ITM University Gwalior

PBL Report File

On

Crime Status Prediction

(MCA-304)

Title:

Enhancing Crime Status Prediction in Los Angeles

Submitted To:

Dr. Sanjay Jain
Professor, Dept. Of CSA

Submitted By:

Prashant Singh
MCAN1CA2201

Table of Contents

1. Abstract

This study focuses on refining crime status prediction through an ensemble methodology applied to extensive datasets obtained from Catalog.Data.Gov, specifically targeting Los Angeles crime incidents since 2020. The research methodology comprises meticulous data collection, rigorous preprocessing, exploratory data analysis, model selection, and comprehensive model evaluation.

Initial challenges included data inaccuracies and privacy-preserving measures in location data, necessitating thorough cleaning and transformation processes. Exploratory data analysis uncovered crucial insights, including the 'Status' attribute's limited correlation, crime code distributions, area-wise crime counts, and temporal patterns.

Addressing class imbalance within 'Status,' the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset. Model evaluation highlighted the superiority of Random Forest models employing 10 and 20 Decision Trees, alongside KNN, demonstrating consistent high accuracy, balanced precision-recall trade-offs, and notable F1 scores in crime status prediction.

2. Introduction

In the realm of crime data analysis, the synergy of machine learning (ML) models has emerged as a transformative force, enriching our understanding of criminal activities and their underlying patterns. Within the United States, the persistent challenges posed by crime imperil the safety and security of its vast population, thereby emphasizing the urgent need to employ sophisticated analytical methodologies. While the national crime database serves as a foundational repository, this study posits that there exists an untapped reservoir of potential, awaiting exploration through the utilization of a novel ensemble machine learning model. This pioneering model endeavors not only to predict but also to comprehensively understand the ultimate outcomes of reported crimes. Situated within the context of the National Crime Victimization Survey (NCVS) in Los Angeles, California, this approach represents a pioneering venture into nuanced and precise crime prediction, aiming to redefine the landscape of crime resolution.

The imperative to integrate predictive modeling into the analytical landscape stems from the inherent complexity ingrained within crime data. It is this multifaceted nature that necessitates the adoption of an ensemble approach, one that amalgamates the strengths of diverse machine learning algorithms. By embracing this methodology, the study seeks to refine law enforcement strategies, optimizing their efficacy and laying the groundwork for a proactive and adaptable response to the ever-evolving panorama of criminal incidents.

Over the years, the field of crime data analysis has witnessed a surge in innovative methodologies, primarily driven by the adoption of machine learning (ML) algorithms. A seminal work by Abba Babakura et al. (2014) underscored the significance of data pre-processing and feature selection in refining models for crime prediction. Notably, Naïve Bayes emerged as a superior predictor over Back Propagation, displaying exceptional prowess in forecasting crime categories across diverse U.S. states [1]. Furthermore, the work by Lawrence McClendon and Natarajan Meghanathan (2015) spotlighted the pivotal role of machine learning in discerning intricate patterns in crime occurrence. Through the application of ML algorithms via Weka, their study facilitated strategic resource allocation, empowering law enforcement with invaluable insights [2].

The realm of spatial analysis within criminology, as elucidated by Cui-cui Sun et al. (2014), revealed subtle nuances in the spatial distribution of crimes and the associated influencing factors [3]. Similarly, the study conducted by Nitin Nandkumar Sakhare and Swati Atul Joshi (2015) harnessed ML techniques to scrutinize crime data from Vancouver. Their work recognized the untapped potential to bolster predictive accuracy, thereby underscoring the transformative capabilities of ML [4].Shaobing Wu et al. (2020), Miquel Vaquero Barnadas (2016), and Kanimozhi N et al. (2021) each contributed groundbreaking studies that demonstrated remarkable achievements in crime prediction leveraging ML algorithms. These studies underscored the transformative potential of ML techniques in attaining remarkably high prediction accuracy. Their work unlocked invaluable insights into crime-prone regions, optimal resource allocation strategies, and even real-time predictions regarding the occurrence and types of crimes [5, 6, 7].

3. Literature Review

Paper	*Authors*	*Journal*	*Technology Used*
Improved Method of Classification Algorithms for Crime Prediction [1]	Abba Babakura, Md Nasir Sulaiman, Mahmud A. Yusuf	ISBAST, IEEE (2014)	Naïve Bayes, Back Propagation
Using Machine Learning Algorithms to Analyze Crime Data [2]	Lawrence McClendon and Natarajan Meghanathan	MLAIJ (2015)	Weka, Spatial and Temporal Analysis
Detecting Crime Types Using Classification Algorithms [3]	Cui-cui Sun, Chun-long Yao, Xu Li, Kejun Lee	Journal of Digital Information Management (2014)	Spatial autocorrelation, Spatial interaction models, Spatial choice models, Analysis of mobility triads
Classification of Criminal Data using J48 Algorithm [4]	Nitin Nandkumar Sakhare, Swati Atul Joshi	IFRSA International Journal of Data Warehousing & Mining (2015)	Machine Learning, K-nearest neighbour (KNN), Boosted decision tree
Crime Prediction Using Data Mining and Machine Learning [5]	Shaobing Wu, Changmei Wang, Haoshun Cao, and Xueming Jia	Springer Nature Switzerland AG (2020)	Data mining, Machine learning, Random Forest, Neural network, Bayesian network
MACHINE LEARNING APPLIED TO CRIME PREDICTION [6]	Miquel Vaquero Barnadas	Telecom BCN (27-09-2016)	Data mining, Naive Bayes classifier, Unstructured database (Mongo DB), Named Entity Recognition (NER), Coreference Resolution
CRIME TYPE AND OCCURRENCE PREDICTION USING MACHINE LEARNING ALGORITHM [7]	Kanimozhi N, Keerthana N V, Pavithra G S, Ranjitha G, Yuvarani S	ICAIS, IEEE (2021)	Machine learning, Naïve Bayes classification

List Of Reviewed Papers

The research conducted by Abba Babakura, Md Nasir Sulaiman, and Mahmud A. Yusuf, titled "Improved Method of Classification Algorithms for Crime Prediction," represents a pivotal stride in the domain of crime prediction methodologies. Their study, published in ISBAST by IEEE in 2014, delves into the intricate realm of predicting crime categories across various states within the United States. Recognizing the multifaceted nature of crime prediction, the authors shed light on the challenges inherent in this field, stemming from the intricate interplay of socioeconomic factors, demographics, and law enforcement practices, all influencing the complex phenomenon of crime. Central to their work is the introduction of a novel approach leveraging classification algorithms, primarily Naïve Bayes and Back Propagation, to predict crime categories based on an array of extracted features obtained from crime data. These features encompass an assortment of socioeconomic indicators, demographic information, and law enforcement data. The study benchmarks their proposed method against two other approaches: a simple baseline method and a method employing a support vector machine (SVM) classifier. Remarkably, the authors unveil that their proposed methodology outperforms both the baseline and SVM-based methods, signifying its efficacy and superiority in predicting crime categories. The document emphasizes the pivotal role of data preprocessing and feature selection in refining the accuracy of crime prediction models. The authors articulate the significance of cleansing and curating the dataset by eliminating anomalies and selecting pertinent features before initiating the training process. Their argument highlights the criticality of these preparatory steps, asserting that the accuracy and efficiency of crime prediction models hinge significantly upon the quality and relevance of the input features. [1]

In the realm of crime analysis and visualization, the comprehensive review authored by Lawrence McClendon and Natarajan Meghanathan, presented in MLAIJ in 2015, stands as a significant milestone. This document meticulously explores the intricate landscape of crime pattern analysis through the lens of spatial and temporal data, spotlighting a case study centered on Maryland State, USA. Their study constitutes an in-depth examination of crime data, aiming to unravel distinct patterns within this complex landscape. To facilitate a comprehensive analysis, the authors categorize related works based on their focus on spatial or spatial-temporal aspects, setting the stage for a meticulous exploration of crime patterns. The core of their work lies in the discernment of crucial patterns unearthed through visualization methodologies. The authors deploy a visualization system to uncover key insights within Maryland State, shedding light on critical aspects such as cities exhibiting the highest frequency of crimes, the temporal dimensions of crime occurrence, and the most prevalent types of crimes within specific temporal contexts. Notably, their findings underscore that cities like Baltimore, Prince George's County, and Montgomery County serve as focal points for frequent criminal activities. Furthermore, their insights reveal distinct temporal patterns, highlighting that violent crimes tend to surge during weekend nights, whereas property crimes peak on weekdays during daylight hours. Categorically, robbery and aggravated assault emerge as the predominant types in the realm of violent crimes, whereas theft and burglary prevail within property crime categories. [2]

