Automate Crime News Segmentation, Knowledge-Based Generation and Public Opinion Mining for Social Awareness

by
Ahmed Bin Sabit
19241027
Asif Ahsan
21141081
Zarin Tasnim
21141085

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
June 2021

DECLARATION

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

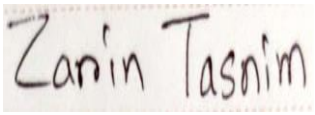
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. I/We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ahmed Bin Sabit
19241027

Asif Ahsan
21141081



Zarin Tasnim
21141085

APPROVAL

The thesis/project titled “Automate Crime News Segmentation, Knowledge-Based Generation and Public Opinion Mining for Social Awareness” submitted by

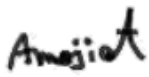
1. Ahmed Bin Sabit (19241027)
2. Asif Ahsan (17301195)
3. Zarin Tasnim (16301179)

of Spring, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on June 02, 2021.

Examining Committee:

Supervisor:

(Member)



Annajiat Alim Rasel
Senior Lecturer
Department of Computer Science and Engineering
BRAC University

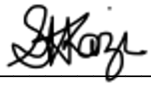
Co-Supervisor:

(Member)

Dr Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University

Departmental Head:

(Chair)



Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

ABSTRACT

In the context of the crime scenario in Bangladesh, people face enormous crime situations which can be eradicated by applying some technological solution. For this a research work needs to be conducted by news segmentation and knowledge base generation. This study aims to determine how authority can target crime scenarios based on public opinion for increasing social awareness. Based on this context there will be an algorithm which will sort out the data sheet as well as the mined data will figure out solutions for respective crime scenarios. These data will be collected from news articles from which a data library will be generated by using natural language processing.

After analyzing all data, it will automatically notify the authority with a precise report. Further research is needed to identify other factors that could strengthen the effectiveness of this report.

Keywords: Natural Language Processing, Machine Learning, Data Mining

DEDICATION

We want to dedicate our work to our parents, without whom we could never come so far in life.

ACKNOWLEDGEMENT

Firstly, all praise to the Great Allah for whom our thesis has been completed without any major interruption.

Secondly, to our supervisor Annajiat Alim Rasel and our co-supervisor Dr. Md. Golam Rabiul Alam for their kind support and advice in our work. He helped us whenever we needed help. And finally to our parents without their thorough support it may not be possible. With their kind support and prayer, we are now on the verge of our graduation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURE

LIST OF ACRONYMS

The next list describes several symbols & abbreviation that will be later used within the body of the document.

BART	Hyper Parameter Optimization [algorithm name of BART]
CNN	Convolutional Neural Network
ETC	Extra Tree Classifier
FIG	Figure
LDA	Linear Discriminant Analysis
LGBM	Light Gradient Boosting Machine
LSTM	Long Short Term Memory
NLP	Natural Language Processing
OCR	Optical Character Recognition
PredPol	Predictive Policing
SVC	Superior vena cava
SVM	Support Vector Machine
UCR	Uniform Crime Report
UNODC	United Nation Office on Drug and Crime
RFC	Random Forest Classifier

GLOSSARY

Thesis:	Automate crime news segmentation, knowledge-based generation and public opinion mining for social awareness.
Glossary:	Natural Language Processing, Machine Learning, Data Mining, Long Short Term Memory, Convolutional Neural Network, Hyper Parameter Optimization, Crime, AI Determined Hotspot, Data Mining, Map Data, Crime Prediction, Data Mining, Open Data, Regression, Decision Trees, Instance Based Learning, Law Enforcement Agency, Achievement Rate of the Investigation, Classification and Regression Problems

Chapter 1

INTRODUCTION

1.1 PROBLEM STATEMENT

If we simplify the problem, “automate crime news segmentation, knowledge-based generation and public opinion mining for social awareness” is the analyzed and detailed information about the current crime scenario in Bangladesh. For progressively exact, why, how, where and when it is happening. This report is important for the authorities like police and other associates to grasp this crime condition furthermore as a summary of the crime scenario and appropriate steps according to the case so they will control crime casualties and eradicate crime situations in a very uniformed way. Additionally, people will get a legit judgment from the government judicial division.

In Bangladesh, crime has been said as a big issue and is an unmistakable reality of latest social orders. The expense of crime fluctuates from an immediate expense to backhanded expenses. Crime as a social issue can limit individuals' opportunities within the network. It can create extensive dread inside the network overriding national security, work, [11] typical cost for basic items, destitution, and wellbeing. Casualties of crime may endure long-term mental injury. The fear rate will block outside speculations even as it ends in the derision of neighborhoods or perhaps whole segments of town. The longing to possess a way of security and secure requests various open strategies and administrative mediations to battle crime. Be that because it may, the difficulty of the crime itself is so far persevering.

Starting by taking a gander at the foremost evident results of the crime, the lawful ones. Under the society's equity framework, the administration has a choice to rebuff people for abusing the law.

This discipline, for the foremost part, comes in one in every of two structures. Less serious is the loss of freedoms or having individual rights disavowed. Essentially, by overstepping the law that has relinquished a little of the privileges of being a Bangladeshi resident.

Bangladesh has been confronting the issues of crime, guiltiness and defilement with its expanded structures, from the earliest starting point of the nation's autonomy. The current examination is an endeavor to portray the far reaching and all-encompassing recorded investigation of crime in contemporary Bangladesh with extraordinary references to crime patterns and their related causations. The investigation is fundamentally spellbinding in nature and the information of the examination has been gathered from different auxiliary sources like police insights, books, diaries, periodicals and related writing. With the end goal of the investigation, the crime information from 1972-2009 has been broken down. The significant discoveries show that the pace of crime has expanded in a consistent manner from 2003-2008 and the most noteworthy number of wrongdoings were submitted in 2008 (1.58 lakh), however; it was diminished a little in 2009. It likewise shows the carried-out crime of the years (2003-2007) where 1.27 lakh crime was submitted in 2003, 1.20 lakh in 2004, 1.26 in 2005, 1.30 in 2006, and 1.57 in 2007. It's a matter of worry that the measurements spoke to just the crime that is being accounted for to the police. The other significant discovery of the investigation states, a large portion of the normal crime are the monetary violations; be that as it may, the patterns of non-financial crime particularly political wrongdoing are alarming. In Bangladesh, to diminish crime the prime helpers like destitution, joblessness, political support, broken home, affectation by companions, and obliviousness must be tended to. Also, different compelling measures including open mindfulness, successful laws and strategy, police change, appropriate equity framework must be guaranteed.

Therefore, there are algorithms which are being utilized to forestall issues in “automate crime news segmentation, knowledge-based generation and public opinion mining for social awareness” like Naive Bayes, [LSTM] 1[2], [CNN] 2[3], [BART] 3[4], SCRAPING [5] and some more. These calculations are effective in some specific fields. Be that as it may, they have a few restrictions as well. Those parts are additionally our anxiety, with the goal that anyone can discover and execute progressively proficient approaches to sort out cloud security issues for a superior and solid web understanding.

Abbreviations: Long Short Term Memory[1], Convolutional Neural Network[2], Hyper Parameter Optimization[3]

1.2 RESEARCH OBJECTIVE

This analysis report and data gives explicit insights concerning crimes which are listed or unlisted in police diaries mainly scrapped from the news as well as submitted by the victims or witnesses. It gathers data about explicit crime in a few distinctive categories. It is data that anybody can get off chance. It gathers data about explicit crime in a few distinctive categories once that they wish. It may not give an asset that audits despise crime measurements in Bangladesh but it will give a clear glance of the current crime scenario and the hotspot of all crime. [2]

The main objective of the research would be building up an implicit model for prediction. For that model, we need planning. Police will not share such complicated data of criminal databases as it will hamper their highly classified intel. Therefore, we need to collect it from news articles and public data inputted in a system. For public data, there has to be a process which validates crime

as if it has occurred or not. This is one of the tough objectives. Then we need to scrap those data organize them and have to use predicting algorithms on it to reach our main objective.

There are some objectives that can be implemented from the database we will collect. This criminal database has “lat long” values. Those values can be used to generate a heat map. Therefore, an AI can determine the most hotspot for crimes. Then that information can be used for public safety. People can be warned as a hot zone for regular crime, maybe they can choose their residence in a safe place and all.

This report is a combination of three different sections. First section is an Algorithm will read newspaper articles every day and among all news it will detect crime news only which is our primary topic for our research work in the second section, another AI algorithm will categories different types of crimes by their lethality, crime context and punishment set by the constitution with a machine learning approach. Thirdly, data mining [6] which is characterized as a procedure used to remove the lacking from the second segment. It infers breaking down information designs in huge groups of information utilizing at least one programming language such as Python. Information mining has applications in numerous fields, similar to science and research. As a use of information mining, organizations can study their clients and grow increasingly viable methodologies identified with different business capacities and thus influence assets in a progressively ideal and astute way. This causes organizations to be nearer to their goal and settle on better choices. Information mining includes successful information assortment and warehousing. For dividing the information and assessing the likelihood of future occasions, information mining utilizes refined algorithms. Information mining is otherwise called knowledge

discovery in data. In addition, this report will help the authority to understand the current crime condition as well as overview of the crime scenario and appropriate steps according to the situation so that they can control crime casualties and eradicate crime situations in a uniformed way.

Keywords: Crime, AI Determined Hotspot, Data Mining, Map Data

Chapter 2

LITERATURE REVIEW

Examining the utilization of automata crime news segmentation public opinion mining for social awareness from Harrendor, S Heiskane, M and Maldy, S [7]. [UNODC[4]] reviews crime patterns and activities in criminal equity frameworks and assembles the principal data on recorded crime just as the assets of the criminal equity framework.

The report attempts to give greater solidity so as to gather a ten years' record of crime.

Firstly, factual information on crime and criminal equity are normally not accessible until after the applicable year.

Country level information on police recorded crime are frequently discharged after the move of the year, but insights on later phases of the crime report are increasingly deferred. As an outcome, reports of this sort are continually giving outcomes that do not allow the present year or the past one yet will reveal insight into the circumstance 3-4 years back in time. 3-4 years would be unreasonably long for a forward-thinking appraisal of the current situation. So, a checked improvement required significantly more progressed measurable systems. To improve their measurement, the UNODC has been chipping away at a help that is proving to be fruitful in the

long term. The report involves eight sections. They are intended to manage every single focal issue tended to in the polls. Initially, police recorded crime.

Second, dealing with crime. Next, complex crime gets more priority than traditional crime measurements and arrangement. Fifthly, submit it to the equity system.

Next discussion on the punitive nature of criminal equity systems.

Seventhly, an introduction on jail populaces of the world shuts the examination of criminal equity data. Finally, discusses challenges with crime and criminal equity insights, contending for the significance of further upgrades in the territory.

