**CHAPTER 1**

**INTRODUCTION**

**1.1 OUTLINE OF THE PROJECT**

The concept of a smart city has been derived as one of the means to improve the lives of the people living within the city by taking smart initiatives in a variety of domains like urban development, safety, energy and so on. One of the factors that determine the quality of life in the city is the crime rate therein. Although there could be a lot of technological advancement in the city but the basic requirement of citizens’ safety still remains. Crime continues to be a threat to us and our society and demands serious consideration if we hope to reduce the onset or the repercussions caused by it. Hundreds of crimes are recorded daily by the data officers working working alongside the law enforcement authorities throughout the United States. Many cities in the United States have signed the Open Data initiative, thereby making this crime data, among other types of data, accessible to the general public. The intention behind this initiative is increasing the citizens’ participation in decision making by utilizing this data to uncover interesting and useful facts. The city of Chicago is one amongst the many to have joined this Open Data movement. The data scientists and engineers working alongside the Chicago Police Department (CPD) have recorded over 100,000 crime cases in the form on police complaints they have received. With the help of this historical data, many patterns can be uncovered. This would help us predict the crimes that may happen in the future and thereby help the city police better safeguard the population of the city.

* 1. **MOTIVATION**

The motivation behind taking up this topic for the research is that every aware citizen in today’s modern world wants to live in a safe environment and neighborhood. However it is a known fact that crime in some form, exists in our society. Although we cannot control what goes on around us, we can definitely try to take a few steps to aid the government and police authorities in trying to control it. The CPD has made the police complaints data from the year 2003 to 2018 (current year) available to the general public. Hence, taking inspiration from the facts stated above, we decided to process this data provided and analyze it to identify the trends in crime over the years as well as make an attempt to predict the crimes in the future.

**1.3 PROBLEM FORMULATION**

The problem being tackled in this research can be best explained in two distinct parts. Performing exploratory analysis of the data to mine patterns in crime. The first step in determining the safety within different areas in the city is analyzing the spread and impact of the crime. We utilize this provided crime dataset by the CPD and perform exploratory analysis on it, to observe existing patterns in the crime throughout the city of Chicago. We study the crime spread in the city based on the geographical location of each crime, the possible areas of victimization on the streets, seasonal changes in the crime rate and the type, and the hourly variations in crime.

Building a prediction model to predict the type of crime that can take place in 2 the city, in the future. After observing the patterns of crime from the historical data as explained previous, the next thing is to predict the crimes that can occur in the future. ∙ Our goal is to build a prediction model that treats this problem as a multiclass classification problem, by classifying the unseen data into one of the crime categories (classes) thereby predicting the crime that can occur. ∙ This is expected to help the police plan their patrol and effectively contribute to building a smarter city. For the first part, we will make use of various data analytics tools along with Spark for initial data preprocessing, to analyze the spread of the crime in the city. For the second part, in order to build a prediction model, we build upon the existing research work and improve their results by experimenting with different types of algorithms. Summarizing, we present the experimental results using graphs and statistics.

**Chapter 2**

**Literature survey**

**An active solution for crime investigation**

Uddin, Osemengbe O., P. S. O. Uddin

This paper focuses on the design a framework that would be automated to trigger alarm for timely solution for prevention, arrest and investigation of crimes and provides a comparision between the predicted analysis

**Crime prediction based on crime types and using spatial and temporal criminal hotspots**

Tahani Almanie,Rsha Mirza and Elizabeth Lor

This paper focuses on finding spatial and temporal criminal hotspots. It analyses two different real-world crimes data sets for Denver, CO and Los Angeles, CA and provides a comparison between the two datasets

**Using Machine Learning Algorithms to Analyze Crime Data**

NatarajanMeganathan

Data mining and machine learning have become a vital part of crime detection and prevention. In this research, we use WEKA, an open source data mining software, to conduct a comparative study between the violent crime patterns from the Communities and Crime Unnormalized Dataset provided by the University of California-Irvine repository and actual crime statistical data for the state of Mississippi that has been provided by neighborhoodscout.com. We implemented the Linear Regression, Additive Regression, and Decision Stump algorithms using the same finite set of features, on the Communities and Crime Dataset. Overall, the linear regression algorithm performed the best among the three selected algorithms. The scope of this project is to prove how effective and accurate the machine learning algorithms used in data mining analysis can be at predicting violent crime patterns.

**An intelligent analysis of a city crime data using data mining**

Malathi.ADr. S. Santhosh BabooAnbarsi.A

Through this the clustering/classification based model to anticipate crime trends. Visual and intuitive criminal and intelligence investigation technics can be developed for crime patterns.

**A proposed data mining profiler model to fight security threats in Nigeria**

Jackson AkpojaroPrincewill AigbeUgochukwu Onwude belu

Crimes cannot be solved by a mere workout, and large data is to be guided by a good mathematical framework. Mining techniques such as pattern recognition, machine learning, artificial intelligence, genetic algorithm, linear regression and more can be well used. To present a good architectural frame work of data, to analysis in a fast way, and to impact the possibility of crimes.

**2.2 SUPERVISED LEARNING**

Supervised learning is the most common form of machine learning scheme used in solving the engineering problems. It can be thought as the most appropriate way of mapping a set of input variables with a set of output variables. The system learns to infer a function from a collection of labeled training data. The training dataset contains a set of input features and several instance values for respective features. The predictive performance accuracy of a machine learning algorithm depends on the supervised learning scheme. The aim of the inferred function may be to solve a regression or classification problem. There are several metrics used in the measurement of the learning task like accuracy, sensitivity, specificity, kappa value, area under the curve etc. In this work, the aim is to classify the patients as healthy or ill based on the past medical records. Before solving any engineering problem, it is vital that it is necessary to choose a suitable algorithm for the training purpose based on the type of the data. The selection of a method depends primarily on the type of the data as the field of machine learning is data driven. The next important aspect is the optimization of the chosen machine learning algorithms.

**2.3** **Classification task**

Classification task is a classical problem in the field of data mining which deals with assigning a pre-specified class to an unknown data. A learning model is built based on the relationship between the predictor attribute values and the value of the target . The challenge is to correctly predict the class based on learning of past data. In machine learning, this kind of classification problems are referred to as supervised learning. Hence, we need to provide a data set containing instances with known classes and a test data set for which the class has to be determined. The success of the classification ability largely depends on the quality of data provided for learning and also the type of machine learning algorithm used. For example, the classification techniques can be used to predict the fraud customers in a bank who apply for a loan or classify mangoes whether they are good or bad and lots of other real time applications. The most common type of classification problem is binary classification, where the target has two possible values like good or bad, yes or no etc. There are several methods for measuring the classification performance like confusion matrix, lift curve, receiver operator characteristics etc.

**2.4. Optimization**

Every machine learning algorithm has a specific technique of learning and is based on the values of their parameters. When an algorithm is applied to solve a classification problem with a different set of parameters, the classification accuracy also differs abruptly in each case. The challenge in machine learning to find the most suitable parameter values of the algorithms that solves an engineering problem to the best possible way in terms of performance metrics. Therefore, one has to fine tune the algorithm parameters that best suits the problem. There are several optimization techniques like genetic algorithm, particle swarm optimization, Tabu search methods etc. The focus of the study is to calibrate the algorithm parameters using design of experiment method.

