Course Project

Spatio Temporal Crime Classification

CS5803 : Natural Language Processing

Tahir Ahmed(CS20MTECH14007)

Utkarsh Surwade(AI20MTECH11004)

Vaibhav Gaydhane(AI20MTECH11002)

I. Crime Classification

I.I Problem Description and Formulation

Through this project we aim to achieve the task of spatio-temporal analysis of crime incidents, through which we would be able to predict the occurrences of different crimes given the spatial features such as address, location coordinates and date and time of the incidents. We use several classification methods and approaches to achieve this task, by training the classifier using the input features such as address, location coordinates, date and time, and the target feature being the category of crime that is to be predicted. We then compare the different classifiers with respect to a set of metrics such as accuracy, precision etc.

I.II Introduction

Crime analysis is an important field or a branch of Forensic studies or Forensic Data Analysis. (FDA)**[1]** which focuses on studying the different patterns and making decisions with respect to investigational aspects of the crime incidents. This study helps throughout the investigation and identification of crimes. Crime classification involves the task of prediction of the crime patterns with respect to certain spatio-temporal features which include attributes such as places or locations , date and time etc, analysing such features enables us to understand the future patterns and occurrences of such incidents across the set of spatial features and enables the system to make predictions with respect to category of crimes that might be most likely to occur.

I.III Motivation

Crime is one such factor that needs to be greatly identified, observed and restricted for the sustainability of society, promoting better and safe living standards and environments. There is a great need for a better and ever growing system that governs and controls such activities. Such a system involves various governing and authoritarian forces such as the Department of Police, Crime Investigation Departments, Special Task forces etc. But it is a fact that Crime is one such activity which outlasts in numbers, it might happen anywhere , anytime and under such a case it adds up to high overhead to such governing forces and makes their functioning more difficult. Hence in this scenario the technology can prove to be a great aid for them and a gateway to offload such an overhead. The advancements in technology domains such as ML, AI can provide techniques such as analysis of patterns and knowledge discovery that can be used for making predictions, understanding future actions and help in future decision making abilities. It can enable them to take proactive measures to restrict and support better efficiency.

II. Dataset Details

For this task we choose the publicly available dataset from the Kaggle repository**[2]**. The Dataset in focus represents the crime incident data in the area of San-Francisco, United States (USA). It Contains the crime data over the period of 2003-2015. The dataset contains the following set of attributes:

- Dates - The timestamp of the crime incident

- Category - category of the crime incident (The Target Feature for prediction)
- Descript - detailed description of the crime incident.
- DayOfWeek - the day of the week
- PdDistrict - name of the Police Department District
- Resolution - Information regarding the resolution status of the crime.
- Address - The approximate street address of the crime incident
- X - Longitude
- Y - Latitude

No of Input Parameters : 8

No of Target Parameters : 1

Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
13-05-2015 23:53	WARRANTS	WARRANT ARR	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.4258917	37.7745986
13-05-2015 23:53	OTHER OFFENSE	TRAFFIC VIOLA	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.4258917	37.7745986
13-05-2015 23:33	OTHER OFFENSE	TRAFFIC VIOLA	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREEN	-122.424363	37.80041432
13-05-2015 23:30	LARCENY/THEFT	GRAND THEFT F	Wednesday	NORTHERN	NONE	1500 Block of LOMBA	-122.4269953	37.80087263
13-05-2015 23:30	LARCENY/THEFT	GRAND THEFT F	Wednesday	PARK	NONE	100 Block of BRODER	-122.4387376	37.77154117
13-05-2015 23:30	LARCENY/THEFT	GRAND THEFT F	Wednesday	INGLESIDE	NONE	0 Block of TEDDY AV	-122.4032524	37.7134307
13-05-2015 23:30	VEHICLE THEFT	STOLEN AUTOM	Wednesday	INGLESIDE	NONE	AVALON AV / PERU A	-122.423327	37.72513804
13-05-2015 23:30	VEHICLE THEFT	STOLEN AUTOM	Wednesday	BAYVIEW	NONE	KIRKWOOD AV / DON	-122.3712743	37.72756407
13-05-2015 23:00	LARCENY/THEFT	GRAND THEFT F	Wednesday	RICHMOND	NONE	600 Block of 47TH AV	-122.508194	37.77660126
13-05-2015 23:00	LARCENY/THEFT	GRAND THEFT F	Wednesday	CENTRAL	NONE	JEFFERSON ST / LEAV	-122.4190877	37.80780155
13-05-2015 22:58	LARCENY/THEFT	PETTY THEFT FR	Wednesday	CENTRAL	NONE	JEFFERSON ST / LEAV	-122.4190877	37.80780155

Fig : Part of the Actual Dataset.

II.I Dataset preprocessing

Before the dataset is used for training the classifiers, it is supposed to be formatted and preprocessed through a set of processes. These processes include the standard NLP processes. The "DayOfWeek" column is a categorical column which needs to be encoded or represented using embeddings to be inputted to the ML algorithms. The **"Category"** column that is the target feature to be predicted also needs to be encoded into integer label type. We drop the columns Address, Descript, Resolution as they no longer help in this process for predictions.Location is already captured using the Latitude and Longitude hence the Address column is dropped.

II.II Data Encoding / Embedding

The "**dayofWeek"** column is encoded using integer labels and also the Category column is encoded to an integer label, this label mapping is saved into a dictionary structure and dumped locally to be used later in predictions using trained models.