Criminal investigation report and crime design from “Criminal equity arrangement Exploration Institute: Portland crime Data” [10]. Crime examples can be distinguished, almost continuously, when online web based life is observed. Crime can happen anyplace whenever. Past Insights do not precisely distinguish the crime power of a particular location. More exact outcomes can be drawn from online life. Results from geographic information investigation led on different tweets gave an away from the criminal patterns in a few unique urban areas. The crime power day-wise is firmly related with wrongdoing estimates from cops, which finally show the theory. The Ferguson shooting relevant examination clearly isolates the city's shielded and dangerous model. To be continuously careful, we analyzed the specific twitter accounts which tweet pretty much the wrongdoing circumstances happening in the city reliant on sheriff data and envisioned. The Outcomes amassed from this assessment were positive. An impelled inclination examination estimation will help in isolating wickedness executioners from tweets inside a specific territory. Video-to-content planning, picture-to-content preparing, and data from diverse online sources would in like manner help improve precision. This kind of study would help with lighting up others

in regards to the Crime plan both inside and around their territory, in the long run assisting them with staying in a protected zone.

Checking diverse online life outlets. For speculation, we follow R. Chandramouli' "Developing Social Media Dangers: Technology and Policy Viewpoints" via web-based networking media. One of the key features of web based life is that it enables anyone to energetically impart slants from any bit of the world. This property engages us to get astoundingly esteem proficient unbiased ends. Nevertheless, it goes with an expense. At first, not all comments are of high bore or important.

Second, it licenses customers/associations to post fake audits. This inclinates a prerequisite for recognizing significant comments from electronic long range interpersonal communication. Various progressing assessments investigated the troubles on the idea of comments. They found that a pleasing review should depict the features of the things furthermore, the experts/cons of the features. A progressively elaborative review that gives the all-out nuances of the thing will undoubtedly be seen as high gauge. Another examination by Ghose Etalon reviewed steadiness researched factors related to the pundit, for model, examiner characteristics and investigator history. For our topic, crime, new segmentation and knowledge base generation and public opinion mining for social awareness we listed the top most popular 30 crimes till now. For having public opinion, we will make a survey about crime. Our main purpose is to focus on the causes of crime and try to make society aware about crime. To prevent it, we used to keep our yard lights on and open the entryway when the doorbell rang, regardless of whether we didn't have a clue who was there. We don't do that on the planet we live in now. Be that as it may, there are numerous approaches to reclaim control and forestall crime in the locale. It just takes correspondence, responsibility and time. Work with nearby open offices and different associations on taking care of normal issues. Set up a Local Watch or a network watch, working with police. Ensure the

boulevards and homes are sufficiently bright. Report any crime or dubious activity rapidly to the police. There's even a free application for that: McGruff Versatile, open on iTunes or on the Google Play store. Volunteer to help clean up the region. Call the city work environments or neighborhood waste the official's association and schedule a dumpster for the event. At that point get litter together. Give care about where an individual life with each other. Organize to help clean and improve stops in the general region. All around kept play gear and an ideal park can pull in enough people to cripple criminal tasks. Request that the area government keep up the parks, expeditiously fix vandalism or other mischief, grasp a school, what's more, help understudies, work force and staff advance a sentiment of the system through the commitment in a wide extent of ventures and works out. Work with the school to develop sedate free, gun free zones on the occasion that they don't start now exist. Guide youths who need positive assistance from adults through tasks like Elder kin and more established kingmaker a system against violence contention. Join talk, move, painting, drawing, singing, instrument acting, and other imaginative articulations. Get adolescents required to structure it and suggest prizes. Make it a happiness, neighborhood merriment. Starting now and into the foreseeable future, we will gather more information from the online news gateway or the paper to develop the library for the language preparation reason and information mining will be going on from now for the individuals' conclusion. We will discharge a review structure to gather the significant information from the arbitrary cerebrum thing about the crime situation. From that we can without much of a stretch make an information list for making the mindfulness pattern.

The literature on fear of crime has filled quickly over the most recent thirty years. This paper analyzes the purposes behind this development and endeavors to put some structure on the work to date. The insufficiencies of proportions of dread of crime are talked about and elective

methodologies proposed. Elective illustrative hypotheses are looked at and systems for lessening dread investigated. As opposed to broad communications accounts and prevalent views, quantitative information demonstrates that the connection among exploitation and dread of crime is feeble. Endeavors to clarify this feeble relationship have, up until now, been unsuitable. This article endeavors to clarify the frail relationship by contending that the effect of exploitation is intervened by the convictions of the person. Specifically, casualties regularly utilize certain convictions or "methods of balance" to persuade themselves that their specific exploitation was not hurtful. By characterizing their exploitation thusly, they keep away from the dread and other negative responses that occasionally go with exploitation. This article records the five significant strategies of balance utilized by casualties, presents proof for their reality, and talks about their pertinence to victimology. The effect of exploitation encounters and crime-related factors on act-specific dread of crime are re-exploited. Seen danger and weakness to crime were required to intercede the impact of segment and crime-related factors on dread. The aftereffects of this investigation recommend that dread of property misfortune is more reasonable by crime-related factors than is dread of fierce exploitation. Perceptual factors lessen the immediate effect of exploitation encounters and neighborhood crime percentage on each sort of dread of crime. Nonetheless, specific segments and crime-related factors affect dread of property misfortune and dread of vicious crime. The paper closes with recommendations for future exploration on the social determinants of dread of crime among the old. Bangladesh has been confronting the issues of crime, culpability and defilement with its broadened structures, from the earliest starting point of the nation's autonomy. The current examination is an endeavor to portray the exhaustive and all-encompassing authentic investigation of crime in contemporary Bangladesh with extraordinary references to crime patterns and their related causations. The investigation is mostly unmistakable

in nature and the information of the examination has been gathered from different auxiliary sources like police insights, books, diaries, periodicals and related literature. For the reason for the investigation, the crime information from 1972-2009 has been dissected. The significant discoveries show that the pace of crime has expanded in a consistent manner from 2003-2008 and the most elevated number of crimes were submitted in 2008 (1.58 lakh), however; it was diminished somewhat in 2009. It likewise shows the perpetrated crime of the years (2003-2007) where 1.27 lakh crimes were committed in 2003, 1.20 lakh in 2004, 1.26 in 2005, 1.30 in 2006, and 1.57 in 2007. It's a matter of worry that the measurements spoke to just the violations that are being accounted for to the police. The other significant discoveries of the examination states, the majority of the basic crimes are financial violations; nonetheless, the patterns of non-monetary violations, particularly political crime, are alarming. In Bangladesh, to lessen crime the prime helpers like neediness, joblessness, political support, broken home, affectation by companions, and obliviousness must be tended to. Furthermore, different viable measures including public mindfulness, successful laws and strategy, police change, appropriate equity framework must be guaranteed. The pattern in announced crime in Bangladesh since the freedom war has been expanded with slight change (Figure-2). As indicated by police measurements, the quantity of all out offenses for both fierce and local misdemeanors have been diminishing gradually. In 1972, there were roughly 18000 brutal offenses. In 2009 the brutal offense was just 4331-a gigantic decline. Vandalism related misdemeanors have diminished at a comparably an enormous rate. In 1972, the complete number of property offenses was 39633. By 2009the number of offenses had tumbled to 14689 local misdemeanors. We realize that the pace of crime in Bangladesh is expanding step by step. In any case, in police measurements, we see the quantities of violations are diminishing because of non-announcing propensity of crime. This reality exhibits the

disappointment and inadequacies of the criminal equity arrangement of Bangladesh in detailing crimes by the people in question. As a result of this lack, the police Crime Rate (per Lakh). Crime in Bangladesh Social Science Review, (December), 2018 organization cannot record the specific number of crimes and to develop proper crime typologies. There are numerous purposes behind the non-detailing of crime. Casualties may consider the violations are of significance and plan to try not to humiliate the guilty party, (ii). Wishing to evade the exposure and burden of calling the police, have consented to the crime, as in betting and some sexual offenses, might be scared by the criminal. Because of informal arrangement techniques utilized by the police, others' classes of crime were most noteworthy in Bangladesh after the freedom period. As indicated by police measurements, in 1972 the quantity of all our other crimes was 32306. Then again, the quantity of different offenses was 87022 out of 2009 - a huge expansion.

Keywords: Crime Prediction, Data Mining, Open Data, Regression, Decision Trees, Instance Based Learning

Abbreviations: United Nation Office on Drug and Crime[4]

Chapter 3

3.1 PROPOSED MODEL

Since the bad guys in the world are growing dramatically, we need to step up to avoid the accidents or at least try to minimize them where possible. Although the law enforcement department is working heart and soul, the crime rates are rising day by day. Keeping in mind, software still steps

up to the rescue, so we have come up with a complex model. Under the Ministry of Home Affairs of the Government of Bangladesh, we have received a dataset from the Bangladesh Police. The aim of our model is to predict necessary details about a crime that analyzes the victim's initial data, most importantly the crucial clue behind it about the perpetrator. Our model pre-processes both Label Encoder and One Hot Encoder data, extracts the features and labels, splits the data into training and test data ratios of 80:20. We will train 80% of the data and we will monitor more than 20% of the data in order to gain greater precision in the final results. In addition, we have introduced seven algorithms for machine learning. Such as: Logistic Regression, Support Vector Classifier (SVC), AdaBoost Classifier, Naïve Bayes, Random Forest Classifier, Decision Tree Classifier and K-Nearest Neighbor to find the best fit for our dataset. However, we applied Hyper Parameter Tuning to find the best parameters that have the best result to boost the outcome of our model. Finally, with the confusion matrix and precision, we compare the classifiers.

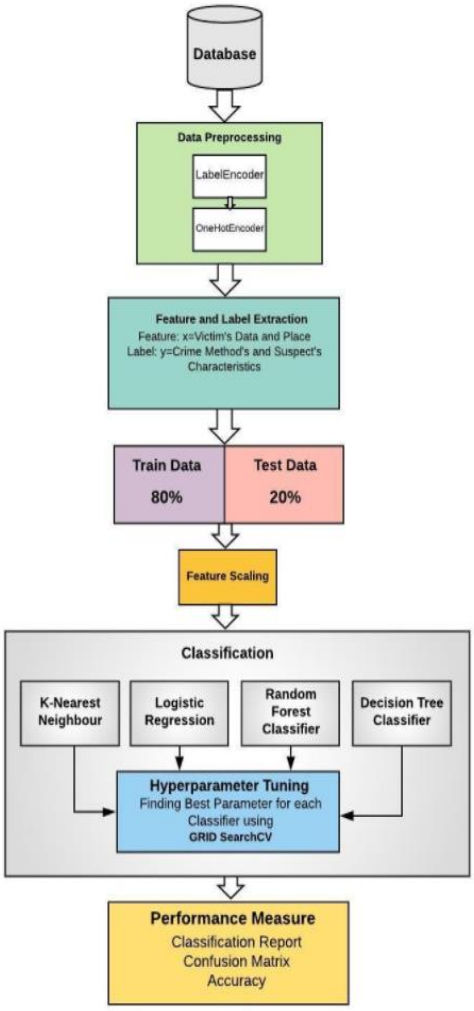


Figure 1: Proposed Model

3.2 METHODOLOGY

The main purpose of the proposed crime detection process is dependent on a few steps. Here the process collects criminal data from inputs. It may follow various types of input methods such as Natural Language Processing from news articles, raw data input, data based on public opinion etc. Then this method will cross check the input data with the previous data and it will pre-process the data for crime feature extraction and detection. Then it goes through the data mining techniques

for detecting the crime pattern. If the pattern is known, it decides and analyses the result otherwise a new pattern recognition process will run and enlist it in the dataset.

