**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | MSc in Data Analytic |
| **Assessment Title:** | Crime Data of Garda Division, Ireland from 2003 to 2019 |
| **Lecturer Name:** | David McQuaid ,Dr. Muhammad Iqbal, Taufique Ahmed, Sam Weiss |
| **Student Full Name:** | Unnati Bhardwaj |
| **Student Number:** | 2023404 |
| **Assessment Due Date:** | 12/11/2023 |
| **Date of Submission:** | 12/11/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

# Group ID - MSc in Data Analytics

Author: Unnati Bhardwaj

e-mail: 2023404@student.cct.ie

Student ID:2023404

***Abstract:****This dateset represent a comprehensive analysis of crime in Ireland spanning from 2003 to 2019,with a focus on different Garda regions. The datasets were collected from public source called kaggle.com.The study aims to to transform a quarterly crime data into annual trends to reveal long-term crime patterns. Our analysis encompasses data exploration,descriptive statistics and data visualization across various crime categories, shedding the light on the regional crime distribution and trends. Additionally, statistical analysis employs the Poisson and Binomial distributions to gain insights into specific offence types following the data visualization. To facilitate future predictions, we incorporate supervised machine learning technique, specifically decision trees and random forest, offering insights into crime dynamics*

**Introduction:**

The Central Statistics Office (CSO) publishes statistics on Recorded Crime on a quarterly basis. The Recorded Crime statistical release and associated tables provide detail on the number and type of crime incidents recorded by An Garda Síochána. This report dig into an extensive datasets, that covers the period from 2003 to 2019, with a prime focus on converting the quarterly data into annual format. This conversion is important for simplifying the data and facilitating a clearer understanding of long-term crime trends and regional variations. The analysis begins with an exploration of key columns such as Region,Garda Division, Offence,type of offence, and quarterly data from 2003 to 2019.

In the data pre processing phase, we convert quarterly data into annual aggregates to enhance the clarity and effectiveness of visualizations. This will allow us to effectively present various types of crimes, create time series plots to track crime trends, and examine the distribution of crimes across different regions.

Moreover, the report includes an analysis of Discrete distributions, specifically the Poisson and Binomial distributions, to quantify the frequency of crime occurrence within the fixed interval of times. The statistical approach provides deeper insights into the nature and regularity of criminal activities.

The latter part of the report is dedicated to predictive analysis using machine learning techniques, specifically, employing the supervised learning methods,including Random Forest and Decision Tree, to forecast future crime trends. These methods are well suited for this purpose due to their ability to handle complex datasets and provide insightful predictions above various crime categories.

**Methodology:** The methodology for analyzing the extensive datasets on crime statistics in Ireland,spanning from 2003 to 2019, is structured around a systematic data processing approach. The key steps involved below:

*Importing Libraries:*The analysis begins with the implementing the important python libraries that facilitate data manipulation,analysis and visualization. These include:

Pandas(pd)=for data manipulation and analysis useful for handling tabular data

Numpy(np)= for numerical computing.

Matplotlib(plt)=for creating static, interactive and animated visualization in python

Seaborn(sns)=A statistical Data Visualization library built on the top of matplotlib,offering a higher-level interface for drawing attractive and informative statistical graphics.

Warnings:It is used to control the issue of warnings

Statistics:This is for calculating the mathematical statistics of numeric data like mean,median,mode

Scipy’s poisson:This is a part of SciPy library. It is for the probability distribution of a number of events occurring in a fixed interval of time or space.

Scipy’sBinomial= It is for the discrete probability distribution of the number of successes in a sequence of n independent experiment.