III. Proposed Methodology

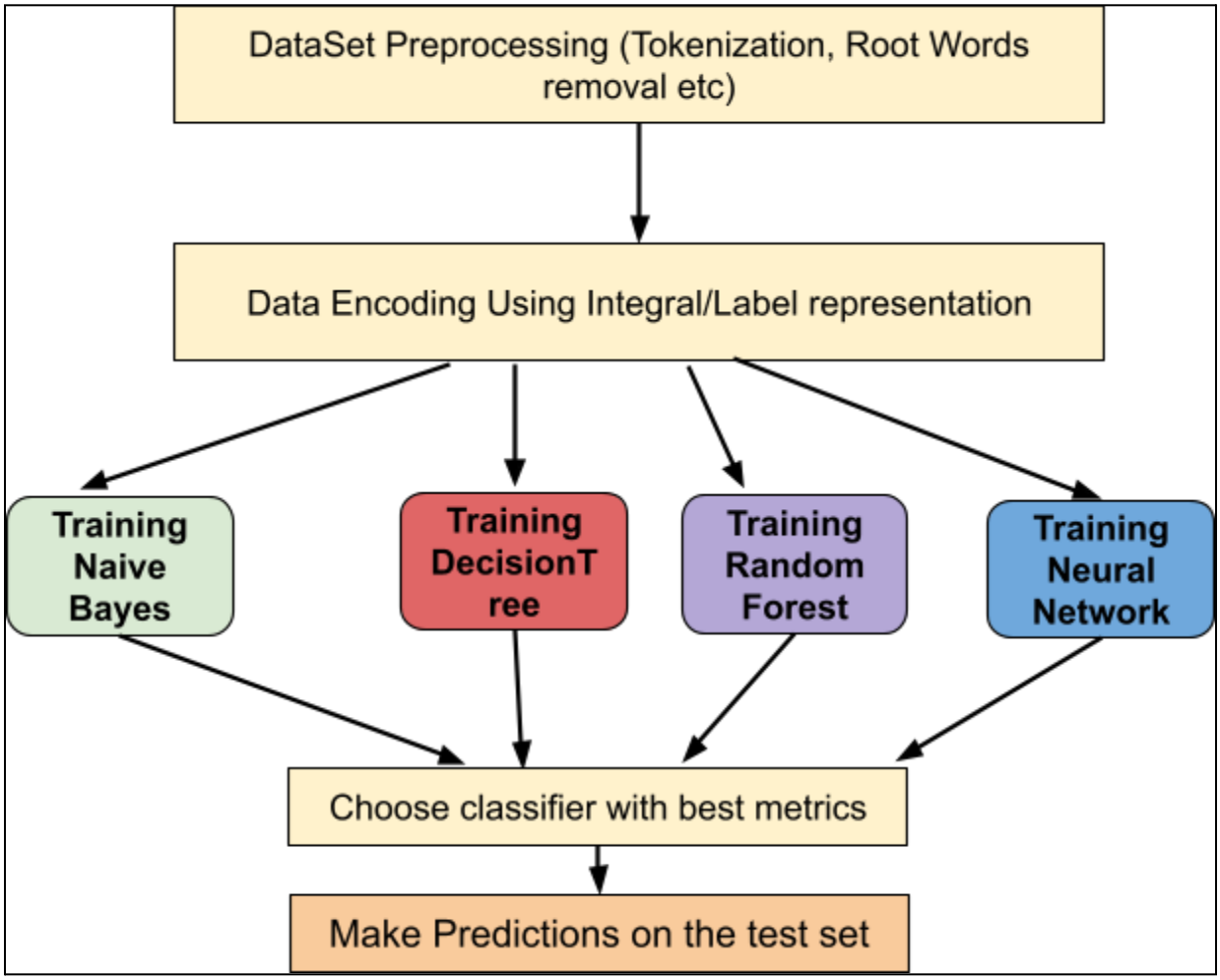


Fig : Proposed Methodology

For this task we propose the methodology as illustrated above, where the dataset is first preprocess to suit the task using suitable processes of NLP and other standards, then the data is fed across a set of classifiers for training, the output metrics then derived from the classifiers such as the accuracy, precision etc , the best classifier is chosen and then the predictions are made on testing set. We can then make comparisons across the classifiers and plot the results.

IV. Preliminary Results/ Observations:

IV.I Processed Dataset

The dataset after the preprocessing is as shown below. After the preprocess
No of Input Features : 9 (Date, Month, Year, Hours,Minutes, DayofWeek,PDistrict,X,Y)
No of Target Features : 1 (Category)

Date	Month	Year	Hours	Minutes	DayofWeek	PDistrict	X	Y	Category
13	5	2015	23	53	3	0	-122.425892	37.7745986	0
13	5	2015	23	53	3	0	-122.425892	37.7745986	1
13	5	2015	23	33	3	0	-122.424363	37.80041432	1
13	5	2015	23	30	3	0	-122.426995	37.80087263	2
13	5	2015	23	30	3	1	-122.438738	37.77154117	2
13	5	2015	23	30	3	2	-122.403252	37.7134307	2
13	5	2015	23	30	3	2	-122.423327	37.72513804	3
13	5	2015	23	30	3	3	-122.371274	37.72756407	3
13	5	2015	23	0	3	4	-122.508194	37.77660126	2
13	5	2015	23	0	3	5	-122.419088	37.80780155	2
13	5	2015	22	58	3	5	-122.419088	37.80780155	2
13	5	2015	22	30	3	6	-122.487983	37.73766665	1

Fig : Initial Preprocessed dataset.

IV.II Visualizations / EDA of the Dataset in Focus:

The different crimes present in the dataset can be vizualized as the wordcloud shown below:

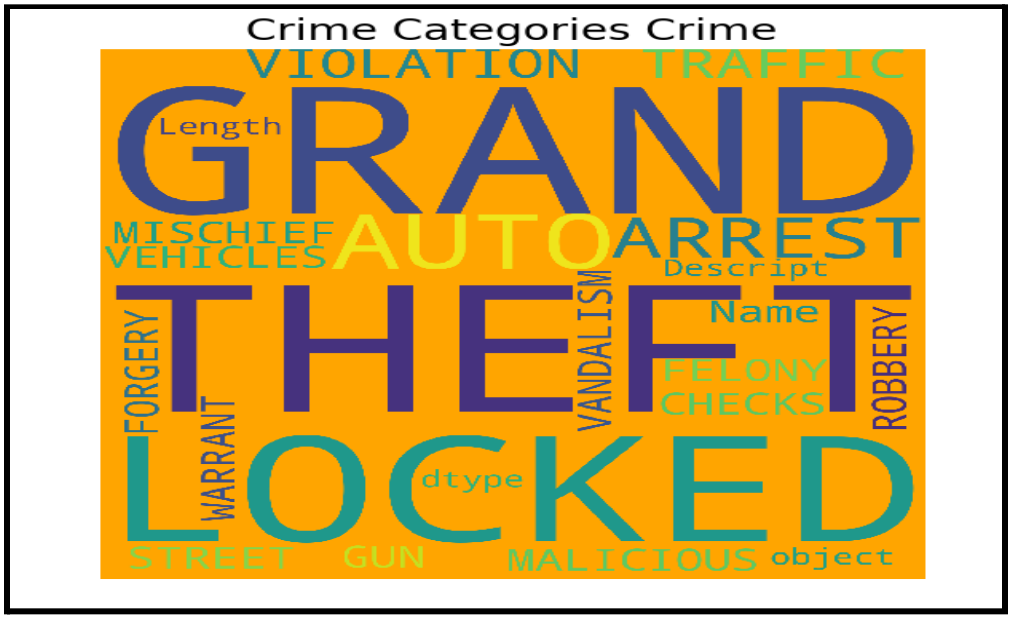
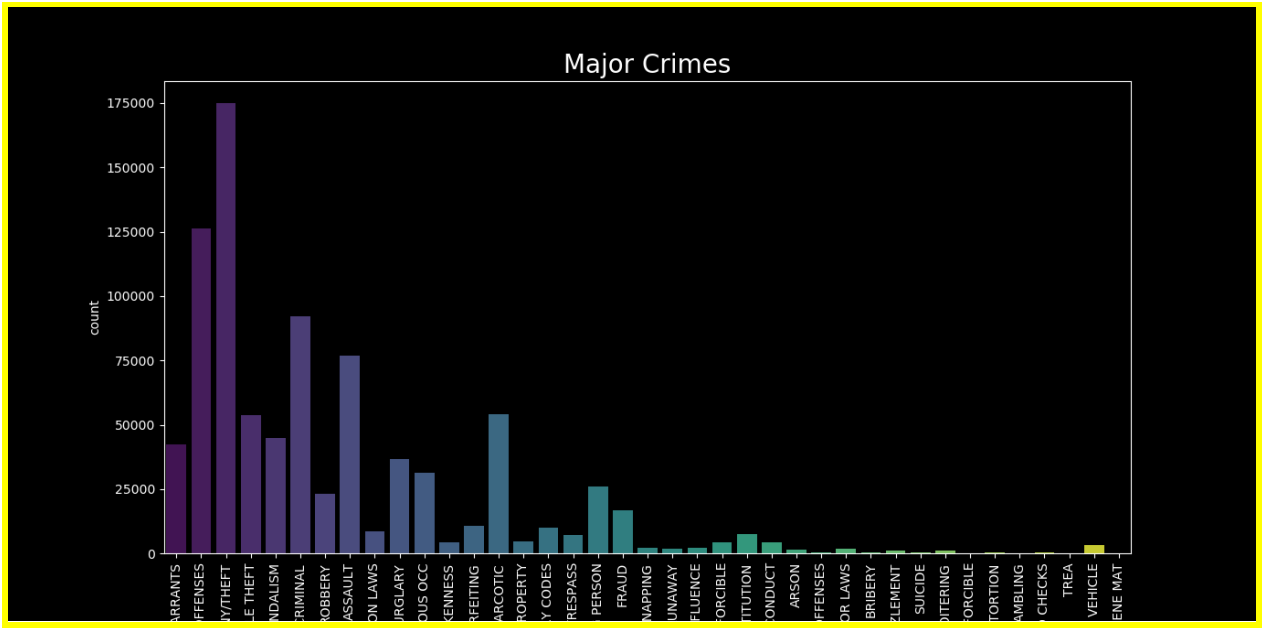


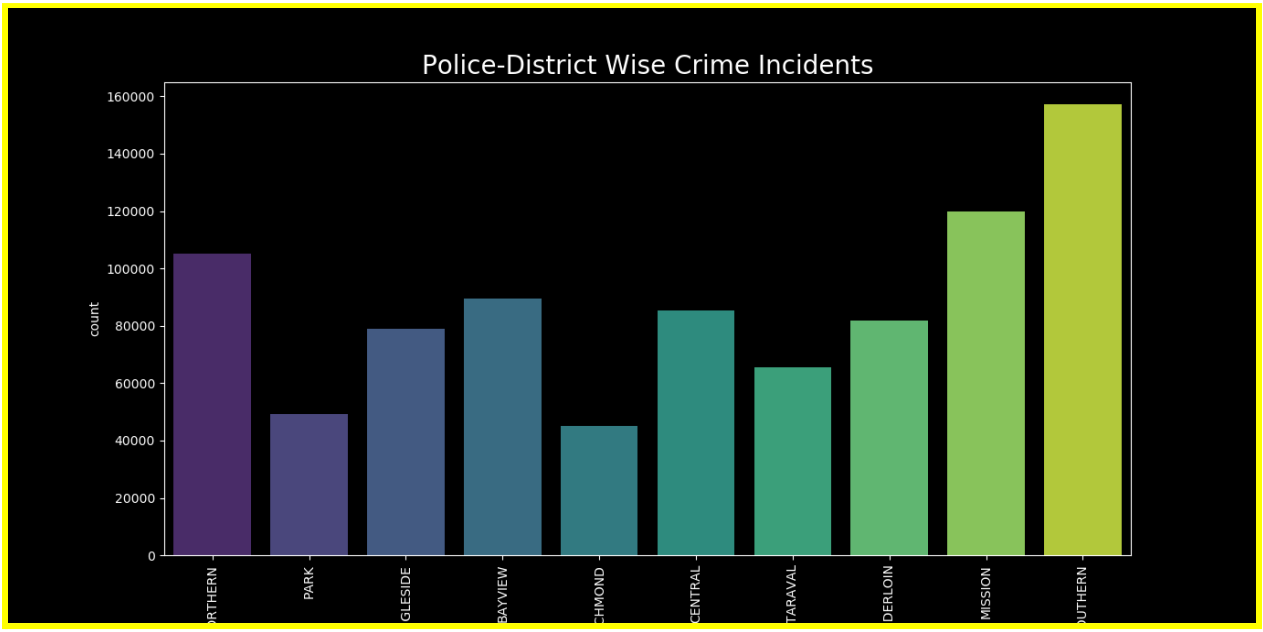
Fig : Word cloud of the crime categories in dataset

Category-Wise Count

The following represents the various categories of crime, with respect of their counts in the dataset. The highest density is shown by the crime category of ***"LANCERY/THEFT".***



Police-District Wise Crime Incidents



Crimes based on days of week

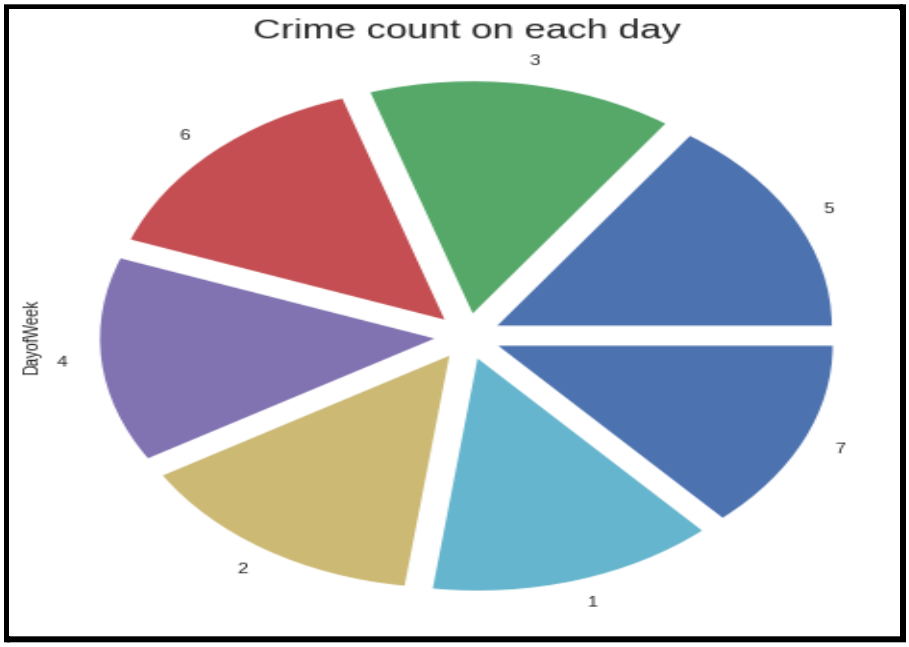


Fig : Crimes Per Day of Week

Crime Occurrences per day in each hour:

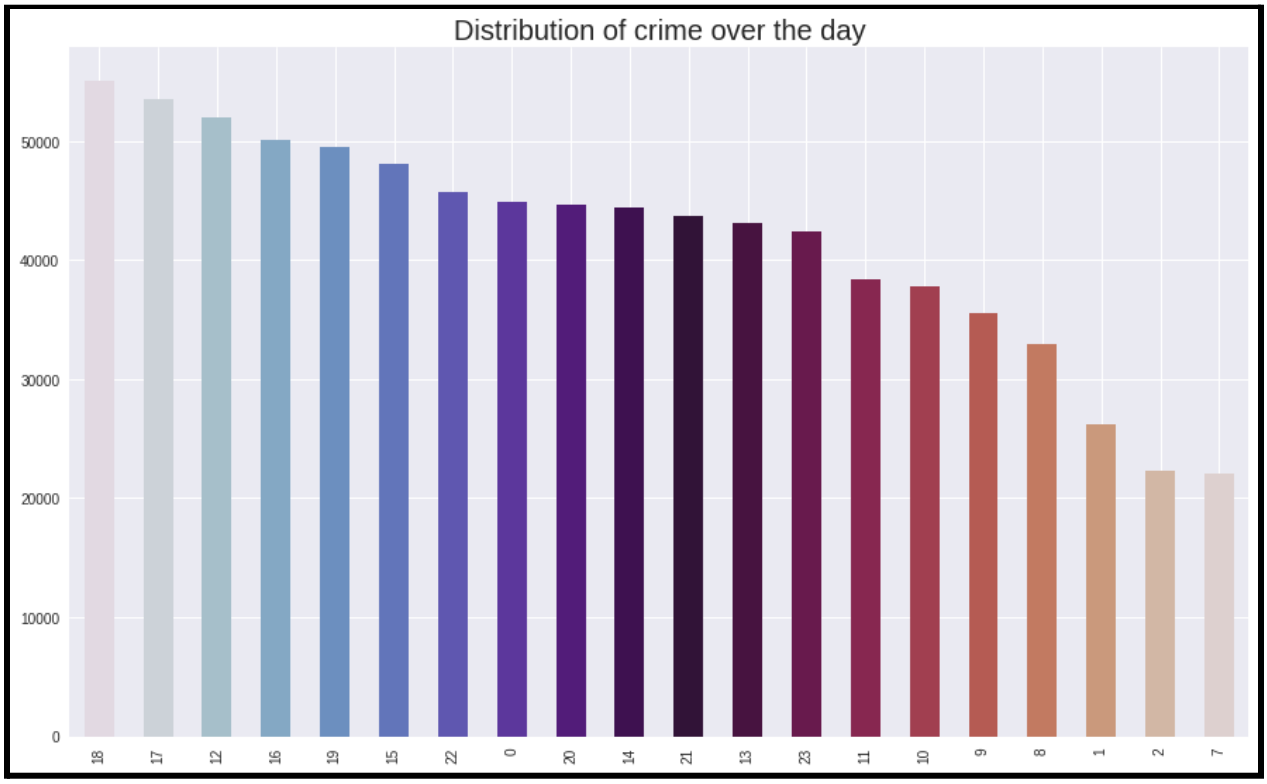


Fig : Crime Distribution Over the Hours of a day

V. Experimentation and Analysis

V.I Preliminary Training Accuracies achieved :

The following are the initial set of accuracies that are obtained by training the models and making predictions on the preprocessed dataset as shown above.

Classifiers	Accuracy	Performance
Random Forest	~30.28 %	**Highest**
Naive Bayes	~7.72 %	**Lowest**
Decision Tree	~23.28 %	**Mediocre**
Neural Network (FNN)	~14.47%	**Low**

```
----------CLASSIFICATION REPORTS AND ACCURACY---------

Classifier                              Accuracy
Naive Bayes                              7.729058709640681
RandomForest                             30.288707932350096
Decision Tree                            23.285120437332726
Neural Network                           14.47297989863903
```

Observations :