The main feature of the model is divided into seven steps which are dataset, processing, extraction, mining, detection, decision making and analysis. After applying these we will get a brief knowledge of the crime scenario based on time, date, crime type, reason, affected people and all other previous and futuristic data.

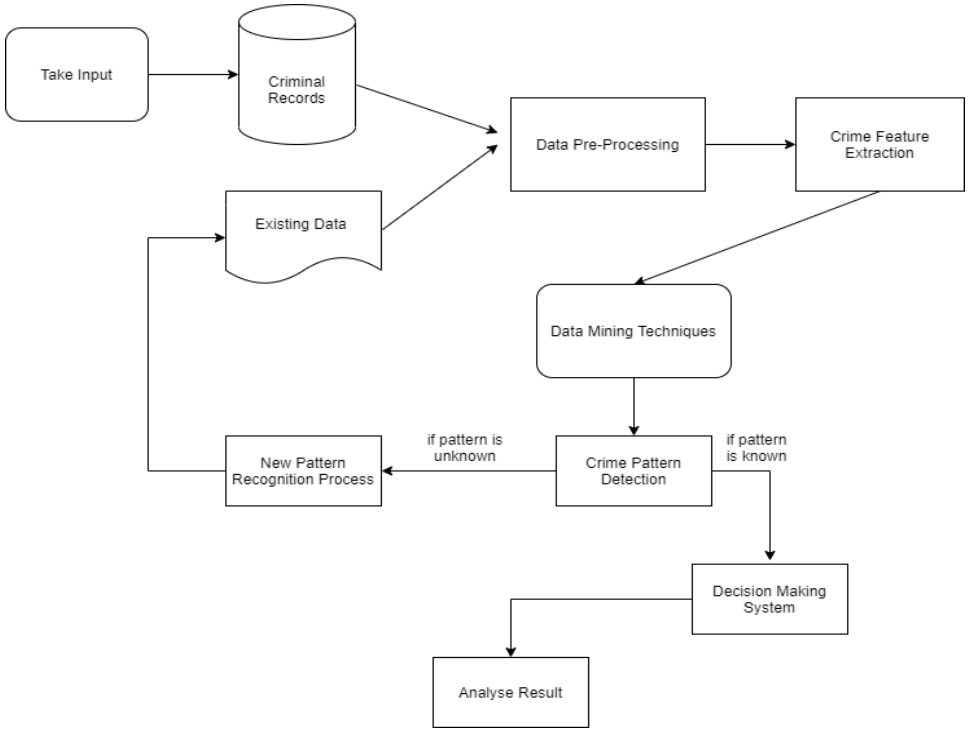


Figure 2: Methodology

Chapter 4

WORK PLAN

4.1 WORKFLOW MODEL

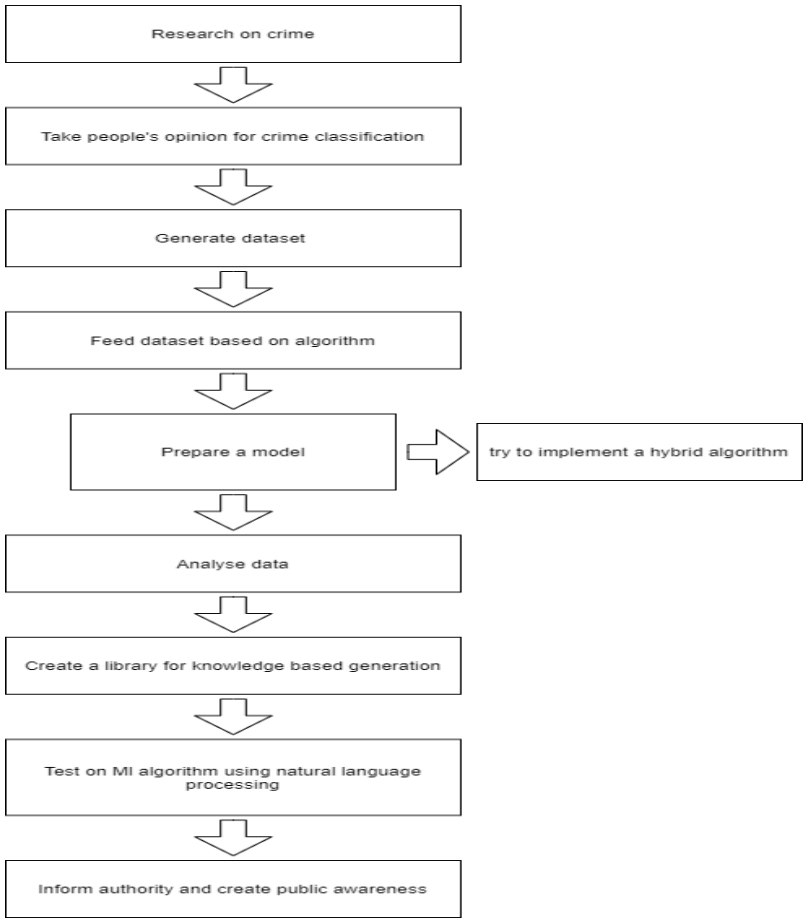


Figure 3: Workflow Model

4.2 DATA PROCESSING

Data preprocessing is a technique for data mining that requires converting raw data into a comprehensible format. Real-world data is often incomplete, unreliable, and/or missing, and is likely to contain several errors in some habits or patterns. Preprocessing of data is an established method of solving such problems. Preprocessing of data prepares raw information for further processing.

There are many ways that can process data for implementation. However, here the Data Cleaning method is used.

4.3 DATA CLEANING METHOD

The data can have many irrelevant and missing parts. To handle this part, data cleaning works fine. It includes handling missing data, noisy data and in some contexts Label Encoder and One Hot Encoder has been applied.

4.4 MISSING DATA

When any information is lacking in the data, this condition occurs. It is possible to handle it in different ways. There are some of them are described after this.

4.5 IGNORE THE TUPLES

Only when the dataset we have is very big and several values inside a tuple are missing is this method appropriate.

4.6 FILL THE MISSING VALUES

There are different ways to do this assignment. You may opt to manually fill the missing values by the mean attribute or by the most likely value.

4.7 LEVEL ENCODING

It is possible to achieve Label Encoding in Python using the Sklearn Library. Sklearn offers a very powerful method for encoding into numerical values the levels of categorical features. The Label Encoder encodes labels with a value between 0 and n-classes-1, where the number of independent labels is n. It assigns the same value as previously assigned if a mark is replicated.

Here the value of the Label Encoder is used in our model and applied to the categorical features.

4.8 ONE HOT ENCODING

One hot encoding is a method that converts categorical variables into a type that could be provided to ML algorithms to do a better prediction job.

Four categorical features are applied to these two techniques: victim sex, victim race, crime and location.

4.9 TRAIN-TEST SPLIT

Data is divided into training data and testing data (and often into three: train, validate and test) in machine learning in most scenarios and matches our model on train data in order to make predictions on test data. The training dataset is a component of the actual dataset that we use for model training. From this knowledge, the model sees and learns. On the other hand, test data is the

sample of data used to provide an objective study of a final model that fits into the training dataset. The Test dataset offers the perfect norm used for model evaluation. When the model is fully educated, it is used.

For this training and research, a certain ratio is picked. The ratio is usually 80:20. In addition, we have also selected 80:20 ratios, 80 percent training data and 20 percent testing details. As a result, 1176 data is allocated for training and 294 data for research is allocated. In addition, with 8 significant classifiers, we trained this data and carefully checked with the test data that robustly delegate our work.

Chapter 5

DATA SET

Out of the focuses introduced in the outline, our present standings so far are, right off the bat, we considered different Algorithms, realized which calculations are for the most part being utilized in crime report analysis. We have additionally executed a program to test data mining calculation on collected data. The conceived program can create a dataset that comprises record traits, calculation utilized, and time taken to encode and decode, increase in document size.

A preview of the dataset from 0 to 99 count with its 17 column data is presented in the diagram which is taken from BOSTON POLICE DEPARTMENT, [Table 1]

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOUR	UCR_PART	STREET	Lat	Long
0	I192074488	619	Larceny	LARCENY ALL OTHERS	C11	351	NaN	2017-08-23 00:00:00	2017	8	Wednesday	0	Part One	ADAMS ST	42.300605	-71.059230
1	I192073511	1107	Fraud	FRAUD - IMPERSONATION	B3	435	NaN	2017-02-21 00:01:00	2017	2	Tuesday	0	Part Two	ARMANDINE ST	42.284315	-71.074108
2	I192073187	2629	Harassment	HARASSMENT	C6	231	NaN	2017-07-13 00:00:00	2017	7	Thursday	0	Part Two	E FIFTH ST	42.333989	-71.032606
3	I192072907	802	Simple Assault	ASSAULT SIMPLE - BATTERY	D4	171	NaN	2017-09-01 16:32:00	2017	9	Friday	16	Part Two	HARRISON AVE	42.335119	-71.074917
4	I192072900	1102	Fraud	FRAUD - FALSE PRETENSE / SCHEME	D4	151	NaN	2017-09-01 00:00:00	2017	9	Friday	0	Part Two	WARREN AVE	42.345163	-71.071291
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	I182093039	3201	Property Lost	PROPERTY - LOST	D4	630	NaN	2017-12-17 00:00:00	2017	12	Sunday	0	Part Three	BEACON ST	42.346426	-71.106114
96	I182092552	3201	Property Lost	PROPERTY - LOST	B2	259	NaN	2017-09-27 00:00:00	2017	9	Wednesday	0	Part Three	MAGNOLIA ST	42.314784	-71.071610
97	I182092368	3201	Property Lost	PROPERTY - LOST	B3	944	NaN	2017-09-01 12:00:00	2017	9	Friday	12	Part Three	AMES WAY	42.289658	-71.086344
98	I182091879	616	Larceny	LARCENY THEFT OF BICYCLE	D14	795	NaN	2017-12-01 07:00:00	2017	12	Friday	7	Part One	COMMONWEALTH AVE	42.349007	-71.138601
99	I182090944	1102	Fraud	FRAUD - FALSE PRETENSE / SCHEME	D14	802	NaN	2017-10-04 21:58:00	2017	10	Wednesday	21	Part Two	CAMBRIDGE ST	42.354409	-71.135404

Table 1: Sample Set of Data

Chapter 6

BACKGROUND ANALYSIS

6.1 PREVIOUS WORK

6.1.1 PredPol

PredPol[5] utilizes an AI calculation to compute its forecasts. Historical occasion datasets are utilized to prepare the calculation for each new city (in a perfect world 2 to 5 years of information). It at that point refreshes the calculation every day with new occasions as they are gotten from the office. This data comes from the organization's records and the executive's framework (RMS)[6]. PredPol utilizes only 3 information focuses – crime type, crime area, and crime date/time – to make its expectations. No, recognizable data is at any point utilized. No segment, ethnic or financial data is at any point utilized. This wipes out the opportunities for protection or social equality infringement seen with other insight driven policing models. Forecasts are shown as red boxes on a web interface utilizing Google Maps. Each case is a 150×150-meter square. The cases address the most noteworthy danger regions for every day and the relating shift: day, swing, or night shift. Officials are told to invest generally 10% of their shift energy watching PredPol boxes. PredPol operation has both mission planning and location management.