The scholarly work authored by Cui-cui Sun, Chun-long Yao, Xu Li, and Kejun Lee, outlined in the Journal of Digital Information Management in 2014, constitutes a seminal contribution to the domain of crime analysis through spatial statistical methodologies. This document serves as a comprehensive guide, meticulously exploring the realm of statistical analysis concerning spatial crime data. At its core, this scholarly endeavor delves into an extensive array of methodologies aimed at modeling spatial crime data, encompassing descriptive spatial statistics, visualization techniques, and spatially informed regression models. The document not only elucidates the relevance of spatial analysis in criminology but also sheds light on the methodologies used to study the distribution of crime across spatial dimensions and the movement patterns of offenders. An integral aspect discussed within the document is the profound relevance of spatial analysis for criminology. Emphasizing the geographical referencing of crime data, the authors highlight the significance of attributes embedded within these datasets, allowing researchers to discern the spatial arrangement of crime events and the underlying patterns in criminal behavior. Moreover, the document meticulously unpacks various types of spatial crime data, including information pertaining to crime event locations, offender and victim characteristics, and attributes related to crime targets, offering a holistic perspective on the dimensions of spatial crime data and their sampling techniques. In its exploration of spatial structure specification, the document introduces the pivotal concept of spatial autocorrelation—a measure quantifying the clustering of crime events in space. This sets the stage for an extensive review of spatially informed regression models, empowering researchers to model the intricate relationship between crime occurrences and associated factors while accounting for spatial autocorrelation. Furthermore, the document encapsulates the intricate realm of analyzing movement patterns within crime data. It delves into the length of the journey-to-crime and explores methodologies such as spatial interaction models, spatial choice models, and the analysis of mobility triads. These sophisticated analyses underscore the depth of understanding achievable through spatial statistical methodologies in the realm of crime and criminal justice. By highlighting the practical applications of these methodologies, the document underscores their relevance and potential in shaping the landscape of crime analysis and criminal justice practices. Overall, this scholarly endeavor serves as a pivotal resource, unveiling the profound potential of spatial statistical methodologies in deciphering the intricacies of crime patterns and offender behavior. [3]

The study authored by Nitin Nandkumar Sakhare and Swati Atul Joshi, published in the IFRSA International Journal of Data Warehousing & Mining in 2015, presents an insightful investigation into leveraging machine learning algorithms for crime prediction in Vancouver, Canada. Focused on the analysis of Vancouver's crime data spanning a significant 15-year period, the authors adeptly employed two distinct classification algorithms: K-nearest neighbor (KNN) and boosted decision tree. Their study's essence lay in dissecting and processing the dataset via two different approaches to discern predictive patterns. The first approach involved allocating unique numerical identifiers to individual neighborhoods and crime categories, while the second approach employed binary representations for neighborhood and day-of-the-week variables. The authors scrutinized the predictive accuracy of these models, revealing that the achieved crime prediction accuracy ranged between 39% and 44%. This critical insight, while showcasing the potential utility of machine learning in predicting crime trends, also underscored the need for further research and improvements to augment predictive accuracy. The authors candidly delineated the limitations inherent in their study, notably emphasizing the use of a singular dataset and a constrained set of features. They judiciously advocated for future research endeavors to explore the broader landscape by employing diverse datasets, incorporating varied features, and extending the application of machine learning to predict crime across multiple cities. [4]

The research by Shaobing Wu, Changmei Wang, Haoshun Cao, and Xueming Jia, published in Springer Nature Switzerland AG in 2020, delves into the realm of crime prediction through the adept utilization of data mining and machine learning techniques within YD County, China. Employing a dataset spanning from September 1, 2012, to July 21, 2015, the authors meticulously trained three distinct machine learning algorithms – random forest, neural network, and Bayesian network – to discern patterns in crime occurrences. Their insightful analysis revealed that the random forest algorithm emerged as the most effective, boasting an impressive accuracy rate of 90%. The study identified several pivotal factors that served as strong predictors of crime in YD County. These factors encompassed elements such as population size, demographic distribution by age, prevalence of violent crimes, drug-related offenses, property crimes, and the occurrence of crimes involving individuals with specific criminal records. These discerning factors were pivotal in developing precise crime prediction models. Furthermore, the authors advocated for the integration of temporal and spatial dimensions into the predictive models, emphasizing the significance of accounting for geographic areas and temporal periods exhibiting higher crime rates. This nuanced understanding of temporal and spatial crime patterns enhanced the predictive accuracy of their models, enabling a more comprehensive grasp of the intricate nature of criminal activities. The findings from this study offer a substantial contribution to the domain of crime prediction. By showcasing the potential of data mining and machine learning methodologies in crafting accurate predictive models, this research lays a robust foundation for leveraging these techniques to fortify law enforcement strategies. The implications of this work extend to aiding law enforcement agencies in resource allocation and devising targeted crime prevention strategies, thereby enhancing the overall efficacy of crime prevention endeavors.. [5]

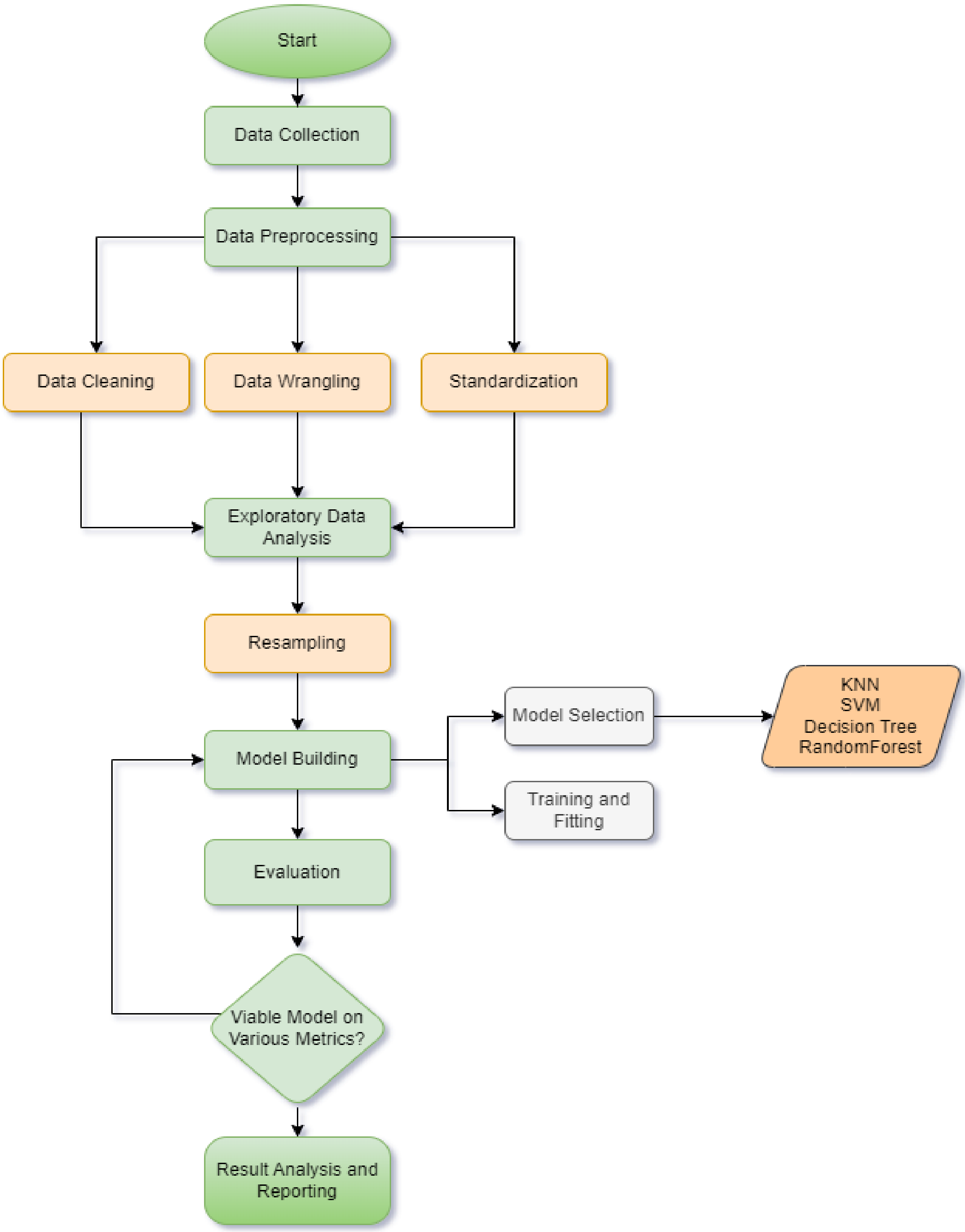
Miquel Vaquero Barnadas's work, presented at Telecom BCN in 2016, introduces an innovative system aimed at crime analysis and prediction by harnessing the power of data mining. This comprehensive system integrates diverse data sources, including crime records, news articles, and social media posts, to construct a predictive model. At its core, the system employs a Naive Bayes classifier, showcasing remarkable efficacy by achieving an accuracy rate exceeding 80% in crime prediction. The document underscores the challenges intrinsic to crime analysis and prediction, emphasizing the volume, incompleteness, and inconsistency of data as formidable hurdles. Despite these challenges, the authors highlight the system's adaptability, demonstrating its ability to handle incomplete and inconsistent datasets while emphasizing the pivotal role of a robust training set in determining predictive accuracy. One of the system's notable capabilities lies in its proficiency in predicting regions susceptible to high crime probabilities and visually presenting crime-prone areas. This insightful information holds tremendous value for law enforcement agencies, aiding in the strategic allocation of resources and the formulation of targeted crime prevention strategies. The work presents a holistic overview of crime analysis, encompassing data mining methodologies to collect and analyze information from varied sources. The adoption of a Naive Bayes classifier serves as a cornerstone for categorizing crime data into distinct types. Moreover, the utilization of an unstructured database, such as Mongo DB, along with advanced techniques like Named Entity Recognition (NER) and Coreference Resolution, enhances the system's efficiency in extracting relevant entities from crime articles. [6]

