Author: **Tanmay, Dhruv, Kunal and Soham**
Date: **30 November 2024**

CSP 571: DPA – Project Report

Title: Chicago Crime Analysis and Predictive Modeling

Link: https://github.com/tanmayypramanick/Chicago-Crime-Analysis-and-Predictive-Modeling

1. Introduction

Every city faces unique challenges when it comes to crime, and Chicago is no different. This project dives into **Chicago's crime data** to uncover patterns and **predict the type of crime** based on **historical trends**. By leveraging machine learning, the goal is to analyse and forecast crime effectively, supporting public safety efforts and demonstrating how data science can be a powerful tool to tackle real-world urban challenges.

Why This Matters:

- Helps make communities safer by identifying areas prone to specific crimes.
- Shows how machine learning and data science can address urban problems in innovative ways.

2. Data Overview and Preparation

Dataset

The data for this project was sourced from the **Chicago Data Portal** using the **Socrata API**. Due to the large dataset size (over 1.9 GB), we worked with a sampled dataset of **100,000 records**. Each record provided valuable details like crime type, location, date, and additional features like FBI codes and community areas.

Preprocessing Steps

To prepare the dataset for analysis:

1. **Filtered Rare Classes**: Crime types with very few occurrences were removed to ensure consistent model performance.

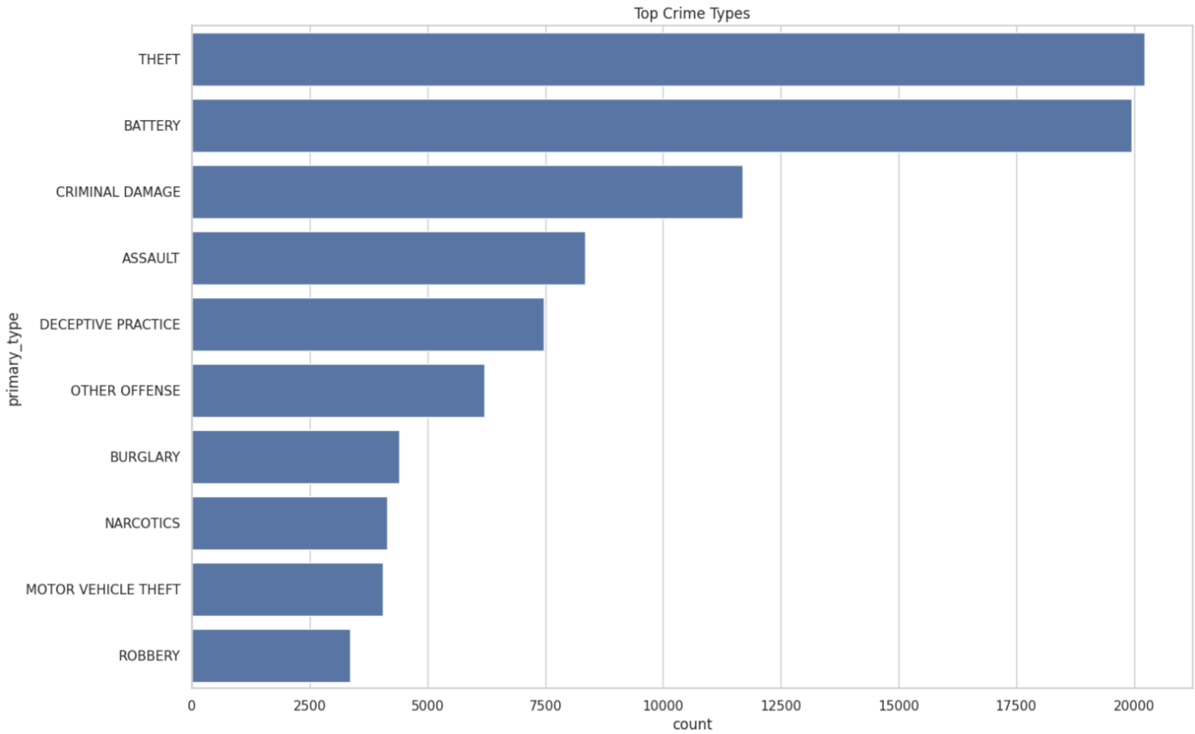
2. **Standardized Numeric Features**: Features like coordinates and community area codes were scaled using **StandardScaler**.
3. **Encoded Categorical Features**: Features such as crime descriptions were encoded with **OneHotEncoder**, resulting in a dataset with **5,000 rows and 20,003 features**.