PredPol[5] set a mission for every shift, beat, and day of a week. The mission is to categorize crime. PredPol showed the highest number of locations for events along with the mission. PredPol can identify the "dosage" of PredPol boxes. PredPol likewise makes patrol heat maps that permit order staff to check whether any spaces of their purview are being under patrolled. PredPol's yield is a guide with 500' x 500' boxes featuring spaces of conceivable crime problem areas. The gauge comprises the best 20 boxes in position request shown on a guide and accessible for coordinated

police patrol. There is no specific analysis for PredPol[5] accuracy. In a study in 2015 a company found that the "model effectively anticipated 4.7% of crimes; a prescient exactness 2.2 times greater" than existing methods. Despite its limits to a particular subset of crime, and regardless of the way that the greatest "boost" went from a 2.1% achievement rate to a 4.7% achievement rate, the investigation has PredPol[5] to guarantee that it prompts forecasts "twice as exact as those made through existing accepted procedures," and that "crime experts utilizing PredPol[5] were 100% more powerful – anticipating twice as numerous violations – as investigators utilizing just problem area planning and knowledge models. A Forbes report in 2015 said that almost 60 departments used PredPol[5].

6.1.2 KeyCrime

KeyCrime is an IT instrument, center of an inventive examination strategy that plans to accomplish, inside its field, one of the furthest points of the intellectual hexagon handled by the psychological science: computerized reasoning. This program can store and dissect up to 12.000 data for each crime: from the most nonexclusive, like date and spot, to the most itemized one that additionally considers the culprit. KeyCrime is going to be introduced to other Police's base camp as a result of its creative potential and its flexibility: by adjusting only a couple boundaries, it very well may be utilized to settle each sort of sequential crime, for instance inappropriate behavior in the city. Decision Support Software is basically KeyCrime arrangement which is intended to help the investigative efforts of Police powers by coordinating with data so they can settle on the correct choices. Then collect data which guides Law Enforcement Agency clients through the screening, ensuring that total information is gathered. Also, key crime considers more than 1.5 Million variable mixes going from the broader (e.g.: date, time, a spot of the occasion) to the unmistakable

Utilizing the Spatial, Temporal, and Behavioral qualities of connected occasions, assists clients with interpreting the time stretches, target types, and geological spaces of the arrangement so that officials can assume the following conceivable occasion. Crime Prevention and Repression helps to understand the criminal mentality, KeyCrime embodies hypotheses for recognizing sequential crooks, investigating and breaking down the information of every criminal occasion.it is a common conviction that KeyCrime is superior to other programming whose investigation standards depend on science models that couldn't measure up to the ones utilized by this imaginative instrument.

The information distributed by Milan's Police settle outline that the measure of violations addressed with KeyCrime went from 27% (2007, first year of testing) to 57% (2013) with pinnacles of 80% for what concerns burglaries in drug stores, which are the organizations with the most noteworthy pace of thefts (269 out of 2013). These outcomes acquainted a solid obstacle with an abatement of these occasions, as enrolled in the initial 10 months of the current year that, for instance, showed a 25% decline in burglaries in drug stores. The complete lesson is 21%. (- 57% in 2015). The outcomes were prompt and momentous. In 2013 they addressed 74% of the burglaries (124 violations dissected in banks). Likewise, in the initial 10 months of the current year there has been a lessening of 48% in this sort of wrongdoing. (- 80% in 2015)

6.2 PRESENT WORK

In our sector, we categorize crimes district-wise, monthly and yearly. For this, we use DecisionTreeClassifier, BernoulliNB, ExtraTreeClassifier, KNeighborsClassifier, GaussianNB, LinearDiscriminantAnalysis, LinearSVC, NearestCentroid, Random Forest, LGBMClassifier, LSTM Architecture. DecisionTreeClassifier is an algorithm in the scikit-learn library. For classification and regression problems supervised learning methods are needed. BernoulliNB

algorithm is suitable for discrete data which is designed for Binary or Boolean features. KNeighborsClassifier does not use training data points for generalization. It is used for classification and regression. GaussianNB It is a simple classification technique but has high functionality which makes the model ready to make predictions. LabelSpread algorithm is used for learning from labeled training data. Random forests create decision trees on randomly selected data samples, get a prediction from each tree, and select the best solution utilizing voting.

Keywords: Law Enforcement Agency, Achievement Rate of the Investigation, Classification and Regression Problems

Abbreviations: Predictive Policing[5], Optical Character Recognition[6]

Chapter 7

DATA EXTRACTION

Being up-to-date with all the relevant crime data and related data from victims as well as suspects, law enforcement authorities are always trying to find out the black sheep behind every incident. The highest efficiency has not been updated, despite getting all the data to some degree. Machine learning has some excellent data analysis capabilities, and Supervised Learning offers the opportunity to predict a cabalistic indication that enables us to grow to another aspect of information technology. In this case, when people are found by a criminal activity, they go to the nearest law enforcement organization to report and first the information gathers are mostly; where it happened, how many victims were affected, what the crime occurred, what the victims' race and

sex were. We have chosen this useful knowledge to figure out a trend, in addition to predicting the criminal traits behind this perpetration, by calculating the dataset.

7.1 Features: (Data here is for example) / (Victim's information)

Victim Age	Victim Sex	Victim Race	Crime	Area
35-40	Male	Black	Murder	Badda
15-20	Female	Latino	Terrorism	Lalbagh

Table 2: Data Extraction Features

7.2 LABELS: (Data here is for example) / (Criminal's Characteristics and Method)

Age Range	Sex	Race	Method
21-30	Male	Black	Electrocution
41-50	Male	White	Deadly Weapon

Table 3: Data Extraction Labels

Chapter 8

ALGORITHMS

Here a selection of a few related algorithms are explained for the better understanding of the problem statement.

8.1 DECISION TREE

The estimation of the decision tree tries to take care of the problem by using tree representation. A characteristic relates to each inner hub of the tree, and each leaf hub corresponds with a class name.

Decision Tree Calculation Pseudo code:

a. Place the best quality of the dataset at the foundation of the tree.

b. Split the preparation set into subsets. Subsets ought to be made so that every subset contains information with a similar incentive for a trait.

c. Repeat stage 1 and stage 2 on every subset until you discover leaf hubs in every one of the parts of the tree. In decision trees, for anticipating a class name for a record we begin from the base of the tree. We look at the estimations of the root property with the record's trait. Based on examination, we pursue the branch relating to that worth and hop to the following hub.

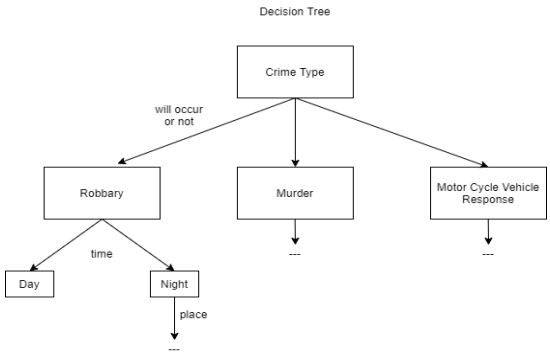


Figure 4: Decision Tree

To continue the contrast characteristic and qualities of our record and other inward hubs of the tree until we arrive at a leaf hub with predicted class esteem. As we are probably aware of how the tree of choice shown can be used to foresee the objective class or the meaning. We're currently seeing how we can render the Decision tree model.

DecisionTreeClassifier is an algorithm in the scikit-learn library. For classification and regression problems supervised learning methods are needed. The idea of decision tree has formed for this concept though it is primarily used for classification problems. It is developed like a tree-organized classifier, where inward nodes address the highlights of a dataset, branches address the choice principles and each leaf hub addresses the result.

Algorithm:

$$\text{Info}(D) = -\sum_{i=1}^{m} pi \log_2 pi$$

Pi is the probability that an arbitrary tuple in D belongs to class Ci.

$$\text{Info}_A(D) = \sum_{j=1}^{V} \frac{|Dj|}{|D|} \text{ X Info}(D_j)$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

Figure 5: DTC Algorithm

Here,

1. Info(D) = average amount of information. it changes until it identifies the class label of a tuple in D
2. D_j/D = weight of the jth partition.
3. $\text{Info}_A(D)$ =expected information that is required to classify a tuple from D based on the partitioning by A.
4. Gain(A), is chosen as the splitting attribute at node N () when the attribute A with the highest information gain.

8.2 LINEAR SVM

One of the fastest AI models for grouping multi-class datasets is the linear vector model. It suits well for taking care of various problems down to earth. Behind the estimate, which finds the best suited line or hyper plane to separate the knowledge into two groups, the short-sighted thinking works.

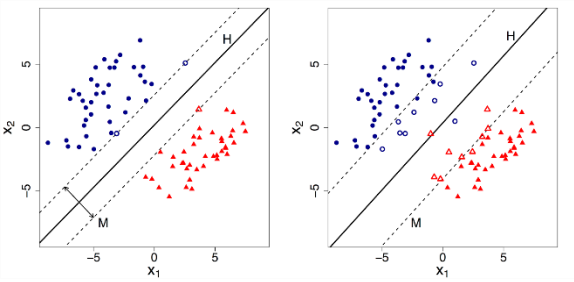


Figure 6: SVM

In Linear SVM, it is easy to think about simply drawing a line and segregating the class. In any event, we need to note that a large number of lines can be drawn to separate the data into two classes, but we need to find the right one for the unique information index to isolate the two classes in the most optimal way imaginable. It depends on the support vector, as per linear SVM, to find the best fitting rows. The emphasis that is closest to the line is Bolster Vectors and we get the Advantage by computing the good ways from support vectors to line. By taking the largest Edge among the other hyper planes, we'll choose the ideal hyper plane. With any number of classes, Linear SVM can draw a multi-class order response. Tt can likewise manage huge information and multi measurement information greatly.

8.3 LinearSVC

Support vector machine is a powerful and versatile machine learning model that can perform linear and non-linear classification, regression, and even external recognition. LinearSVC is the linear version of the SVM model.

8.4 LabelSpreading

LabelSpreading is a scikit-learn library algorithm that also belongs to the semi-supervised learning algorithm. To make use of both labeled and unlabeled training data this algorithm is used. Semi-supervised learning algorithms are the subset of supervised learning algorithms and it is used for learning from labeled training data. Algorithm,

$$\mathcal{L} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

$$l_{ij} = -\frac{1}{\sqrt{deg(v_i)}\sqrt{deg(v_j)}} \quad if\ v_i \in NN(v_j)$$

Figure 7: LS Algorithm

8.5 LinearDiscriminantAnalysis

Linear Discriminant Analysis is a machine learning algorithm that is used for classification. By using the mean and standard deviation it calculates summary statistics for the input features by class label which represent the model learned from the training data.