**CHAPTER 3**

**AIM AND SCOPE OF THE PROJECT**

**3.1 AIM**

In the contemporary world we know that technology takes its lion’s share in every available sector. One of them that should be considered in the box is crime rate that is being increased in many countries. Based on this the city of Chicago has taken a step forward and it open sources the crime data that is snowballing from the last three decades. So now, we are going to consider the data and the research carried out on it to reduce the crime rate. As per the data that is given by the city of Chicago organization many of the data scientists have worked and are constantly thriving out to build a better block of analytics on the crime structure and are inventing best strategies to make use of the available data in the best ever possible way. Let us consider all the research and summarize the best approach to reduce or envisage the crime scenarios in the city. This survey consists of different technologies like data mining, machine learning and its role on crime applications.

**3.2 EXISTING SYSTEM**

* For the accessing of any information, one has to dig up piles of data manually.
* More time is needed in searching the required information.
* Bulks of hard copies are required to be referred.
* Access is not available anywhere.
* The increasing crime in Chicago is a big issue for us. To solve this issue, various systems are developed but they cannot find the area where the crime will happen.
* But it is not possible to keep all the records in memory or in file because of large data.
* The large amount data maintenance is the big problem is searching and predicting a particular data manually is not possible from the files.

**3.3 DISADVANTAGES**

* It'll take long time to get the perfect information as one has to analyze all the information.
* As large amount of data is needed maintenance is difficult.
* To collect the perfect data, there are many restrictions to collect the required data.
* One has to be very keen towards the data as to visualize all the data have to be perfect.

**3.4 MODEL ARCHITECTURE**

Machine learning algorithm statistical models

Exploratory Data Analysis

Clean Data

Data is Processed

Raw Data is Collected

Real World

Communicate visualizations report findings

Build Data Product

**Fig 3.4: system architecture**

**3.5 MODULES INVOLVED**

1. Collecting Data set -

Getting a large data set from which we have to predict and check its accuracy level

1. Removing of Duplicates.

As it is a large data set there will be more duplicates where it is of no use and it occupies more space, so this data must be removed or eliminated to get the analyse and prediction in good accurate level.

1. Analysis of the Data set.

The data set should be well analysed in such a way that it should be well understood by the analyst and the person trying to understand. So the analysit have to analyse the data and represent it in undestandable manner.

1. Location using Coordinates.

As it is easy to identify a place where the crime had happened by locating the places in map, so using folium we will generate the place using coordinates

1. Predicting the data and Accuracy level.

Now the data which we do not know is predicted and well analysed, also the accuracy level is tested to show that what we analysed or predicted is good and improvised compared to the previous model.

**3.6 Predictive Analytics**

Predictive Analytics Predictive analytics is the technique of analyzing the past or historical data in order to predict the future outcome. It is different from data mining. As explained in the Figure, predictive analysis starts by capturing relationships between the different variables in the data. After that, hypothesis is developed based on these results. Following this, based on the outcome of the previous steps, a model is built in order to test this hypothesis Figure: Data mining and predictive analytics.

There are numerous advantages of using Predictive Analytics in general. For example, an organization can study it’s internal data to identify trends in profit, so that they can adopt the necessary steps to possibly replicate that in the future. It is also a useful technique for professionals in the marketing industry as it can help decide which campaign successfully generated revenue and business. For the purpose of this research, Predictive Analytics is helpful for the following reasons:

∙ It will help us identify the progression in crime throughout the years 4

∙ It will help us closely observe the variables having highest correlation with the predictor or target variable

∙ Visualizing the data can even help map potential outliers, which can then be effectively handled during data preprocessing

∙ The analysis can bring up some interesting facts from the past which might prove to be useful to the CPD while planning their patrol or strategies against crime

**Chapter 4**

**ANALYSIS OF DATA SET AND ALGORITHMS USED**

**4.1 PROPOSED SYSTEM**

In the proposed system, we are introducing the application which will predict the crime that criminals can do in the future and also we can predict the crimal using the previous data on clues.

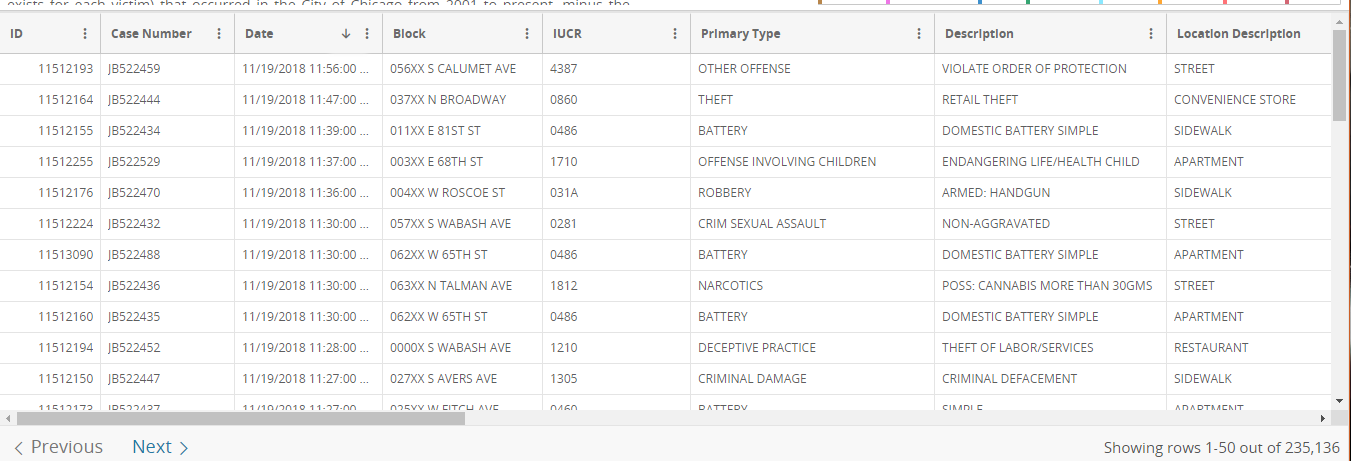
This prediction is based on attributes like criminal record, education, occupation, friend circle, family background and other various factors.

We will store all the previous record of criminal and by mining his previous record we can calculate the possibility and prediction of the crime he is likely to perform.

**4.2 DATA COLLECTION**

The data that is going to be collected is the data that is officially being posted by the chicago police officers and that will be updated for every week with the crimes that occured for that particular week. In this way in the website <https://data.cityofchicago.org/Public-Safety/Crimes-2001-present/ijzp-q8t2/data> the data will be present past 10 years.

As the data is huge at it will take lot of time to load, so we only took the data from 2012, as the crimes stated increasing rapidly from that year and from that data we predicted the data in the best accurate way.



Generating information from official chicago crime data website

**4.1 DATA SET ANALYSIS**

***4.1.1 Description of columns in the Dataset***

***ID***- Unique identifier for the record.

***Case Number*** - The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.

***Date*** - Date when the incident occurred. this is sometimes a best estimate.

***Block*** - The partially redacted address where the incident occurred, placing it on the same block as the actual address.

