Report On

Data Analystic On Crime Data

Submitted in partial fulfillment of the requirements of the Mini project in Semester VI of Third Year Computer Engineering

by
Hemangi Jadhav (Roll No. 65)
Pranjal Mane (Roll No. 68)
Manasvi Mhatre(Roll No. 69)

Supervisor
Prof. Speril Demello

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(2020-21)

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

CERTIFICATE

This is to certify that the project entitled “Data Analystic On Crime Data” is a bonafide work of "Hemangi Jadhav(Roll No. 65), Pranjal Mane(Roll No. 68), Manasvi Mhatre(Roll No. 69)" submitted to the University of Third Year Computer Engineering.

Supervisor

Prof. Speril Demello

Internal Examiner

External Examiner

Dr Megha Trivedi
Head of Department

Dr. H.V. Vankudre
Principal

ACKNOWLEDGEMENT

We have immense pleasure in expressing our sincerest and deepest sense of gratitude towards our guide Prof. Speril Demello for the assistance, valuable guidance and co-operation in carrying out this Project successfully. We have developed this project with the help of Faculty members of our institute and we are extremely grateful to all of them. We also take this opportunity to thank Head of the Department Prof. Dr Megha Trivedi for providing the required facilities in completing this project. We are greatly thankful to our parents, friends and faculty members for their motivation, guidance and help whenever needed.

Abstract

To be better prepared to respond to criminal activity, it is important to understand patterns in crime. In our project, we analyze crime data from San Fransisco . At the outset, the task is to predict which category of crime is most likely to occur given a time and place . The use of AI/ML in predicting crimes or an individual's likelihood for committing a crime has promise but is still more of an unknown. The biggest challenge will probably be "proving" to politicians that it works. When a system is designed to stop something from happening, it is difficult to prove the negative. Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop. Improvements in crime prevention technology will likely spur increased total spending on this technology. We also attempt to make our classification task more meaningful by merging multiple classes into larger classes. Finally, we report and reflect on our results with different classifiers, and dwell on avenues for future work.

Contents

1.Problem Statement

Many important questions in public safety and protection relate to crime, and a better understanding of crime is beneficial in multiple ways: it can lead to targeted and sensitive practices by law enforcement authorities to mitigate crime, and more concerted efforts by citizens and authorities to create healthy neighborhood environments. With the advent of the Big Data era and the availability of fast, efficient algorithms for data analysis, understanding patterns in crime from data is an active and growing field of research.

With the rapid urbanization and development of big cities and towns, the graph of crimes is also on the increase. This phenomenal rise in offences and crime in cities is a matter of great concern and alarm to all of us.

There are robberies, murders, rapes and what not. The frequent and repeated thefts, burglaries, robberies, murders, killings, rapes, shoplifting, pick pocketing, drug- abuse, illegal trafficking, smuggling, theft of vehicles etc., have made the common citizens to have sleepless nights and restless days.

They feel very insecure and vulnerable in the presence of anti-social and evil elements. The criminals have been operating in an organized way and sometimes even have nationwide and international connections and links.

1.1 Goal :

Much of the current work is focused in two major directions:

· Predicting surges and hotspots of crime, and

· Understanding patterns of criminal behavior that could help in solving criminal investigations.

1.2 Objective The objective of our work is to:

· Predicting crime before it takes place.

· Predicting hotspots of crime.

· Understanding crime pattern.

· Classify crime based on location.

· Analysis of crime in Indore.

1.3 Methodology

Machine learning

The term machine learning refers to the automated detection of meaningful patterns in data. In the past couple of decades it has become a common tool in almost any task that requires information extraction from large data sets.We are surrounded by a machine learning based technology : search engines learn how to bring us the best results (while placing pro_table ads), anti-spam software learns to filter our email messages, and credit card transactions are secured by a software that learns how to detect frauds.

2 . Implementation

The implementation of the project is done with the help of python language. To be particular, for the purpose of machine learning Anaconda is being used.

Anaconda is one of several Python distributions. Anaconda is a new distribution of the Python. It was formerly known as Continuum Analytics. Anaconda has more than 100 new packages. Anaconda is used for scientific computing, data science, statistical analysis, and machine learning. On Python technology, we found out Anaconda to be easier. Since it helps with the following problems:

· Installing Python on multiple platforms.

· Separating out different environments.

· Dealing with not having correct privileges.

