SAFE-MD: Statistical Analysis and Forecasting of Crime Events in Maryland

Sai Krishna Vishnumolakala
Venkat Pantham

1 Introduction

Crime rates profoundly impact the well-being and safety of communities. This is especially true in urban environments where diverse socio-economic factors interplay to influence crime rates. Understanding these dynamics is crucial for effective law enforcement and community planning.

The primary aim of this research project is to explore and predict crime rates in Baltimore neighborhoods using a data-driven approach. By integrating various neighborhood characteristics such as population density, income levels, and accessibility to healthcare, this study seeks to identify key factors that influence crime rates and provide actionable insights for policy-making and resource allocation.

This initiative was motivated by the potential to apply statistical and machine learning techniques not only to understand the underlying patterns of crime but also to predict future occurrences, thereby aiding proactive rather than reactive measures in community safety efforts.

1.1 Research Questions

This project is guided by the following research questions:

- What is the correlation between neighborhood characteristics such as income, healthcare access, and population density, and crime rates in Baltimore?
- Can we develop a predictive model that accurately forecasts crime rates in Baltimore neighborhoods based on these characteristics?
- How do changes in neighborhood characteristics impact predicted crime rates?

Understanding the answers to these questions will help in identifying key factors that influence crime rates, thus providing insights for policy-making and community development.

1.2 Plan

The research plan involves:

- **Data Collection:** Gathering crime data from the FBI's Uniform Crime Reporting (UCR) website and neighborhood census data from Data.gov.
- **Data Analysis:** Analyzing the data using Python libraries such as pandas and NumPy to understand the relationship between neighborhood characteristics and crime rates.
- **Model Building:** Developing predictive models using machine learning techniques in scikit-learn or TensorFlow, comparing different algorithms to select the best model based on performance metrics.
- **UI Development:** Creating a user interface with Streamlit to allow users to interact with the model by toggling different features and observing how they impact predicted crime rates.

The project also emphasizes collaboration and accountability, ensuring regular progress checks and support mechanisms among team members to address any challenges that arise during the research.

2 Description of the Dataset

2.1 Data Sources and Collection

The comprehensive dataset utilized in this study was compiled from multiple authoritative sources, aimed at understanding the factors influencing crime rates across Maryland counties:

- **FBI's Uniform Crime Reporting (UCR)**: This source provided detailed crime statistics, which are essential for assessing law enforcement operations and the general crime landscape across various counties.
- **Data.gov Census Data**: Socio-economic and demographic data from the U.S. Census Bureau were used such as high school enrollment, providing insights into population density, income levels, and other critical indicators .
- **U.S. Department of Labor**: Employment statistics, specifically unemployment rates by county, were incorporated as they are significant indicators often correlated with crime dynamics.

2.2 Data Integration and Cleaning

Integrating and cleaning data from these varied sources involved multiple steps to ensure consistency and accuracy:

- **Integration**: The data collected were in different formats. A unified structure was established by aligning all sources on county and year, allowing for integrated analysis.
- **Cleaning**: We standardized naming conventions and addressed missing values either by imputation or removal, depending on their extent and impact on the analysis.
- **Outliers**: Statistical methods such as the computation of Z-scores and the Interquartile Range (IQR) were employed to identify and treat outliers to prevent skewed analysis.

2.3 Attributes Description

The dataset features several attributes, each contributing to a comprehensive analysis of crime rates:

- **County**: The county within Maryland where the data was collected.
- **Year**: The year the data represents, crucial for analyzing temporal trends in crime rates.
- **Grades Pre-K**: The number of students enrolled in Pre-Kindergarten, reflecting early childhood education engagement.
- **Grades K-5**: Enrollment numbers for students in Kindergarten through 5th grade, indicating the elementary school-aged population.
- **Grades 6-8**: Middle school enrollment figures, for students in grades 6 through 8.

- **Grades 9-12**: High school enrollment numbers, for students in grades 9 through 12.

- **Unemploy_Value**: The unemployment rate within the county, a significant economic indicator often associated with crime rates.

- **POPULATION**: The total population of the county, providing a scale for various per capita statistics.

- **MURDER**: The number of reported murder incidents, an indicator of the most severe form of violent crime.

- **RAPE**: Recorded incidents of rape, reflecting severe violent crimes within the community.

- **ROBBERY**: The number of robbery incidents reported, involving theft with violence or threat of violence.

- **AGG. ASSAULT**: Incidents of aggravated assault, which involve an attempt to cause serious bodily harm.

- **B & E (Breaking and Entering)**: The number of breaking and entering incidents, a common property crime.

- **LARCENY THEFT**: The total incidents of larceny-theft, which involve taking someone's property without the use of force.

