Crime Data Analysis of Los Angeles

1st Ashwini Kasbekar
Computer Science
University of North Carolina,Charlotte
Charlotte,USA
akasbeka@uncc.edu

2nd Eshan Bhatt
Computer Science
University of North Carolina,Charlotte
Charlotte,USA
ebhatt1@uncc.edu

3rd Manasi Prabhune
Information Technology
University of North Carolina,Charlotte
Charlotte,USA
mprabhun@uncc.edu

4th Nikita Nalawade
Computer Science
University of North Carolina,Charlotte
Charlotte,USA
nnalawad@uncc.edu

5th Prutha Shirodkar
Computer Science
University of North Carolina,Charlotte
Charlotte,USA
pshirodk@uncc.edu

6th Ripal Bhavsar
Computer Science
University of North Carolina,Charlotte
Charlotte,USA
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

I. 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. The missing data values are filled with the highest frequency values of that specific attribute i.e. we have considered the mode values.

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 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 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 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.

II. 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].

III. 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.

A. *Data preprocessing*

1) 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.

2) 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.

3) 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 viewed below:

Crime Code Intervals		
From	To	#Records
110	194.6	61
194.6	279.2	1907
279.2	363.8	2706
363.8	448.4	1841
448.4	533	3487
533	617.6	0
617.6	702.2	1508
702.2	786.8	840
786.8	871.4	150
871.4	956	619

Figure 1. Binned Crime Code Intervals

Weapon Used Code Intervals		
From	To	#Records
0	51.5	9108
51.5	103	426
103	154.5	311
154.5	206	183
206	257.5	187
257.5	309	218
309	360.5	44
360.5	412	2101
412	463.5	0
463.5	515	541

Figure 2. Binned Weapon Used Code Intervals

4) Creating Dummy Variables: It is necessary for all independent variables to be numeric for the purpose of analysis. Hence, we have created dummies for Area_ID, Victim Sex, Victim Descent, Status Code, Time Occurred, Age and Binned Weapon Used Code.

