Crime Rate Prediction Project Report

Team Members

Rana Walid Bishr - 2305288

Abdallah Emad Ibrahim - 2305292

Omar Abdelhady Mohamed - 2305480

Youssef Ashraf Abd Elmohsen - 2305102

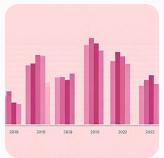
Omar Nabil Gaafar - 2305494

Abd Elrahman Adel Shahin - 2305432

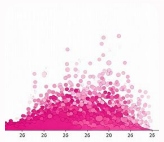
Project Overview

Analyzing San Francisco crime data to understand patterns.

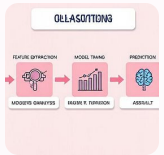
Cleaning and preprocessing the dataset for accuracy.



Exploring and visualizing crime patterns effectively.



Applying clustering techniques to discover hidden groupings.



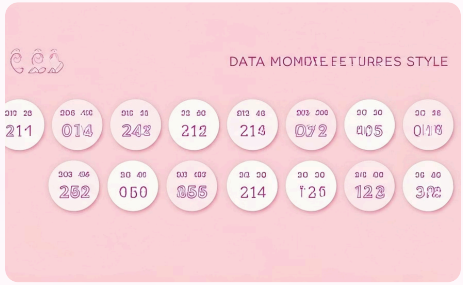
Using classification models to predict crime categories.

1. Data Understanding

Notebook: Data_Understanding.ipynb

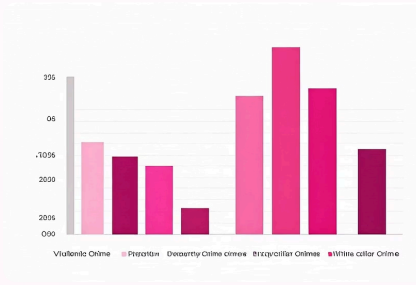
Dataset Structure

Inspecting dataset using .head(), .info(), and .describe() to understand its shape and content.



Main Features

Identifying key dataset attributes such as Dates, Category, PdDistrict, Address, and location coordinates (X, Y).



Class Distribution

Analyzing how crime categories are distributed across the dataset for insight into balance.

Missing Data Check

Verifying the absence of missing values to ensure data completeness and quality.

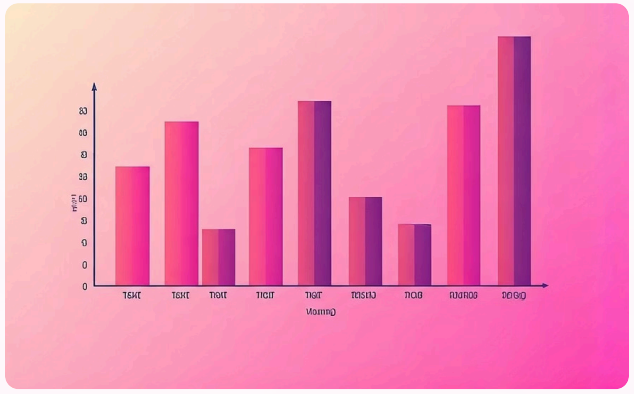
This phase provided crucial insight into the data's composition, size, and balance.

2. Data Preprocessing

Notebook: Data_Preprocessing.ipynb

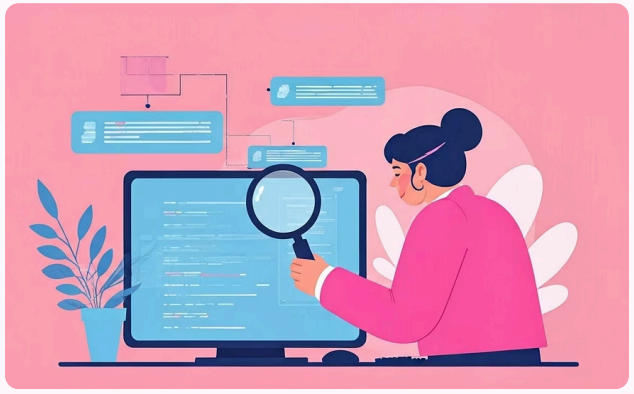
Feature Engineering

Extracted temporal features such as Year, Month, Day, and Hour from the Dates column to enhance the dataset's informational value.



