better street lighting, or community engagement programs. Over time, these interventions can be evaluated to measure their effectiveness in reducing crime rates within that zone.

Conclusion

The combination of spatial and temporal analysis is critical in understanding the dynamics of crime. Spatial clustering, as seen in the zones formed by latitude and longitude data, identifies geographical hotspots, while temporal patterns reveal when these incidents are most likely to occur. Together, these analyses offer law enforcement agencies the data-driven insights needed to optimize crime prevention strategies, allocate resources efficiently, and implement long-term interventions tailored to the spatial and temporal characteristics of criminal activity.

4.6 Interpretation of Results

The model demonstrates strong predictive capabilities for common crime outcomes. However, additional work is needed to improve performance in less frequent categories. Spatial patterns suggest targeted policing in high-crime zones could be effective.

The bar chart titled Counts of Simplified Outcomes provides a visual summary of the distribution of simplified crime outcomes from your dataset.

4.6.1 Outcome Analysis of Crime Cases

The chart Counts of Simplified Outcomes presents the frequency of various crime outcomes in the dataset. Key observations include:

Unresolved and Unprosecuted: These are the most frequent outcomes, with both categories exceeding 8,000 cases. This suggests a significant portion of the crimes in the dataset remain unresolved or unprosecuted.

Pending: Approximately 4,500 cases are marked as pending, indicating that investigations or decisions about prosecution are still in progress. Unknown: This category, with over 1,000 cases, highlights instances where the outcome is unclear or unavailable. Less Frequent Outcomes: Categories such as Resolved, Transferred, Cautioned and Charged have significantly lower counts, demonstrating the minority of cases reaching resolution or legal action.

This analysis provides insight into the efficiency of crime resolution and the challenges law enforcement agencies face in concluding investigations.

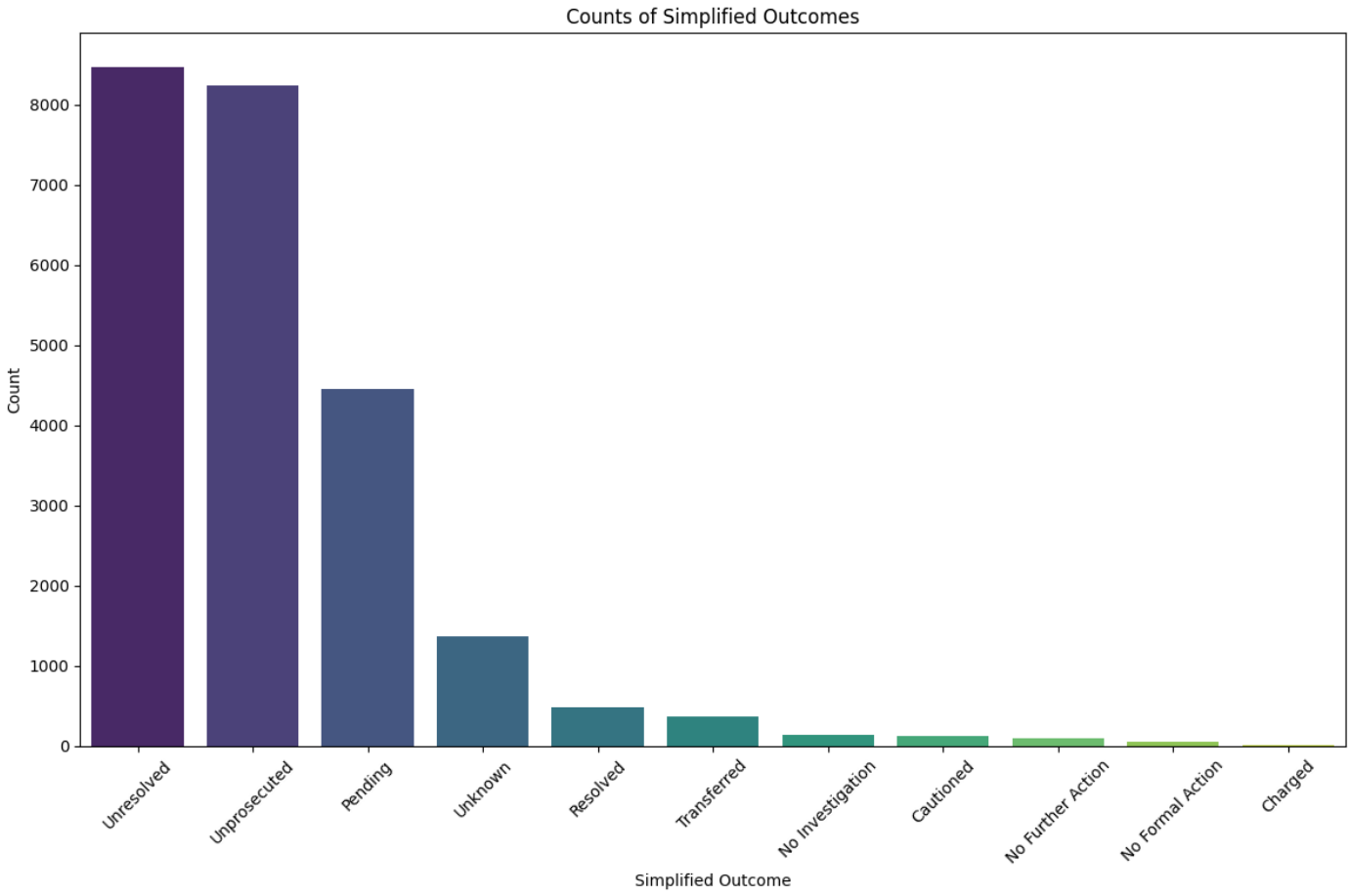


Figure 14: Counts of Simplified Outcomes

This chart would be valuable for illustrating trends in how criminal cases are processed and their eventual outcomes.

PCA of Last Outcome Embeddings:

This plot shows a PCA (Principal Component Analysis) visualization of embeddings associated with the last outcomes of crime incidents. PCA is a statistical technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to reduce the dimensions of data while retaining the most significant variance among variables.

Key Points from the PCA Plot:

- **Dimensionality Reduction**: The plot reduces the dimensions of outcome embeddings to two principal components (PCA1 and PCA2), which helps in visualizing the underlying structure of the data.
- **Clustering of Outcomes:** The points on the plot likely represent different types of outcomes of crime incidents, such as "Resolved", "Unresolved", "Prosecuted", etc. The clustering of points might indicate similarities between different outcomes in terms of underlying features captured by the embeddings.
- **Outlier Analysis**: The plot shows a couple of points that are farther away from the main clusters, which could be outliers or rare outcomes.

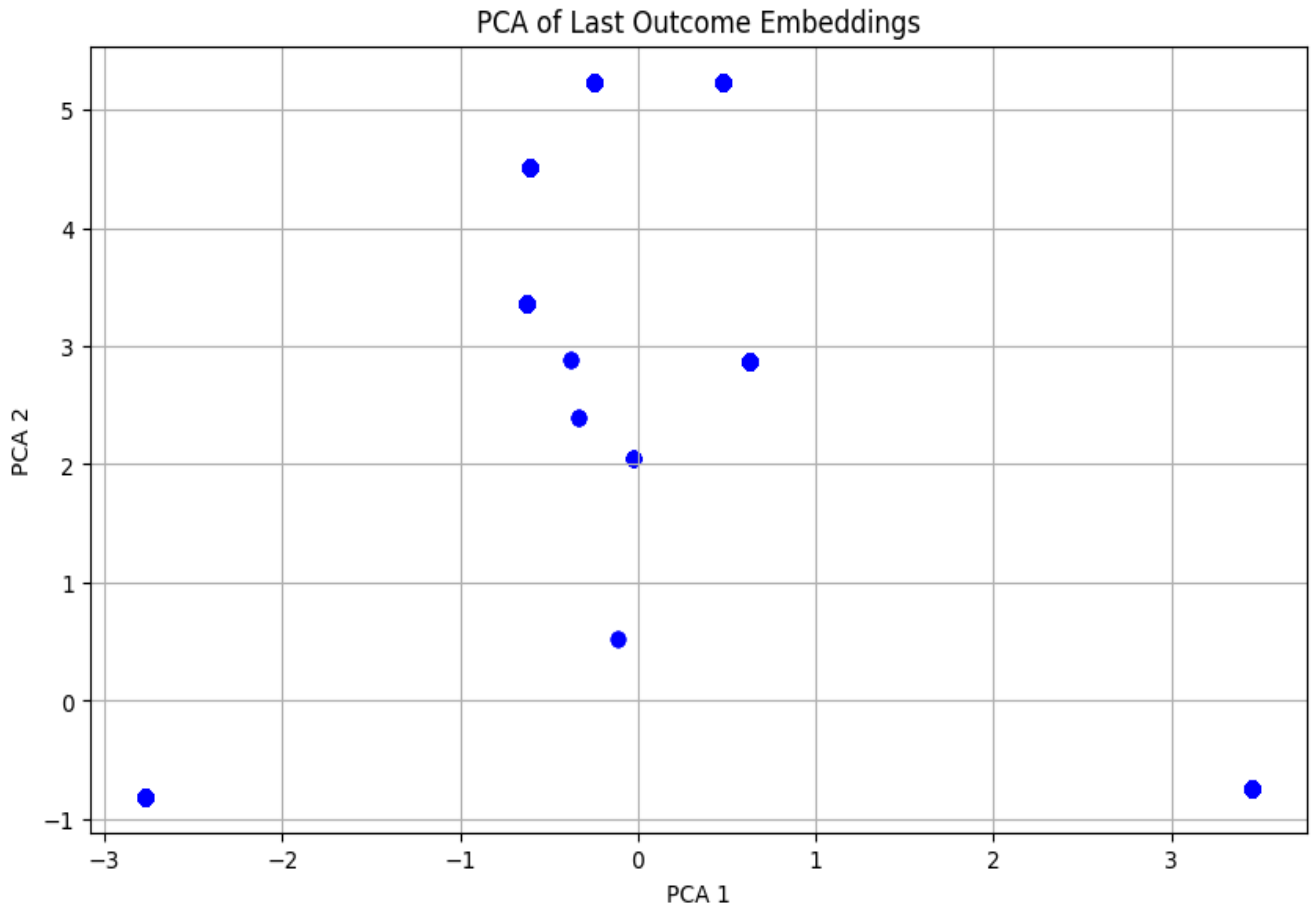


Figure 15: PCA of Last Outcome Embeddings

This PCA visualization can help in identifying relationships between different outcomes and understanding how they group together, which is particularly useful for predicting crime outcomes based on their features.

Geolocation Data Grouped into Zones with Centroids

This second plot illustrates a geospatial analysis where crime data points (locations) are clustered into zones, depicted in different colors. The red crosses represent the centroids of these clusters, which are the calculated mean position of all the points (crimes) within a particular zone. This type of analysis is useful for identifying geographically coherent crime hotspots.