B. *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.

1) Area wise crime count

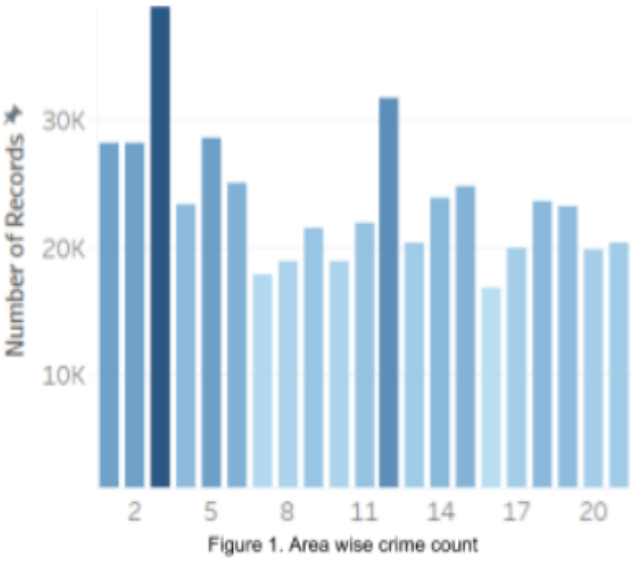


Figure 1. Area wise crime count

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

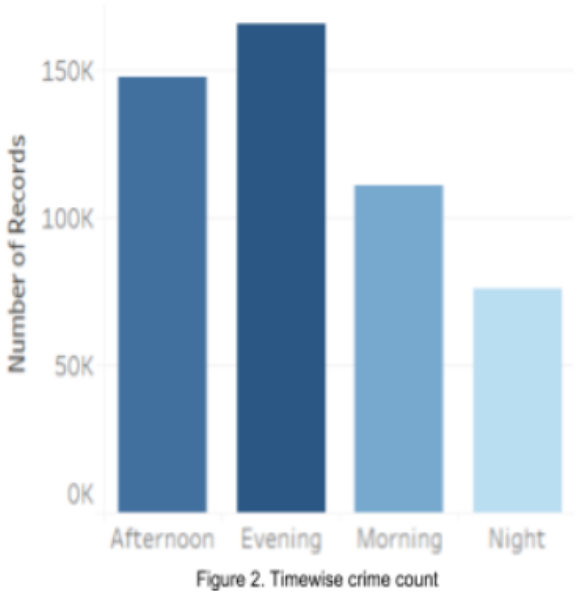


Figure 2. Timewise crime count

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

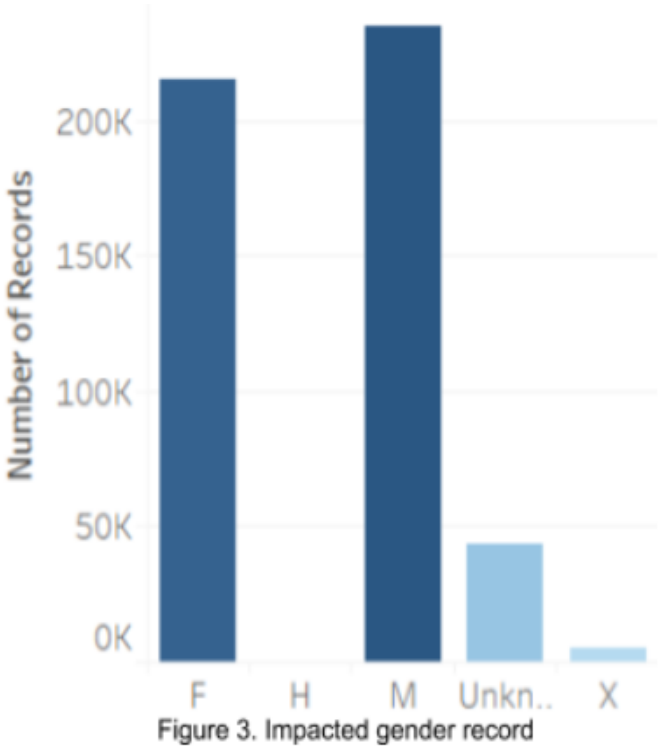


Figure 3. Impacted gender record

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.

3) Count of weapons

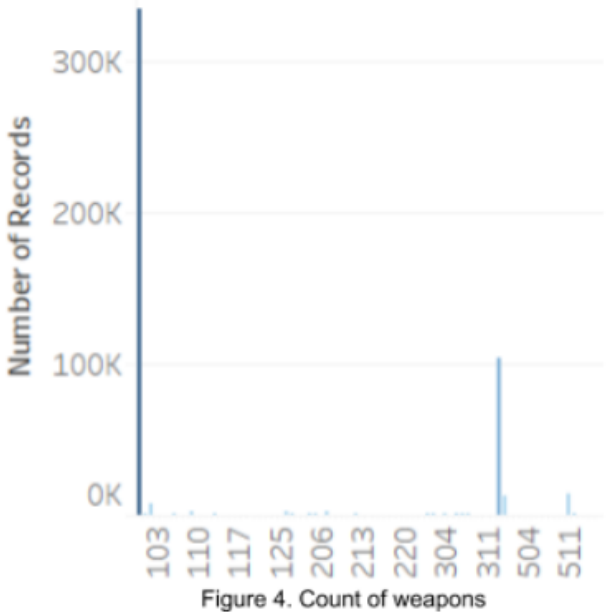


Figure 4. Count of weapons

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.