3. Data Visualization and Exploration

To better understand the dataset and its patterns, we explored the data using visual tools:

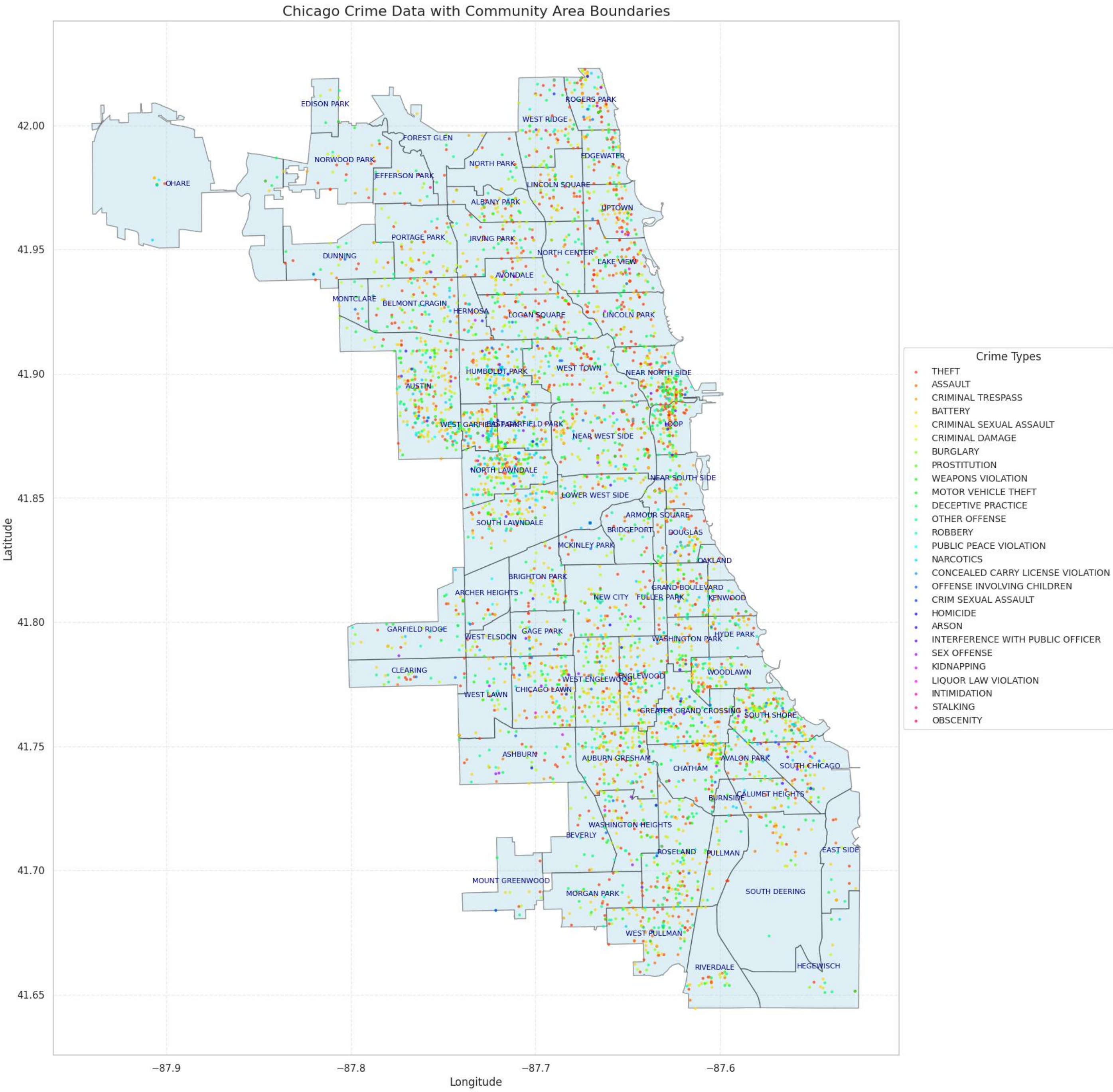
Key Insights

1. **Top Crime Types**: Theft and battery were the most frequently reported crimes, as shown in a bar chart.