Categorical Encoding

Converted categorical variables including Category and PdDistrict into numerical format using LabelEncoder to facilitate modeling.



Data Cleaning

Eliminated irrelevant columns like Descript and Resolution to improve the dataset's quality and focus.

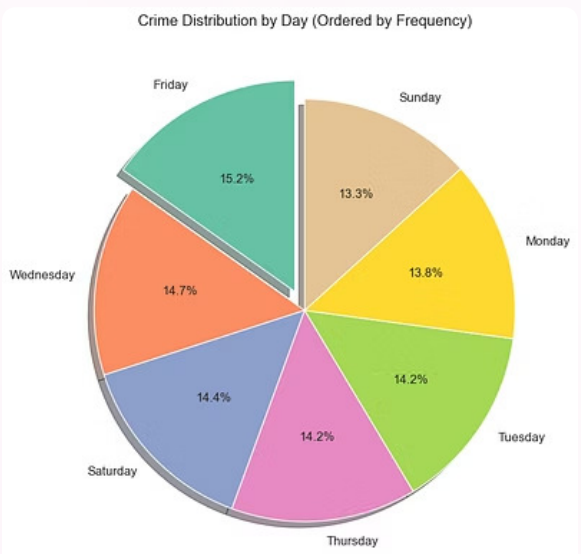
This preprocessing phase ensured that the dataset was clean, consistent, and properly prepared for subsequent machine learning algorithms.

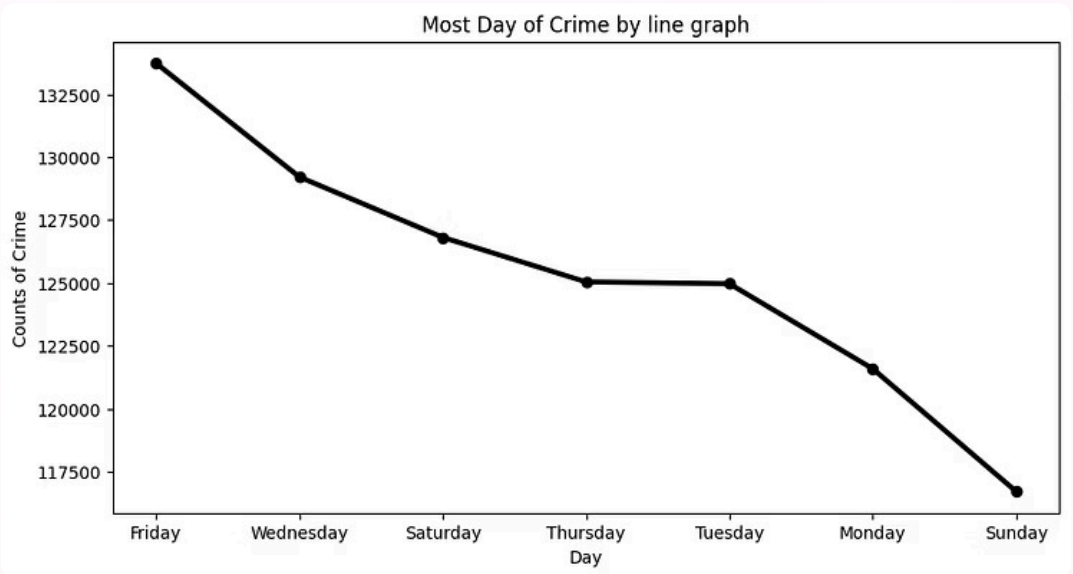
3. Exploratory Data Analysis (EDA)