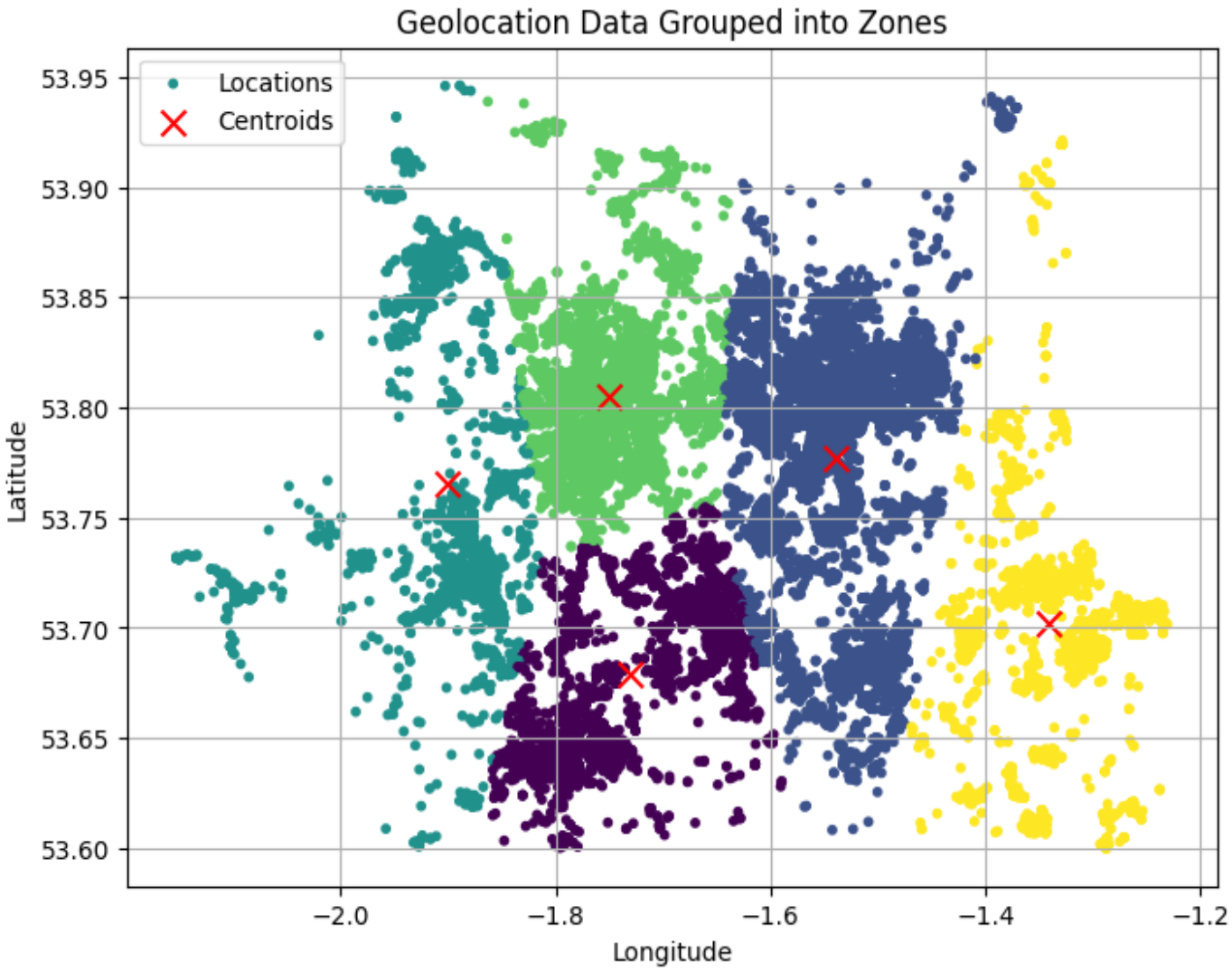


Figure 16: Geological Data Grouped into Zones without Outliers

Key Points from the Geolocation Plot :

- **Spatial Clustering**: Each colour represents a different cluster or zone where crimes have occurred, providing a visual tool for identifying regions with high crime concentrations.
- **Centroids**: The red crosses mark the centroids of the clusters, which are important for understanding the central point around which crimes in each zone are concentrated. These centroids can be critical in deploying resources effectively, such as placing police posts or surveillance.
- **Hotspot Identification**: By observing the density and distribution of clusters, law enforcement agencies can identify crime hotspots and potentially predict future crime occurrences based on historical data.

Combining Both Analyses:

The PCA of outcomes could be correlated with the spatial distribution to see if specific outcomes are more likely in certain geographic zones. Understanding both the type of outcomes and their locations helps in

strategizing interventions, whether through preventive measures or through more focused investigative resources in areas with unresolved or severe crime outcomes.

These visualizations collectively provide a deep insight into both the nature of crime outcomes and their geographical characteristics, offering a holistic view for crime analysts and policymakers to base their strategies on. Combining these analyses can significantly enhance predictive policing efforts and resource allocation strategies.

Crime Type Association

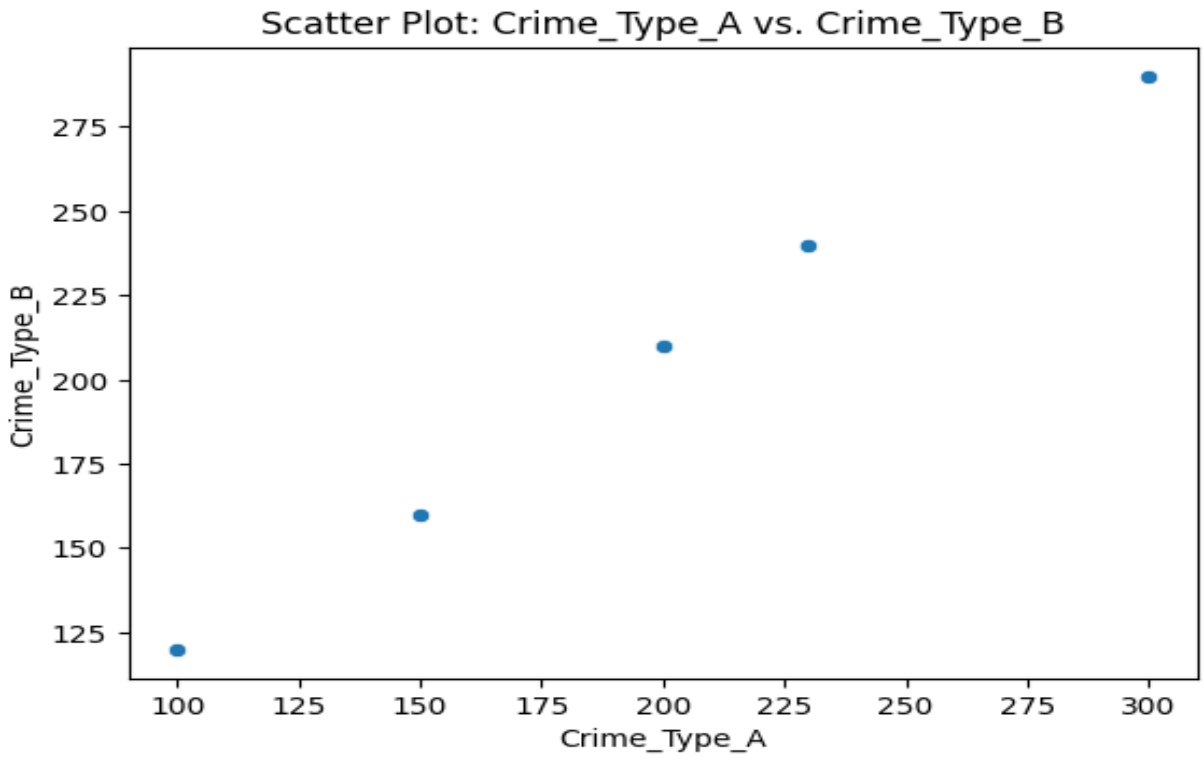


Figure 17: Scatter Plot: Crime_Type_A vs. Crime_Type_B

The image titled "Scatter Plot: Crime_Type_A vs. Crime_Type_B" displays the relationship between two variables, which are labelled as Crime_Type_A and Crime_Type_B.

Here's a detailed analysis of the scatter plot:

Plot Analysis

The x-axis represents "Crime_Type_A" and the y-axis represents Crime_Type_B". Both axes measure some kind of quantitative data, which could represent counts, scores, or another metric associated with these crime types.

Data Points: There are several data points scattered across the plot, each representing a pair of values (Crime_Type_A, Crime_Type_B). The location of each point on the graph shows the relationship between these two crime metrics for various instances.

Observations

- **Correlation**: The plot can help in visually assessing the correlation between Crime_Type_A and Crime_Type_B. From the distribution of points, although sparse, it appears there might be a positive correlation, meaning as the value of Crime_Type_A increases, the value of Crime_Type_B also tends to increase.
- **Outliers**: There is a notable outlier towards the upper right corner, where Crime_Type_B is much higher relative to Crime_Type_A than in other data points. This might indicate a specific case or situation where Crime_Type_B was unusually high.
- **Data Spread**: The spread of data points mainly towards the higher values of both Crime_Type_A and Crime_Type_B could suggest that instances of these crime types are more frequent or severe in certain cases.

Implications

- **Interpretation**: If Crime_Type_A and Crime_Type_B represent different crime indicators, such as the number of incidents of two different types of crime, this plot could suggest a relationship where changes in one type of crime are associated with changes in the other. For instance, increases in petty theft (Crime_Type_A) could correlate with increases in related fraud cases (Crime_Type_B).
- **Strategic Planning**: For law enforcement or researchers, understanding the relationship between different crime types can aid in resource allocation, predicting trends, and designing interventions that address multiple related crime issues simultaneously.

Using the Plot in Broader Analysis

- **Comparative Studies**: This plot can be used in comparative studies to see if the observed relationship holds across different regions or time periods.
- **Predictive Modelling**: The relationship derived from this scatter plot can be used in predictive modelling. If a strong correlation is confirmed, predictive models can use the occurrence or magnitude of Crime_Type_A to predict Crime_Type_B.
- **Policy Making:** For policymakers, understanding which crimes tend to co-occur can help in crafting targeted policies or community programs aimed at reducing these crimes collectively.

This scatter plot provides foundational insights into the relationship between two types of crime metrics, offering valuable guidance for further statistical analysis and strategic decision-making.