· Getting up and running with specific packages and libraries.

This data was scraped from kaggel . Implementation of the idea started from the San Fransisco city itself so as to limit an area for the prediction and making it less complex. The data was sorted and converted into a new format of timestamp, longitude, latitude, which was the input that machine would be taking so as to predict the crime rate in particular location or city.

The entries was done just to make the machine learn what all it has to do with the data and what actually the output is being demanded. As soon as the machine learnt the algorithms and the process, accuracy of different algorithms were measured & the algorithm with the most accuracy is used for the prediction kernel .

2.1 Implementation details

Prediction

For the purpose of proper implementation and functioning several Algorithms and techniques were used. Following are the algorithm used:

KNN (K-Nearest neighbors)

A powerful classification algorithm used in pattern recognition K nearest neighbors stores all available cases and classifies new cases based on a similarity measure (e.g. distance function).One of the top data mining algorithms used today. A non-parametric lazy learning algorithm (An Instance based Learning method).

KNN: Classification Approach

· An object (a new instance) is classified by a majority votes for its neighbor classes.
· The object is assigned to the most common class amongst its K nearest neighbors.(measured by distance function)

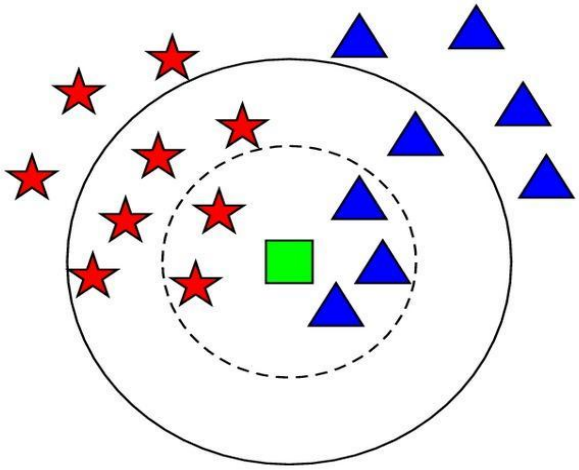


Fig. Principle diagram of KNN

Prediction has following Structure

1. Read and explore the train data
2. Understanding the Data
3. Target Variable
4. Read the test data
5. Understanding the features and the problems
6. Clean the Data
7. Visualize the Data
8. Model the Data using KNN

Understanding the data

It's important to understand all the columns before we move further. Train data has the following columns:

1. Dates - timestamp of the crime incident
2. Category - category of the crime incident (only in train.csv). This is the target variable you are going to predict.
3. Descript - detailed description of the crime incident (only in train.csv)
4. DayOfWeek - the day of the week
5. PdDistrict - name of the Police Department District
6. Resolution - how the crime incident was resolved (only in train.csv)
7. Address - the approximate street address of the crime incident
8. X - Longitude
9. Y - Latitude

Software Details :

· Anaconda Distribution (v5.1)

· Python (3.6.5)

· Packages Used:

  o Flask (0.12.2)

  o Pandas (0.22.1)

  o Numpy (1.14.2)

  o Sklearn (0.19.1)

  o Geopy (1.13.0)

· HTML 5

· CSS 3

Hardware Details :

· Operating system: Windows 7 or newer, 64-bit macOS 10.9+, or Linux.

· System architecture: 64-bit x86, 32-bit x86 with Windows or Linux.

· CPU: Intel Core 2 Quad CPU Q6600 @ 2.40GHz or greater.

· RAM: 4 GB or greater.