Notebook: EDA_Collection.ipynb

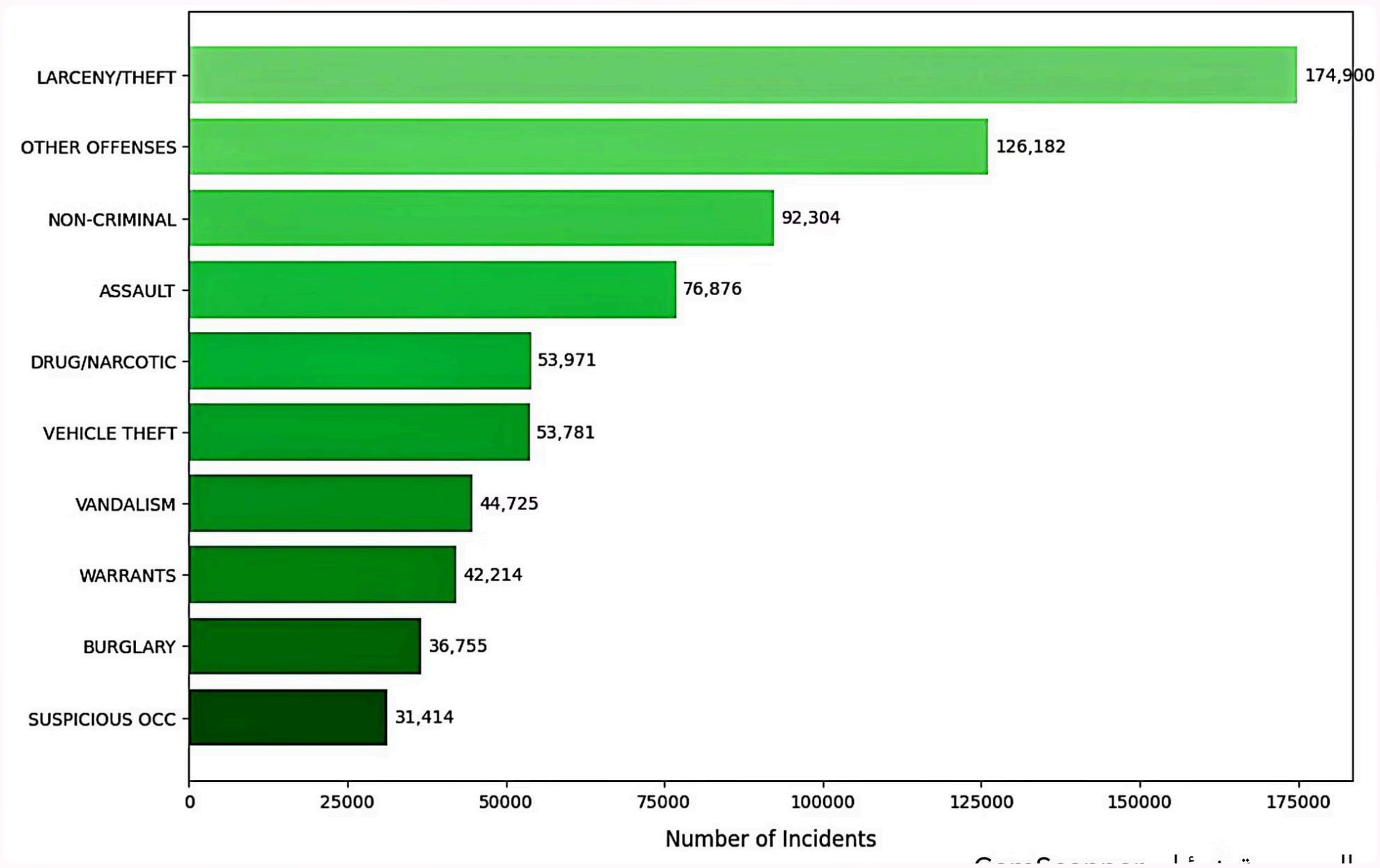
This stage focused on visualizing crime patterns to uncover trends and correlations. Key visualizations included:

- Bar charts showing the frequency of each crime category
- Heatmaps revealing temporal trends in crime rates by hour and day
- Geographic visualizations (scatter plots) highlighting crime distribution across districts