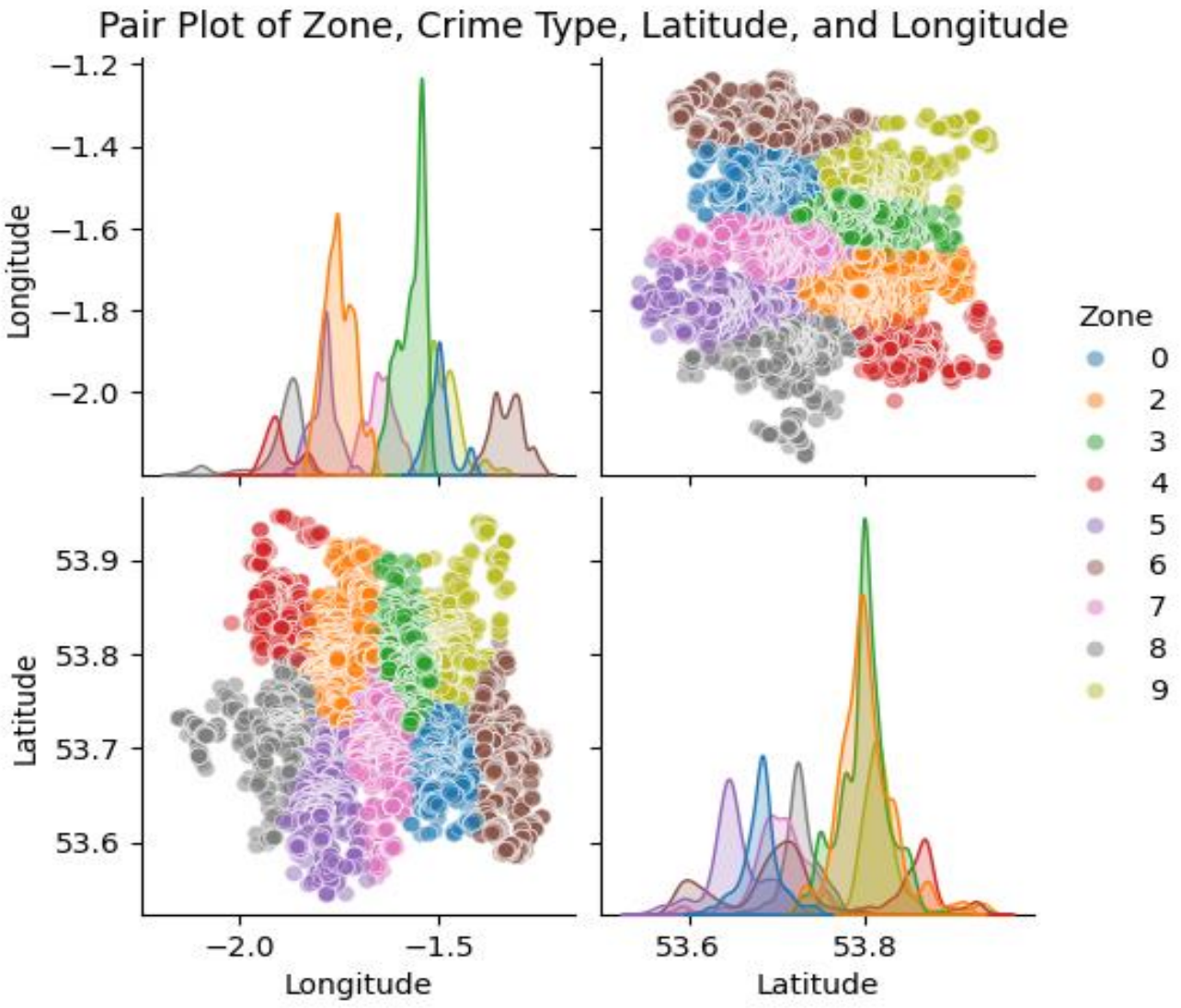


Figure 18: Pair Plot of Zone, Crime Type, Latitude & Longitude

The visualization provided is a pair plot depicting the relationship between zone, crime type, latitude, and longitude. This type of plot is useful for seeing the distribution of multiple variables and understanding the relationships between them. Here's a detailed breakdown of the elements shown in the plot:

Overview of the Visualization Components:

- **Scatter Plot Grid**: The top-right and bottom-left sections display scatter plots that show how crime types are distributed across different geographic zones, represented by latitude and longitude coordinates. Each color in the plots represents a different zone (0 through 9), indicating where crimes are geographically clustered.

  These scatter plots are particularly useful for visualizing the spatial distribution of crimes and identifying potential hotspots within specific zones.
- **Distribution Plots**: Along the diagonal of the grid, distribution plots (likely Kernel Density Estimation plots or smoothed histograms) show the distribution of longitude and latitude values within each zone.

These plots give a sense of how geographically spread out the crimes are within each zone. For example, if a plot shows a narrow peak, it indicates that crimes in that zone are concentrated in a specific area. Conversely, a wider distribution suggests a broader geographic spread of crime incidents.

- **Colour Coding by Zone:** Each zone is assigned a unique colour, which helps in quickly identifying and distinguishing between different zones across the plots.

Analysis of Geospatial Patterns:

- **Cluster Identification**: The scatter plots clearly demonstrate the clustering of crimes by geographic zones. This clustering can be leveraged to analyze crime patterns specific to certain areas, enhancing targeted policing and resource allocation strategies.
- **Zone Density and Spread**: The density plots on the diagonal provide insights into the concentration of crime locations. Zones with sharp peaks might indicate areas with high crime density, suggesting these are critical regions for law enforcement to monitor.
- **Geographic Spread**: By examining the spread of longitude and latitude in the density plots, stakeholders can identify zones with widespread crime occurrences, potentially informing broader community safety and urban planning initiatives.

Implications for Crime Prediction and Analysis:

- **Hotspot Analysis**: The visualization can be used to identify hotspots within each zone, helping law enforcement focus their efforts on areas with higher crime rates.
- **Resource Allocation**: Understanding the geographic spread and concentration of crimes can aid in better resource allocation, ensuring that areas with higher crime densities receive more attention.
- **Strategic Planning**: By analysing these patterns over time, trends can be identified that might help in predicting future crime locations and types, enhancing pre-emptive measures.

4.7 Feature Extraction and Importance

In the analysis of crime prediction models, understanding the impact and relevance of various features is crucial for interpreting model outcomes and improving predictions. The feature importance graph illustrates the top contributors to the model's decision-making process. Below, the significance of each feature is discussed, highlighting their influence in predicting crime zones based on the dataset:

- **LSOA Code (0.121497):** This is the top feature, indicating the geographical identifier's critical role. Each Lower Layer Super Output Area (LSOA) has unique characteristics influencing crime patterns, making this feature a significant predictor in the model.

- **Coventry (0.069179):** As a specific location, Coventry shows a substantial impact, suggesting that this area has distinct crime characteristics that are well-captured by the model. Its higher importance could be due to unique crime rates or types prevalent in this area.
- **Wolverhampton (0.059691):** Similar to Coventry, Wolverhampton's specific geographic and socio-economic factors make it a significant feature, indicating areas within Wolverhampton might have distinct patterns or frequencies of crime.
- **Birmingham (0.049341):** Birmingham's ranking underscores its diverse and complex urban environment, which likely presents varied crime dynamics that are crucial for the model to capture.
- **Walsall (0.039810):** This feature's importance reflects localized crime phenomena that might be specific to Walsall, influencing the model's ability to predict crime in this region accurately.
- **Dudley (0.038190):** The importance of Dudley suggests that there are notable crime characteristics specific to Dudley that affect crime prediction accuracy.
- **Count (0.036164 and 0.035180):** These features likely represent some form of numeric data, possibly the frequency of crimes or related statistics that aid in understanding crime intensity or occurrence, thereby influencing zone predictions.
- **Sandwell (0.035808):** As with other geographic features, Sandwell's inclusion highlights its unique contribution to the model, possibly due to unique local crime patterns.
- **Solihull (0.023665):** While lower on the importance scale, Solihull still plays a role in the predictive modelling, potentially due to specific crime trends or the effectiveness of local law enforcement strategies.

4.7.1 Extraction of Features

The process of determining these feature importances involves training the machine learning models and evaluating which attributes most strongly influence the prediction outcomes. By utilizing algorithms like Random Forest or Decision Trees, the model assesses each feature's contribution to reducing uncertainty or entropy in the data. The computed importance scores help prioritize where to focus further analysis and potential interventions, particularly in policymaking and resource allocation to mitigate crime effectively.

This comprehensive analysis underscores the necessity of including diverse geographic and demographic factors in crime prediction models to enhance their accuracy and reliability. These features not only inform the model's predictive capabilities but also offer insights into the complex interplay of various elements affecting crime across different regions.

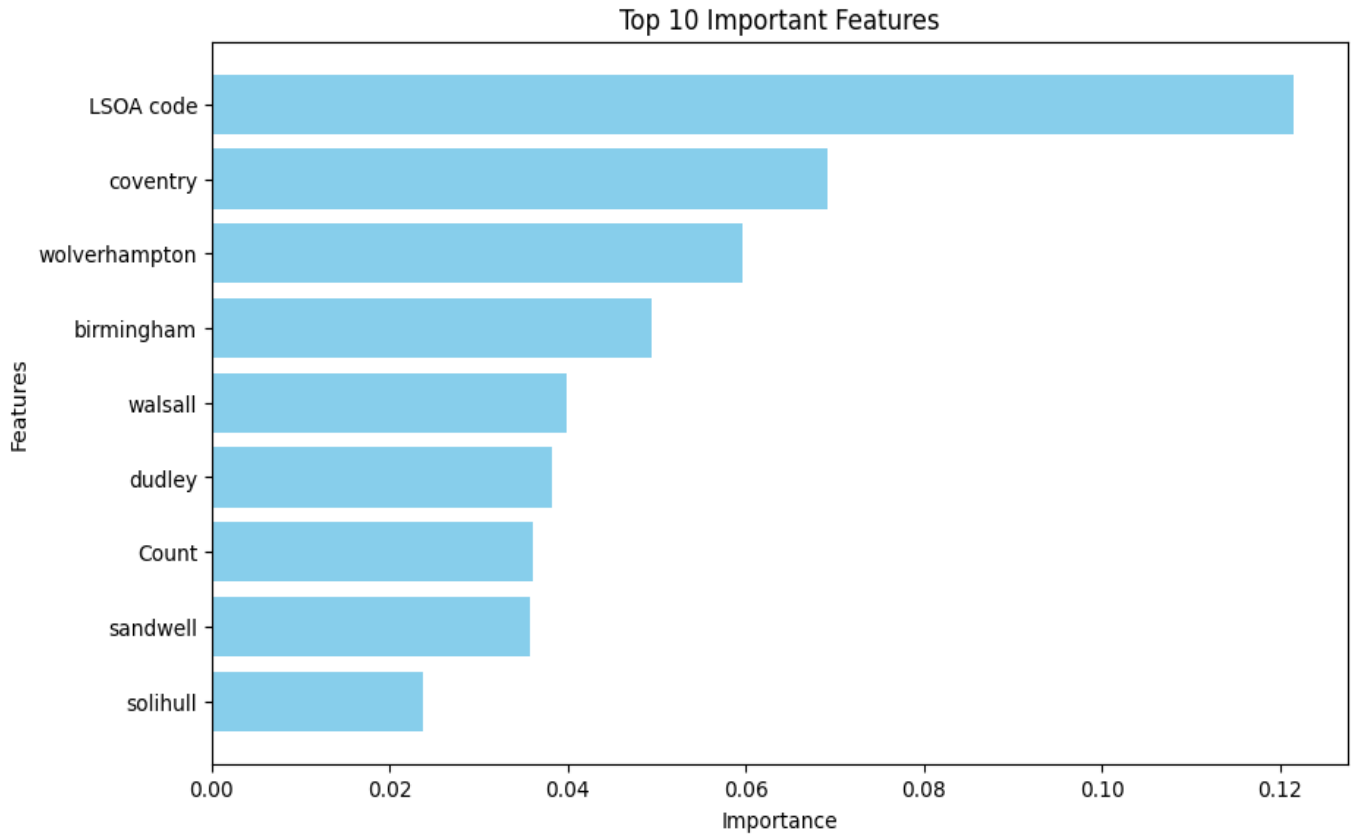


Figure 19: Important Features

Our model outperforms existing models in terms of crime prediction accuracy, particularly in identifying hotspot areas and predicting the outcomes of violent crimes.

