Crime Data Analysis of Los Angeles

1st Ashwini Kasbekar   
*Computer Science*  
*University of North Carolina,Charlotte* Charlotte,USA  
[akasbeka@uncc.edu](mailto:akasbeka@uncc.edu)

2nd Eshan Bhatt  
*Computer Science*  
*University of North Carolina,Charlotte* Charlotte,USA  
[ebhatt1@uncc.edu](mailto:ebhatt1@uncc.edu)

3rd Manasi Prabhune  
*Information Technology*  
*University of North Carolina,Charlotte* Charlotte,USA  
[mprabhun@uncc.edu](mailto:mprabhun@uncc.edu)

4th Nikita Nalawade  
*Computer Science*  
*University of North Carolina,Charlotte* Charlotte,USA  
[nnalawad@uncc.edu](mailto:nnalawad@uncc.edu)

5th Prutha Shirodkar  
*Computer Science*  
*University of North Carolina,Charlotte* Charlotte,USA  
[pshirodk@uncc.edu](mailto:pshirodk@uncc.edu)

6th Ripal Bhavsar  
*Computer Science*  
*University of North Carolina,Charlotte* Charlotte,USA  
[rbhavsar@uncc.edu](mailto:rbhavsar@uncc.edu)

***Abstract*— This paper focuses on analyzing the crime incidents occurred in Los Angeles. We performed exploratory data analysis to analyze various attributes and carried out experiments using different models such as Naive Bayes, k-Nearest Neighbors, Random Tree and Single Tree classifiers. Using this data analysis, we researched on various attributes like type of crime, area of crime, time, victim gender, victim age, most used weapons etc. The results of these experiments could be used to raise awareness about crime prone locations and can help agencies to predict crime in specific locations at specific time.**

***Keywords—classifier, crime classification, analysis***

# Introduction

The report provides an analysis of the dataset that reflects on the crime incidents in Los Angeles in the year 2015, 2016 and 2017.The data is obtained from original crime reports typed on paper and thus might contain some inaccuracies.These inaccuracies were removed by using some data pre-processing techniques.

The report mainly focuses on predicting the age group of the victim, the weapon used and the area of the crime using known values of other attributes. We have used four models of classification to predict the age group of the victim, the weapon namely k- nearest neighbors classification, Naive Bayesian classification and Single Tree and Random Tree. The k- nearest neighbors algorithm[2] is one of the simplest, non-parametric, lazy learning classification algorithm. It works on the dataset in which the data points are separated into several classes to predict the classification of a new sample point. The Naive Bayesian classifier[3] is a simple probabilistic classifier based on Bayes’ theorem with the independent assumptions between predictors. Naive Bayesian model is without any complicated iterative parameter estimation which makes it particularly useful for very large datasets. Single tree divides the data set into smaller data sets based on the descriptive features until you reach a small enough set that contains data points that fall under one label. Classification problems for single trees are often binary. The goal is to create a model that predicts the value of a target variable based on several input variables. The Random Tree[4] is used for classification as well as regression problems. It works very similar to decision tree except for each split only a random subset of attributes is available.

A comparative analysis on the results obtained from the four classification algorithms is then used determine the best technique for predicting the victim’s age, weapon used and area of crime.

# RELATED WORK

There has been great amount of work conducted related to crime. Large datasets have been reviewed to extract information about crime location, type of crime etc. There exist various map applications showing exact time locations for crime type. Even with crime location identified, there is no information that includes crime occurrence with techniques which can predict crime occurrence in future. For the dataset “ Crime in Los Angeles from 2010”, there has been less research to analyze dataset to predict crime. An exploratory analysis has been conducted in the past to study the attributes of this dataset [1]. In our study, we conducted experiments to create data mining models to classify age, weapon and type of crime.

III. DATASET

The dataset is on Crime Data for the years 2015, 2016 and 2017 in Los Angeles. The attributes considered are Date Reported, Date Occurred, Area ID, Crime Code, Victim Sex, Victim Descent, Premise Code, Weapon Used Code, Status Code, Time Occurred 1 and Age with respect to the crime incident. The Date Reported and Date Occurred attributes provide information about the date of occurrence of the crime incident and the date when it was reported. The Area ID consists of numbers from 1-21. The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21. The Crime Code attribute indicates the type of the crime committed. The Victim Sex attribute consists of three attributes namely F - Female, M - Male and gender X. The Victim Descent attribute consists of the victim’s descent code namely Descent Code: A - Other Asian, B - Black, C - Chinese, D - Cambodian, F - Filipino, G - Guamanian, H - Hispanic/Latin/Mexican, I - American Indian/Alaskan Native, J - Japanese - Korean, L - Laotian, O - Other, P - Pacific Islander, S - Samoan, U - Hawaiian, V - Vietnamese, W - White, X - Unknown, Z - Asian Indian. The attribute Premise Code gives us information about the type of structure, vehicle, or location where the crime took place. The attribute Weapon Used Code gives us information about the weapon used in a particular crime. The attribute Status Code gives information about the status of the crime, IC being the default value. The various Status Codes are IC - Invest Cont, AA - Adult Arrest, AO - Adult Other, JA- Juv Arrest. The attribute Time Occurred gives information about the time of occurrence of the crime in 24-hour military time. The attribute Age indicates the age of the victim [2].

