📝 Project Report: Crime Rate Prediction

Course Title	Data Mining and Analytics
Term	Spring 2025
Professors	Dr. Magda Madbouly, Dr. ReemEssameldin
Project Title	Crime Rate Prediction using Data Mining Techniques
Dataset Used	SF Crime Rate Prediction
Team ID	5

San Francisco Crime Data Visualization

👥 Team Members

Student Name	ID	Task
Abdelrahman Khaled Mohamed (Team Leader)	23011315	Visualization,Streamlit app
Mai Mohamed zaki	23011574	Classification Models
Nour Hosam Fahmy	23011593	Hierarchical Clustering
Shahd Hesham Ahmed	23010156	K-Medoids
Shahd Ashraf Elzieny	23011296	Evaluation metric
Habiba Hamdy Ezzat	23011250	Data Preprocessing

👀 Why we chose this Project?

- In a world where **crime** doesn't sleep, we set out to illuminate the patterns behind it. Using real-world crime data from **San Francisco**, we applied clustering and classification techniques to uncover hidden structures and predict where crimes are likely to occur. This project allowed us to turn raw data into insight and combining machine learning with social purpose to support safer communities through **smarter decisions**.

🧠 Understanding the Project and Setting Objectives:

The **goal** of this project is to explore and predict crime occurrences using historical San Francisco crime data. The main objectives include:

- To **understand** and explore the underlying patterns in crime data using unsupervised learning (**clustering**).
- To **apply** clustering techniques such as **K-Medoids** and **Hierarchical Clustering** to discover natural groupings of incidents.
- To **build** classification models (**Decision Tree and Random Forest**) that can predict the cluster or group of a new crime case based on its features.
- To **evaluate** the effectiveness of these models using performance metrics and visualization techniques..

🧲 Data Collection:

The dataset used is the **SF Crime Rate Dataset** which includes **878,049 rows** and **9 main columns** such as:

- `Dates`, `Category`, `Descript`, `DayOfWeek`, `PdDistrict`, `Resolution`, `Address`, `X`, and `Y`.

⚙️ Data Preprocessing:

- **Removed** 0 missing values and eliminated 2323 duplicates.
- **Converted** `Dates` column to `datetime`.
- **Extracted** time features: `Hour`, `Month`, `Year`, `Day`, `n_days` **(Feature Engineering)**.
- **Applied** `LabelEncoder` to `Category`, `DayOfWeek`, and `PdDistrict` **in order to make them suitable for machine learning algorithms**.
- **Removing outliers(X,Y).**
- **We applied scaling to selected columns using** `StandardScaler` **to normalize the values and prepare the data for modeling.**
- **The final** dataset included only encoded and numerical features for modeling.

✂️ Dataset Splitting:

- Used `train_test_split()` to divide data into training and testing sets with an 80/20 ratio.

🔨 Application of Mining Techniques:

Clustering:

- Applied **K-Medoids** clustering to crime features.
- Used **Agglomerative Clustering** and **Hierarchical Clustering** for further spatial analysis.

Classification:

- **Models Used:** Decision Tree Classifier and Random Forest Classifier.
- **Features:** `X`, `Y`, `day`, `Month`, `Hour`, `n_days`, `PdDistrict_encoded`.
- **Target:** `New_Labels` **derived from the clustering step (**We generated this new target because the original crime categories showed weak correlation with our numerical features—by first grouping similar incidents via clustering, we created a stronger, more learnable label for downstream classification)**.**

📏 Evaluation Metrics:

Clustering:

- **Silhouette Score**: Evaluated the cohesion and separation of clusters (better to be High).
- **Davies-Bouldin Index** (better to be low) and **Calinski-Harabasz Score** (better to be High): Used to measure cluster compactness and separation.

Classification:

- **Metrics Used:** Accuracy Score, Confusion Matrix, Classification Report.
- Example: Random Forest outperformed Decision Tree with higher accuracy and precision.

📊 Visualization (Matplotlib & Seaborn):

- Histogram: Displayed the top 15 most frequent crime categories.
- Pie Chart: Crime distribution across police districts.
- Line Plot: Illustrated number of crimes by hour of the day.
- Correlation Heatmap: Showed relationships between numerical features.
- 3D Plot: Visualized clusters in 3D using PCA to better understand the grouping of crime data.

To enhance data understanding and present insights effectively, we developed an interactive Streamlit dashboard, available at:

San Francisco Crime Data Visualization

The **dashboard** is organized into multiple tabs, each focusing on a different dimension of the **crime** dataset. It allows users to explore the data dynamically through visual interactions.

1. **Crime Categories**

- Users can view the most frequent **crime categories** using different chart types:
    - **Bar Chart**
    - **Horizontal Bar**
    - **Treemap**

2. **District Distribution**

- This tab displays the **distribution** of crimes across San Francisco's **police districts**.
- Available visualization types:
    - **Pie Chart**
    - **Bar Chart**
    - **Interactive Map:** allows users to **switch between multiple map styles** such as:
        - `open-street-map`
        - `carto-darkmatter`
        - `carto-positron`

  - This **enhances** the user's ability to explore the data using different **visual backgrounds** depending on their preference or environment.
- Useful to **spot geographic crime concentration** and identify dangerous areas.

3. **Time Analysis**

- This section visualizes how crimes vary over time, allowing filtering by:
  - **Hour of Day**
  - **Day of Week**
  - **Month**
- This helps in discovering **peak crime hours** and temporal trends (e.g., crimes happen more frequently at night or on weekends).

4. **Additional Insights**

- Cross-analysis between **crime categories** and **districts**.
- Visualization type:
  - **Heatmap**: Shows concentration of categories per district
  - **Stacked Bar Chart**: Displays how multiple categories are distributed within each district

📒 Documentation:

The complete data mining pipeline was documented in the provided Jupyter Notebook, which includes:

- **Data loading** and **cleaning.**
- **Feature engineering** and **encoding.**
- **Visualizations** and **clustering.**
- **Modeling** and **evaluation.**

✅ Conclusion & Insights:

We analyzed San Francisco crime data using clustering and classification techniques. KMedoids helped reveal hidden crime patterns by grouping incidents based on time, location, and district. We then built models to predict the cluster of new incidents using their features.

- **Top Crime Categories:** `LARCENY/THEFT`.
- **Crime Hotspots:** Found primarily in central districts like `SOUTHERN` and `MISSION`.
- **Peak Crime Hours:** Evening and night, between **6 PM to 2 AM**.
- **Clustering** provides meaningful geographical crime groups that can support urban planning and law enforcement.
- **Prediction Accuracy:** Random Forest achieved higher accuracy than Decision Tree.
