Project Report
Analysis and Prediction of Crimes in Boston

CS6220 Data Mining Techniques-Spring 2019

Akriti Bhat, Amrit Bhatia

1. Abstract

Crime continues to remain a severe threat to all communities and nations across the globe. Alongside the sophistication in technology and processes that are being exploited to enable highly complex criminal activities. Data mining, the process of uncovering hidden information from Big Data, is now an important tool for investigating, curbing and preventing crime and is exploited by both private and government is 1 in 34. Based on FBI crime data, Boston is not one of the safest communities in America. Relative to Massachusetts, Boston has a crime rate that is higher than 97% of the state's cities and towns of all sizes. Hence, a topic related to analysis and prediction of crimes in Boston piqued our interest as it is something that could be extremely relevant to the current trends and useful for the society.

This project provides a framework would could be used by various government or non-government agencies to find out how the crime changed over the years, areas which are more affected, and the statistics for crime in various areas of Boston. Law enforcement agencies like Boston Police could be benefitted from this project, by increasing the patrolling in unsafe areas. In future, this project could be used for other cities and can be expanded to other countries as well.

2. Introduction

Boston is the capital of Massachusetts. Its many colleges and universities have made Boston a higher education international hub. Many students from all over the world come to this city to study law, medicine, engineering, and business. Moreover, the city has around 2,000 startups which has earned it the title of world leader in innovation and entrepreneurship. The monuments like the Freedom Trail, Isabella Stewart Gardner museum, MFA, Copley Square; the famous Red Sox game make it a tourist attraction year round.

As any major city, crime can be a big threat to Boston. Through this project we wish to analyse the crimes in the city of Boston. The chance of becoming a victim of either violent or property crime in Boston is a major problem. There are some specific areas of Boston in which crime is an issue like Roxbury, Mattapan and Dorchester. Since Boston is a hub for education and tourist attraction, it is important to keep it safe. The projects like these would help narrow down the areas and days, where and when the enforcement agencies need to be more careful. The statistics generated from the project could also be used by government to decide if they need an expansion of their team.

Boston was more violent than New York and Seattle, but less violent than Chicago and Las Vegas, according to numbers from the FBI, based on crimes committed in 2015. Nationally, Boston ranked 47 out of 79 cities with populations of 250,000 or more.

The news has gotten better for Boston since then. More recent numbers from the Boston Police Department, or BPD, show that violent crime, as well as property crime, has continued to drop, and has been steadily dropping for several years.

Robberies for 2016 were down five percent and auto thefts were down nine percent. The city has seen a steady decline in violent and property crime, which was down seven percent from 2015, continuing a long-term trend. Through this project, we also tend to validate these theories and understand the patterns better.

Literature Survey:

Bogahawatte and Adikari[2] proposed an approach in which they highlighted the usage of data mining techniques, clustering and classification for effective investigation of crimes and criminal identification by developing a system named Intelligent Crime Investigation System (ICSIS) that could identify a criminal based up on the evidence collected from the crime location.
Agarwal et al. [3] used the rapid miner tool for analyzing the crime rates and anticipation of crime rate using different data mining techniques. Their work done is for crime analysis using the K-Means Clustering algorithm.

Yu et al.[4] have discussed the preliminary results of a crime forecasting model developed in collaboration with the police department of a United States city in the Northeast. Their approach is to architect datasets from original crime records by employing ensemble of data mining classification techniques for crime forecasting.

Above mentioned are just a few of the many related works for crime analysis and prediction. We make use of this literature to find the best possible way in which it can be implemented for the city of Boston.

3. Datasets

For the kind of problem we are addressing there is a dataset available on Kaggle[1] for Crimes in Boston: https://www.kaggle.com/ankkur13/boston-crime-data. This is a dataset containing records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. The dataset was made available by Boston Police Department.

4. Methodology

We have approached this project in four steps:

4.1 Data cleaning and preprocessing : This step involves data cleaning, data reduction (dimensionality reduction using attribute subset selection) and statistical data analysis.
Data cleaning tasks included are filling in missing data, smoothing-out noisy data, removing outliers and artifacts, using label encoder to transform string values into label

formats. Data reduction is performed using Principal Component Analysis for reducing the features and sampling has been used for reducing the number of instances.