Algorithm,

$$P(Y=x|X=x) = (Plk * fk(x)) / sum(Pll * fl(x))$$

$$Plk = nk/n$$

$$Dk(x) = x * (muk/siga^2) – (muk^2/(2*sigma^2)) + ln(Plk)$$

Figure 8: LD Process

8.6 LOGISTIC REGRESSION

One of the most well-known relapse models is Logistic Regression. In addition, in the AI sector, calculation is commonly used for arranging data sets. It pursues a controlled organizational teaching strategy. Despite the fact that it is essentially the same as with another well-known relapse model (Linear Regression), the essential difference between the equations lies in generating the desired yield. Generally, it offers free qualities in Linear relapse and anticipates the estimates of the journalist ward as a result by preparing the model with the preparation of datasets. Interestingly, the same approach is carried out to anticipate a parallel option in logistic relapse (subordinate qualities). Here, it can be numerical or all out free qualities or components. Besides, in Multinomial Logistic relapse, it is utilized for downright ward factors with multiple classes. The general condition of Logistic relapse is:

$$\mathbf{\log(p(X)/(1-p(X))) = \beta 0 + \beta 1 X}$$

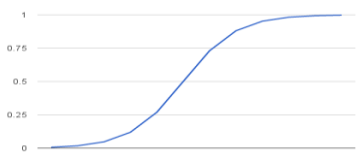


Figure 9: Logistic Regression

Where, p(X) speaks to the needy variable, X speaks to the autonomous variable. β0 speaks to the capture and β1 speaks to the slant co-proficient.

Logistic Regression utilizes a more mind-boggling skill called Sigmoid Feature than straight. It produces an S-formed bend that can take any real value number and produce a double yield somewhere in the 0 and 1 range, but never precisely at those containment points. From the preparation of details, the coefficients (beta characteristics b) of the logistic regression calculation must be assessed. For that, one needs to use the most extreme estimate of likelihood. The needy vector pursues distribution to Bernoulli here. For the default class and a value near 0 for various classes in the model, the best coefficients will predict a value remarkably close to 1. In addition, we use Gradient Descent in Logistic relapse to lower the cost value and make it progressively competent.

8.7 NAïVE BAYES

For model forecasting, Naive Bayes is a simple but effective calculation. This estimate originates from the hypothesis of Bayes' possibility of ordering protests. Naive Bayes classifiers accept strong, or credulous, autonomy between data focus characteristics. For a parallel and multi classification order, the Naïve Bayes calculation can be related. It could be prepared on a small uninformed set that, in addition to stage, could end up being a major. Furthermore, it is much faster and more climbable. There we found several propelled relations between the variables of the element. Ventures to ascertain Prediction:

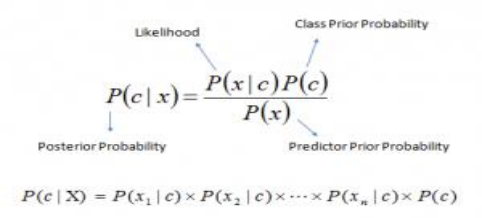


Figure 10: Naïve Bayes

Step1: Converting the informational index into a recurrence table

Step2: Creating probability table by finding the probabilities

Step3: Using Naive Bayesian condition to compute the back likelihood for each class. The class with the most astounding back likelihood is the result of the forecast.

8.8 RANDOM FOREST

Random Forest is a controlled measurement of the arrangement settled on many trees of choice. Any decision tree generated using a subset of attributes used to characterize a predefined sub-tree is a collection of selection trees. In a predefined model, those decision trees vote on the most competent method of organizing input information, and the Random Forest bootstraps those votes to choose the best estimate. Random Forest is completed to counteract over-fitting, a common defect of the decision tree. Random Forest creates a forest where it haphazardly arranges the nodes and components. The more trees in the forest, the better the results that can be produced. If a planning dataset is not likely to lead to goals and highlights in the decision tree, a collection of principles will be prepared that can be used to meet expectations.

It is easy-to-use and more flexible supervised learning algorithms used both for classification and regression are known as Random Forest. A forest is created from trees. The forest is more robust when it has more trees. Random forests create decision trees on randomly selected data samples, get a prediction from each tree, and select the best solution by means of voting. The formula for the random forest:

$$RFfi_i = \frac{\sum_j normfi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} normfi_{jk}}$$

Figure 11: Random Forest

1. RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model
2. normfi sub(ij)= the normalized feature importance for i in tree j

8.9 LGBMClassifier

With the capability of handling large-scale data, LightGBM is a framework for radiant boosting with the use case of tree-based learning algorithms. LGBMClassifier is used for its Faster training speed and higher efficiency, lower memory usage, and better accuracy which Supports parallel, distributed, and GPU learning.

8.10 LSTM ARCHITECTURE

Long short-term memory (LSTM) is created for short-term memory problems. In the field of deep learning this is an architecture for artificial recurrent neural networks (RNN).

8.11 ADaBOOST

AdaBoost, short for Adaptive Boosting, is an algorithm for machine learning that solves two-class problems and is used to train weak classifiers to construct a strong algorithm. AdaBoost is adaptive in the sense that subsequent constructed classifiers are tweaked in favor of certain instances that are misclassified by the previous classifiers.

A training set **S = (X1, Y1) ….., (Xm, Ym)** is taken as input, where each instance **Xi belongs** to a domain or instance space X, and each label **Yi** belongs to a finite label space Y.

In every round m = 1; M, a weak or base learning algorithm is called by AdaBoost which accepts a sequence of training examples S as an input along with a distribution or set of weights over the training example. A weak classifier is computed by a weak learner given such an input, **{-1, +1} ∈ ht** where **ht** has the form **ht : X→R**. Once the weak classifier has been received, AdaBoost chooses a parameter that intuitively measures the importance that it assigns to ht.

Boosting is to use the weaker learner by repeatedly calling the weaker ones over the training examples on various distributions to form a highly accurate prediction law. Initially, all the weights are set evenly, but in each round the weights of incorrectly classified examples are increased such that the observations that the previous classifier wrongly predicts get greater weight on the next iteration.

8.12 K-NEAREST NEIGHBOR

The classifier used by KNN acts as a pursuit. Preparing is a simple task for reserving all models of planning. The predefined steady K indicates the quantity of stored passages that are closest to the model being characterized in the region to be used when playing out the order. The grouping

occurs by casting a ballot. The indicated range limits the region of each factor's persistent knowledge space.

The class ID that is legitimately returned by the classifier is converted to a label for display on the working graphical board. The K-NN classifier returns a simple approximation that recognizes a mark.

KNearestClassifier is an algorithm from the scikit-learn library. It is also known as KNN which is a lazy model. This algorithm does not use training data points for generalization. It is used for classification and regression. The formula:

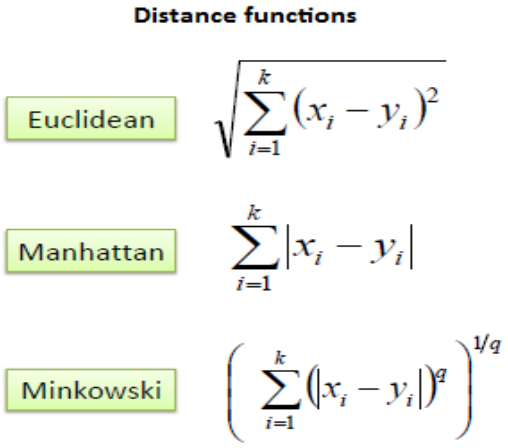


Figure 12: K-Nearest Neighbor

Here,

1. With a majority vote, the case is classified.
2. The case is assigned to the class most common amongst its K nearest neighbors measured by a distance function.
3. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

8.13 BernoulliNB

BernoulliNB is an Implementation in the scikit-learn library that is a variant of the Bernoulli Nayeb Bayes Algorithm. This Algorithm is suitable for discrete data which is designed for Binary or Boolean features.

Algorithm:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Figure 13: BernoulliNB

The parameters σ_y and μ_y are estimated using maximum likelihood.

8.14 ExtraTreesClassifier

ExtraTreesClassifier is an algorithm from the scikit-learn library. It is a troupe learning strategy generally dependent on choice trees. ExtraTreesClassifier, as RandomForest, randomizes certain choices and subsets of information to limit over-gaining from the information and overfitting.

The formula:

$$Gain(S, A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2(p_i)$$

Figure 14: ETC

where c is the number of unique class label and Pi is the proportion of rows with output label is i.

8.15 GaussianNB

Naive Bayes has another variant called Gaussian Naive Bayes and for its implementation, GaussianNB was built. It has high functionality with a simple classification technique. It is a simple classification technique but has high functionality which makes the model ready to make predictions.

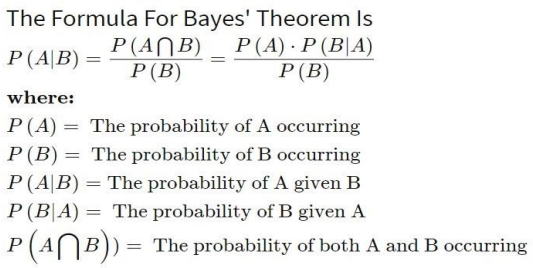


Figure 15: GaussianNB

8.16 LinearDiscriminantAnalysis

Linear Discriminant Analysis is a machine learning algorithm that is used for classification. By using the mean and standard deviation it calculates summary statistics for the input features by class label which represent the model learned from the training data.

8.17 NearestCentroid

NearestCentroid is a mean value. While encountering machine learning, there is the nearest centroid classifier or nearest prototype classifier which is a classification model. This assigns to observations the label of the class of training samples. And the closest observation means is known as NearestCentroid.

8.18 NATURAL LANGUAGE PROCESSING

The area of artificial intelligence has always envisaged that computers are capable of imitating the human mind's functioning and abilities. Language is recognized as one of the most important human accomplishments that has driven humanity's progress. Therefore, it is not shocking that there is a lot of work being done in the form of natural language processing to incorporate language into the field of artificial intelligence (NLP). We see the work today manifesting itself in the likes of Alexa and Siri.

NLP consists mainly of understanding natural language (human to machine) and generation of natural language (machine to human).

This article deals primarily with understanding natural language (NLU). There has been a surge in unstructured data in the form of text, videos, audio, and photos in recent years. NLU helps to extract valuable text information, such as social media information, customer surveys, and complaints.

