Crime Classification Project

Overview

In this project, student teams will analyze the Crimes dataset released by the San Francisco Police Department to explore factors influencing different types of crimes and build a crime category prediction model. Through data preprocessing, feature engineering, model building, and evaluation, students will compare the performance of multiple algorithms (e.g., XGBoost) and produce insightful conclusions along with policy or social management recommendations.

This project emphasizes:

- Extracting valuable insights from real-world crime data;
- Mastering the design and evaluation of machine learning classification models;
- Developing the ability to translate analytical findings into actionable social recommendations.

Project Requirements

Datasets

- **Dataset Source:**
  Crime data: https://www.kaggle.com/competitions/sf-crime/data?select=train.csv.zip
- **Description:**
  This dataset contains records of hate crime incidents collected by the San Francisc Police Department. Each record represents a single incident and includes fields describing the time, location, law enforcement unit, type of crime and relevant legal classifications.
- **Schema:** This dataset records detailed information for each crime incident, including the timestamp of the crime (Dates), the category of the crime (Category, only in train.csv), a detailed description of the incident (Descript, only in train.csv), the day of the week the crime occurred (DayOfWeek), the name of the Police Department District (PdDistrict), the resolution status of the crime (Resolution, only in train.csv), the approximate street address of the incident (Address), and the geographic coordinates in longitude (X) and latitude (Y), useful for analyzing crime patterns and predicting incident categories.

Tasks

A. Data Understanding and Preprocessing

- Import and explore the dataset;
- Check and clean missing or invalid values (dates, duplicates, etc.);

- Encode textual categorical variables (e.g., Bias Motive Description);
- Create derived variables (e.g., Arrest Indicator, Incident Quarter);
- Optionally, filter by year or county for deeper analysis.

B. Feature Engineering and Model Building

- Select features to build a classification model for the category of crime;
- Compare at least two algorithms (e.g., Decision Tree, Logistic Regression, Random Forest, XGBoost);
- Use cross-validation for model evaluation;
- Split data into train/test, and evaluate model performance using Accuracy, Macro-F1, and Confusion Matrix;
- Perform hyperparameter tuning for model optimization;
- Visualize important features, model structure, or key relationships between variables.

C. Interpretation and Insights

- Identify the most influential features affecting crime type;
- Analyze major correlations between bias motives and offense categories;
- Provide actionable social or policy recommendations, such as:
  - Community education programs targeting specific bias motives;
  - Resource allocation strategies for high-risk areas;
  - Preventive measures addressing multicultural and demographic diversity.
- (Optional) Extend the analysis over time, such as observing monthly or quarterly trends in hate crime occurrences.

Expected Learning Outcome

1. Technical Skills
- Master data cleaning and feature engineering methods;
- Gain familiarity with classification algorithms and evaluation metrics;
- Learn to use visualization to interpret model results.

2. Analytical & Creative Thinking
- Understand the connection between social issues and data patterns;
- Interpret model outputs strategically and propose innovative, actionable insights.

3. Collaboration & Communication
- Work effectively in teams to complete a data analysis project;
- Communicate model design, findings, and social significance clearly.