3. Code

```
 # for some basic operations
import numpy as np
import pandas as pd

# for visualizations
import matplotlib.pyplot as plt
import seaborn as sns
import folium
import squarify

# for providing path
import os
print(os.listdir("../input"))

# reading the dataset

data = pd.read_csv('../input/Police_Department_Incidents_-_Previous_Year__2016_.csv')

# check the shape of the data
data.shape

# checking the head of the data

data.head()

# describing the data

data.describe()
# checking if there are any null values
```

```
data.isnull().sum()
# filling the missing value in PdDistrict using the mode values

data['PdDistrict'].fillna(data['PdDistrict'].mode()[0], inplace = True)

data.isnull().any().any()
y = data['Category'].value_counts().head(25)

plt.rcParams['figure.figsize'] = (15, 15)
plt.style.use('fivethirtyeight')

color = plt.cm.magma(np.linspace(0, 1, 15))
squarify.plot(sizes = y.values, label = y.index, alpha=.8, color = color)
plt.title('Tree Map for Top 25 Crimes', fontsize = 20)

plt.axis('off')
plt.show()

# description of the crime

from wordcloud import WordCloud

plt.rcParams['figure.figsize'] = (15, 15)
plt.style.use('fast')

wc = WordCloud(background_color = 'orange', width = 1500, height =
1500).generate(str(data['Descript']))
plt.title('Description of the Crime', fontsize = 20)

plt.imshow(wc)
plt.axis('off')
plt.show()

# Regions with count of crimes
```

```
plt.rcParams['figure.figsize'] = (20, 9)
plt.style.use('seaborn')

color = plt.cm.spring(np.linspace(0, 1, 15))
data['PdDistrict'].value_counts().plot.bar(color = color, figsize = (15, 10))

plt.title('District with Most Crime',fontsize = 30)

plt.xticks(rotation = 90)
plt.show()

# Regions with count of crimes

plt.rcParams['figure.figsize'] = (20, 9)
plt.style.use('seaborn')

color = plt.cm.ocean(np.linspace(0, 1, 15))
data['Address'].value_counts().head(15).plot.bar(color = color, figsize = (15, 10))

plt.title('Top 15 Regions in Crime',fontsize = 20)

plt.xticks(rotation = 90)
plt.show()
```

```
# Regions with count of crimes

plt.style.use('seaborn')


data['DayOfWeek'].value_counts().head(15).plot.pie(figsize = (15, 8), explode = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1))

plt.title('Crime count on each day',fontsize = 20)
```

```
plt.xticks(rotation = 90)
plt.show()
```

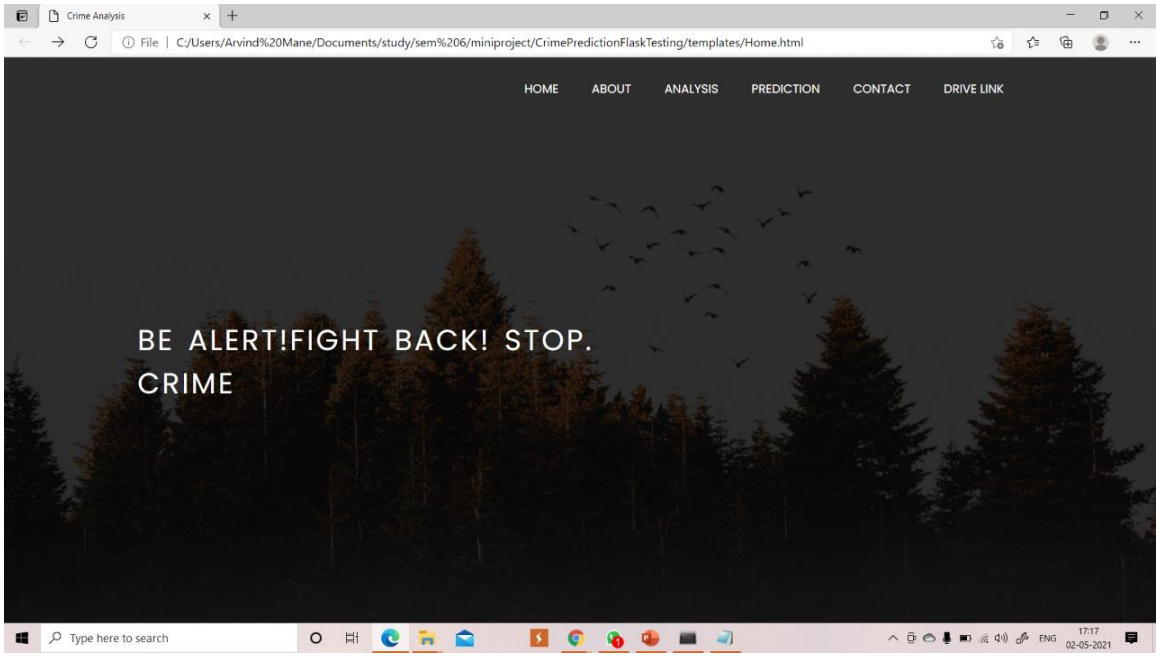
```
# Regions with count of crimes

plt.style.use('seaborn')

color = plt.cm.winter(np.linspace(0, 10, 20))
data['Resolution'].value_counts().plot.bar(color = color, figsize = (15, 8))

plt.title('Resolutions for Crime',fontsize = 20)
plt.xticks(rotation = 90)
plt.show()
```

```
data['Date'] = pd.to_datetime(data['Date'])

data['Month'] = data['Date'].dt.month

plt.style.use('fivethirtyeight')
plt.rcParams['figure.figsize'] = (15, 8)

sns.countplot(data['Month'], palette = 'autumn',)
plt.title('Crimes in each Months', fontsize = 20)

plt.show()
# checking the time at which crime occurs mostly

import warnings
warnings.filterwarnings('ignore')

color = plt.cm.twilight(np.linspace(0, 5, 100))
data['Time'].value_counts().head(20).plot.bar(color = color, figsize = (15, 9))
```

