Final Paper for Crime Prediction Using Machine Learning

Team: *Cool Team*

Members: *Aidan Tan, Cindy Dong, Herin Seo, Joseph Choe, Allie Kim, David Bae, Benjamin Yeh, Jonah Lee*

Abstract

This study explores the application of machine learning techniques to predict crime types in Los Angeles using a comprehensive dataset spanning from 2020 to the present. Employing a K-Nearest Neighbors (KNN) algorithm, we aim to forecast crime occurrences based on various factors including location, time, and victim demographics. Our research addresses the growing need for data-driven approaches in law enforcement and public safety strategies. The model achieved an accuracy of approximately 47.24% in predicting crime types, using victim age, time, and area as key features. While this accuracy is modest, it demonstrates the potential for machine learning as a supplementary tool in crime analysis and prevention strategies. Key findings include significant correlations between specific premises and crime types, as well as variations in crime patterns across different areas of Los Angeles. The study revealed that certain locations are more prone to specific types of crimes, and that crime patterns vary significantly across different neighborhoods. Our analysis also highlighted the importance of temporal factors in crime occurrence, with distinct patterns emerging for different times of day and days of the week. We emphasize the critical importance of ethical considerations in the application of such models, particularly regarding potential biases in crime data and the risk of reinforcing stereotypes. This research contributes to the field of predictive policing by providing insights into the complex dynamics of urban crime and offering a methodological framework for future

studies. Our results suggest that while machine learning can be a valuable asset in crime prevention strategies, it must be implemented with careful consideration of its current limitations and broader societal implications.

Introduction

Crime prediction has emerged as a crucial area of study at the intersection of data science, criminology, and public safety. As urban environments become increasingly complex, traditional methods of crime analysis have proven insufficient in capturing the nuanced dynamics of criminal behavior. The advent of big data and advanced machine learning techniques offers a promising solution to this challenge, enabling researchers and law enforcement agencies to develop more sophisticated and accurate methods of forecasting criminal activities.

This study focuses on the application of machine learning to crime prediction in Los Angeles, utilizing a rich dataset of crime incidents from 2020 to the present. Our research is motivated by the potential of data-driven approaches to enhance public safety measures, inform resource allocation in law enforcement, and contribute to more effective crime prevention strategies.

The ability to predict crime has long been a goal of law enforcement agencies and criminologists. Traditional approaches have relied heavily on historical data and human expertise to identify patterns and trends. However, these methods often struggle to capture the complex, multifaceted nature of criminal behavior and its relationship to various environmental and social factors. Machine learning offers a more nuanced approach, capable of analyzing large volumes of data and identifying subtle patterns that may not be apparent to human observers.

Our study employs a supervised learning approach, specifically a K-Nearest Neighbors (KNN) algorithm, to classify crime types based on various input features. This method was chosen for its effectiveness in handling complex, multi-dimensional data and its ability to provide interpretable results. The KNN algorithm works by classifying new data points based on the majority class of their k-nearest neighbors in the feature space, making it well-suited for our crime prediction task.

The primary objective of this research is to develop a machine learning model capable of predicting the type of crime likely to occur in specific situations in Los Angeles. By analyzing patterns in crime types, area characteristics, time of occurrence, and victim profiles, we aim to create a tool that can identify high-risk situations and inform targeted prevention efforts. Specific objectives include analyzing the relationship between crime types and various factors such as location, time, and victim demographics; developing and optimizing a KNN model to predict crime types based on these factors; evaluating the accuracy and reliability of the model using various performance metrics; exploring the potential applications of the model in real-world scenarios; and addressing ethical considerations and potential biases in crime prediction models.

While our study offers significant potential benefits for public safety and law enforcement strategies, we acknowledge the ethical concerns associated with predictive policing. These include the risk of perpetuating existing biases in crime reporting and enforcement, potential misuse of predictions for discriminatory practices, and privacy concerns related to the use of personal data. Our research incorporates measures to address these risks, including explicit discussion of data limitations and potential biases, focus on using predictions for community awareness and support rather than punitive measures, and careful interpretation of results to avoid reinforcing stereotypes or stigmatizing communities.

