



HEC-HACKATHON PROJECT

Project Title

Crime Prediction Model Using San Francisco Crime Data by Classification Techniques

Team Members:

Name: Asif Ullah (**TEAM-LEADER**)

Muhammad Huzaifa Iftikhar

Muhammad Atif Laghari

Ahsan Ali

Naveera Fatima

Alishba

1. Project Background and Introduction

Crime has always been a significant issue in society, impacting safety, well-being, and the economy. With the advancement of technology, especially in data science and machine learning, it has become possible to analyze and predict crime patterns based on historical data. This project focuses on building a predictive model using the San Francisco crime dataset from Kaggle to forecast whether a crime is likely to be violent or nonviolent based on features like time, location, and crime type.

2. Problem Statement

Law enforcement agencies often lack accurate tools to predict and prevent crime before it happens. Traditional systems do not efficiently use historical data to forecast future incidents. This project aims to build a machine learning-based prediction system that classifies crimes into violent and nonviolent categories to help improve resource planning and public safety.

3. Objectives

- To perform exploratory data analysis (EDA) on the San Francisco crime dataset
- To preprocess and engineer features from raw crime data
- To apply and compare different classification algorithms:
 - Naïve Bayes
 - Random Forest
 - Gradient Boosting Decision Tree (XGBoost)
- To evaluate models using accuracy, precision, and recall
- To develop a model that can predict whether a crime is violent or nonviolent
- To visualize trends based on time (hour, day, month) and location

4. Methodology

- Dataset: Kaggle – San Francisco Crime Classification
- Tools & Technologies: Python, pandas, matplotlib, seaborn, scikit-learn, XGBoost
- Steps:
 1. Load and clean dataset
 2. Feature engineering (time, date, location-based features)
 3. Categorical encoding
 4. Classification using ML algorithms
 5. Evaluation (confusion matrix, precision, recall)
 6. Visualization and final model selection

5. Expected Results

The expected outcome is a machine learning model that can:

- Accurately classify crimes as violent or nonviolent

- Help visualize crime patterns over time and space
- Provide decision support to law enforcement

6. Tools and Resources Required

- Python 3.x
- Jupyter Notebook / VS Code
- Libraries: pandas, matplotlib, seaborn, scikit-learn, xgboost
- Dataset: <https://www.kaggle.com/datasets/san-francisco/crime-classification>

7. Timeline

| Hours | Task |
|-------|---|
| 1-3 | Dataset understanding & EDA |
| 3-6 | Preprocessing & feature engineering |
| 6-10 | Model training and evaluation |
| 10-12 | Result interpretation & visualization |
| 12-14 | Final report and presentation preparation |