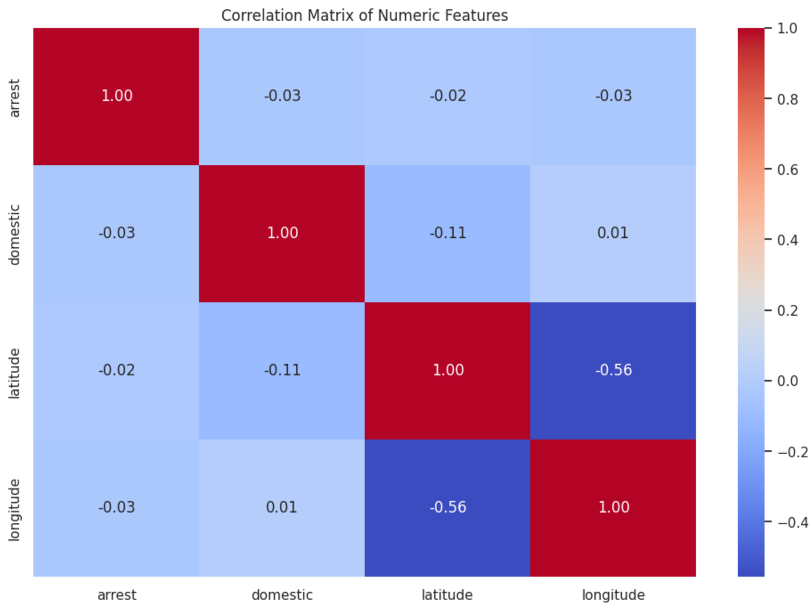
2. **Geospatial Analysis**: Crimes were mapped to Chicago's community boundaries, revealing clusters of criminal activity in specific neighborhoods.