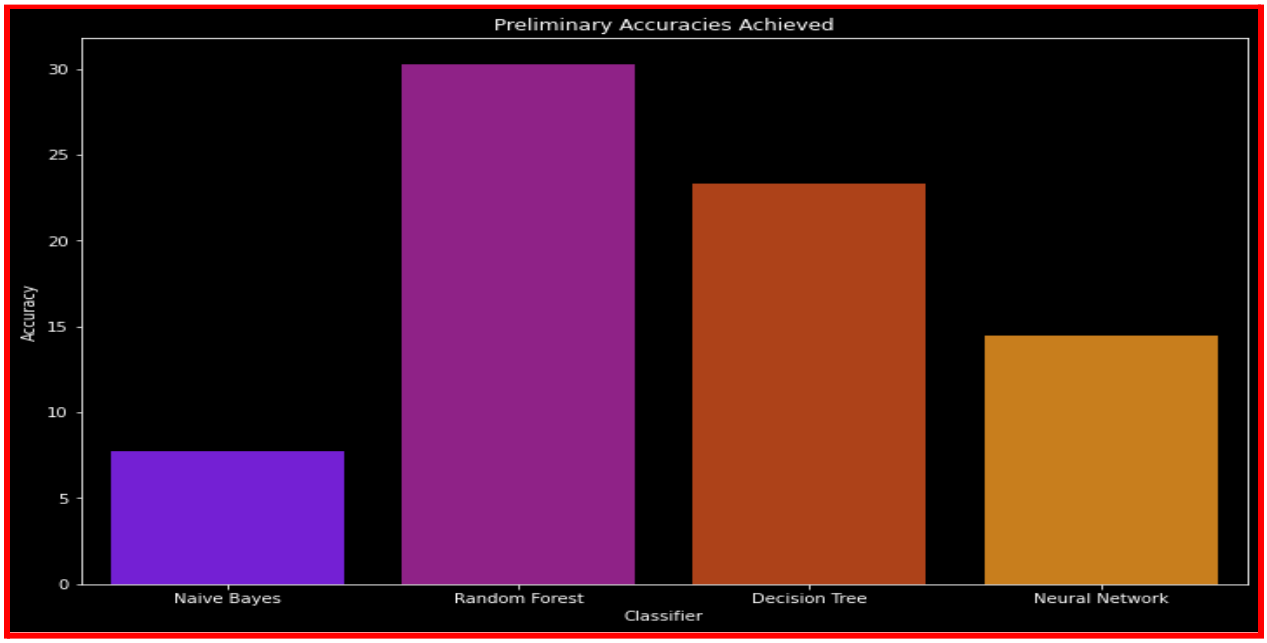


Fig : Accuracies achieved for classifiers

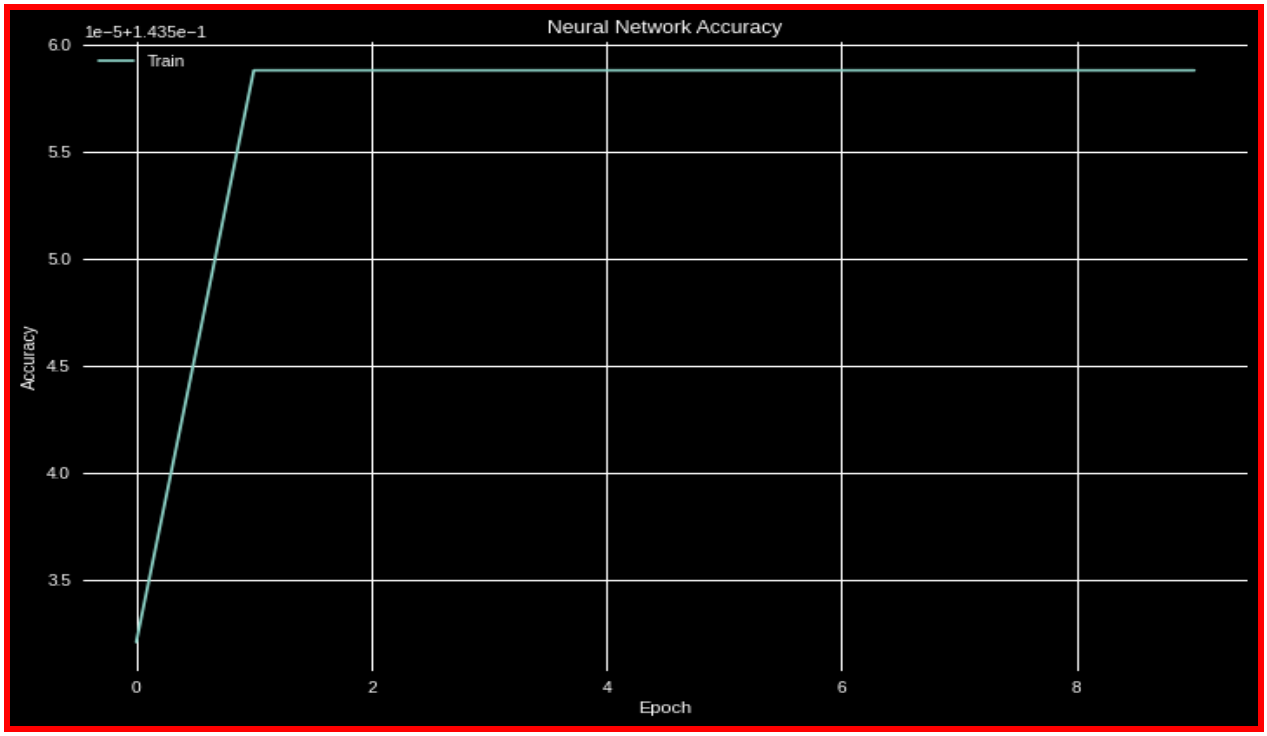


Fig : Neural Network Training Model Accuracy

We can see that the accuracy for all the classifiers are low i.e between 7 to 30 %. The Random forest classifier scored the highest accuracy of ~30.26 %. The main cause behind the low accuracy of the models is the ***disproportionate*** dataset, where a few classes have high density compared to the other classes as shown in the category wise crimes plot of the dataset. Hence further ahead, we look at improving the accuracy of the classifiers through certain approaches.

V.II Improving the accuracy

Crime category aggregation:

To improve the accuracy, the dataset is to proportionate with respect to the crime categories that is the target feature for our prediction. One way could be to possibly add synthetic or artificial data representing the crime incidents, but this leads to decrease in the efficiency of the model, because the classifier would train upon non- realistic data which is not our focus. Hence now, we try to aggregate the different classes of crimes under a more generalized class. We create two broad classes representing the ***Violent crimes*** and ***Nonviolent crimes***, All the sub crime categories that fall under violent crimes are placed under the Violent crime category, and similarly the Non-Violent category. The data set is re-processed for this purpose and an aggregated dataset is created as follows.

Non-Violent Crimes[23] -- >
(WARRANTS,OTHER_OFFENCES,LANCERY/THEFT,VEHICLE THEFT, NON_CRIMINAL, SUSPICIOUS OCC, DRUNKENNESS, FORGERY/COUNTERFEITING, STOLEN PROPERTY, SECONDARY CODES, TRESPASS, MISSING PERSON, FRAUD, KIDNAPPING, RUNAWAY, DRIVING UNDER THE INFLUENCE, LIQUOR LAWS, EMBEZZLEMENT, LOITERING, SEX OFFENSES NON FORCIBLE, GAMBLING, RECOVERED VEHICLE, BRIBERY)

Violent Crimes[15] --->
(VANDALISM, ROBBERY, ASSAULT, WEAPON LAWS, BURGLARY, DRUG/NARCOTIC, SEX OFFENSES FORCIBLE, PROSTITUTION, DISORDERLY CONDUCT, ARSON, FAMILY OFFENSES, SUICIDE, EXTORTION, TREA, PORNOGRAPHY/OBSCENE MAT)

Date	Month	Year	Hours	Minutes	DayofWeek	PDistrict	X	Y	Category
13	5	2015	23	53	3	0	-122.426	37.7746	1
13	5	2015	23	53	3	0	-122.426	37.7746	1
13	5	2015	23	33	3	0	-122.424	37.80041	1
13	5	2015	23	30	3	0	-122.427	37.80087	1
13	5	2015	23	30	3	1	-122.439	37.77154	1
13	5	2015	23	30	3	2	-122.403	37.71343	1
13	5	2015	23	30	3	2	-122.423	37.72514	1
13	5	2015	23	30	3	3	-122.371	37.72756	1
13	5	2015	23	0	3	4	-122.508	37.7766	1
13	5	2015	23	0	3	5	-122.419	37.8078	1
13	5	2015	22	58	3	5	-122.419	37.8078	1
13	5	2015	22	30	3	6	-122.488	37.73767	1
13	5	2015	22	30	3	7	-122.412	37.783	0
13	5	2015	22	6	3	0	-122.433	37.78435	1
13	5	2015	22	0	3	3	-122.398	37.72993	1
13	5	2015	22	0	3	3	-122.384	37.74319	1
13	5	2015	22	0	3	7	-122.413	37.78393	0

Fig : Aggregated Dataset with (Violent and Non-Violent crimes)

The category distribution for this dataset is visualized below. This dataset is now more balanced and proportionate. We now use this dataset to train the classifiers.

