Crime Spotter: Intelligent Crime Hotspot Detection

Student : Aishwarya Hiremath

Seidenberg School of Computer Science & Information Systems

INTRODUCTION

- The project aims to improve crime prediction and prevention in urban areas.
- Advanced machine learning (ML) models are developed using the NYPD arrest dataset from NYC Open Data.
- The focus is on accurately identifying crime hotspots and analyzing temporal trends.
- Law enforcement agencies can use the findings for better resource allocation and proactive crime prevention.
- The project compares ML models with traditional statistical methods to assess accuracy in detecting crime patterns.
- The research seeks to demonstrate that ML models provide superior accuracy in crime detection.
- The project could revolutionize crime prevention strategies in urban environments.
- It offers a blueprint for implementing similar initiatives in other cities.
- The ultimate goal is to contribute to safer urban environments through data-driven insights.

OBJECTIVES

Utilize machine learning (ML) techniques to accurately predict crime hotspots and analyze spatial and temporal trends in criminal activity using the NYPD arrest dataset from NYC Open Data.

METHODOLOGY

1. Data Collection and Preprocessing:

- The project uses the NYPD arrest dataset from NYC Open Data.
- Unnecessary columns were removed, and relevant columns were renamed for clarity.
- New columns like 'Description' are added for standardized offense descriptions and 'Class' for categorizing offenses. Temporal features including 'year', 'day', 'day_name', and 'month_name' from the 'Arrest_Date' field were Derived
- Created an 'Indicator' column to detect crime hotspots based on offense type and added the 'No_days' column to track crime recency.

2. Exploratory Data Analysis (EDA):

- Conducted detailed data analysis using visualization tools like Seaborn and Matplotlib to uncover patterns, trends, and insights.

3. Feature Encoding:

- Utilized the Label Encoder in Python to convert categorical variables into numerical values, ensuring compatibility with machine learning models.

4. Data Splitting:

- The dataset was split into training and testing sets using an 80:20 ratio with the `train_test_split` library.

5. Stratified K-Fold Cross-Validation:

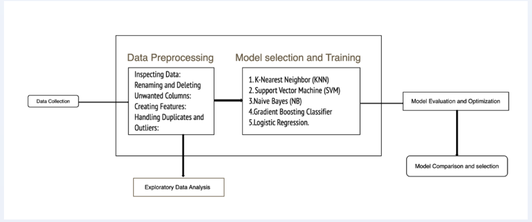
- Performed Stratified K-Fold Cross-Validation to ensure that each fold of the dataset maintains the proportion of each class, enhancing the robustness of the model evaluation.

6. Model Training:

- Implemented and trained various machine learning algorithms, including K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Gradient Boosting Classifier, and Logistic Regression.

7. Performance Evaluation:

- Evaluated model performance using metrics such as accuracy, precision, recall, F1-score, and confusion matrix to assess their effectiveness in crime prediction.



RESULTS

Naive Bayes:

- The model achieved an average validation accuracy of 0.77.
- It performs well on Green and Yellow classes, but struggles with Orange and Red, showing high precision for Orange but low recall, and perfect recall for Red with lower precision.
- The confusion matrix reveals that while Blue, Green, and Yellow classes are largely correctly classified, Orange and Red instances are misclassified across different classes.

Support Vector Machine (SVM):

- The model achieved a high average validation accuracy of 0.99.
- It performs exceptionally well on Blue, Green, and Yellow classes, with moderate performance on Orange, but struggles with the Red class, failing to predict it correctly.
- The confusion matrix shows strong performance for Blue, Green, and Yellow, while the Red class has significant misclassifications and the Orange class has some confusion with Yellow.

KNN

- The model achieved a high average validation accuracy of 97%.
- Precision, recall, and F1-scores are exceptional across most classes, with slightly lower recall for the Red class.
- The confusion matrix shows the model's strong performance in correctly classifying most instances, with a few misclassifications between classes.

Gradient Boosting:

- The model achieved perfect average validation and final fold accuracies of 1.00, indicating flawless classification on validation and final data.
- The classification report shows 100% precision, recall, and F1-scores across all classes, demonstrating the model's perfect learning of the dataset patterns.
- Despite the perfect performance, the model may be overfitting, as it might not generalize well to new, unseen data, suggesting a need for further validation on independent datasets.

Logistic Regression:

- The model achieved an average validation accuracy of 0.97.
- The classification report highlights near-perfect precision, recall, and F1-scores for Green, Red, and Yellow classes, with Blue and Orange also performing very well but Blue having slightly lower recall.
- The confusion matrix analysis shows minimal misclassifications across classes, with Blue, Green, Orange, and Yellow being correctly classified in the majority of instances.

CONCLUSION

- Machine learning models like KNN and Logistic Regression showed high accuracy and reliable performance, Gradient Boosting Classifier experienced overfitting, limiting their practical use.
- Future work should focus on refining models for better accuracy and generalization, exploring additional features, and incorporating diverse and real-time data to enhance the robustness of crime hotspot detection systems.
- The study highlights the potential of machine learning to improve public safety by providing valuable insights for law enforcement and public safety agencies through effective crime hotspot detection.

REFERENCES

✦XU ZHANG, LIN LIU, LUZI XIA, AND JIAKAI JI. "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots". October 2, 2020.

✦Yadhunath Ramshankar, Srikanth Srivenkata, Sudheer Arvind and Palaniswamy Suja. "Identification of Criminal Activity Hotspots using Machine Learning to Aid in Effective Utilization of Police Patrolling in Cities with High Crime Rates". IEEE, 12 March 2020.

✦Marcus Pinto, Hsinrong Wei, Kiyatou Konate, and Ida Touray. "Delving into Factors Influencing New York Crime Data with the Tools of Machine Learning".ACM, NY, USA. 01 October 2020

✦Sphamandla May, Omowunmi Isafiade, and Olasupo Ajayi. "An Enhanced Naïve Bayes Model for Crime Prediction using Recursive Feature Elimination". Association for Computing Machinery New York, NY, United States. 25 February 2022

ACKNOWLEDGEMENTS & CONTACT

I would like to express my heartfelt gratitude to Professor Krishna Bathula at Pace University for his invaluable guidance, support, and mentorship throughout this project. His expertise and encouragement were instrumental in the successful completion of this work.

For any inquiries or further information regarding this project, please feel free to contact me at:

Email: salimath.aishu@gmail.com

LinkedIn: https://www.linkedin.com/in/aishwarya-hiremath-4987a22a4/

GitHub: https://github.com/ah11441n