*Data loading:*After setting up the necessary libraries, the next step is to load the datasets. The datasets, titled “IRELAND\_CRIME\_GARDA\_DIVISON\_wise\_2003-2019.csv”,provided detailed crime statistics across different Garda division over the specified period. The initial exploration of the datasets, such as examining the first five rows and understanding the structure of the data, is performed using the “head()” method in Pandas. This method applied to the Data Frame, conventionally named “df” in Pandas to get a glimpse of the dataset’s structure, including columns and initial entries

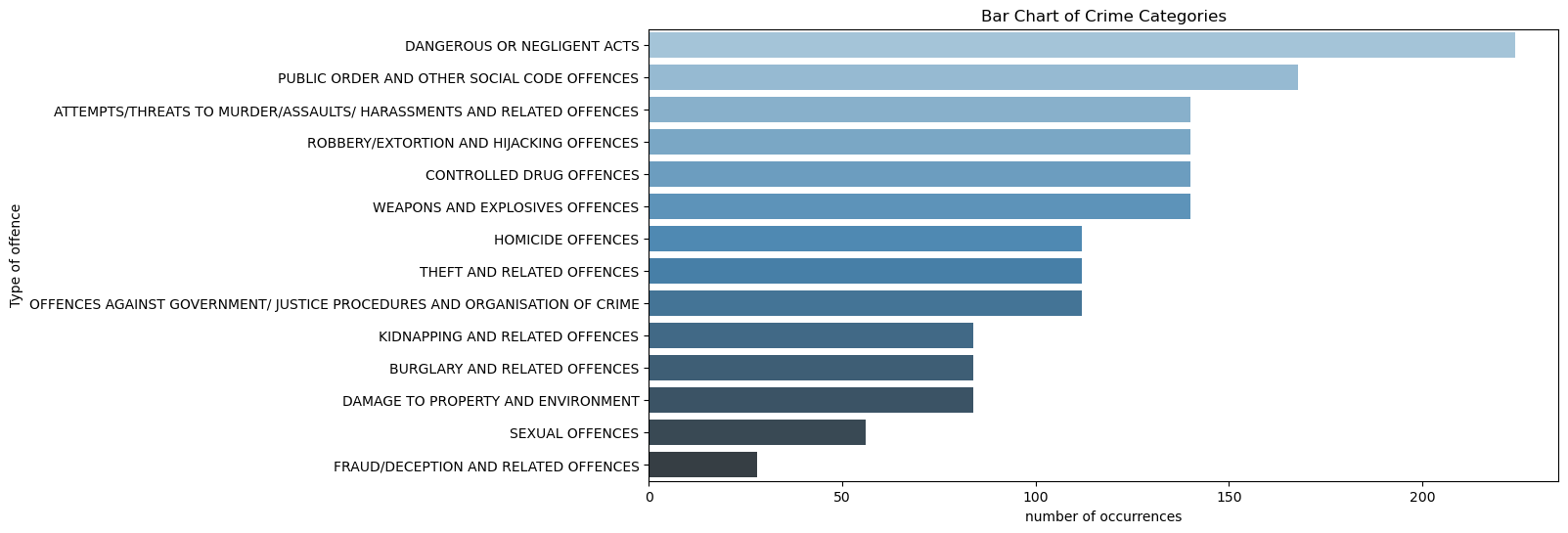
*Data Cleaning and Transformation:* The next phase in our analysis involves refining the datasets to ensure accuracy and relevance in our findings. The important step in this process is the removal of specific columns:2019Q1,2019Q2,2019Q3.This decision is based on the absence of the data for the quarter(Q4) of 2019. Eliminating these incomplete data points for 2019 enables a more consistent and reliable analysis, as it aligns the datasets to complete annual cycles from 2003 to 2018. Such type of visualization is useful because it will provide a clear visualizations and insights of the patterns and types of crimes committed during the period. This step enhances the integrity and ensures the proper visualization of crime trends over the years.

Now after dropping the columns the next part is transforming the Quarters into years. The data used for this analysis comprises quarterly data points spanning the year 2003 to 2018. Each quarter’s data is represented in columns with the format “YYYQn”, where YYYY is year, and n is the quarter(1 to 4). To convert quarterly data into annual data, we utilized Python programming with the Pandas library.The transformation process involves defining the year range,do the Iterative Summation. After that the process was iterated for each year and than remove the original Quarterly columns.

