# Crime Data Analysis and Prediction in Los Angeles

Akshay Ravi, Brandon Palomino
Department of Computer Science, San Jose State University, USA
[$]San Jose, California 95192, USA
akshay.ravi@sjsu.edu, brandon.palomino@sjsu.edu

***Abstract*—Crime rates have been increasing at an alarming rate in different parts of our country. The need for different tools to keep track of where they occur is crucial for keeping record of past incidents. In the process, data scientists have utilized this info to make predictions on future incidents. This project aims to help solve this issue by building different machine learning algorithms and applying it on real crime data to classify crimes using both binary and multi-class classification. Crimes were mapped into violent and non-violent for binary class, while they were mapped to kidnap, intenet to murder, crime against employers, human trafficiking, crime against religion, injuries, crime against public peace and others. It was observed that multi-class classification outperformed binary-class classification while using the same feature set and same Machine Learning Models. Results were compared on the basis of accuracy and Random Forest performed the best with an average accuracy of 96% over both types of classifications.**

***Keywords*—data, features, preprocessing, crime, machine learning, artificial intelligence, classification, prediction, KNN, Neural networks.**

## I. Introduction

Cities in all parts of the world face one common problem, i.e. crimes. Normal people are the most affected by this as they lose lives and livelihoods. In earlier days, police forces used to resort to taking action on criminals after crimes were reported. This proved to be helpful in bringing them to justice but it did not help in reducing crimes. To prevent such crimes, it would be helpful for the police to know where a crime is more likely to take place. This information will help them decide where to position forces. With the advent of machine learning algorithms that help with classification and prediction, crimes can be predicted with sufficient data of past crimes.

In the last decade, the collection of data has moved from papers to documents in the cloud. Every police department maintains a record of crimes committed in their area, which includes accused and victim information, event description, time of the crime, weapon used, etc. This data is available online and any individual can perform an analysis of it. With more and more data being available everyday, classification and prediction algorithms can learn better and improve their accuracy.

In this project, we aim to classify and predict the crimes that happen in a particular city, Los Angeles, CA, and offer insights into where a crime can take place and forecast the number of crimes in the upcoming years. This project has the potential to reduce a significant number of crimes.

## II. Related Work

Before choosing the models to use on the dataset, we first reviewed the literature in this subject and summarized it below to understand the methods that have been used so far and their shortfalls. This helped us brainstorm different ways to classify and predict crimes effectively.

In [1], the authors used two classification methods, namely Naive Bayesian and Decision Tree to predict crime in US states using WEKA, and open source tool in JAVA. Iqbal et al [1] applied the decision tree method by building the tree from the root until it reached a stopping condition. They chose 12 out of a dataset of 28 attributes and added a new nominal attribute called ‘Crime Category’ with three values, ‘Low’, ‘Medium’, and ‘High’. These values were decided using the percentage of violent crimes per population i.e. ‘Violent Crimes Per Pop’. For Decision Tree, the Accuracy, Precision and Recall were 83.9519%, 83.5% and 84%. Whereas the accuracy, precision and recall values for Naive Bayesian are 70.8124%, 66.4% and 70.8%, respectively. Although the accuracy was surprisingly very high using the decision tree method, a downfall was that a small change in data resulted in a big change in the structure. [1] concluded that potential extensions were to further apply other classification algorithms on the crime data set and evaluate their prediction performances.

In [3], shojaee et al. [3] proposed a model where crimes were distinguished as critical and non-critical and the updated crime set was used to predict using multiple methods. Using KNN rendered an accuracy of approximately 87%. This method performed well for the given preprocessed data but a lot of data was lost while they classified crimes on a very general basis.

The authors of [2] focussed more on preprocessing the data by filling the missing data first. Usually researchers fill data manually. This takes a long time although it can be accurate. [2] used 3 algorithms i.e. Maximum class filling algorithm (filling the missing values with the maximum frequency class attribute value), roulette filling algorithm (random selection of a class value based on its frequency) and GBWKNN filling algorithm (assigning weights to each individual based on the KNN graph) to obtain a better real crime dataset. For classification, they used Decision Tree’s C4.5 algorithm, Naive Bayesian algorithm and K-nearest neighbours (KNN) algorithm. They observed that the highest accuracy of 72.95%

was achieved by combining the GBWKNN filling algorithm (K=70) and KNN classification algorithm. However, the issue with this approach was that it was heavily reliant on a good fitting algorithm. So, if the data and features changed a little bit, the fitting algorithm did not fill the data well and in turn, the prediction accuracy was affected.

