Crime Prediction through Data Analysis and Statistical Learning

Team Number: 17
Team Members: Afreen, Kethavath Kousalya, Ambaldhage Tarun, Meher Dhavala

Introduction

Criminal activity has remained a concern for many years, affecting both quality of life and economic progress. It's essential to remain vigilant, especially when visiting areas with higher crime rates. The crime rates can vary significantly depending on various factors such as location, time of year, socioeconomic conditions, and law enforcement efforts.

Problem Definition

Given a dataset containing information about past criminal incidents, socioeconomic factors, and law enforcement efforts, the task is to develop machine learning classification models that accurately predict the occurrence of crimes in specific locations and time periods. Based on the given factors, our goal is to utilize methodologies in data analysis and machine learning classification models that can study how crimes happen and improve how accurately we predict where and when they might occur (#1).

Importance: Effective crime prediction can significantly enhance the ability of law enforcement agencies to prevent crimes before they occur, optimize resource allocation, and improve community safety. By anticipating crime hotspots and times, agencies can deploy officers more strategically and efficiently, potentially reducing crime rates and increasing public trust.

The primary objective is to develop machine learning models that predict crime occurrences accurately. Key research questions include:

1. How do different socioeconomic factors influence crime rates?(time and place)
2. What is the impact of law enforcement efforts on crime prediction accuracy?(crimes solved, unsolved)
3. How can machine learning models be used to identify crime hotspots and times?

Literature Survey

A significant amount of data exists concerning criminal activity; this extensive crime-related data provides valuable insights for analyzing criminal patterns. This real-time analysis will provide law enforcement with up-to-date insights for quick decision making(#4). There are various approaches for forecasting crime data using Machine Learning and Deep Learning techniques. Models such as

Random Forest, Decision Tree, Logistic Regress have been utilized to analyze crime data and identify patterns that can be used to predict criminal activity. These models were implemented on the data and compared the accuracy of the prediction [1]. Deep learning approaches, including CNNs and RNNs, are also applied to handle large, complex datasets for better prediction.

By utilizing the Logistic Regression algorithm for forecasting crime types, time, and locations, we plan to apply data preprocessing techniques such as random undersampling to mitigate the class imbalance present in the dataset [3] and feature selection methods similar to those outlined in the study based on historical data, aiming to boost our prediction accuracy [4]. And most of the current research in crime data analysis focuses primarily on datasets with crime locations, often overlooking crucial factors such as the type and timing of the crime(#2).

As a result, we intend to use the classification models to uncover correlations between crime incidents and these crucial factors, including the crime location, victim age, gender, type of crime and others. The dataset we are going to use for our model is taken from a free source kaggle(#7). Integration of different predictors optimizes model performance and accuracy(#3). While the sensitivity of this data could be a privacy risk and the data may also lead to algorithmic complexity and potential biases in prediction models, the advantage is an improved crime prediction system, resulting in lower crime rates(#5) and safer communities(#6).

Using the classification models, we can examine spatial trends of criminal activity to highlight areas with elevated risk or crime zones in the city [2]. A few studies highlight the significance of crime analysis, the difficulties in accurately predicting future crimes, and encourage the development of hybrid approaches to enhance accuracy in handling increasing volumes of crime data [5].

[6] The spatio-temporal analysis methods can help us in examining our crime data, revealing comparable crime patterns and characteristics relevant to the area. And the findings on seasonal and periodical patterns of crime can guide us in recognizing comparable trends in our project and help in targeted crime prevention strategies.

The midterm success check involves researching related papers, collecting diverse data, and preprocessing it effectively for modeling. The final success check includes model selection, training and testing, and data visualization, along with comprehensive documentation for reproducibility and transparency in research outcomes. Progress is measured by tracking the completion of activities, ensuring timely execution and quality outcomes that are aligned with project goals.(#9)

Proposed Method

1. **Intuition**: Our proposed method aims to surpass the current state of the art by integrating comprehensive data collection, advanced feature selection techniques, and diverse machine learning algorithms. By leveraging domain knowledge and machine learning algorithms like Decision Tree, Random Forest and Logistic Regression to capture intricate feature interactions
2. **Approaches:**

   Algorithms:
   We utilized three machine learning algorithms such as Decision Trees, Random Forests, Logistic Regression to predict crime occurrences. Visualizations were created using heatmaps, feature importance plots, and confusion matrices to understand the data and model performance.

- Decision Trees split data into branches based on feature values, making decisions at each node. They are simple, interpretable, but prone to overfitting.
- Random Forests are an ensemble method that builds multiple decision trees and merges their predictions, enhancing accuracy and robustness, thus reducing overfitting.
- Logistic Regression models the probability of a binary outcome using a logistic function to transform a linear combination of input features, providing probabilistic outputs interpreted as confidence levels. It is simple, effective, and less prone to overfitting, assuming a linear relationship between features and log-odds, trained by maximizing the likelihood function.

These approaches were chosen for their respective strengths: Decision Trees and Random Forests excel at identifying important features and patterns, Logistic regression is effective for binary classification, providing interpretability, probability estimation, suitability for linearly separable data, scalability, identification of feature importance , but may not perform well in highly non-linear relationships.

Visualizations:
To interpret the model results and understand the data, we used various visualization techniques:

- Heatmaps to display the number of crimes based on the crime types.
- Ggplots to visualize crimes based on area, status of the crime.
- Feature importance plots to identify the most important features.
- Confusion matrices to assess model performance.

These visualizations help in understanding the underlying patterns in crime data and the effectiveness of our predictive models. The combination of these approaches enhances the overall predictive power and reliability of the crime prediction system, providing valuable insights for law enforcement agencies.