Author: **Tanmay, Dhruv, Kunal and Soham**
Date: **30 November 2024**



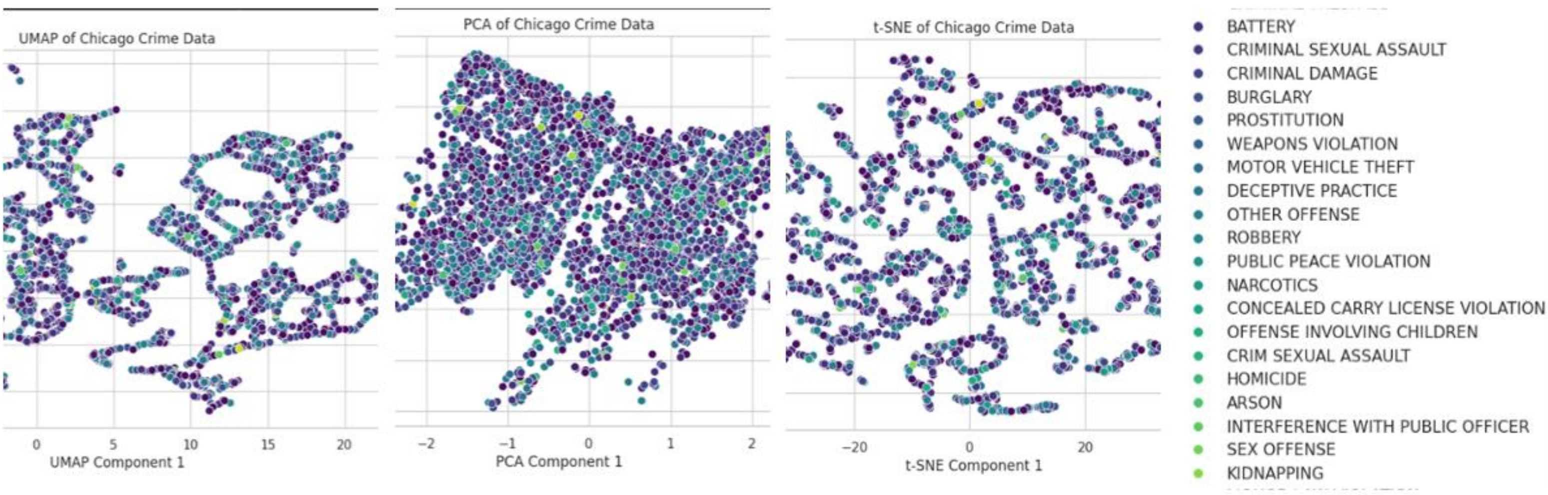
Visualization Highlights

- A **correlation heatmap** helped identify relationships between community areas and crime types.
- A **crime map overlay** showed which areas were most affected, providing actionable insights for public safety planning.



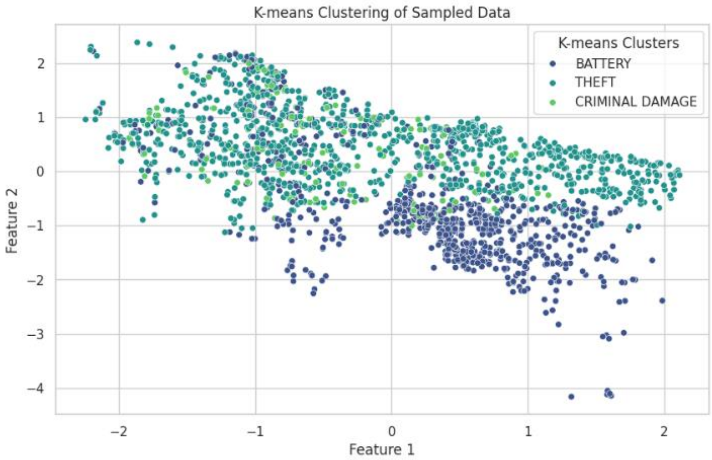
Key Steps:

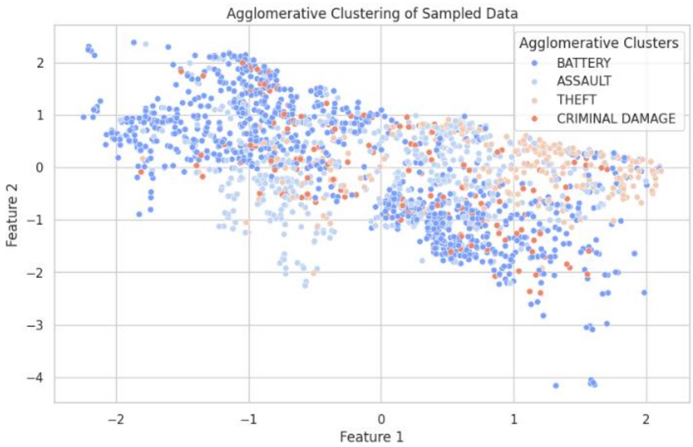
- **Dimensionality Reduction:** Applied techniques such as t-SNE, UMAP, and PCA to visualize high-dimensional data.
    - **PCA**: Highlighted variance across features to reduce noise.
    - **t-SNE/UMAP**: Clustered data into visually interpretable segments to detect similarities.