The authors of [5] focussed on investigating the capability of Deep Learning methods to forecast hotspot areas in an urban environment where crimes of certain types were more likely to occur in a defined future window. Due to the lack of processing power, neural networks were used with only 9 neurons in a single hidden layer. The Deep Learning methods were fed with the minimum amount of data containing only spatial, temporal and crime type information. Across all methods, most accuracies averaged around 88% accuracy. While it is true that they provided high accuracy, a fallout from this was that they were biased due to tests happening within a small area radius. Models better understand the order of "hotness" in a dual output setting where the second output is the number of crimes that occurred in the same future window. In other words, the incorporation of temporal semantics was required to predict crime fluctuations. Although preprocessing wasn't discussed, the paper acknowledged that it is important for the overall crime classification process.

In [4], the authors used several general prediction models to calculate the future forecasting of potential crime happening and proposed face detection after the learning algorithms as well. For this particular paper, 10 different models of analysis like decision tree, KNN, Naive Bayes, Regression Model, SVM, and Random Forest regressor were studied. The accuracies were 59.15%, 66.69%, 87.0%, 42%, 84.37% and 97% respectively. Although the accuracy was surprisingly very high for random forest, a con was that newly added data would interfere with the location of future crime incidents. [4] proposed a system to monitor individuals closely and detect them using face recognition. This would give details on individuals who were more likely to commit a crime. In other words, the flow chart of the proposed system first processed the data, determined a threat detection level, then classified the threat, simulated a scenario, and then finally briefed the authorities with a 60 word description. This method required huge processing power and a lot of data gathering to be able to identify criminals. Moreover, hardware was also needed to detect faces live.

## III. Proposed Method

This project followed a traditional approach when preparing to apply machine learning algorithms onto a large dataset as shown in the flowchart below. The data was first collected from the data source, then preprocessed, cleaned and transformed to the required format, analyzed using exploratory data analysis and forecasted. Further, features were selected for classification and the data was trained on multiple Machine Learning models and trained on a large set of data. Each model was then compared and the results were studied and analyzed to conclude. Each of the steps are described in detail in further sections of this report.

## IV. Data Collection

This project followed a traditional approach when preparing to apply machine learning algorithms onto a large dataset as shown in the flowchart below. The data was first collected from the data source, then preprocessed, cleaned and transformed to the required format, analyzed using exploratory data analysis and forecasted. Further, features were selected for classification and the data was trained on multiple Machine Learning models and trained on a large set of data. Each model was then compared and the results were studied and analyzed to conclude. Each of the steps are described in detail in further sections of this report.

## V. Data Preprocessing

This is the process of applying certain data transformations on data so it can be used properly by ML algorithms. These include data cleaning, feature subset selection, data transformation, data reduction, etc. Data cleaning is the process of transforming the dataset to a state that is suitable for data analysis and data processing through machine learning algorithms. Tools such as OpenRefine and Spark offer users the necessary support to clean datasets. "Google Colab" was chosen to perform the cleaning process.

When reviewing the dataset closely, there were inconsistencies with the ages of some of the victims. Some were listed as negative numbers, which is impossible since a victim can never have that age. Therefore, cases where the victim's age is less than 0 were removed from the dataset.

In regards to the LAT and LON fields in the dataset, it is important to note that Los Angeles is located at LAT 34 and LON -118 on the world map. Any other LAT and LON values will result in a different location on the map. Values that lied outside the range of possible LA locations were removed from the dataset.

When investigating the area field, there were two columns marked "Area", with the same data. To remove this redundancy, one of the area fields was removed from the dataset.

Also, in regards to the victim's set, there were other genders like "nan, 'H', 'X', '-', 'N' " that were neither male nor female. However it would not be right to remove these rows from the dataset. Hence, these genders were converted to 'O' for this project.

The dataset also contained NULL values in the columns, "Vict Descent","Mocodes","Crm Cd 1", "Crm Cd 2", Crm Cd 3", Crm Cd 4", "AREA", "Cross Street", "Premis Cd", "Premis Desc", "Weapon Used Cd", "Weapon Desc", and "Status". These were replaced by "N/A","Unknown"., etc to preserve the integrity of the data, while also not losing the rows.

In addition, duplicate crime data were removed as well.