***IUCR*** - The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e.

***Primary Type*** - The primary description of the IUCR code.

***Description*** - The secondary description of the IUCR code, a subcategory of the primary description.

***Location Description*** - Description of the location where the incident occurred.

***Arrest*** - Indicates whether an arrest was made.

***Domestic*** - Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.

***Beat*** - Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74.

***District***- Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.

***Ward*** - The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.

***Community Area*** - Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6.

***FBI Code*** - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap\_crime\_sums/crime\_types.html.

***X Coordinate*** - The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.

***Y Coordinate*** - The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.

***Year*** - Year the incident occurred.

***Updated On*** - Date and time the record was last updated.

***Latitude*** - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

***Longitude*** - The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

***Location*** - The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

**4.2 ROAD MAP**

* Load the dataset.
* Understand the variables.
* Explore the dataset: - Data type conversions if needed, Fill in the missing values.
* Statistical analysis: - Descriptive and Inferential analysis (If needed).
* Exploratory data analysis: - Mostly focusing on Data visualization (Univariate, Bivariate and multivariate analysis).
* Understanding the EDA and Statistical analysis from which a lot of insights are possible but choosing the right insights for research purpose is the key.
* Feature Engineering if possible

***4.2.1 Advance analytics***

Supervise Methods:- Predictive models

Unsupervise Methods :- Cluster analysis

**4.3 IMPORT NECESSARY PACKAGES**

import pandas as pd # Data manipulation and analysis pacakge

import numpy as np *# Math package*

import matplotlib.pyplot as plt *# Visualization package*

import seaborn as sns *# Advance Stats viz package build top on matplotlib*

import plotly.offline as pyo *# Open source Interactive viz package*

import plotly.figure\_factory as ff *#wrapper functions that create unique charts*

*Note*: Before importing ,make sure that the packages are installed in the Environment to avoid any errors.

*To install packages :*

Go to Command prompt and type :pip3 install “package name” (ex: pip3 install plotly)

***4.3.1 Load the dataset***

* df = pd.read\_csv('E://Chicago\_Crime/Chicago\_Crimes\_2012\_to\_2018.csv')

Reads the Chicago Crime dataset which is originally a csv (comma separated value) as pandas dataframe.

*Note:*Make sure the path of .csv file is correct to avoid errors . OR We can also try a small hack - Copy the dataset in the working directory/folder and simply use the command “pd.read\_csv(‘./Chicago\_Crimes\_2012\_to\_2018.csv')”

## *4.3.2 A sneak peek into the dataset*

## df.shape

Displays the size of the data frame(1456714, 23):

The given data frame contains more than 14.5 million records/observations with 23 columns/variables

## df.columns

## Displays what are the columns present in the dataset.

## df.head(3)

## Displays top 3 records in the dataset. If not specified any number inside the head() function, by default it displays top 5 records.

## *4.3.3 Remove Duplicate data in the dataset*

## Remove the duplicate observations if any. In the given crime dataset,

## “Case Number” should be unique.

## df = df.drop\_duplicates(subset='Case Number')

## If multiple observation are found with same case number , they will be deleted.

## *4.3.4 Observe data type and other details of variables*

* df.info()

Gives the details of all variables in the dataset like datatype, number of records under each column.. etc. Here we can observe if there are missing records.

* df.isnull().sum()

Gives the total number of missing records under each column/variable.

* (df.isnull().sum()/df.shape[0])\*100

We can also observe what is percentage of missing values in each column.

* df.dropna(inplace=True)
* df.shape

we can drop/delete those observation/rows which are having some missing values.

**4.4 DESCRIPTIVE ANALYSIS**

As the most of the variables are categorical type, we do descriptive analysis in line with that.

* + def cat\_des(data):
  + print('Categorical Descriptive analysis on {}'.format(data.name))
  + print('Frequency')
  + print(data.value\_counts())
  + print(' -------------------------------------------------------------------------‘)
  + print('Proportion')
  + print((data.value\_counts()/data.shape[0])\*100)
  + print(' ------------------------------------------------------------------------‘)

*Note:*Make sure about indentation (No brackets are provided for the function block in Python) i.e the function in python will be defined using a keyword “def”. Maintain some space before the code under the function which helps Python to understand the code written after function definition is the part of the function. This is called indentation.

* cat\_des(df['Arrest'])
* cat\_des(df['Primary Type'])
* cat\_des(df['FBI Code'])
* cat\_des(df['Year'])
* cat\_dec(df[‘Location Description’])

We are calling the earlier function on various variables to know the frequency and proportion of values under the variables.

FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap\_crime\_sums/crime\_types.html.

#### 4.4.1 Analysis that can be drawn from the above analysis

***Year:*** On going through observation, the percentage of crime rate has been coming down year by year. Especially after 2016, it came down drastically means the Chicago Police department might have come up with some new policies and strict laws.

***Primary Type:*** Most of the crimes are related to ‘Theft’.

***FBI Code:*** Under crime category 6, the crime cases are reported more.

***Location Description*:** Most of the crimes are happened in the Streets indicates there could be frequent mob activities.

***Domestic:*** Most of the crimes are done by non-residents only.

***Arrest:*** No arrests are taken place during most of the crimes.

***District:*** District 11 is very prone to crimes.

##### Things which i want to know from Univariate analysis*:* From block i want to know which area had been frequently reported.

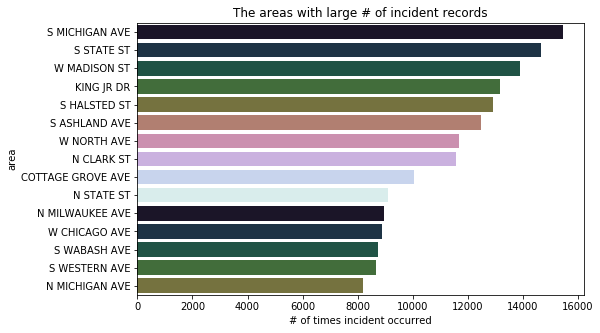
* Area = []
* def block(data):
* for i in data:
* Area.append(i.split(' ')[-3:])
* block(data = df['Block'])
* area\_st = []
* for area in Area:
* area\_st.append(' '.join(area))
* df['area\_st'] = area\_st
* print(df['Block'].head())
* print("\n\n After discarding prefix of the block name\n\n")
* print(df['area\_st'].head())
* areas = df['area\_st'].value\_counts()
* print(areas[1:5])
* areas = areas.reset\_index()
* areas.rename(columns={'index':'area','area\_st':'# of times incident occurred'},inplace=True)
* print(areas[1:5])

The above code will give the number of crimes/incidents happened in each area.

We can also present these details visually using different plots.