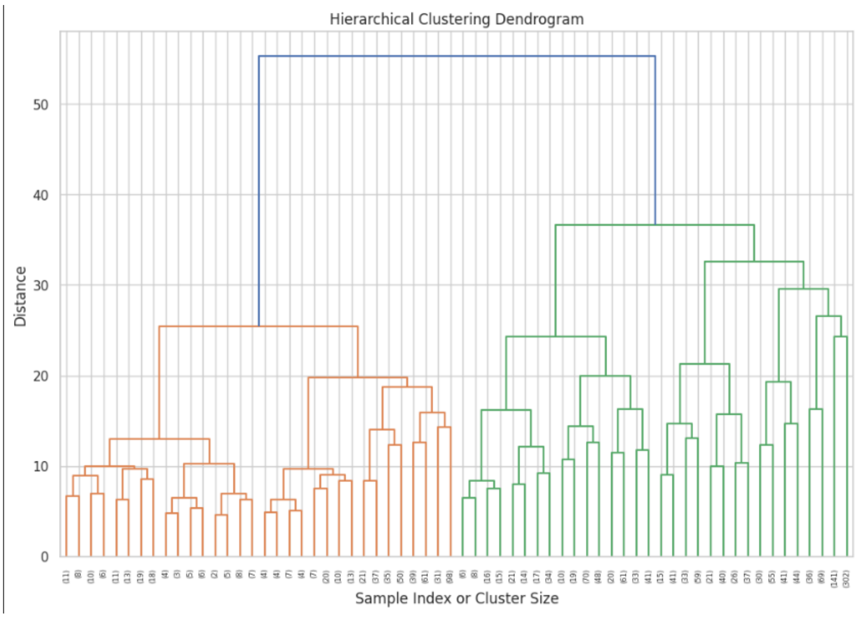


Author: **Tanmay, Dhruv, Kunal and Soham**
Date: **30 November 2024**

- **Unsupervised Learning Techniques:**
  - **K-Means Clustering:** Identified crime hotspots based on crime type and location.
  - **Hierarchical Clustering:** Visualized nested groupings of community areas with high crime rates.
  - **Agglomerative Clustering:** Explored subgroup relationships among crime types.







4. Methodology

4.1. Model Selection and Cross-Validation

We used the **Random Forest Classifier** as our baseline model because of its accuracy and ability to handle high-dimensional data.

- **Cross-Validation**: Used **StratifiedKFold** to ensure balanced splits of the data, achieving a baseline accuracy of **99%**.

4.2. Hyperparameter Tuning

To further improve the model, we optimized key parameters such as n_estimators, max_depth, and min_samples_split using **GridSearchCV**.

- **Best Model Parameters**:
  - max_depth: None
  - n_estimators: 50
  - min_samples_split: 5
- **Validation Accuracy**: 99.5%.

5. Results and Insights

Performance

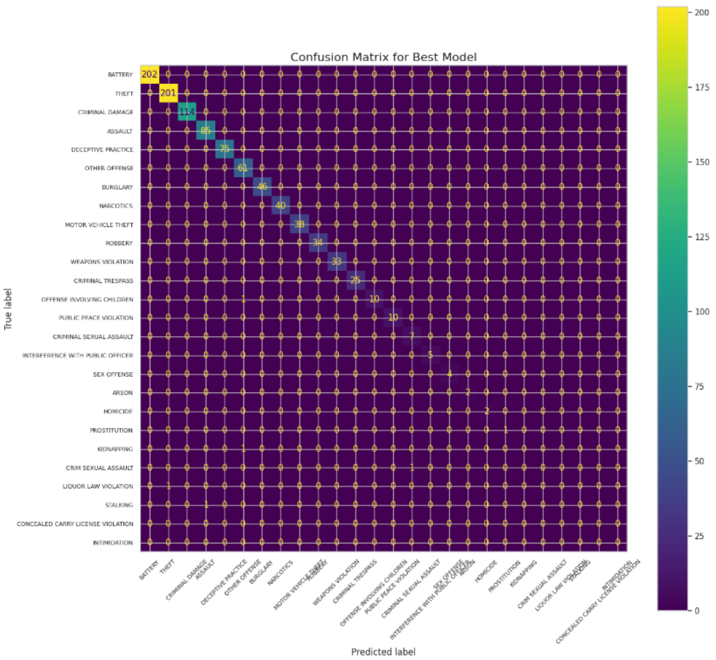
- **Overall Accuracy**: Achieved 99% accuracy across multiple experiments.
- **Misclassifications**: Out of 1,000 validation samples, only 5 were misclassified. These typically occurred in closely related crime categories like **Stalking** being predicted as **Assault**.