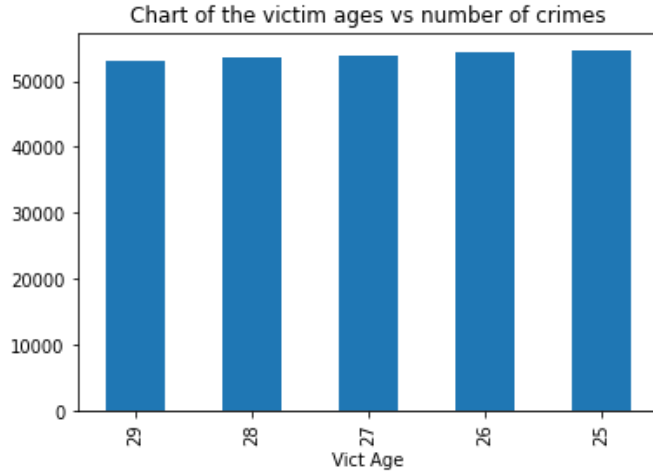
Applying all these data cleaning techniques helped ensure that the dataset would be processed throughout each ML algorithm properly.

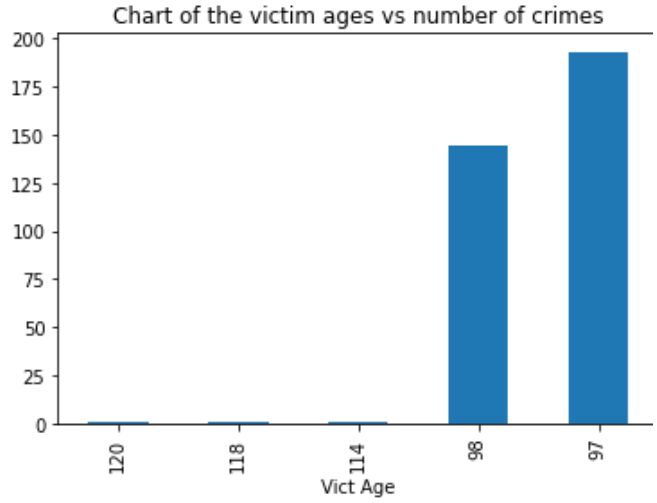
## VI. Exploratory Data Analysis

This is the process of analyzing the data to understand it better. Since the data set is large and there are a lot of features, it is important to note which features impact the crime rate. Plotting graphs to vizualize the data and its various counts is one of the most effective and popular ways to read data. Hence, we calculated, plotted, and included our understanding below.

*A. Number of crimes for different ages*

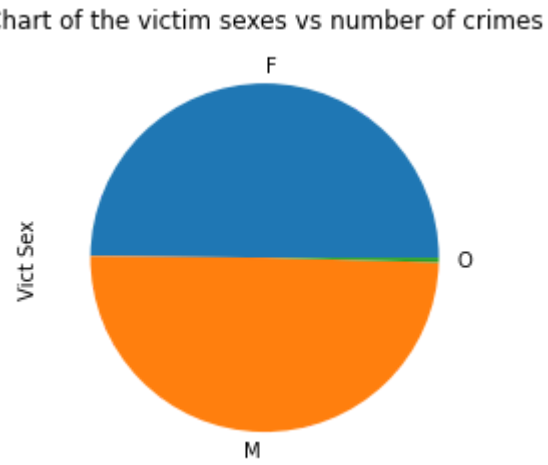
It was observed that crimes occurred most to people in their 20s and least to the extremely elderly in their 90s. The first chart shows the top 5 victim age groups and the second chart shows the least 5 victim age groups.



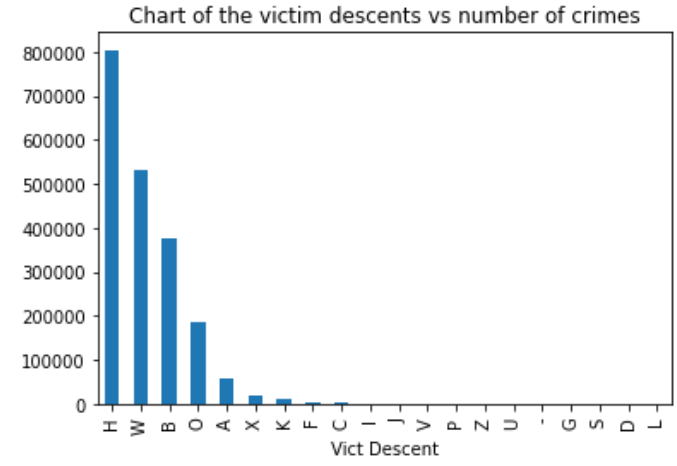

*B. Number of crimes for different sexes*

It was observed that males and females were equally likely affected.