The paper titled "Crime Type and Occurrence Prediction Using Machine Learning Algorithm" presented at ICAIS, IEEE, introduces an innovative machine learning algorithm designed to predict crime type and occurrence by leveraging temporal and spatial data. Employing the Naïve Bayes classification method, the algorithm is trained using crime data sourced from Denver, Colorado, achieving an impressive accuracy rate of 93.07%. Distinguishing itself from prior models, the algorithm's unique strength lies in its ability to handle both nominal and real-valued attributes. This versatility empowers the algorithm to predict not only categorical crime types, such as robbery or assault, but also continuous crime occurrences within specific areas. Notably, the authors emphasize the algorithm's suitability for real-time predictions, enabling proactive anticipation of likely future crime types.The algorithm's workflow involves transforming temporal and spatial data into a feature set, which serves as the basis for training the Naïve Bayes classifier. Leveraging the classifier's probabilistic nature and the assumption of feature independence simplifies the training process and accelerates prediction speed. During evaluation using the Denver crime dataset, the algorithm demonstrated a remarkable 93.07% accuracy in predicting crime types. This high accuracy rate underscores the algorithm's proficiency in correctly identifying crime types in the majority of cases. [7]

4. Proposed work

Crime prediction presents a formidable challenge owing to the intricate nature of criminal activities influenced by a myriad of factors. Current models predominantly rely on individual classification algorithms, potentially limiting their efficacy in capturing the diverse and complex patterns inherent in crime data. The central objective of this research is to develop an ensemble model, strategically amalgamating the strengths of various classification algorithms, thereby enhancing the precision of crime status predictions. This ensemble approach is envisaged to mitigate the shortcomings of individual models, providing a holistic and dependable prediction framework.

Methodology:



a. Data Collection

In this project, we employ crime incident data extracted from the Catalog.Data.Gov site, focusing on incidents in the City of Los Angeles dating back to 2020. The dataset, sourced from the Gov Crime Data from Los Angeles, CA, USA, encompasses 834.320 observations with 28 attributes, forming the basis for developing an ensemble crime status prediction model. However, this dataset presents challenges, including inaccuracies transcribed from original paper reports and privacy-preserving measures in location data. [8]

The primary dataset, named "Crime_Data_from_2020_to_Present.csv," includes attributes such as incident number (DR_NO), report date (Date Rptd), occurrence date (DATE OCC), time of occurrence (TIME OCC), area details (AREA and AREA NAME), crime details (Crm Cd and Crm Cd Desc), victim information (Vict Age, Vict Sex, Vict Descent), and location coordinates (LAT and LON), among others.

CSV: Crime_Data_from_2020_to_Present.csv

Attribute	Dtype
DR_NO	int64
Date Rptd	object
DATE OCC	object
TIME OCC	int64
AREA	int64
AREA NAME	object
Rpt Dist No	int64
Part 1-2	int64
Crm Cd	int64
Crm Cd Desc	object
Mocodes	object
Vict Age	int64
Vict Sex	object
Vict Descent	object
Premis Cd	float64
Premis Desc	object
Weapon Used Cd	float64
Weapon Desc	object
Status	object
Status Desc	object
Crm Cd 1	float64
Crm Cd 2	float64
Crm Cd 3	float64
Crm Cd 4	float64
LOCATION	object
Cross Street	object
LAT	float64
LON	float64

b. Data Pre-processing

Pre-processing, encompassing data cleaning, categorical encoding, and standardization, is crucial when working with datasets, especially in machine learning tasks like crime status prediction using the NCVS dataset from LA City. Data cleaning is essential to handle missing values, outliers, and inconsistencies that could mislead the model's learning process, ensuring the integrity and accuracy of the information. Categorical encoding transforms categorical variables into numerical equivalents, enabling algorithms to interpret and learn from them effectively. Standardization normalizes the range of numerical features, preventing any particular feature from dominating the model due to its scale, thus ensuring fair and balanced learning. These pre-processing steps collectively enhance the dataset's quality, enabling the machine learning model to learn patterns and relationships accurately, resulting in more reliable predictions regarding crime status in LA City.

i. Data Cleaning

In the process of preparing the crime prediction dataset sourced from NCVS focusing on LA City crimes from 2020 onwards, a comprehensive data cleansing journey unfolded. To streamline the dataset for enhanced predictive analysis, columns deemed irrelevant for the prediction task, such as 'DR_NO' and 'Crm Cd Desc', were meticulously dropped, ensuring a focus on pivotal and impactful features that could significantly contribute to the predictive models. Additionally, a careful curation process took place, wherein rare crime codes with occurrences below a predefined threshold of 1000 were eliminated. This strategic removal aimed to spotlight prevalent crimes, effectively mitigating potential noise that might obscure the predictive patterns. Furthermore, specific values within the 'Status' feature were deliberately removed, a decision likely influenced by their perceived insignificance or to prevent potential bias in subsequent modelling endeavours.

ii. Data Wrangling

The data underwent a transformative phase, especially in handling temporal aspects. The date columns, namely 'Date Rptd' and 'DATE OCC', underwent a conversion journey: first into datetime objects and then into numerical values. This conversion facilitated the modeling process by enabling algorithms to effectively interpret temporal patterns within the dataset. Moreover, to ensure the continuity and completeness of the dataset, a robust approach was adopted to handle missing values. Both forward and backward fill methods were employed to seamlessly fill in null values, thus ensuring a continuous flow of data without compromising its integrity or losing critical information.

Categorical variables played a pivotal role in this data pre-processing odyssey. 'Status', a categorical feature, underwent a transformation using the LabelEncoder technique. This transformation converted categorical variables into numerical equivalents, a crucial step allowing algorithms to comprehend and extract meaningful insights from these variables.

iii. Standardization

Standardization emerged as a fundamental practice to harmonize the numerical columns selected for scaling, including 'Date Rptd', 'DATE OCC', and others. Leveraging the StandardScaler, these numerical features were brought onto a unified scale, effectively averting any feature from unduly dominating the model due to its scale or magnitude. This meticulous standardization process significantly contributed to the model's equilibrium and accuracy in discerning patterns within the dataset.