4.2 Classification: For classification we are used sklearn's algorithms for Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier and K-Nearest Neighbors. Along with these classifiers we have used features obtained from Principle Component Analysis and Correlation Analysis using Pearson Coefficient. Use of these features has improved our models. We also implemented K-Means Clustering and created 2 clusters based on month and location of crime and offense code group and location of crime. Clustering has also been performed on location in the form of 3, 5 and 10 clusters.

4.3 Pattern Identification: We have used Apriori algorithm and FP- growth algorithm for mining frequent patterns or association rules. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

FP stands for frequent pattern. In the first pass, FP-growth algorithm counts the occurrences of items (attribute-value pairs) in the dataset of transactions, and stores these counts in a 'header table'. In the second pass, it builds the FP-tree structure by inserting transactions into a trie. Items in each transaction have to be sorted by descending order of their frequency in the dataset before being inserted so that the tree can be processed quickly. Items in each transaction that do not meet the minimum support requirement are discarded. If many transactions share most frequent items, the FP-tree provides high compression close to tree root. Recursive processing of this compressed version of the main dataset grows frequent item sets directly, instead of generating candidate items and testing them against the entire database (as in the apriori algorithm).

Our goal of using these models is to find all possible crime frequent patterns regardless of the committed crime type.

4.4 Prediction: Our aim is to predict the potential crime type in a specific location within a particular time in the future. We examine Gradient Boosting classifier, Decision tree classifier, K-Nearest Neighbors and Random Forest classifier, and then choose the model that gives the best accuracy in prediction.

5. Code

The dataset which we are referring, is provided by Boston Police Department and has details of 290,156 unique incidents reported between 2015-2-18.

The code is divided into six jupyter notebooks. We started the project by basic

analysis of the data, understanding the data, finding the missing values and sampling the data. We did principal component analysis of the data set. Since the data is in string format, we used LabelEncoder to transform the data and used PCA module provided by sklearn - decomposition library.

Using the features obtained from PCA, we predicted the crime type using various models. For prediction, we used modules from sklearn ensemble and sklearn tree libraries. We also discovered association rules using Apriori and FP-Growth algorithms and found similar results from both the techniques. The last step was visualization, based on District, Street, Year, Month, Hour, Day of the Week, Offense Code and UCR part. We used matplotlib library for plotting the graphs.

6. Results

We divided the study into six sections to understand the impact of all the features closely. Below are the results for each section:

6.1 Basic Analysis :

The dataset has 327820 with 6.8% missing data. Total columns are 17. There 17 columns determine the incident number, the details of the time of incident and the location and the type of crime that was committed.

(i) INCIDENT_NUMBER: Registered Crime Identification Number

(ii) OFFENSE_CODE: Code for the crime

(iii) OFFENSE_CODE_GROUP: Group of the crime code

(iv) OFFENSE_DESCRIPTION: Detailed description of the incident

(v) DISTRICT: The crime district code

(vi) REPORTING_AREA: Code of the reporting area

(vii) SHOOTING: Yes if shooting happened, else blank

(viii) OCCURRED_ON_DATE: Date of the incident in YYYY-MM-DD HH:MM:SS format

(ix) YEAR: Year of the incident

(x) MONTH: Code of the month

(xi) DAY_OF_WEEK: The day of week of the incident

(xii) HOUR: the hour of the incident

(xiii) UCR_PART: Uniform Crime Reports Code to which the crime belongs

(xiv) STREET: The street of the crime

(xv) LATITUDE: Recorded Latitude of the crime location

(xvi) LONGITUDE: Recorded Longitude of the crime location

(xvii) LOCATION: Combination of Latitude and Longitude

After preprocessing, we dropped rows with null values, the remaining row length is 304,659.

6.2 Feature Selection:

The maximum variance is obtained by features offense code and offense code group when Principal Component Analysis is used. These features have hereby been used along with models as shown in Table 1 to obtain the respective accuracy scores. For correlation analysis using Pearson Coefficient gives shooting, year and UCR part as the three most correlated features. The Decision Tree

Classifier is used along with these 3 features to see how it performs as compared to PCA.

6.3 Prediction Model :

Model	Precision
Gradient Boosting Classifier	88%
Decision Tree	85%
KNeighborsClassifier	89%
Random forest Classifier	90.5%

Table 1

Feature Selection	Accuracy
PCA + Decision Tree	0.88
Pearson features + DTC	0.4545

Table 2

We observed that Random forest Classifier performs better in this case than Decision Tree because it doesn't suffer the instability problems of decision trees. Also PCA performs better than Pearson Coefficient for Feature Selection as shown in Table 2.

