Project Title: Crime Prediction and Analysis in Chicago Using Machine Learning

Kerem Aydın - 090200315

1. Introduction

In recent years, Chicago has become one of the major cities where rising crime rates and public safety issues have been intensely discussed. As one of the largest and most populous cities in the United States, Chicago also stands out with its crime statistics. Predicting crimes and taking preventative measures in advance is of critical importance for city administrations and law enforcement agencies. However, crime prediction is not limited to analyzing historical data alone; it also requires considering environmental factors, economic conditions, and socio-cultural elements. At this point, machine learning and artificial intelligence offer effective tools for analyzing crime data, predicting crime rates, and developing strategies to prevent crimes.

Crime data from Chicago reveals that a wide range of factors can influence criminal activity. These factors include variables such as time, location, weather conditions, holidays, and social events. In this project, we aim to analyze crime data in Chicago to examine the key factors influencing crime rates, develop models related to these factors, and attempt to predict future crime rates. The project will utilize machine learning algorithms to develop crime prediction models and evaluate their accuracy.

2. Project Objectives

The primary objective of this project is to understand crimes in Chicago and use machine learning techniques to predict future criminal activities. The detailed goals of the project are outlined below:

- **Data Collection and Preprocessing**: The crime dataset contains detailed information on the types, timing, and geographical locations of crimes in Chicago. This data will be collected in the initial phase of the project and then preprocessed for analysis. Missing values will be identified and either appropriately filled or removed. Additionally, erroneous data will be cleaned to prepare it for modeling.
- **Data Analysis and Exploratory Data Analysis (EDA)**: EDA will be conducted to uncover relationships and meaningful patterns in the data. This will include an analysis of crime types, regions with high crime density, and temporal trends in crime rates. EDA will help identify the factors that influence crime rates and highlight the regions with higher criminal activity.
- **Feature Engineering**: In this step, new features will be created in addition to the existing ones in the dataset to enhance model accuracy. For instance, new features can be introduced that relate to specific time windows with high crime intensity. Environmental factors such as weather, holidays, and social events will also be linked to crime incidents.
- **Model Development and Classification**: Various machine learning algorithms will be applied in this project to predict crime types and classify future crime occurrences. These models will be developed to estimate changes in crime rates and the likelihood of future crimes.

- **Time Series Analysis and Forecasting**: Based on historical data, future crime probabilities will be forecasted using time series analyses and other prediction techniques. These forecasts will provide insights into when and where crime is likely to increase.

3. Dataset and Resources

This project will utilize the publicly available **Chicago Crime Data** dataset to analyze criminal activities in the city. The dataset includes detailed information about crimes committed in Chicago since 2001. It contains data such as types of crimes, dates and times of occurrence, and geographic locations of the incidents. Additionally, the locations of crimes are specified using GPS coordinates.

The dataset includes the following information:

- Types of crimes (e.g., theft, homicide, assault, violence, etc.)
- Date and time information of the crimes
- Neighborhoods and geographic areas where crimes occurred
- Relationships between crimes and environmental factors
- **Date**: The date and time when the crime occurred
- **Primary Type**: The general category of the crime (e.g., THEFT, BATTERY)
- **Description**: A more specific description of the offense
- **Location Description**: Description of the location where the incident took place
- **Arrest**: Whether an arrest was made or not
- **Domestic**: Whether it was a case of domestic violence
- **Beat, District, Ward, Community Area**: Administrative and geographical location indicators
- **Latitude / Longitude**: Geographical coordinates
- **Year**: The year in which the incident occurred
- **Updated On**: The last update date of the data

4. Research Questions

In this project, the following research questions will be formulated to analyze and predict crimes in Chicago:

1. **Temporal Changes in Crime**: How do crime rates vary over time? For example, on which days of the week and at what times of the day are crime rates highest? In which seasons are crimes more prevalent?
2. **Geographical Distribution and Regional Factors**: In which regions of Chicago do crimes concentrate? How are these concentrations related to demographic, economic, and social factors?
3. **Crime and Environmental Factors**: How do environmental factors such as weather conditions, public holidays, and social events affect crime rates?
4. **Future Crimes**: Based on historical data, how can future crime rates be predicted?

5. **Crime Reduction Strategies**: What measures might be effective in reducing crime rates in high-crime areas? How might the model simulate the impact of such interventions on future predictions?

5. Methodology

Data cleaning and preprocessing are foundational steps in any data analysis process. The purpose of these steps is to make the data suitable for modeling and improve the model's accuracy. The following operations will be performed on the Chicago crime dataset:

1. Handling Missing Values

Missing values are quite common in datasets. Since they can lead to inaccurate results during model training, they must be handled properly. Some of the techniques used for dealing with missing values include:

- **Deleting Missing Records**: If missing values do not affect a significant portion of the dataset, the rows containing them may be removed entirely.
- **Filling in Missing Data**: Missing entries can be filled using techniques such as mean, median, or nearest neighbor imputation. Especially for categorical variables, the most frequent category or values derived from similar records can be used.