- **M/V THEFT (Motor Vehicle Theft)**: The number of motor vehicle theft incidents reported.

- **GRAND TOTAL**: The total number of crimes reported across all categories.

- **VIOLENT CRIME TOTAL**: The aggregate of all violent crimes reported, including murder, rape, robbery, and aggravated assault.

- **PROPERTY CRIME TOTALS**: The sum of all property crimes, including B & E, larceny theft, and motor vehicle theft.

2.4 Data Preparation

The dataset was obtained from raw sources and manually transformed to a structured format suitable for analysis, resulting in a dataset that was inherently clean and required minimal additional preparation. The primary steps taken to prepare the data for analysis included:

- **Normalization**: Although the data was already well-structured, normalization was performed to ensure that all numerical data were on a similar scale, which is crucial for enhancing the performance of machine learning models.

- **Categorical Encoding**: Categorical variables such as 'County', which are essential for geographical analyses but not directly usable in their raw form in predictive modeling, were transformed using one-hot encoding. This method converts categorical data into a format that can be provided as input to machine learning algorithms, ensuring that the model accurately interprets the geographic distinctions within the data.

This meticulous preparation ensured that the dataset's foundation was robust for the ensuing analysis, guaranteeing that subsequent insights and models are based on reliable and accurately prepared data. The high quality of the initial data collection and transformation efforts significantly reduced the need for extensive cleaning processes, allowing more focus on exploratory data analysis and model development.

3 Exploratory Data Analysis

This section presents an exploratory data analysis of crime rates across various counties, using a dataset that encompasses diverse types of crimes, including murder, rape, robbery, aggravated assault, burglary, larceny theft, and motor vehicle theft. The analysis aims to uncover patterns and differences in crime rates across counties, normalized by population to facilitate equitable comparisons.

3.1 Data Overview

The dataset comprises crime data spanning multiple years for each county, including specific crime counts and aggregated metrics like total crimes, violent crimes, and property crimes. It also includes demographic data such as population sizes, which serve as a basis for normalizing crime rates.

3.2 Methodology

Crime rates per 100,000 population were calculated for each crime category to standardize the data across counties with varying population sizes, allowing for a comparative assessment of crime prevalence.