***4.4.1.1 Bar Plots on Areas***

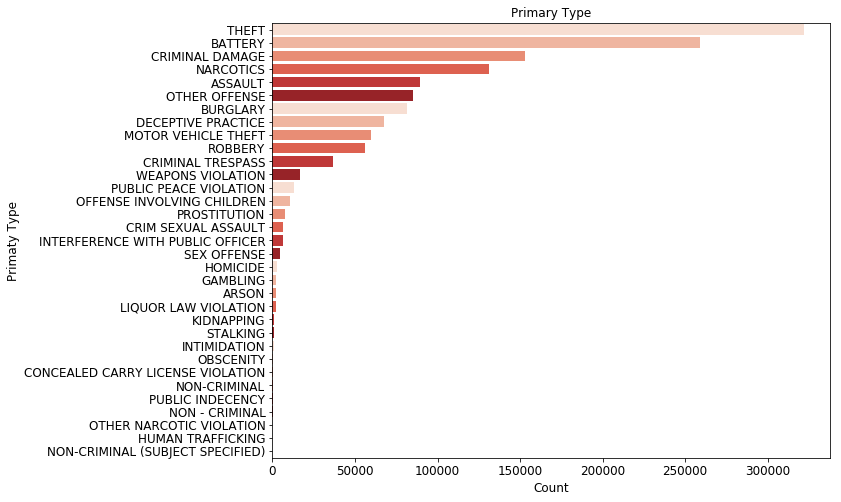
* areas = areas[0:15]
* plt.figure(figsize=(8,5))
* cl = sns.color\_palette("cubehelix",10)
* sns.barplot(x ='# of times incident occurred',y='area',data=areas, palette=cl )
* plt.rcParams.update({'font.size': 10})
* plt.title('The areas with large # of incident records')



***Bar Plots regarding areas***

***4.4.1.2 Bar plots on Primary Type***

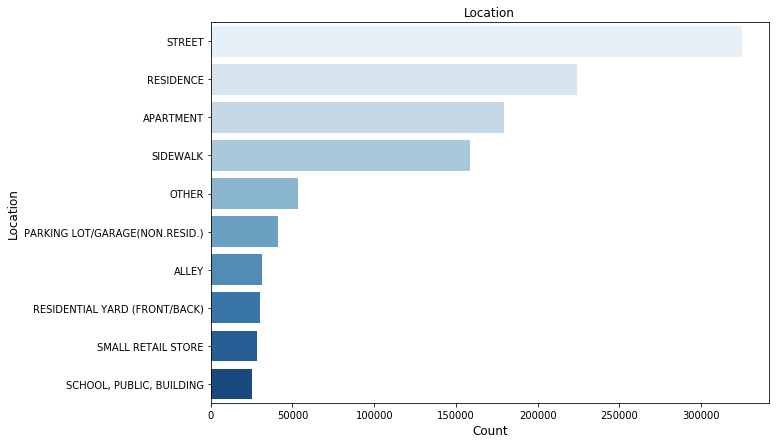
* plt.figure(figsize=(10,8))
* palette = sns.color\_palette("Reds")
* #sns.set\_palette(palette)
* sns.countplot(y='Primary Type',data=df, palette=palette,order=df['Primary Type'].value\_counts().index )
* #plt.rcParams.update({'font.size': 12})
* plt.rc('axes', titlesize=12)
* plt.xlabel( 'Count',fontsize=12)
* plt.ylabel( 'Primaty Type',fontsize=12)
* plt.title('Primary Type')



***Bar Plots regarding primary type***

***4.4.1.3 Bar plots on Location :***

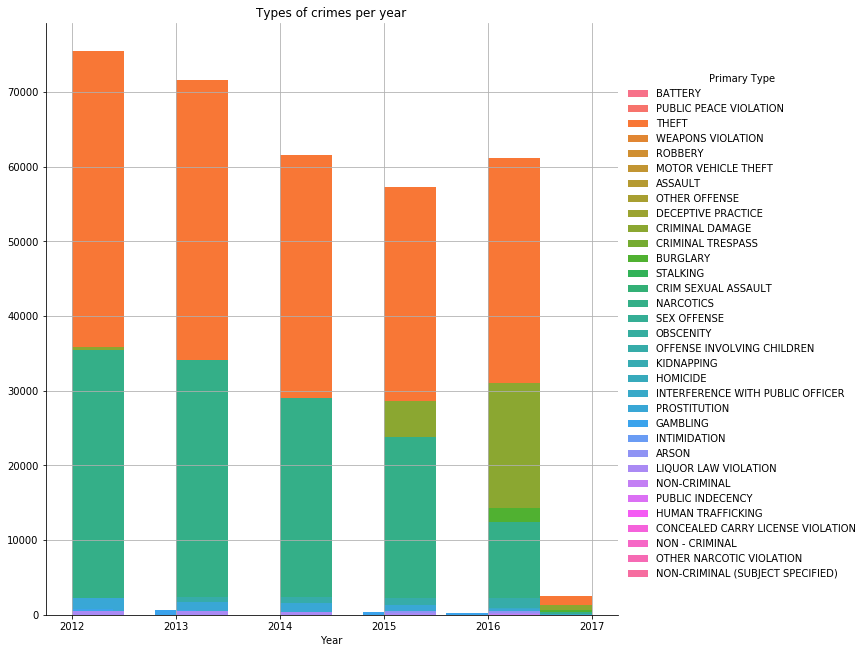
* location = df['Location Description'].value\_counts().reset\_index()
* location = location[0:10]
* location = location.rename(columns={'index':'Location','Location Description':'Freq'})
* location
* plt.figure(figsize = (10, 7))
* sns.barplot(y ="Location", x = "Freq", data = location, palette="Blues")
* plt.xlabel( 'Count',fontsize=12)
* plt.ylabel( 'Location',fontsize=12)
* plt.title('Location')



***Bar Plots regarding location***

***4.4.1.4 Yearly Distribution of Various Crimes***

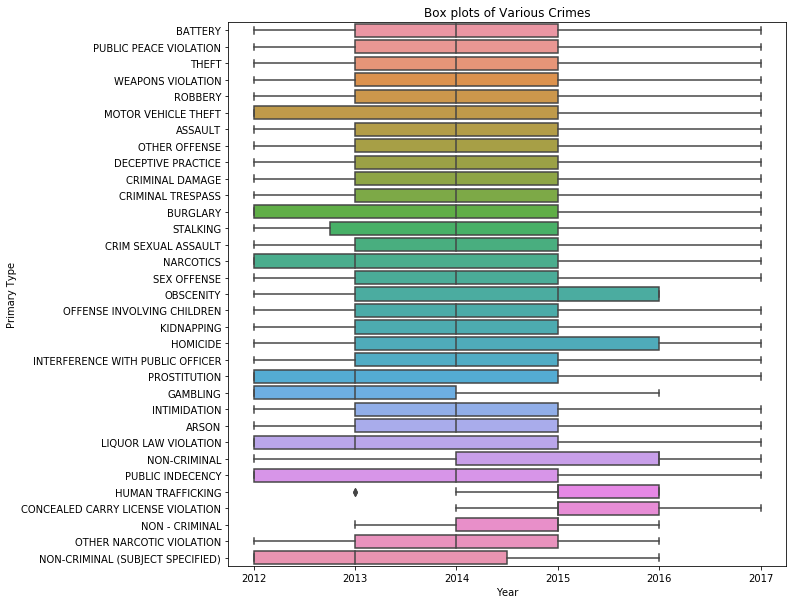
* plt.figure(2,figsize=(10,5))
* sns.FacetGrid(df,hue='PrimaryType',height=9).map(plt.hist,'Year').add\_legend()
* plt.xticks(u)
* plt.title("Types of crimes per year")
* plt.grid()
* plt.show()