This study on crime prediction in Los Angeles employs machine learning techniques to analyze a dataset of crime incidents from 2020 to the present, revealing important insights into urban crime patterns. Utilizing a K-Nearest Neighbors (KNN) algorithm, our research achieved a predictive accuracy of 47.24% for classifying crime types. Key findings indicate a general increase in crime frequency throughout the day, with theft peaking at midday, and highlight specific high-crime areas such as Pacific and Central. The analysis also identifies temporal "hotspots" where certain crimes are more likely to occur, providing valuable information for law enforcement resource allocation.

While our model demonstrates potential in enhancing public safety measures, the accuracy rate suggests significant room for improvement in predictive capabilities. The challenges faced in accurately predicting less common crime categories underscore the complexities of using machine learning in this context. As we explore the methodology, results, and implications of our findings in the following sections, readers will gain a deeper understanding of both the advantages and limitations of predictive policing. In the following sections, we provide a detailed description of our dataset, including its sources, contents, and limitations. We then explain our analytical approach, including data preprocessing, feature engineering, and model development. The results section presents our findings, including model performance metrics and key insights derived from the analysis.

Data Wrangling/EDA

Our dataset was obtained from Data.gov and reflects incidents of crime in the City of Los Angeles from 2020 to the present. It includes columns such as 'Premis Desc' (premises description), 'Crm Cd Desc' (crime type description), and 'AREA NAME' (the geographical area

where the crime occurred), along with other details like dates, times, and coordinates. This dataset is highly relevant for analyzing crime patterns based on type, location, and other contextual factors. It can provide valuable insights into which types of crimes are more prevalent in specific areas or premises, identifying potential hotspots and trends. These insights can inform public safety measures, law enforcement strategies, and the allocation of community resources.

However, the dataset presents certain challenges that must be addressed. Managing the large volume of data efficiently, ensuring data cleanliness, and handling missing or ambiguous entries—such as unclear premises and crime descriptions—are key tasks. Furthermore, effectively visualizing the data to highlight trends and patterns without oversimplifying complex relationships remains a significant hurdle. Resolving these challenges is critical to ensuring the dataset can be leveraged effectively for our analysis.

Methods

Our team aims to answer the following prediction question: Given various factors such as location, time, and victim demographics, can we accurately predict the type of crime most likely to occur in a specific situation? By analyzing patterns in crime types, area characteristics, time of occurrence, and victim profiles, we seek to develop a predictive model that can identify high-risk situations. This model could be utilized by law enforcement agencies, community safety organizations, or even integrated into public safety apps to help individuals make informed decisions about their personal safety. For example, the model could alert users to potentially dangerous areas at specific times or provide real-time risk assessments based on current location and conditions. Additionally, we will explore correlations between crime types and various factors to uncover insights that can enhance public awareness and inform targeted crime

prevention strategies. Ultimately, our goal is to create a tool that empowers both individuals and communities to proactively address safety concerns and reduce the likelihood of victimization.

The dataset we are working with consists of reported crime incidents in Los Angeles from 2020 to the present. Each observation represents a unique crime report, capturing details such as the time of occurrence, type of crime, location, and victim demographics. However, we acknowledge that inaccuracies may exist due to manual transcription of the original reports. Moreover, missing or anonymized data, such as locations represented as coordinates (0°, 0°), presents challenges that must be addressed during data preprocessing.

To address the problem, we will use a supervised learning approach, specifically approaching it as a classification problem. Our primary objective is to build a model that can predict crime types based on the given circumstances. This model is intended to provide actionable insights to residents, researchers, or public safety officials by profiling potential crime risks under specific conditions. By focusing on patterns rather than causation, we aim to use this model as a tool for community awareness rather than punitive measures.

We will use the K-Nearest Neighbors (KNN) algorithm for this task. KNN is well-suited for classification problems as it predicts the class of an observation based on the majority class of its nearest neighbors in the feature space. Key steps in using KNN include encoding categorical variables, scaling numerical features to ensure meaningful distance-based calculations, and experimenting with different values of k and distance metrics to optimize model performance. This iterative process will help us refine the model to improve accuracy and reliability.

For data cleaning, we will handle missing numeric values by imputing the mean or median, while categorical variables will be imputed with the mode or labeled as "Unknown." Categorical features such as victim sex, area, and premise description will be one-hot encoded to

prepare them for the model. We will select features such as time of occurrence, area, premise description, victim sex, and victim age, which will be binned into meaningful groups if necessary. If dimensionality reduction becomes necessary to address redundancy in highly correlated variables, we will apply Principal Component Analysis (PCA) to reduce the feature space.