# Business Use Case

Mrs. D’souza, is the Deputy Commissioner of Los Angeles Police Department and she wants to understand the current crime scenario in the different areas of Los Angeles. Being the Deputy Commissioner, she needs to understand different aspects of the committed crimes, based on which she can take the necessary measures to reduce the crime rate.

Mrs. D’souza has provided us with the data of committed crimes for the years 2010-2018 and has asked us to use this data along with the analytical skills of the team, to generate insights using which Mrs. D’souza can take the necessary actions to minimize the crime rate in Los Angeles.

1. *Hypothesis 1*

The area in the city where the crime took place.

In this the major areas of the Los Angeles city which are prone to crime zones.The dataset has various area Id’s mentioned which will be used collectively to come up with the major crime prone areas.This will help the deputy commissioner understand the areas where the crime rate is high and so extra measures and strong policies can be set up to handle crime in such areas.

1. *Hypothesis 2*

The weapon with which the crime was committed.

In this hypothesis the most used and dangerous weapons are found out.The dataset has weapon used code which will be used in finding out the weapons which need to be banned or needs some action to be taken by the commissioner in the city.This will also help people to know the weapons the criminals are using and so take precautions towards these.

1. *Hypothesis 3*

The type of crime and the time at which it was committed.

In this hypothesis,the different crime is found out.This also includes finding out the time at which these type of crimes are committed.The dataset includes crime code and time occurred at which is used to find out these two.This will help the commissioner to understand different types of crimes and also the serious one which needs immediate attention.The time will help both the authorities and people to take precautions.

# Methodology

Finding relationships between various attributes of crime can help to predict type of crime occurring on different locations. In our approach, we aim to focus on classifying three major attributes i.e. victim age, most used weapon and type of crime. We tried to extract patterns based on these attributes. With various models we classified these variables to predict potential crimes. In this section, we explain dataset preparation, data analysis and model generation.

## Data preprocessing

### Data Cleaning and Reduction: There are few missing values in our attributes. However, we observed that all attributes which exist are not our key attributes. Hence, we decided to drop few of them. The attributes we dropped are Area name, MO Code, Crime Code Description, Premise Description, Weapon Description, Status Description, Address, Cross Street and location as these were descriptive attributes for numeric code available. Thus, we performed reduction to preprocess our data further. We performed data imputation to clean data for handling missing values. We analyzed various papers and replaced missing values by M as males were the most impacted victim gender. We replaced missing values of victim descent by calculating mode.

* + 1. *Data Transformation and Discretization:* Victim age was discretized into 4 categories namely Kid, Teen, Adult and Senior. Time occurred was discretized into 4 categories namely morning, afternoon, evening and night. This discretization helps to classify specific category of the attribute.
    2. *Data Binning:* Data binning is the method of grouping continuous values in a smaller number of bins.We created bins for attributes which had a large number of distinct values*.* We created 10 bins for the attributes Crime Code and Weapon Used Code each. The bins for Crime Code and Weapon Used Code can be referred in Figure 1 and Figure 2:

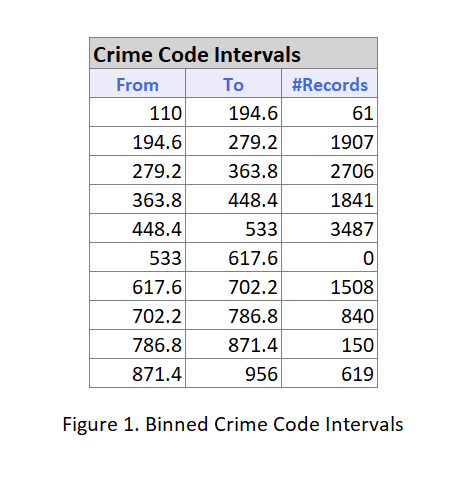


Table 3. Binned Crime Code Intervals

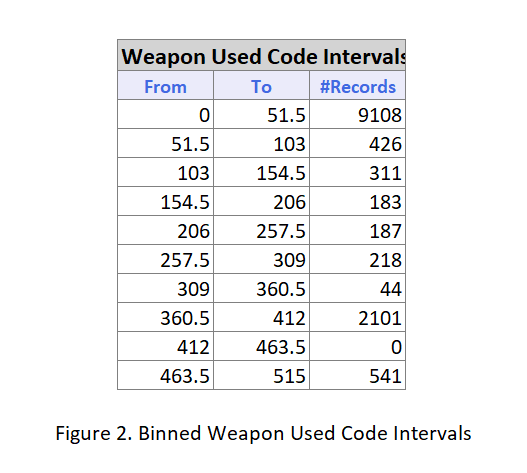


Table 4. Binned Weapon Used Code Intervals

* + 1. *Creating Dummy Variables:* It is necessary for all independent variables to be numeric for analysis. Hence, we have created dummies for Area\_ID, Victim Sex, Victim Descent, Status Code, Time Occurred, Age and Binned Weapon Used Code.