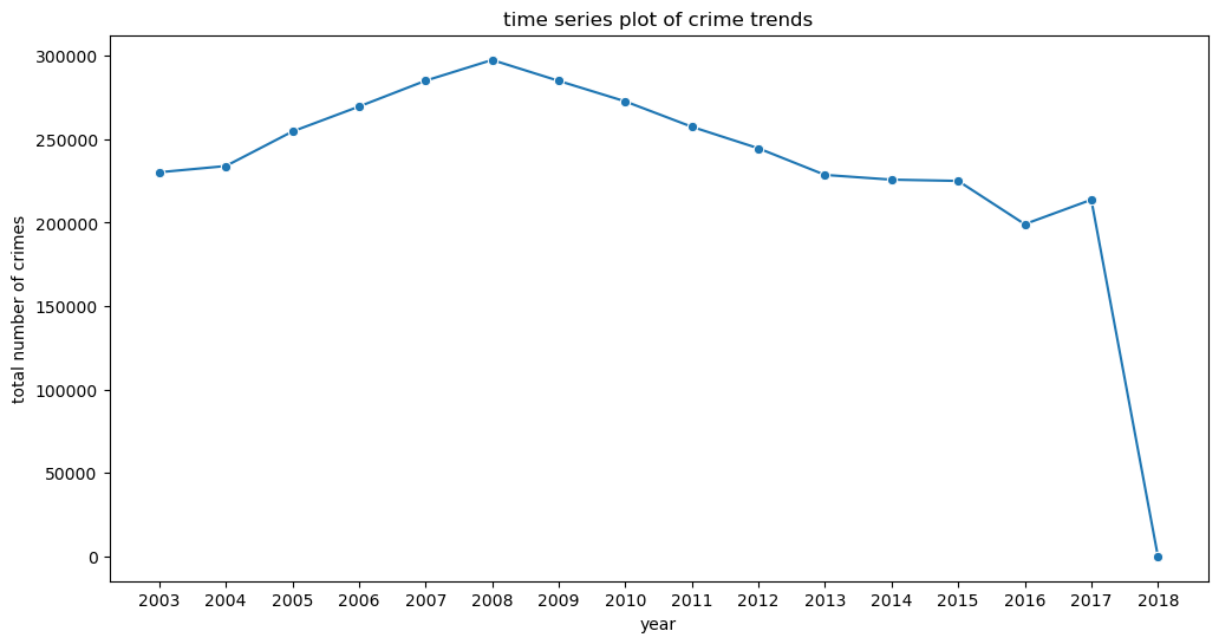
*Descriptive analysis:* After the necessary cleaning and transformation steps were applied to our datasets, we conducted a descriptive analysis. The descriptive statistics provide several key insights into the data. Notably, the “count” for each year in this datasets is 1624, indicating that there are 1624 observations for each year with no missing value. The “mean” represents the average number of crimes reported annually. For instance, the mean for the first year is 428.48 which suggests that the average is around 428 crimes were reported that year.It serves as a basic indicator of the overall crime level for each year. The “Standard deviation” gives an understanding of the variability or spread of crime numbers around the mean. A higher standard deviation shows a wider dispersion of values.For example,the std for the first year is about 369.04 suggesting a variation in the number of reported incidents around the average.The “minimum value” indicates the least number of crimes reported in any given year, providing the lowest crime levels encountered in the datasets. The “median” offers a middle point in our crime data distribution, representing the value where half of the years have higher crime reports and half have lower. And lastly “mode”, in this datasets is notably 0, it shows that many types of offenses were not occurred in several years. This is s important insight, indicating the absence or rarity of certain crimes over the observed period.

