EDS 6340 Introduction to Data Science
Project Proposal

October 2, 2024

Project Group Number: 2

Name of the Student	UH ID
Yuzhen Hu	2299391
Murtaza Mustafa	2415417
Duggimpudi, Bala Meghana	2275900
Rehan, Muhammad Asad	1951934

Project Title: "Predicting Community Crime Rates Using Socio-Economic and Law Enforcement Data"

Dataset Information:

Dataset Characteristics	Multivariate
Subject Area	Social Science
Associated Tasks	Regression
Feature Type	Real
Number of Instances	2215
Number of Features	125
Has Missing Values?	Yes

Dataset: `https://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized`

Quick Summary of the Project:

The "Communities and Crime Unnormalized" dataset provides a rich collection of socio-economic, demographic, and law enforcement statistics from various U.S. communities, aimed at understanding factors influencing crime rates. In this project, we will develop a regression model to predict crime rates in different communities based on the 125 features provided, which include variables like population demographics, economic status, and law enforcement resources. The project will involve data preprocessing to handle missing values, converting categorical variables into numerical forms, and feature selection using correlation analysis to identify the most influential factors. Multiple regression algorithms, such as linear regression, decision trees, and random forests, will be employed and evaluated to determine the most effective model. The final model will then be tested on unseen data to assess its accuracy and robustness in predicting community crime rates. This study aims to provide insights into which socio-economic and law enforcement characteristics significantly impact crime rates and how they can be leveraged to inform public policy decisions.

Project Pipeline

To build a machine learning pipeline for predicting community crime rates based on the provided dataset, we will follow these steps:

1. **Data Collection:** The dataset has been sourced from the UCI Machine Learning Repository. Click here to download the data.
2. **Data Preprocessing:** Clean the dataset by handling missing values using appropriate imputation techniques. Convert all categorical features into numerical values using techniques like one-hot encoding. Normalize or standardize the numerical features to ensure they are on a similar scale for optimal model performance.
3. **Feature Selection:** Identify and select features that have a significant correlation with crime rates using statistical analysis and correlation plots. This step will help focus on the most influential factors, reducing the risk of overfitting and improving model accuracy.
4. **Model Selection:** Choose suitable regression algorithms (e.g., Linear Regression, Decision Trees, Random Forest, Support Vector Regression) to model the relationship between the selected features and community crime rates.
5. **Model Training:** Split the dataset into training and testing sets to train the selected regression models. Use cross-validation techniques to ensure that the model's performance is generalized across different subsets of the data.
6. **Model Evaluation:** Assess the performance of the models using evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. This will help identify the most effective model for predicting crime rates.
7. **Results Interpretation:** Interpret the model's results to understand which socio-economic and law enforcement features have the most significant impact on crime rates. Use the findings to provide insights that can help in policy-making and community resource allocation.

By following these steps, we aim to develop a robust regression model that can effectively predict crime rates in communities based on socio-economic and law enforcement variables.

Tools, Libraries, and Frameworks:

- NumPy, Pandas for Data Preprocessing
- Matplotlib, Seaborn, for Visualization
- Correlation-based Feature Selection
- Scikit-learn for Model Building and Evaluation