6.4 Clustering :

K-Means was used for clustering. As shown in Figure 1, the location in the form of longitude and latitude is plot on axes and month/offense code group is represented by the colorbar. The Figure 1 plots are based on 3 clusters. In figure 2, 10 clusters have been created based on location of crime. Latitude and Longitude on y and x axes represent the location.

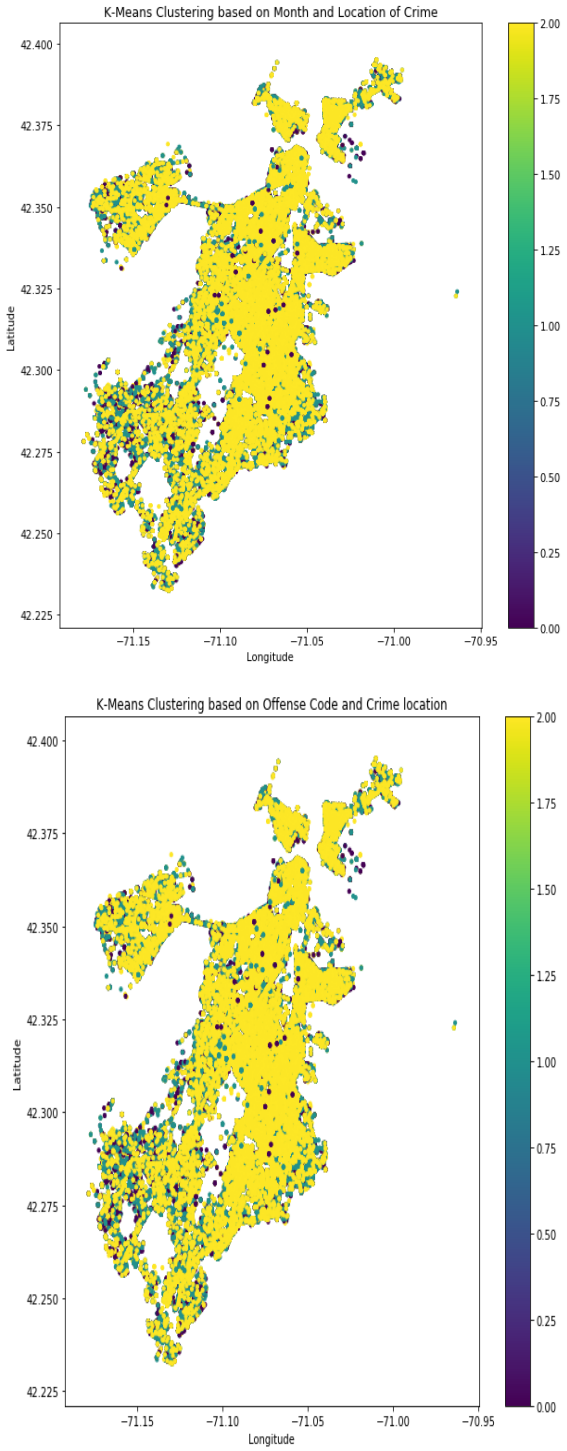


Figure 1

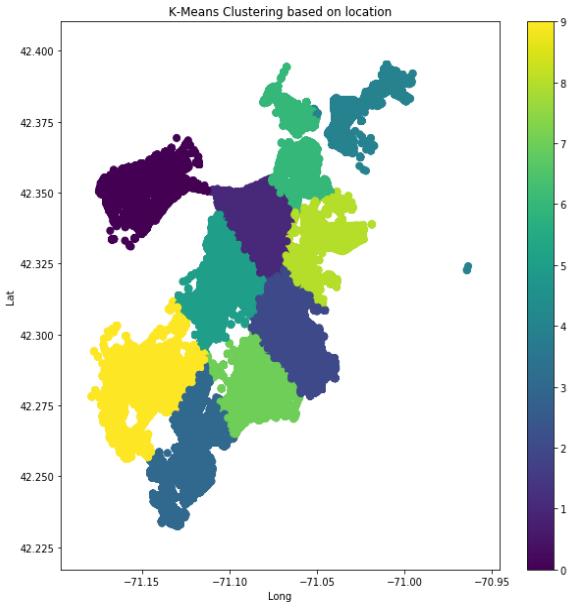


Figure 2

6.5 Association Rules :

Similar results obtained from applying Apriori and FP Growth. It is observed that Motor Vehicle Accident Response is associated with UCR Part Three. UCR Part Two Crimes are mostly during the day time and UCR Part Three Crimes are associated with Night time.

```
{Day} ---> {Friday}:  conf = 0.157, sup = 0.104
{Friday} ---> {Day}:  conf = 0.676, sup = 0.104
{Day} ---> {Sep}:  conf = 0.157, sup = 0.104
{Sep} ---> {Day}:  conf = 0.663, sup = 0.104
{Night} ---> {Part Three}:  conf = 0.504, sup = 0.115
{Part Three} ---> {Night}:  conf = 0.225, sup = 0.115
{Part Three} ---> {Motor Vehicle Accident Response}:
{Motor Vehicle Accident Response} ---> {Part Three}:
{Day} ---> {Tuesday}:  conf = 0.152, sup = 0.101
{Tuesday} ---> {Day}:  conf = 0.694, sup = 0.101
{Day} ---> {Part One}:  conf = 0.185, sup = 0.123
