ITM UNIVERSITY
GWALIOR • MP • INDIA
"CELEBRATING DREAMS"

ITM University Gwalior

PBL Synopsis File

On

Crime Status Prediction

(MCA-304)

Problem Statement:

Can we predict what will be the end status for any crime?

Submitted To:

Dr. Sanjay Jain
Professor, Dept. Of CSA

Submitted By:

Prashant Singh
MCAN1CA2201

Table of Contents

1. Abstract

This machine learning-driven analysis aims to revolutionize the prediction of final status codes for reported crimes, utilizing data from the National Crime Victimization Survey (NCVS) within the Gov Crime Data in Los Angeles, CA, USA. Employing an ensemble approach in predictive modelling, the study seeks to significantly enhance anticipatory capabilities regarding the resolution of criminal incidents. This advancement contributes to the refinement of law enforcement strategies, offering a proactive tool for policymakers. Recognizing the pivotal role of understanding the trajectory of reported crimes, this PBL underscores the importance of informed policy formulation and intervention initiatives.

The preliminary exploration has uncovered valuable insights, providing a solid foundation for this study to extend into the predictive realm. Through the integration of machine learning algorithms within an ensemble framework, this PBL aspires to empower law enforcement agencies and policymakers with a sophisticated tool for crime prevention and intervention. The anticipated outcomes hold promise for shaping targeted, efficient, and well-informed strategies to ensure public safety and well-being in Los Angeles and comparable regions facing similar challenges.

2. Introduction

In the dynamic landscape of crime data analysis, the integration of machine learning (ML) models emerges as a potent force to deepen our comprehension of criminal activities. Amidst significant crime challenges in the United States impacting the safety of its extensive population, harnessing advanced analytical techniques becomes imperative. While the national crime database serves as a foundational resource, this paper contends that untapped potential lies in leveraging a novel ensemble machine learning model to predict and comprehend the final outcomes of reported crimes. Focused on data sourced from the National Crime Victimization Survey (NCVS) in Los Angeles, CA, USA, this approach represents a pioneering stride toward a more nuanced and accurate understanding of crime resolution. This PBL recognizes that the ensemble approach, combining the strengths of diverse machine learning algorithms, is paramount in addressing the multifaceted nature of crime data. By seamlessly integrating predictive modeling into the analytical landscape, this study endeavors to refine and optimize law enforcement strategies, thereby fostering a proactive and adaptive response to emerging criminal incidents. The ensuing sections will delve into the methodology, expected outcomes, and significance of this ensemble-driven machine learning approach in shaping the future of crime data analysis and proactive law enforcement.

3. Literature Review

The landscape of crime data analysis has witnessed a surge in innovative methodologies, particularly those driven by machine learning (ML) algorithms. In "Improved Method of Classification Algorithms for Crime Prediction" by Abba Babakura et al. (2014), the authors propose a classification algorithm-based approach to predict crime categories in different U.S. states. Notably, Naïve Bayes outperforms Back Propagation, underscoring the significance of data pre-processing and feature selection in refining crime prediction models. [1]

Lawrence McClendon and Natarajan Meghanathan's work (2015) contributes a comprehensive review, focusing on spatial and temporal data analysis with a case study in Maryland State, USA. Leveraging machine learning via Weka, the study discerns patterns in crime occurrence, aiding law enforcement in strategic resource allocation. [2]

The study by Cui-cui Sun et al. (2014) delves into the statistical analysis of spatial crime data, emphasizing the relevance of spatial analysis in criminology. Spatial autocorrelation and various regression models are explored, offering a nuanced understanding of crime distribution across space and factors influencing it. [3]

Nitin Nandkumar Sakhare and Swati Atul Joshi's paper (2015) investigates the use of ML, employing K-nearest neighbour and boosted decision tree algorithms to analyze Vancouver crime data. Acknowledging limitations, the study advocates for further research to enhance predictive accuracy, highlighting the potential of ML in predicting crime types. [4]

The document by Shaobing Wu et al. (2020) centres on data mining and ML for crime prediction in YD County, China. Employing random forest, neural network, and Bayesian network algorithms, the study attains a remarkable 90% accuracy. Factors like population size, age distribution, and crime types emerge as significant predictors, showcasing the potential for informed decision-making by law enforcement. [5]

Miquel Vaquero Barnadas' work (2016) introduces a comprehensive system for crime analysis and prediction, utilizing data mining, a Naive Bayes classifier, and unstructured databases. The system achieves over 80% accuracy, offering valuable insights into crime-prone regions and aiding resource allocation. [6]

