# 

**IS424 Data Mining and Business Analytics**

**Project Report**

**Title: Predictive Analysis of Crime Hotspots in the US**

**Section : G3**

**Group : Team 8**

**Team Members :**

* **Bryan Bramaskara**
* **Nanda Gian Yong Xin**
* **Nguyen Tuan Minh**
* **Nguyen Vu Truong An**
* **Sebastian Hong Kai Xuan**

**Table of Content**

[**1. Project Abstract**](#_v3javl3lqsab) **3**

[1.1 Problem Statement](#_tmztv275vjrm) 3

[1.2 Motivation](#_jg2vbzmbkskk) 3

[**2. Literature Review**](#_z90wrxwg4nnf) **4**

[2.1 Source 1: Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures (Al-Janabi, Kadhim & Fatlawi, Hayder, 2010)](#_z83lgue0ixxl) 4

[2.2 Source 2: A Review of Data Mining Applications in Crime (Hassani, Hossein & Huang, Xu & Silva, Emmanuel & Ghodsi, Mansi, 2016)](#_hwg5z0ha8lkb) 4

[2.3 Source 3: An Adaptive Approach For Analyzing And Predicting The Crime Location (R, S., & D, E., 2017)](#_smytk1agvska) 4

[**3. Dataset**](#_d0k5q3yrrmlq) **4**

[**4. Data Preparation and Exploratory Data Analysis**](#_z22n1yuocnzz) **5**

[4.1 Data Preprocessing and Data Transformation](#_36bu4z8fywgt) 5

[4.1.1 Handling Missing Values](#_u2hsw9ouca6d) 5

[4.1.2 Outliers Removal](#_v0b2gejyw16c) 6

[4.1.3 Log Transformation](#_sinfsqevfhk6) 6

[4.1.4 Standardization](#_mro7ox5gc5y7) 8

[4.2 Correlation Analysis](#_tnwry9boofcb) 8

[4.3 Data Preparation for Regression Analysis using Multiple Linear Regression Model](#_x9ak0tmqljfu) 9

[4.4 Data Preparation for Binary Classification](#_featiaoc6hlo) 11

[4.5 Data Preparation for Cluster Analysis](#_vi1ac3rf41b1) 11

[**5. Methodologies**](#_qsms9vl55i2) **12**

[5.1 Regression Analysis with Multiple Linear Regression Model](#_zc2ngvgk3760) 12

[5.2 Regression Analysis with Ensemble Methods](#_cypfs7v6aoh7) 13

[5.3 Binary Classification with Ensemble Methods](#_anhmvdc91sia) 15

[5.4 Binary Classification with Other Classical Machine Learning Models](#_g5sautgucw8) 16

[5.5 Cluster Analysis](#_l3lyqgcw42df) 17

[**6. Results and Discussion**](#_28pj5iwxn9ew) **20**

[6.1 Regression Analysis with Multiple Linear Regression Model](#_isp7zy2cwee0) 20

[6.2 Regression Analysis with Ensemble Methods](#_hp08nd687gjp) 22

[6.3 Binary Classification with Ensemble Methods](#_i23teun7a6kh) 23

[6.4 Binary Classification with Other Classical Machine Learning Models](#_4lkczbje569l) 24

[6.5 Cluster Analysis](#_57d0l8ck1ui) 25

[**7. Conclusion and Future Work**](#_ep7blyxmqpp5) **31**

[7.1 Cross Comparison of Models](#_kuyzbr3l88f0) 31

[7.2 Future Works](#_l8bzbiwjkw6v) 31

[7.3 Conclusion](#_49kvsa5q98z9) 33

[**8. References (APA 7 Format)**](#_45uhumvhtps0) **34**

[**9. Appendix**](#_gi995833tw0g) **36**

# 1. Project Abstract

The various communities in the US have diverse demographic makeups. This project aims to conduct an in-depth analysis of the characteristics of the residents of the communities, i.e. social economic data and law enforcement resource allocation, to predict the potential emergence of crime through ensemble learning methods.

## 1.1 Problem Statement

Through utilizing various data mining techniques, we want to take on the challenge of manipulating the crime dataset to discover embedded patterns in different types of attributes present. This enables insight generation to form the basis of predictive modelling, which can be used to assist risk assessment methodologies to aid in the prevention of future crimes.

Below are a few tasks which we have defined when exploring this dataset (non-exhaustive list):

* To understand the contributing factors that can affect crime rates.
* To discover inter and intra-relationships between crime records within the same cluster and crime records from different clusters. This is useful in helping law enforcement officers to focus their efforts on specific areas with specific crime types.
* Predict whether a particular community or state will have a high or low crime rate based on external factors and variables analysed in the analysis.
* Predict the level of association of the analysed model with each state’s characteristics in the present as the dataset was donated in 2011.

## 1.2 Motivation

Looking at global crime rates, the Americas have a rate that is higher than the global average, with the US prison population making up more than 20% of the global prison population, in spite of the fact that the US population accounts for less than 5% of the global population (Statistica Research Department, 2020). These offenders come from various backgrounds and have different crime scenes. With such variability in crime records, there is a need to uncover hidden patterns and use them to identify areas prone to criminal activity to prevent crimes from occurring (“Predictive Policing: Stopping Crime Before it Starts”, 2020).

Although there has been rigorous work dedicated to improve the accuracy of predictive policing, several drawbacks have been identified, such as being overly data-driven so much so that accuracy is compromised. Therefore, our project aims to minimize this drawback by using data mining techniques and methodologies to inculcate context into our predictions. Through Predictive Analyses, we hope to improve resource deployment and lower crime rates, which will translate into lower costs for the police.

# 2. Literature Review

## 2.1 Source 1: *Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures (Al-Janabi, Kadhim & Fatlawi, Hayder, 2010)*

This conference paper puts forward a data warehouse framework for crime data analysis and detection using different data mining techniques, such as Classification, Prediction, Link Analysis and Association. The paper aims to assist in making criminal forecasts, finding reasons and relationships behind crimes, and mapping criminal networks. Its data mining techniques are applied on a repository in which different sources of data are integrated. From which Data Marts are created to draw out depending on the requirements of the analysis. We can leverage on the Data Mart architecture highlighted in this paper to further explore the crime dataset thoroughly.

## 2.2 Source 2: *A Review of Data Mining Applications in Crime (Hassani, Hossein & Huang, Xu & Silva, Emmanuel & Ghodsi, Mansi, 2016)*

This paper presents a concise review and efficiency of various data mining techniques that have been implemented for crime analysis. It offers insights on data mining techniques most frequently adopted for crime analysis, against the backdrop of Big Data. With reference to this paper, we can apprise ourselves on research that has already been carried out and use it to further our understanding of the crime dataset, along with rigorous analysis.

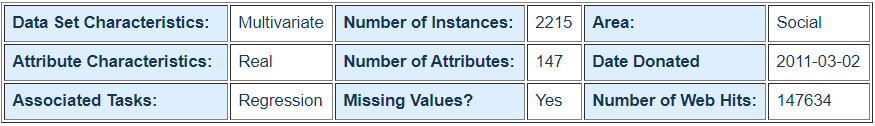
## 2.3 Source 3: *An Adaptive Approach For Analyzing And Predicting The Crime Location (R, S., & D, E., 2017)*

This paper divides data mining approaches commonly employed in crime analysis into different categories and provides in-depth explanation of each. It lists out the different approaches used in past researches and also includes expert opinions on some methodologies. From this, we can use it to avoid pitfalls in our own research and draw on insights stated in this paper for novel ways to explore the dataset by using methodologies not explicitly touched on in class.

# 3. Dataset

The dataset for this project is retrieved from the following link: <https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized>. It describes the distribution of crime rates in different states in the US, along with attributes covering socioeconomic status such as median income and the number of violent crimes per population.

With reference to Figure 1, the crime dataset consists of many different types of attributes, including numerical and nominal variables. It has a total of 2215 observations with 147 attributes. There are columns with many missing values, such as NumInShelters and murdPerPop. These columns will be handled in Section 4 where data preparation and preprocessing will be performed.