{Part One} ---> {Day}:  conf = 0.657, sup = 0.123
{Day} ---> {Wednesday}:  conf = 0.153, sup = 0.102
{Wednesday} ---> {Day}:  conf = 0.695, sup = 0.102
{Day} ---> {Part Two}:  conf = 0.306, sup = 0.203
{Part Two} ---> {Day}:  conf = 0.687, sup = 0.203
{Day} ---> {Part Three}:  conf = 0.505, sup = 0.335
{Part Three} ---> {Day}:  conf = 0.653, sup = 0.335
```

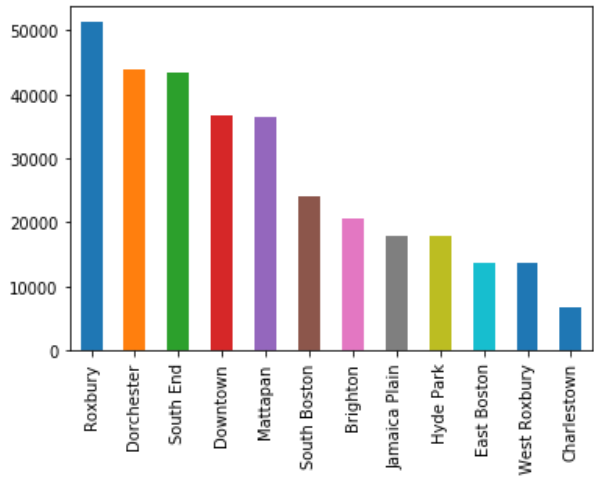
Apriori Results

```
{Day} ---> {Part Three}:  conf = 0.482, sup = 0.32
{Part Three} ---> {Day}:  conf = 0.649, sup = 0.32
{Day} ---> {Part Two}:  conf = 0.316, sup = 0.21
{Part Two} ---> {Day}:  conf = 0.687, sup = 0.21
{Day} ---> {B2}:  conf = 0.154, sup = 0.102
{B2} ---> {Day}:  conf = 0.656, sup = 0.102
{Day} ---> {Wednesday}:  conf = 0.154, sup = 0.102
{Wednesday} ---> {Day}:  conf = 0.693, sup = 0.102
{Day} ---> {Part One}:  conf = 0.197, sup = 0.131
{Part One} ---> {Day}:  conf = 0.662, sup = 0.131
{Day} ---> {Tuesday}:  conf = 0.152, sup = 0.101
{Tuesday} ---> {Day}:  conf = 0.694, sup = 0.101
{Part Three} ---> {Motor Vehicle Accident Response}:  c
{Motor Vehicle Accident Response} ---> {Part Three}:  c
{Night} ---> {Part Three}:  conf = 0.489, sup = 0.113
{Part Three} ---> {Night}:  conf = 0.229, sup = 0.113
{Day} ---> {Friday}:  conf = 0.153, sup = 0.101
{Friday} ---> {Day}:  conf = 0.667, sup = 0.101
```