Top 10 Most Frequent Crime Types San Francisco



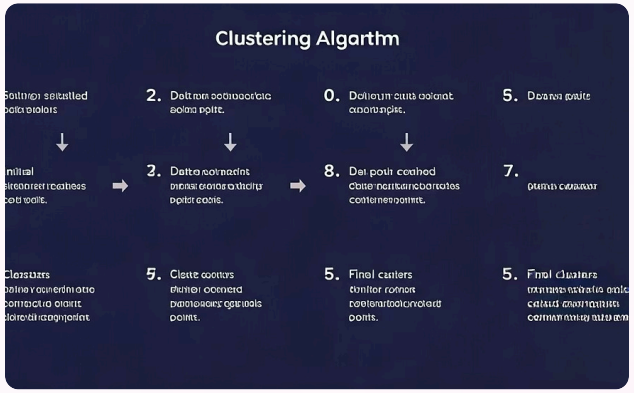
Crime Category

- Identification of the top and bottom occurring crime categories

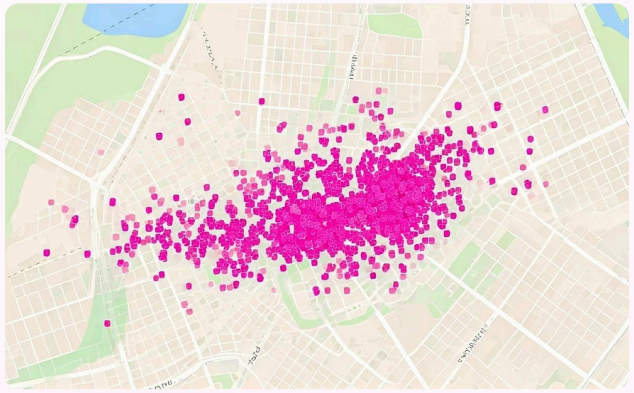
EDA helped in drawing valuable insights and guided decisions in the modeling phase.

4. Clustering Analysis

Visualizing crime data clusters by location and time.



Different clustering algorithms reveal unique patterns.



Clustering helps uncover hidden groupings within raw crime data.

Clustering Analysis Implementation

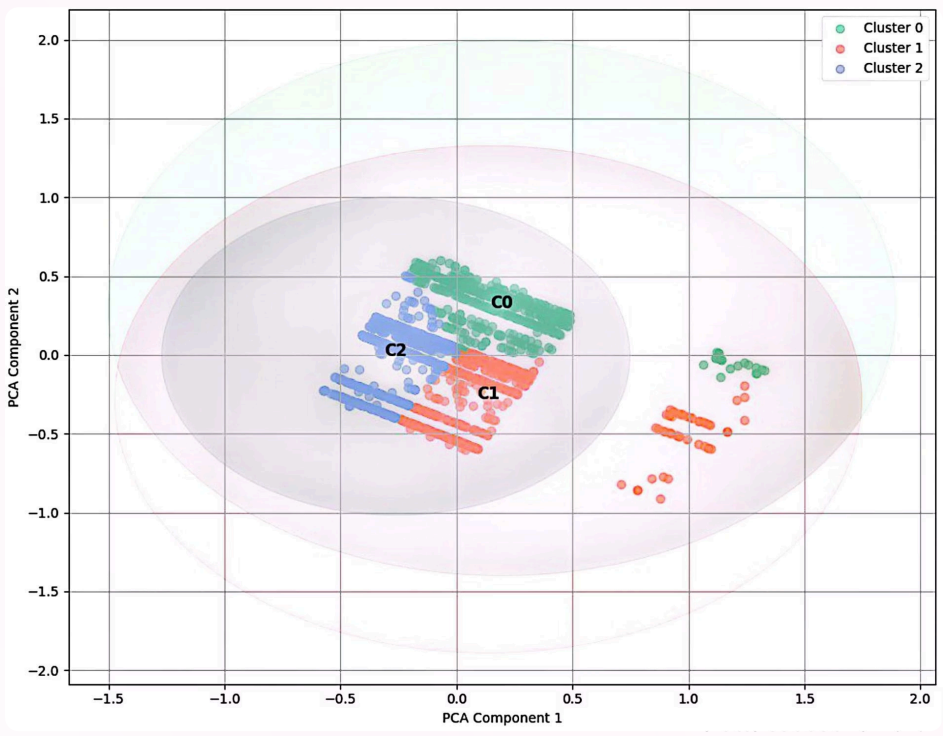
A. K-Means Clustering

Notebook: Kmeans.ipynb

- Applied the Elbow Method to determine the optimal number of clusters
- Used KMeans to group crimes, mainly based on geographic coordinates
- Plotted cluster assignments to visualize spatial groupings

code:

```
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(X_scaled)
labels = kmeans.labels_
df['Cluster'] = labels
```