*Figure 1: Overview of crime dataset*

# 4. Data Preparation and Exploratory Data Analysis

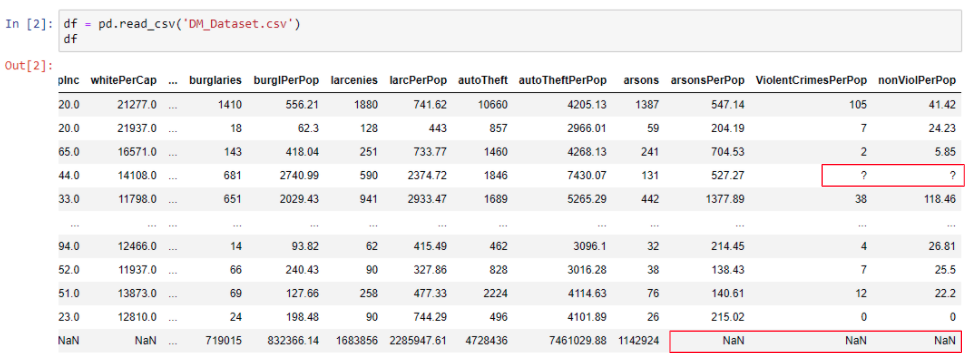
Data cleaning is required as not all columns and rows are suitable for analysis. Firstly, missing values that are indicated with ‘?’ in the dataset are handled in 4.1.1. Additionally, outlier analysis is performed, where rows with Z-score that are not within ±3 standard deviation are removed. Log transformation is also carried out to make highly skewed columns less skewed, which is valuable in making patterns in the data more visible and interpretable. Lastly, standardization is performed to place all variables on a common scale for easy and fair comparison in our analyses. After all data preparation steps are carried out, the dataset contains 1593 rows and 49 attributes.

## 4.1 Data Preprocessing and Data Transformation

### 4.1.1 Handling Missing Values

There are missing values for several columns in some rows, some represented with ‘?’ values and some with NaN values (see Figure 2). To ensure all missing values in the dataset are represented in the same way, ‘?’ values are modified to np.NaN by changing the relevant columns to numeric data type, except for the State attribute (see Figure 3).

Then, depending on the state that a community belongs to, the NaN values are imputed to average values of other communities residing within the same state. In the case where there are no other communities belonging to the same state containing non-null values, the NaN values are replaced with the national average instead, which is computed from the entire dataset across all communities (see Figure 4).



*Figure 2: Overview of dataset showing missing values*



*Figure 3: Modification of all column types to numeric data type*



*Figure 4: Imputation of missing values with the state average or national average*

### 4.1.2 Outliers Removal

Observations with Z-score > 3 or Z-score < -3 are removed as they are interpreted as outliers. With reference to Figure 5, only observations within ±3 standard deviations are kept. After removing the outliers, the number of rows is reduced from 2215 to 1593.

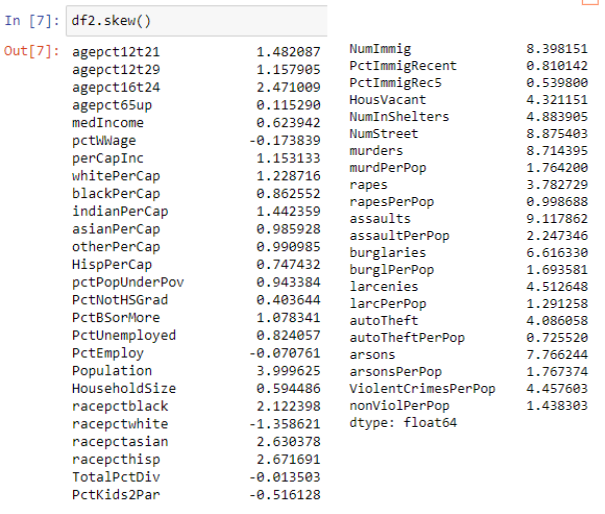


*Figure 5: Removal of outliers*

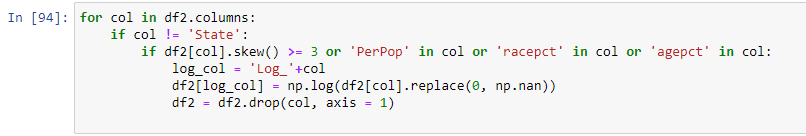
### 4.1.3 Log Transformation

The degree of skewness of each attribute is investigated. For this project, having a skewness value of within ±1 is considered to be acceptable; any value beyond the lower and upper limits are highly skewed. Based on Figure 6, there are multiple columns with significantly high skewness, such as NumInShelters and murders. Therefore, log transformation is implemented to reduce the overall skewness of the dataset.

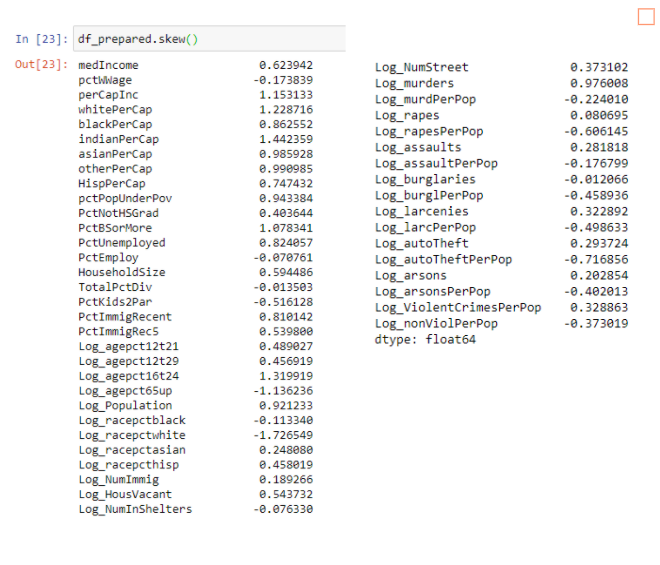
Log transformation is conducted, where all non-null values are replaced with their log values and all ‘0’ values are replaced with np.NaN values (see Figure 7). After successful transformation, Figure 8 shows the new skewness values of all columns, where all the values are now approximately conformed to normality and less skewness is observed.



*Figure 6: Skewness before transformation*



*Figure 7: Log transformation*

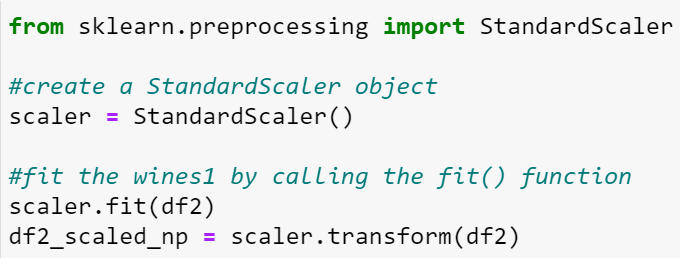


*Figure 8: Skewness after transformation*

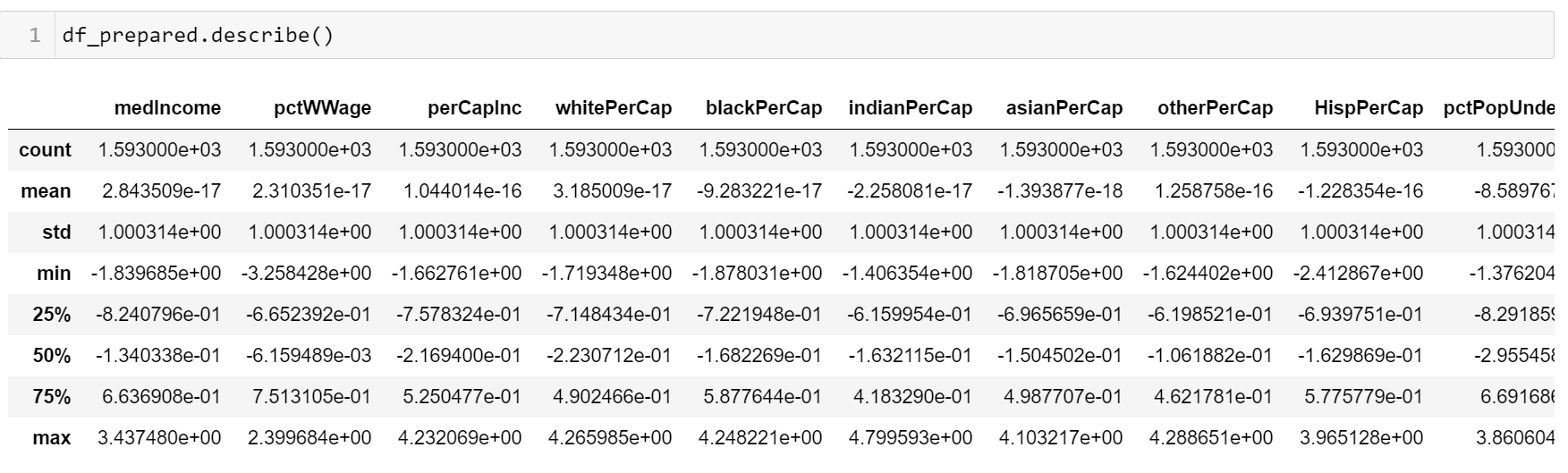
### 

### 4.1.4 Standardization

Using StandardScaler from Scikit-Learn to standardize the attributes and ensure the dataset follows a normal distribution, StandardScaler is first initialized and then fitted with the dataset (see Figure 9). After standardization is performed, all variables now share a common scale (see Figure 10). The dataset is now fully prepared and cleaned, and ready for further analysis.