Feature Importance

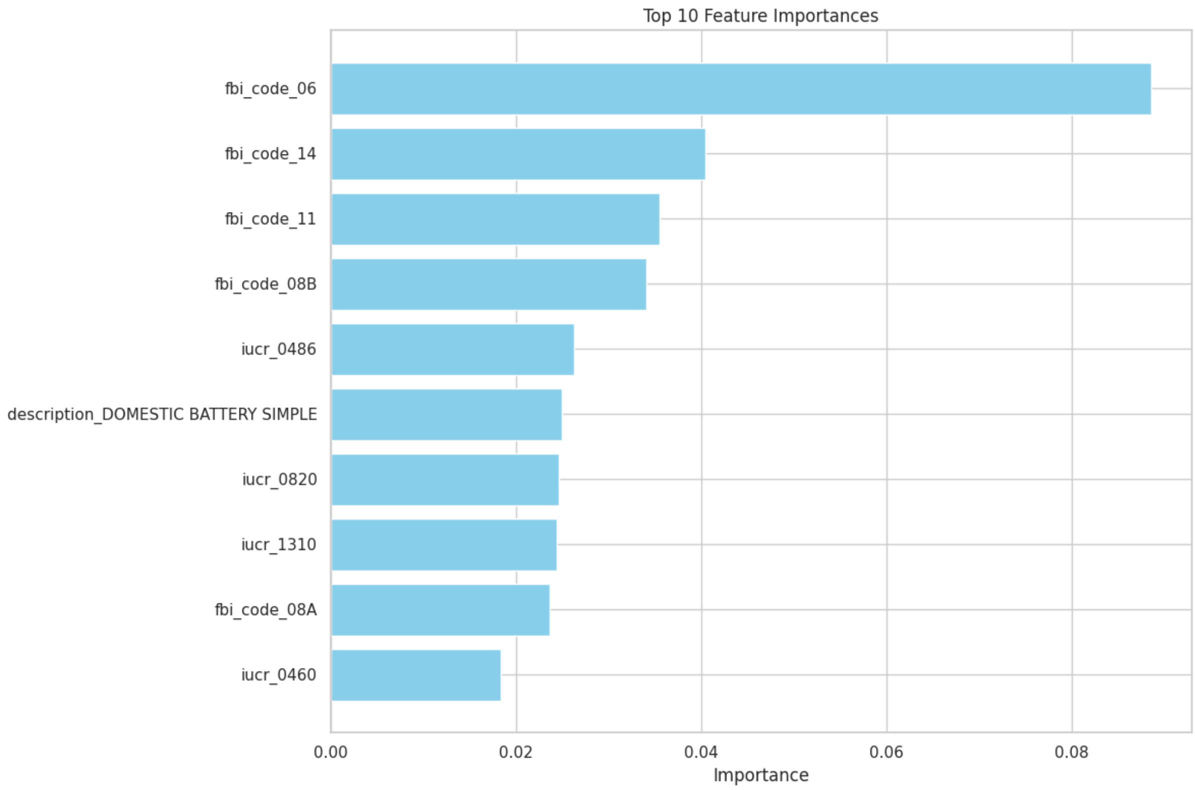
Analyzing feature importance revealed that FBI codes and community area information were the most influential factors in predicting crime types.