***Grid view of Various Crimes Per Year***

***4.4.1.5 Box Plot on Various Crimes***

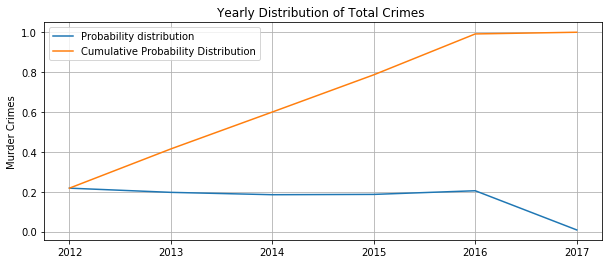
* plt1=plt
* plt1.figure(3,figsize=(10,10))
* sns.boxplot(x='Year', y='Primary Type', data=df)
* plt.title("Box plots of Various Crimes")
* plt1.xticks(u)
* plt1.show()



***Box Plot On Various Crimes***

***4.4.1.6 Yearly Distribution total Crimes***

* murder=df['Year'][df['Primary Type']=="ASSAULT"]
* u,c=np.unique(murder,return\_counts=True)
* unique,counts=np.unique(df['Primary Type'],return\_counts=True)
* plt.figure(1,figsize=(10,4))
* pdf=c/np.sum(c)
* cdf=np.cumsum(pdf)
* plt.title("Yearly Distribution of Total Crimes ")
* plt.plot(u,pdf,label='Probability distribution')
* plt.plot(u,cdf,label="Cumulative Probability Distribution")
* plt.ylabel("Murder Crimes")
* plt.grid()
* plt.xticks(u)
* plt.legend()
* plt.show()



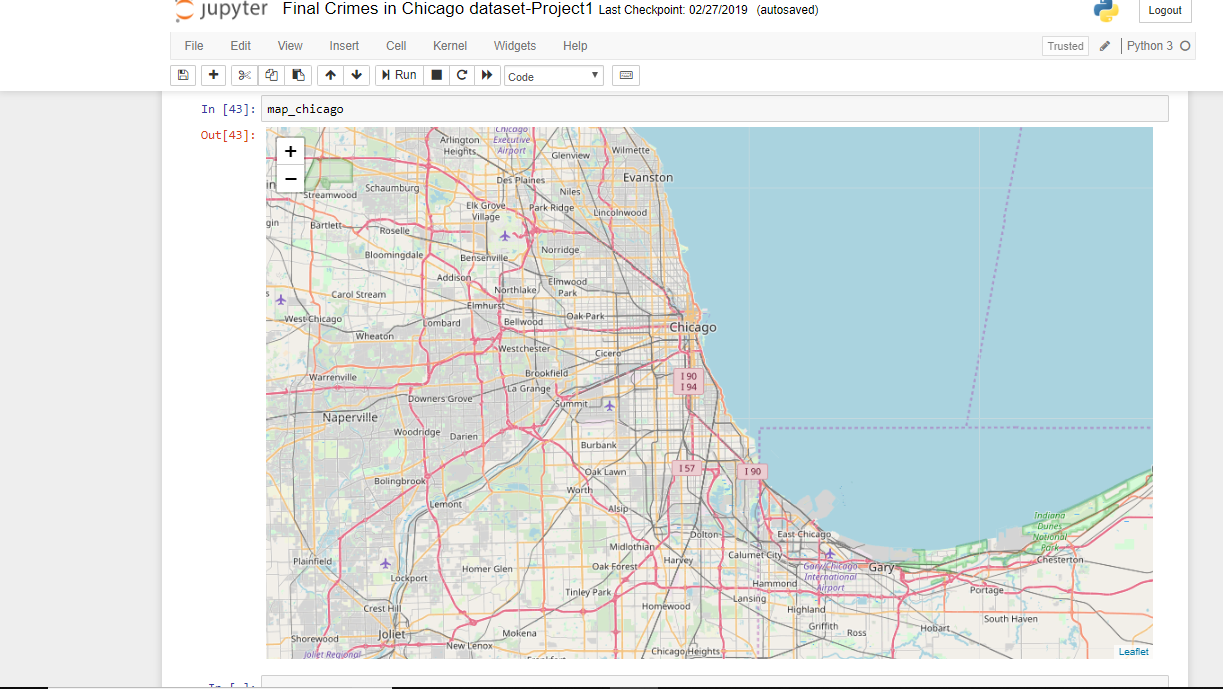
**Yearly Distibution Of Crimes**

**4.5 BUILDING MAPS**

Folium is a powerful Python library that helps you create several types of Leaflet maps. The fact that the Folium results are interactive makes this library very useful for Viz

# The latitude of Chicago, IL, USA is 41.881832, and the longitude is -87.623177

* import folium
* map\_chicago = folium.Map(location=[41.8,-87.6],zoom\_control=12) # tiles = "Stamen Terrain"
* map\_chicago



Locating using co-ordinates

**4.6 ALGORITHM USED**

MULTINOMIAL NAIVE BAYES

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

Incremental fit on a batch of samples.

This method is expected to be called several times consecutively on different chunks of a dataset so as to implement out-of-core or online learning.

This is especially useful when the whole dataset is too big to fit in memory at once.

This method has some performance overhead hence it is better to call partial\_fit on chunks of data that are as large as possible (as long as fitting in the memory budget) to hide the overhead.

**CHAPTER 5**

**RESULT AND DISCUSSION**

**5.1 UNSUPERVISED LEARNING and VISUALIZATION**

As part of this we do data mining of areas where crimes has been occurred. We take the help of WordCloud for this task.

* area\_cr=[]
* for i,j in enumerate(df['area\_st']):
* area\_cr.append(j)
* from wordcloud import WordCloud

import necessary packages. If not already installed, go to cmd and type

-pip3 install wordcloud for the package installation.

Now Generate the WordCloud of areas of crimes.