CSV: Cleaned_Data.csv

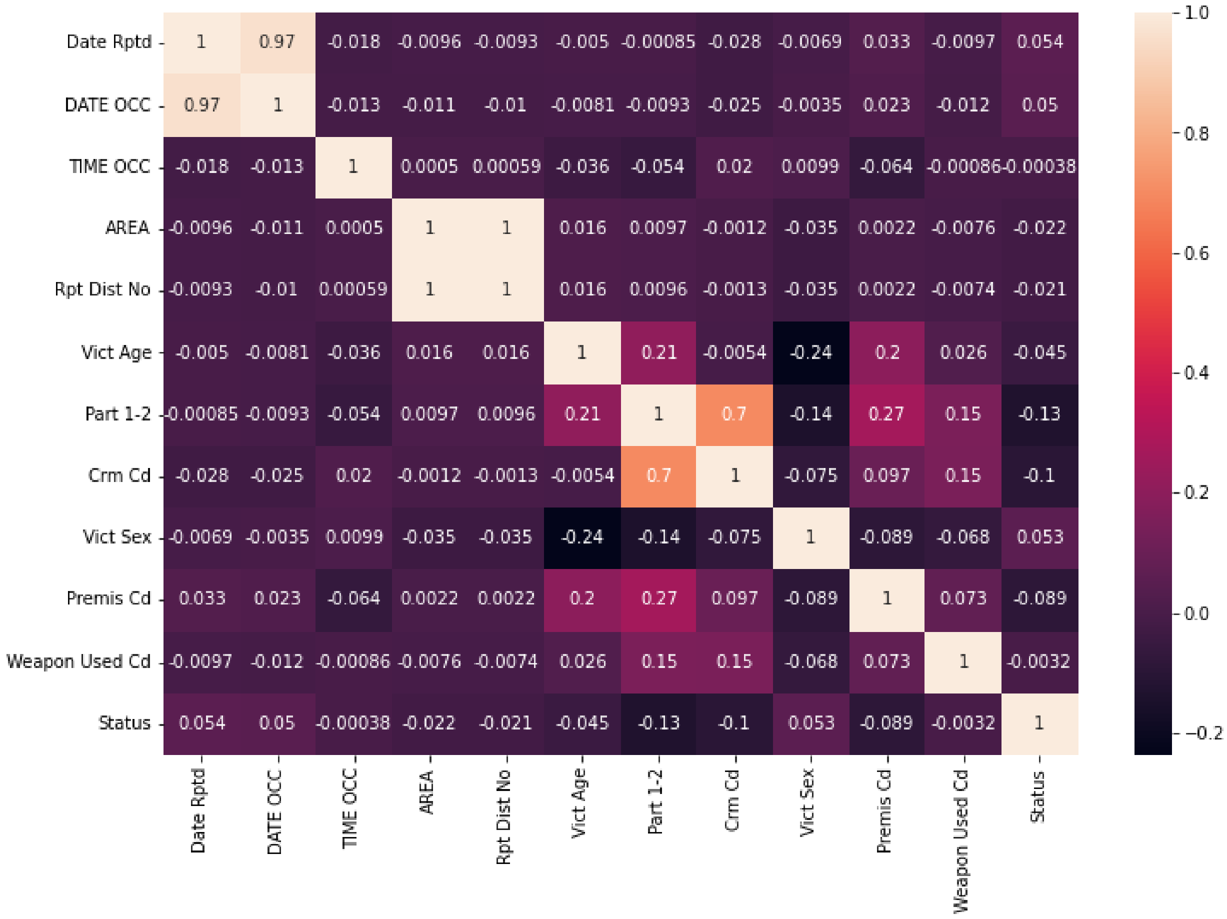
Attributes	Dtypes
Date Rptd	object
DATE OCC	object
TIME OCC	int64
AREA	int64
Rpt Dist No	int64
Part 1-2	int64

Crime Code	int64
Vict Age	int64
Vict Sex	object
Premis Cd	float64
Weapon Used Cd	float64
Status	Object

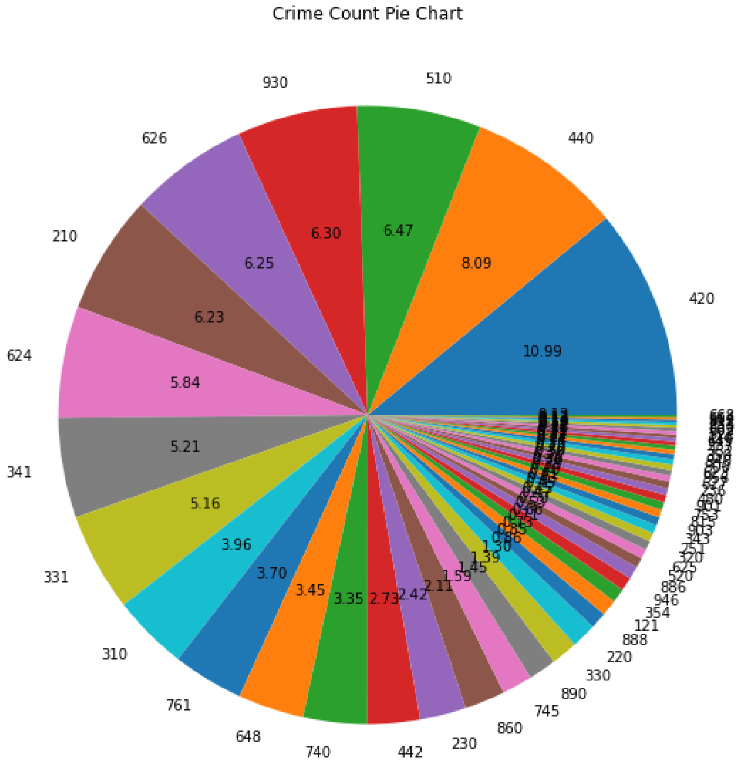
The resultant dataset, named "Cleaned_Data.csv," comprises 812,361 entries across 12 attributes, forming a refined foundation for building an ensemble crime status prediction model.

c. Exploratory Data Analysis

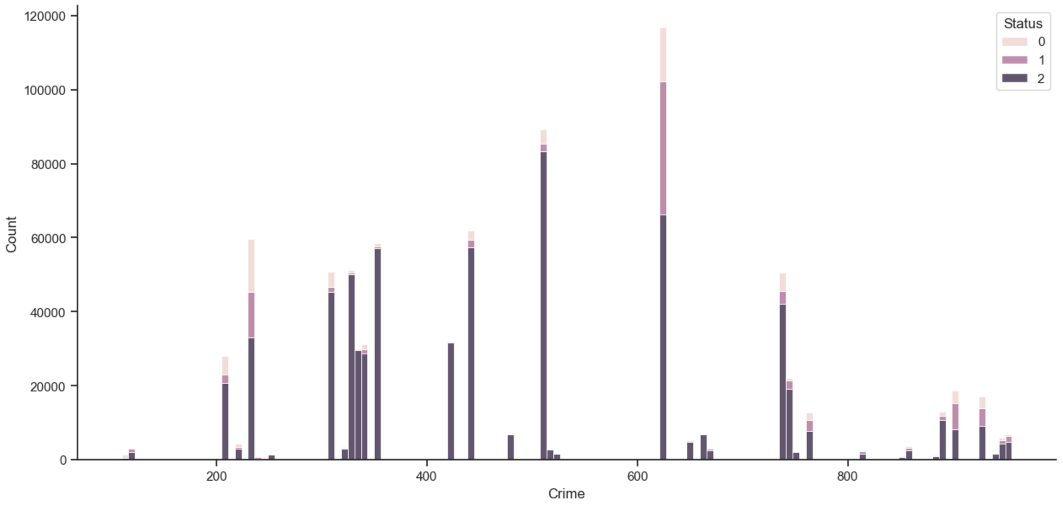
In exploring the correlation heatmap of the dataset attributes, it's evident that the 'Status' column exhibits notably minimal correlation with other attributes. The correlation heatmap, a visual representation showcasing the strength and direction of relationships between variables, reveals that 'Status' exhibits very weak correlation coefficients with other columns. This observation suggests that 'Status' has limited linear association or dependency on the other features within the dataset. The lack of significant correlation implies that the 'Status' attribute behaves largely independently concerning the variations or patterns present in the remaining dataset attributes. Consequently, when considering the predictive or explanatory power of other features on the 'Status' column, it's crucial to explore alternative methodologies or non-linear relationships, as its association with other attributes appears to be minimal based on the linear correlation analysis depicted in the heatmap.