*Figure 9: Initialization and fitting of StandardScaler to the dataset*



*Figure 10: Statistical details of attributes after standardization*

## 4.2 Correlation Analysis

To identify the strength and direction of association between variables, correlation analysis is carried out. In addition, correlation analysis is also useful in the selection of relevant variables for regression and cluster analyses. The Log\_ViolentCrimesPerPop attribute is chosen as the target variable as it represents the total percentage of crimes per population for each community, which is what this project aims to analyse. The correlation between each of the remaining variables and the target variable are then computed (see Figure 11). The findings derived from the correlation analysis will be used to prepare the data for regression and cluster analyses later on.



*Figure 11: Correlation values of all variables with Log\_ViolentCrimesPerPop*

## 4.3 Data Preparation for Regression Analysis using Multiple Linear Regression Model

The crime dataset is preprocessed before using it for Linear Regression 1 and 2 (2 rounds of linear regression are performed).

**Linear Regression 1**

Firstly, attributes containing more than 50% null values are removed from the dataset. This is because when the proportion of missing values is too high, the results obtained will have less natural variation and this could result in a less effective model (“How to Deal with Missing Data”, n.d.). With reference to Table 1, the 4 attributes listed are removed as they contain more than 50% of null values. Data imputation methods were taken into consideration, however, they are not applicable to the crime dataset as the attributes shown in Table 1 are unique to each state. For instance, imputing missing values using the mean or median of each attribute is not feasible since such values should be unique to each state only and if data imputation was done, it would result in an inaccurate representation of the states in America.

Secondly, the non-numerical attribute, State, is also removed from the dataset as it is not suitable to be used in regression analysis.

Thirdly, Log\_nonViolPerPop is removed as it has a strong correlation value of 0.7940 with Log\_ViolentCrimesPerPop. The higher the absolute correlation value, the more similar the variables are. Hence, Log\_nonViolPerPop is removed while Log\_ViolentCrimesPerPop is kept as the target variable.

Overall, 43 attributes are used in Linear Regression 1.

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Number of null values** | **Percentage of null values (%)** |
| Log\_NumInShelters | 923 | (923/1593)\*100 = 57.9410 |
| Log\_NumStreet | 1242 | (1242/1593)\*100 = 77.9661 |
| Log\_murders | 812 | (812/1593)\*100 = 50.9730 |
| Log\_murdPerPop | 812 | (812/1593)\*100 = 50.9730 |

*Table 1: Attributes with more than 50% of null values*

**Linear Regression 2**

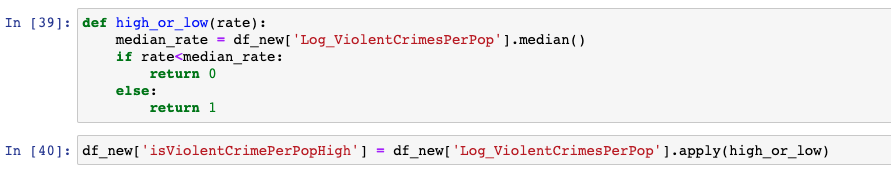
Building on the preprocessing done for Linear Regression 1, further data cleaning is required for Regression 2. Attributes with absolute correlation values of less than 0.4 with Log\_ViolentCrimesPerPop are removed as they signify weak correlation. Table 2 shows the attributes that remain and their respective correlation values with Log\_ViolentCrimesPerPop. After removal, there are 10 predictor variables (see Table 2) and 1 target variable (Log\_ViolentCrimesPerPop).

|  |  |
| --- | --- |
| **Attribute** | **Correlation Value** |
| Log\_assaultPerPop | 0.4137 |
| Log\_NumImmig | 0.4175 |
| Log\_rapes | 0.4673 |
| Log\_HousVacant | 0.5397 |
| Log\_burglaries | 0.5528 |
| Log\_arsons | 0.5660 |
| Log\_assaults | 0.5936 |
| Log\_Population | 0.6222 |
| Log\_larcenies | 0.6296 |
| Log\_autoTheft | 0.6340 |

*Table 2: Attributes with absolute correlation values greater than 0.4*

## 4.4 Data Preparation for Binary Classification

In our analysis, binary classification is implemented to classify each record to either have ‘high’ or ‘low’ crime rate. A record is considered to have ‘high’ crime rate if its corresponding value of Log\_ViolentCrimesPerPop is higher than the median crime rate; a record has ‘low’ crime rate if its value of Log\_ViolentCrimesPerPop is lower than the median. The class ‘high’ is assigned to a value of 1 while the class ‘low’ is assigned to 0 (see Figure 12). Additionally, similar to the preparation steps done in regression analysis, columns with more than 50% of missing values are also omitted.



*Figure 12: Binary classification based on median crime rate*

## 4.5 Data Preparation for Cluster Analysis

Similar to the data preparation for regression analysis in 4.3, Cluster Analysis 1 consists of 43 attributes while Cluster Analysis 2 consists of 11 attributes.

To elaborate in detail, this analysis used the dataset that has been prepared earlier in section 4.3. The State attribute contains the code of the states in America, which will be useful in drawing insights from the cluster analysis based on different states. Therefore, cluster analysis is performed using 4 different copies of data:

* A dataset that drops attributes with more than 50% of missing values and with “State” attribute preserved in Cluster Analysis 1.
* A dataset that drops attributes with more than 50% of missing values and “State” attribute to perform Cluster Analysis 1.
* A dataset that drops attributes with more than 50% of missing values, attributes with absolute correlation values < 40%, with “State” column preserved in Cluster Analysis 2.
* A dataset that drops attributes with more than 50% of missing values, attributes with absolute correlation values < 40% and “State” attribute to perform Cluster Analysis 2.

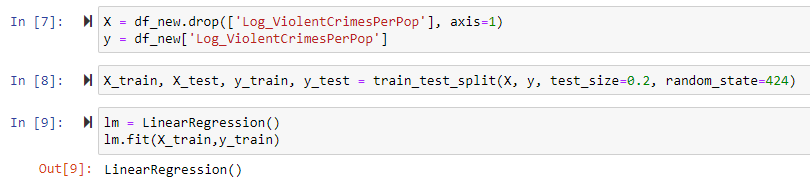
# 5. Methodologies

## 5.1 Regression Analysis with Multiple Linear Regression Model

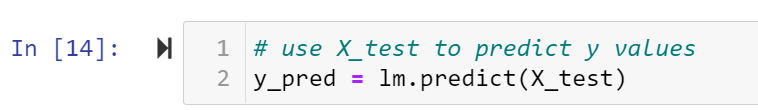
To find out which attributes have a greater impact on crime rates, we adopted Regression Analysis using the Multiple Linear Regression Model. The target variable is Log\_ViolentCrimesPerPop for both Linear Regression 1 and 2. However, the predictor variables used in both runs of linear regression differ: Linear Regression 1 runs the analysis with 43 attributes while Linear Regression 2 runs the analysis with 11 attributes. The main difference between the 2 runs of regression analysis is that Linear Regression 2 only contains attributes that have a moderate to strong correlation with Log\_ViolentCrimesPerPop while Linear Regression 1 includes all relevant attributes after data cleaning is done. By running 2 rounds of regression analyses, we want to determine the most important factors that affect crime rates in the US, as well as to investigate whether there is a significant difference in results if weakly correlated variables are removed.