In "Crime Type and Occurrence Prediction Using Machine Learning Algorithm" (2021), Kanimozhi N et al. propose an ML algorithm for predicting crime types and occurrences. Noteworthy for its ability to handle nominal and real-valued attributes, the algorithm, based on Naïve Bayes, achieves a high accuracy of 93.07%. Its versatility in handling both categorical and continuous crime types positions it as a promising tool for real-time predictions. [7]

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

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:

- ***Data Collection***: The initial phase involves the utilization of a comprehensive dataset encompassing a spectrum of features, including temporal information, spatial data, demographic details, and historical crime records. This rich dataset serves as the foundation for training and evaluating the ensemble model.
- ***Pre-processing***: A meticulous data pre-processing step is undertaken to cleanse and refine the dataset. This involves addressing missing values, identifying and handling outliers, and ensuring the dataset's compatibility with diverse classification algorithms.
- ***Feature Selection***: The subsequent step focuses on enhancing model efficiency by identifying the most pertinent features. Techniques such as recursive feature elimination and correlation analysis are employed to isolate and prioritize features crucial for accurate predictions.
- ***Ensemble Model Construction***: The crux of this research lies in the development of an ensemble model that amalgamates multiple machine learning classification models. This ensemble may include, but is not limited to, Naïve Bayes, Decision Trees, Random Forest, Support Vector Machines, and Gradient Boosting. By integrating these diverse models, the ensemble approach seeks to leverage the unique strengths of each algorithm.
- ***Voting Mechanism***: A sophisticated weighted voting mechanism will be implemented to aggregate predictions from individual models within the ensemble. This mechanism takes into account the relative strengths and weaknesses of each model, ensuring a balanced and informed decision-making process.

By intricately weaving these methodological steps together, this research endeavours to transcend the limitations of individual classification algorithms and create a robust ensemble model poised to revolutionize the landscape of crime status predictions. The ensuing sections will delve into the expected outcomes, significance, and potential applications of this pioneering ensemble-driven approach in advancing crime prediction methodologies.

5. Dataset

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

To facilitate meaningful analysis and enhance the effectiveness of an ensemble crime status prediction model, the extraction process involves dealing with mixed data, combining numerical and categorical information. Techniques like one-hot encoding, label encoding, and feature scaling are applied to manipulate the data for accurate and insightful analysis.

In the data cleansing phase, redundant attributes such as 'Crm Cd Desc,' 'AREA NAME,' 'Premis Desc,' 'Weapon Desc,' 'Status Desc,' and 'Cross Street' are removed. Additionally, certain attributes like 'DR_NO,' 'Mocodes,' 'LAT,' 'LON,' 'LOCATION,' 'Vict Descent,' and 'Crm Cd 1-4' are eliminated due to limited relevance for the ensemble crime status prediction model.

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

Furthermore, crimes with a count below the threshold value of 1000 are excluded to streamline the dataset and enhance its relevance for ensemble model development. The resultant dataset, named "Cleaned_Data.csv," comprises 816,247 entries across 12 attributes, forming a refined foundation for building an ensemble crime status prediction model.

6. Expected Outcomes

Ensemble Model

The ensemble crime status prediction model will be crafted by integrating the strengths of Naïve Bayes, Decision Trees, Random Forest, Support Vector Machines, and Gradient Boosting algorithms. The model's architecture will leverage a weighted voting mechanism to aggregate predictions, considering the distinctive strengths and weaknesses of each underlying algorithm. This collaborative approach is designed to mitigate the limitations of individual models, resulting in a more comprehensive and accurate crime status prediction tool.

Outcomes

Improved Prediction Accuracy: The implementation of the ensemble crime status prediction model is expected to yield a substantial enhancement in prediction accuracy when compared to the performance of individual models. By amalgamating the strengths of diverse classification algorithms, including Naïve Bayes, Decision Trees, Random Forest, Support Vector Machines, and Gradient Boosting, the ensemble model aims to capture a broader range of patterns within crime data, leading to more precise and reliable predictions.

Robustness: The ensemble approach seeks to fortify the predictive model's robustness by incorporating a variety of classification algorithms. This diversity is anticipated to contribute to a more resilient framework, capable of delivering consistent and accurate predictions across various types of crimes. The ensemble model's adaptability to different crime scenarios is crucial for its effectiveness in real-world applications.

Evaluation

Cross-Validation Techniques: To thoroughly assess the ensemble crime status prediction model, cross-validation techniques will be employed. Metrics such as accuracy, precision, recall, and F1-score will be utilized to gauge the model's performance under different scenarios and ensure its reliability across diverse datasets.

Comparison with Individual Models: The performance of the ensemble model will be systematically compared against that of individual models to ascertain the effectiveness of the proposed approach. This comparative analysis aims to highlight the added value of the ensemble strategy in terms of accuracy and predictive capability.

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.