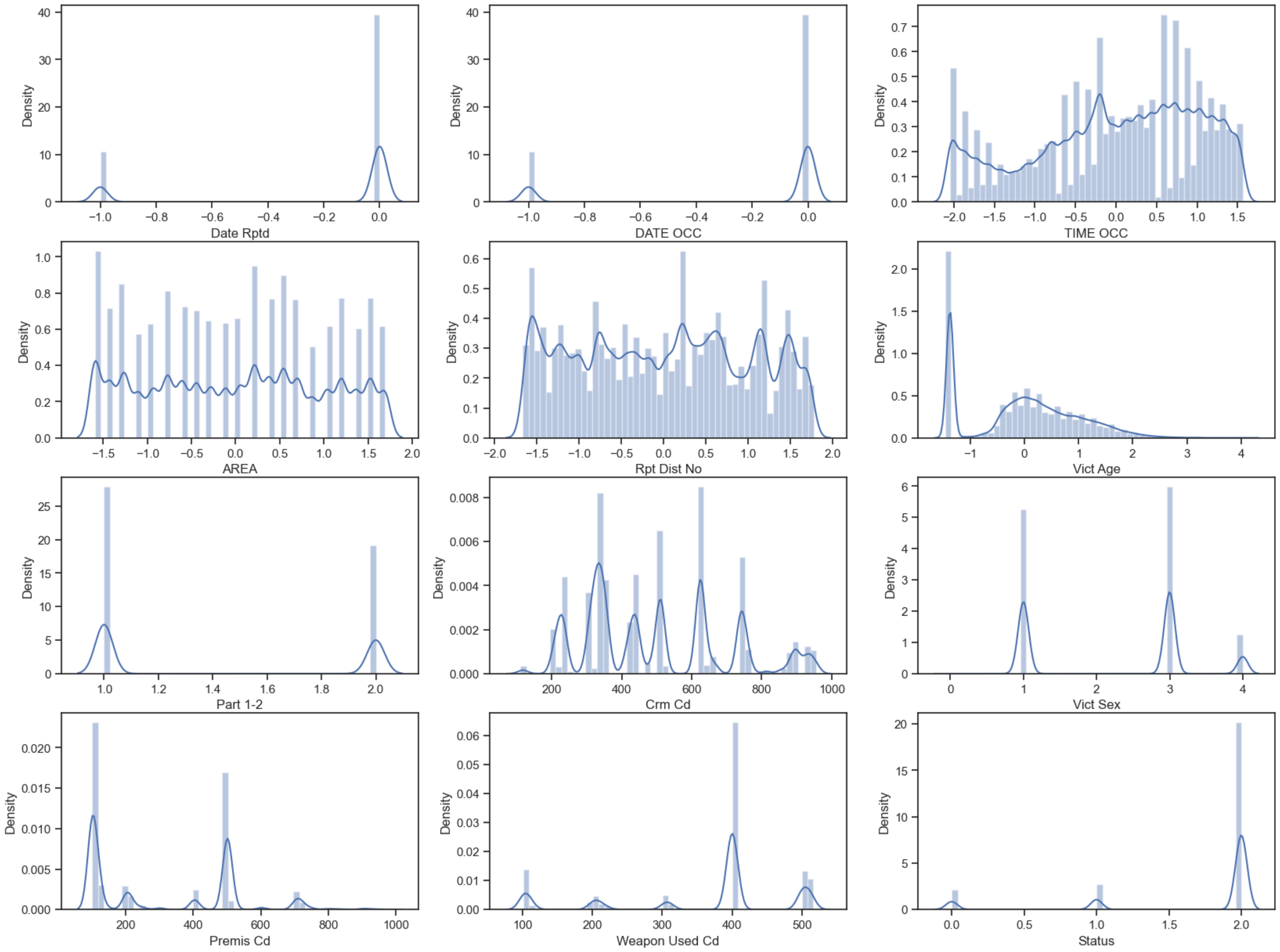
The pie chart showcasing the distribution of crime codes indicates a clear hierarchy in terms of frequency or occurrence within the dataset. Crime code 420 emerges as the most prevalent, constituting approximately 10.99% of the total reported crimes, signifying a substantial proportion of the incidents. Following closely, crime code 440 accounts for 8.09% of the reported incidents, illustrating a noteworthy presence in the dataset. Additionally, crime codes 510, 930, and 636 contribute 6.47%, 6.3%, and 6.35%, respectively, marking them as notable categories in terms of reported occurrences. Moreover, code 210 maintains a significant share, representing about 6.23% of the reported crimes. The remaining crime codes collectively constitute the rest of the distribution, implying a more dispersed representation with relatively lower frequencies compared to the top-ranking codes. This hierarchical breakdown provides valuable insights into the distribution of different crime codes, highlighting the most prevalent incidents and their respective contributions to the overall dataset.



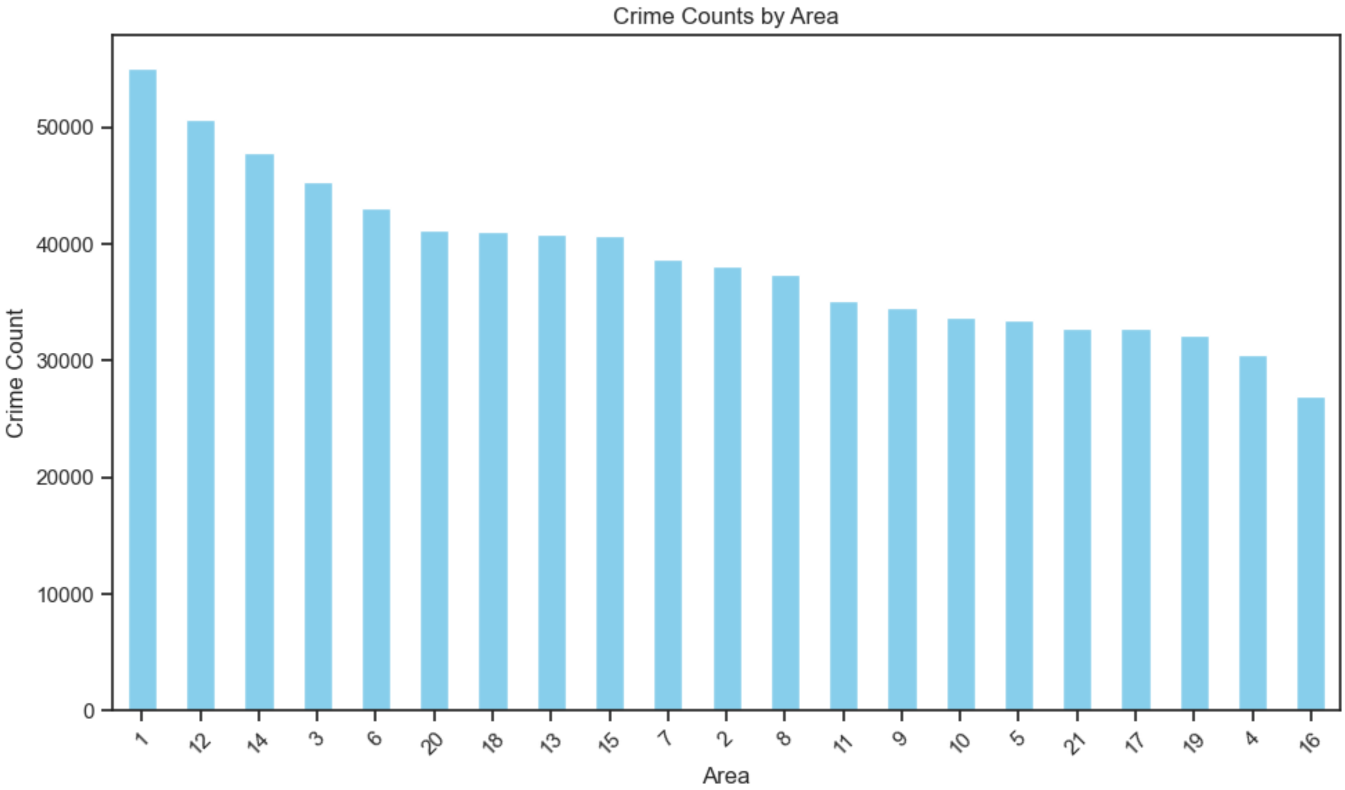
The bar chart representing the distribution of crime codes offers a clear visual hierarchy of the most prevalent incidents within the dataset. Among these codes, the highest count is observed for a specific crime code, highlighting its prominence compared to others. Following this prominent code, several other crime codes exhibit varying degrees of frequency, each representing a distinct portion of reported incidents. The descending bars in the chart provide an easily interpretable sequence, indicating the decreasing frequencies of different crime codes. Notably, this graphical representation effectively delineates the distribution of crime codes, emphasizing the significant variations in occurrence among different categories. Overall, the bar chart serves as a comprehensive visual guide, enabling a quick grasp of the hierarchy of crime codes based on their respective frequencies within the dataset.