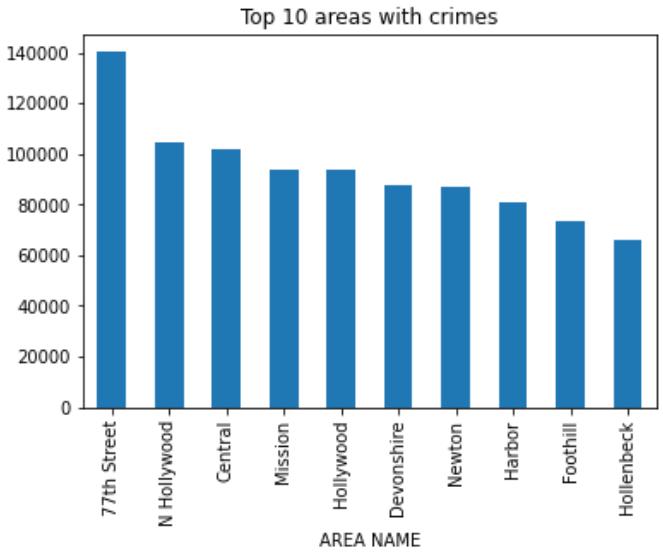

*C. Number of crimes for different victim descents*

It is usually claimed that race is a significant contributor to the reason for crimes. In order to verify that, the number of crimes for victim descents was plotted and it was observed that the top 5 races that are victims are the hispanics, whites, blacks, others and asians in that order. The hispanics were however the most affected by crimes by a significantly big margin. While such data might not be conclusive because the attacker's information is not available, it still tilts towards suggesting that race is an important factor in many crimes.



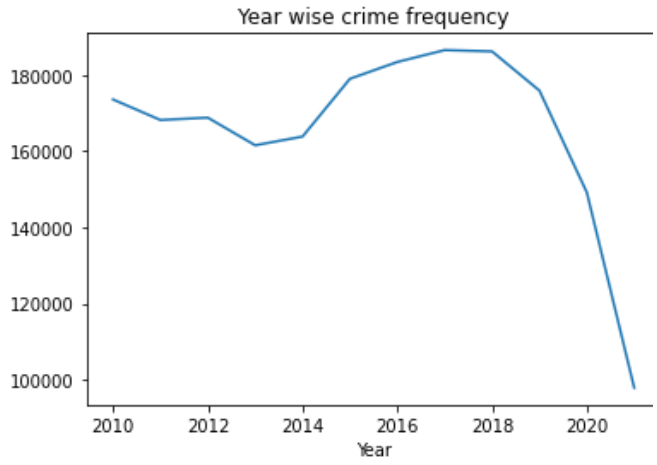

*D. Number of crimes in different areas*

Locality is an important measure for crimes. Rightly so, the plot showed that the infamous 77th Street,Hollywood and Central areas were the top crime hubs.