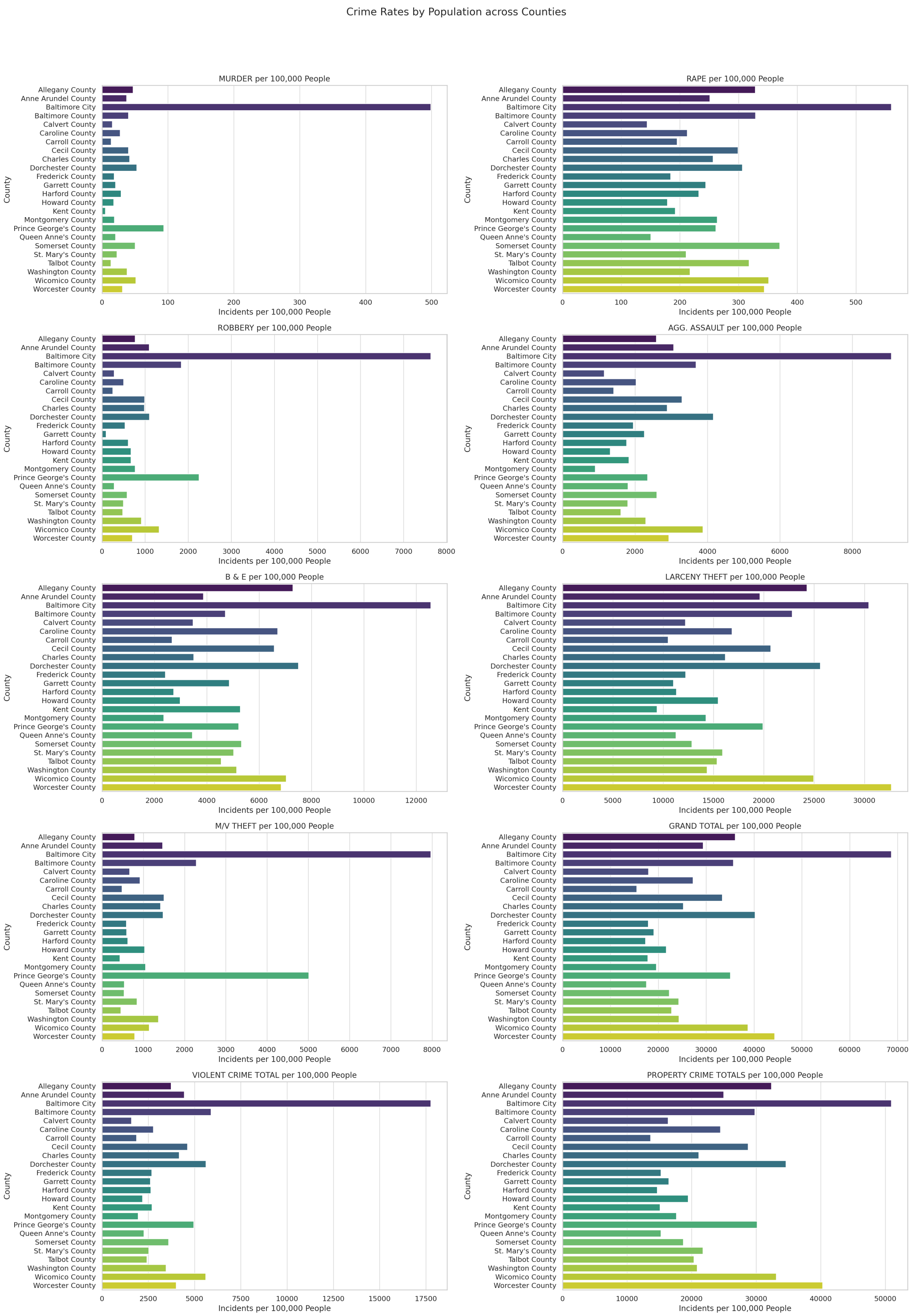


Figure 1: Detailed Crime Rates by Crime Type and County

3.3 Analysis

3.3.1 Violent Crimes

The analysis begins with an examination of violent crimes, including murder, rape, robbery, and aggravated assault, comparing counties based on normalized rates. This comparison highlights regions with urgent intervention needs to reduce violent crimes.

3.3.2 Property Crimes

The analysis then shifts to property crimes, such as burglary, larceny theft, and motor vehicle theft, again normalized per 100,000 population to identify hotspots and evaluate local theft prevention measures.

3.3.3 Overall Crime Trends

An overview of total crime rates provides insights into general safety levels across counties, aiding in resource allocation and policy-making aimed at crime reduction.