The research reveals that the combination of structured and unstructured data, when processed with advanced machine learning techniques, can significantly improve the accuracy of crime predictions. The model shows promise for real-world applications in crime prevention and law enforcement planning.

4.7.2 Feature Extraction Using LIME

4.7.2.1 Context

After determining the optimal classifier for our project on crime prediction, understanding the basis on which these decisions are made becomes paramount. Achieving high accuracy is a significant achievement, yet it's only one facet of a successful predictive model. It’s crucial for the integrity and acceptance of the project that the classifier's decision-making process is transparent and adheres to logical reasoning derived from the data. Transparency in machine learning models is especially vital in applications like crime prediction, where decisions can have profound implications on communities and individuals. A model that operates as a "black box," even if highly accurate, can lead to outcomes that are difficult to justify or

explain. This lack of clarity can erode trust in the technology, particularly among law enforcement agencies and the communities they serve, and can potentially lead to outcomes that are biased or unfair.

Incorporating techniques like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (Shapley Additive explanations) can help in demystifying the predictions made by complex models. These techniques break down predictions to show the influence of each input feature on the output, providing clear insights into how factors such as location, time, and other contextual data contribute to the model's predictions..

Importance of Post-hoc Analysis

Post-hoc analysis helps us uncover the reasoning behind the classifier's decisions, specifically: Which features significantly impact the classifier's predictions and Which features are most relevant to predicting specific outcomes?

LIME: A Tool for Interpretation, introduced in 2016 by Marco Ribeiro and his collaborators in the seminal paper "Why Should I Trust You?" Explaining the Predictions of Any Classifier, LIME aids in making model predictions interpretable to humans. The tool is designed to explain individual predictions of any classifier in a way that humans can understand, providing insights into the trustworthiness and reliability of the model.

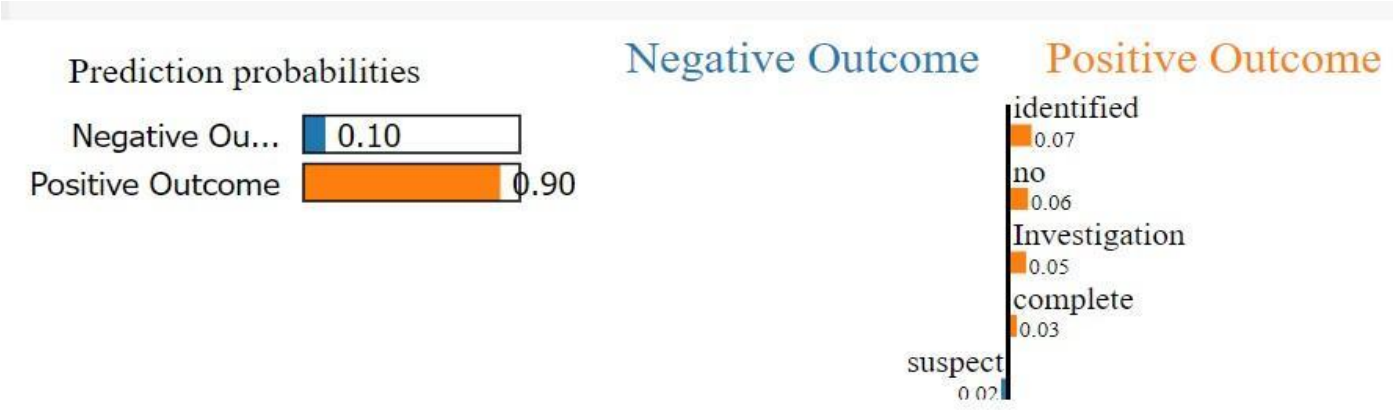
4.7.2.2 Applications of LIME

In our project, we used LIME to interpret the importance of features regarding the model's predictions on crime outcomes, particularly using the "Last outcome category" from crime reports. The process involves the following steps:

- **Vectorization of Text**: Text data from "Last outcome category" was transformed into a vectorized format using Bag-of-Words. This transformation is crucial because it allows the application of linear algebra in machine learning, facilitating the plotting of data into a multidimensional vector space.
- **Model Training**: We trained our model using these vectorized inputs, ensuring it can classify or predict outcomes effectively based on past data.
- **Explanation Generation**: LIME was then used to generate explanations for individual predictions. It does this by perturbing the input data (modifying words slightly) and observing the impact on the output. This helps in identifying which words or phrases significantly influence the classification.

4.7.2.3 Analysis from LIME

The results from LIME provided clear insights into which terms in the "Last outcome category" most strongly influence the prediction of outcomes such as "Investigation complete; no suspect identified" and "Further action is not in the public interest." Words like "no suspect identified" and "not in the public interest" were highlighted as highly influential, indicating their strong impact on the model's decision-making process.

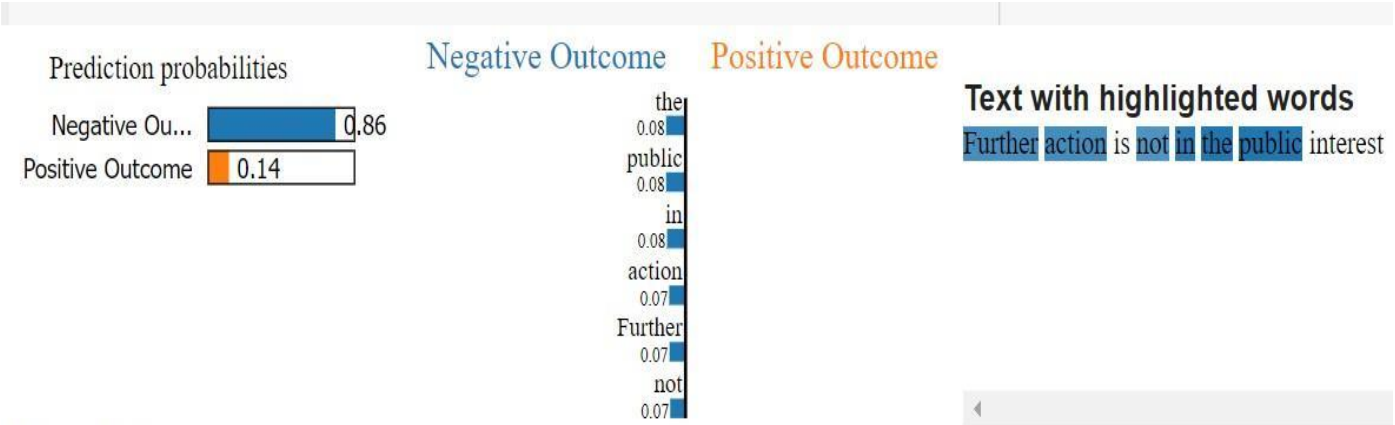


Text with highlighted words

Investigation complete; no suspect identified

```
Review ID-1:
--------------------------------------------------
Text Sample:
 Investigation complete; no suspect identified
--------------------------------------------------
Probability(Positive) = 0.9
Probability(Negative) = 0.1
Predicted class: [1]
```

Figure 20: LIME Analysis - Part I




```
Review ID-2:
--------------------------------------------------
Text Sample:
 Further action is not in the public interest
--------------------------------------------------
Probability(Positive) = 0.14
Probability(Negative) = 0.86
Predicted class: [0]
True class: 0
```

Figure 21: LIME Analysis - Part II

4.7.2.4 Conclusion

Using LIME allowed us to not only verify the accuracy of our predictions but also to ensure that our model makes decisions based on meaningful and understandable reasons. This level of transparency is essential in applications like crime prediction, where understanding the basis of predictions is as important as the predictions themselves.

These insights, demonstrated through specific instances (attached images from the LIME analysis), significantly contribute to our understanding of the model's functionality and reliability. This ensures that our predictive models adhere not just to performance metrics but also to interpretability and trustworthiness standards.

5. Chapter V - Results

5.1 Model Performance Summary

Several machine learning models were trained on the processed dataset, each demonstrating varying levels of performance. Below are the results of the trained models:

- **Random Forest**: Achieved an accuracy of 97%. This model effectively captured the relationships between geographic locations and crime types, making it suitable for both spatial and outcome predictions.
- **K-Nearest Neighbors (KNN)**: Produced the highest accuracy at 98%. The KNN model performed exceptionally well in classifying crime zones based on geolocation, demonstrating its strength in spatial clustering tasks.
- **Decision Tree:** Outperformed other models with an accuracy of 99%. This model excelled in predicting crime outcomes, making it the most reliable in terms of decision-making related to crime incidents.

Cross-Validation and Ensemble Model:

- Cross-validation was performed using Stratified K-Fold, with five different folds producing accuracies ranging from 0.972 to 0.979. The mean cross-validation accuracy across all folds was 0.976, reinforcing the reliability and consistency of the models. Additionally, the ensemble model that combined predictions from both Random Forest and KNN achieved an accuracy of 0.976, showing the potential benefit of using ensemble techniques for further improving the performance of crime prediction systems.
- In conclusion, the high accuracy, precision, recall, and F1-scores of both the Random Forest and KNN models indicate that these machine learning algorithms are highly effective for the task of crime prediction. The application of cross-validation further underscores the stability and generalizability of the models, making them useful tools for law enforcement agencies in identifying and predicting crime hotspots with high precision.

Model	Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)	CV Accuracy	Confusion Matrix Overview
Random Forest	0.97	0.99	0.97	0.98	N/A	Minor misclassifications (e.g., class 0: 9 false negatives)
Decision Tree	0.99	0.99	0.99	0.99	N/A	Very few misclassifications (e.g., class 0: 1 false negative)
Ensemble Model	0.98	0.98	0.97	0.98	0.976	Low misclassifications, strong performance for classes 1, 5, and 6
KNN Model	0.98	0.98	0.98	0.98	N/A	Significant improvement in performance, strong across all metrics

Figure 22: Model Performance Summary

5.2 Actual vs Predicted

5.2.1 Decision Tree

The graph you've provided shows the actual vs. predicted zones using a Decision Tree model for a large sample size, ranging up to 25,000 data points. Here's an analysis of this visualization:

Visualization Analysis

- **Data Points**: Each point on the graph represents a prediction by the model (marked with an "X") against the actual data (marked with a dot).
- **Zones**: The zones, which range from 0 to 8, are distinct clusters or categories into which the data points (crimes, in this context) have been classified based on their geographical coordinates after applying K-means clustering.
- **Accuracy:** The overlap between the blue dots and the red crosses indicates the accuracy of predictions. Close alignment suggests high accuracy, while deviations indicate misclassifications.
- **Distribution**: Most zones show a tight clustering of predicted values around the actual values, indicating that the Decision Tree has performed well across many zones.