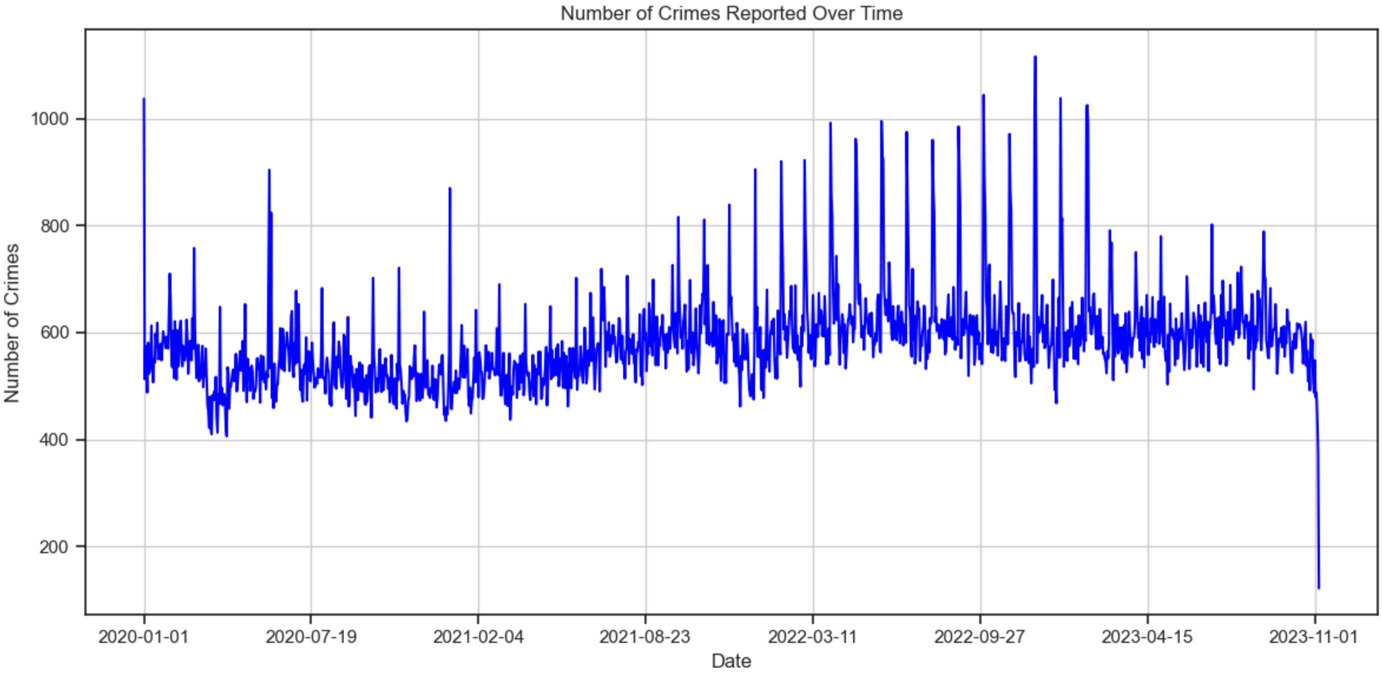


Creating count graphs for all attributes offers a comprehensive visual overview of the distribution of values within each column of the dataset. These count graphs provide a detailed depiction of the frequency or occurrence of distinct values present in the various attributes. By visualizing the counts for each attribute, one can easily discern the spread and concentration of different categories or values within the dataset. This extensive set of count graphs enables a quick assessment of the data distribution across multiple attributes, aiding in identifying any skewed distributions, dominant categories, or outliers present within the dataset. The visual representation of counts for each attribute serves as a valuable exploratory tool, facilitating a deeper understanding of the data's composition and patterns across different variables

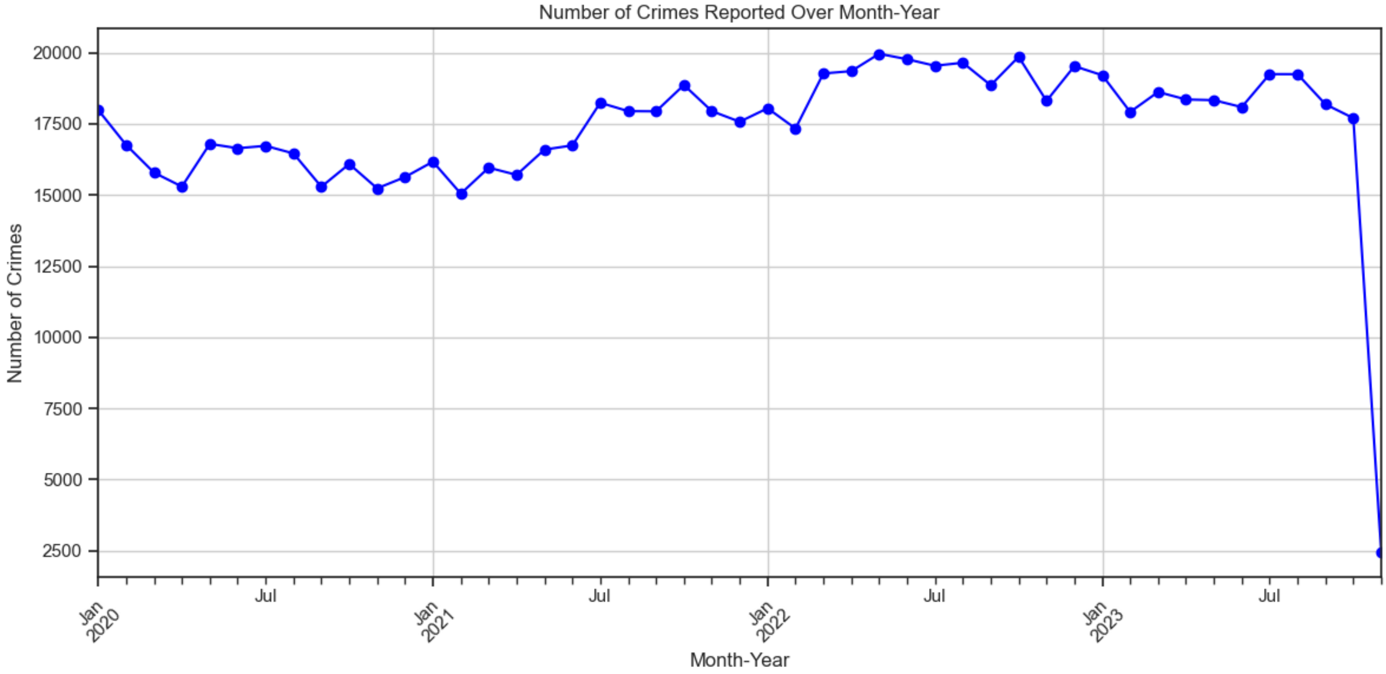


The crime counts by area graph presents an insightful portrayal of crime occurrences across various districts within the region. The graph highlights a hierarchy based on the frequency of reported incidents, with area 1 emerging as the district with the highest count of crimes, signifying its prevalence compared to others. Following closely, areas 12 and 14 exhibit notable crime counts, suggesting substantial incident rates within these regions. Moreover, districts 3, 6, 20, and 18 present themselves as significant contributors to reported crimes, showcasing distinct levels of occurrences within their respective areas. As the graph progresses, it reveals a descending order of crime counts among other districts, each representing varying extents of reported incidents. Notably, this graphical representation effectively delineates the distribution of crime counts across different areas, emphasizing the varying levels of criminal activities across these districts. Overall, the crime counts by area graph offers a clear depiction of the hierarchy of crime occurrences, enabling a swift understanding of the comparative frequency of reported incidents across different regions within the locality..