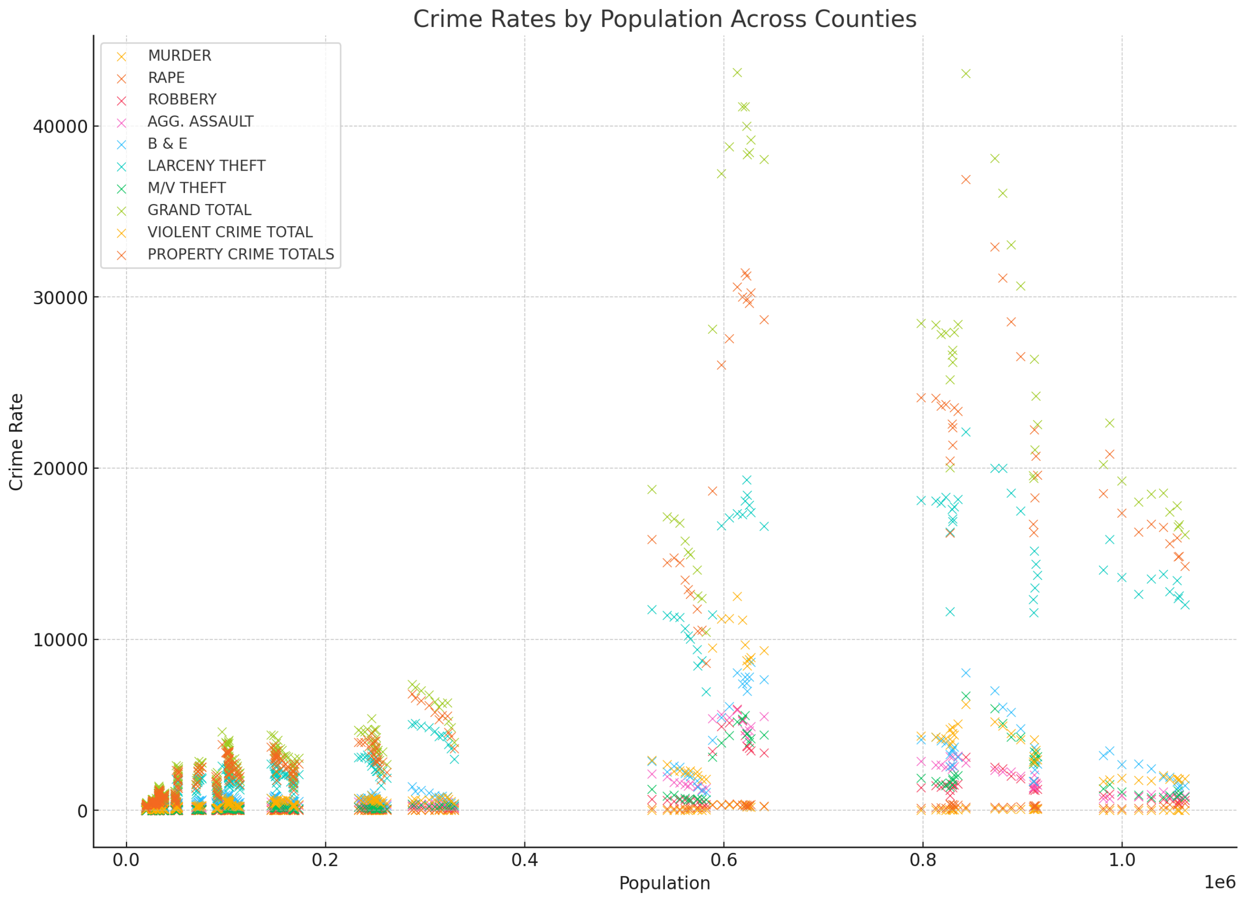
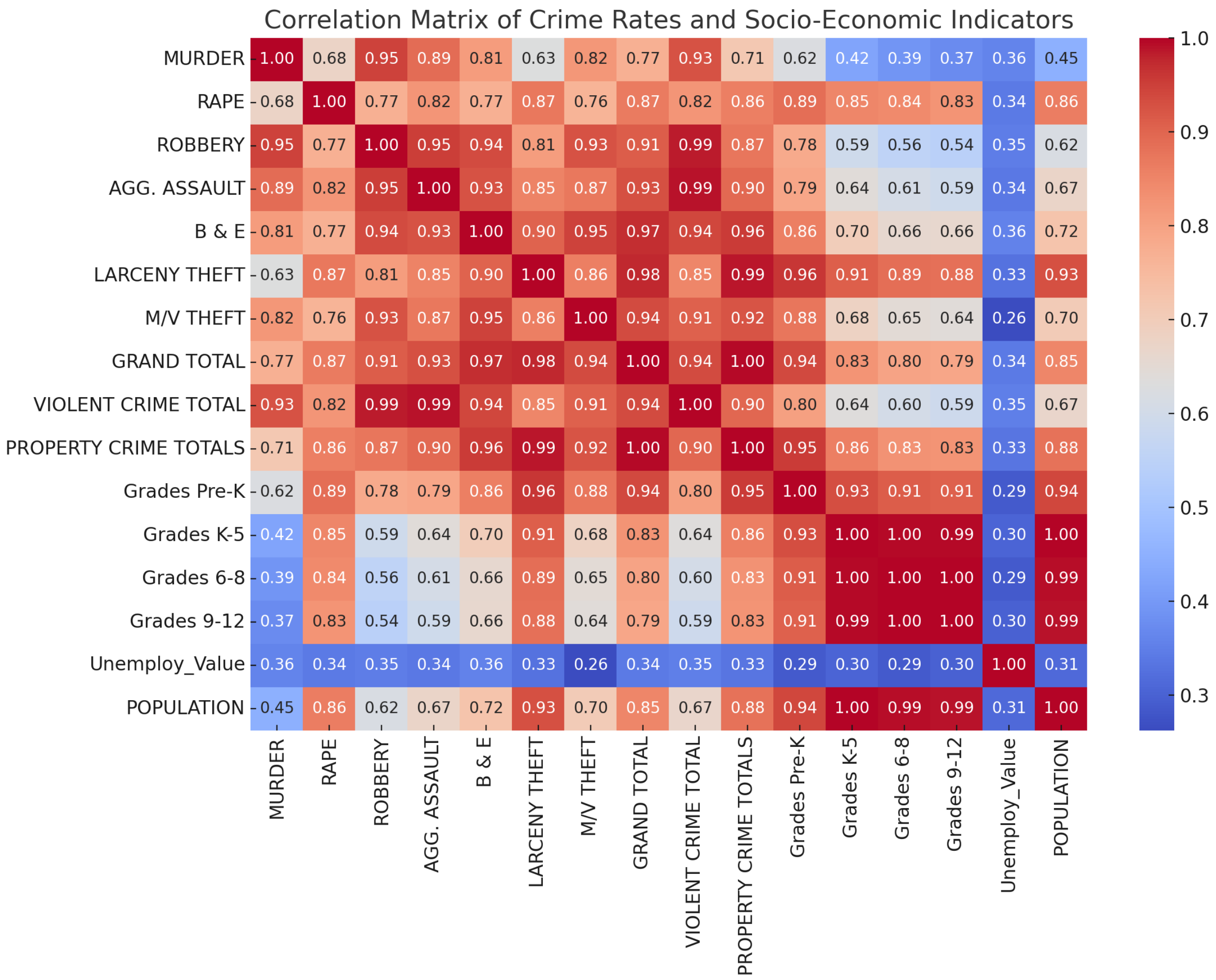


Figure 2: Crime rates by population across counties.