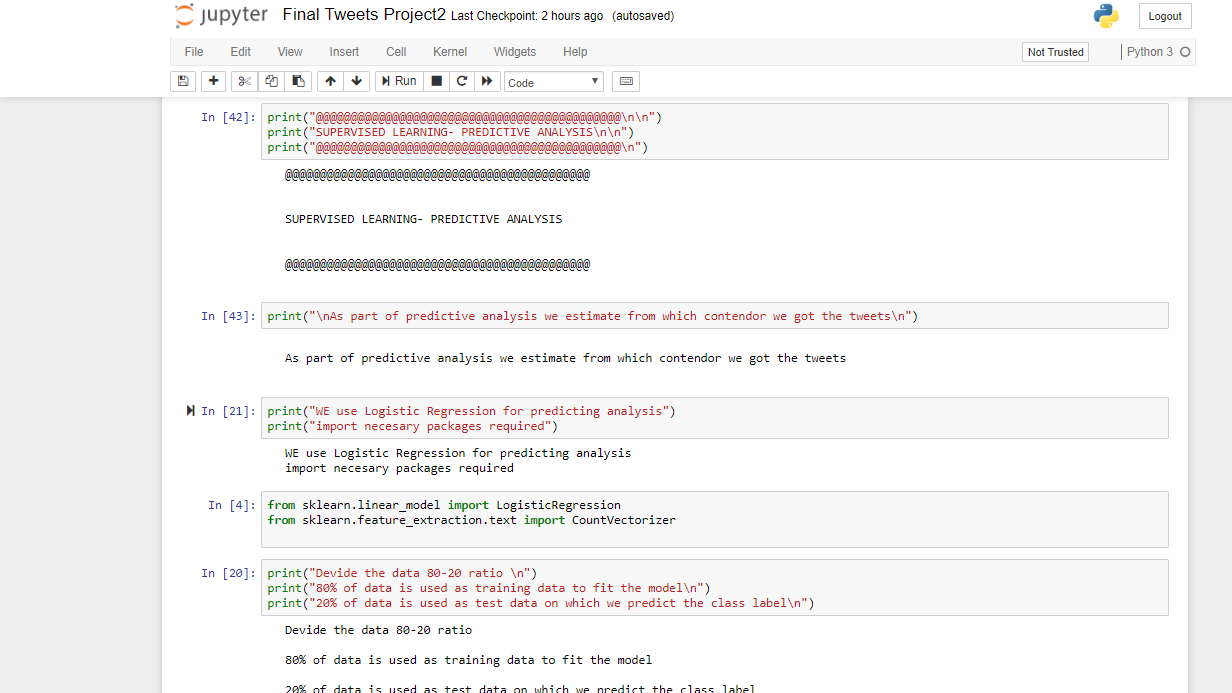


WordCloud of various cities regarding crimes

As per above visualization it seems like, S MICHIGAN AVE is top on the row by virtue of crime rate Most of the crimes are happening in that place, So special care to be taken in S MICHIGAN AVE. Finally we analysed that the areas like 'S MICHIGAN AVE' are very prone to crimes.

**5.2 SUPERVISED LEARNING- PREDICTIVE ANALYSIS**

As part of predictive analysis, if the inputs of crime happened are given, we predict what might be the district in which crime has happened. As there are more than 2 districts in the dataset , we used Multinomial Naive Bayes for predictive analysis. Import the necessary package required for this task.



Dividing the data into 80-20 ratio

* from sklearn.naive\_bayes import MultinomialNB

Divide the data 80-20 ratio .80% of data is used as training data to fit the model. 20% of data is used as test data on which we predict the class label.

* Y\_data=df['District']
* Y\_train=Y\_data[0:1135000]
* Y\_test=Y\_data[1135000:]

We devided the class lables into train and test.

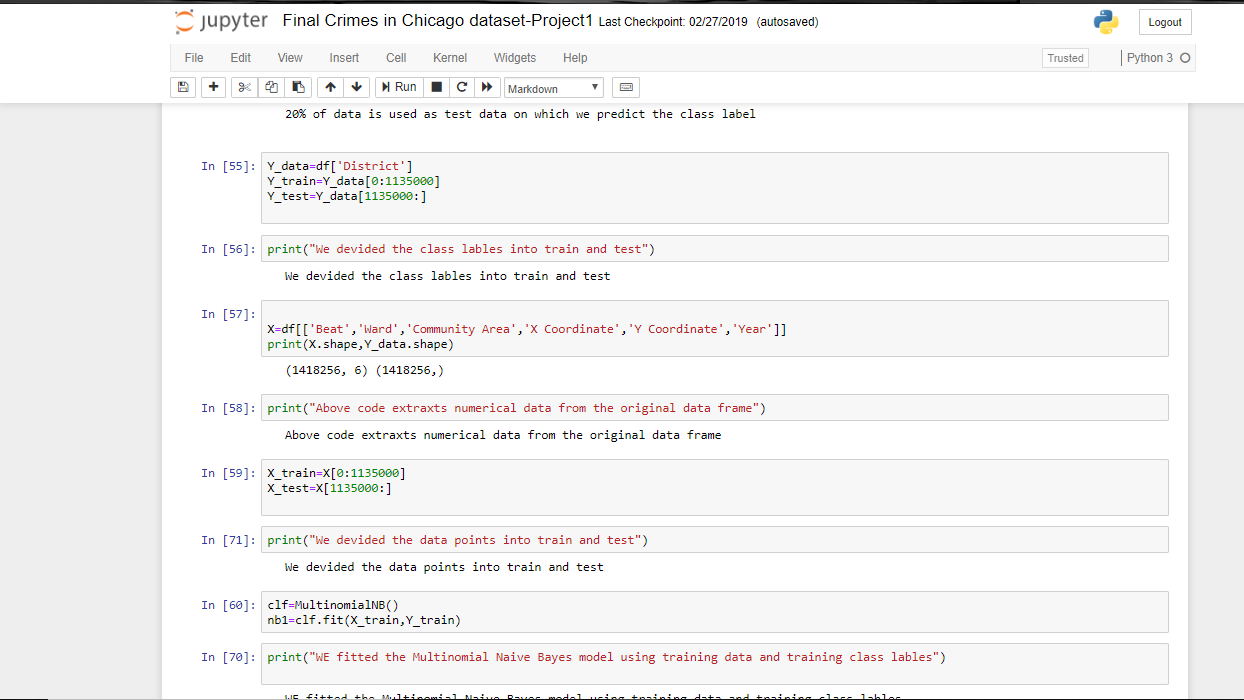
* X=df[['Beat','Ward','Community Area','X Coordinate','Y Coordinate','Year']]
* print(X.shape,Y\_data.shape)

Above code extracts numerical data from the original data frame

* X\_train=X[0:1135000]
* X\_test=X[1135000:]

We devided the data points into train and test

* clf=MultinomialNB()
* nb1=clf.fit(X\_train,Y\_train)



Dividing the data into train data and test data

We fitted the Multinomial Naive Bayes model using training data and training class labels as per the above code.

* Y1\_pred=nb1.predict(X\_test)

Above code is to predict the labels from test data.