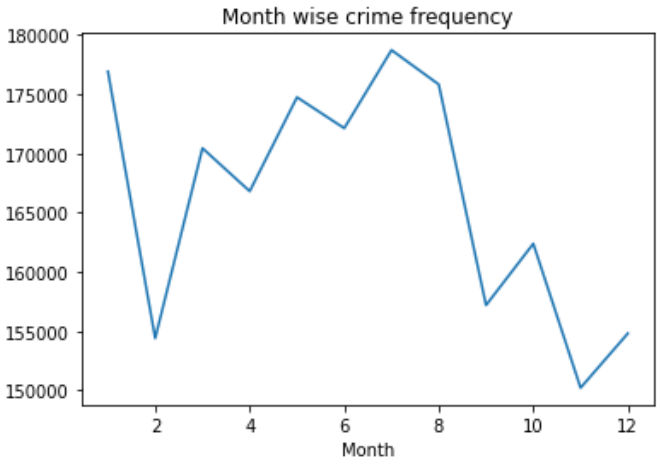



*E. Number of crimes date wise*

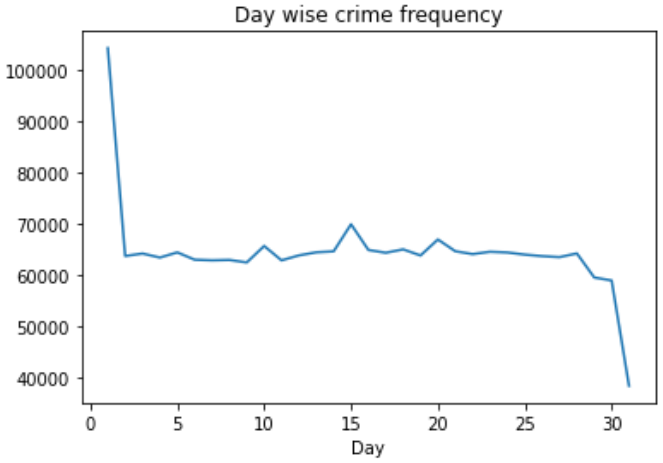
The plot of year wise crime frequency showed that the number of crimes have come down drastically in the last few years. The years between 2016-2019 saw some of the highest number of crimes committed in LA. The rest of the years were however very similar and there is no specific trend since we do not see a plateau,incline or decline.




In the month wise data,the observable point is that crimes are generally low in December and are high in January and July.

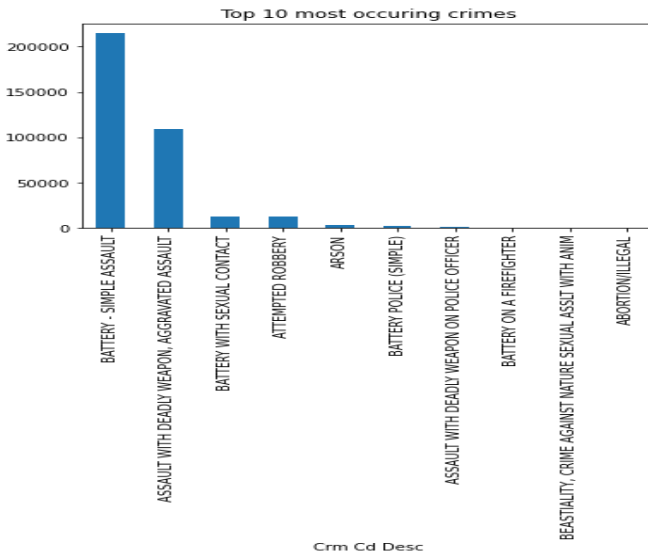


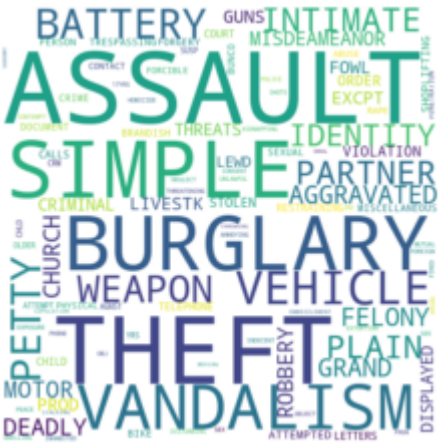

After digging deeper by examining the day-wise crime frequency, it is evident that the 1st of every month has almost 1.5 times the number of crimes that are committed on other days. Secondly, the last few days of the months usually see some of the lowest crimes in the month. This could be explained by the tighter law enforcement during the last few days of the month.




*F. Different types of crime frequencies*

Battery assault is the most occurring crime while Abortion and crimes against nature and animals are the least occuring crimes.




## G. *Word cloud for premis description*

Since the premise description is a text field, word cloud is a more appropriate way of vizualizing the frequency of the premises.”Street”,”Parking” and “Apartment” and “Dwelling” stood out as the hubs.

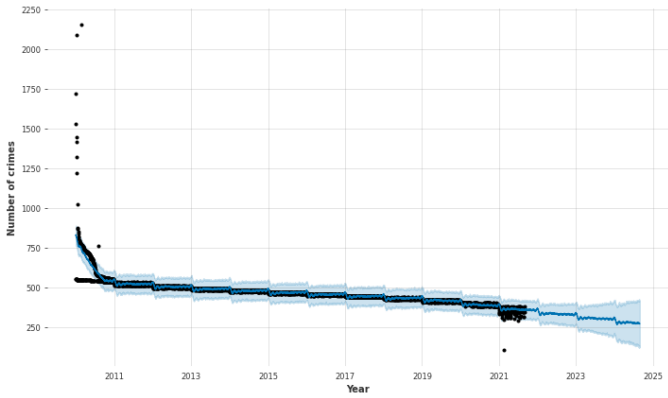

# VII. Time Series Forecasting

## A. *Data Preparation for Forecasting*

For time series analysis, the date column has to be used alongside another feature. Because of this, the number of crimes happening per day was chosen and predicted using various techniques.