To evaluate the model, we will use an accuracy metric with a success threshold of 80%. Additionally, evaluation metrics such as precision, recall, and F1 score will provide a comprehensive view of the model's performance. We will create confusion matrices to visualize the relationship between predicted and actual crime types. To demonstrate the model's real-world application, we will analyze specific scenarios, such as predicting crime risks at noon in Hollywood for a 30-year-old victim at a residence.

However, we recognize several risks and limitations inherent to this project. Crime prediction is inherently complex due to the socio-economic, cultural, and systemic factors that influence crime patterns but are not captured in the dataset. To mitigate this, we will explicitly communicate the limitations of our model and avoid making causational inferences. Bias in crime data, which may disproportionately represent certain demographics or areas, is another critical consideration. To address this, we will acknowledge and discuss these biases to prevent conclusions that perpetuate stereotypes. Missing or anonymized data fields may also impact prediction accuracy, so we will carefully impute missing values and assess their impact. Finally, we recognize the potential misuse of the model, such as unintentional contributions to racial profiling or stigmatization. To prevent this, we will ensure that the model is focused solely on raising awareness or supporting community-based initiatives.

The results of this project will be communicated effectively through visualizations such as heatmaps and bar charts, highlighting crime risks for specific premises or areas. Confusion matrices will illustrate the model's predictions against actual crime types, while a detailed written analysis will contextualize the results and their implications. Importantly, we will explicitly acknowledge the limitations and risks to ensure that the findings are interpreted ethically and responsibly.

Results

Our study addresses the following prediction question: Given various patterns in crime types, area names, and times of occurrence, how can we predict high-risk situations to help individuals avoid dangerous areas and make informed safety decisions? In order to further study and answer this question we created a line graph, heatmap, KDE density plot, and a confusion matrix covering various variables which include TIME OCC (time at which the crime took place), Crm Cd Desc (crime type), Area Name (location of crime), Vict Age (victim age), Premis Desc (location description).

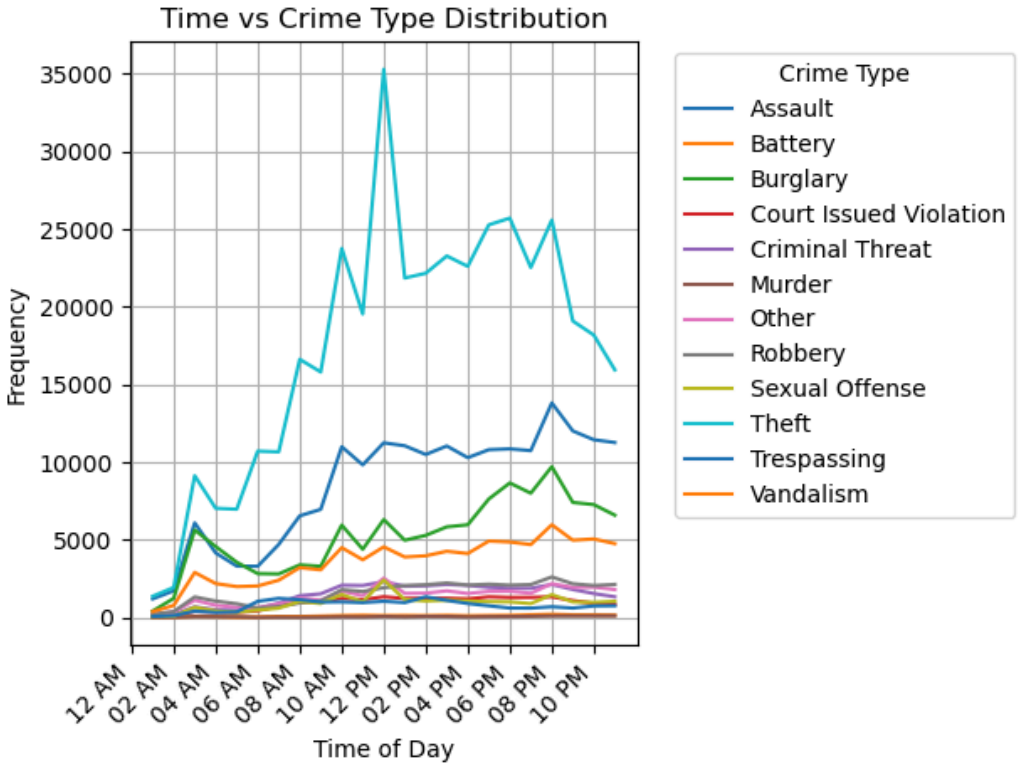

Figure 1: Time (TIME OCC) vs. Crime Type (Crm Cd Desc)