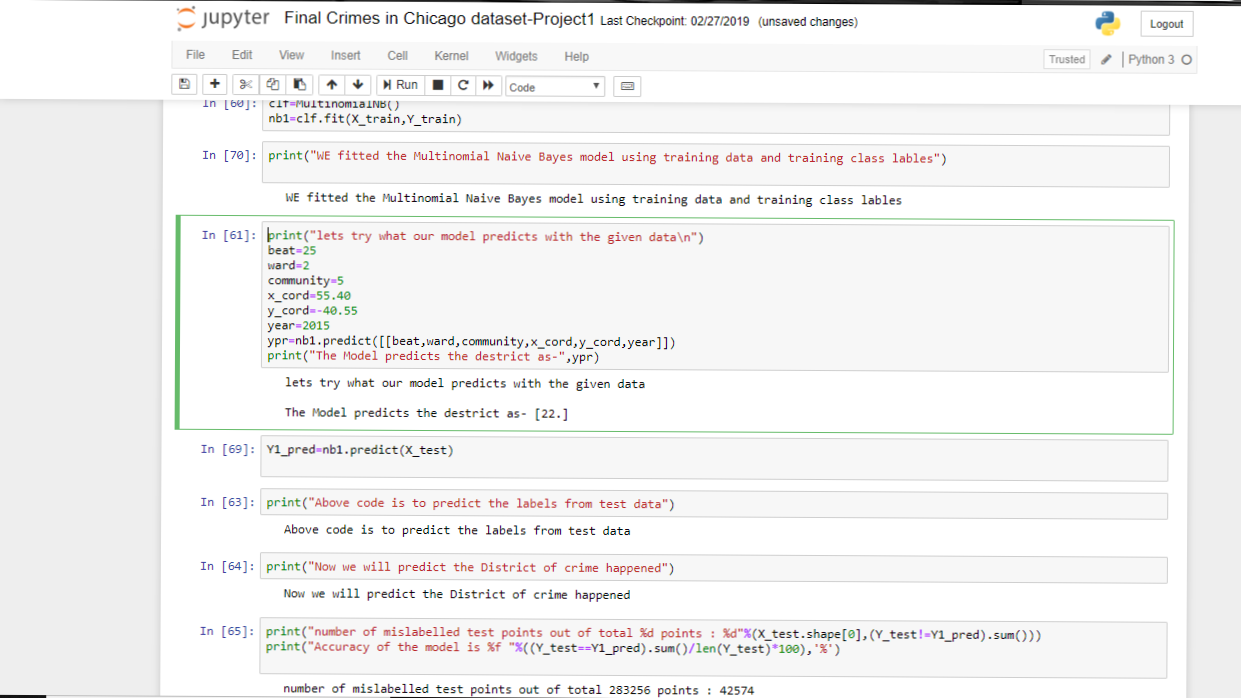
**5.3 PERFORMANCE ANALYSIS**

Now we will predict the number of mis-predictions and accuracy as follows:

* print("number of mislabelled test points out of total %d points : %d"%(X\_test.shape[0],(Y\_test!=Y1\_pred).sum()))
* print("Accuracy of the model is %f "%((Y\_test==Y1\_pred).sum()/len(Y\_test)\*100),'%')

Number of mis-predicted test points out of total 283256 points : 42574 Accuracy of the model is 84.969780 %. From the above , we can analyse that around 15% of crimes are wrongly judged in view of destricts. As a whole the model has predicted with 85% accuracy

* print("Some of the wrongly predicted data is \n")
* print(" Actual Destrict ---- Predicted Destrict --- Primary Type of crime is\n\n")



Predicting the data from the data

We can pick up some mis- predicted data points as follows:

* for i in range(0,500):
* if (Y\_test.iloc[i]!=Y1\_pred[i]): # y\_test is dataframe and y\_pred is numpy array
* print(i,np.array(Y\_test)[i],"-----",Y1\_pred[i],"------",df['Primary Type'][1135000:].iloc[i],"\n")

The Predictive Analysis has been done on the Crime data and predicted with good accuracy

* Y1\_pred=nb1.predict(X\_test)

Above code is to predict the labels from test data.

Now we will predict the number of mis-predictions and accuracy as follows:

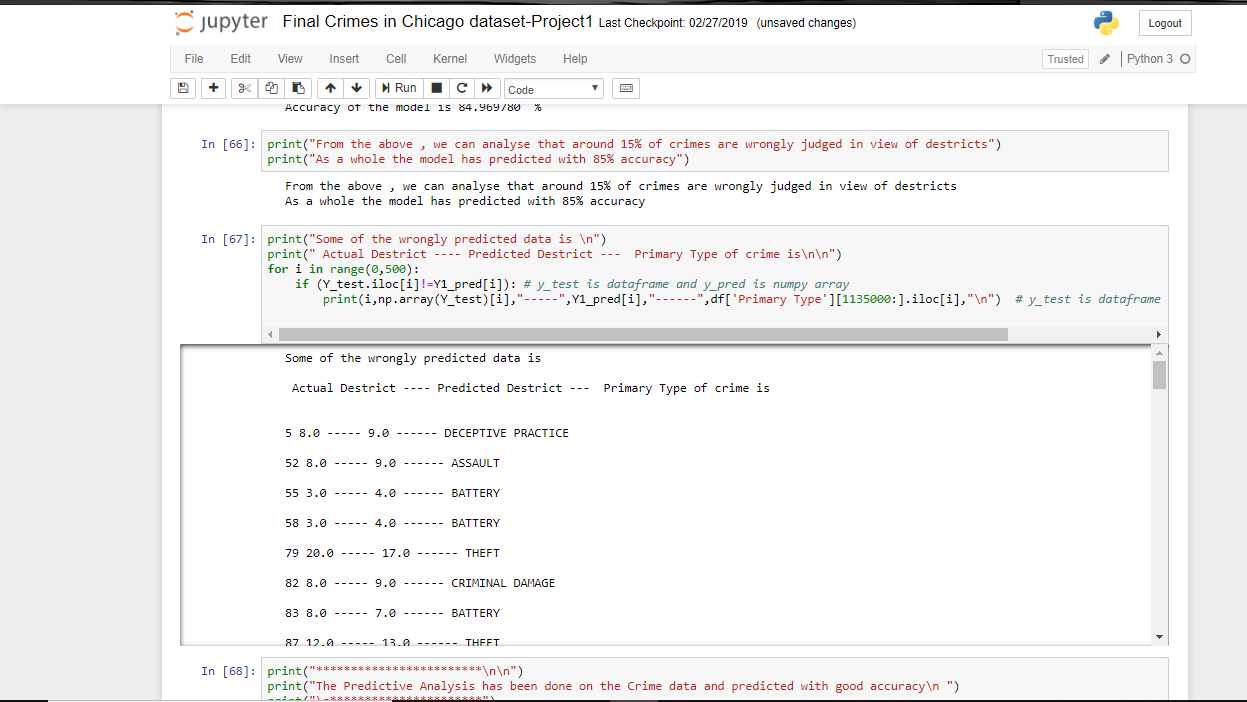
* print("number of mislabelled test points out of total %d points : %d"%(X\_test.shape[0],(Y\_test!=Y1\_pred).sum()))
* print("Accuracy of the model is %f "%((Y\_test==Y1\_pred).sum()/len(Y\_test)\*100),'%')

Number of mis-predicted test points out of total 283256 points : 42574 Accuracy of the model is 84.969780 %. From the above , we can analyse that around 15% of crimes are wrongly judged in view of destricts. As a whole the model has predicted with 85% accuracy

* print("Some of the wrongly predicted data is \n")
* print(" Actual Destrict ---- Predicted Destrict --- Primary Type of crime is\n\n")

We can pick up some mis- predicted data points as follows:

* for i in range(0,500):
* if (Y\_test.iloc[i]!=Y1\_pred[i]): # y\_test is dataframe and y\_pred is numpy array
* print(i,np.array(Y\_test)[i],"-----",Y1\_pred[i],"------",df['Primary Type'][1135000:].iloc[i],"\n")



Some of the wrongly predicted data

The Predictive Analysis has been done on the Crime data and predicted with good accuracy.

**CHAPTER 6**

**CONCLUSION AND FUTURE WORK**

**6.1 CONCLUSION**

Our goal in this project has been to cover techniques and approaches that promise to directly enable opinion-oriented information seeking systems, and to convey to the reader a sense of our excitement about the intellectual richness and breadth of the area. This project suggests the answer that by providing awareness and education.