## B. *Prophet*

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality along with holiday effects. It handles outliers well and is known for taking care of missing values as well. In addition, seasonality based predictions are covered by Prophet as well.




We chose the ‘DATE OCC’ column and extracted the number of unique dates from it. We then calculated the number of crimes using value_counts(). We trained the prophet model with the time dataframe in mind. We then predicted the number of crimes over the next 3 years. Based on the results of the prophet model, the graph shows a downward trend in the number of crimes in LA. If we were to continue projecting future data for several more years, we could potentially see less crime in LA in the foreseeable future. Note that other factors in real time could influence whether we do see a decrease in crime (i.e. law changes, less policy security, etc).

# VIII. Feature Selection

For any classification and prediction problems using Machine Learning, choosing features is one of the most important factors in generating a good model. It is important to choose representative features for the data, else the model would not learn much.

Here the features chosen were 'Rpt Dist No','Crm Cd','Vict Age','Vict Sex','Vict Descent','Premis Cd','LAT','LON’.

There are 3 umbrella factors that play a role in a crime, the first being location of the crime, second being the victim information like age,sex and descent and the third being the type of crime.The above features were chosen carefully to encompass these 3 categories. “Rpt Dist no”, “Premis Cd”, “LAT” and “LON” included the location description. “Vic Sex”,”Vict Desc” and “Vict Age” included the victim’s

information. “Crm Cd” described the type of crime that occured.

The above features would enable the model to get enough information to learn and predict/classify correctly.

## IX. Experimental Evaluation with Machine Learning Models

This is where the machine learning algorithms and their essential parameters were decided. For the purpose of this project, several algorithms, detailed below, were used to perform the two types of classification, namely Binary and Multi class classification.

### A. Binary and Multiclass Classification

Binary class refers to classification problems that have two class labels. Common problems that involve binary class classification are issuing emails as spam or not spam. On the other hand, multi class classification refers to problems that have more than two class labels. Common examples that are defined as multi class include predicting a set of values given previous information.

For binary class classification, the predicted value ‘y’ has to be split into binary values. We categorized the crimes in the dataset into two different groups: violent or not violent. This choice was done based on whether a weapon was used in the crime or not. If it was used, then it was considered a violent crime and if not, it was non-violent.

For multi-class classification, to classify crimes based on multiple labels, we used LAPD’s crime coding document to map the ranges in crime types. Crime were classified into kidnap, intent to murder or felony, crime against employers, human trafficing, crime again religion, injuries, crime against public peace, or anything else. These were mapped to numbers from 0-7.

### B. Machine Learning models

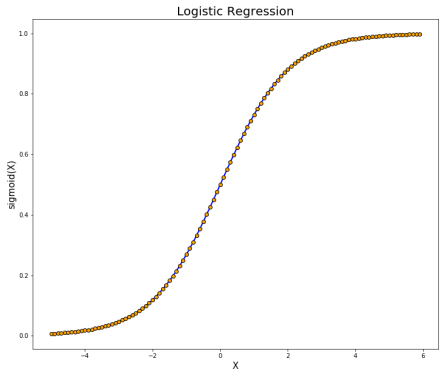
Popular machine learning algorithms from both classification problems include Logistic Regression, Naive Bayes, k-nearest Neighbors, Decision Trees, and Support Vector Machines. For the purpose of the given dataset, each algorithm was utilized under a different classification problem.

After choosing significant features from the dataset and classifying the data, the test and train data could be split. We used sklean’s train-test-split to make our test dataset and train dataset and the standard scaler to scale our values to smaller values.

In regards to feature selection, features are relatively the same as binary, except crime type mapping was added for multi class classification. Once again, sklean’s train-test-split was used to split it into test dataset and train dataset and the standard scaler was used to scale the values.

