A logo for university computing

Auto-generated description

**Evaluation cover page**

|  |  |
| --- | --- |
| *Student's Full Name:* | Taylan Ozgur Ozkan |
| *Student Number:* | 2024140 |
| *Module Title:* | Machine Learning (10 ETCS) |
| *Evaluation Title:* | CA1 Project |
| *Assessment Due Date:* | 21/04/2024. |
| *Presentation Date* | 26/04/2024. |

**Declaration**

By submitting this review, I confirm that I have read CCT's policy on academic misconduct and understand the implications of submitting work that is not mine or that does not appropriately reference material taken from a third party or other source.

I declare that it is my own work and that all third-party material has been properly referenced.

I further confirm that this work has not previously been submitted for evaluation by me or anyone else at CCT College Dublin or any other higher education institution.

Contents

[Introduction: 2](#_Toc165063948)

[Data Review: 2](#_Toc165063949)

[Data Cleansing: 3](#_Toc165063950)

[Data processing: 3](#_Toc165063951)

[Hyperparameter Tuning: 4](#_Toc165063952)

[Analysis and Findings: 5](#_Toc165063953)

[Conclusion and recommendations: 5](#_Toc165063954)

CRIMES IN USA

# Introduction:

This report was prepared to examine crime rates in the USA and the distribution of these crimes over time. The data is based on a data set containing crime statistics collected from various states and cities. The main method used in the project is the application of machine learning techniques. The main goal of the report is to make effective analyzes on crime rates using data science and machine learning methods and to make recommendations for policy makers with these analyses. At the end of the report, the analysis techniques applied and the findings obtained will be presented in a way that will contribute to the development of strategies that can be used in the fight against crime.

# Data Review:

The data set used within the project is based on crime reports collected from different cities and states of the USA. The data set includes various information such as crime types, the date the crime was committed, the city and state where the incident occurred. Before starting the analysis of the data, a preliminary analysis was made through the Pandas library using the Python programming language.

As a first step, the data set was read with the pd.read\_csv function and a DataFrame was created. This DataFrame provides a basic structure for performing operations on data.

Initial inspection of the DataFrame was done by reviewing the first five rows using the df.head() method. This step is important to quickly understand the overall structure of the dataset and the main columns it contains. Additionally, a basic statistical summary of the data set was obtained by using the df.describe() function. This summary includes statistical information such as means, standard deviations, minimum and maximum values for numerical columns and gives an idea of the overall distribution of the data set.

For a more detailed analysis of the data set, the df.info() method was used. This method displays the data type of each column in the data set and the number of non-null values it contains. This information is critical to determining which columns contain missing data and which data types are used.

# Data Cleansing:

This process includes steps such as correcting missing or incorrect data, removing unnecessary data, and making data types suitable for analysis.

• **Editing Date Data**: The 'Date' column containing crime dates has been converted to datetime format so that it can be used as a time series in analysis. This conversion was performed by using the pd.to\_datetime function and marking incorrect date formats as NaT (Not a Time) with the coerce option.

• **Missing Data Management**: Missing values in columns containing location information such as 'City' and 'State' have been changed to 'Unknown' to preserve the integrity of the analysis. This process was done in place using the fillna method. Additionally, empty values in the 'Keyword' and 'Summary' columns are filled with empty strings ('') to avoid problems during text processing.

• **Removal of Duplicate Data**: Duplicate records in the data set were removed using the drop\_duplicates method because they could lead to misleading data analysis.

These cleaning steps ensured that the dataset was created as an accurate and clean basis for machine learning models to be applied in later stages. As a result, these operations performed on the data set increased the reliability of the analysis and prevented potential errors.

# Data processing:

• **Feature Engineering**: Some columns in the data set have been transformed to create more meaningful and functional features. For example, the 'Date' column is divided into two separate columns, year and month, in order to analyze the distribution of crimes by years and months. This process was performed using Python's datetime features.

• **Editing of Categorical Variables**: Categorical variables in the data set have been converted to numerical format so that they can be directly processed by machine learning models. Columns such as 'City' and 'State' are represented by numerical labels, each corresponding to a different category. This conversion was made using the LabelEncoder method.

• **Handling of Missing Values**: Although basic adjustments were made for missing values in the previous cleaning steps, advanced processing was applied for other missing values that emerged during analysis. In particular, gaps that could cause problems during the training of the model were eliminated with appropriate imputation techniques.

• **Data Normalization**: Numerical variables, especially those of various scales, are normalized to make the model more stable and faster to learn. This step is critical to prevent numerical differences in the data set from being misinterpreted by the model. Methods such as StandardScaler were preferred for normalization.

# Hyperparameter Tuning:

In this study, the following steps regarding hyperparameter tuning were performed:

**Model Selection**: A classification model was created using RandomForestClassifier.

**Parameter Determination**: Hyperparameters of the model were determined. These parameters included n\_estimators (number of trees), min\_samples\_split, min\_samples\_leaf and max\_depth.

**Model Training**: First, the model was trained with the determined hyperparameters and some initial results were evaluated.

**Hyperparameter Optimization**: To optimize the performance of the model, a search was performed to find the optimal combination of hyperparameters using the RandomizedSearchCV method. This process made the process more efficient by testing a certain number of randomly chosen combinations of parameters.

**Evaluation of Results**: After the RandomizedSearchCV process, the best hyperparameters of the model and the best score obtained with these parameters were determined.

This process is critical to improving the model's predictive performance because choosing the right hyperparameters directly affects the model's ability and generalizability to learn from your data. Hyperparameter tuning is essential to maximize the potential of your model and avoid problems such as overfitting or underfitting.

# Analysis and Findings:

After the processing and preparation of the data set was completed, in-depth analyzes were carried out on the crime data using various machine learning techniques. These analyzes aim to increase the capacity to predict crime while examining the types of crimes and the context of the time and place in which they occur. Here are some important analysis methods applied in this process and the findings obtained:

• **Classification Models**: Various classification models have been used to predict crime types. These models include Decision Trees and Random Forest. Each model was optimized by experimenting with different parameter settings on the training data. Model performance was evaluated with the cross-validation method and the model that gave the best results was selected.

• **Regression Analysis**: Regression models were used to estimate crime numbers and crime rates based on various demographic and socio-economic factors. To increase the accuracy of the model, the relationships between variables in the data set were examined in detail.

• **Time Series Analysis**: Time series analysis techniques have been applied to examine the change in crime rates over time. These analyzes have been used especially to determine seasonal effects and trends in crime rates. The findings revealed that crime rates increased or decreased in certain periods.

• **Determination of Important Features**: During model installation, feature importance analysis was performed to understand which features are more predictive of crime prediction. This analysis was performed specifically using the feature importance outputs of the Random Forest model.

# Conclusion and recommendations:

Although the analyzes and machine learning models applied allowed us to better understand and predict crime rates in the United States, they were not sufficient. The findings obtained do not enable evaluation or prediction. More data is needed for this analysis, and the current dataset has made it difficult to make an analysis and prediction due to both insufficient columns and insufficient data. Although it is possible to understand the data better with the applied techniques, as a result more data is needed.