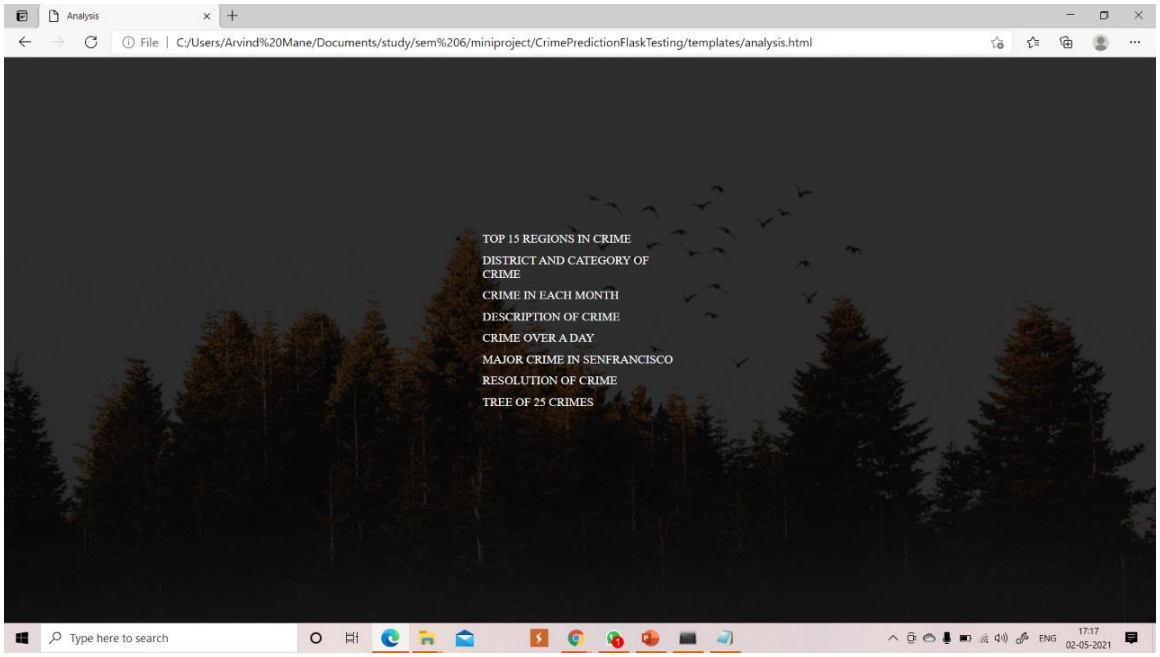
```
plt.title('Distribution of crime over the day', fontsize = 20)
plt.show()
df = pd.crosstab(data['Category'], data['PdDistrict'])
color = plt.cm.Greys(np.linspace(0, 1, 10))

df.div(df.sum(1).astype(float), axis = 0).plot.bar(stacked = True, color = color, figsize =
(18, 12))
plt.title('District vs Category of Crime', fontweight = 30, fontsize = 20)

plt.xticks(rotation = 90)
plt.show()
```