KMeans Clustering (2D PCA with Circles)

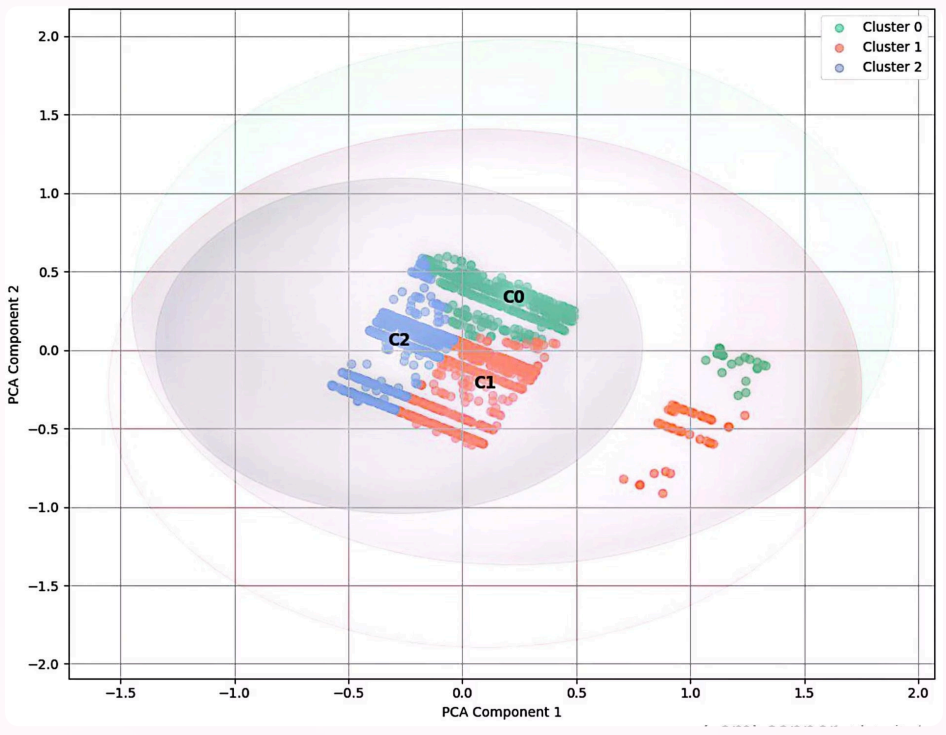
B. K-Medoids Clustering

Notebook: K_Medoids_Clusters.ipynb

- Implemented KMedoids to address K-Means sensitivity to outliers
- Compared cluster stability and accuracy with K-Means results

code:

```
kmedoids = KMedoids(n_clusters=3, random_state=42)
kmedoids.fit(X_scaled)
df['Cluster'] = kmedoids.labels_
```



K-Medoids Clustering (2D PCA)