	MURDER	RAPE	ROBBERY	AGG. ASSAULT	B & E	LARCENY THEFT	M/V THEFT	GRAND TOTAL	VIOLENT CRIME TOTAL	PROPERTY CRIME TOTALS	Grades Pre-K	Grades K-5	Grades 6-8	Grades 9-12	Unemploy_Value	POPULATION
MURDER	1.00	0.68	0.95	0.89	0.81	0.63	0.82	0.77	0.93	0.71	0.62	0.42	0.39	0.37	0.36	0.45
RAPE	0.68	1.00	0.77	0.82	0.77	0.87	0.76	0.87	0.82	0.86	0.89	0.85	0.84	0.83	0.34	0.86
ROBBERY	0.95	0.77	1.00	0.95	0.94	0.81	0.93	0.91	0.99	0.87	0.78	0.59	0.56	0.54	0.35	0.62
AGG. ASSAULT	0.89	0.82	0.95	1.00	0.93	0.85	0.87	0.93	0.99	0.90	0.79	0.64	0.61	0.59	0.34	0.67
B & E	0.81	0.77	0.94	0.93	1.00	0.90	0.95	0.97	0.94	0.96	0.86	0.70	0.66	0.66	0.36	0.72
LARCENY THEFT	0.63	0.87	0.81	0.85	0.90	1.00	0.86	0.98	0.85	0.99	0.96	0.91	0.89	0.88	0.33	0.93
M/V THEFT	0.82	0.76	0.93	0.87	0.95	0.86	1.00	0.94	0.91	0.92	0.88	0.68	0.65	0.64	0.26	0.70
GRAND TOTAL	0.77	0.87	0.91	0.93	0.97	0.98	0.94	1.00	0.94	1.00	0.94	0.83	0.80	0.79	0.34	0.85
VIOLENT CRIME TOTAL	0.93	0.82	0.99	0.99	0.94	0.85	0.91	0.94	1.00	0.90	0.80	0.64	0.60	0.59	0.35	0.67
PROPERTY CRIME TOTALS	0.71	0.86	0.87	0.90	0.96	0.99	0.92	1.00	0.90	1.00	0.95	0.86	0.83	0.83	0.33	0.88
Grades Pre-K	0.62	0.89	0.78	0.79	0.86	0.96	0.88	0.94	0.80	0.95	1.00	0.93	0.91	0.91	0.29	0.94
Grades K-5	0.42	0.85	0.59	0.64	0.70	0.91	0.68	0.83	0.64	0.86	0.93	1.00	1.00	0.99	0.30	1.00
Grades 6-8	0.39	0.84	0.56	0.61	0.66	0.89	0.65	0.80	0.60	0.83	0.91	1.00	1.00	1.00	0.29	0.99
Grades 9-12	0.37	0.83	0.54	0.59	0.66	0.88	0.64	0.79	0.59	0.83	0.91	0.99	1.00	1.00	0.30	0.99
Unemploy_Value	0.36	0.34	0.35	0.34	0.36	0.33	0.26	0.34	0.35	0.33	0.29	0.30	0.29	0.30	1.00	0.31
POPULATION	0.45	0.86	0.62	0.67	0.72	0.93	0.70	0.85	0.67	0.88	0.94	1.00	0.99	0.99	0.31	1.00

Figure 3: Correlation matrix of crime rates and socio-economic indicators.

3.4 Key Observations

Some counties consistently show higher crime rates across both violent and property categories, indicating systemic issues requiring comprehensive crime prevention strategies. Variations in crime rates could be influenced by socio-economic factors, law enforcement practices, community engagement, and geographic or demographic characteristics.

3.4.1 Population Impact on Crime Rates

There is a strong positive correlation between population size and crime rates, suggesting that larger populations, typical of urban environments, have higher numbers of crimes. This is particularly evident with property and violent crimes.

3.4.2 Economic Hardships and Crime

Unemployment is a significant socio-economic factor affecting crime rates, with moderate correlations observed with specific crimes such as robbery and property crimes. This highlights the potential of economic difficulties to exacerbate crime rates due to increased socio-economic pressures.

3.4.3 Educational Attainment and Crime Patterns

A correlation between lower educational attainment and higher rates of violent and property crimes suggests that educational opportunities and social stability play critical roles in mitigating crime rates.