**Data Visualization:** Following the descriptive analysis and the removal of unnecessary columns, we move on the data visualization phase. Our goal is visually represent the data using various graphical techniques, including bar charts to depict different crime categories, time series plots to illustrate crime trends over time, and heat maps to visualize the geographical distribution of crimes. These visualization plays a crucial role in uncovering valuable insights from the data. By doing these visualization, we aim to extract meaningful patterns, trends and geographical variations in the crime data, providing a more comprehensive understanding of the underlying information on the analysis

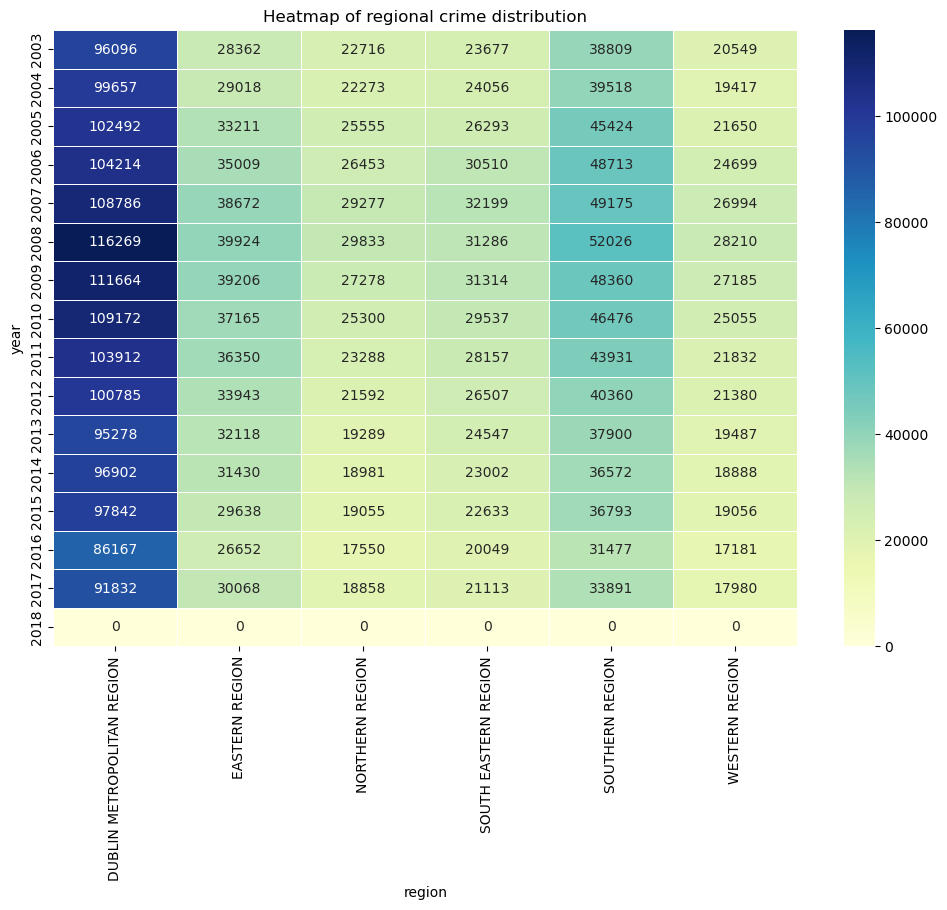
*Bar chart* :The bar chart presented a distinct visual representation of the frequency of various Types of Offenses. It is evident that certain categories are significantly more prevalent than others,highlighting areas that may necessitate a more targeted allocation of attention and resources. Upon a closer examination of the bar chart, it becomes apparent that “Dangerous or Negligent Acts” stands out as the most frequently occurring category.Conversely, the category “Fraud/Deception and Related Offenses” registers a relatively low frequency, falling within the range 0 to 50 across all the years. This observation shows the importance of focusing efforts and resources on addressing the categories with higher frequency while considering the potential strategies to address less frequent categories in a more tailored manner.