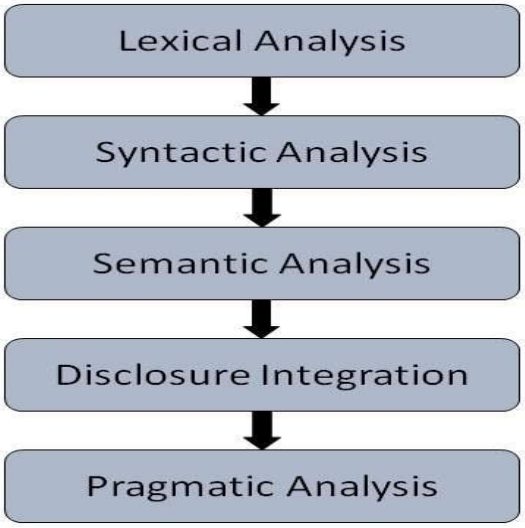


Figure 16: NLP

Chapter 9

DATA MANIPULATION AND MACHINE LEARNING

In data training and learning segmentation, calculation is done crime wise, district wise and UCR wise where eight (8) algorithm is tested to check which one gives better accuracy, support, precision, recall and f1-score. The max, min and mean value is also calculated and macro average and weighted average are calculated.

9.1 DecisionTreeClassifier

9.1.1 CRIME WISE

Here this test case contains a "fun_DecisionTreeClassifier (X_train, Y_train)" method which contains the crimes on x-tree to predict its occurrence accuracy. The max f1-score is 48 and the lowest one is 10 where the mean accuracy is 22% [fig17].

```
mean: 0.20072765215508578
max: 0.4771875403903322
min: 0.10136986301369862
              precision    recall  f1-score   support

           1       0.30      0.31      0.30      7878
           2       0.31      0.31      0.31      5850
           3       0.18      0.18      0.18      5325
           4       0.13      0.13      0.13      4133
           5       0.14      0.13      0.13      4229
           6       0.51      0.45      0.48      4106
           7       0.11      0.12      0.11      3424
           8       0.10      0.10      0.10      3197
           9       0.13      0.14      0.14      2755
          10       0.24      0.25      0.25      2400
          11       0.17      0.18      0.17      2380
          12       0.10      0.10      0.10      2362

    accuracy                           0.22     48039
   macro avg       0.20      0.20      0.20     48039
weighted avg       0.22      0.22      0.22     48039

accuracy 0.21865567559691085
```

Figure 17: DTC Crime Wise Result

9.1.2 DISTRICT WISE

Following the previous workflow, if x-tree has the value of districts, that will give an accuracy for the possible crime incidents by analyzing the previous data. Here the max f1-score is 21, min is 1 and the accuracy is 15% [fig18].

```
mean: 0.10759166483360157
max: 0.20852158993422937
min: 0.011372867587327378
              precision    recall  f1-score   support

         1.0       0.24      0.14      0.18      8939
         2.0       0.08      0.07      0.07      2986
         3.0       0.25      0.18      0.21     10130
         4.0       0.04      0.06      0.05      1478
         5.0       0.09      0.11      0.10      2897
         6.0       0.07      0.09      0.08      2279
         7.0       0.04      0.07      0.05      1570
         8.0       0.15      0.16      0.16      5854
         9.0       0.19      0.21      0.20      5543
        10.0       0.02      0.05      0.03       752
        11.0       0.13      0.19      0.16      3622
        12.0       0.01      0.03      0.01       246

    accuracy                           0.15     46296
   macro avg       0.11      0.11      0.11     46296
weighted avg       0.17      0.15      0.15     46296

accuracy 0.14705374114394332
```

Figure 18: DTC District Wise Result

9.1.3 UCR WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 54, min f1-score of 30 and accuracy of 45% [fig19].

```
mean: 0.4154988605597054
max: 0.5422987830739616
min: 0.3004797208896642
              precision    recall  f1-score   support

         1.0       0.31      0.29      0.30      7173
         2.0       0.42      0.39      0.40     10800
         3.0       0.52      0.57      0.54     14667

    accuracy                           0.45     32640
   macro avg       0.42      0.41      0.42     32640
weighted avg       0.44      0.45      0.44     32640

accuracy 0.4470281862745098
```

Figure 19: DTC UCR Wise Result

9.2 BernoulliNB

9.2.1 CRIME WISE

Here this test case contains a “fun_BernoulliNB (X_train, Y_train)” method which contains the crimes on x-tree to predict its occurrence accuracy by crime. The max f1-score is 26 and the lowest one is 0 where the accuracy is 16% [fig20].

```
mean: 0.0603988076643411
max: 0.2612469743760955
min: 0.0
              precision    recall  f1-score   support

           1       0.19      0.42      0.26      3768
           2       0.78      0.14      0.24     32475
           3       0.23      0.12      0.16      9916
           4       0.00      0.00      0.00         0
           5       0.00      0.00      0.00         0
           6       0.00      0.00      0.00         0
           7       0.00      0.00      0.00         0
           8       0.05      0.09      0.07      1880
           9       0.00      0.00      0.00         0
          10       0.00      0.00      0.00         0
          11       0.00      0.00      0.00         0
          12       0.00      0.00      0.00         0

    accuracy                           0.16     48039
   macro avg       0.10      0.06      0.06     48039
weighted avg       0.59      0.16      0.22     48039

accuracy 0.1553321259809738
```

Figure 20: BernoulliNB Crime Wise Report

9.2.2 DISTRICT WISE

From the previous workflow of BernoulliNB, if x-tree has the value of districts, that will give an accuracy for the possible crime incidents by analyzing the previous data. Here the max f1-score is 27, min is 0 and the accuracy is 16% [fig21].

```
mean: 0.029774019506895302
max: 0.2724610518917653
min: 0.0
              precision    recall  f1-score   support

         1.0       0.00      0.00      0.00         0
         2.0       0.00      0.00      0.00         0
         3.0       0.95      0.16      0.27     43867
         4.0       0.00      0.00      0.00         0
         5.0       0.00      0.00      0.00         0
         6.0       0.00      0.00      0.00         0
         7.0       0.00      0.00      0.00         0
         8.0       0.06      0.15      0.08      2429
         9.0       0.00      0.00      0.00         0
        10.0       0.00      0.00      0.00         0
        11.0       0.00      0.00      0.00         0
        12.0       0.00      0.00      0.00         0

    accuracy                           0.16     46296
   macro avg       0.08      0.03      0.03     46296
weighted avg       0.90      0.16      0.26     46296

accuracy 0.158545014688094
```

Figure 21: BernoulliNB District Wise Report

9.2.3 UCR WISE

Here if the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 66, min f1-score of 0 and accuracy of 49% [fig22].

```
mean: 0.21940605534341515
max: 0.6574243577349487
min: 0.0
              precision    recall  f1-score   support

         1.0       0.00      0.00      0.00         0
         2.0       0.00      0.57      0.00         7
         3.0       1.00      0.49      0.66     32633

    accuracy                           0.49     32640
   macro avg       0.33      0.35      0.22     32640
weighted avg       1.00      0.49      0.66     32640

accuracy 0.48973651960784315
```

Figure 22: BernoulliNB UCR Wise Report

9.3 ExtraTreeClassifier

9.3.1 CRIME WISE

This test case contains a “fun_ExtraTreeClassifier (X_train, Y_train)” method which contains the crimes on x-tree to predict its occurrence accuracy by crime. The max f1-score is 47 and the lowest one is 9 where the accuracy is 21% [fig23].

```
mean: 0.19066558718305585
max: 0.46730191880955485
min: 0.08994548758328286
              precision    recall  f1-score   support

           1       0.30      0.30      0.30      8146
           2       0.32      0.31      0.31      5917
           3       0.17      0.17      0.17      5191
           4       0.11      0.11      0.11      3968
           5       0.14      0.13      0.13      4338
           6       0.49      0.44      0.47      4030
           7       0.10      0.10      0.10      3405
           8       0.09      0.09      0.09      3231
           9       0.12      0.12      0.12      2804
          10       0.24      0.25      0.24      2401
          11       0.15      0.15      0.15      2376
          12       0.09      0.10      0.09      2232

    accuracy                           0.21     48039
   macro avg       0.19      0.19      0.19     48039
weighted avg       0.21      0.21      0.21     48039

accuracy 0.2100376777201857
```

Figure 23: ETC Crime Wise Report

9.3.2 DISTRICT WISE

From the previous workflow of BernoulliNB, if x-tree has the value of districts, that will give an accuracy for the possible crime incidents by analyzing the previous data. Here the max f1-score is 21, min is 1 and the accuracy is 15% [fig24].

```
mean: 0.10769803762782203
max: 0.20934069059673474
min: 0.011466011466011467
              precision    recall  f1-score   support

         1.0       0.24      0.14      0.18      9052
         2.0       0.08      0.07      0.07      2947
         3.0       0.25      0.18      0.21     10224
         4.0       0.04      0.06      0.05      1431
         5.0       0.09      0.11      0.10      2884
         6.0       0.07      0.10      0.08      2263
         7.0       0.04      0.07      0.05      1541
         8.0       0.15      0.16      0.16      5860
         9.0       0.19      0.21      0.20      5569
        10.0       0.02      0.05      0.02       697
        11.0       0.13      0.18      0.15      3592
        12.0       0.01      0.03      0.01       236

    accuracy                           0.15     46296
   macro avg       0.11      0.11      0.11     46296
weighted avg       0.17      0.15      0.16     46296

accuracy 0.14759374459996544
```

Figure 24: ETC UCR Wise Report

9.3.3 UCR WISE

Here if the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 54, min f1-score of 28 and accuracy of 44% [fig25].

```
mean: 0.4040677776867619
max: 0.5363786354372231
min: 0.27851830031666547
              precision    recall  f1-score   support

         1.0       0.29      0.27      0.28      6994
         2.0       0.41      0.39      0.40     10719
         3.0       0.52      0.56      0.54     14927

    accuracy                           0.44     32640
   macro avg       0.41      0.40      0.40     32640
weighted avg       0.43      0.44      0.44     32640

accuracy 0.4384497549019608
```

Figure 25: ETC UCR Wise Report

9.4 KNeighborsClassifier

9.4.1 CRIME WISE

Here this test case contains a “fun_KNeighborsClassifier (X_train, Y_train)” method which contains the crimes on x-tree to predict its occurrence accuracy. The max f1-score is 37 and the lowest one is 7 where the mean accuracy is 22% [fig26].

```
mean: 0.17789879808095255
max: 0.36739692475166685
min: 0.07404505386875614
              precision    recall  f1-score   support

           1       0.47      0.24      0.32     15920
           2       0.38      0.27      0.31      8169
           3       0.16      0.16      0.16      5384
           4       0.09      0.13      0.11      2906
           5       0.09      0.13      0.11      2718
           6       0.37      0.36      0.37      3718
           7       0.08      0.14      0.10      1928
           8       0.06      0.11      0.07      1732
           9       0.09      0.14      0.11      1784
          10       0.23      0.32      0.27      1762
          11       0.10      0.23      0.14      1071
          12       0.05      0.14      0.08       947

    accuracy                           0.22     48039
   macro avg       0.18      0.20      0.18     48039
weighted avg       0.30      0.22      0.24     48039

accuracy 0.22082058327608817
```