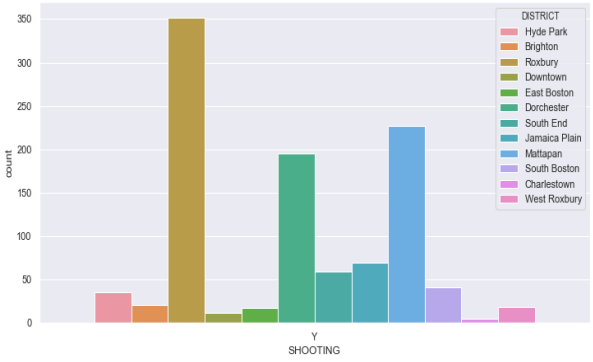
FP Growth Results

6.6 Observations from Visualizations :

1. Roxbury district has the highest crime rate from 2015-2018, followed by Dorchester.

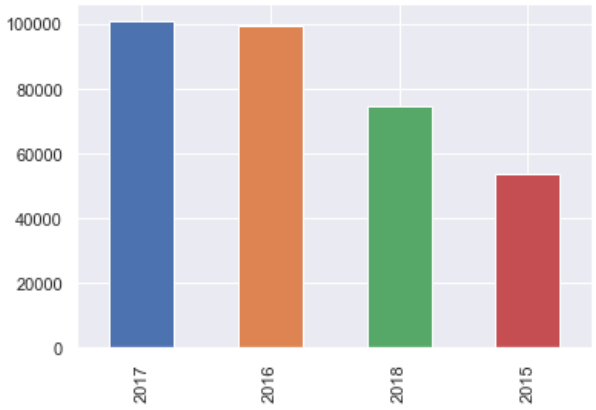


2. Shooting rate is highest in Roxbury as well, followed by Mattapan.

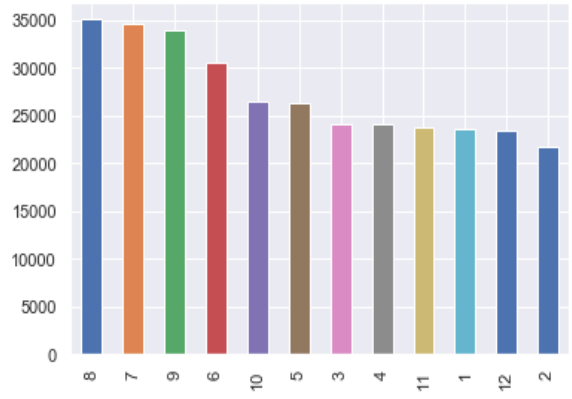


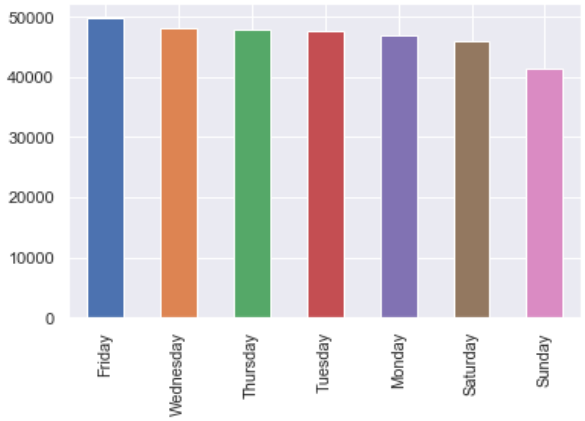
3. Washington street from Roxbury has the highest crime and shooting rate among all the streets.

4. Crime and shooting rates from 2015 to 2017. Crimes rate in 2016 was almost double than observed in 2015, but almost remained constant in 2017. The reason for the high numbers in 2016 & 2017 can be the political changes. The crimes started going down again in 2018.

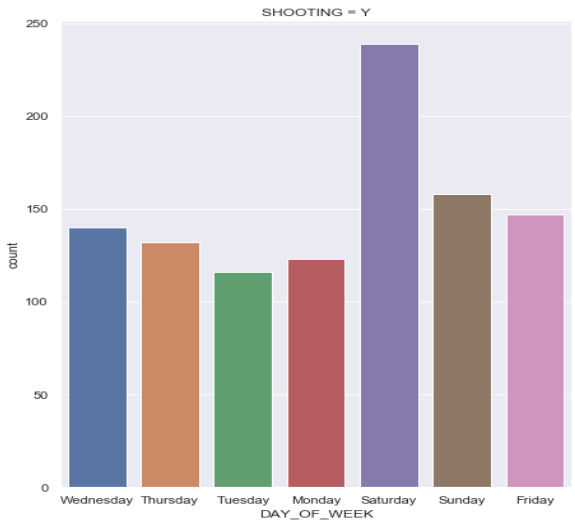


5. The crimes rates have been considerably higher from July to September. As per the studies by researchers, such as Gerhard J. Falk, this is connected to July being the summer month. The count is almost evenly divided among the day of the week, which implies the day has no impact on the number of crimes.