In Figure 1, the graph depicts a line graph analyzing crime type and how frequent that crime type is at a specific time of the day. In this graph, our algorithm uses the cleaned data with crime type sorted into 12 categories. The most prominent finding is that theft is the most frequently occurring crime, followed by trespassing, then burglary as the top 3. We can also see that for the top crime, theft, it most commonly occurs at 12PM, mid-day. There is also an overall trend of crime generally increasing in frequency as the day progresses from morning to night (12AM to 12PM).

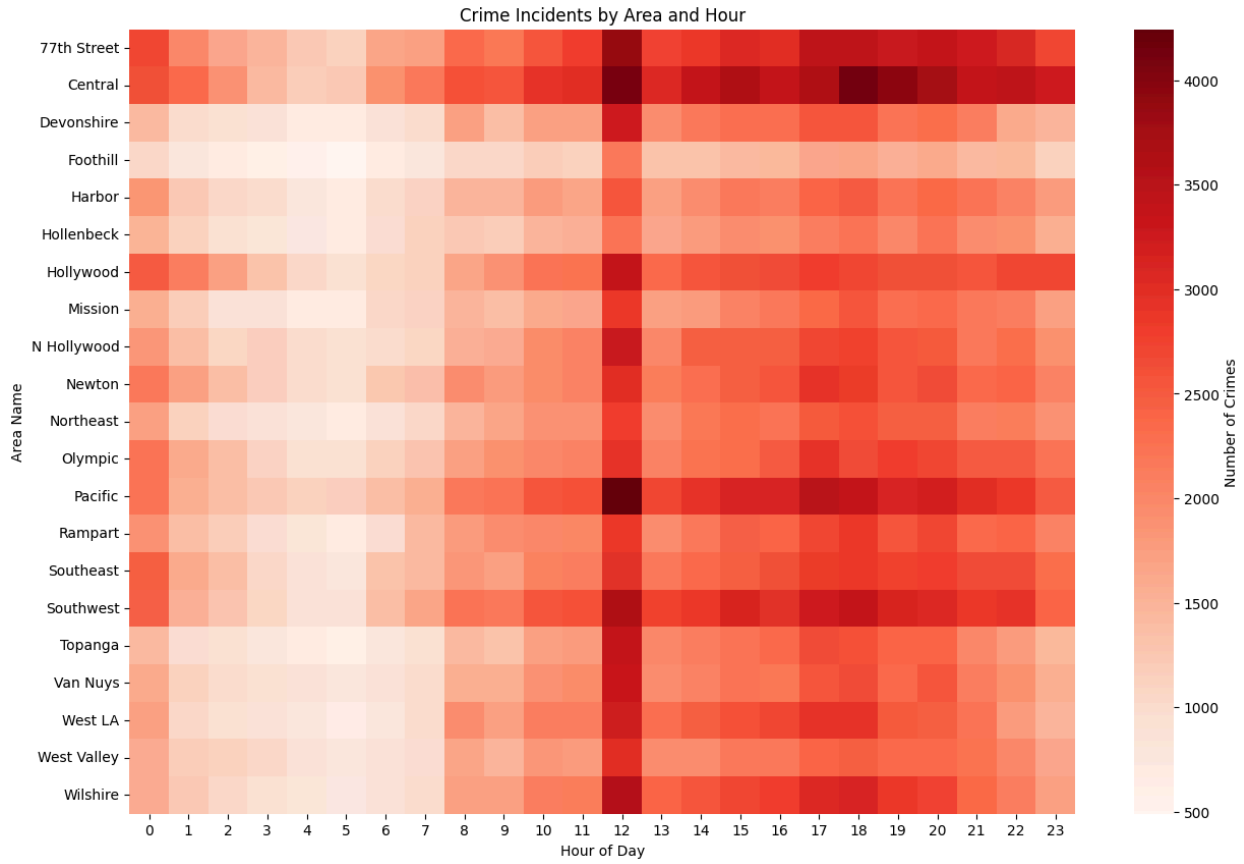

Figure 2: Heatmap of Crime Incidents by Area and Hour

The graph above, Figure 2, presents a heatmap visualizing the frequency of crime incidents across different areas of Los Angeles and hours of the day. This visualization is based on our cleaned and processed dataset, which includes information on crime locations and times of occurrence. The heatmap uses a color gradient from light to dark red to represent the number of crimes, with darker shades indicating a higher frequency of incidents. The x-axis represents the 24 hours of the day, while the y-axis lists the various areas within Los Angeles.

One of the most striking observations from this heatmap is the variation in crime patterns across different areas and times. Some areas appear to have consistently higher crime rates throughout the day like Pacific and Central, as indicated by darker red blocks across their rows. Conversely, other areas show lower overall crime rates, represented by lighter shades. There's a