Figure 26: KNeighbors Crime Wise Report

9.4.2 DISTRICT WISE

From the previous workflow of BernoulliNB, if x-tree has the value of districts, that will give an accuracy for the possible crime incidents by analyzing the previous data. Here the max f1-score is 21, min is 1 and the accuracy is 15% [fig27].

```
mean: 0.10439159603725612
max: 0.2108940947362392
min: 0.010291595197255574
              precision    recall  f1-score   support

         1.0       0.29      0.14      0.19     10928
         2.0       0.09      0.07      0.08      3502
         3.0       0.25      0.18      0.21     10104
         4.0       0.03      0.05      0.04      1308
         5.0       0.08      0.11      0.09      2489
         6.0       0.06      0.10      0.08      1873
         7.0       0.03      0.06      0.04      1196
         8.0       0.13      0.15      0.14      5236
         9.0       0.17      0.20      0.19      5096
        10.0       0.02      0.06      0.03       607
        11.0       0.15      0.19      0.17      3776
        12.0       0.01      0.03      0.01       181

    accuracy                           0.15     46296
   macro avg       0.11      0.11      0.10     46296
weighted avg       0.18      0.15      0.16     46296

accuracy 0.14655693796440297
```

Figure 27: KNeighbors District Wise Report

9.4.3 UCR WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 54, min f1-score of 29 and accuracy of 44% [fig28].

```
mean: 0.40343994704634717
max: 0.5404911838790931
min: 0.2939498573787719
              precision    recall  f1-score   support

         1.0       0.30      0.29      0.29      6737
         2.0       0.38      0.37      0.38     10127
         3.0       0.54      0.54      0.54     15776

    accuracy                           0.44     32640
   macro avg       0.40      0.40      0.40     32640
weighted avg       0.44      0.44      0.44     32640

accuracy 0.4392463235294118
```

Figure 28: KNeighbors Crime Wise Report

9.5 GaussianNB

9.5.1 CRIME WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 26, min f1-score of 0 and accuracy of 15% [fig29].

```
mean: 0.06279445013894926
max: 0.26442028384098265
min: 0.0
              precision    recall  f1-score   support

           1       0.17      0.41      0.24      3442
           2       0.55      0.17      0.26     18317
           3       0.00      0.25      0.00        24
           4       0.00      0.00      0.00         0
           5       0.00      0.00      0.00         0
           6       0.00      0.00      0.00         0
           7       0.00      0.00      0.00         0
           8       0.08      0.09      0.08      2899
           9       0.72      0.09      0.16     23357
          10       0.00      0.00      0.00         0
          11       0.00      0.00      0.00         0
          12       0.00      0.00      0.00         0

    accuracy                           0.15     48039
   macro avg       0.13      0.08      0.06     48039
weighted avg       0.58      0.15      0.20     48039

accuracy 0.14502799808488936
```

Figure 29: GaussianNB Crime Wise Report

9.5.2 DISTRICT WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 26, min f1-score of 0 and accuracy of 16% [fig30].

```
mean: 0.060728738676580775
max: 0.25505648278548526
min: 0.0
              precision    recall  f1-score   support

         1.0       0.00      0.00      0.00         0
         2.0       0.00      0.00      0.00         0
         3.0       0.58      0.16      0.26     26018
         4.0       0.00      0.00      0.00         0
         5.0       0.00      0.00      0.00         0
         6.0       0.00      0.00      0.00         0
         7.0       0.00      0.00      0.00         0
         8.0       0.07      0.14      0.09      2954
         9.0       0.35      0.17      0.23     12270
        10.0       0.00      0.00      0.00         0
        11.0       0.15      0.15      0.15      5054
        12.0       0.00      0.00      0.00         0

    accuracy                           0.16     46296
   macro avg       0.10      0.05      0.06     46296
weighted avg       0.44      0.16      0.23     46296

accuracy 0.16297304302747537
```

Figure 30: GaussianNB District Wise Report

9.5.3 UCR WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 65, min f1-score of 0 and accuracy of 48% [fig31].

```
mean: 0.21940605534341515
max: 0.6574243577349487
min: 0.0
              precision    recall  f1-score   support

         1.0       0.00      0.00      0.00         0
         2.0       0.00      0.57      0.00         7
         3.0       1.00      0.49      0.66     32633

    accuracy                           0.49     32640
   macro avg       0.33      0.35      0.22     32640
weighted avg       1.00      0.49      0.66     32640

accuracy 0.48973651960784315
```

Figure 31: GaussianNB UCR Wise Report

9.6 RandomForestClassifier

9.6.1 CRIME WISE

To check the Random Forest Classifier algorithm, the dataset is divided into 5 folds of 90 candidates and the possible combination for testing is 450 fits. After testing, the accuracy is 24% [FIG32].

```
Fitting 5 folds for each of 90 candidates, totalling 450 fits

GridSearchCV(cv=5, estimator=RandomForestClassifier(), n_jobs=-1,
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_depth': [4, 5, 6, 7, 8],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'n_estimators': [200, 500, 100]},
             verbose=3)

CV_rfc.best_params_
{'criterion': 'entropy',
 'max_depth': 8,
 'max_features': 'log2',
 'n_estimators': 500}

Best score is 0.2379797559262054
```

Figure 32: RFC Crime Wise Report

9.6.2 DISTRICT WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 21, min f1-score of 01 and accuracy of 15% [fig33].

```
mean: 0.11106690871038642
max: 0.2121326584573544
min: 0.018880647336480108
              precision    recall  f1-score   support

         1.0       0.17      0.16      0.16      5549
         2.0       0.05      0.07      0.05      1792
         3.0       0.24      0.19      0.21      8970
         4.0       0.04      0.06      0.05      1276
         5.0       0.09      0.11      0.10      2727
         6.0       0.07      0.09      0.08      2399
         7.0       0.05      0.07      0.06      1784
         8.0       0.20      0.16      0.18      7455
         9.0       0.23      0.20      0.21      7199
        10.0       0.03      0.05      0.04      1171
        11.0       0.19      0.17      0.18      5476
        12.0       0.01      0.03      0.02       498

    accuracy                           0.15     46296
   macro avg       0.11      0.11      0.11     46296
weighted avg       0.17      0.15      0.16     46296

accuracy 0.152172973907033
```

Figure 33: RFC District Wise Report

9.6.3 UCR WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 63, min f1-score of 28 and accuracy of 50% [fig34].

```
mean: 0.43447849609494815
max: 0.6345228711983257
min: 0.28053272881836216
              precision    recall  f1-score   support

         1.0       0.23      0.37      0.28      4002
         2.0       0.35      0.44      0.39      7829
         3.0       0.73      0.56      0.63     20809

    accuracy                           0.51     32640
   macro avg       0.43      0.46      0.43     32640
weighted avg       0.58      0.51      0.53     32640

accuracy 0.5096200980392157
```

Figure 34: RFC UCR Wise Report

9.7 LGBMClassifier

9.7.1 CRIME WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 27, min f1-score of 0 and accuracy of 19% [fig35].

```
mean: 0.10164643167919406
max: 0.274986638161411
min: 0.002018163471241171
              precision    recall  f1-score   support

         1.0       0.11      0.22      0.15      2685
         2.0       0.00      0.15      0.01        78
         3.0       0.56      0.18      0.27     22581
         4.0       0.01      0.17      0.02       149
         5.0       0.02      0.26      0.04       329
         6.0       0.02      0.28      0.04       235
         7.0       0.00      0.09      0.00        35
         8.0       0.15      0.17      0.16      5213
         9.0       0.34      0.22      0.27      9316
        10.0       0.01      0.32      0.01        34
        11.0       0.25      0.22      0.24      5635
        12.0       0.00      0.17      0.00         6

    accuracy                           0.20     46296
   macro avg       0.12      0.20      0.10     46296
weighted avg       0.40      0.20      0.24     46296

accuracy 0.19682045965094178
```

Figure 35: LGBMC Crime Wise Report

9.7.2 DISTRICT WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 27, min f1-score of 0 and accuracy of 19% [fig36].

```
mean: 0.10164643167919406
max: 0.274986638161411
min: 0.002018163471241171
              precision    recall  f1-score   support

         1.0       0.11      0.22      0.15      2685
         2.0       0.00      0.15      0.01        78
         3.0       0.56      0.18      0.27     22581
         4.0       0.01      0.17      0.02       149
         5.0       0.02      0.26      0.04       329
         6.0       0.02      0.28      0.04       235
         7.0       0.00      0.09      0.00        35
         8.0       0.15      0.17      0.16      5213
         9.0       0.34      0.22      0.27      9316
        10.0       0.01      0.32      0.01        34
        11.0       0.25      0.22      0.24      5635
        12.0       0.00      0.17      0.00         6

    accuracy                           0.20     46296
   macro avg       0.12      0.20      0.10     46296
weighted avg       0.40      0.20      0.24     46296

accuracy 0.19682045965094178
```

Figure 36: LGBMC District Wise Report

9.7.3 UCR WISE

If the x-tree has UCR data which contains the reporting time, it gives a max f1-score of 66, min f1-score of 17 and accuracy of 51% [fig37].

```
mean: 0.33500177492264616
max: 0.6605723370429253
min: 0.17102258310073584
              precision    recall  f1-score   support

         1.0       0.10      0.52      0.17      1297
         2.0       0.10      0.51      0.17      2039
         3.0       0.94      0.51      0.66     29304

    accuracy                           0.51     32640
   macro avg       0.38      0.52      0.34     32640
weighted avg       0.85      0.51      0.61     32640

accuracy 0.5110906862745098
```

Figure 37: LGBMC UCR Wise Report

9.8 LSTM ARCHITECTURE

9.8.1 CRIME WISE

The LSTM architecture is calculated by the value loss and value accurate rate where the lowest time is taken 423ms/steps and the data was divided into 11912 steps of 4 times. Here the accuracy of crime wise data is 28% and the lowest loss is 1.32% [FIG38].

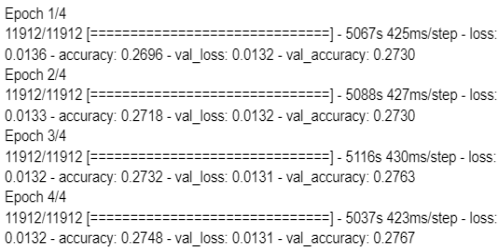


Figure 38: LSTM Crime Wise Report

Chapter 10

TERMS OF RESULT

10.1 WEIGHTED AVERAGE

Weighted arithmetic mean is like arithmetic mean. But the difference is in arithmetic mean, all data points contribute equally to the final average, but here some data points have a significant contribution than others. The idea of weighted mean assumes a part in graphic insights and furthermore happens in a broader structure in a few different spaces of arithmetic. The mathematical equation for this term is given here as a figure. [fig39]

$$\sigma_{\bar{x}} = \left(\sqrt{\sum_{i=1}^{n} w_i}\right)^{-1}$$

Figure 39: Weighted AVG

10.2 PRECISION AND RECALL

Precision or positive predictive value is the fraction of relevant instances among the retrieved instances and Recall or Sensitivity is the relevant instances that were retrieved. Both are related to relevance. The mathematical equation for these terms are

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Figure 40: Precision & Recall

10.3 F-SCORE

The F-score or F-measure is the measurement of a test's accuracy is basically the calculation of true positive results divided by the number of total positive results. The mathematical equation for this term is here.