Derived new features from existing dataset attributes have been innovatively analyzed to assess their impact on crime rates. These novel features have significantly contributed to generating visualizations that enhance the comprehension of crime patterns in particular areas.

Data Collection and Exploration:
We begin by gathering historical crime data of Los Angeles from kaggle, encompassing essential variables such as crime type, location, time of occurrence, demographics, and weapons used. Our initial step involved exploring this dataset to comprehend its data types, identified missing values, and derived summary statistics.

Before data preprocessing and feature extraction, the size of the data is about 170MB and the number of records in the dataset are around 170000. The data is then processed by removing duplicate rows and null value rows which resulted in 20MB of data and 20000 of records. Crime data, being temporal, includes timestamps for each incident, necessitating analysis of temporal trends, seasonality, and patterns for effective prediction modeling. Class imbalance in crime datasets, where certain crime types are much less frequent than others, has been addressed to ensure proper model training and evaluation.

Experiments/Evaluation

Description of Testbed:

- **Dataset:** Free source from Kaggle, including past criminal incidents, socioeconomic factors, and law enforcement efforts.
- **Questions:** What are the key factors influencing crime rates? How accurately can we predict crime occurrences?

Detailed Description of Experiments:

- Conducted experiments using Decision Trees, Random Forests, Logistic Regression.
- Performed data preprocessing techniques such as random undersampling to mitigate class imbalance.
- Utilized feature selection methods to boost prediction accuracy.
- Evaluated model performance using metrics like accuracy, precision, recall, and F1-score.
- Visualized results using heatmaps, feature importance plots, and confusion matrices.

Results:

The results indicate that the Random Forest Classifier is the most effective model for predicting crime occurrences in this dataset, followed closely by the Decision Tree Classifier. The Logistic Regression Classifier, while still effective, did not perform as well as the other two models.

Conclusions and Discussion:

In conclusion, our Crime Analysis and Prediction System, leveraging advanced machine learning techniques and spatial-temporal analysis, provides valuable insights for law enforcement agencies. The system's goal is to enhance public safety and security by accurately predicting crime occurrences. Future work will focus on improving model accuracy, addressing potential biases, and exploring additional data sources for a more comprehensive analysis, ensuring ongoing refinement and effectiveness in crime prediction and prevention strategies.

Activities:

Activity	**week 1**	**week 2**	**week 3**	**week 4**
Project Research	Everyone			
Proposal Documentation	Everyone			
Data Collection		Tarun		
Data Pre-Processing		Afreen, Kousalya		
Data Splitting		Meher		
Model Selection		Tarun	Kousalya	
Model Training			Afreen, Tarun	
Model Testing and Evaluation			Kousalya	Meher
Data Visualization				Meher, Afreen
Final Report Documentation				Everyone

All team members contributed equally

Limitation and future work:

- The real-world complexity of crime dynamics and social factors presents ongoing challenges.
- Future work should incorporate additional data sources and explore advanced forecasting techniques to further enhance model robustness and accuracy.

References:

[1] Malik, K., Pandey, M., Khan, A., Srivastav, M., Gera, A., & Chauhan, S. S. (2024, March). Crime Prediction by Comparing Machine Learning and Deep Learning Algorithms. In 2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 215-219). IEEE.

[2] ToppiReddy, H. K. R., Saini, B., & Mahajan, G. (2018). Crime prediction & monitoring framework based on spatial analysis. *Procedia computer science*, *132*, 696-705.

[3] Hossain, S., Abtahee, A., Kashem, I., Hoque, M. M., & Sarker, I. H. (2020). Crime prediction using spatio-temporal data. In Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1 (pp. 277-289). Springer Singapore.

[4] Kumar, A., Verma, A., Shinde, G., Sukhdeve, Y., & Lal, N. (2020, February). Crime prediction using K-nearest neighboring algorithm. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (pp. 1-4). IEEE.

[5] Shamsuddin, N. H. M., Ali, N. A., & Alwee, R. (2017, May). An overview on crime predictions ethods. In *2017 6th ICT International Student Project Conference (ICT-ISPC)* (pp. 1-5). IEEE.

[6] Li, Z., Zhang, T., Yuan, Z., Wu, Z., & Du, Z. (2018, August). Spatio-temporal pattern analysis and prediction for urban crime. In *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)* (pp. 177-182). IEEE.

[7] Alves, L. G., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. Physica A: Statistical Mechanics and its Applications, 505, 435-443

[8] Elluri, L., Mandalapu, V., & Roy, N. (2019, June). Developing machine learning based predictive models for smart policing. In 2019 IEEE International Conference on Smart Computing (SMARTCOMP) (pp. 198-204). IEEE.

[9] Raza, D. M., & Victor, D. B. (2021, March). Data mining and region prediction based on crime using random forest. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 980-987). IEEE.

[10] Wang, B., Yin, P., Bertozzi, A. L., Brantingham, P. J., Osher, S. J., & Xin, J. (2019). Deep learning for real-time crime forecasting and its ternarization. Chinese Annals of Mathematics, Series B, 40(6), 949-966.

[11] Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Visual Computing for Industry, Biomedicine, and Art, 4(1), 9.

[12] Yuki, J. Q., Sakib, M. M. Q., Zamal, Z., Habibullah, K. M., & Das, A. K. (2019, July). Predicting crime using time and location data. In Proceedings of the 7th International Conference on Computer and Communications Management (pp. 124-128).

[13] Saraiva, M., Matijošaitienė, I., Mishra, S., & Amante, A. (2022). Crime prediction and monitoring in porto, portugal, using machine learning, spatial and text analytics. ISPRS International Journal of Geo-Information, 11(7), 400.

[14] Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. IEEE Access.