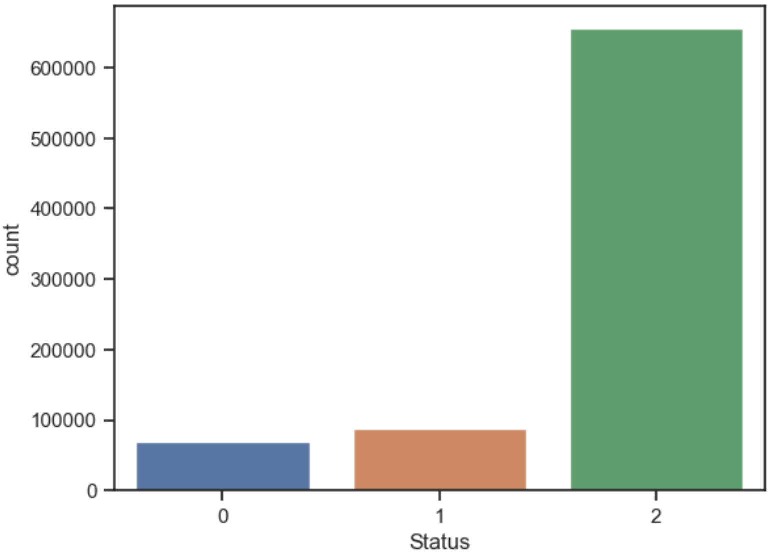




The graph illustrating the count of crimes per day reveals a consistent, gradual increase in reported incidents over time, mirroring the general upward trend observed in criminal activities. However, intertwined within this steady rise, the graph notably displays sporadic and substantial spikes at certain intervals. These spikes denote instances of remarkably high counts of reported crimes on specific days, standing out prominently against the backdrop of the overall rising trend. While the overall pattern depicts a gradual escalation in criminal incidents, these pronounced spikes serve as distinct anomalies, showcasing exceptional surges in reported crimes on particular days. These sharp and abrupt deviations from the consistent trend highlight specific instances where the number of reported incidents significantly exceeds the typical daily count, suggesting potential periods of heightened criminal activity or notable events that precipitated these surges. Despite the overarching gradual increase, these pronounced spikes punctuate the graph, signifying remarkable fluctuations in reported crime counts on select days throughout the observed timeframe. The month-year graph depicting the trend of reported crimes over time illustrates a discernible pattern characterized by a generally increasing trajectory interspersed with intermittent, albeit negligible, declines. The visual representation showcases a prevailing upward trend in the number of reported crimes over the specified time period. While the overall trend indicates a consistent rise in criminal activities, the graph does exhibit sporadic and minor downward movements at certain junctures. These occasional dips, although present, appear to be relatively inconsequential in the larger context of the escalating trend. Despite these intermittent fluctuations, the overarching trend delineates a clear and persistent increase in reported crimes over the observed period. This portrayal emphasizes the prominence of the rising pattern in crime occurrences, accentuated by occasional, marginal deviations that do not significantly deter the overall upward trajectory in reported incidents.



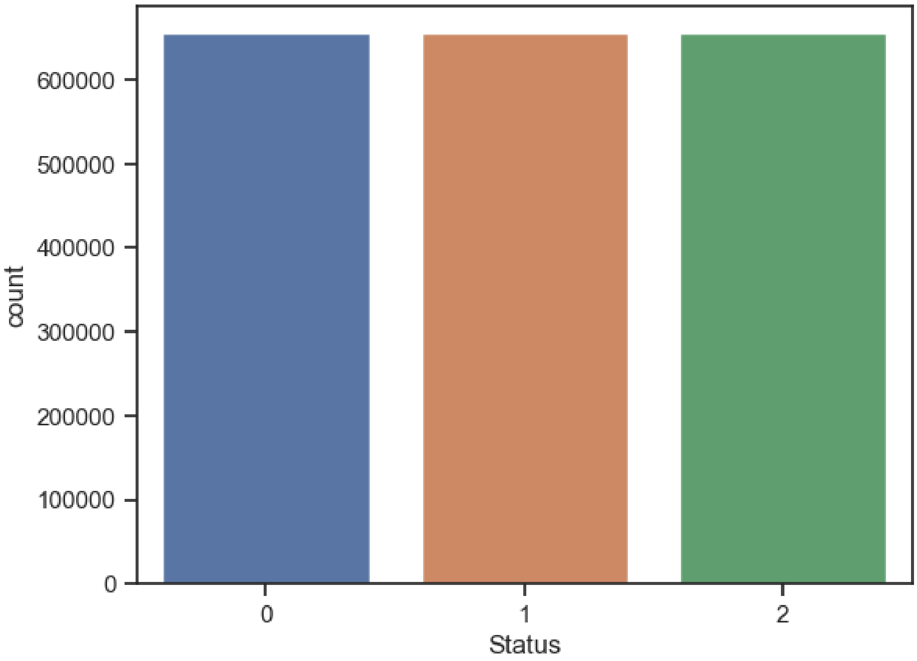
The count graph representing the output classes within the 'Status' attribute exhibits a distinct distribution among the three classes. Class 3 emerges as the predominant category, significantly outweighing the other two classes with a count exceeding 600,000 incidents. In contrast, the remaining two classes showcase a notably lower count, each recording a number of occurrences surpassing 100,000 but substantially lesser when compared to the dominant Class 3. This graphical representation effectively emphasizes the substantial imbalance in the distribution of classes within the 'Status' attribute. Class 3 stands out prominently as the overwhelmingly dominant category, overshadowing the prevalence of the other two classes, which demonstrate comparatively lower frequencies. This visual depiction underscores the significant disparity in the occurrence of different classes within the 'Status' attribute, with one class markedly prevailing over the others in terms of reported incidents.



i. Resampling

Recognizing the substantial disparity in the dataset's class distribution within the 'Status' attribute, a concern naturally arises regarding the imbalance among the three classes, notably with Class 3 overshadowing the others by a considerable margin. To address this issue, Synthetic Minority Over-sampling Technique (SMOTE) was employed to mitigate the imbalance. SMOTE, a widely-used technique in machine learning, works by generating synthetic samples for the minority classes, thereby equalizing the representation of all classes. By creating synthetic instances that resemble the existing minority class samples, SMOTE helps in rebalancing the dataset without duplicating existing data points. This process aims to enhance the learning process of the model, particularly in scenarios where the imbalance in class distribution might lead to biased predictions or decreased accuracy. By augmenting the dataset with synthetic samples, SMOTE facilitates a more balanced representation of all classes, fostering improved model performance and reducing the risk of the model favoring the majority class due to the disparity in class distribution. Ultimately, SMOTE's application in oversampling the dataset helps in creating a more equitable representation of classes, potentially leading to better predictive capabilities and enhanced model generalization.

```
from imblearn.over_sampling import SMOTE

X=df.iloc[:,:-1]
Y=df.iloc[:,-1]

smote = SMOTE(random_state=42)

X_resampled, y_resampled = smote.fit_resample(X, Y)

X=pd.DataFrame(X_resampled)
Y=pd.DataFrame(y_resampled,columns=['Status'])

X.reset_index(drop=True,inplace=True)
Y.reset_index(drop=True,inplace=True)

df=pd.concat([X, Y],axis=1)
```



d. Model Selection and Building

Utilizing classification models in crime status prediction enables the categorization of criminal incidents into distinct classes, aiding law enforcement in understanding and predicting the nature and outcome of reported crimes. These models analyze various features associated with crimes, such as location, time, demographics, and offense type, to predict the status or outcome of an incident, assisting in resource allocation, risk assessment, and proactive measures to prevent criminal activities.

Each classifier possesses unique characteristics and constraints. KNN (K-Nearest Neighbors) relies on proximity-based learning, suited for simple decision boundaries but may suffer from computational inefficiency with large datasets. SVM (Support Vector Machine) handles complex decision boundaries but may struggle with larger datasets due to computational demands. Decision Trees are interpretable and handle non-linear relationships but are prone to overfitting. MLP (Multi-Layer Perceptron) excels in learning complex patterns but requires careful tuning and is sensitive to hyperparameters and data scaling.