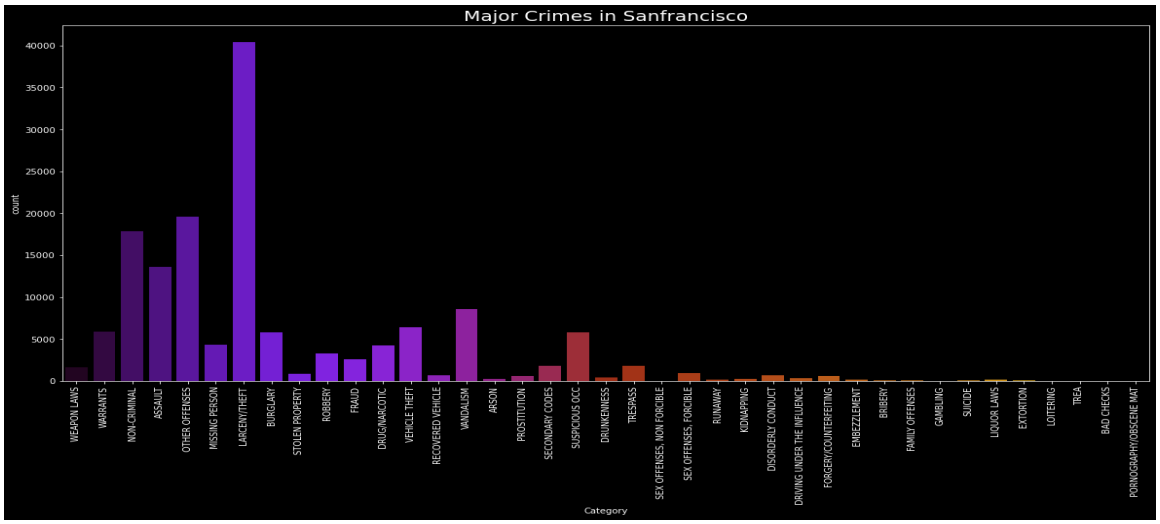
4. output

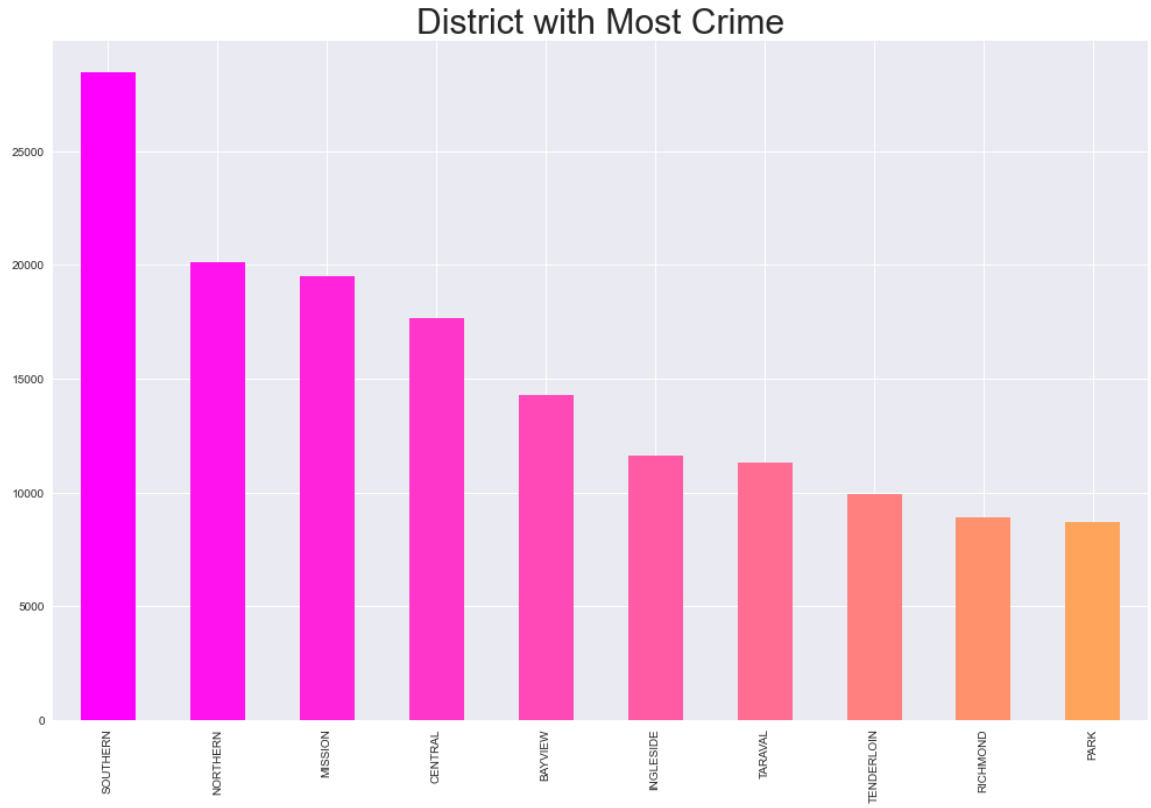
Home page

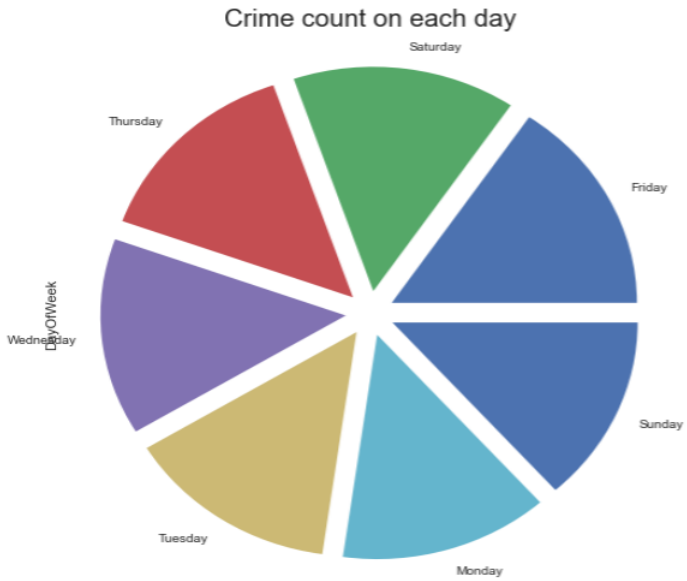


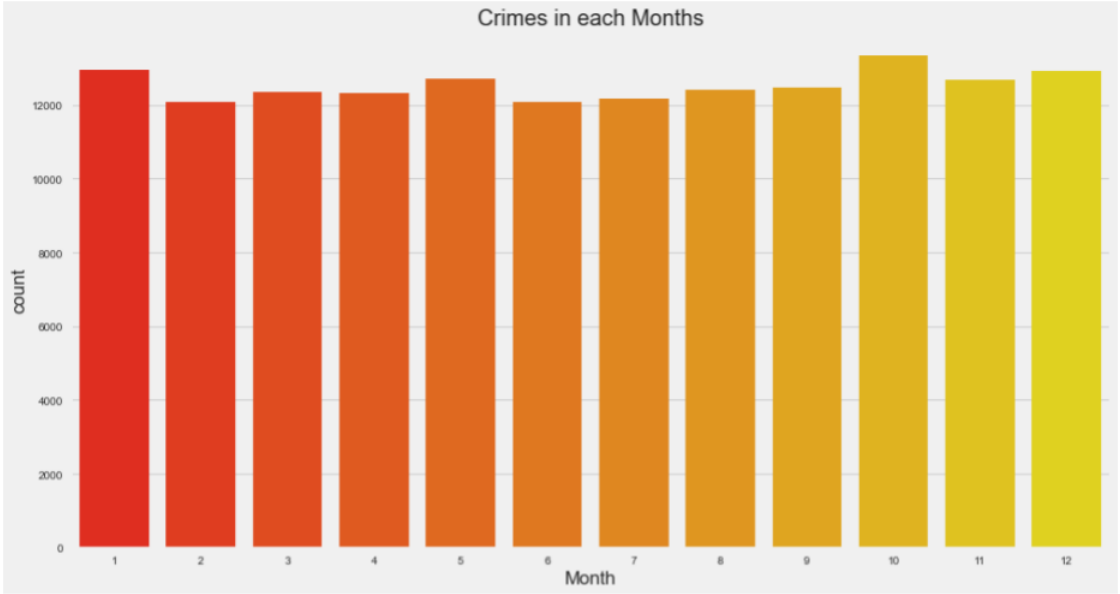
Analysis



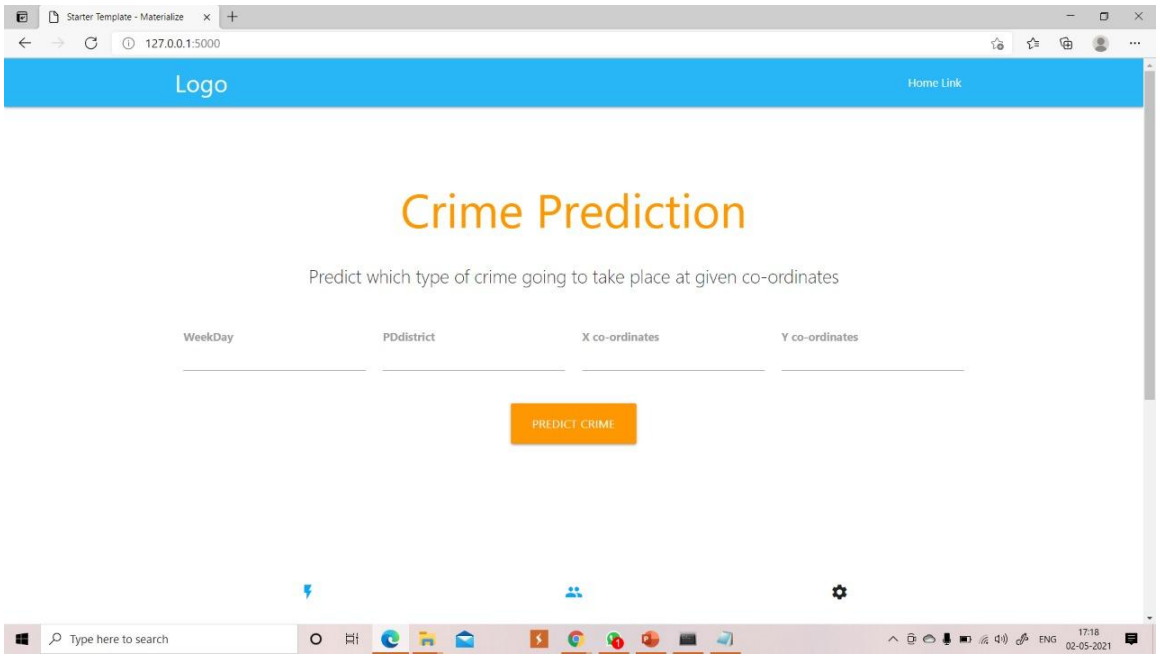

Major Crimes in Sanfrancisco
count
Category
WEAPON LAWS
WARRANTS
NON-CRIMINAL
ASSAULT
OTHER OFFENSES
MISSING PERSON
LARCENY/THEFT
BURGLARY
STOLEN PROPERTY
ROBBERY
FRAUD
DRUG/NARCOTIC
VEHICLE THEFT
RECOVERED VEHICLE
VANDALISM
ARSON
PROSTITUTION
SECONDARY CODES
SUSPICIOUS OCC
DRUNKENNESS
TRESPASS
SEX OFFENSES, NON FORCIBLE
SEX OFFENSES, FORCIBLE
RUNAWAY
KIDNAPPING
DISORDERLY CONDUCT
DRIVING UNDER THE INFLUENCE
FORGERY/COUNTERFEITING
EMBEZZLEMENT
BRIBERY
FAMILY OFFENSES
GAMBLING
SUICIDE
LIQUOR LAWS
EXTORTION
LOITERING
TREA
BAD CHECKS
PORNOGRAPHY/OBSCENE MAT