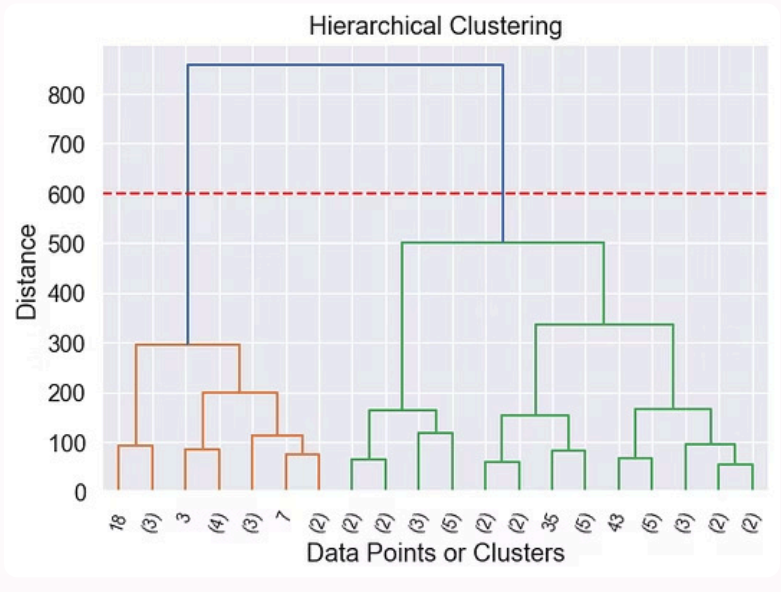
C. Hierarchical Clustering

Notebook: Hierachical_Clusters.ipynb

- Used AgglomerativeClustering to build a dendrogram-based hierarchy
- Determined the number of clusters from the dendrogram
- Applied hierarchical clustering to the same feature set

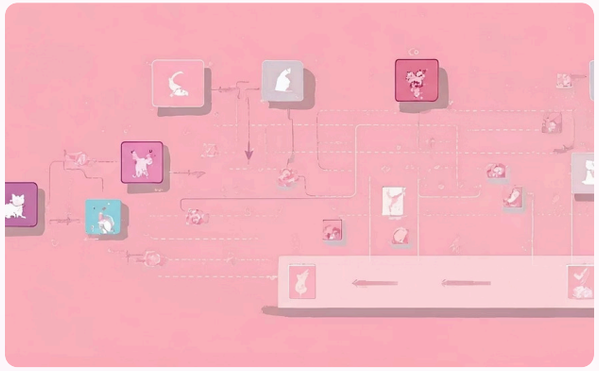
code:

```
Hierar_model = AgglomerativeClustering(n_clusters=800, metric='euclidean', linkage='average')
Hierar_labels = Hierar_model.fit_predict(X_train)
hierar_Data=X_train.copy()
hierar_Data['Hierarchical_Cluster'] = Hierar_labels
hierar_Data.head()
```



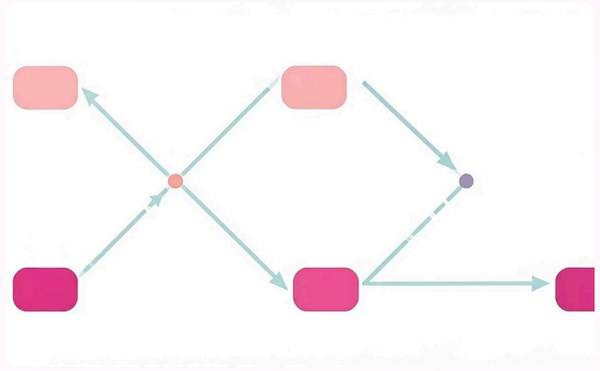
Clustering revealed crime hot spots and behavioral patterns that were not obvious from raw data.

5. Classification Models



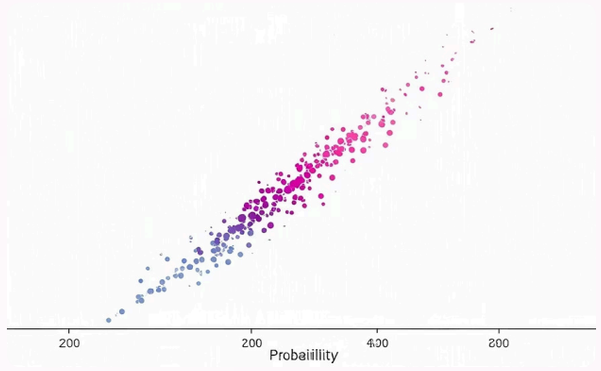
Random Forest Classifier

The best performing model with high accuracy and balanced class results.



Decision Tree Classifier

A model that uses tree-like rules to classify crime categories.



Logistic Regression

Used to predict binary outcomes based on feature probabilities.