noticeable trend in the temporal distribution of crimes. Many areas seem to experience an increase in criminal activity during noon through late afternoon and into the evening hours, as evidenced by the darker shades in the right half of the heatmap. This could suggest that certain types of crimes are more likely to occur during these times. Additionally, we can observe some 'hotspots' where certain areas at specific times show particularly high crime rates (Pacific at 12 pm) represented by the darkest red squares. These hotspots could be valuable for law enforcement to target their resources more effectively.

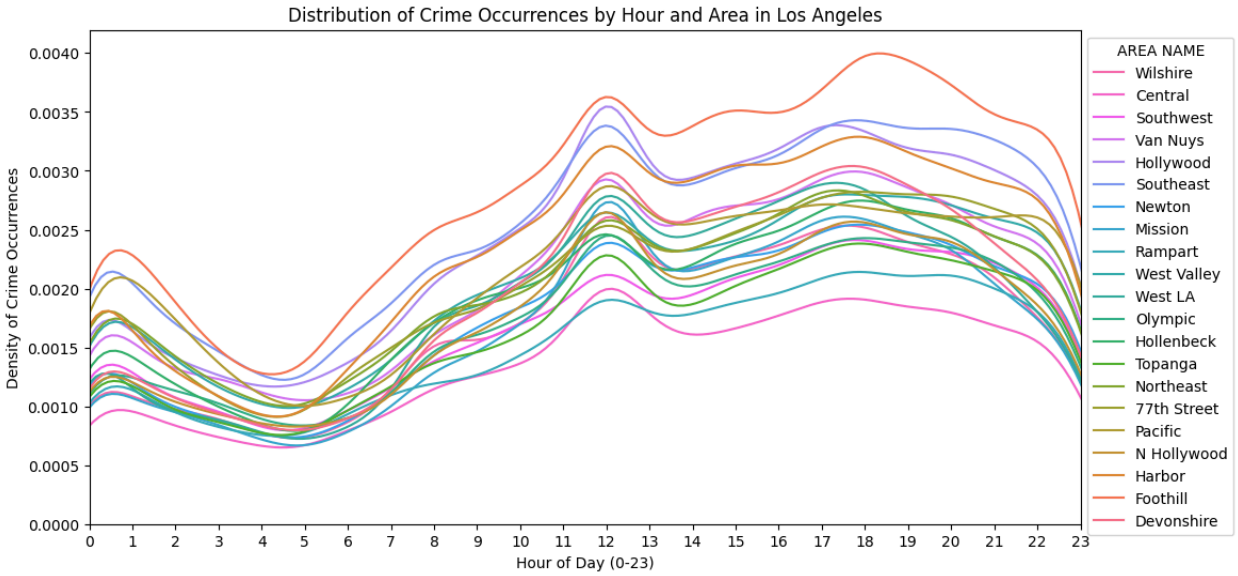

Figure 3: KDE Plot of Crime Incidents by Density of Crime Occurrences and Hour

Figure 3 presents a kernel density estimation (KDE) plot that illustrates the distribution of crime occurrences by hour and density of crime occurrences in Los Angeles, based on a processed dataset of crime incidents. The x-axis represents the 24 hours of the day, while the y-axis shows the density of crime occurrences, with different colored lines indicating various areas. Key observations reveal a general decrease in crime density from 1 AM to 5 AM, followed by a significant increase from 5 AM to 11 AM, culminating in a sharp jump at 11 AM. After this peak, there is a noticeable decrease in density from 12 PM to 1 PM; however, crime rates remain

relatively high throughout the afternoon and early evening until around 10 PM to 11 PM, where a visible drop occurs. This pattern highlights distinct temporal trends in criminal activity, with certain hours exhibiting higher densities that could inform targeted law enforcement strategies. Overall, this KDE plot serves as an insightful tool for understanding the complex dynamics of crime occurrences across different times of day in Los Angeles.