**Linear Regression 1**

There are 43 attributes in this analysis after removing attributes with more than 50% null values, attributes that are non-numerical and attributes that are similar to the target variable. Firstly, the predictor variables and target variable are defined in Figure 13. Using train\_test\_split from sklearn, the dataset is split into train and test datasets, with the test dataset having a size of 0.2 compared to the original dataset. Random state is standardized to 424 to ensure that repeating the analysis will return the same results. Next, the linear regression classifier is initialized and fitted to the train dataset. Then, the model is used to predict the y values based on x test data (see Figure 14).



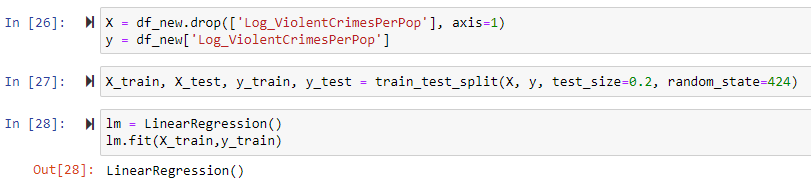
*Figure 13: Linear Regression Classifier 1*

**

*Figure 14: Using the model to predict y values from x test data*

**Linear Regression 2**

There are 11 attributes in Linear Regression 2 after removing attributes with absolute correlation values of less than 0.4 (weak correlation). Then, the predictor variables and target variable are defined in Figure 15. Using train\_test\_split from sklearn, the dataset is split into train and test datasets, with the test dataset having a size of 0.2 compared to the original dataset. Random state is standardized to 424 to ensure that repeating the analysis will return the same results. Next, the linear regression classifier is initialized and fitted to the train dataset. The model is then used to predict the y values based on x test data (see Figure 16).



*Figure 15: Linear Regression Classifier 2*



*Figure 16: Using the model to predict y values from x test data*

## 5.2 Regression Analysis with Ensemble Methods

Regression analysis is carried out using ensemble methods to predict the value of the target variable Log\_ViolentCrimesPerPop based on values of the predictor variables. To reduce variance and improve the stability and accuracy of regression analysis, bagging and boosting are implemented together with regression analysis. In addition, random forest is also used to increase the accuracy of the prediction results obtained. Random forest is run 2 times - with and without random search cv to determine whether there is any significant difference in results between the 2 instances.

The ensemble methods performed are:

* Bagging
* Boosting
  + Gradient Boosting: XGBoost , LightGBMBoost
  + Adaptive Boosting: AdaBoost
* Random Forest
  + Without Random Search CV
  + With Random Search CV

Using train\_test\_split from sklearn, the dataset is split into train and test datasets, with the test dataset having a size of 0.2 compared to the original dataset. Random state is standardized to 424. Additionally, R2 will be used as a measure of accuracy for the fitted regressor to evaluate the performance of the model. For better evaluation of accuracy, 10-fold cross validation is also utilized using cross\_val\_score in sklearn.

Table 3 shows the code snippets used to initialize the regressors for each of the 6 models.

|  |  |
| --- | --- |
| **Model** | **Regressor Model** |
| Bagging | from sklearn.ensemble import BaggingRegressor  bagging\_regressor = BaggingRegressor() |
| XGBoost | import xgboost as xgb  data\_dmatrix = xgb.DMatrix(data=X,label=y)  xg\_reg = xgb.XGBRegressor(objective ='reg:linear', colsample\_bytree = 0.3, learning\_rate = 0.1,  max\_depth = 3, alpha = 10, n\_estimators = 200) |
| LightGBM | from lightgbm import LGBMRegressor  LGBM\_reg = LGBMRegressor() |
| AdaBoost | from sklearn.ensemble import AdaBoostRegressor  ada\_rgr = AdaBoostRegressor() |
| Random Forest | from sklearn.ensemble import RandomForestRegressor  regressor = RandomForestRegressor() |
| Random Forest tuned with Random Search |  |

*Table 3: Initialization of regressors with ensemble methods*

## 5.3 Binary Classification with Ensemble Methods

Binary classification is carried out using ensemble methods to predict the level of crime rate of each observation (whether it is high or low) given the values of the predictor variables. To reduce variance and improve the stability and accuracy of the analysis, bagging and boosting are implemented together with classification. In addition, random forest is also used to increase the accuracy of the prediction results obtained. Random forest is run 2 times - with and without random search cv to determine whether there is any significant difference in results between the 2 instances.

The ensemble methods performed are:

* Bagging
* Boosting
  + Gradient Boosting: XGBoost, LightGBMBoost
  + Adaptive Boosting: AdaBoost
* Random Forest
  + Without Random Search CV
  + With Random Search CV

Using train\_test\_split from sklearn, the dataset is split into train and test datasets, with the test dataset having a size of 0.2 compared to the original dataset. Random state is standardized to 424. Additionally, accuracy score will be used to evaluate the performance of the model. For better evaluation of accuracy, 10-fold cross validation is also utilized using cross\_val\_score in sklearn.

Table 4 shows the code snippets used to initialize the classifiers for each of the 6 models.

|  |  |
| --- | --- |
| **Model** | **Classifier Model** |
| Bagging | from sklearn.ensemble import BaggingClassifier  bagging\_clf = BaggingClassifier() |
| XGBoost | data\_dmatrix = xgb.DMatrix(data=X,label=y)  import xgboost as xgb  xg\_classifier = xgb.XGBClassifier(learning\_rate = 0.01) |
| LightGBM | from lightgbm import LGBMClassifier  LGBM\_clf = LGBMClassifier() |
| AdaBoost | from sklearn.ensemble import AdaBoostClassifier  ada\_clf = AdaBoostClassifier() |
| Random Forest | from sklearn.ensemble import RandomForestClassifier  rf\_clf= RandomForestClassifier() |
| Random Forest tuned with Random Search |  |

*Table 4: Initialization of classifiers with ensemble methods*

## 5.4 Binary Classification with Other Classical Machine Learning Models

Binary classification is also carried out using other classical machine learning models. In addition to traditional methods such as decision trees and logistic regression, other methods such as perceptron and support vector machine are also applied to solve the classification problem. Perceptron is a single-layer neural network that is often used for the supervised learning of binary classifiers, while support vector machine helps to split the data into 2 groups of points in space via either a linear or non-linear hyperplane. In essence, both methods are very useful in tackling classification problems.

The classical methods are implemented using Python ML tool called Scikit-Learn:

* Decision Tree: sklearn.tree.DecisionTreeClassifier
* Logistic Regression: sklearn.linear\_model.LogisticRegression
* Perceptron: sklearn.linear\_model.Perceptron
* Support Vector Machine (SVM): sklearn.svm

Similar to ensemble methods, the dataset is also split into 80% train data and 20% test data, with random state set to 424. 10-fold cross validation is also implemented for better evaluation of the results.

Table 5 shows the code snippets used to initialize the classifiers for each of the 6 models.

|  |  |
| --- | --- |
| **Model** | **Classifier Model** |
| Decision Tree | from sklearn.tree import DecisionTreeClassifier  dtree = DecisionTreeClassifier(criterion = 'entropy', max\_depth = 2) |
| Logistic Regression | from sklearn.linear\_model import LogisticRegression  logreg = LogisticRegression() |
| Perceptron | from sklearn.linear\_model import Perceptron  PPT = Perceptron() |
| SVM (linear) | from sklearn import svm  SVM\_linear = svm.SVC(kernel='linear') |
| SVM (nonlinear) | from sklearn import svm  SVM\_rbf = svm.SVC(kernel='rbf') |

*Table 5: Initialization of classical classifiers*

## 5.5 Cluster Analysis

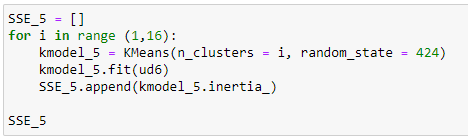
Cluster analysis is an unsupervised learning method performed to find out the inter and intra relationship among and within clusters by grouping them with similar features. With clustering, the shared characteristics of the communities belonging to each cluster can be clearly seen and analysed. Through cluster analysis, this project aims to derive the key contributing factors resulting in a community having a low or high crime rate.