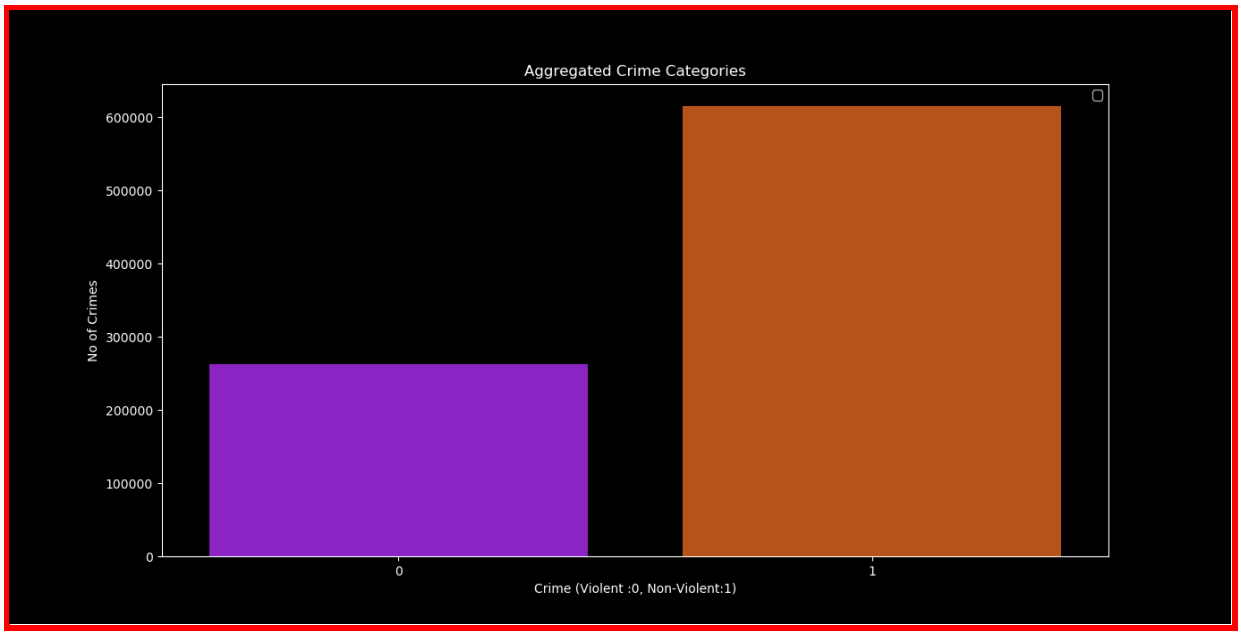


Fig : Aggregated Crime Categories Distribution

The new accuracies achieved after the aggregation is as follows:

After the dataset is more balanced across the crime categories by aggregating them. The dataset is used to train again the same set of classifiers. We obtain the following accuracies and performances by each classifier as shown below.

Classifiers	Accuracy	Performance
Random Forest	~68.89 %	**High**
Naive Bayes	~69.94 %	**High**
Decision Tree	~62.46 %	**Good**
Neural Network (FNN)	~70.29 %	**Highest**

```
----------CLASSIFICATION REPORTS AND ACCURACY---------

Classifier                          Accuracy
Naive Bayes                          69.94305563464496
RandomForest                         68.89926541768692
Decision Tree                        62.46569101987358
Neural Network                       70.29838847446045
```

We can see a very high and significant rise in the accuracies of all the classifiers. The following plot shows the accuracies of all the classifiers.