## Exploratory Data Analysis

To analyze and get entire view of data, we performed statistical data analysis on the entire dataset using Tableau to visualize various attributes. In this section, we explain various visualizations and analysis on attributes based on those visualizations.

### Area wise crime count

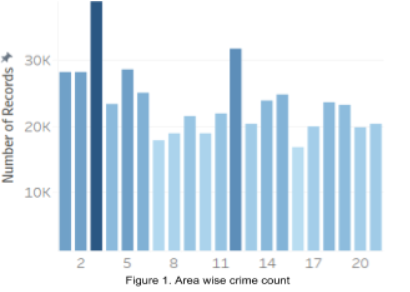


Figure 1. Area wise crime crime count

Analysis: From the visualization in Figure 1, we can observe area codes with count of crime records. Darker colors in the visualization indicate more count of crime in those areas. Lighter shades indicate that those areas are less impacted by crime. This visualization gives us details of crime prone areas.

*2) Timewise crime count*

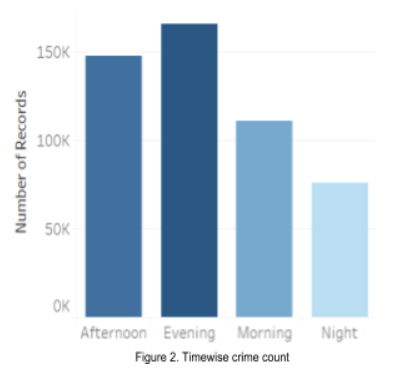
**

Figure 2. Time wise crime crime count

Analysis: The bar graph in Figure 2 gives the time divisions (Afternoon, Evening, Morning and Night) with the total count of crimes. We can observe that most crimes occurred in the evening (around 170K) and least crimes occurred at night.

*3) Impacted gender details*

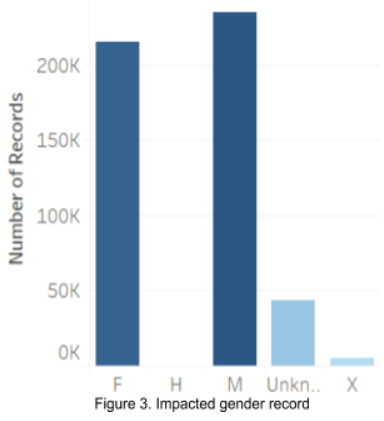
**

Figure 3. Impacted gender record

Analysis: From the visualization in Figure 3, we can observe the count of impacted victims gender wise. Darker the color of graph, higher is the count of respective gender victims. We can see from the above graph that M are mostly the target victims for all criminal activities followed by F. Gender H has no victim record.

*4) Count of weapons*

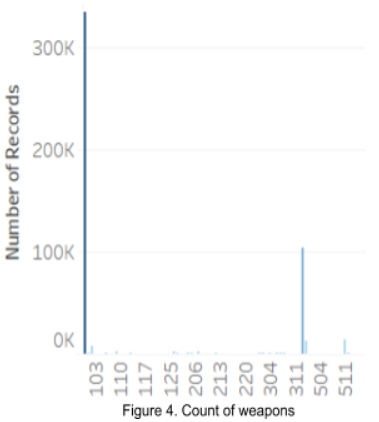
**

Figure 4. Count of weapons

Analysis: From the visualization in Figure 4 we can see the details of count of weapons used. Weapons are indicated with a specific code. Darker the color, greater the use of specific weapon.

*5) Most used weapons*

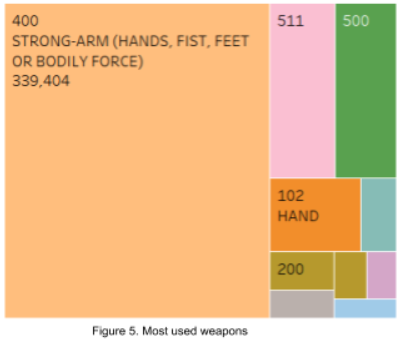
**

Figure 5. Most used weapons

Analysis: We can see from Figure 5, a tree map visualization of most used weapons. We have plotted a tree map for top 10 weapons used. Different colors indicate different weapon codes and we also have a label along with it indicating weapon name and count of number of times the weapon is used. On hovering on the respective portion of tree map, weapon code along with details pop up. In this visualization, we observe that Strong-arm (Hands, fist, feet or bodily force) is used mainly to attack victims instead of a sharp tools or other weapons

*6) Descent wise victim count*

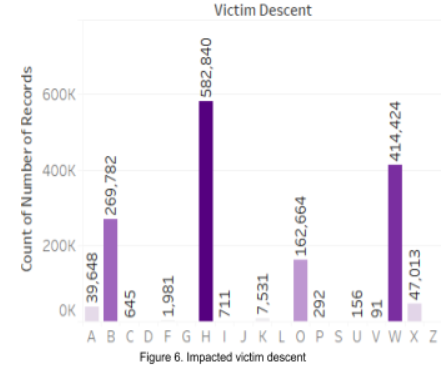


Figure 6. Impacted victim descent

Analysis: Figure 6 gives a visualization of count of victim descent. Darker the shade of graph greater is the count of impacted victims. As we can see from the plot in Figure 6, H - Hispanic/Latino/Mexican are the descent with highest victim count (582,840) followed by W - Whites. C - Cambodian, J - Japanese, L - Laotian, S - Samoan, Z - Asian Indian are the descents with no victims.