2. Detecting and Correcting Erroneous Data

Incorrect values in the dataset (e.g., invalid GPS coordinates or illogical crime type labels) must be cleaned appropriately. Outlier analysis can be used to detect such anomalies. Values to be removed should be chosen in a way that does not significantly distort the overall data distribution.

3. Encoding Categorical Data

Since machine learning algorithms work more effectively with numerical data, categorical variables must be encoded. The following methods will be used:

- **One-Hot Encoding**: Especially for categorical features like crime types, each category will be converted into a binary column (0 or 1).
- **Label Encoding**: If categories are ordinal (e.g., crime severity: low, medium, high), numerical labels can be assigned accordingly.

4. Processing Temporal Information

Crime data is associated with timestamps. To analyze how crime is distributed across different time periods:

- **Extracting Time Features**: The timestamp data will be broken down into subcomponents such as year, month, week, day, and hour. This is essential for understanding temporal trends in crime activity.

5. Processing Geographic Information

The dataset may contain geographical coordinates (GPS) for the location of crimes. These coordinates will be used to identify crime hotspots. GPS data is often visualized through interactive maps, helping us understand the regional distribution of crimes.

Feature Engineering

Not all features in the dataset may have a direct impact on model performance. Some features might be irrelevant or carry very little information. Feature engineering is the process of transforming the data into a more meaningful form and creating new features to enhance model performance. The following steps will be carried out:

1. Transformation of Categorical Variables

- **Temporal Factors**: Time information in the dataset (e.g., day, week, hour) can be highly valuable for understanding crime patterns. Factors such as weekdays vs. weekends, holidays, and weather conditions will also be considered.
- **Local Factors**: The neighborhood or area where a crime occurs is a crucial feature. The socio-economic status of different neighborhoods may significantly impact crime rates. Therefore, interaction features using socio-economic data for each area will be created.

2. Creation of New Features

New features will be derived by establishing relationships between the existing ones in the dataset. For example:

- **Crime Density**: The interaction between high-crime areas and environmental factors (e.g., weather, special events).
- **Seasonal Crime Trends**: Features that reflect how crime rates respond to seasonal changes will be introduced.

3. Feature Selection

Some features in the dataset may lead to overfitting rather than improving model accuracy. Therefore, selecting more meaningful features will help boost model performance. For feature selection, the following techniques will be applied:

- **Correlation Analysis**: Strong correlations between features will be identified, and redundant columns carrying similar information will be removed.

- **Feature Importance Ranking**: Using models like Random Forest, feature importance scores will be calculated to select the most influential features.

Models for Crime Prediction

Different machine learning models will be used for crime prediction. Each model may perform differently depending on the features of the data and the problem definition. Therefore, comparing these models is important. The models to be used and the evaluation metrics are as follows:

1. Models to Be Used

- **Logistic Regression**: A simple yet effective model used for classifying types of crimes.
- **Random Forest**: A robust ensemble model that provides high accuracy by combining multiple decision trees. It also offers feature importance rankings, enhancing interpretability. It is resistant to noisy data.
- **Support Vector Machines (SVM)**: A strong classification model suitable for both linear and non-linear datasets. It will be used to classify crimes.
- **XGBoost / LightGBM**: These boosting-based models correct errors in previous iterations. They are fast, memory-efficient, and suitable for large datasets. They perform especially well with imbalanced data.
- **Gradient Boosting**: Another ensemble model that makes more accurate predictions by correcting previous errors. Variants like LightGBM or XGBoost can further enhance its performance.
- **Neural Networks**: Deep learning models capable of capturing complex patterns, ideal for larger and more complex datasets.

2. Performance Evaluation

Model performance will be assessed using metrics such as accuracy, precision, recall, and F1 score. Additionally, the following metrics will be considered:

- **Precision**: Indicates how many of the predicted crimes were actually correct. High precision means fewer false positives.
- **Accuracy**: The overall correctness of the model's predictions.
- **Precision/Recall**: Shows how effectively the model detects actual crimes. High recall means fewer false negatives.
- **F1 Score**: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.
- **ROC-AUC**: Useful for evaluating classification performance, especially with imbalanced datasets.
- **RMSE / MAE**: Root Mean Square Error and Mean Absolute Error will be used to evaluate error in time series models.

To improve model performance, **hyperparameter optimization** (e.g., grid search or random search) will be applied, and care will be taken to avoid **overfitting or underfitting**.

6. Expected Outcomes

This project aims to develop a model capable of predicting crime rates in Chicago. The comparison of different models will provide insights into which one produces the most accurate predictions and will enhance the overall accuracy of crime forecasting. Additionally, valuable information will be obtained about how crime evolves over time and which areas are likely to experience higher crime rates.

These findings will help law enforcement agencies and city administrations anticipate future crimes and develop strategic plans to enhance public safety. The models will not only support decision-making but also assist in allocating resources more effectively in high-risk areas, ultimately contributing to crime prevention efforts.

7. References

1. https://data.cityofchicago.org