*Time series Plot(crime trends):* The plot shows the overall trends of crime which happened in all those years. This can help in understanding whether crime rates are increasing, decreasing or stable over the time and can be helpful for making the policy and resource allocation. The trend analysis reveals that in 2003, the total number of reported crimes stood at approx 240,000. Later on, there was a noteworthy upward trend in crime rated, peaking at nearly 290,000 in 2008. However, following this peak, there was a gradual decline in the number of reported crimes, ultimately reaching a lower point in 2018, where count appears to near zero. This perspective of crime trends provides essential insights of policy making and resource allocation.

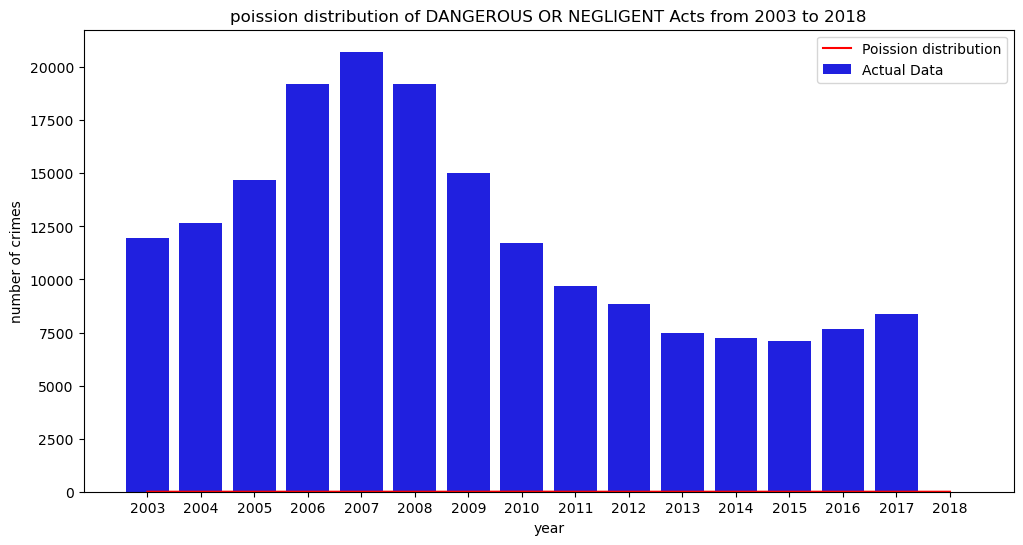


*Heat-map(Regional Distribution):* The heat-map gives a visual representation of how crime is distributed across different regions over the years. This can be useful for identifying the regions with consistently high or low crimes rated and understanding the regional trends and patterns. In this heat-map the years are listed along the y-axis and regions are distributed on x-axis. The color intensity in each cell correlates with the crime rate; darker shades indicates higher crime rates, while lighter shades indicate lower crime rates. The color scale on the right provides a reference for the number of crimes, ranging from 0 to 100,000.In this matrix, the Dublin metropolitan region shows the highest crime rates, particularly in the early years and the western region also showing high crime rates but does not surpass the Dublin Metropolitan Region.