Notebook Reference

Notebook: Classification.ipynb

Evaluation Metrics

- Accuracy Score
- Confusion Matrix
- Classification Report (Precision, Recall, F1-score)

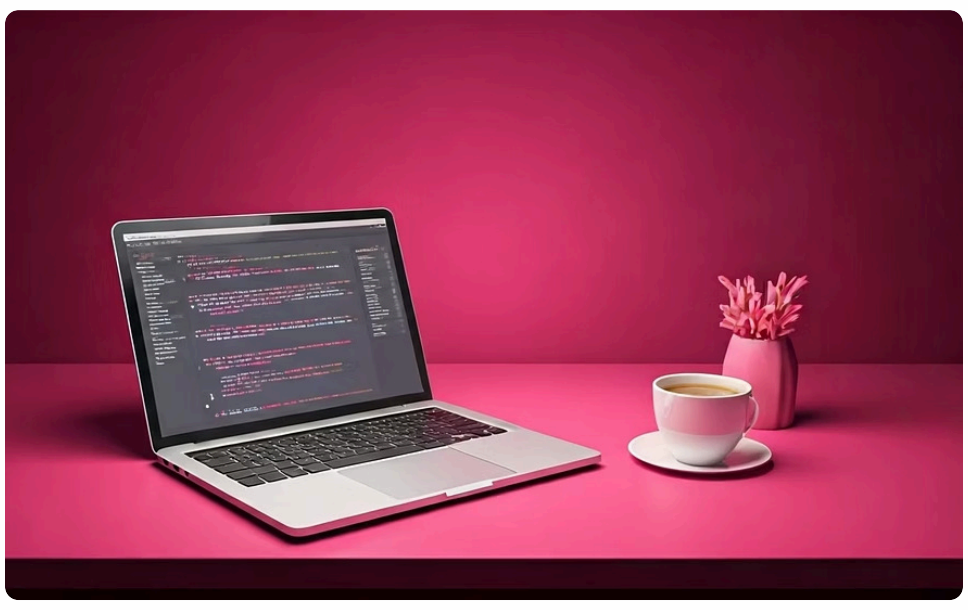
Classification Results Visualization

Data Preparation

Data Preparation

Feature engineering and scaling

Model Training



Model Training

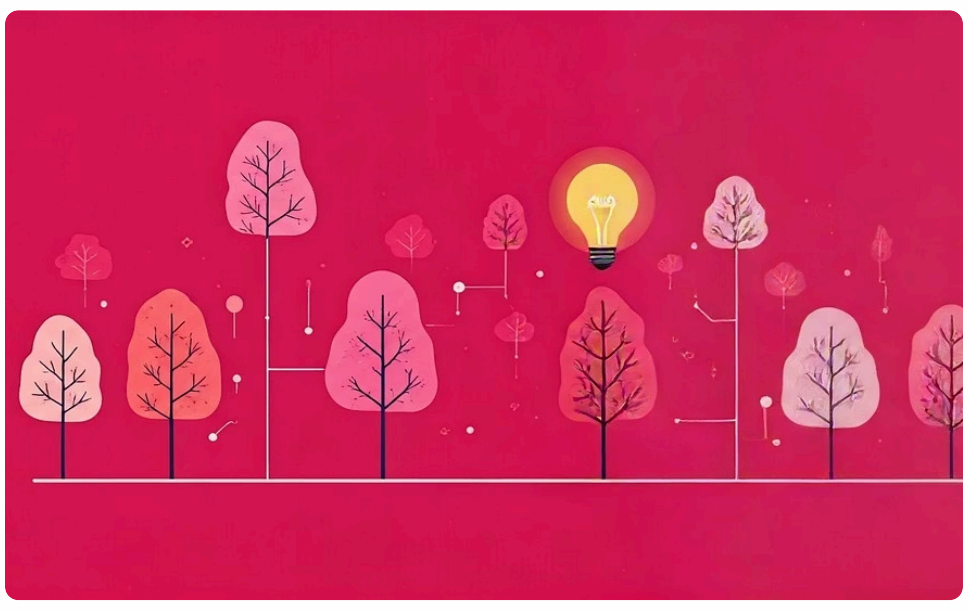
Multiple algorithms with cross-validation