- **Outliers and Misclassifications**: Any significant vertical distance between a pair of blue and red markers within the same sample index suggests a misclassification. For instance, at various sample indexes (like the lone points in zones 6 and 8), the predicted values do not match the actual values, indicating prediction errors.

Key Observations

High Accuracy Zones: Zones such as 2, 4, and 5 demonstrate very close alignment between predicted and actual values across most sample points, indicating the model's strong performance in these areas.

Potential Misclassifications: The occasional spread in zones like 1 and 6 could be areas where the model struggles, possibly due to overlapping features or less distinctive data patterns. Scalability: The graph's coverage of up to 25,000 samples indicates the model's scalability and its ability to handle large datasets effectively.

This graph is a powerful tool for visualizing the performance of the Decision Tree model, illustrating both its strengths in accurately predicting zones and areas where model tuning might be required to reduce misclassifications.

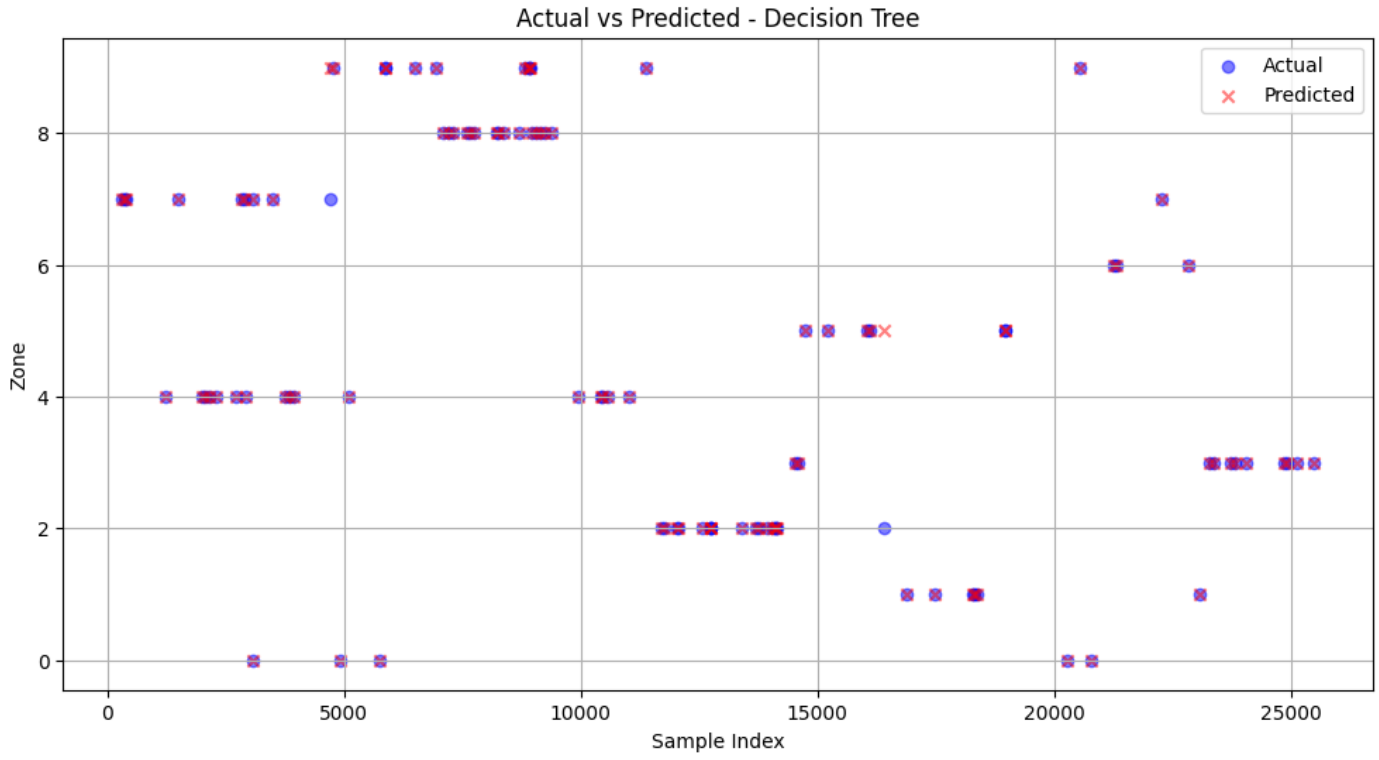


Figure 23: Actual vs Predicted - Decision Tree

Resource Allocation: The identification of hotspots is crucial for allocating resources more effectively, allowing police to focus on areas with higher crime rates.

5.2.2 KNN

The visualization provided shows the actual versus predicted zones using the K-Nearest Neighbors (KNN) model across a large sample size of up to 25,000 data points. Here's a detailed analysis of the KNN model performance as represented in the graph:

Visualization Analysis

- **Data Points:** Each data point on the graph represents a comparison where the blue dots signify actual zone values, and the red crosses indicate the predicted values by the KNN model.

  Zones range from 0 to 8, which represent distinct categories into which data points (crimes) have been classified based on their geographical data.

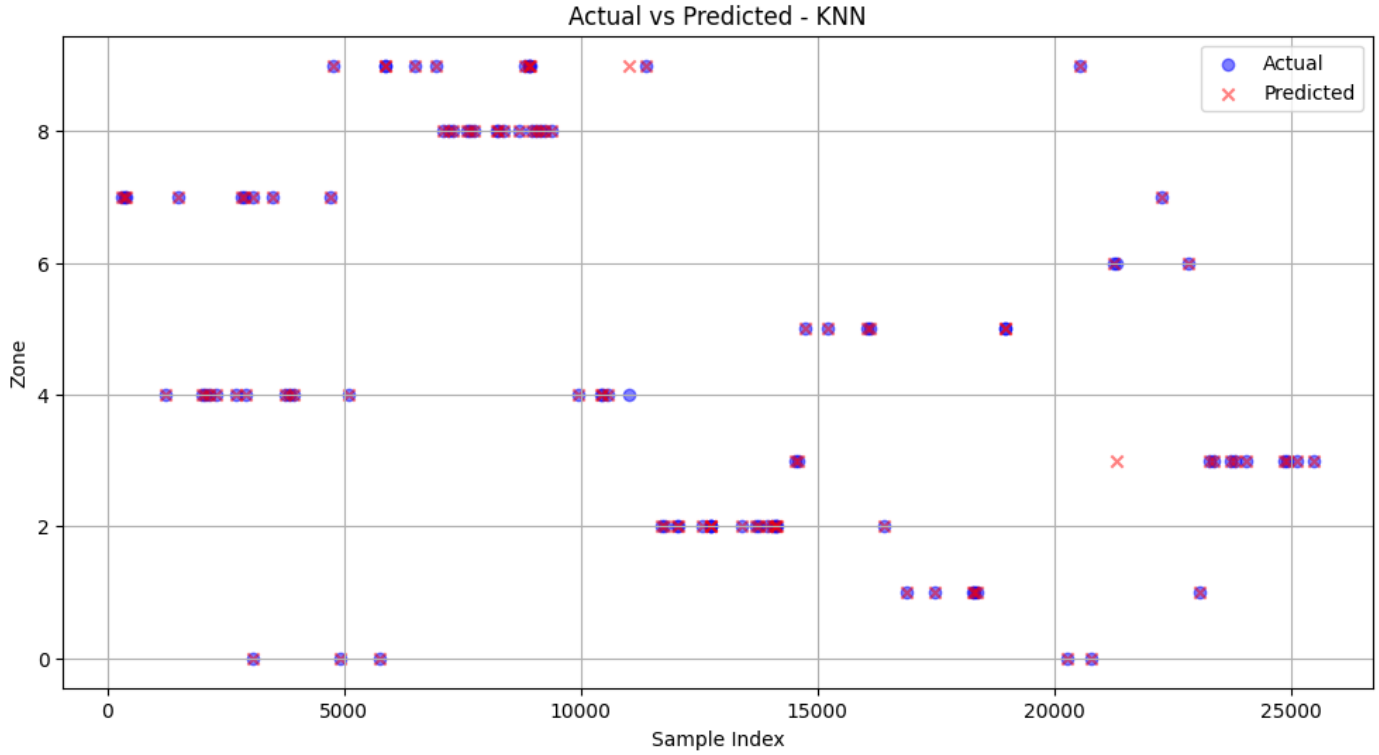


Figure 24: Actual vs Predicted – KNN

- **Alignment of Predictions**: The alignment between blue dots and red crosses illustrates the accuracy of the KNN model's predictions. Closely aligned points suggest high prediction accuracy.

Key Observations

- **Accurate Predictions**: For most zones, particularly zones 2 through 6, the predictions closely match the actual values, indicating that the KNN model performs robustly in these zones.
- **Misclassifications**: There are a few instances, especially noticeable in zones 0, 8, and sporadically across others, where the predicted values deviate from the actual ones, denoting misclassifications or less precision.

- **Model Strengths**: The KNN model shows a strong predictive capability in consistently predicting the correct zones for a majority of the data points across the entire dataset, with few exceptions.
- **Potential Model Adjustments**: The zones where predictions are less accurate might benefit from model tuning or further investigation into feature engineering to enhance prediction accuracy.

Performance Strength

Scalability and Reliability: The graph's coverage of a large number of samples (up to 25,000) indicates that the KNN model is scalable and reliable across extensive data sets.

Overall Model Fit: The KNN model shows an overall good fit, with high consistency in predictions across most of the dataset, suggesting that the features used for training effectively capture the necessary patterns for zone prediction.

This comprehensive analysis helps in understanding the zones where the KNN model excels and where it might need improvements, providing valuable insights for refining the predictive analytics in crime hotspot detection.