4) Most used weapons

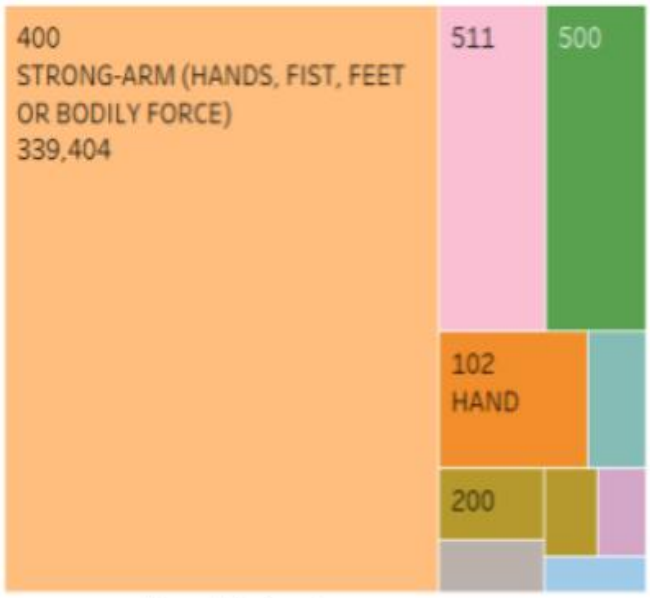


Figure 5. Most used weapons

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

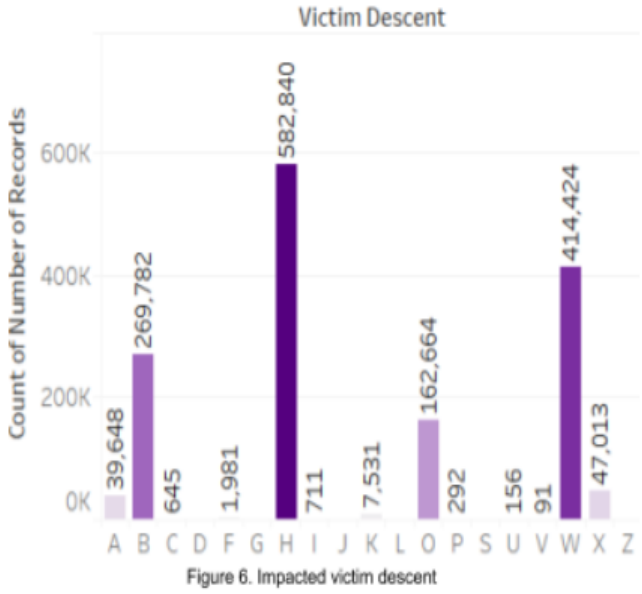


Figure 6. Impacted victim descent

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

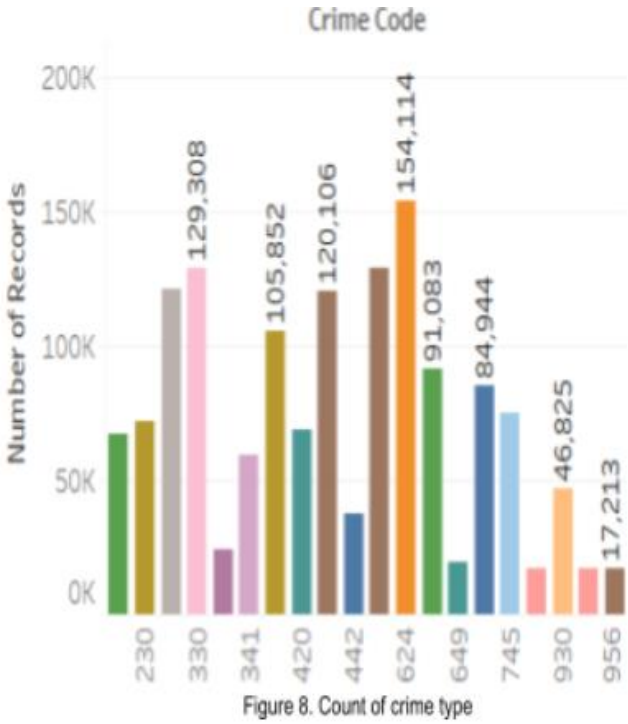


Figure 8. Count of crime type

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.

IV. 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.

A. 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.

B. 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.

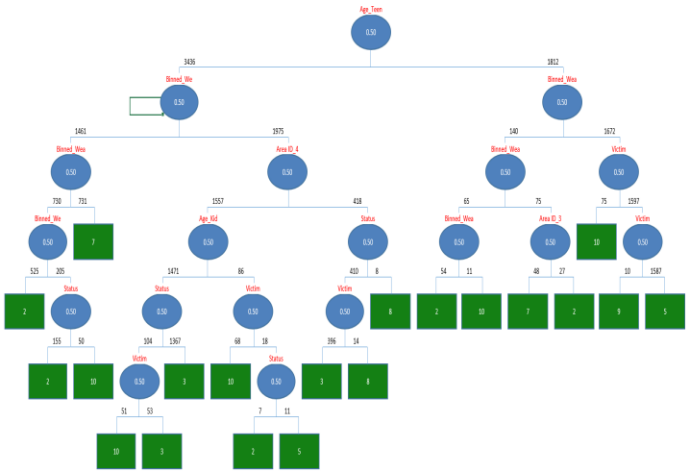
C Random Trees

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.