**Cluster Analysis 1**

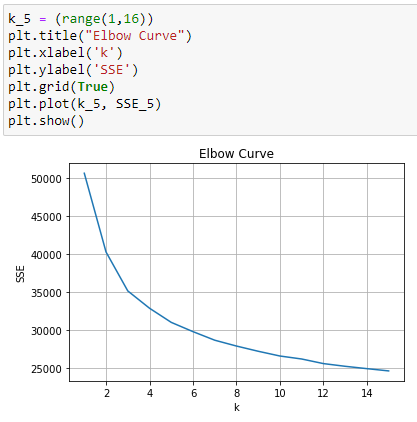
Cluster Analysis 1 is performed using 43 attributes, similar to Linear Regression 1. Firstly, a simple for-loop is run to obtain the Sum of Squared Errors (SSE) for different k values (see Figure 17). The SSE and k values are then plotted onto a graph called the Elbow Curve, which is used to derive the most suitable number of clusters or in other words, the optimal value of k.

In the selection of the number of clusters, the Elbow Method is employed by observing the decrease in SSE score with a range of cluster sizes, followed by picking the elbow of the curve as the number of clusters to be used. From Figure 18, any cluster size bigger than or equivalent to 3 is suitable. However, it can be seen that k=5 is the optimal cluster size as there is a steep decrease in SSE value from k=3 to k=5 of approximately 4000 units. As cluster size increases further from 6 to 15, the decrease in SSE value is less significant and proportionate as compared to the cluster size. For instance, as k increases from 6 to 15, SSE is reduced by approximately 6000 units only. Therefore, using any value of k that is greater than 5 will not be able to return meaningful insights from cluster analysis.

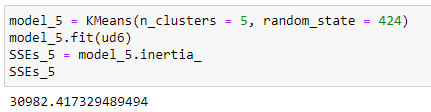
When k=5, the model is able to achieve an SSE score of 30982.4173 (see Figure 19). Using k=5, initial centroids are selected first, followed by cluster centers. This is repeated multiple times until the SSE remains relatively constant.



*Figure 17: K-means clustering for-loop code snippet*



*Figure 18: Elbow Curve plotting SSE against cluster size*



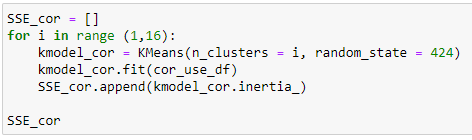
*Figure 19: SSE score for k=5 in Cluster Analysis 1*

**Cluster Analysis 2**

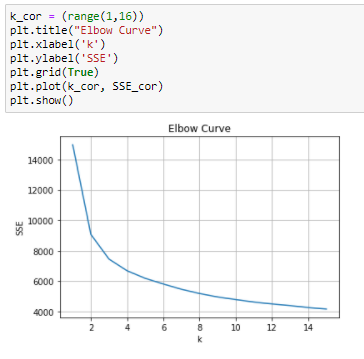
Cluster Analysis 2 is performed with 13 attributes. Similar methodology implemented in Cluster Analysis 1 is also used in Cluster Analysis 2. Firstly, a simple for-loop is run to obtain the SSE values for different k values (see Figure 20). The SSE and k values are then plotted onto a graph called the Elbow Curve, which is used to derive the most suitable number of clusters or in other words, the optimal value of k.

In the selection of the number of clusters, the Elbow Method is employed by observing the decrease in SSE score with a range of cluster sizes, followed by picking the elbow of the curve as the number of clusters to be used. From Figure 21, k=3 is the optimal cluster size as there is a steep decrease in SSE value from k=1 to k=3 of approximately 8000 units. As cluster size increases further from 4 to 15, the decrease in SSE value is less significant and proportionate as compared to the cluster size. Therefore, using any value of k that is greater than 3 will not be able to return meaningful insights from cluster analysis.

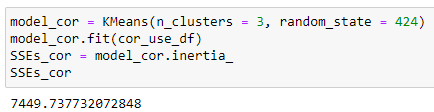
When k=3, the model is able to achieve an SSE score of 7449.7377 (see Figure 22). Using k=3, initial centroids are selected first, followed by cluster centers. This is repeated multiple times until the SSE remains relatively constant.



*Figure 20: K-means clustering for-loop code snippet*



*Figure 21: Elbow Curve for cluster numbers vs SSE score*



*Figure 22: SSE score for k=3 in Cluster Analysis 2*

# 6. Results and Discussion

## 6.1 Regression Analysis with Multiple Linear Regression Model

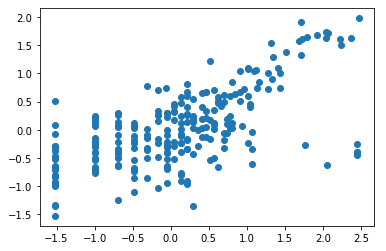
**Linear Regression 1**

After learning the train data, the model is used to predict y values based on x test values. The scatter plot in Figure 23 shows the predicted y values plotted against the actual y test values. From the scatter plot, it can be seen that for smaller values of y test, the data points are plotted vertically on top of each other, which means that several predicted y values that are different are supposed to have the same y values according to the test data. Overall, there appears to be a positive linear relationship between predicted y values and actual y test values.

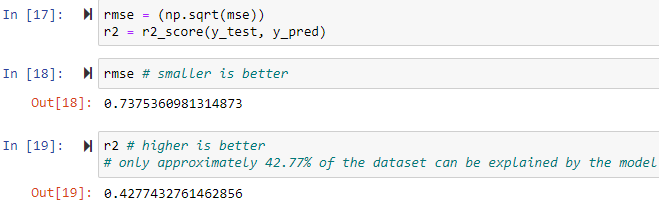
Based on Figure 24, Root Mean Squared Error (RMSE) and R-Squared (R2) are 0.7375 and 0.4277 respectively. RMSE is considerably large, which means that the regression model failed to account for important features in the dataset. R2 is considerably small, which means that only 42.77% of the dataset can be explained by the model.

With reference to Figure 25, 10-fold cross validation score is calculated to be 0.4222, which is close to the R2 value. Therefore, approximately 42.22%~42.77% of the dataset can be explained by the model. Since cross validation score prevents overfitting while ensuring a good estimate of performance, it is able to give a less biased estimate of the model as compared to RMSE and R2. Hence, in our evaluation of the regression model, we will place greater emphasis on the cross validation score. As both R2 and cross validation score are relatively low, most attributes in the dataset do not follow a linear regression model.

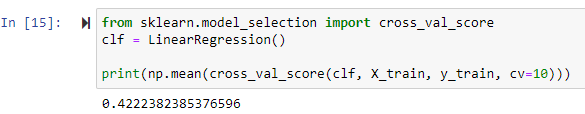
Looking at the residual plot in Figure 26, there is no violation of any assumption of linear regression.



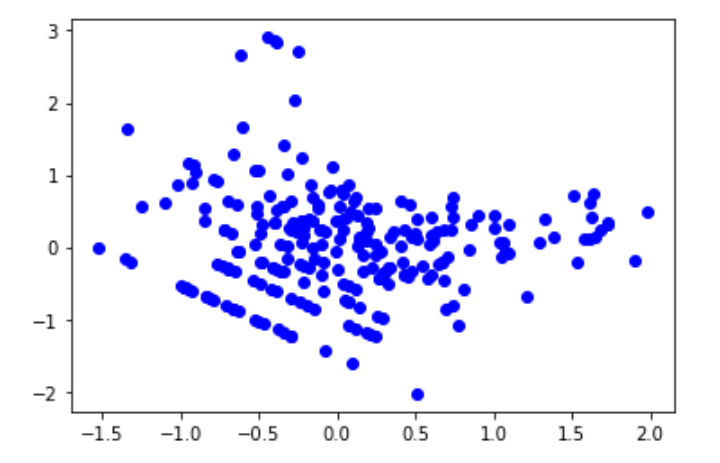
*Figure 23: Scatter plot of y\_pred against y\_test of Linear Regression 1*

**

*Figure 24: RMSE and R2**of Linear Regression 1*

**

*Figure 25: 10-fold cross validation score of Linear Regression 1*

**

*Figure 26: Residual plot of Linear Regression 1*

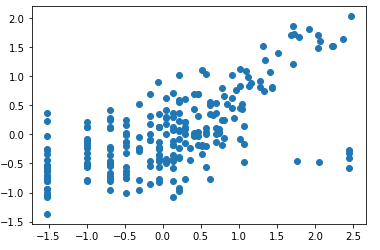
**Linear Regression 2**

After learning the train data, the model is used to predict y values based on x test values. The scatter plot in Figure 27 shows the predicted y values plotted against the actual y test values. From the scatter plot, it can be seen that for smaller values of y test, the data points are plotted vertically on top of each other, which means that several predicted y values that are different are supposed to have the same y values according to the test data. Overall, there appears to be a positive linear relationship between predicted y values and actual y test values.