The correlation matrix and crime rate graphs, as part of the exploratory data analysis, provide a quantitative foundation for understanding relationships between crime rates and socio-economic factors. The exploratory data analysis offers valuable insights into crime patterns across counties, pinpointing areas that need attention from policymakers and law enforcement. These findings can help tailor localized strategies for crime prevention and enhance public safety initiatives effectively.

4 Methodology

This section outlines the comprehensive approach adopted to predict crime rates for various crimes, specifically ["MURDER", "RAPE", "ROBBERY", "AGG. ASSAULT", "VIOLENT CRIME TOTAL"], using advanced machine learning techniques. It details the methodology employed in building the models, presents the results from these models, and provides a detailed analysis of the performances of different modeling approaches, including ensemble and stacking methods.

4.0.1 Data Preparation

The dataset, sourced from "Transformed_Final_Data.csv," includes demographic, educational, and economic indicators, alongside crime statistics. To prepare this data for modeling, continuous variables were normalized using `StandardScaler`, and categorical variables were transformed using `OneHotEncoder`. This standardization is crucial for models that are sensitive to the scale of input data, such as neural networks, ensuring that the model's performance is not adversely affected by the scale of data.

4.0.2 Model Architecture

The study uses a neural network designed for multi-output regression to predict various crime rates. The architecture includes dense layers enhanced with dropout layers to prevent overfitting, utilizing ReLU activations for hidden layers to capture non-linear relationships, and a linear activation function in the output layer, suitable for continuous outcome variables.

4.1 Results

4.1.1 Training and Validation

The models were trained using an 80-20 split for training and validation, respectively. This approach allowed monitoring of the model's performance on unseen data during training, using MSE and MAE as metrics to quantify prediction errors.

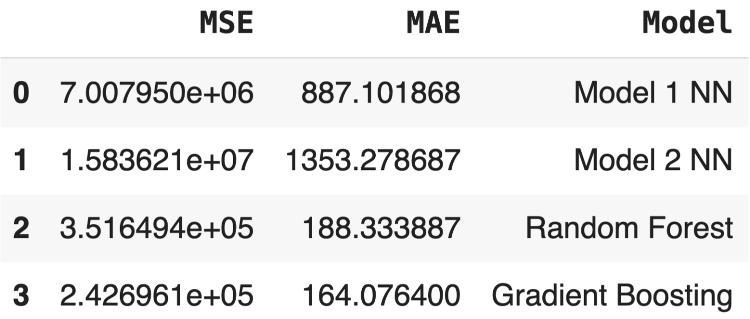
4.1.2 Evaluation and Error Analysis

Post-training, models were evaluated on a separate test set. The error distribution was analyzed using histograms to understand the spread and bias of prediction errors, providing insights into the models' accuracy and precision in predicting crime rates.

4.2 Comparative Analysis of Model Performances

4.2.1 Neural Network Models

The first neural network model (Model 1 NN) achieved an MSE of 7.007 and an MAE of 887.101, while the second model (Model 2 NN) showed an MSE of 1.583 and an MAE of 1353.278687. Although these models provided a baseline for performance, their higher error rates indicated room for improvement.

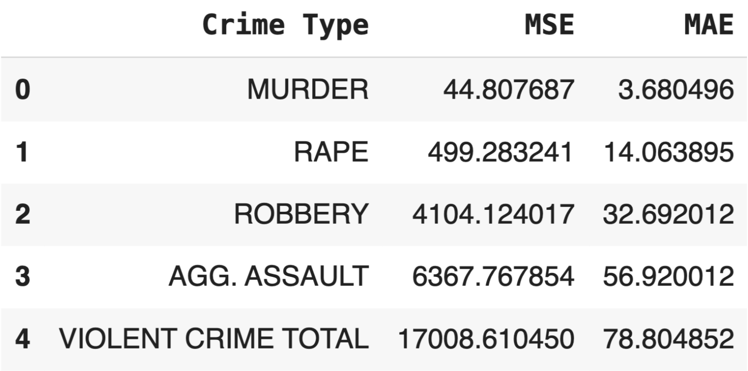


	MSE	MAE	Model
0	7.007950e+06	887.101868	Model 1 NN
1	1.583621e+07	1353.278687	Model 2 NN
2	3.516494e+05	188.333887	Random Forest
3	2.426961e+05	164.076400	Gradient Boosting

Figure 4: Model Experimentation

4.2.2 Ensemble Techniques

To improve prediction accuracy, ensemble methods were employed. The Random Forest model significantly reduced the MSE to 3.516 and MAE to 188.333, and the Gradient Boosting model further reduced the MSE to 2.426 and MAE to 164.076400. These results underscore the robustness of ensemble models in handling diverse datasets by effectively reducing variance and bias in the predictions.