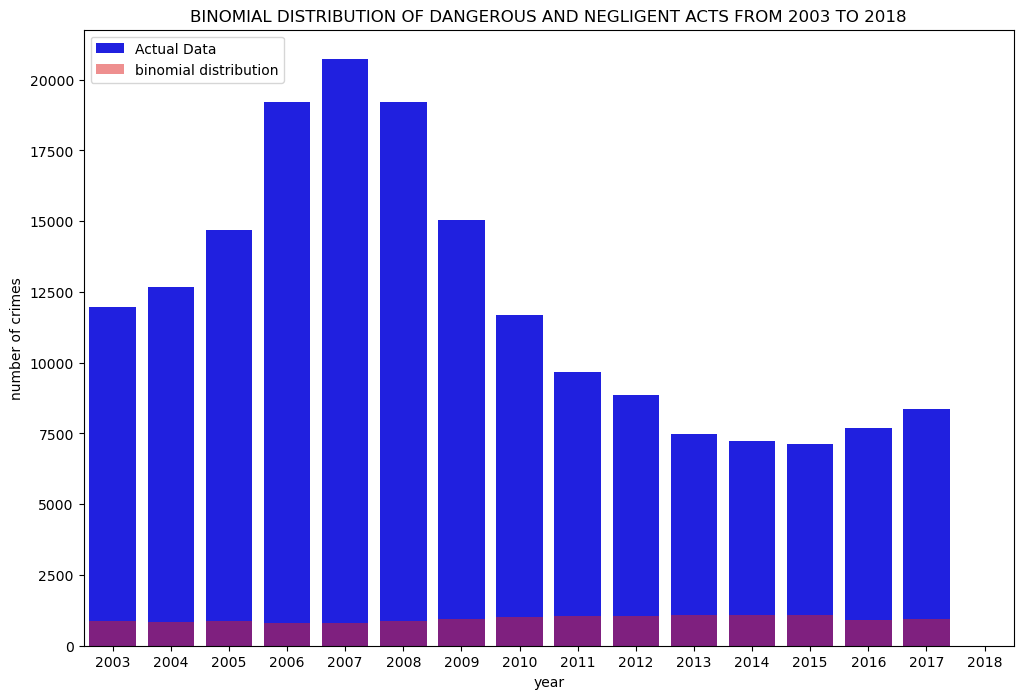


**Discrete Distribution:** We choose the discrete distribution not the continuous distribution because in this type of datasets crimes are discrete events that can be countered as a whole numbers.For example, you can have 3,5,100 instances of crime, but not 3.5 or 15.7 instances. We choose two distribution Poisson distribution and binomial distribution. In this visualization both the charts are used to compare actual crime data against the theoretical distribution. Together, these analysis provide a comprehensive statistical perspective, helping to understand the occurrence patterns of dangerous or negligent acts over time.

*Poisson Distribution:* The datasets indicates that incidents involving dangerous or negligent acts are a recurring issue within the time frame observed. To analyze the frequency of these incidents, we employ the Poisson distribution, which is apt for examining the number of times an event occurs within a specified interval. The Poisson distribution analysis for “Dangerous or Negligent Acts” over the years 2003 to 2018 has been visualized.in this plot: The blue bar represent the actual number of dangerous or negligent act reported each year.The red line represents the Poisson distribution fit, scaled to compare with actual data. The poison line provides a benchmark to evaluate whether the actual data aligns with what would be expected statistically, based on the mean.



*Binomial Distribution:*The binomial distribution analysis for Dangerous or negligent Acts over the years 2003 to 2018 has been visualized. In this plot The blue bar represent the actual number of Dangerous or Negligent Acts reported each year. The burgundy portion of the bars represents the expected number of crimes according to the binomial distribution. The binomial distribution provides a theoretical estimates based on the probability of a specific occurrence of a crime in a sequence of n independent experiment which in this case it is time periods.



**Machine Learning:**

*Framework:* Our research focused on a datasets encompassing detailed crime reports that include the nature of the offense as well as the corresponding region and Garda division over a multi-year period. The primary objective was to employ the supervised learning models to categorize the offenses based on the available features.

To this end, we deployed two well-established machine learning algorithms: the decision tree classifier and the random forest classifier. The comparative performance of these classifier was evaluated to ascertain the more effective algorithms for our specific datasets.

Furthermore, we undertook a visualization of the data to gain deeper insights and facilitate a more intuitive understanding of the crime patterns. In addition, we performed hyper parameter tuning for that algorithms which will show the best result, aiming to optimize their performance that govern the learning process. The outcomes of these analytical processes provided a clear direction for the selection of the most suitable machine learning algorithm, enhancing the predictive accuracy of our crime classification model.