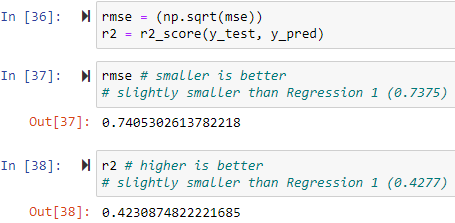
Based on Figure 28, RMSE and R2 are 0.7405 and 0.4231 respectively. RMSE is slightly larger than that of Linear Regression 1 while R2 is slightly smaller than that of Linear Regression 1. Looking at the R2 value, only 42.31% of the dataset can be explained by the model.

With reference to Figure 29, 10-fold cross validation score is calculated to be 0.4542, which is higher than that of Linear Regression 1. Therefore, approximately only 42.31%~45.42% of the dataset can be explained by the model. As both R2 and cross validation score are relatively low, most attributes in the dataset do not follow a linear regression model.

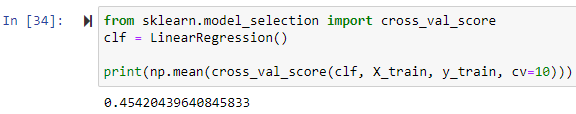
Looking at the residual plot in Figure 30, there is no violation of any assumption of linear regression.



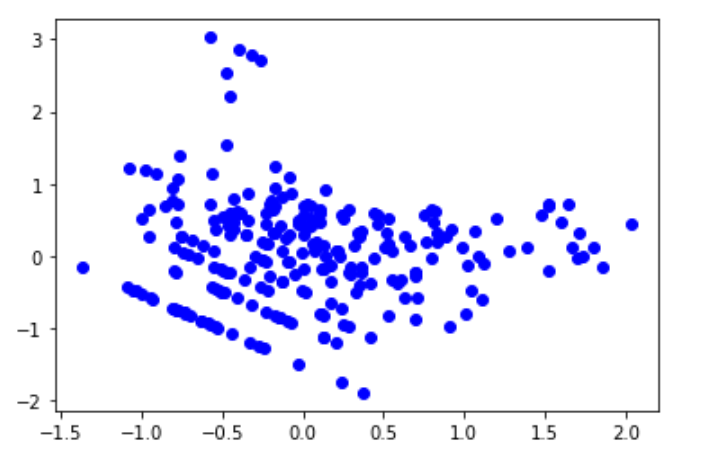
*Figure 27: Scatter plot of y\_pred against y\_test of Linear Regression 2*

**

*Figure 28: RMSE and R2**of Linear Regression 2*

**

*Figure 29: 10-fold cross validation score of Linear Regression 2*

**

*Figure 30: Residual plot of Linear Regression 2*

## 6.2 Regression Analysis with Ensemble Methods

According to the R2 scores obtained from the test set, tuned Random Forest, XGBoost and bagging regressors are the top 3 best performing models (see Table 6). However, the bagging regressor showed the greatest difference in cross validation score and R2 score, with the latter being higher. Hence, it is highly likely that the model is overfitted. The AdaBoost regressor consistently gave the lowest scores for both cross validation score and R2 score. This is possibly due to the fact that the model is underfitted and failed to capture the salient patterns of the dataset. The Random Forest regressor tuned with Random Search CV is the best performing model overall as it achieved the highest cross validation score and R2 score, which means that there is minimal or no underfitting or overfitting. XGBoost and untuned Random Forest regressors gave the approximate median scores for both cross validation score and R2 score.

|  |  |  |
| --- | --- | --- |
| **Model** | **Cross Validation Accuracy** | **R2 Score on Test Set** |
| Bagging | 0.3928 | 0.4824 |
| Random Forest | 0.4058 | 0.4330 |
| Random Forest tuned with Random Search | 0.4535 | 0.5154 |
| AdaBoost | 0.3552 | 0.3954 |
| LightGBM | 0.3970 | 0.4381 |
| XGBoost | 0.4329 | 0.5071 |

*Table 6: Cross validation scores and R2 scores for regression analysis with ensemble methods*

## 6.3 Binary Classification with Ensemble Methods

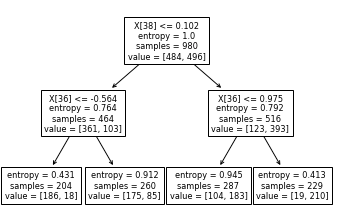
According to the accuracy scores obtained from the test set, tuned Random Forest and LightGBM classifiers are the top 2 best performing models (see Table 7). However, the LightGBM classifier showed the greatest difference in cross validation score and accuracy score, with the latter being higher. This could be because the model focuses too much on noise in the dataset and is overfitted. The Random Forest classifier tuned with Random Search CV achieved the highest scores for both cross validation score and accuracy score, which shows that there is minimal or no underfitting or overfitting. The AbaBoost classifier consistently gave the lowest cross validation score and accuracy score because it is underfitted and failed to capture the underlying patterns of the dataset. XGBoost and bagging classifiers gave the approximate median scores for both cross validation score and accuracy score.

|  |  |  |
| --- | --- | --- |
| **Model** | **Cross Validation Accuracy** | **Accuracy on Test Set** |
| Bagging | 0.7044 | 0.7429 |
| Random Forest | 0.7273 | 0.7347 |
| Random Forest tuned with Random Search | 0.7444 | 0.7918 |
| AdaBoost | 0.6807 | 0.6980 |
| LightGBM | 0.6905 | 0.7918 |
| XGBoost | 0.7167 | 0.7510 |

*Table 7: Cross validation scores and accuracy scores for binary classification with ensemble methods*

## 6.4 Binary Classification with Other Classical Machine Learning Models

After running the classical machine learning models, based on cross validation score, Decision Tree (see Figure 31) emerged as the best performing model with a score of 0.7428 (see Table 8). Logistic Regression scored well with an accuracy score of 0.7469 (see Table 8) but only managed to achieve a cross validation score of 0.7119 in Table 8, indicating that the classifier is overfitted and only performed well on certain test sets. SVM\_rbf achieved a higher score after 10-fold cross validation than SVM\_linear (see Table 8), indicating that the hyperplane cutting through 2 groups of binary classification points in space is most likely non-linear.



*Figure 31: Decision Tree plotted by sklearn.tree*

|  |  |  |
| --- | --- | --- |
| **Model** | **Cross Validation Accuracy** | **Accuracy on Test Set** |
| Decision Tree | 0.7428 | 0.7061 |
| Logistic Regression | 0.7119 | 0.7469 |
| Perceptron | 0.6047 | 0.6857 |
| SVM\_linear | 0.7078 | 0.7469 |
| SVM\_rbf | 0.7151 | 0.7184 |

*Table 8: Cross validation scores and accuracy scores for binary classification with classical models*

## 6.5 Cluster Analysis

**Cluster Analysis 1**

The heat map shows the coordinates of the cluster centers of each cluster, which is denoted using a colour gradient scale ranging from -2.0 to +2.0, with the colour gradually darkening from light yellow to dark blue as the value of the coordinate increases (see Appendix B). The heat map illustrates Cluster 0 and Cluster 4 with the lowest crime rates. However, the coordinate of the medIncome attribute representing median income differs significantly between the 2 clusters. From this, it can be inferred that median income may not be a significant characteristic to predict the crime rate in a community.

On the flipside, Cluster 1 and Cluster 2 have the highest crime rates. However, the coordinate of the medIncome attribute in Cluster 2 is significantly lower than that in Cluster 1, where the values are -1.10 and -0.54 respectively. This indicates that there may be other attribute(s) that contributed more significantly to the high crime rates.

Cluster 3 produces approximately median values for most of the attributes, where the coordinate of the medIncome attribute is 0.32, which is a comparatively moderate level of income. However, upon closer analysis, a notable difference is that Cluster 3 has a comparatively high crime rate for arsons.

Table 9 shows the distinct characteristics of each cluster and the overall crime rate compared to other clusters. From the heat map in Appendix B, it can also be deduced that Cluster 2 represents the group of communities in a state with relatively high crime rate for raping, assault, burglary, larcenies and auto theft as compared to Clusters 0, 3 and 4.