Visualizations

1. A **confusion matrix** showed excellent model performance, with most crimes accurately predicted.



2. A **feature importance chart** highlighted key predictors, helping interpret the model results.



6. Experiments and Improvements

Author: **Tanmay, Dhruv, Kunal and Soham**
Date: **30 November 2024**

Experiment 1: Feature Selection

We used Random Forest to identify and retain the most important features. This reduced noise and improved interpretability while maintaining high accuracy (**99.4%**).

```
Validation Accuracy with Feature Selection: 0.994
Classification Report with Feature Selection:
                                  precision    recall  f1-score   support

                           ARSON       1.00      1.00      1.00         2
                         ASSAULT       0.99      1.00      0.99        85
                         BATTERY       1.00      1.00      1.00       202
                        BURGLARY       1.00      1.00      1.00        46
             CRIM SEXUAL ASSAULT       0.00      0.00      0.00         1
                 CRIMINAL DAMAGE       1.00      1.00      1.00       114
         CRIMINAL SEXUAL ASSAULT       0.88      1.00      0.93         7
               CRIMINAL TRESPASS       1.00      1.00      1.00        25
              DECEPTIVE PRACTICE       1.00      1.00      1.00        75
                        HOMICIDE       1.00      1.00      1.00         2
INTERFERENCE WITH PUBLIC OFFICER       1.00      0.80      0.89         5
                      KIDNAPPING       0.00      0.00      0.00         1
             LIQUOR LAW VIOLATION       1.00      1.00      1.00         1
             MOTOR VEHICLE THEFT       1.00      1.00      1.00        38
                       NARCOTICS       1.00      1.00      1.00        40
      OFFENSE INVOLVING CHILDREN       0.91      0.91      0.91        11
                   OTHER OFFENSE       0.97      1.00      0.98        61
                    PROSTITUTION       1.00      1.00      1.00         1
           PUBLIC PEACE VIOLATION       0.91      1.00      0.95        10
                         ROBBERY       1.00      1.00      1.00        34
                     SEX OFFENSE       1.00      0.75      0.86         4
...
                        accuracy                           0.99      1000
                       macro avg       0.86      0.85      0.85      1000
                    weighted avg       0.99      0.99      0.99      1000
```

Experiment 2: Regularization

To address potential overfitting:

1. **Lasso (L1)** selected only the most relevant features by shrinking irrelevant ones to zero.
2. **Ridge (L2)** reduced overfitting while preserving all features.

- Both models achieved **100% validation accuracy**, showing the effectiveness of regularization techniques.

```
Validation Accuracy (Lasso): 1.00
Classification Report (Lasso):
                               precision    recall  f1-score   support

                        ARSON       1.00      1.00      1.00         2
                      ASSAULT       0.99      1.00      0.99        85
                      BATTERY       1.00      1.00      1.00       202
                     BURGLARY       1.00      1.00      1.00        46
          CRIM SEXUAL ASSAULT       0.00      0.00      0.00         1
              CRIMINAL DAMAGE       1.00      1.00      1.00       114
      CRIMINAL SEXUAL ASSAULT       0.88      1.00      0.93         7
            CRIMINAL TRESPASS       1.00      1.00      1.00        25
           DECEPTIVE PRACTICE       1.00      1.00      1.00        75
                     HOMICIDE       1.00      1.00      1.00         2
INTERFERENCE WITH PUBLIC OFFICER       1.00      1.00      1.00         5
                   KIDNAPPING       0.00      0.00      0.00         1
         LIQUOR LAW VIOLATION       1.00      1.00      1.00         1
          MOTOR VEHICLE THEFT       1.00      1.00      1.00        38
                    NARCOTICS       1.00      1.00      1.00        40
    OFFENSE INVOLVING CHILDREN       1.00      0.91      0.95        11
                OTHER OFFENSE       0.97      1.00      0.98        61
                 PROSTITUTION       1.00      1.00      1.00         1
         PUBLIC PEACE VIOLATION       1.00      1.00      1.00        10
                      ROBBERY       1.00      1.00      1.00        34
                  SEX OFFENSE       1.00      1.00      1.00         4
...
                     accuracy                           1.00      1000
                    macro avg       0.87      0.87      0.87      1000
                 weighted avg       0.99      1.00      0.99      1000
```

7. Challenges and Future Work

Challenges

- **High Dimensionality**: Handling over 20,000 features required careful preprocessing and feature selection.
- **Class Imbalances**: Rare crimes like **Stalking** or **Kidnapping** had very few examples, impacting model predictions for those categories.

Future Directions

- **Real-Time Predictions**: Extend the model to analyze incoming crime reports in real time.
- **Temporal Insights**: Incorporate time-based trends (e.g., day vs. night crimes) to improve accuracy.

- **Deep Learning Models**: Experiment with neural networks for further performance improvements.

8. Conclusion

This project demonstrated how machine learning can effectively analyze and predict crime patterns in a large urban dataset. The results not only provide actionable insights for public safety but also highlight the power of data science in addressing real-world challenges.

By leveraging crime data, geospatial analysis, and machine learning, we take a step closer to smarter, safer cities.