Moreover, by this we can give a better outlook on the way how we solve a crime analysis and also by this process we have seen the nearest accurate way of analyzing a crime happened

**6.2 FUTURE WORK**

The data set can be changed in various ways and also can be used in a real time scenarios for getting predictios from the data which we have.as we see the performance of the model have been increased from the previous once.

Furthermore the data is directly proportional to the data that we have and that we are going to generate.The more data we have the best we can predict the data ina very accurate manner.

**REFERENCES**

[1] Uddin, Osemengbe O., P. S. O. Uddin, “Data Mining: An Active Solution for Crime Investigation”, in International Journal of Computer Science and Technology, Vol. 5, SPL - 1, Jan - March 2014.[2] Lawrence McClendon and Natarajan Meghanathan, “Using Machine Learning Algorithms To Analyze Crime Data”, Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1, March 2015.[3] Prajakta Yerpude and Vaishnavi Gudur. “ Predictive Modelling Of Crime Dataset Using Data Mining”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.4, July 2017.[4] Malathi. A , Dr. S. Santhosh Baboo and Anbarasi. A,” An intelligent Analysis of a City Crime Data Using Data Mining”, 2011 International Conference on Information and Electronics Engineering IPCSIT vol.6 (2011) © (2011) IACSIT Press, Singapore.Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.4, July 2017.  
[5] Malathi. A , Dr. S. Santhosh Baboo and Anbarasi. A,” An intelligent Analysis of a City Crime Data Using Data Mining”, 2011 International Conference on Information and Electronics Engineering IPCSIT vol.6 (2011) © (2011) IACSIT Press, Singapore.  
[6] Gourav Govindaswamy, Vinod Kumar Kethineni, Santhosh Kumar P, “ A Survey on Crime Data Analysis Using Data Mining Techniques”, Gourav Govindaswamy.et.al. Int. Journal of Engineering Research and Application www.ijera.com ISSN: 2248-9622, Vol. 7,Issue8,(Part-6)August2017,pp.30-34.  
[7]www.internetlivestats.com  
[8] Alberto De Marco, Giulio Mangano, Giovanni Zenezini, “Digital Dashboards for Smart City Governance: A Case Project to Develop an Urban Safety Indicator Model”, Journal of Computer and Communications, 2015, 3, 144-152.  
[9] Tahani Almanie, Rsha Mirza and Elizabeth Lor, “Crime Prediction Based On Crime Types And Using Spatial And Temporal Criminal Hotspots”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.4, July 2015.

**APPENDIX**

1. SAMPLE CODE

Remove Duplicate data in the dataset

## df = df.drop\_duplicates(subset='Case Number')

Descreptive Analysis:

* cat\_des(df['Arrest'])
* cat\_des(df['Primary Type'])
* cat\_des(df['FBI Code'])
* cat\_des(df['Year'])
* cat\_dec(df[‘Location Description’])

*Univariate Analysis*

* Area = []
* def block(data):
* for i in data:
* Area.append(i.split(' ')[-3:])
* block(data = df['Block'])
* area\_st = []
* for area in Area:
* area\_st.append(' '.join(area))
* df['area\_st'] = area\_st
* print(df['Block'].head())
* print("\n\n After discarding prefix of the block name\n\n")
* print(df['area\_st'].head())
* areas = df['area\_st'].value\_counts()
* print(areas[1:5])
* areas = areas.reset\_index()
* areas.rename(columns={'index':'area','area\_st':'# of times incident occurred'},inplace=True)
* print(areas[1:5])

Bar plots on Primary Type

* plt.figure(figsize=(10,8))
* palette = sns.color\_palette("Reds")
* #sns.set\_palette(palette)
* sns.countplot(y='Primary Type',data=df, palette=palette,order=df['Primary Type'].value\_counts().index )
* #plt.rcParams.update({'font.size': 12})
* plt.rc('axes', titlesize=12)
* plt.xlabel( 'Count',fontsize=12)
* plt.ylabel( 'Primaty Type',fontsize=12)
* plt.title('Primary Type')

Bar plots on Location

* location = df['Location Description'].value\_counts().reset\_index()
* location = location[0:10]
* location = location.rename(columns={'index':'Location','Location Description':'Freq'})
* location
* plt.figure(figsize = (10, 7))
* sns.barplot(y ="Location", x = "Freq", data = location, palette="Blues")
* plt.xlabel( 'Count',fontsize=12)
* plt.ylabel( 'Location',fontsize=12)
* plt.title('Location')

Yearly Distribution total Crimes

* murder=df['Year'][df['Primary Type']=="ASSAULT"]
* u,c=np.unique(murder,return\_counts=True)
* unique,counts=np.unique(df['Primary Type'],return\_counts=True)
* plt.figure(1,figsize=(10,4))
* pdf=c/np.sum(c)
* cdf=np.cumsum(pdf)
* plt.title("Yearly Distribution of Total Crimes ")
* plt.plot(u,pdf,label='Probability distribution')
* plt.plot(u,cdf,label="Cumulative Probability Distribution")
* plt.ylabel("Murder Crimes")
* plt.grid()
* plt.xticks(u)
* plt.legend()
* plt.show()

UN SUPERVISED LEARNING

* area\_cr=[]
* for i,j in enumerate(df['area\_st']):
* area\_cr.append(j)
* from wordcloud import WordCloud

SUPERVISED LEARNING- PREDICTIVE ANALYSIS

* Y\_data=df['District']
* Y\_train=Y\_data[0:1135000]
* Y\_test=Y\_data[1135000:]

We devided the class lables into train and test.

* X=df[['Beat','Ward','Community Area','X Coordinate','Y Coordinate','Year']]
* print(X.shape,Y\_data.shape)

Above code extracts numerical data from the original data frame

* X\_train=X[0:1135000]
* X\_test=X[1135000:]

We devided the data points into train and test

* clf=MultinomialNB()
* nb1=clf.fit(X\_train,Y\_train)

WE fitted the Multinomial Naive Bayes model using training data and training class lables as per the above code.

* Y1\_pred=nb1.predict(X\_test)

Above code is to predict the labels from test data.

Now we will predict the number of mis-predictions and accuracy as follows:

* print("number of mislabelled test points out of total %d points : %d"%(X\_test.shape[0],(Y\_test!=Y1\_pred).sum()))
* print("Accuracy of the model is %f "%((Y\_test==Y1\_pred).sum()/len(Y\_test)\*100),'%')

number of mis-predicted test points out of total 283256 points : 42574 Accuracy of the model is 84.969780 %

From the above , we can analyse that around 15% of crimes are wrongly judged in view of destricts

As a whole the model has predicted with 85% accuracy

* print("Some of the wrongly predicted data is \n")
* print(" Actual Destrict ---- Predicted Destrict --- Primary Type of crime is\n\n")

We can pick up some mis- predicted data points as follows:

* for i in range(0,500):
* if (Y\_test.iloc[i]!=Y1\_pred[i]): # y\_test is dataframe and y\_pred is numpy array
* print(i,np.array(Y\_test)[i],"-----",Y1\_pred[i],"------",df['Primary Type'][1135000:].iloc[i],"\n")