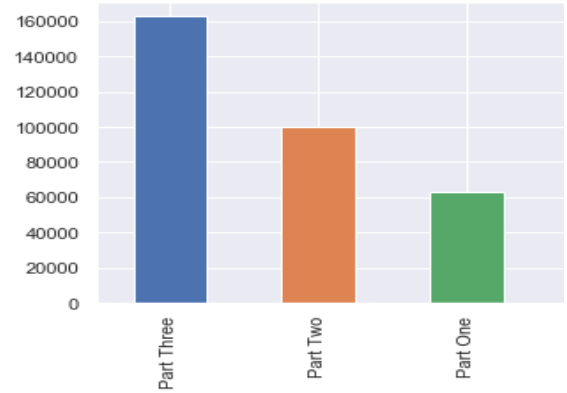




The shooting count is higher on the weekends, but shoots up on Saturday.



6. The crimes are divided into various offence code groups and UCR-Parts. Part Three UCR crimes are the most frequent. Majorly the crimes belong to Motor Vehicle Accident Response Offense Code group.



	Sum
OFFENSE_CODE_GROUP	
Motor Vehicle Accident Response	38134.0
Larceny	26670.0
Medical Assistance	24226.0
Investigate Person	19176.0
Other	18612.0
Drug Violation	17037.0
Simple Assault	16263.0
Vandalism	15810.0
Verbal Disputes	13478.0
Towed	11632.0

7. Discussion

From our experiments, we realised Linear Regression is not suitable for this dataset because of non-linear values. After predicting using features obtained from PCA, the model obtained better prediction results. The accuracy was improved by using various classifiers. Random forests outperforms decision tree because it consists of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees. This is shown in Table 1.

The results obtained for prediction of month are poor and adding Principal Component Analysis(PCA) worsened the results. But at the same time prediction of OFFENSE_CODE_GROUP generated good results, the maximum accuracy of Random Forest used with PCA being 90% approximately. The classifiers when used independently produce less accurate predictions than when used with PCA. Also, PCA proves to be better for feature selection than Pearson Coefficient as the accuracy of Decision Tree Classifier with PCA is 88 percent while that with 3 most correlated features for Offense Code Group prediction is 45 percent only.

8. Future Work

In future, a larger dataset could be used to predict for the entire state and the country but currently such dataset is not available. Other classification models like Naive Bayes, Support Vector Machines can be applied on the given data.

The results generated could be used by various agencies to understand the changes in the crime rates and use the model to check the possibility of crimes in various areas of Boston. The patterns would change over the years, so the data needs to be updated to understand if the measures taken are effective or not. In the future we would like to integrate the system with an interactive user interface, like a web-portal, so that is accessible everywhere, even on the mobile phones and users would select multiple parameters, on the basis of which they can see customized results.

9. Conclusions

From the experiments we conclude that for predicting correctly, we need to sample the data and reduce the number of features. We

did PCA and correlation analysis to get better results. Linear Regression is not possible on the data. The data needs to be preprocessed to generate models.

10. References

[1] FBI Data for the Boston Statistics: https://www.northeastern.edu/thescope/2017/06/21/boston-crime-map-how-safe-is-your-neighborhood/

[2] Kaumalee Bogahawatte and Shalinda Adikari, "Intelligent Criminal Identification System", Proceedings of 8th IEEE International Conference on Computer Science and Education, pp. 633-638, 2013.

[3] Jyoti Agarwal, Renuka Nagpal and Rajni Sehgal, "Crime Analysis using K-Means Clustering", International Journal of Computer Applications, Vol. 83, No. 4, pp. 1-4, 2013.

[4] Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, "Mining Location-based Social Networks for Criminal Activity Prediction", Proceedings of 24th IEEE International Conference on Wireless and Optical Communication, pp.185-190, 2015.

[5] Sathyadevan, Shiju & S, Devan & S Gangadharan, Surya. (2014). Crime Analysis and Prediction Using Data Mining. 10.1109/CNSC.2014.6906719.

[6] Gerhard J. Falk "The Influence of the Seasons on the Crime Rate"

[7]Association Rule Learning on Wikipedia