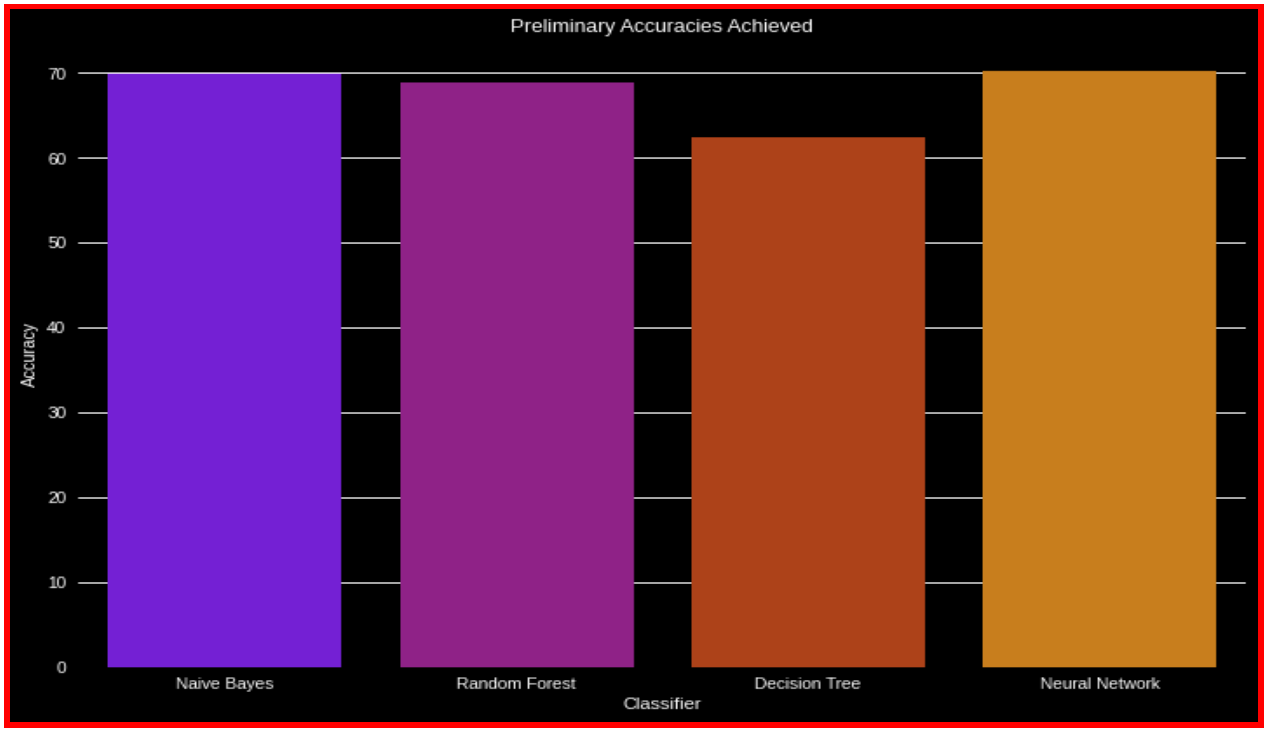


Fig : Accuracy of all classifiers for aggregated dataset

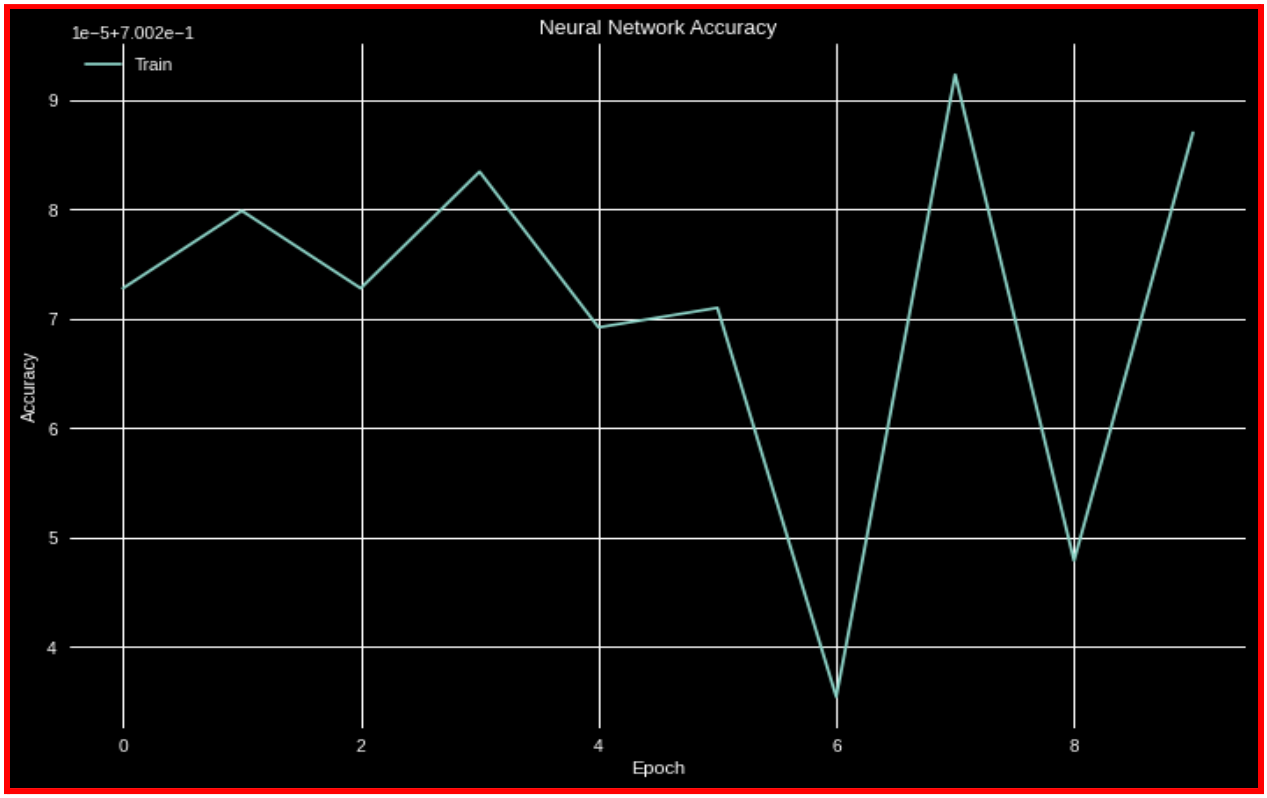


Fig : Training curve for the neural network

V.III Final Predictions on test data :

Hence we see that the Neural Network achieves the highest accuracy of ~70%. Hence it is chosen to make predictions. The NN classifier is now tested on a completely new test dataset that is unseen till now. The following are the predictions generated by the NN classifier.

Date	Month	Year	Hours	Minutes	DayofWeek	PDistrict	X	Y
10	5	2015	23	59	7	0	-122.4	37.73505
10	5	2015	23	51	7	0	-122.392	37.73243
10	5	2015	23	50	7	1	-122.426	37.79221
10	5	2015	23	45	7	2	-122.437	37.72141
10	5	2015	23	45	7	2	-122.437	37.72141
10	5	2015	23	40	7	3	-122.459	37.71317
10	5	2015	23	30	7	2	-122.426	37.73935
10	5	2015	23	30	7	2	-122.413	37.73975
10	5	2015	23	10	7	4	-122.419	37.76516
10	5	2015	23	10	7	5	-122.414	37.79889
10	5	2015	23	0	7	2	-122.409	37.74679
10	5	2015	23	0	7	4	-122.411	37.76105
10	5	2015	23	0	7	4	-122.411	37.76105

Fig : The Test Dataset

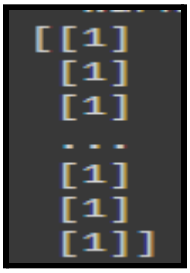


Fig Prediction by the NN Model**(Model With Highest Accuracy)**

The final model metrics are obtained by evaluating the model as follows. This shows a final accuracy of ~70% on the completely unseen dataset.

```
False
5488/5488 [==============================] - 5s 841us/step - loss: 0.5979 - accuracy: 0.7030
Model Loss : 0.597938060760498
Model Accuracy 0.7029838562011719
```

VI. Geo-Spatial Visualizing the Predictions of various classifier models:

Given the ***spatio-temporal*** data from the test dataset, the following are the predicted output classes of a few classifiers plotted across the San-Francisco geographical map to visualize the crime activity and incidence occurrence density.

1. Decision Tree Classifier: [62.46 %]

The trained Decision Tree Classifier with accuracy 69.94% predicted the categories as shown below.