#### 1. Logistic Regression

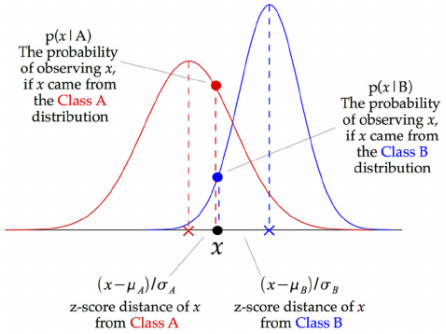
Logistic regression is often used to conduct predictive analysis. This is only applicable in situations where the dependent variable is binary. A relationship between the dependent variable and another independent variable in the data is required.




https://towardsdatascience.com/logistic-regression-explained-9ee73cede081

#### 2. Gaussian Naive Bayes Classifier

Naive Bayesian Classifier is a method in which the classifiers features are assumed to be independent of each other i.e. they contribute to the probability individuality.




https://www.researchgate.net/figure/Illustration-of-how-a-Gaussian-Naive-Bayes-GNB-classifier-works-For-each-data-point_fig8_255695722

#### 3. Bernouli Naive Bayes Classifier

Bernoulli is a classification algorithm based on Bayes theorem which gives the likelihood of occurrence of event or data. It is a probabilistic classifier for all different classes.

```
TrainBernoulliNB(C, D)
1  V ← ExtractVocabulary(D)
2  N ← CountDocs(D)
3  for each c ∈ C
4  do N_c ← CountDocsInClass(D, c)
5     prior[c] ← N_c/N
6     for each t ∈ V
7     do N_ct ← CountDocsInClassContainingTerm(D, c, t)
8        condprob[t][c] ← (N_ct + 1)/(N_c + 2)
9  return V, prior, condprob

ApplyBernoulliNB(C, V, prior, condprob, d)
1  V_d ← ExtractTermsFromDoc(V, d)
2  for each c ∈ C
3  do score[c] ← log prior[c]
4     for each t ∈ V
5     do if t ∈ V_d
6           then score[c] += log condprob[t][c]
7           else score[c] += log(1 − condprob[t][c])
8  return arg max_{c∈C} score[c]
```

https://nlp.stanford.edu/IR-book/html/htmledition/the-bernoulli-model-1.html

#### 4. Random Forest Classifier

Random Forest is an ensemble algorithm that builds many decision trees in a group-like structure. Each tree-like model represents a decision making step that can be defined both visually and explicitly. It aggregates the total score from each set of trees. As a result, an accurate prediction can be made from it.

*5. K-Nearest Neighbors*

KNN algorithm is an analysis method which predicts the class of unlabeled instances by using the classes of nearby neighbors.

*6. Neural networks*

Neural networks are modelled inherently to work with both binary and multi-class classification problems. They are a collection of connected nodes called neurons that group together to create perceptrons. In other words, it mimics brain activity and thought process.

*7. One vs Rest Classifier*

One vs rest classifier algorithm is a method for using binary classification algorithms for multi class algorithms. Due to this, we only conducted this test on a multi class classification basis since it cannot be compared to a binary couterpart. The method involves splitting a dataset into binary problems. Thus, conversion is a necessity to apply this classifier.

## X. Results

All experiments on the dataset were performed once with each algorithm under each type of classification. Results were evaluated on the basis of the accuracy performance on each algorithm. Accuracy is the ratio of how correctly labels were predicted relative to all predictions that were made initially. A low accuracy can show that an algorithm doesn't perform accurately given the data whereas a high accuracy leads to successful tests. However, keep in mind that high accuracy models could show signs of overfitting or bias.

### *A. Binary class classification results*

| ML algorithm | Accuracy Score |
|---|---|
| Logistic Regression | 62.90% |
| Gaussian Naive Bayes Classifier | 62.35% |
| Bernoulli Naive Bayes Classifier | 67.12% |
| Random Forest Classifier | 94.89% |
| K-Nearest Neighbours | 92.35% |
| Neural Networks | 84.82% |

It is clear from the results table for binary class classification that both random forest and k-nearest neighbors performed the best among all six algorithms. While both logistic regression and gaussian naive bayes classifier performed the worst with a difference of around 30% in terms of accuracy, this makes sense because random forest and KNN work well with large datasets such as this one.

Algorithms like logistic regression and gaussian naive bayes did not provide high accuracy due to the large number of fields that define a crime. A better accuracy could have been generated if less data was being processed by the algorithm. Even with cleaning parts of the dataset, there are noticeable differences amongst all six algorithms.