*7) Premise wise Male and Female victim count*

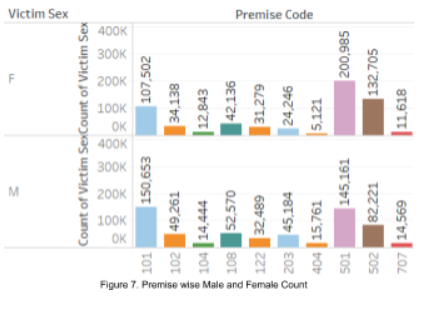


Figure 7. Premise wise Male and Female Count

Analysis: The graph in Figure 7 shows premise wise count for male and female victims. We can see from above graph that premise 501 had highest male and female victims as compared to other premises. We can also observe that ratio of male to female victims would follow a pattern as they are in proportion.

*8) Details on type of crime*

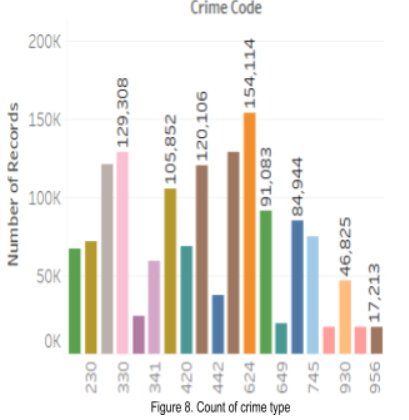


Figure 8. Count of crime type

Analysis: Figure 8 gives the details of type of crime occurring. Here we can see that crime code 624 - Simple assault occurred maximum times with record count of 154,114 followed by 510 - stolen vehicle. Thus, we get details of all crimes through this visualization. In the above graph in Figure 8, different colors represent different crime codes. Hovering on this data, we get detailed count of specific crime type with their number of records.

# Model building

In order to extract frequent patterns from the Crime Dataset of Los Angeles, we applied different modeling techniques like classification, association, clustering on the dataset. After thorough study on the dataset, we formed research questions and hypothesis which we considered as the base for developing models. We examined every model then choose classification as the modeling technique as it gave the best accuracy in prediction. We used the Naive Bayes, K Nearest Neighbor, Single Tree and Random Tree Classification techniques for examining the type of crime, type of weapon used and the age group of affected victims.

We constructed the models using XLMiner which provides a platform for performing data mining in Excel. We divided the dataset randomly into 60% of data as a training set and 40% of data as a testing set. For each model, we chose 3 different class labels based on the hypothesis we formed during the preliminary study of the dataset. We selected the crime dataset features of Area ID, Victim Sex, Victim Descent, Time Occurred, Age, Binned Weapon Used Code for the Binned Crime Code class label. For the class label of Weapon Used Code, we selected the features of Area ID, Victim Sex, Time Occurred, Age and Binned Crime Code. We selected the features of Area ID, Victim Sex, Victim Descent, Binned Crime code and Binned Weapon Used Code.

Classification Trees in XLMiner are especially useful to classify/predict outcomes. They generate simple rules that can easily be translated to a natural query language. The decision trees work by binary recursive partitioning i.e. they keep on classifying a record by checking whether it meets the criteria at a node or not [4]. In this section, we provide a brief description of each model used.

## Naïve Bayesian Classifier

Naïve Bayesian classifier is a widely used supervised learning algorithm. This model considers the independent effect between attribute values. This model is our ideal choice as our dataset has features which are independent of each other. After generating the model in XLMiner, we studied the summary report which includes the confusion matrix and error report for the above mentioned 3 class labels.

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

* 1. *K- Nearest Neighbor*

K-nearest neighbor algorithm is a pattern recognition algorithm. In this algorithm, the input consists of k closest training examples from the feature space. An object is classified by majority vote of its neighbors. The object is assigned to the class most common among its k nearest neighbors. We have selected the value of k as 1. Hence, in this case, the object is simply assigned to the class of a single nearest neighbor. After generating the model in XLMiner, we studied the summary report which includes the confusion matrix and error report for the above mentioned 3 class labels.

In the k-Nearest Neighbor prediction method, the Training Set is used to predict the value of a variable of interest for each member of a target data set. The structure of the data generally consists of a variable of interest (i.e., amount purchased), and a number of additional predictor variables (age, income, location).

1) For each row (case) in the target data set (the set to be predicted), locate the k closest members (the k nearest neighbors) of the Training Set. A Euclidean Distance measure is used to calculate how close each member of the Training Set is to the target row that is being examined. 2) Find the weighted sum of the variable of interest for the k-nearest neighbors (the weights are the inverse of the distances. 3) Repeat this procedure for the remaining rows (cases) in the target set. 4) XLMiner allows them to selection of a maximum value for k, builds models in parallel on all values of k (up to the maximum specified value), and performs scoring on the best of these models.

Computing time increases as ‘k’ increases, but the advantage is that higher values of k provide smoothing that reduces vulnerability to noise in the Training Set. Typically, k is in units of tens of units, rather than in hundreds or thousands.