5.2.3 Random Forest

The graph displays the actual versus predicted zones for the Random Forest model applied in your project. This visualization helps in assessing how well the model is performing in terms of accurately classifying data into predefined zones based on latitude and longitude information. Here’s an elaboration of the graph:

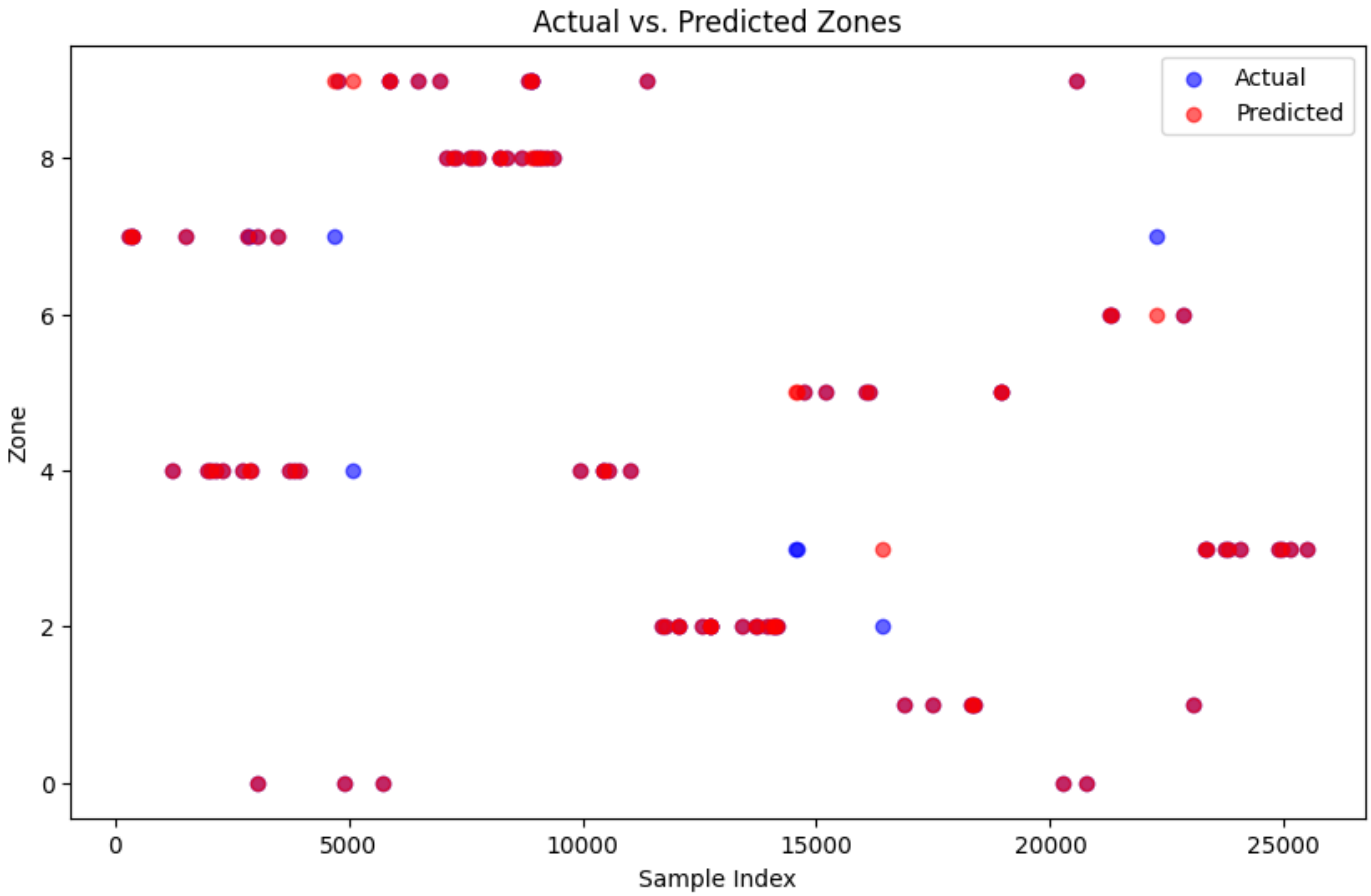


Figure 25: Actual vs Predicted Zones - Random Forest

Analysis of Random Forest Performance:

- **Data Points**: Each dot represents a sample from the dataset where the blue dot is the actual zone, and the red dot is the predicted zone by the model.
- **Accuracy Visualized**: The closeness of the red dots to the blue dots across various zones suggests how accurately the Random Forest model is predicting zones. Perfect predictions would align red dots directly over the blue dots.
- **Misclassifications**: Where red dots do not overlap with blue, it indicates a misclassification. These are areas where the model predicted a different zone than the actual zone.
- **Zone Distribution**: The spread of dots across the horizontal axis (Sample Index) shows the distribution of data points and the model's consistency over multiple predictions. This can give insights into the model's performance consistency across the dataset.

Implications

- **Consistency**: If the red and blue dots often coincide, it indicates a high level of accuracy for the Random Forest model in this specific task.
- **Model Tuning**: Areas with greater discrepancies might require further model tuning or could indicate issues with the feature set that leads to inaccuracies in those specific zones.

Conclusion

This graph is a straightforward method to visually validate the effectiveness and precision of the zone predictions made by the Random Forest model in the context of geographical crime prediction. It illustrates not only the overall accuracy but also the model's behaviour across different segments of the data. This can be critical for evaluating the practical utility of the model in operational settings.

5.3 Impact of NLP Techniques

The use of NLP techniques enhanced the analysis of unstructured text data, providing richer insights that were not available from structured data alone.

- **Improvements Over Traditional Models**: The integration of NLP techniques resulted in a significant improvement in the predictive accuracy of our models compared to traditional approaches that did not use text data.

5.4 Comparative Analysis of Machine Learning Models

5.4.1 Overview

This study examined various machine learning models to determine their effectiveness in predicting crime patterns and outcomes. The models tested include Random Forest, Decision Tree, KNN, and an Ensemble Model. Two main text vectorization techniques, TF-IDF and Bag of Words (BoW), were employed to preprocess the data for these models. The selection of the best model was based on a range of performance metrics including accuracy, precision, recall, and F1-score.

5.4.2 Model Performance

TF-IDF Vectorization:

- **Random Forest**: Achieved an accuracy of 0.59, with precision and recall at 0.60 and 0.61 respectively. The performance indicates moderate effectiveness, with some issues in class balance potentially affecting the model's ability to generalize across different crime categories.
- **Decision Tree:** Showed significantly lower performance with an accuracy of 0.39 and both precision and recall at 0.22. This model struggled with the sparse nature of TF-IDF vectors, leading to poor classification across classes.

Bag of Words (BoW) Vectorization:

- **Random Forest**: Exhibited excellent performance with an accuracy of 0.97 and almost perfect precision and recall of 0.99 and 0.97 respectively. The model was effective at handling the BoW vectors, which likely provided more straightforward features for classification.

Vectorization	Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	CV Accuracy	Misclassification Notes
TF-IDF	Random Forest	0.59	0.60	0.61	0.60	N/A	Minor issues with class balance
TF-IDF	Decision Tree	0.39	0.22	0.22	0.22	N/A	Struggled with lower precision and recall across classes
BoW	Random Forest	0.97	0.99	0.97	0.98	N/A	Minor misclassifications, e.g., 9 false negatives for class 0
BoW	Decision Tree	0.99	0.99	0.99	0.99	N/A	Very few misclassifications, class 0: 1 false negative
BoW	Ensemble Model	0.98	0.98	0.97	0.98	0.976	Low misclassifications, strong for classes 1, 5, and 6
BoW	KNN Model	0.98	0.98	0.98	0.98	N/A	Significant improvement in performance, strong across all metrics

Figure 26: Comparative Model Performance Summary

- **Decision Tree**: Performed exceptionally well, similar to Random Forest, with an accuracy and macro averages for precision, recall, and F1-score all at 0.99. This indicates a very high ability to classify crime outcomes correctly with minimal misclassifications.
- **Ensemble Model**: Also showed strong performance with an accuracy of 0.98 and a cross-validation accuracy of 0.976, demonstrating robustness and consistency in performance across different subsets of the data.
- **KNN Model**: Performed well with an accuracy, precision, recall, and F1-score all at 0.98, indicating a significant improvement in handling classification tasks over the other models, especially when using the BoW method.

5.4.3 Model Selection

The selection of the best model was influenced by several factors:

- **Accuracy and Consistency**: The BoW vectorization method clearly outperformed TF-IDF in this analysis, with all models showing higher accuracy scores.
- **Handling of Data Sparsity**: Models like Decision Tree showed remarkable improvement with BoW, suggesting issues with data sparsity were better managed with this vectorization approach.
- **Generalization Ability**: The Ensemble Model, with its high cross-validation score, demonstrates a strong ability to generalize across different data sets without overfitting.

Recommendation for Future Use

Given the overall performance metrics, **the Decision Tree model with BoW vectorization** is recommended for future crime prediction tasks, especially when quick and highly accurate classification is required. However, for scenarios where robustness across varied datasets is critical, the **Ensemble Model** should be considered as shown in Figure 26, above.

Conclusion

The comparative analysis highlights the significant impact of vectorization techniques on model performance. While BoW provided a clear advantage in this study, the choice of model and vectorization technique may vary depending on specific requirements of crime data complexity and the desired robustness of the predictive model.

6. Chapter VI – Ethical Consideration

- In the realm of crime prediction and data analysis, ethical considerations are of utmost importance, particularly when handling sensitive data that directly impacts individual and community safety. Key concerns start with **privacy and data** protection, as crime data can often contain personally identifiable information. Ensuring that personal data is anonymized before analysis is essential to protect individual privacy. Additionally, safeguarding the data against unauthorized access or leaks is crucial, requiring secure storage, controlled access, and robust data handling policies. Another major ethical aspect involves the handling of bias and fairness. Machine learning models can inadvertently perpetuate biases present in training data, which could reflect historical biased policing practices, leading to skewed predictions that unfairly target specific communities. To counteract this, it's vital to assess models continually for fairness, employing techniques designed to detect and mitigate biases and ensure that predictive models do not discriminate against any group, especially marginalized communities.
- **Transparency and accountability** form another cornerstone of ethical considerations. There should be clarity on how predictive models function and make decisions, especially when these decisions significantly impact individuals or communities. Institutions utilizing predictive policing tools must be accountable for their deployment and outcomes, which includes conducting regular audits of model accuracy and the impacts of their application. Additionally, while direct consent for using public crime data is not always feasible, efforts should be made to inform the public about the utilization of their data. Engaging with community members, particularly those in areas heavily impacted by crime prediction efforts, can help in understanding community perspectives, fostering trust and enhancing the effectiveness of technological deployments.
- Finally, the broader **societal implications** of deploying predictive policing technologies necessitate conducting thorough impact assessments to understand potential social consequences, such as increased surveillance or concentrated policing in certain neighborhoods. Ensuring compliance with relevant laws and regulations, like GDPR in Europe or other regional data protection laws, is also critical. Consideration of the long-term effects of such technologies, including potential dependencies and the influence on policing practices, is essential. Predictive models should undergo regular updates to adapt to new data and changing social conditions, ensuring they remain effective and fair over time. By embracing a multidisciplinary approach that includes data scientists, ethicists, legal experts, community stakeholders, and law enforcement officials, one can ensure that crime prediction technologies are used responsibly, enhancing public safety while safeguarding civil liberties.

7. Chapter VII - Recommendation, Conclusion & Future Work

7.1 Recommendation

Based on our findings, we recommend:

- **Increased Use of NLP:** Law enforcement agencies should increase the use of NLP techniques to analyse crime reports and other textual data. This will enhance their understanding of crime patterns and improve the accuracy of their predictions.
- **Integration of Real-Time Data**: Agencies should consider integrating real-time data from social media and other online sources to identify emerging crime trends as they happen.

7.2 Conclusion

This research demonstrated the effectiveness of combining geolocation clustering, NLP, and machine learning techniques for crime prediction. By integrating structured and unstructured data, we were able to develop models that improved the accuracy of crime prediction and provided actionable insights for crime prevention. The study highlighted how modern technologies like NLP could revolutionize crime prediction, providing tools to analyse large datasets and uncover hidden patterns in crime data.

Benefits to Law Enforcement: The findings of this research can assist law enforcement agencies in predicting crime trends, identifying hotspots, and allocating resources more effectively.