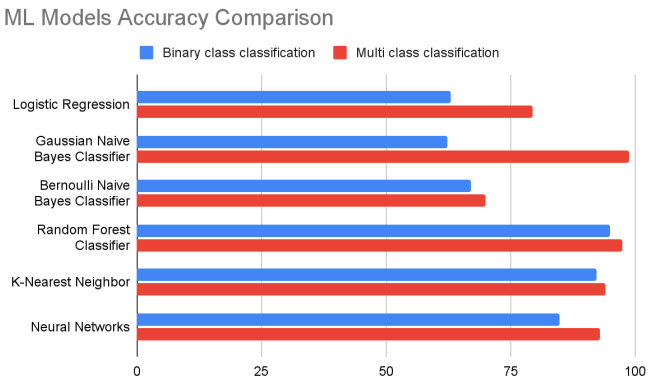
### *B. Multi-class classification results*

| ML algorithm | Accuracy Score |
|---|---|
| Logistic Regression | 79.40% |
| Gaussian Naive Bayes Classifier | 98.84% |
| Bernoulli Naive Bayes Classifier | 69.98% |
| Random Forest Classifier | 97.47% |
| K-Nearest Neighbours | 94.03% |
| Neural Networks | 92.86% |

The results obtained after performing multi-class classification ML algorithms show relatively higher accuracy when compared to binary classification. While random forest classifier and k nearest neighbors outputted a higher accuracy again, gaussian naive bayes and neural networks have shown an increased accuracy of around 25% each. This is likely due to ML algorithms being able to classify different components of a crime directly.

While both logistic regression and Bernoulli naive bayes received a lower accuracy in comparison to the other algorithms, we believe there was not a higher accuracy loss due to less noise in the data.

### *C. Comparative Analysis*




Overall, we can see that there was an improvement across all algorithms under multi-class when compared to its binary counterpart. These results come as no surprise since multi-classification techniques have seen better results when the imputed dataset is of greater scale than the ordinary.

## XI. Conclusion and Future Work

### *A. Conclusion*

In summary, this project aimed to showcase analysis and breakdown of a large-scaled dataset such as the crime data set provided from Los Angeles crime records. Many large cities are becoming infested in crimes, meaning there needs to be a

way to predict where the next one will occur is crucial to reduce violence.

The project experimented with a variety of preprocessing and cleaning techniques as well as utilizing different types of machine learning algorithms. Python libraries from Tensorflow and scikit-learn were incorporated into our project and have helped us ease into the process of training and testing the data into a prediction model.

The experimental results proved that there were different performance accuracies recorded among all the algorithms. It is evident that Random forest and KNN performed consistently well over both classification sets. In addition, Random forest was better with an overall average accuracy of 96%. In the end, we were able to conclude that multi-class classification algorithms performed with a higher accuracy than binary classification. Hopefully, this project will be a stepping stone for future research in other cities and their crime data.

### *B. Future Work*

For future work, we hope that experiments can be made in other data processing platforms such as Apache Spark or Apache Hadoop. Implementing machine learning models on them would provide more insight since they were designed to work with large datasets. With them, data can be distributed in a manner. Finally, a deployment of the prediction model would be useful to provide a visualization on where potential crime would take place in a given city's location.

## ACKNOWLEDGMENT

We would like to thank Dr. Wu for assisting and guiding us throughout this project, as well as giving us the opportunity to share our experiences and thought process throughout the project and the course. We would also like to thank our classmates for their thought provoking questions and suggestions that helped us improve this project.

## REFERENCES

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] Sun, Yao, C.-L., Li, X., & Lee, K. (2014). Detecting crime types using classification algorithms. Journal of Digital Information Management, 12(5), 321-327.

[3] Shojaee, Somayeh & Mustapha, Aida & Sidi, Fatimah & A. Jabar, Marzanah. (2013). A Study on Classification Learning Algorithms to Predict Crime Status. International Journal of Digital Content Technology and its Applications. 7. 361-369. 10.4156/jdcta.vol7.isssue9.43

[4] Shah, Neil, Bhagaat, Nandish, Shah, Manan. (2021). Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Shat et al. Visual Computing for Industry, Biomedicine, and Art, 16(5) 421-492

[5] Stalidis, Panagiotis, Semertzidis, Theodoreos, Daras, Petros. (2019) Examining Deep Learning Architectures for Crime Classification and Prediction. Forecasting 2021, 3, 741-762