*C Random Trees*

The Random Trees ensemble method of XLMiner works by training multiple weak classification trees using a fixed number of randomly selected features, then taking the mode of each class to create a strong classifier. Number of randomly selected features controls the fixed number of randomly selected features in the algorithm. The default setting is 4. [3] The minimum records in terminal node is chosen as 1 for all the class labels. XLMiner stops splitting a node (during tree growth) when the number of records in the Training Set in the node is below this selected value of Minimum record in terminal node. For class label Binned Crime Code, 7 random features were chosen. For the class label Weapon Used code, 8 random features were chosen and for Age, 8 random features were chosen.

The random trees method (random forests) is a variation of bagging. This method works by training multiple weak regression trees using a fixed number of randomly selected features (sqrt[number of features] for classification and number of features/3 for prediction), then takes the average value for the weak learners and assigns that value to the strong predictor. Typically, the number of weak trees generated could range from several hundred to several thousand depending on the size and difficulty of the training set. Random trees are parallelizable, since they are a variant of bagging. However, since random trees selects a limited amount of features in each iteration, the performance of random trees is faster than bagging.

*D. Single Tree*

The Single Tree for Classification feature of XLMiner model is generated by providing the parameters of

input variables, single Output variable which is the class label, minimum number of records in terminal node and maximum levels in tree to be displayed. There exists a Prune Tree option which is selected by default when a Validation Set exists. When this option is selected, XLMiner prunes the tree using the Validation Set. (Pruning the tree using the Validation Set reduces the error from over-fitting the tree using the Training Set.) [5]. Further, we selected the options for generating Full Tree, Best Pruned Tree and Minimum Error Tree. Full tree is (grown using training data) to grow a complete tree using the Training Set. Best pruned tree is (pruned using validation data) to grow a tree with the fewest number of nodes, subject to the constraint that the error be kept below a specified level (minimum error rate plus the standard error of that error rate).

Minimum error tree is (pruned using validation data) to produce a tree that yields the minimum classification error rate when tested on the Validation Set [5].

The Best Pruned Tree for Binned Crime Code class label can be referred from Figure 9:

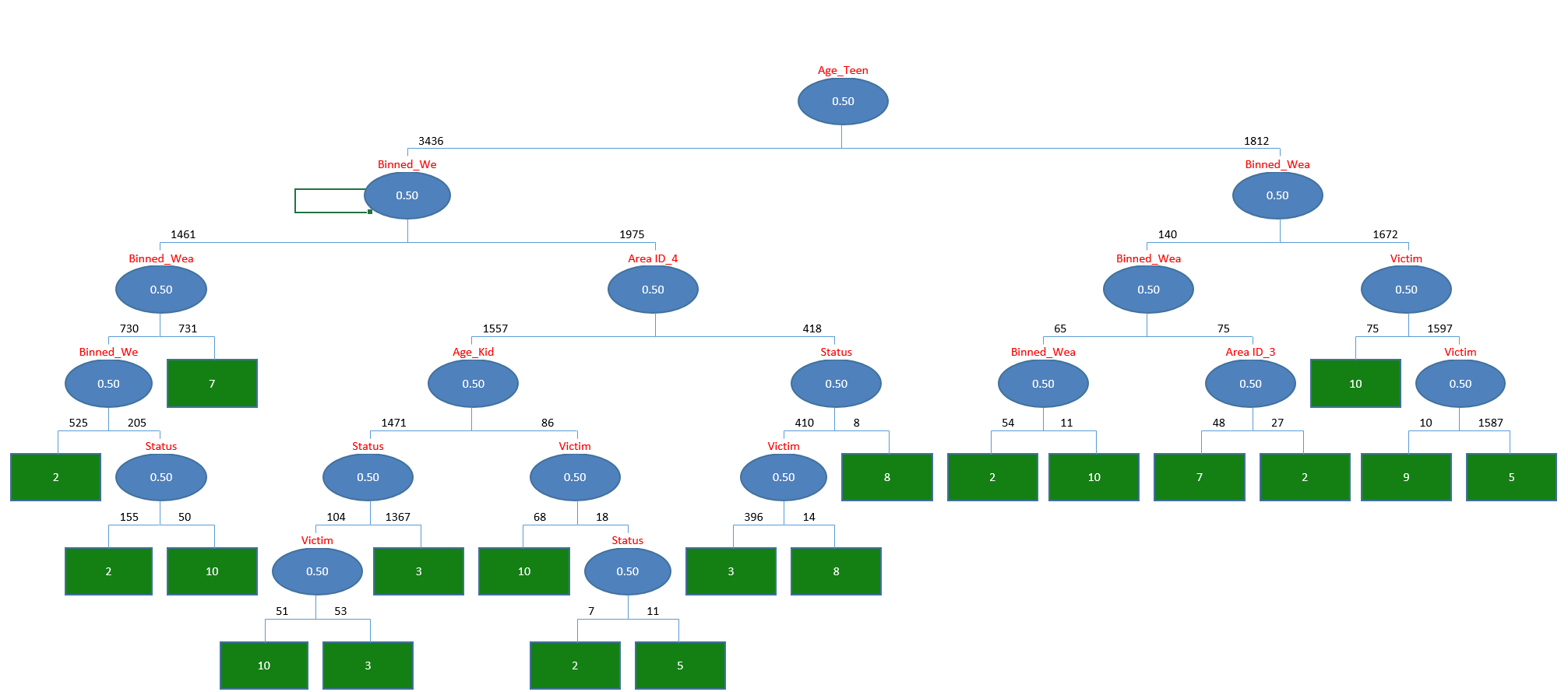


Figure 9. Best Pruned Tree

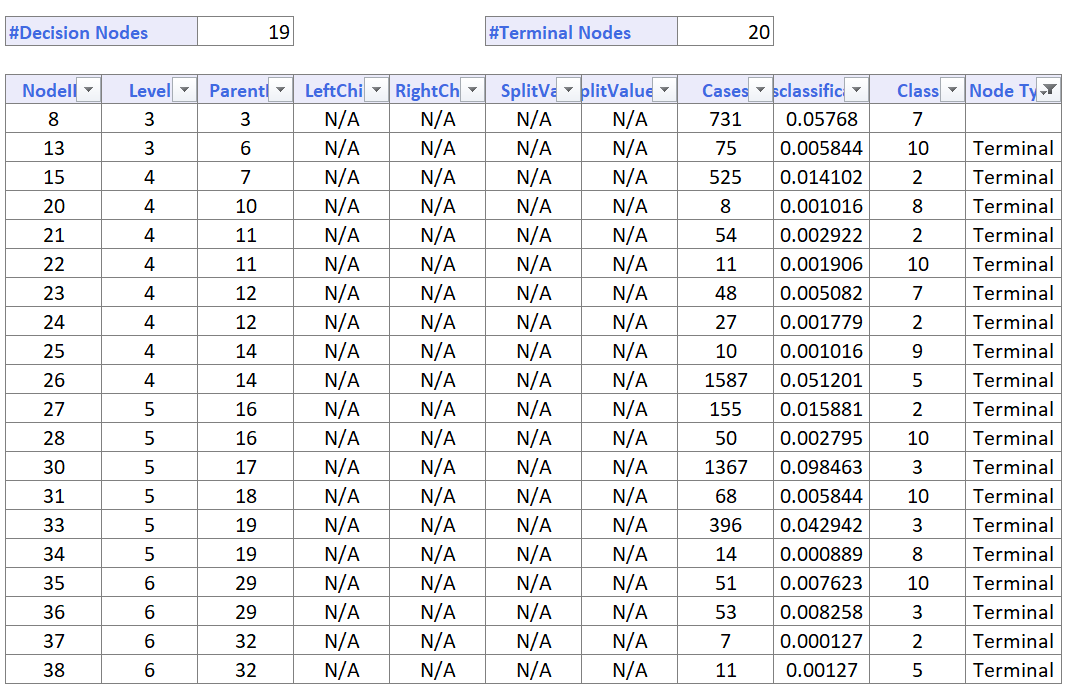


Table 3

The Best Pruned Tree for Binned Weapon Used Code class label can be referred from Figure 10:

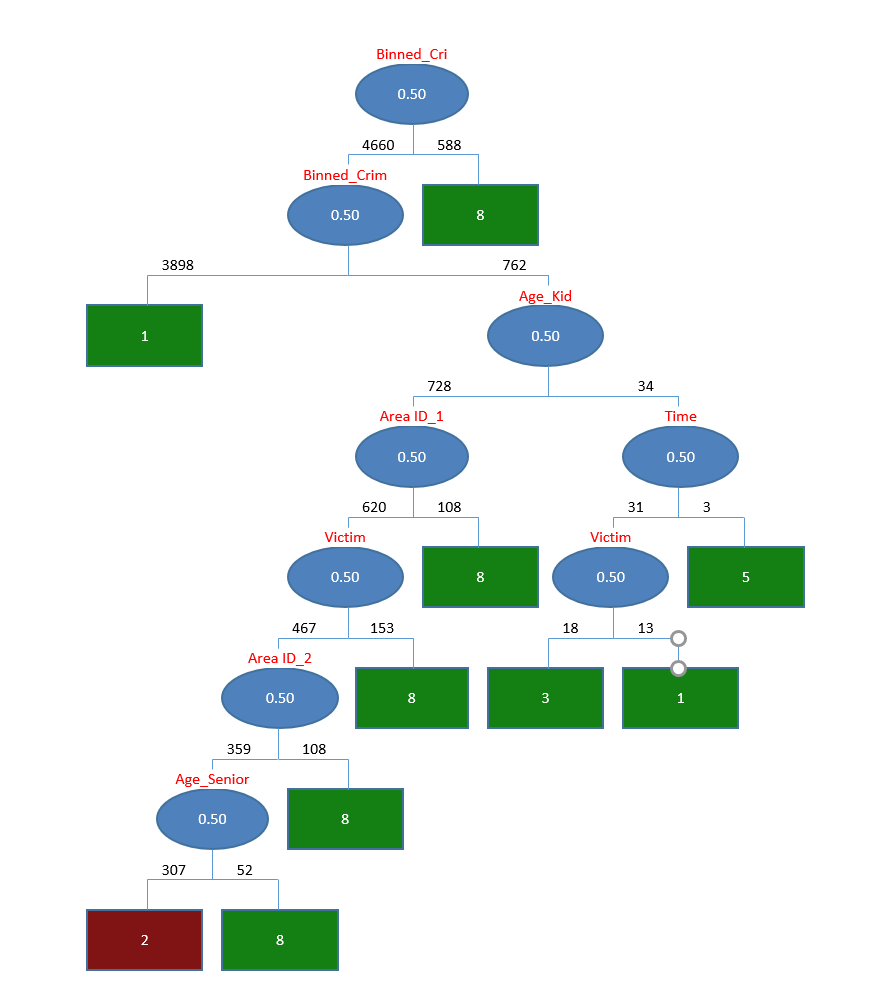


Figure 10. Best Pruned Tree for Binned Weapon Code

The Best Pruned Tree Rules for class label, Weapon Used Code can be referred to Table 4:

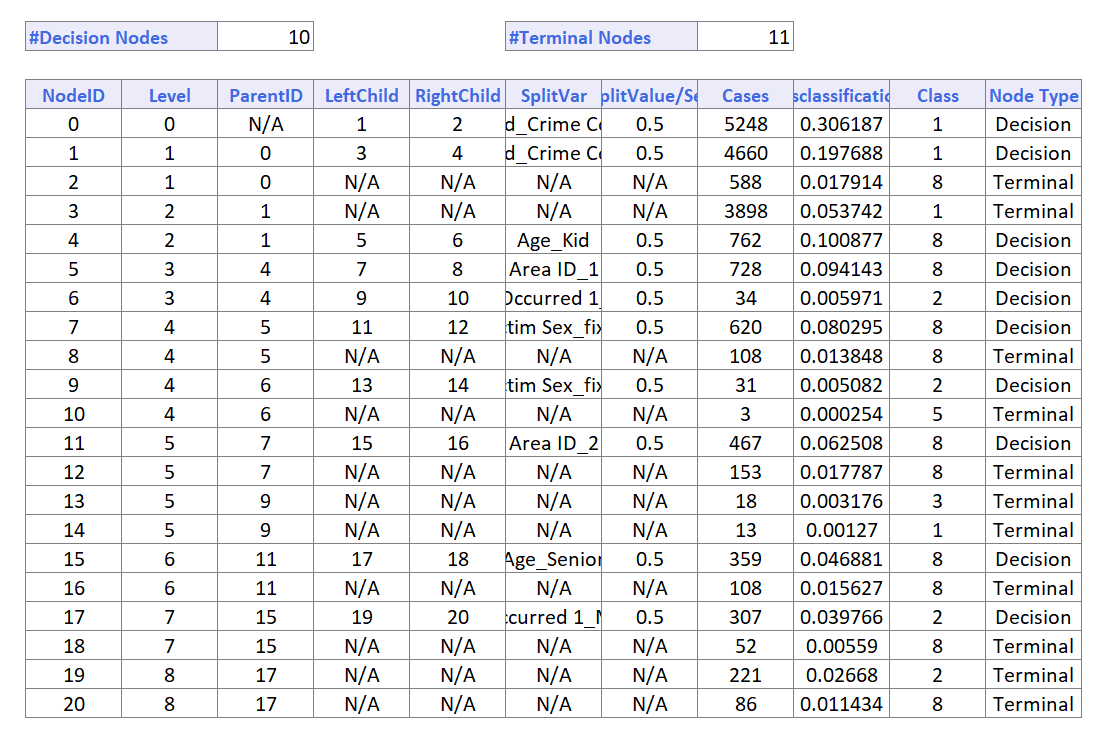


Table 4. Best Pruned Tree Rules

The Best Pruned Tree for Age can be referred from Figure 11:

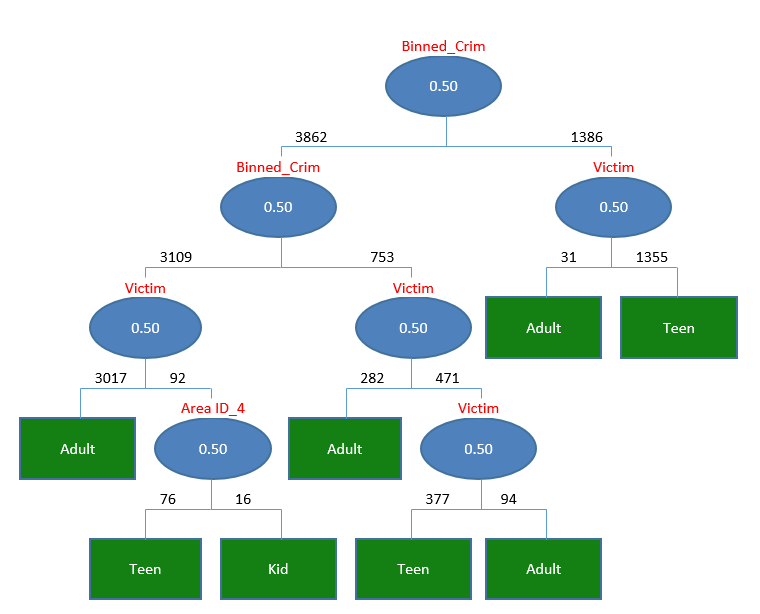


Figure 11. Best Pruned Tree for Age

The Best-Pruned Tree Rules can be viewed in Table 5:

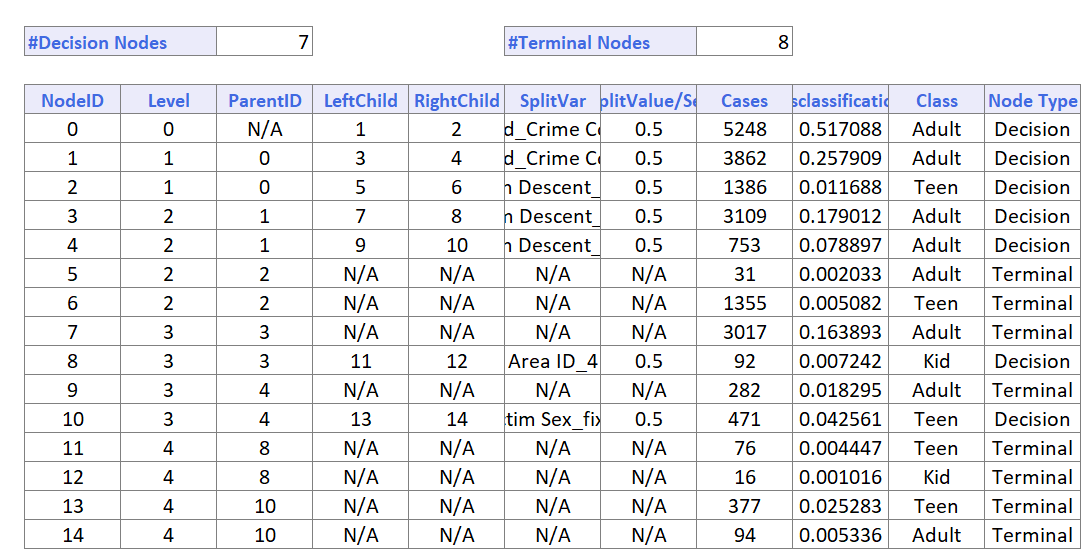
****

Table 5. Best Pruned Tree Rules

1. EVALUATION

We have worked on dataset using 4 classification algorithms Naive Bayes classifier, k-Nearest Neighbors, Random Tree and Single Tree classifier.

Below are the result tables for all the classifiers:

|  |  |  |  |
| --- | --- | --- | --- |
|  | k-Nearest Neighbors | Naive Bayes | Random Tree |
| Age | 14.91 | 24.05 | 22.71 |
| Weapon | 13.37 | 18.87 | 17.34 |
| Crime | 25.38 | 35.82 | 34.76 |

Table 6. Result Table

Referring to Table 6,

If we use Age as the target variable, k-Nearest Neighbors gives the least error rate of 14.91%.

If we use Weapon as the target variable, k-Nearest Neighbors gives the error rate 13.37%.

If we use Crime as the target variable, k-Nearest Neighbors gives the error rate of 25.38%.

We have also implemented Single Tree for all the three target variables but could not achieve good results for error rate.

1. CONCLUSION

We performed exploratory data analysis on preprocessed data and studied various attributes through graphs and visualizations plotted to understand Los Angeles crime dataset. We explored patterns in our attributes using model building techniques like Naive Bayes, k-Nearest Neighbor, Random Tree and Single Tree Classifiers.

After analysing the results obtained from the above model building techniques we can conclude that k-Nearest Neighbors gave the best result with least error rate of 14.91% for Age , 13.37% for Weapon and 25.38% for Crime.

VII. RECOMMENDATIONS

The analysis thus shows us that Area 3 and Area 12 are prone to crime.Crime rates are high in the evening time. The adult age group is most prone to being victims of crime incidents followed by teens and then kids.To reduce the crime rates , it is highly recommended that awareness programs be carried out according to various age groups.Self-defense programs for people of all age groups is a must.

The highest percentage of crimes are performed using physical strength/verbal threat followed by knife or blade upto 6 inches followed by gun/pistol/revolver. It is thus highly recommended that license on guns should not be easily made available. To add to it ,the regulations on usage of various hazardous weapons should be made more stringent.

VIII. FUTURE SCOPE

As a part of future work, we plan to implement various models on this dataset to increase prediction and accuracy on the dataset. We also plan to include various other factors and find their relationships with crime rate. This data could be useful for various agencies to predict crime.

##### References

1. <https://www.kaggle.com/rineelreddy/basic-visualization>
2. <https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7>
3. <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>
4. https://en.wikipedia.org/wiki/Random\_forest
5. <https://catalog.data.gov/dataset/crime-data-from-2010-to-present>
6. <https://www.solver.com/using-classification-tree>
7. <https://www.slideshare.net/dataminingtools/xl-miner-classification>
8. <https://www.solver.com/xlminer/help/classification-tree-example>