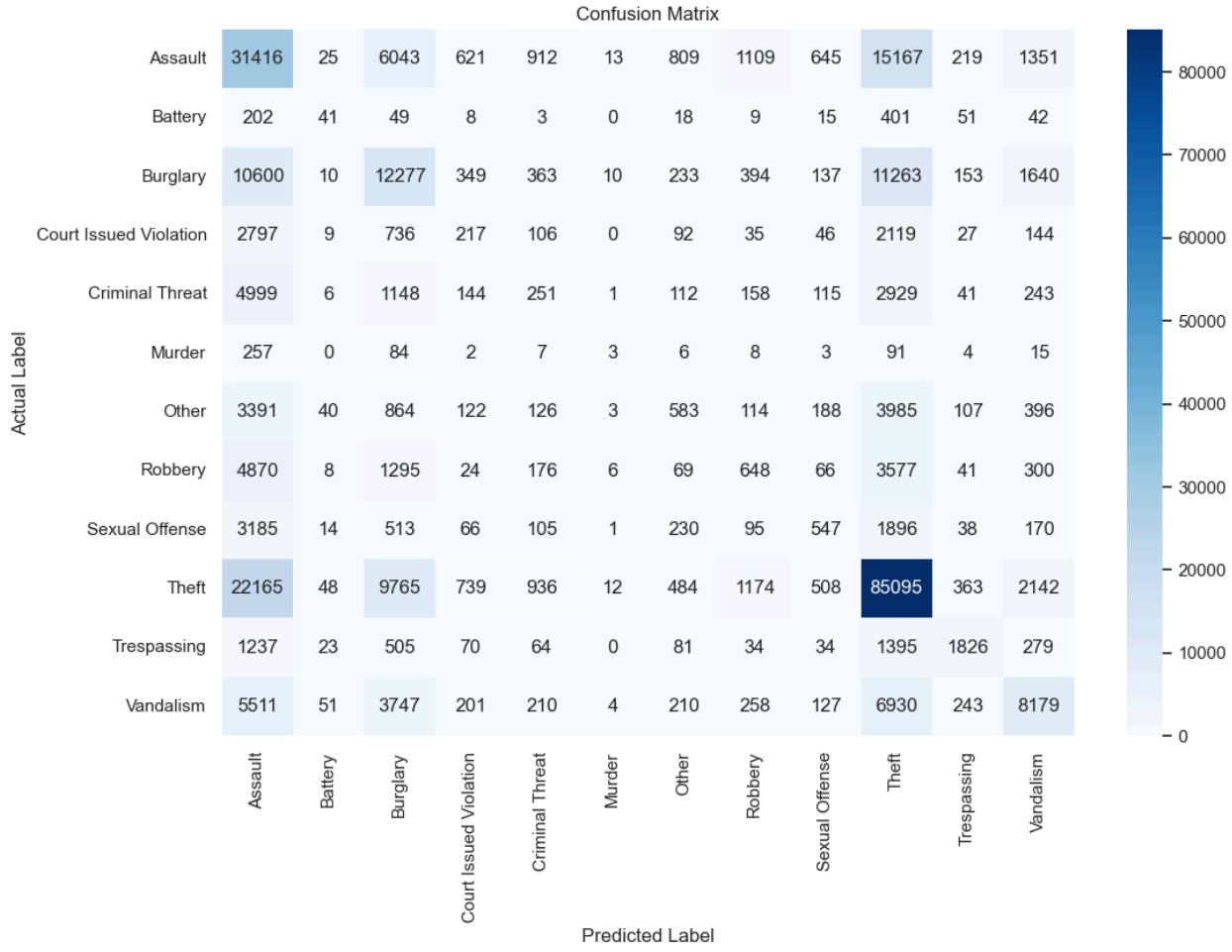

Figure 4: KNN Model Confusion Matrix for Crime Category Prediction

In figure 4, the model performs well in predicting certain types of crimes, with the highest number of true positive predictions for Theft (Figure 4). Assault and Burglary also have relatively high accuracy, suggesting that these crimes were predicted fairly well. However, the model struggles with less common or more nuanced categories such as Murder, Criminal