In addition, Clusters 1 and 2 have higher crime rates compared to the other clusters, with low values for the medIncome attribute. This accounts for its higher unemployment rate and therefore, higher rate of population under poverty. Statistics show that the state of Mississippi has the 2nd lowest median income in the United States in 2020 (Williams, 2020). This proves that the results obtained from Cluster Analysis 1 aligns with the present context that Clusters 0 and 2 have comparatively low income and Cluster 2 has the lowest income out of all clusters. In retrospect, the model is able to correctly group some communities in Mississippi states to these 2 clusters (see Table 10).

Furthermore, Cluster 0 has relatively low income and the highest number of white race. Even though the analysis of the model was done using data donated in 2011, the result of the cluster analysis for Cluster 0 is still relevant and applicable in the current context. This can be proven as the state Maine has the highest percentage of white race in the United States of 94.6% (United States - White Population Percentage by State) with the lowest violent crime rate (i.e. rank 50th out of 50 states) in the whole United States (Stebbins, 2020) while Ohio has a relatively high number of white race of 81.9% (United States - White Population Percentage by State) and ranked 35th out of 50 states (Stebbins, 2020), causing this cluster to have the lowest violent crime rate overall.

Lastly, Clusters 3 and 4 have comparatively high income, with the latter having the highest income compared to all clusters. This is supported statistically whereby Illinois state was ranked as the 17th richest state in the United States, followed by California at 14th, New Jersey at 8th and Massachusetts at 7th (Knueven, 2019). In our analysis results, all these states are grouped together in Clusters 3 and 4. In particular, New Jersey in Cluster 4 has comparatively high income according to the cluster analysis and this is supported by the fact that New Jersey was ranked to be one of the safest states (45th out of 50 states) with the lowest violent crime rate in 2020 (Stebbins, 2020).

Hence, it is clear that the results obtained from Cluster Analysis 1 is highly applicable even in today’s context in 2020, and this has been supplemented by various sources in the paragraphs above.

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **Distinct Characteristics** | **Remarks** |
| 0 | * **Relatively low** income compared to Clusters 3 and 4 * **Small** proportion of residence who obtained bachelor degree compared to Clusters 3 and 4 * **Largest** no. of white race * **Smallest** no. of black race | **Lowest** crime rate |
| 1 | * **Low** income compared to Clusters 3 and 4 * **High** population under poverty compared to Clusters 0, 3 and 4 * **Comparatively high** rate of unemployed compared to Clusters 0, 3 and 4 * **Small** percentage of 2 parents living with their kid under the same house compared to Clusters 0, 3 and 4 * **High** divorce rate compared to Clusters 0 and 4 * **More densely** populated compared to Clusters 0, 2 and 4 * **Largest** no. of black race * **Smallest** no. of white race * **Largest** no. of immigrants * **Largest** no. of vacant houses | **Highest** crime rate |
| 2 | * **Lowest** income * **Highest** population under poverty * **Smallest** proportion of residence who obtained bachelor degree * **Highest** unemployment rate * **Highest** divorce rate * **Less densely** populated state compared to Clusters 1 and 3 * **Large** no. of black race compared to Clusters 0 and 4 * **Small** no. of white race compared to Clusters 0 and 4 | **Comparatively High** crime rate for raping, assault, burglary, larcenies and auto theft |
| 3 | * **Relatively high** income compared to Clusters 0, 1 and 2 * **Large** proportion ofresidence who obtained bachelor degree compared to Clusters 0, 1 and 2 * **Low** unemployment rate compared to Clusters 0, 1 and 2 * **Large** no. of immigrant compared to Clusters 0 and 2 * **Large** no. of asian race compared to Clusters 0 and 2 * **Large** no. of hispanic race compared to Clusters 0, 2 and 4 | **Comparatively High** crime rate for arson |
| 4 | * **Highest** income * **Largest** proportion of residence who obtained bachelor degree * **Smallest** population under poverty * **Highest** employment rate * **Highest** percentage of 2 parents living with their kid under the same house * **Lowest** divorce rate * **Comparatively small** no. of black race compared to Clusters 1, 2 and 3 | **Low** Crime Rate |

*Table 9: Summary of derived distinct characteristics of different clusters*

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **States Statistic** | **Communities Count** |
| 0 | Mississippi (9.5%), Maine (8.8%), Ohio (8%), Pennsylvania, etc. | 274 |
| 1 | California (23.8%), Texas (10.9%), Florida (6.2%) , North Carolina, etc. | 193 |
| 2 | Texas (13.2%), Mississippi (6.2%), North Carolina (6.2%), Oklahoma, etc. | 258 |
| 3 | California (23.6%), New Jersey (13.5%), Mississippi (7.3%), Texas, Massachusetts, etc. | 288 |
| 4 | New Jersey (18.8%), Pennsylvania (9.9%), Illinois (9%), Maine , Connecticut, etc. | 212 |

*Table 10: Cluster demographic with states*

**Cluster Analysis 2**

Similarly, the heat map in Cluster Analysis 2 shows the same set of attributes for different clusters as Cluster Analysis 1 (see Appendix C). The main difference between the 2 analyses is that Cluster Analysis 1 uses 5 clusters while Cluster Analysis 2 uses 3 clusters. The clusters in this analysis are separated by extreme values, where Cluster 0 has the lowest crime rate while Cluster 1 has the highest crime rate. The crime rate of Cluster 2 is in between Clusters 0 and 1.

Table 11 shows the distinct characteristics of each cluster and the overall crime rate compared to other clusters. It is observed that when a state has low income, a dense population and has a large number of immigrants and vacant houses, this will translate to a higher crime rate. Vacant houses will result in a larger number of burglaries which is depicted in the heat map (see Appendix C) and this relation can be also be seen in the heat map in Cluster Analysis 1 (see Appendix B) where the value for Log\_HousVacant and Log\_burglaries attributes are similar (e.g. in Cluster 0 - Cluster Analysis 1, Log\_HousVacant=-0.53 and Log\_burglaries=-0.51).

Similar to Cluster Analysis 1, the states in each cluster represent the distinct cluster characteristics (see Table 12).

Of the 2 cluster analyses, Cluster Analysis 1 provided more useful insights that explained the hidden characteristics of each cluster that contributed to the level of crime rate in a specific state. The results from Cluster Analysis 2 further supported the different attributes contributing to a high crime, which is also supported in Cluster Analysis 1. Therefore, there is a significant difference between the 2 analyses performed: Cluster Analysis 1 provided a more holistic consideration of all attributes while Cluster Analysis 2 gave a more specific and narrowed analysis.

Table 13 shows the key factors that contribute to high crime rates. The impact column shows the degree of impact these attributes had that resulted in higher or lower crime rates.

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **Distinct Characteristics** | **Remarks** |
| 0 | * **Relatively high** income * **Least densely** populated state * **Smallest** no. of immigrants * **Smallest** no. of vacant houses | **Lowest** crime rate |
| 1 | * **Lowest** income * **Most densely** populated state * **Largest** no. of immigrants * **Largest** no. of vacant houses | **Highest** crime rate |
| 2 | * **Relatively low** income compared to Cluster 1 * **More densely** populated state compared to Cluster 0 * **Relatively large** no. of immigrants compared to Cluster 1 * **Relatively large** no. of vacant houses compared to Cluster 1 | **Moderate** crime rate |

*Table 11: Summary of derived distinct characteristics of different clusters*

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **States Statistic** | **Communities Count** |
| 0 | New Jersey (9.5%), Texas, Pennsylvania, Maine, Ohio, Mississippi, Massachusetts, etc. | 518 |
| 1 | California (23.7%), Texas, Florida, Mississippi, etc. | 215 |
| 2 | California (14.8%), New Jersey, Mississippi, Texas, Massachusetts, Florida, Ohio, etc. | 492 |

*Table 12: Cluster demographic with states*

|  |  |  |
| --- | --- | --- |
| **No.** | **Contributing Attributes** | **Impact** |
| 1 | Population in a community/state | High |
| 2 | Population living under poverty | High |
| 3 | Divorce Rate | High |
| 4 | Percentage of kid living with 2 parents under the same house | Low |
| 5 | Number of Immigrants | High |
| 6 | Young adult population (age 12-29) | Relatively High |
| 7 | Number of black race resident | High |
| 8 | Number of white race resident | Low |
| 9 | Number of vacant houses | High |
| 10 | Percentage of Unemployed | High |

*Table 13: Contributing factors of high crime rate*

# 7. Conclusion and Future Work

## 7.1 Cross Comparison of Models

Random Forest tuned with Random Search CV achieved the highest scores for both Classification and Regression. This is attributed to the non-linear dataset being explored. As the total number of instances of the dataset is 1593 with 49 attributes, the analysis is able to run efficiently. Furthermore, Random Forest generates result models which are less prone to overfitting.