Employing multiple classifiers aids in validating the crime prediction model by assessing diverse learning approaches. Each model perceives patterns differently, offering varied perspectives on the data. By comparing their performance metrics and identifying consensus among predictions, we ascertain the model's robustness and generalize its predictive capabilities across different learning paradigms. This approach helps in mitigating biases or limitations inherent in a single model, enhancing the model's reliability and adaptability in real-world crime prediction scenarios.

e. Model Evaluation

Evaluating the models solely based on accuracy may overlook crucial aspects of performance. Assessing precision, recall, and F1 score alongside accuracy provides a more comprehensive understanding of a model's effectiveness in crime prediction. Precision measures the accuracy of positive predictions, recall assesses the model's ability to capture actual positives, while the F1 score balances precision and recall. By considering these metrics collectively, a more nuanced appraisal of a model's overall performance emerges. This approach aids in identifying the most suitable model for crime prediction, one that achieves a balance between accurate positive predictions, effective capture of actual positives, and a harmonized precision-recall trade-off, ensuring a more robust and reliable model selection process.

5. Result Analysis

Model	**Accuracy**	**Recall**	**Precision**	**F1**
KNN	84.66	84.66	86.11	84.04
SVM	81.475	81.475	82.235	81.175
MLP	57.1	57.1	57.39	56.57
Decision Tree	78.29	78.29	78.36	78.31
Random Forest With 10 DTs	85.44	85.44	85.47	85.43
Random Forest With 20 DTs	86.13	86.13	86.19	86.14

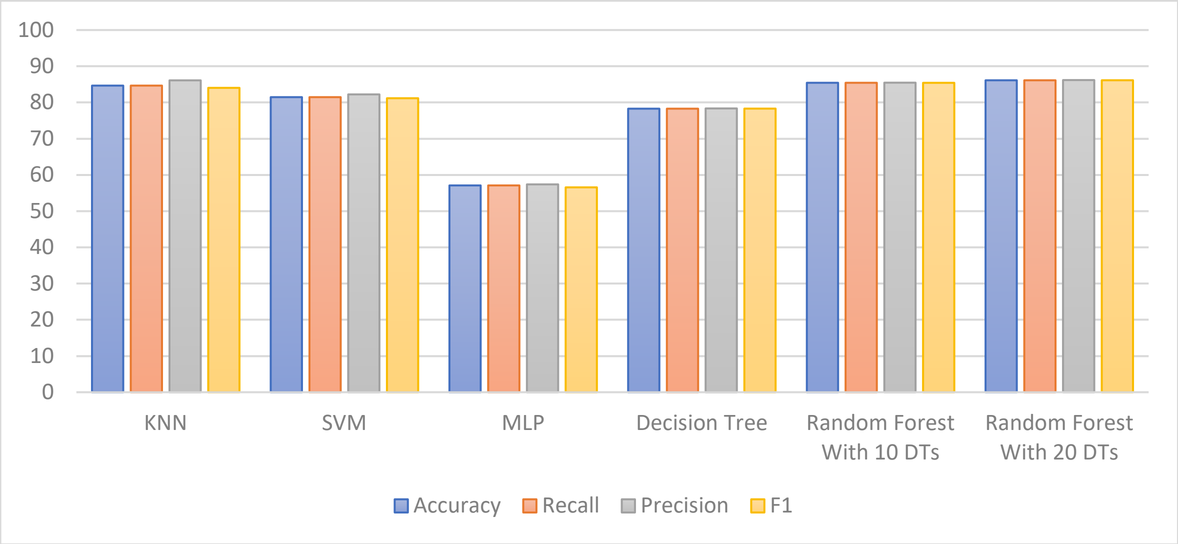
The evaluation results across various metrics for the classification models offer several important inferences:

- **Accuracy and Consistency:** The Random Forest models, especially with 10 and 20 Decision Trees, demonstrate the highest accuracy rates, exceeding 85%. This signifies their consistent and accurate predictions compared to other models. KNN also performs reasonably well with an accuracy of 84.66%. In contrast, SVM and Decision Tree models exhibit slightly lower accuracy scores, while the MLP model shows the lowest accuracy at 57.1%.

- **Balanced Precision and Recall:** Across the models, precision and recall scores generally align closely, indicating a balance between accurate positive predictions and the model's ability to capture actual positives. The Random Forest models, KNN, and SVM show relatively balanced precision-recall trade-offs, with slight variations among them. However, the MLP model demonstrates a notable disparity between precision and recall, indicating a potential bias toward false negatives or positives.

- **F1 Score Consistency:** F1 scores, which harmonize precision and recall, show consistency across models. The Random Forest models, especially with 20 Decision Trees, maintain the highest F1 scores above 86%, signifying a good balance between precision and recall. KNN also exhibits a reasonable F1 score, indicating a balanced performance in identifying true positives while minimizing false positives and negatives. However, SVM, Decision Tree, and notably the MLP model, display lower and more varied F1 scores, suggesting a less balanced trade-off between precision and recall.

- **Model Performance Ranking:** The Random Forest models with 20 and 10 Decision Trees emerge as the top performers across various evaluation metrics, showcasing consistency in accuracy, precision, recall, and F1 scores. KNN follows closely behind, displaying robust and balanced performance. SVM and Decision Tree models exhibit decent but comparatively lower performance in multiple metrics. The MLP model, while showing potential, requires further optimization to enhance its predictive capabilities.



Additionally, a 10-fold analysis on the Random Forest model reveals a mean prediction accuracy of 86.20%, with minimal variance (0.0000016) and a standard deviation of 0.00112122. The range spans from a maximum of 86.42% to a minimum of 86.02%, emphasizing the model's stability and reliability in consistently predicting crime status.

6. Conclusion

The crime status prediction effort utilized a meticulous approach, starting with data collection from Catalog.Data.Gov on Los Angeles crime since 2020. Despite initial challenges like data inaccuracies and privacy measures, rigorous cleaning and standardization prepared the dataset. Exploratory analysis revealed the 'Status' attribute's weak correlation, crime code distributions, area-wise crime counts, and temporal patterns. Notably, class imbalance within 'Status' led to employing SMOTE to balance the dataset. The selection of classification models, each with its unique strengths and limitations, allowed for a diverse approach to crime status prediction. Evaluation of these models across multiple metrics provided a nuanced understanding of their performance.

The conclusion drawn from the model evaluation highlights the superiority of the Random Forest models, particularly those employing 10 and 20 Decision Trees. These models consistently demonstrated high accuracy, balanced precision-recall trade-offs, and notable F1 scores, indicating their robustness in predicting crime status. KNN also portrayed commendable performance, showcasing reliability and balanced predictive capabilities. However, SVM, Decision Tree, and notably the MLP model showed comparatively lower and more varied performance metrics, suggesting areas for further improvement and optimization. In conclusion, the ensemble approach utilizing Random Forest models, especially those with 10 and 20 Decision Trees, alongside KNN, proved to be the most promising for crime status prediction within the dataset. These models showcased consistent and robust performance across various evaluation metrics, offering reliable predictive capabilities crucial for aiding law enforcement in understanding, strategizing, and mitigating criminal activities within the City of Los Angeles.

7. References

[1] M. N. S. M. A. Y. Abba Babakura, “Improved Method of Classification Algorithms for Crime Prediction,” *ISBAST, IEEE,* 2014.

[2] L. M. a. N. Meghanathan, “Using Machine Learning Algorithms to Analyze Crime Data,” *MLAIJ,* 2015.

[3] C.-l. Y. X. L. K. L. Cui-cui Sun, “Detecting Crime Types Using Classification Algorithms,” *Journal of Digital Information Management ,* 2014.

[4] S. A. J. Nitin Nandkumar Sakhare, “Classification of Criminal Data using J48 Algorithm,” *IFRSA International Journal of Data Warehousing & Mining ,* 2015.

[5] C. W. H. C. a. X. J. Shaobing Wu, “Crime Prediction Using Data Mining and Machine Learning,” *Springer Nature Switzerland AG ,* 2020.

[6] M. V. Barnadas, “MACHINE LEARNING APPLIED TO CRIME PREDICTION,” *Telecom BCN ,* 2016.

[7] K. N. V. P. G. S. R. G. Y. S. Kanimozhi N, “CRIME TYPE AND OCCURRENCE PREDICTION USING MACHINE LEARNING ALGORITHM,” *ICAIS, IEEE ,* 2021.

[8] L. A. LAPD, “Data.GOV,” LA Goverment, 11 11 2023. [Online]. Available: https://catalog.data.gov/dataset/crime-data-from-2020-to-present.