Decoding Danger – A Predictive Analysis of the Communities and Crime (UCI) Dataset

Student Name: Deepika Gopi

Student ID: 24099803

Date: 06/01/2025

Course: Data Mining and Discovery

Context: Final Project

Project GitHub Repository: https://github.com/Deepi-boobi02/communities-crime-prediction.git

Contents

1. Executive Summary

The goal of this project was to analyse how socio-economic factors (such as income, education, and family structure) influence violent crime rates in U.S. communities. Using the Communities and Crime dataset from the UCI Machine Learning Repository, I cleaned raw data, performed exploratory data analysis, and built machine learning models to predict crime rates.

Key Finding: My analysis revealed that family structure—specifically the percentage of children living in two-parent households—is the single strongest predictor of community safety, significantly outranking economic variables like median rent or employment rates.

2. Methodology

Data Source

- Dataset Name: Communities and Crime (UCI)
- Source: UCI Machine Learning Repository
- Description: The dataset combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 LEMAS survey, and crime data from the 1995 FBI UCR.

Data Cleaning

The original dataset contained 128 variables and 1,994 communities.

- **Missing Data:** I identified that 27 columns related to police department reporting (e.g., LemasSwornFT) were missing over 80% of their data. These columns were removed to prevent model errors.
- **Imputation:** Minor missing values in demographic columns were filled using the column mean.
- **Final Shape:** The clean dataset used for modelling contained 1,994 rows and 96 feature columns.

Exploratory Data Analysis (EDA)

I examined the distribution of the target variable, ViolentCrimesPerPop. The data is right-skewed, indicating that while most communities have low-to-moderate crime rates, a small subset of communities' experiences disproportionately high crime.

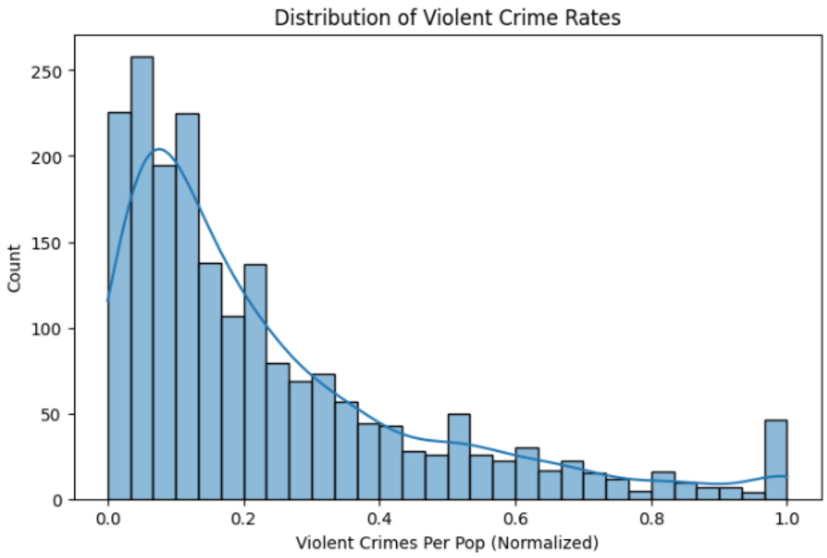

Figure 1: Distribution of Violent Crime Rates (Normalized).

3. Model Performance

I trained two separate machine learning models to predict the crime rate.

Model 1: Linear Regression (Baseline)

- **RMSE:** 0.1337
- **R^2 Score:** 0.6269
- **Interpretation:** This model performed the best, explaining ~63% of the variance in crime rates. This suggests the relationships in the data are largely linear.

Model 2: Random Forest Regressor

- **RMSE:** 0.1373
- **R^2 Score:** 0.6062
- **Interpretation:** The Random Forest model provided valuable feature importance data but slightly overfitted the noise in the training set compared to the linear model.