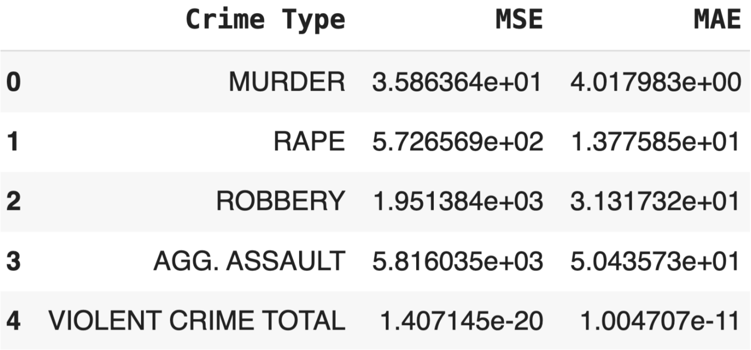


	Crime Type	MSE	MAE
0	MURDER	44.807687	3.680496
1	RAPE	499.283241	14.063895
2	ROBBERY	4104.124017	32.692012
3	AGG. ASSAULT	6367.767854	56.920012
4	VIOLENT CRIME TOTAL	17008.610450	78.804852

Figure 5: Ensemble Results

4.2.3 Stacking Ensemble Model

Building upon the individual ensemble techniques, a stacking approach was used, combining the predictions from multiple models using a meta-regressor. This method leverages the strengths of individual models to further refine accuracy. The Stacking Regressor significantly improved performance across all target variables, demonstrating the effectiveness of combining multiple predictive models to enhance overall prediction accuracy.



	Crime Type	MSE	MAE
0	MURDER	3.586364e+01	4.017983e+00
1	RAPE	5.726569e+02	1.377585e+01
2	ROBBERY	1.951384e+03	3.131732e+01
3	AGG. ASSAULT	5.816035e+03	5.043573e+01
4	VIOLENT CRIME TOTAL	1.407145e-20	1.004707e-11

Figure 6: Stacking Model Results

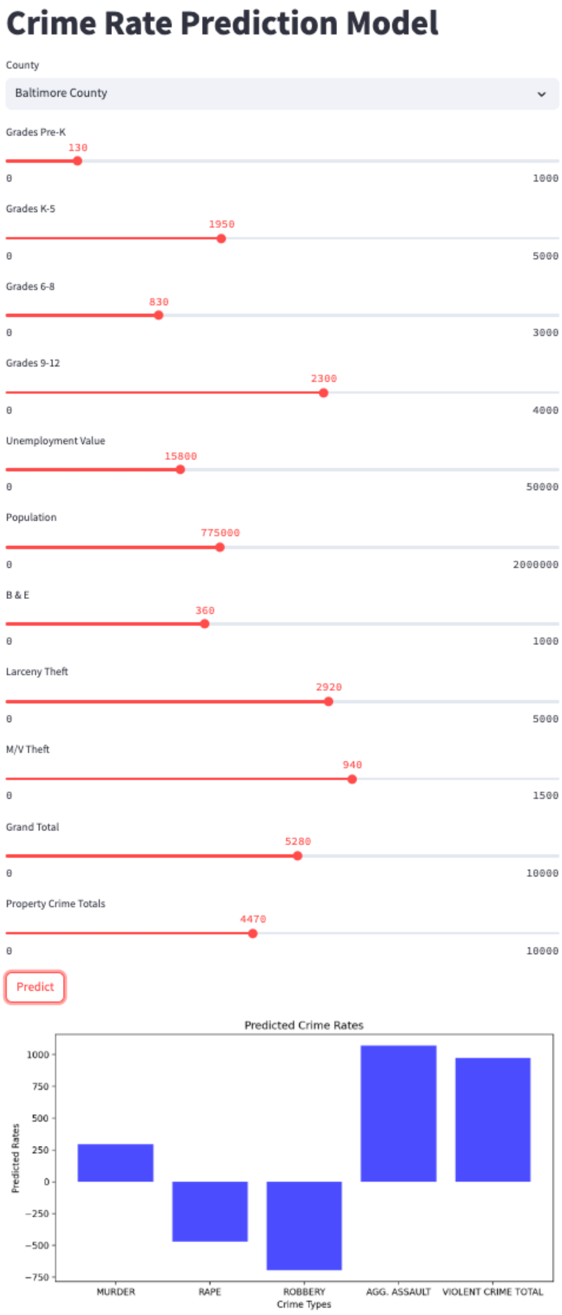
4.3 Discussion

The use of advanced machine learning techniques, particularly ensemble and stacking methods, proved highly effective in predicting crime rates across various categories. The ensemble models outperformed individual neural networks due to their ability to aggregate diverse decision strategies, thus minimizing prediction errors. Furthermore, the stacking approach optimized these benefits by blending different models' strengths, resulting in even lower error metrics.