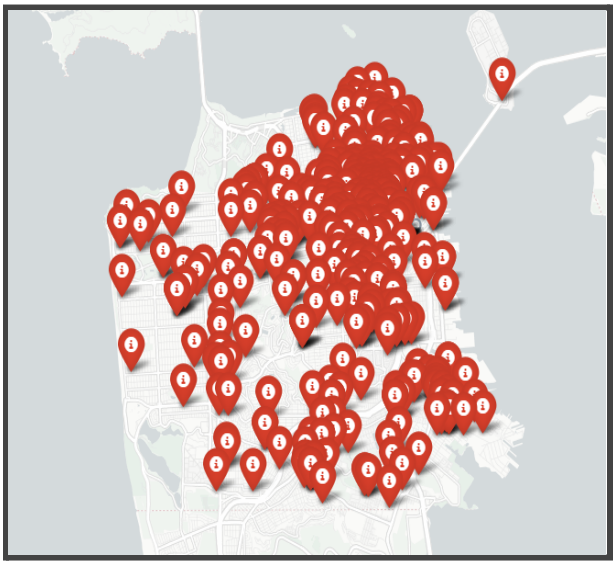


Fig : Violent Crimes Prediction

Fig : Non-violent Crimes Prediction

2. Naive Bayes Classifier [69.94%]

The trained Naive Bayes Classifier with accuracy 69.94% predicted the categories as shown below.

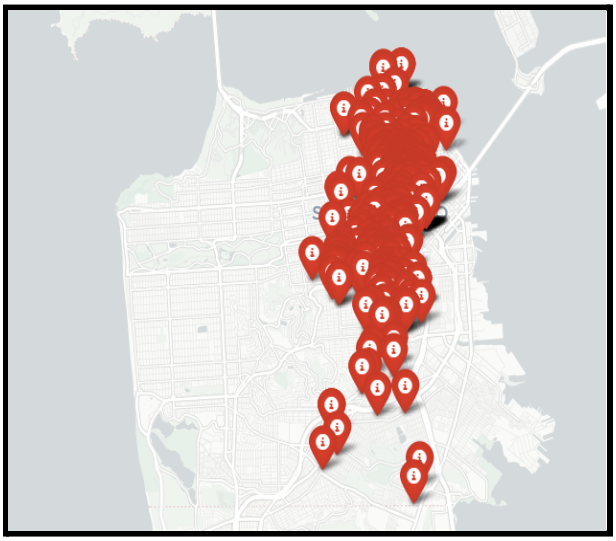


Fig : Violent Crimes Prediction

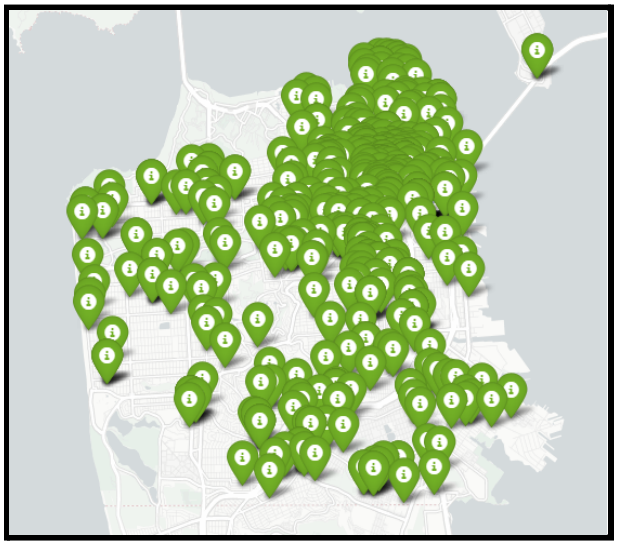
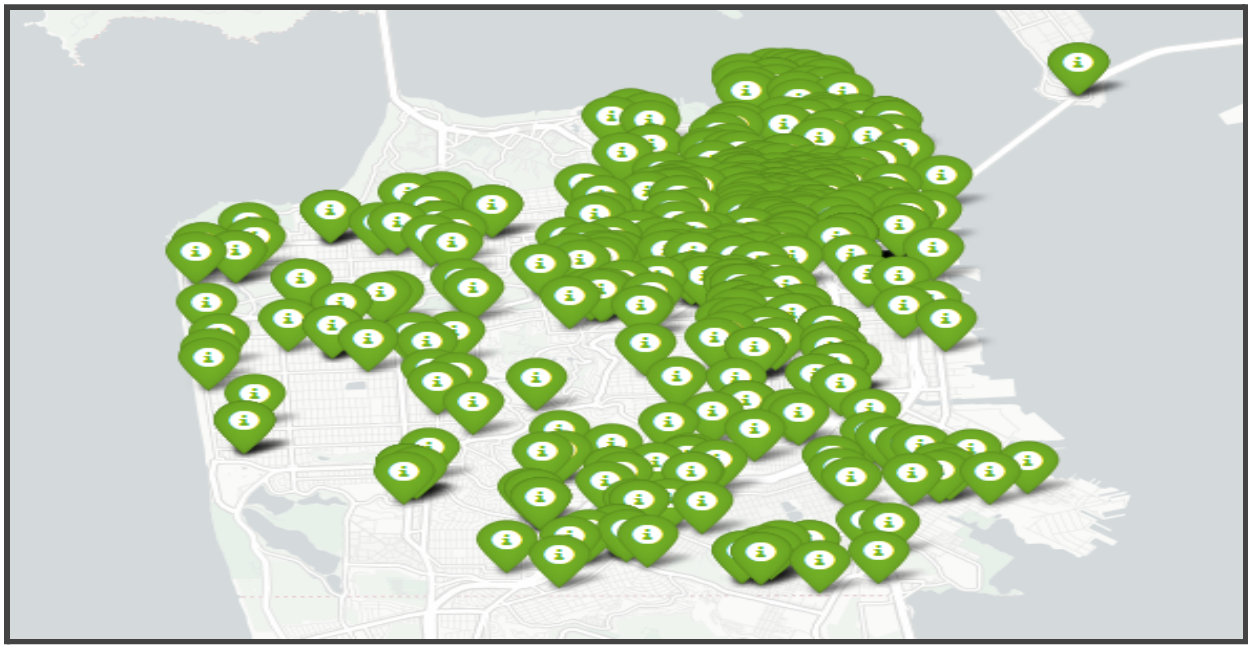


Fig : Non-violent Crimes Prediction

3. Neural Network (Highest Accuracy) **[Used In Final Predictions]**

This was the chosen model to predict on the test dataset with highest accuracy, it generated the following visualizations.



From the above plots, the main observation that can be drawn is that all the classifiers predict a huge crime activity (Violent / Non-Violent) crime in the ***North-Eastern*** part of the province. Hence it can be determined that most large amounts of criminal activity can be focussed into the ***North-Eastern*** part of the state.

<Project Code Base Can be Found on Colab on the following link :>
https://colab.research.google.com/drive/1p8X3Ohny6g_bSWr8_F33LPqlJOzzb5-z?usp=sharing

References

[1] Forensic data analysis - Wikipedia
[2] SF crime classification | Kaggle
[3] http://cs229.stanford.edu/proj2015/228_report.pdf
[4]https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9318972
[5] https://www.mdpi.com/2076-3417/10/22/8220/pdf