Evaluation

Evaluation

Performance metrics comparison

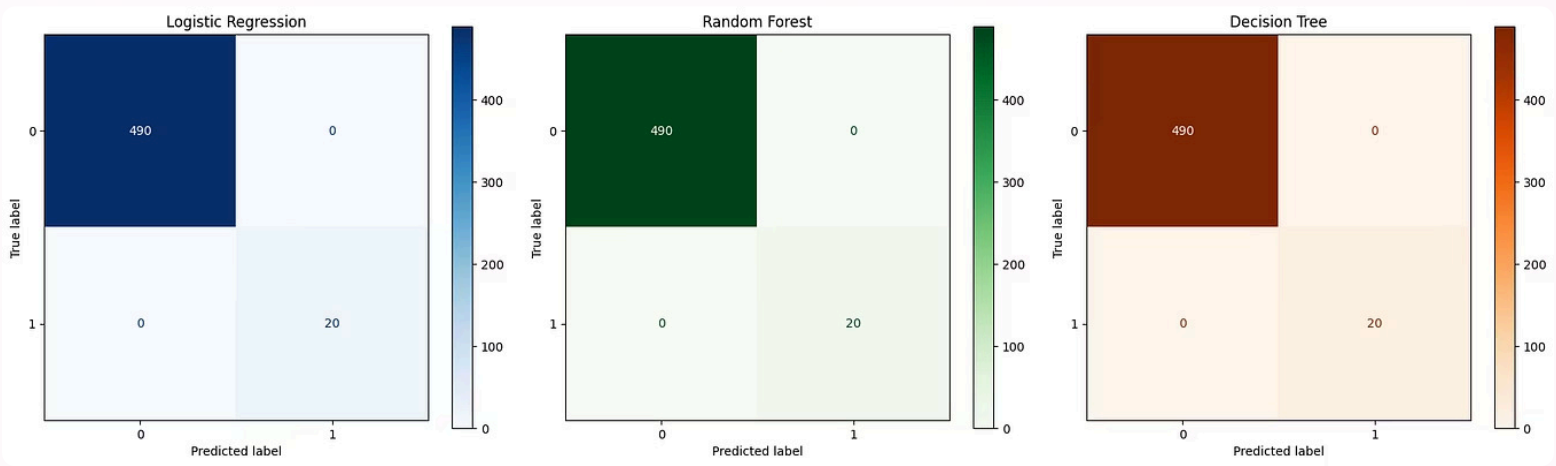
Selection



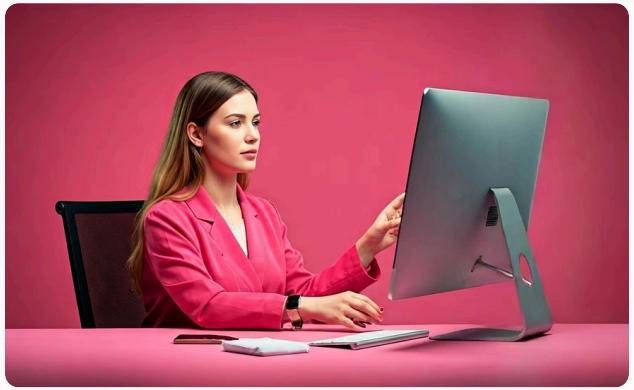
Selection

Random Forest chosen as best performer

The classification results show that our models were able to predict crime categories with varying degrees of success. The Random Forest model consistently outperformed other algorithms, likely due to its ability to handle the complex relationships between features and the multi-class nature of the prediction task.



Conclusion



Data Understanding & Preprocessing

Thorough analysis and preparation of crime data for modeling through cleaning and feature engineering.

Exploratory Analysis & Visualization

Comprehensive EDA revealing crime distribution patterns across time, location, and categories.

Clustering & Classification

Applied clustering algorithms to identify hotspots and prediction models with Random Forest classification.

This project demonstrated effective data mining on crime data, generating insights and predictive models to aid law enforcement resource allocation and crime prevention.