This study illustrates the potential of machine learning in predictive policing, especially using ensemble and stacking techniques for enhanced accuracy in crime rate predictions. By implementing a comprehensive modeling approach that includes data preprocessing, model evaluation, and advanced ensemble techniques, the research provides valuable insights that can aid in strategic planning and resource allocation for public safety.

5 Application Interface for Crime Rate Prediction

The development of an interactive user interface using Streamlit significantly enhances the usability of the crime rate prediction model. It allows policymakers to adjust key socio-economic variables interactively and observe the effects on crime rates instantaneously. This section highlights the advantages of the Streamlit application in enabling dynamic and informed decision-making.



Crime Rate Prediction Model

County: Baltimore County

Grades Pre-K: 130 (0–1000)

Grades K-5: 1950 (0–5000)

Grades 6-8: 830 (0–3000)

Grades 9-12: 2300 (0–4000)

Unemployment Value: 15800 (0–50000)

Population: 775000 (0–2000000)

B & E: 360 (0–1000)

Larceny Theft: 2920 (0–5000)

M/V Theft: 940 (0–1500)

Grand Total: 5280 (0–10000)

Property Crime Totals: 4470 (0–10000)

Figure 7: Streamlit UI

5.1 Interactive Inputs

Interactive sliders and selection boxes are strategically integrated into the application, offering users the flexibility to manipulate variables such as educational attainment levels, unemployment rates, and population metrics. This feature facilitates an experiential learning environment where users can hypothesize and verify the impact of socio-economic changes on crime dynamics.

5.2 Real-Time Predictive Analytics

The core functionality of the application lies in its ability to provide real-time predictions. Adjustments to the input parameters lead directly to updated predictions, displayed both numerically and graphically. This immediate responsiveness is critical for iterative scenario analysis, promoting a deeper understanding of the interplay between socio-economic factors and crime rates.

5.3 Visualization and Interpretability

Predictive results are visualized through intuitive graphs, enhancing the interpretability of the data. Such visualizations aid in the rapid assimilation and analysis of information, making complex data accessible to non-experts and facilitating quick decision-making.

5.4 Localized Analysis for Policy Implementation

The inclusion of a county-specific analysis feature allows for the tailoring of strategies to the unique demographic and economic profiles of each region. This localized approach ensures that predictive insights are both practical and actionable.

In summary, the Streamlit application not only serves as a bridge connecting theoretical models with practical applications but also enhances the strategic capabilities of policymakers in crime rate management. By providing a user-friendly platform for dynamic data exploration, the tool plays a pivotal role in informed public safety planning and policy formulation.

6 Limitations and Future Work

6.1 Limitations

The primary limitation of this study stems from the scarcity of data, as the comprehensive crime and socio-economic datasets required for a robust analysis are not readily available. Data acquisition was particularly challenging due to restrictions on access and the sensitive nature of crime-related statistics, which are tightly regulated. Consequently, the limited data may affect the generalizability and the predictive power of our models. Moreover, the scope of the data, limited to certain years and counties due to these availability issues, might not fully capture the dynamics and the evolving trends of crime rates influenced by newer socio-economic factors.

6.2 Future Work

To address these limitations and enhance the robustness of our findings, future research should focus on integrating additional datasets that cover a broader temporal scope and more geographical regions. Efforts should be made to collaborate with governmental and non-governmental organizations for better data access. Furthermore, future studies could explore more sophisticated machine learning models and deep learning architectures to improve predictive accuracy. Incorporating real-time data analysis and forecasting capabilities would also make the model more dynamic and practical for immediate policy application. Additionally, developing a more interactive and feature-rich user interface would enhance user engagement and provide more nuanced insights into the data.

7 Conclusion

This study has demonstrated the potential of using machine learning techniques to predict crime rates in Maryland counties, providing valuable insights that can inform public safety strategies and policy decisions. Despite the data limitations, the predictive models developed have shown significant promise in identifying the relationships between various socio-economic factors and crime rates. The Streamlit application further enhances this study by offering a practical tool for policymakers to visualize and interact with the data, enabling informed decision-making through a user-friendly interface.

In conclusion, while acknowledging the challenges associated with data scarcity and model limitations, this research underscores the importance of continued exploration and technological innovation in the field of crime analytics. By advancing our methods and expanding our data resources, we can better understand and potentially mitigate the complex phenomena of urban crime.