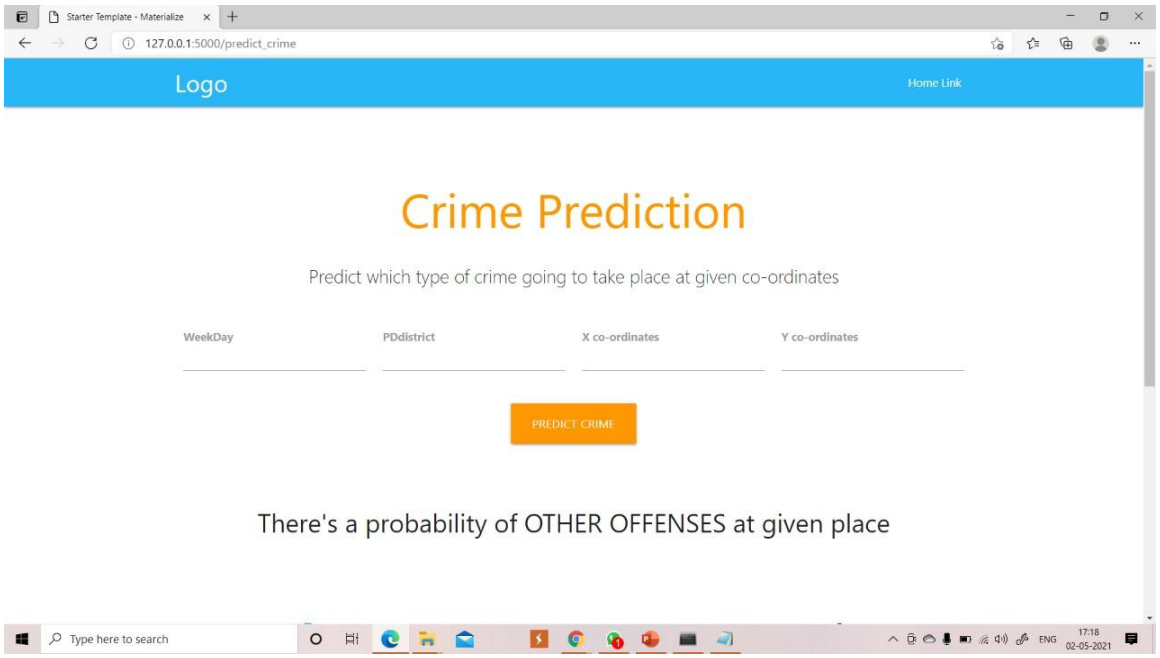

District with Most Crime
SOUTHERN
NORTHERN
MISSION
CENTRAL
BAYVIEW
INGLESIDE
TARAVAL
TENDERLOIN
RICHMOND
PARK


Crime count on each day
Saturday
Thursday
Friday
DayOfWeek
Wednesday
Sunday
Tuesday
Monday


Crimes in each Months
12000
10000
8000
6000
4000
2000
0
count
1
2
3
4
5
6
7
8
9
10
11
12
Month

Prediction


Starter Template - Materialize
127.0.0.1:5000
Logo
Home Link
Crime Prediction
Predict which type of crime going to take place at given co-ordinates
WeekDay
PDdistrict
X co-ordinates
Y co-ordinates
PREDICT CRIME
Type here to search
ENG
17:18
02-05-2021


Starter Template - Materialize
127.0.0.1:5000/predict_crime
Logo
Home Link
Crime Prediction
Predict which type of crime going to take place at given co-ordinates
WeekDay
PDdistrict
X co-ordinates
Y co-ordinates
PREDICT CRIME
There's a probability of OTHER OFFENSES at given place
Type here to search
ENG
17:18
02-05-2021

5.Conclusion and Future Scope

Conclusion

The initial problem of classifying 6 different crime categories was a challenging multi-class classification problem, and there was not enough predictability in our initial data-set to obtain very high accuracy on it. We found that a more meaningful approach was to collapse the crime categories into fewer, larger groups, in order to find structure in the data. We got high accuracy and precision on Prediction. However, the Violent/Non-violent crime classification did not yield remarkable results with the same classifiers – this was a significantly harder classification problem. Thus, collapsing crime categories is not an obvious task and requires careful choice and consideration.

Possible avenues through which to extend this work include time-series modeling of the data to understand temporal correlations in it, which can then be used to predict surges in different categories of crime. It would also be interesting to explore relationships between surges in different categories of crimes – for example, it could be the case that two or more classes of crimes surge and sink together, which would be an interesting relationship to uncover. Other areas to work on include implementing a more accurate multi-class classifier, and exploring better ways to visualize our results.

Future Scope :

we can also predict the estimated time for the crime to take place as a future scope. Along with this, one can try to predict the location of the crime. We will test the accuracy of frequent-itemsets and prediction based on different test sets. So the system will automatically learn the changing patterns in crime by examining the crime patterns. Also the crime factors change over time. By shifting through the crime data we have to identify new factors that lead to crime. Since we are considering only some limited factors full accuracy cannot be achieved. For getting better results in prediction we have to find more crime attributes. Our software predicts the crime an individual criminal is likely to perform.

6. Reference

- https://www.kaggle.com/roshansharma/sanfrancisco-crime-analysis
- https://www.irjet.net/archives/V7/i6/IRJET-V7I624.pdf
- "Datamining Techniques to Analyze and Predict Crimes" S.Yamuna, N.SudhaBhuvaneswari 1 M.Phil (CS) Research Scholar School of IT Science, Dr.G.R.Damodaran College of science Coimbatore 2 Associate Professor MCA, Mphil (cs), (PhD) School of IT Science