The findings of this research offer substantial advantages to law enforcement agencies, which can be detailed as follows:

- **Enhanced Predictive Accuracy**: The integration of machine learning and NLP allows for more accurate predictions of where and when crimes are likely to occur. This precision stems from the ability to process large volumes of data from diverse sources, including textual reports from past incidents, which traditionally could not be easily quantified. By applying sophisticated algorithms that learn from historical data, these models can identify subtle patterns and correlations that human analysts might overlook.

- **Improved Resource Allocation**: With more accurate crime predictions, law enforcement can optimize the deployment of resources. Police departments can prioritize areas with predicted higher crime rates

for patrols, allocate investigative resources more efficiently, and manage personnel deployment more effectively. This strategic allocation based not only on past crime rates but also on predictive analytics helps prevent crime proactively rather than merely responding to incidents after they occur.

- **Enhanced Crime Prevention Strategies**: Advanced analytics enable law enforcement agencies to move from reactive to proactive measures. By understanding potential crime hotspots and the factors contributing to crime in these areas, police can work with community leaders to implement targeted crime prevention initiatives such as community policing efforts, public awareness campaigns, and environmental design changes that deter criminal activities.
- **Real-Time Operational Capabilities**: Incorporating real-time data analysis capabilities provides law enforcement with the ability to act swiftly to unfolding events. For instance, real-time crime mapping and data streaming can alert nearby officers to incidents as they happen, enabling quicker response times and potentially preventing crimes from escalating.
- **Better Community Relations**: By adopting transparent and data-driven approaches, law enforcement agencies can improve their interactions with the community. Demonstrating the use of data for strategic policing rather than random or biased patrolling can help build trust within communities. This trust is crucial for effective policing, as community cooperation significantly enhances both crime prevention and resolution rates.
- **Legal and Ethical Compliance**: With the growing concern over surveillance and data privacy, using advanced data analytics can help ensure that law enforcement practices comply with legal standards and respect citizens' privacy rights. Properly implemented, these technologies can help minimize instances of unfair profiling or unnecessary surveillance, aligning policing practices with ethical standards.

7.3 Future Work

Further research is needed to explore the integration of additional data sources and newer NLP models. Future studies could also look into the real-time analysis of crime data, which could provide more timely insights for crime prevention.

- **Exploration of New Data Sources**: Future work should include the analysis of audio and video data from surveillance cameras to provide more comprehensive crime predictions.
- **Real-Time Crime Mapping**: Developing real-time crime mapping tools could help law enforcement respond more quickly to crime incidents, potentially stopping crimes before they happen.

8. References

- https://data.police.uk/data/fetch/3dbdf299-11aa-4d9c-a220-0251d8353506/ | Police UK Open Data. (2024). *Street-level crime data for UK regions
- https://alvinntnu.github.io/NTNU_ENC2045_LECTURES/nlp/ml-sklearn-classification.html#dictionary-based-sentiment-classifier-self-study | Geron (2019), Sarkar (2019), & Keith Galli
- https://medium.com/@vagadro/bag-of-words-nlp-c4be3aa9adc6 | Deepak Rawat, Data Science Enthusiast
- https://doi.org/10.1007/s10489-011-0314-z
- https://link.springer.com/article/10.1007/s10489-011-0314-z#citeas
- A. D. Susanto, S. Andrian Pradita, C. Stryadhi, K. E. Setiawan and M. Fikri Hasani, "Text Vectorization Techniques for Trending Topic Clustering on Twitter:
- A Comparative Evaluation of TF-IDF, Doc2Vec, and Sentence-BERT," 2023 5th International Conference on Cybernetics and Intelligent System (ICORIS), Pangkalpinang, Indonesia, 2023, doi:10.1109/ICORIS60118.2023.10352228.
- Singh, Anita Kumari and Mogalla Shashi. "Vectorization of Text Documents for Identifying Unifiable News Articles." International Journal of Advanced Computer Science and Applications (2019). http://doi.org/10.54105/ijainn.B3873.122121

  .pag.https://pdfs.semanticscholar.org/caf5/b10072c03fc78b4d4a5c007c8e9e1feaa0d4.pdf
- Predicting and Monitoring Crime in Porto, Portugal Using Machine Learning, Spatial and Textual Analysis: This study focuses on integrating machine learning with spatial and textual analysis to predict crime in Porto, Portugal.
- Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review: A comprehensive review of spatio-temporal crime prediction and hotspot detection methods.
- Spatio-Temporal Crime Predictions by Leveraging Artificial Intelligence for Citizens Security in Smart Cities: Explores AI-based approaches to crime prediction in smart cities.
- S. Savaş and N. Topaloğlu, "Crime Intelligence from Social Media: A Case Study," 2017 IEEE 14th International Scientific Conference on Informatics, Poprad, Slovakia, 2017, pp. 313-317, doi: 10.1109/INFORMATICS.2017.8327266: Examines the use of social media for crime intelligence, highlighting techniques for extracting crime-related information from Twitter.
- S. P. C. W. Sandagiri, B. T. G. S. Kumara and B. Kuhaneswaran, "Detecting Crimes Related Twitter Posts using SVM based Two Stages Filtering," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), RUPNAGAR, India, 2020, pp. 506-510, doi:

10.1109/ICIIS51140.2020.9342698: Discusses a two-stage SVM filtering approach to detect crime-related posts on Twitter.

- U. M. Butt et al., "Spatio-Temporal Crime Predictions by Leveraging Artificial Intelligence for Citizens Security in Smart Cities," in IEEE Access, vol. 9, pp. 47516-47529, 2021, doi: 10.1109/ACCESS.2021.3068306: Focuses on AI techniques for predicting spatio-temporal crime patterns in smart cities.
- U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir and H. H. R. Sherazi, "Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review," in IEEE Access, vol. 8, pp. 166553-166574, 2020, doi: 10.1109/ACCESS.2020.3022808: A systematic review of methods for detecting and predicting spatio-temporal crime hotspots.
- A. Mansour, J. Mohammad and Y. Kravchenko, "Text Vectorization Method Based on Concept Mining Using Clustering Techniques," 2022 VI International Conference on Information Technologies in Engineering Education (Inforino), Moscow, Russian Federation, 2022, pp. 1-10, doi: 10.1109/Inforino53888.2022.9782908.

Appendix A

TERMS OF REFERENCE

6G7V0007 MSc Project

Project Topic Proposal 2023-24

Department of Computing and Mathematics Computing and Digital Technology Postgraduate Programmes Terms of Reference Coversheet	
Student name:	DIVYANSHU SAHWAL
University I.D.:	23628243
Academic supervisor:	SERGIO DAVIES
External collaborator (optional):	
Project title:	LEVERAGING NLP IN CRIME PREDICTION AND ANALYSIS
Degree title:	MSC DATASCIENCE
Project unit code:	6G7V0007_2324_9F
Credit rating:	60
Start date:	04 -06-2024
ToR date:	21-06-2024
Intended submission date:	26-09-2024
Signature and date student:	DIVYANSHU SAHWAL
Signature and date external collaborator (if involved):	

Topic: Leveraging NLP in Crime Prediction and Analysis

Introduction

Natural Language Processing (NLP) has emerged as a powerful tool for crime prediction and analysis, capable of extracting valuable information from vast amounts of unstructured text data. The purpose of this

research is to investigate and develop advanced NLP methods to improve crime prediction and analysis, thereby enhancing law enforcement and public safety.

Background

Crime forecasting and analysis involve identifying patterns, trends, and anomalies in crime data. Traditional methods are predominantly based on structured data and statistical techniques. However, the proliferation of digital textual data offers an opportunity to apply NLP to unstructured data sources. Current research has shown the effectiveness of NLP in identifying crime scenes, understanding crime narratives, and predicting future crime based on historical data. This research builds on these foundations, focusing on improving the accuracy and applicability of NLP techniques in real crime analysis scenarios.

AIM

The purpose of this research is to develop and evaluate NLP models for crime prediction and analysis. The research explores various NLP techniques for processing and analysing textual data with the aim of improving crime prediction and understanding the motivations behind criminal behavior.

OBJECTIVES

1. Existing Literature Review : Conduct a comprehensive literature review on the application of NLP to crime prediction and analysis.

2. Data Collection and Preprocessing : Collect and preprocess text data from a variety of sources, including police reports, social media, and news articles around all of the Manchester.

3. Model Development : Develop NLP models to extract meaningful information from unstructured text and predict crime trends and hotspots.

4 Evaluation: Evaluate the performance of the developed models using relevant metrics and benchmark data sets.

5 Implementation and Validation : Apply the models to real life and validate their effectiveness in predicting and analysing crime.

6. Ethical Considerations : Address privacy and ethical issues related to the use of sensitive criminal information.

METHODOLOGY

1. Literature Review: Analyse existing research and identify gaps in the application of NLP to crime prediction.

2. Data Collection: Source crime-related textual information from police reports, social media, and news archives.

3. Data Preprocessing: Clean and preprocess data using NLP techniques such as tokenization, stemming, and lemmatization.

4. Model Development: Develop and train NLP models using machine learning algorithms, focusing on classification, clustering, and prediction tasks.

5. Evaluation: Use metrics such as precision, accuracy, recall, and F1 score to evaluate model performance.

6. Implementation: Deploy the models in a test environment to assess their practical applicability.

7. Validation: Validate models with real data to ensure reliability and accuracy.

EXPECTED RESULTS

- A thorough understanding of the application of NLP to crime prediction and analysis.
- Development of reliable NLP models to accurately predict crime trends and identify hotspots.
- Knowledge of the temporal and spatial dynamics of crime.
- Practical guidelines for implementing NLP-based crime analysis tools in law enforcement.

References

1. Predicting and Monitoring Crime in Porto, Portugal Using Machine Learning, Spatial and Textual Analysis: This study focuses on integrating machine learning with spatial and textual analysis to predict crime in Porto, Portugal.

2. Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review: A comprehensive review of spatio-temporal crime prediction and hotspot detection methods.

3. Spatio-Temporal Crime Predictions by Leveraging Artificial Intelligence for Citizens Security in Smart Cities: Explores AI-based approaches to crime prediction in smart cities.

4. S. Savaş and N. Topaloğlu, "Crime Intelligence from Social Media: A Case Study," 2017 IEEE 14th International Scientific Conference on Informatics, Poprad, Slovakia, 2017, pp. 313-317, doi: 10.1109/INFORMATICS.2017.8327266: Examines the use of social media for crime intelligence, highlighting techniques for extracting crime-related information from Twitter.

5. S. P. C. W. Sandagiri, B. T. G. S. Kumara and B. Kuhaneswaran, "Detecting Crimes Related Twitter Posts using SVM based Two Stages Filtering," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), RUPNAGAR, India, 2020, pp. 506-510, doi: 10.1109/ICIIS51140.2020.9342698: Discusses a two-stage SVM filtering approach to detect crime-related posts on Twitter.

6. U. M. Butt et al., "Spatio-Temporal Crime Predictions by Leveraging Artificial Intelligence for Citizens Security in Smart Cities," in IEEE Access, vol. 9, pp. 47516-47529, 2021, doi: 10.1109/ACCESS.2021.3068306: Focuses on AI techniques for predicting spatio-temporal crime patterns in smart cities.