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 seen below:

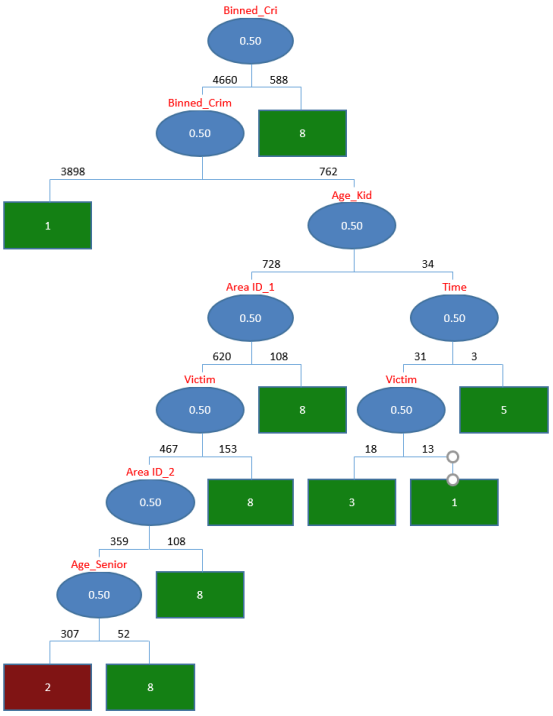


The Best Pruned Tree for Binned Crime Code can be seen below:

#Decision Nodes	19	#Terminal Nodes	20

NodeID	Level	ParentI	LeftChi	RightCh	SplitVa	plitValue	Cases	sclassific	Class	Node Ty
8	3	3	N/A	N/A	N/A	N/A	731	0.05768	7	
13	3	6	N/A	N/A	N/A	N/A	75	0.005844	10	Terminal
15	4	7	N/A	N/A	N/A	N/A	525	0.014102	2	Terminal
20	4	10	N/A	N/A	N/A	N/A	8	0.001016	8	Terminal
21	4	11	N/A	N/A	N/A	N/A	54	0.002922	2	Terminal
22	4	11	N/A	N/A	N/A	N/A	11	0.001906	10	Terminal
23	4	12	N/A	N/A	N/A	N/A	48	0.005082	7	Terminal
24	4	12	N/A	N/A	N/A	N/A	27	0.001779	2	Terminal
25	4	14	N/A	N/A	N/A	N/A	10	0.001016	9	Terminal
26	4	14	N/A	N/A	N/A	N/A	1587	0.051201	5	Terminal
27	5	16	N/A	N/A	N/A	N/A	155	0.015881	2	Terminal
28	5	16	N/A	N/A	N/A	N/A	50	0.002795	10	Terminal
30	5	17	N/A	N/A	N/A	N/A	1367	0.098463	3	Terminal
31	5	18	N/A	N/A	N/A	N/A	68	0.005844	10	Terminal
33	5	19	N/A	N/A	N/A	N/A	396	0.042942	3	Terminal
34	5	19	N/A	N/A	N/A	N/A	14	0.000889	8	Terminal
35	6	29	N/A	N/A	N/A	N/A	51	0.007623	10	Terminal
36	6	29	N/A	N/A	N/A	N/A	53	0.008258	3	Terminal
37	6	32	N/A	N/A	N/A	N/A	7	0.000127	2	Terminal
38	6	32	N/A	N/A	N/A	N/A	11	0.00127	5	Terminal

The Best Pruned Tree for Binned Weapon Used Code class label can be seen below:

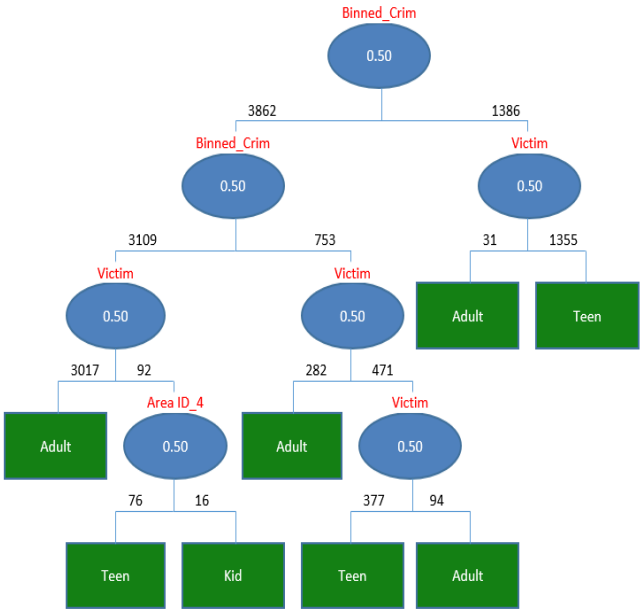


The Best Pruned Tree Rules for class label, Weapon Used Code can be viewed below:

#Decision Nodes	10	#Terminal Nodes	11

NodeID	Level	ParentID	LeftChild	RightChild	SplitVar	plitValue/S	Cases	sclassificati	Class	Node Type
0	0	N/A	1	2	d_Crime C	0.5	5248	0.306187	1	Decision
1	1	0	3	4	d_Crime C	0.5	4660	0.197688	1	Decision
2	1	0	N/A	N/A	N/A	N/A	588	0.017914	8	Terminal
3	2	1	N/A	N/A	N/A	N/A	3898	0.053742	1	Terminal
4	2	1	5	6	Age_Kid	0.5	762	0.100877	8	Decision
5	3	4	7	8	Area ID_1	0.5	728	0.094143	8	Decision
6	3	4	9	10	Occurred 1	0.5	34	0.005971	2	Decision
7	4	5	11	12	tim Sex_fi	0.5	620	0.080295	8	Decision
8	4	5	N/A	N/A	N/A	N/A	108	0.013848	8	Terminal
9	4	6	13	14	tim Sex_fi	0.5	31	0.005082	2	Decision
10	4	6	N/A	N/A	N/A	N/A	3	0.000254	5	Terminal
11	5	7	15	16	Area ID_2	0.5	467	0.062508	8	Decision
12	5	7	N/A	N/A	N/A	N/A	153	0.017787	8	Terminal
13	5	9	N/A	N/A	N/A	N/A	18	0.003176	3	Terminal
14	5	9	N/A	N/A	N/A	N/A	13	0.00127	1	Terminal
15	6	11	17	18	Age_Senio	0.5	359	0.046881	8	Decision
16	6	11	N/A	N/A	N/A	N/A	108	0.015627	8	Terminal
17	7	15	19	20	curred 1_I	0.5	307	0.039766	2	Decision
18	7	15	N/A	N/A	N/A	N/A	52	0.00559	8	Terminal
19	8	17	N/A	N/A	N/A	N/A	221	0.02668	2	Terminal
20	8	17	N/A	N/A	N/A	N/A	86	0.011434	8	Terminal

The Best Pruned Tree for Age can be seen below:



The Best-Pruned Tree Rules can be viewed below:

#Decision Nodes	7	#Terminal Nodes	8

NodeID	Level	ParentID	LeftChild	RightChild	SplitVar	plitValue/S	Cases	sclassificati	Class	Node Type
0	0	N/A	1	2	d_Crime C	0.5	5248	0.517088	Adult	Decision
1	1	0	3	4	d_Crime C	0.5	3862	0.257909	Adult	Decision
2	1	0	5	6	n Descent_	0.5	1386	0.011688	Teen	Decision
3	2	1	7	8	n Descent_	0.5	3109	0.179012	Adult	Decision
4	2	1	9	10	n Descent_	0.5	753	0.078897	Adult	Decision
5	2	2	N/A	N/A	N/A	N/A	31	0.002033	Adult	Terminal
6	2	2	N/A	N/A	N/A	N/A	1355	0.005082	Teen	Terminal
7	3	3	N/A	N/A	N/A	N/A	3017	0.163893	Adult	Terminal
8	3	3	11	12	Area ID_4	0.5	92	0.007242	Kid	Decision
9	3	4	N/A	N/A	N/A	N/A	282	0.018295	Adult	Terminal
10	3	4	13	14	tim Sex_fi	0.5	471	0.042561	Teen	Decision
11	4	8	N/A	N/A	N/A	N/A	76	0.004447	Teen	Terminal
12	4	8	N/A	N/A	N/A	N/A	16	0.001016	Kid	Terminal
13	4	10	N/A	N/A	N/A	N/A	377	0.025283	Teen	Terminal
14	4	10	N/A	N/A	N/A	N/A	94	0.005336	Adult	Terminal

V. 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 1

Referring to Table 1,
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.

VI. 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.

We can conclude that k-Nearest Neighbors gave the best result with least error rate for all the three target variables.

VII. 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://catalog.data.gov/dataset/crime-data-from-2010-to-present
[3] https://www.solver.com/using-classification-tree
[4] https://www.slideshare.net/dataminingtools/xl-miner-classification
[5] https://www.solver.com/xlminer/help/classification-tree-example