**Decision Tree classifier:**

*Data Preparation and Feature Selection:* In the machine learning, our analysis began with the preparation of the datasets, which included crime statistics categorized by region and police division over several years. We selected two categorical features,”Region” and “Garda Division”, along with the series of annual data, to serve as the predictors in our model. The variable type of offence was designated as the target outcome, representing the various crimes we aimed to classify.

*Encoding Categorical Variables:* Machine learning models requires numerical input, categorical variables were transformed using one-hot encoding. This technique creates binary columns for each category within a feature,allowing the model to interpret the data correctly.

*Feature and Target Separation: A*fter that, we segregated the predictors(X) from the outcome variable(y). The predictors consisted of all one-hot encoded feature except those associated with the “TYPE OF OFFENCE”. The outcome variable y retained the categorical crime types.

*Data Splitting:* The datasets was then split into training set, which constitutes 70% of the data, and a testing set, which makes up the remaining 30%. the division allows the model to learn from the larger portion of the data and be validated on the unseen subset to evaluate its predictive performance.

*Model Training: T*he model was trained on the training set with a consistent random state to ensure reproductibility of results.

*Accuracy and classification metrics:* In our analysis the decision tree classifier’s performance was quantified using its accuracy metrics, which stood at approximately 27.45%. this metric reflects the ratio of correctly predicted instances to the total predictions made.

**Random Tree classifier:**

In our study, we applied the random forest classifier to our crime datasets. The random forest, an ensemble learning method that builds multiple decision trees and merges their outcome to improve prediction accuracy and control over-fitting, was expected to enhance the predictive model’s performance. Upon training the random forest classifier with the training set, we evaluate its accuracy on the test set. The model has an accuracy of approximately 37.29%, indicating that it correctly predicted the type of offence in about 37 out of every 100 cases.

**Comparison of two models:** comparing these two model in which the decision tree had an accuracy of approx 27.45% and the random forest classifier showed an improvement in predictive accuracy. The improved performance of the random forest classifier suggests it is a more suitable algorithm for our dataset, providing a more accurate and reliable classification of the type of offense based on the given features.

**Hyper Parameter Tuning:** Hyper Parameter is crucial step in machine learning to optimize the model’s parameters for better understanding. We utilized a technique known as a Grid search to systematically work through multiple combinations of parameter values,cross-validating as we went to determine which tune gives the best performance.

*Parameter selection:* We focused on two key parameters: the number of trees in the forest which is “n\_estimator” and the maximum depth of each tree(“max\_depth”). A simplified parameter grid was defined to limit the computational complexity while still exploring a representative set of options. The grid included a choice between using 100 or 50 trees and either allowing the trees to grow without restrictions their growth at a depth of 10*.*

*Grid search Execution:* The grid search was executed using 3-fold cross-validation approach to ensure the model robustness and generalization,utilizing all variables processing cores to expedite computation*s.*

*Results:*The optimization process through grid search determined the ideal parameters for the random forest classifier. The optimal configuration was found to be maximum tree depth of 10 and 100 estimators, leading to the most effective prediction outcomes for the offence types within our datasets. This parameters set achieved a cross- validated accuracy score of approximately 35.33%

**Conclusion:** With optimized parameters, our random forest classifier showed an enhanced ability to classify crime types accurately. The process of hyper parameter tuning and the resulting improvements underscore the value of meticulous model- refinement. By selecting the random forest model and fine- tuning its parameters, we were able to significantly advance the model predictive accuracy and reliability.

**Reference and Citation:** <https://www.kaggle.com/datasets/zazaucd/recorded-crime-in-ireland>

<https://www.kaggle.com/datasets/sameerkulkarni91/crime-in-ireland>

<https://bedford-computing.co.uk/learning/wp-content/uploads/2015/10/Python-for-Data-Analysis.pdf>

<https://www.garda.ie/en/about-us/organisational-structure/>