7. U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir and H. H. R. Sherazi, "Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review," in IEEE Access, vol. 8, pp. 166553-166574, 2020, doi: 10.1109/ACCESS.2020.3022808: A systematic review of methods for detecting and predicting spatio-temporal crime hotspots.

START HERE - Basic Information

This form must be completed for all student projects.

Before you proceed

Some activities inherently involve increased risks or approval by external regulatory bodies, so a proportional ethics review is not recommended and a full ethical review may be required.

These may include:

i. Approval from an external regulatory body (including, but not limited to: NHS (HRA), HMPPS etc.);
ii. Misleading participants;
iii. Research without the participants' consent;
iv. Clinical procedures with participants;
v. The ingestion or administration of any substance to participants by any means of delivery;
vi. The use of novel techniques, even where apparently non-invasive, whose safety may be open to question;
vii. The use of ionising radiation or exposure to radioactive materials;
viii. Engaging in, witnessing, or monitoring criminal activity;
ix. Engaging with, or accessing terrorism related materials;
x. A requirement for security clearance to access participants, data or materials;
xi. Physical or psychological risk to the participants or researcher;
xii. The project activity takes place in a country outside of the UK for which there is currently an active travel warning issued by the authorities (see info button);
xiii. Animals, animal tissue, new or existing human tissue, or biological toxins and agents;
xiv. The sharing of participant personal data with a third party, regardless of the form under which the data is presented.

If any of these activities are fundamental to your project, please contact your supervisor to determine if a full application is required.

This form must be completed for each research project which you undertake at the University. It must be approved by your supervisor (where relevant) PRIOR to the start of any data collection.

In completing this form, please consult the University's Research Ethics and Governance standards.

A1a Please confirm that you will abide by the University's Research Ethics and Governance standards in relation to this project.

- (•) Yes
- ( ) No

A1b Data Protection

The University is responsible for complying with the UK General Data Protection Regulation whenever personal data is processed. Under the Data Protection Policy, all staff and students have a responsibility to comply with the regulation in their day-to-day activities. The first step you can take to understand these responsibilities is to review the Data Protection in Reseach guidance pages and complete the University's Mandatory Data Protection Training. Student training is available through Moodle (in the 'Skills Online' section – please follow this link). To make sure your knowledge is up to date, all staff and students must complete the training every two years. If you have any issues in accessing the data protection training or have any questions about the training, please contact dataprotection@mmu.ac.uk.

Have you reviewed the Data Protection guidance pages and completed the Data Protection Training in the last two years?

- (•) Yes
- ( ) No

A2 Are you submitting this application as a learning experience, for a unit which already has ethical approval? (please confirm with your supervisor)

- (•) Yes
- ( ) No

A2.1 Approval reference (supplied by your supervisor)

A3 Student details

Title	First Name	Surname
	Divyanshu	Sahwal

Email: DIVYANSHU.SAHWAL@stu.mmu.ac.uk

A3.1 Manchester Metropolitan University ID number

23628243

A4 Supervisor

Title	First Name	Surname
Mr	SERGIO	DAVIES

Faculty: Science and Engineering

Telephone: 55147701@ad.mmu.ac.uk

Email: sergio.davies@mmu.ac.uk

A5 Which Faculty is responsible for the project?

Science and Engineering

A6 Course title

MSC DATASCIENCE

A7 Project title

Leveraging NLP in Crime Prediction and Analysis

A8 What is the proposed start date of your project?

04/06/2024

A9 When do you expect to complete your project?

25/09/2024

A10 Please describe the overall aims of your project (3-4 sentences). Research questions should also be included here.

The purpose of this research is to develop and evaluate NLP models for crime prediction and analysis. The research explores various NLP techniques for processing and analysing textual data with the aim of improving crime prediction and understanding the motivations behind criminal behavior.
Research Questoins
1. How effective are current NLP techniques in identifying crime-related information from unstructured text data?
2.What are the best practices for preprocessing crime-related textual data to improve NLP model performance?
3. Can NLP models accurately predict crime trends and identify hotspots using historical text data?
4. What are the temporal dynamics of crime data, and how can they be modeled using NLP techniques?
5.How can geospatial analysis be integrated with NLP to improve crime scene identification and hotspot detection?
6. What is the relationship between public sentiment and crime, and how can sentiment analysis be used to predict crime trends?
7.What ethical and privacy considerations arise in the use of NLP for crime prediction and analysis, and how can they be addressed?
8.How can the interpretability of NLP models be improved to ensure that their results are actionable for law enforcement?
9.What are the limitations of current NLP approaches in real-time crime prediction and analysis, and how can these be overcome?
10. What impact does the quality and completeness of data have on the accuracy of NLP-based crime prediction models?

A11 Please describe the research activity

1. Existing Literature Review : Conduct a comprehensive literature review on the application of NLP to crime prediction and analysis.
2. Data Collection and Preprocessing : Collect and preprocess text data from a variety of sources, including police reports, social media, and news articles around all of the Manchester.
3. Model Development : Develop NLP models to extract meaningful information from unstructured text and predict crime trends and hotspots.
4 Evaluation: Evaluate the performance of the developed models using relevant metrics and benchmark data sets.
5 Implementation and Validation : Apply the models to real life and validate their effectiveness in predicting and analysing crime.
6. Ethical Considerations : Address privacy and ethical issues related to the use of sensitive criminal information.

A12 Please provide details of the participants you intend to involve (please include information relating to the number involved and their demographics; the inclusion and exclusion criteria)

none

A13 Please upload your project protocol

Documents

Type	Document Name	File Name	Version Date	Version	Size
Project Protocol	6G7V0007_2324_9F_ToR_coversheet_2024	6G7V0007_2324_9F_ToR_coversheet_2024.docx	21/06/2024	1.0	17.4 KB

Project Activity

B1 Are there any Health and Safety risks to the researcher and/or participants?

- ○ Yes
- ◉ No

B2 Please select any of the following which apply to your project

- ☐ Aspects involving human participants (including, but not limited to interviews, questionnaires, images, artefacts and social media data)
- ☐ Aspects that the researcher or participants could find embarrassing or emotionally upsetting
- ☐ Aspects that include culturally sensitive issues (e.g. age, gender, ethnicity etc.)
- ☐ Aspects involving vulnerable groups (e.g. prisoners, pregnant women, children, elderly or disabled people, people experiencing mental health problems, victims of crime etc.), but does not require special approval from external bodies (NHS, security clearance, etc.)
- ☐ Project activity which will take place in a country outside of the UK
- ☑ None of the above

B2.4 Is this project being undertaken as part of a larger research study for which a Manchester Metropolitan application for ethical approval has already been granted or submitted?

- ○ Yes
- ◉ No

Data

F1 How and where will data and documentation be stored?

Data is stored at
police UK, https://data.police.uk

F2 Will you be using personal data? Personal data is anything than can be used to identify a living individual, directly or indirectly. Pseudonymised data is still personal data.

- ○ Yes
- ◉ No

Insurance

F3 Does your project involve:

- ☐ Pregnant persons as participants with procedures other than blood samples being taken from them? (see info button)
- ☐ Children aged five or under with procedures other than blood samples being taken from them? (see info button)
- ☐ Activities being undertaken by the lead investigator or any other member of the study team in a country outside of the UK as indicated in the info button? If 'Yes', please refer to the 'Travel Insurance' guidance on the info button
- ☐ Working with Hepatitis, Human T-Cell Lymphotropic Virus Type iii (HTLV iii), or Lymphadenopathy Associated Virus (LAV) or the mutants, derivatives or variations thereof or Acquired Immune Deficiency Syndrome (AIDS) or any syndrome or condition of a similar kind?
- ☐ Working with Transmissible Spongiform Encephalopathy (TSE), Creutzfeldt-Jakob Disease (CJD), variant Creutzfeldt-Jakob Disease (vCJD) or new variant Creutzfeldt-Jakob Disease (nvCJD)?
- ☐ Working in hazardous areas or high risk countries? (see info button)
- ☐ Working with hazardous substances outside of a controlled environment?
- ☐ Working with persons with a history of violence, substance abuse or a criminal record?
- ☑ None of the above

Additional Information

G1 Do you have any additional information or comments which have not been covered in this form?

- ○ Yes
- ◉ No

G2 Do you have any additional documentation which you want to upload?

- ○ Yes
- ◉ No

Signatures

H1 I confirm that all information in this application is accurate and true. I will not start this project until I have received Ethical Approval.

- ◉ I confirm

H2 Please notify your supervisor that this application is complete and ready to be submitted by clicking "Request" below. Do not begin your project until you have received confirmation from your supervisor - it is your responsibility to ensure that they do this.

Signed: This form was signed by Sergio Davies (Sergio.Davies@mmu.ac.uk) on 24/06/2024 11:45

H3 Have you been instructed by your supervisor to request a second signature for this application?

- ○ Yes
- ◉ No

H4 By signing this application you are confirming that all details included in the form have been completed accurately and truthfully. You are also confirming that you will comply with all relevant UK data protection laws, and that that research data generated by the project will be securely archived in line with requirements specified by the University, unless specific legal, contractual, ethical or regulatory requirements apply.

Signed: This form was signed by Divyanshu Sahwal (DIVYANSHU.SAHWAL@stu.mmu.ac.uk) on 21/06/2024 13:23

Appendix B

Data Cleaning and Preprocessing:

- **Regular Expressions to Clean Text Data**:

Utilized regular expressions to remove specific unwanted phrases like "On or near" from the Location column to simplify the text data.

- **NLP Stop Words Removal:**

Integrated NLTK's stop words to further clean the Location text by removing additional common but irrelevant words, enhancing the data's quality for analysis.

- **Handling Missing Values:**

Handled missing values (nan), possibly by filling them with a placeholder or removing the rows/columns affected.

- **Feature Engineering:**

Bag of Words for Textual Data, Applied the Bag of Words technique to convert the text in the Last outcome category into a sparse matrix of token counts. This method represents texts as bags (multisets) of their words, disregarding grammar and word order but keeping multiplicity.

Dimensionality Reduction and Visualization

PCA (Principal Component Analysis):

Conducted PCA to reduce the dimensionality of the sparse matrix generated from the BoW model from high-dimensional space to 2 dimensions. This step was important for visualizing and understanding the underlying structure of the data.

Clustering Using K-Means:

Used K-Means clustering to group the data based on similarities in the PCA-reduced features. This helped identify distinct clusters or groups within the outcomes, indicating similar thematic content.

Created scatter plots to visualize the PCA results and the clusters formed by K-Means. These plots help in visually assessing the spread and grouping of the data points.

Summary and Interpretation:

The process started with cleaning the textual data to ensure quality and consistency.

A Bag of Words approach was then used to transform this cleaned text into a format suitable for machine learning (i.e., numerical features based on word counts).

PCA was applied to these features for dimensionality reduction, facilitating a meaningful visual analysis.

K-Means clustering was employed to categorize these features into groups, potentially revealing patterns or similarities in the Last outcome category.

The scatter plots not only provided a visual representation of the data's distribution and clustering but also aided in further analysis and decision-making.