However, the model accuracy achieved is limited by the tuning time required to set the hyperparameters. As the parameter tuning relies on experimental results, this project is limited by the amount of time given to conduct enough iterations to achieve the most optimal results. Hence, the results presented by Random Forest tuned with Random Search CV are near optimal but not completely optimal.

From both cluster analyses performed, numerous useful insights pertaining to the characteristics of the clusters of states that contributed to the level of crime rate were gathered. These attributes are:

* Densely Populated State
* High No. of Population Living Under Poverty
* High Rate of Divorce\*
* Low Rate of Children Living with 2 Parents (under the same house)
* High No. of Immigrants
* Relatively High Young Adult Population
* High No. of Black Race Resident\*
* Low No. of White Race Resident\*
* High No. of Vacant Houses
* High Percentage of Unemployed\*

With reference to the above list, attributes with asterisks denote the most distinct attributes that assisted in determining the level of crime rate in a community.

## 7.2 Future Works

**Principal Component Analysis (PCA)**

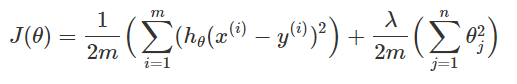
PCA aims to reduce the dimensionality of the data by deriving components to capture the underlying variance of data using orthogonal linear projection. PCA can be viewed as a special scoring method under the SVD algorithm. It produces projections that are scaled with the data variance, and projections of this type are sometimes preferable in feature extraction to the standard non-scaled SVD projections. The PCR idea is to summarize the features by the principle components, which are the combinations with the highest variance (Data Mining - Principal Component (Analysis: Regression) (PCA), n.d.). Although the number of attributes analysed in this project is 49, feature selection is done by manually analysing each attribute through various methods, such as columns containing many missing values. Thus, by using PCA, the first few principal components can be calculated which will help to simplify the model and improve the process of model building and better results can be yielded.

**Collinearity Testing**

While manipulating the predictor variables, it would be helpful to check for correlations between them via the variance inflation factor. A key goal of regression analysis is to isolate the relationship between each predictor variable and the target variable. The regression coefficient represents the mean change in the target variable for every 1 unit change in a predictor variable, paribus ceteris. Hence, by employing collinearity testing before embarking on regression analysis, it can be better ensured that the variables selected are truly independent of each other so that the finals results will be as accurate as possible.

**Regularized Linear Regression**

By doing regularized linear regression, the model is able to provide more explanation about the relationship between evaluation metrics and model complexity. As regularization helps to sort out error terms of the coefficients, complexity is reduced and this generates a model that is not overfitted. Furthermore, regularization of the linear regression model solves the issue of overfitting by reducing the weight given to a particular attribute. As such, this produces a model that retains more features yet does not put undue weight on a particular attribute. Furthermore, regularization is mediated with the parameter ‘λ’ which can be seen in the cost function (see Figure 32).

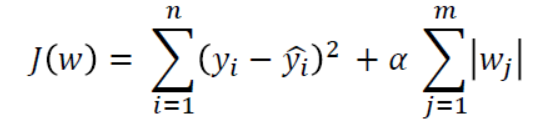


*Figure 32: Cost function of Regularized Linear Regression*

The first term in the cost function is the mean-squared-error term, whilst the additive term multiplies the sum of the square of the parameters (θ) by λ over 2m, where m is the number of training examples (see Figure 32). Since the objective is to minimise J(θ) (min J(θ)), using a large λ will require small values of θj in order to achieve a minima (Upson, 2015).

**Lasso Regression**

Lasso regression is an L1 penalised model where L1 norm of the weight is simply added to the Least-Squares Cost function as shown below (see Figure 33):



*Figure 33: Least-squares cost function*

Increasing the value of ⍺ increases the regularization strength and shrinks the weights of the model (Ph.D., 2020). Therefore, the goal of lasso regression is to determine which predictors are the most important in an analysis. It also minimizes the prediction error. Simultaneously, it enables a simpler model that avoids computing with high dimensional data.

## 7.3 Conclusion

Rising crime rates does not only present itself in the US, but it is also a worldwide phenomenon. However, limited resources and manpower have made it difficult to control crime rates. Henceforth, tackling rising crime rates requires a layered approach, one which consists of conducting analyses to assist in better allocation of resources through a better understanding of each precinct. Despite many works dedicated to predictive policing, there is more to be done to increase the accuracy and efficiency of it. It is also important to minimize the drawbacks of each type of analysis with different models and methods while maximizing the insights that can be drawn from them. In essence, rising crime rate is a worrying trend and this report examines in detail the different methodologies that can be employed to study crime hotspots to value-add past works that have been done to improve the current crime control model, which can be potentially useful in crime prevention.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 8. References (APA 7 Format)

DataCadamia. Data Mining - Principal Component (Analysis: Regression) (PCA). Data Mining - Principal Component (Analysis|Regression) (PCA). <https://datacadamia.com/data_mining/pca>

Federal Bureau of Investigation. (1996). Crime in the United States - 1995: SECTION II Crime Index Offenses Reported . United States.

Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(3), 139–154. <https://doi.org/10.1002/sam.11312>

How to Deal with Missing Data. Master's in Data Science. <https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data/>.

Index Mundi. United States - White Population Percentage by State

<https://www.indexmundi.com/facts/united-states/quick-facts/all-states/white-population-percentage>

Kaur, S., & Singh, W. (2017). Systematic Review of Crime Data Mining, 8(5), 1336–1342.

Knueven, L. (2019, August 12). The average American household earns $61,000 a year. Here are 17 states where the typical resident earns even more. Business Insider. <https://www.businessinsider.com/personal-finance/richest-states-in-the-us-by-median-household-income-2019-8>

Ph.D., B. O. T. (2020, January 8). *LASSO Regression Tutorial*. Medium. <https://towardsdatascience.com/lasso-regression-tutorial-fd68de0aa2a2> .

R, S., & D, E. (2017). An Adaptive Approach For Analyzing And Predicting The Crime Location (thesis). Shodhganga, India.

Redmond, M. (2011, March 2). Communities and Crime Unnormalized Data Set. <https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized>.

Sanders, J. (2020, April 3). Predictive Policing: Stopping Crime Before it Starts. Pappalardo & Pappalardo LLP. <http://pappalardolaw.com/2019/02/predictive-policing-stopping-crime/>.

Statistica Research Department. (2020, January 28). Topic: Crime and punishment around the world. Statista. <https://www.statista.com/topics/780/crime/>.

Stebbins, S. (2020, January 13). Dangerous states: Which states have the highest rates of violent crime and most murders? USA Today. <https://www.usatoday.com/story/money/2020/01/13/most-dangerous-states-in-america-violent-crime-murder-rate/40968963/>

Swadi Al-Janabi, K. B., & Fatlawi, H. K. (2010). Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures.

United States. Bureau of the Census. (1991). *1990 census of population and housing*. Washington, DC: U.S. Dept. of Commerce, Economics and Statistics Administration, Bureau of the Census : For sale by the Supt. of Docs., U.S. G.P.O

Upson, M. (2015, July 26). Regularised linear regression. <https://www.machinegurning.com/rstats/regularised-linear-regression/> .

Williams, W. (2020, September 15). Median Income by State. Investopedia. <https://www.investopedia.com/median-income-by-state-5070640>

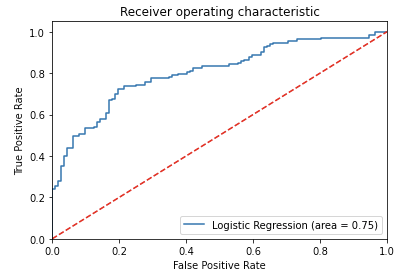
# 

# 

# 

# 9. Appendix

**Appendix A: ROC curve for Logistic Regression**



**Appendix B: Heat map of Cluster Analysis 1**

### 

**Appendix C: Heat map of Cluster Analysis 2**