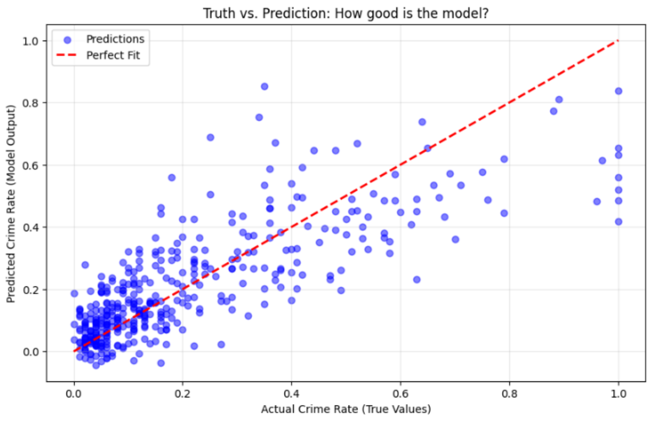

Figure 2: Actual vs. Predicted Crime Rates. The proximity of the blue dots to the red dashed line demonstrates the model's accuracy.

4. Key Insights: What Drives Crime?

Using the Random Forest model, I extracted the "Feature Importance" to understand which specific variables drove the predictions.

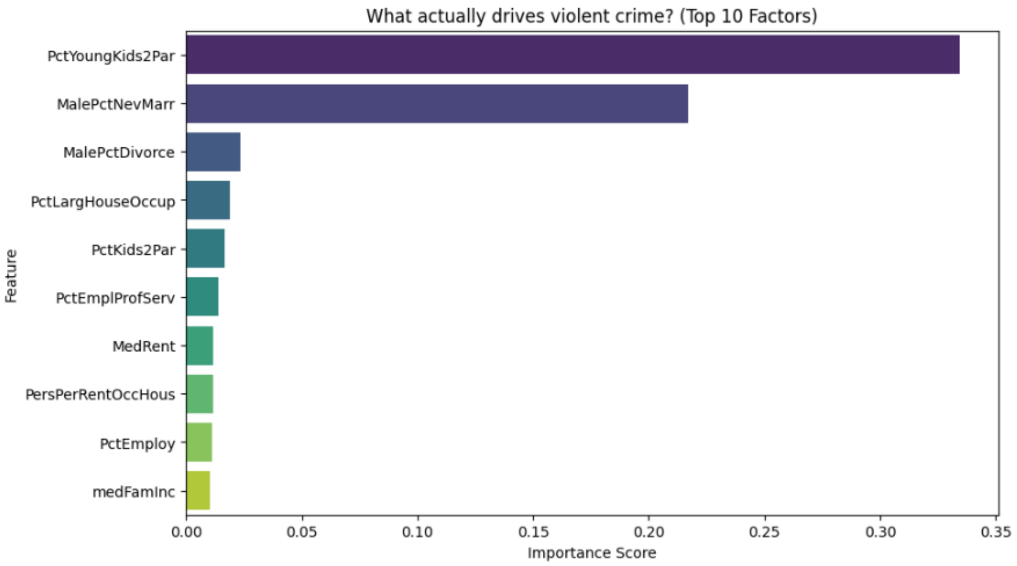

Figure 3: Top 10 Factors Influencing Violent Crime.

Analysis of Drivers:

Family Stability: The top predictor was PctYoungKids2Par (Percentage of kids in two-parent households). The model indicates a strong correlation between family stability and lower crime rates.

Marital Status: The second strongest predictor was MalePctNevMarr (Percentage of males never married), reinforcing the family structure trend.

Economic Impact: While medIncome (Median Income) was a factor, it ranked lower than family dynamics, suggesting that economic wealth alone does not guarantee safety.

5. Conclusion

I successfully built a machine learning pipeline using the **Communities and Crime (UCI)** dataset that predicts community crime rates with **~63% accuracy**. The project highlights that social support structures (family units) may be more critical predictive signals than purely economic indicators.

6. References

1. **Dataset:** Redmond, M. A., & Baveja, A. (2002). *Communities and Crime Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

2. **Original Research:** Redmond, M. A., & Baveja, A. (2002). A Data-Driven Software Tool for Enabling Cooperative Information Sharing Among Police Departments. *European Journal of Operational Research*, 141(3), 660-678.

3. **Tools Used:** Python (pandas, scikit-learn, matplotlib, seaborn).