Threats, and Sexual Offenses, which shows a low number of true positives, indicating difficulty in accurately classifying these instances. Common misclassifications include Burglary being predicted as Theft and Robbery being misclassified as Other, suggesting that these crime types share similarities in their features or descriptions, which may lead to confusion. Overall, while the model effectively identifies some crime categories, it struggles in differentiating between categories with overlapping characteristics and enhancing predictions for less frequently occurring crimes.

Conclusion

This study investigated the potential of using machine learning, specifically the K-Nearest Neighbors (KNN) algorithm, to predict crime types in Los Angeles based on a dataset of crime incidents from 2020 to the present. Our findings highlight significant temporal and spatial patterns in crime occurrences, with certain crime types, such as theft, being more prevalent at specific times of the day. The analysis revealed that crime tends to increase as the day progresses, with theft peaking at noon and other crimes also following distinct time-based trends. The heatmaps and kernel density estimation (KDE) plots further revealed that certain areas of Los Angeles, such as Pacific and Central, experience consistently higher crime rates across different hours of the day, indicating hotspots that can inform law enforcement resource allocation.

The KNN model achieved a modest accuracy of 47.24%, which demonstrates the viability of machine learning for crime prediction. The model performed reasonably well in predicting common crimes like theft, assault, and burglary, but struggled with rarer crime categories, such as murder, sexual offenses, and criminal threats. These less common crime types

were often misclassified or poorly predicted, suggesting that the model's performance could be improved by incorporating more cleaned data or using more advanced algorithms. The challenges in predicting these crimes also point to the complexity of crime patterns, which may require more nuanced models to better differentiate between overlapping crime categories.

One of the key challenges encountered in this study was the quality and completeness of the dataset. Missing or incomplete data, inaccuracies in crime descriptions, and inconsistencies in category names or location reporting all affected the accuracy of the predictions. We also observed that the dataset may be biased, potentially overrepresenting certain areas or demographic groups, which could lead to skewed predictions. Future research could address these limitations by improving data cleaning techniques, applying more sophisticated imputation methods, and considering ways to correct for any biases in the dataset.

Another limitation of this study is the ethical concerns associated with predictive policing. While machine learning models like the one developed in this study have the potential to assist in crime prevention and resource allocation, there is a risk that such models could perpetuate existing biases in crime reporting and enforcement, particularly when it comes to racial or socioeconomic disparities. To mitigate these risks, it is crucial for future work to incorporate fairness-aware machine learning techniques and to ensure that the model is used in ways that support community safety rather than reinforcing stereotypes or discriminatory practices.

Looking forward, there are several areas for improvement and future research. First, more sophisticated machine learning algorithms, such as ensemble methods or deep learning models, could be explored to improve the accuracy and robustness of crime predictions, especially for less frequent crime categories. Another potential area for future work is the integration of

real-time data to enable dynamic crime prediction, providing law enforcement with up-to-date insights into emerging crime trends and hotspots.

In conclusion, this study provides a valuable contribution to the growing body of research on predictive policing and crime analysis. While the results demonstrate the promise of machine learning in identifying crime patterns and informing law enforcement strategies, they also underscore the challenges and ethical considerations that must be addressed. By refining models, improving data quality, and carefully considering the social implications of predictive policing, future research can continue to develop more effective, equitable, and responsible tools for enhancing public safety.

References/ Bibliography

"City of Los Angeles - Crime Data from 2020 to Present." *Catalog*, Publisher data.lacity.org, 29 Nov. 2024, catalog.data.gov/dataset/crime-data-from-2020-to-present?fbclid=IwZXh0bgNhZW0CMTEAAR3zwULhon7uKIrjrk5cmj3fbks-cH6d7hEQV-mEsCTqTkWLIQOicIBSAss_aem_eVtZaueW3VLnZ2igZs9SvQ.