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}.$$

Figure 41: F-Score

10.4 MACRO AVERAGE

Macro-average is the computation of the independent metric for each class and then take the average by treating all classes equally The mathematical equation for this term is:

$$\text{Macro-average precision} = \frac{P1+P2}{2}$$
$$\text{Macro-average recall} = \frac{R1+R2}{2} = .$$

Figure 42: Macro AVG

Chapter 11

TOOLS, SOFTWARE AND LIBRARIES

11.1 ANACONDA

For Python data science we need a specialized application. With this in mind, Anaconda was born and developed by Anaconda Inc. It is used for Python and R programming languages for scientific computing, for example, data science, machine learning applications, data processing, Proportional data, Predictive analysis, etc. so that it can simplify package management and deployment.

11.2 JUPYTER NOTEBOOK

Jupyter Notebook is, it is made for creating and sharing documents that contain live code, equations, visualizations, and narrative text in JSON format. The use cases are data cleaning and

transformation, numerical simulation, statistical modeling, data visualization, machine learning, and far more.

11.3 SPYDER

Spyder is an incredible logical structure written in Python for Python, created and created by researchers, designers, and data analysts. It gives a one-of-a-kind mix of advanced editing, analysis, debugging, and profiling functionality capacities in exhaustive data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

11.4 POWERSHELL

PowerShell is a computerized and ordering system consisting of an ordering line shell and a corresponding pre-ordering language. It is available for Windows, macOS and Linux.

11.5 NUMPY

Numpy is a Python bundle utilized for logical calculations. It likewise comprises N-dimensional exhibit objects to work with. It likewise proves to be useful when utilizing straight polynomial math, Fourier changes, and different irregular worth capacities. Moreover, Numpy likewise goes about as a multi-dimensional holder for a wide range of information and articles. Subjective kinds of information can likewise be depicted making it simpler for Numpy to incorporate with data sets.

11.6 PANDAS (Python Data Analysis Library)

Pandas is a library of software. The Python programming language needs this library for information control and analysis. In Addition, it gives a fluent work path for information designs and mathematical tables controlling activities and time arrangements. It is free programming delivered under the three-proviso BSD permit

11.7 MATPLOTLIB

Matplotlib is a library of python which is utilized for 2D plotting. It is utilized in the production of value figures and diagrams. Plots, histograms, power spectra, dispersed plots, bar diagrams can be made utilizing Matplotlib.

11.8 SCIKIT-LEARN

Scikit-learn is a library in Python that empowers utilizing a few unaided and regulated learning calculations. It is based on top of NumPy, pandas, and matplotlib. Scikit-learn gives various capacities including relapse, arrangement, bunching, model choice, and preprocessing.

11.9 GOOGLE COLLAB

Google Colab (Collaboration) is an item from Google Exploration. Colab permits anyone to compose and execute discretionary python code through the program and is particularly appropriate to AI, information examination, and instruction.

11.10 TENSORFLOW

Tensor Flow is created for the purpose of the machine learning platform. It features end-to-end utilization and is open-source. It has a thorough, adaptable environment of instruments, libraries, and local area assets that allows analysts to push the best in class in ML and engineers effectively assemble and convey ML-fueled applications.

11.11 DOCKER

For OS-level virtualization, a set of platforms is a service product that delivers software packages as the container. This gives the idea of Docker. These containers prohibit contact with each other directly and group their software, libraries, and set up records; they can speak with one another through obvious channels.

11.12 MPL_TOOLKITS

mpl_toolkits are an assortment of aide classes. The matplotlib needs to display multiple images. To make it easier, mpl_toolkits has been introduced.

Chapter 12

DATA ANALYSIS

12.1 COLUMN WITH NULL DATA

There is a number of null data present in the dataset. As null data cannot be used in the training process, those have to be popped out from the data column. For that null data are defined out and shown in fig43.

INCIDENT_NUMBER	0
OFFENSE_CODE	0
OFFENSE_CODE_GROUP	112339
OFFENSE_DESCRIPTION	0
DISTRICT	2238
REPORTING_AREA	0
SHOOTING	351798
OCCURRED_ON_DATE	0
YEAR	0
MONTH	0
DAY_OF_WEEK	0
HOUR	0
UCR_PART	112436
STREET	11207
Lat	22530
Long	22530
Location	0

Figure 43: Null Data

These data are found by calling all the possible fit from the dataset which is (465592, 17). Then all the fits are searched by their index name and value. Later then, the fields with null data are found.

To manipulate data by area, wise some districts are created for mark down. These will be used for plotting the crime area, finding the area where crime may occur and for calculating the accuracy for crime occurrence. District codes are ['B2', 'C11', 'A1', 'E18', 'D4', 'B3', nan, 'C6', 'D14', 'A7', 'E5', 'E13', 'A15', 'External'] whereas the crime data distribution for the district and months is shown in the table 4.

However, to find out when the crime may occur, twelve (12) months are taken in integer value, seven (7) days of week are taken in this orientation ['Friday', 'Thursday', 'Monday', 'Tuesday', 'Sunday', 'Wednesday', 'Saturday'] and time of crime occurrence is mentioned in zero (0) to twenty-three (23) orientation.

District / Month	1	2	3	4	5	6	7	8	9	10	11	12
B2	5466	4935	5656	5346	5132	5766	6438	6461	6603	7392	6664	6605
D4	4281	4196	4509	4571	4545	4899	5594	54776	5538	6247	5761	5798
C11	4784	4335	4703	4377	4131	4815	5492	5391	5478	6098	5647	5511
A1	3833	3489	3757	3842	3832	4248	4604	4856	4739	5553	4900	4790
B3	3948	0	3949	3519	3452	3736	4573	4864	4506	5244	4849	4608
C6	2580	2427	2752	2639	2223	2673	2971	3096	2981	0	3205	3123
D14	2204	2059	2227	2092	2144	2318	2553	2634	2810	2683	2751	2735
E13	0	1899	1954	1956	1838	2128	2212	2244	2267	2532	2456	27486
E18	1960	1861	1559	1883	1836	2028	2193	2240	1717	3271	2441	2338
E5	1655	1397	1456	1480	1404	1556	1704	1813	0	2077	2030	1879
A7	1477	1407	763	1465	1326	1545	1660	1766	1649	1989	0	1768
A15	721	670	0	736	679	816	824	816	816	993	1721	925
External	40	43	22	25	27	31	24	29	19	45	25	41

Table 4: District & Month’s Data

12.2 CRIME AREA MAP

The dataset for machine learning is taken from “BOSTON POLICE DEPARTMENT (BPD)” which is a real dataset based on the initial details surrounding an incident to which BPD officers respond. This is a dataset containing records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. Records in the new system begin in June of 2015. The crime area map is shown in fig44.

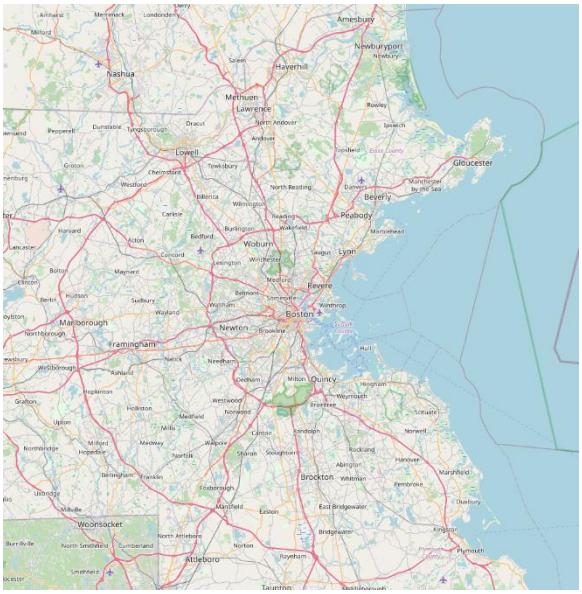


Figure 44: Data Map

12.3 CATEGORIZING BY MAPPING

Total response number of crime occurrence is calculated by the head name and its value. After that the values are summed up and shown in the fig45.

Motor Vehicle Accident Response	41064
Larceny	28906
Medical Assistance	26211
Investigate Person	20420
Other	19860
Drug Violation	18156
Simple Assault	17531
Vandalism	16862
Verbal Disputes	14470
Towed	12453
Investigate Property	12372
Larceny From Motor Vehicle	11889
Property Lost	11043
Warrant Arrests	9728
Aggravated Assault	8751

Figure 45: Data Category & Data Count

The data are taken and the crime area map is generated by the Boston longitude and latitude and by putting all the crime data in a map by category, a clustered map is generated containing all the crime in a dot formation on the whole area shown in fig46.

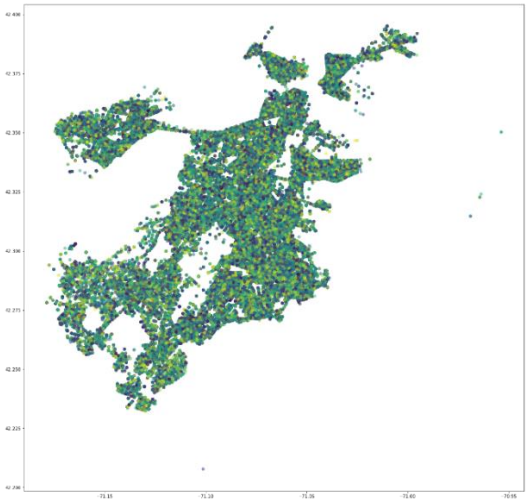


Figure 46: Plotting Data on map

12.4 CRIME OCCURRENCE TIME

The crime happening time is divided into two (2) segments, day and night. The main calculation is done by taking the hour. For example, for January, if the hour value is less than 6 and greater than 18, that is taken as night; where for August, hour is taken if the value is less than 5 and greater than 21; that is called night. Here all the data are oriented by their count and plotted in a diagram [fig47] where “0” is representing the night scenario and “1” is for day crimes.

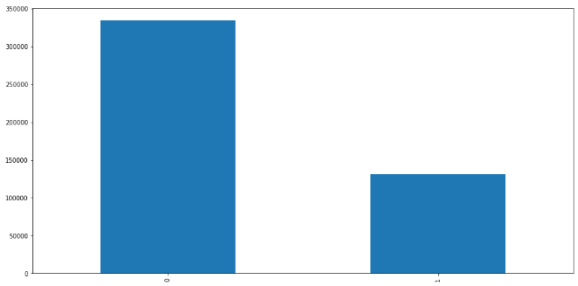


Figure 47: Crime Rate Depending on Time

12.5 K-MEANS MAPPING

The K-Means mapping is created based on a centroid which is a data point at the center of a cluster. The resulting classifier is used to classify (using k = 1 or any number) the data and thereby produce an initial randomized set of clusters. Each centroid is thereafter set to the arithmetic mean of the cluster it defines.

Here the cluster is shown in 2, 3, 5 and 10 classified centroids shown in the figure accordingly.

12.5.1 2-CLUSTER

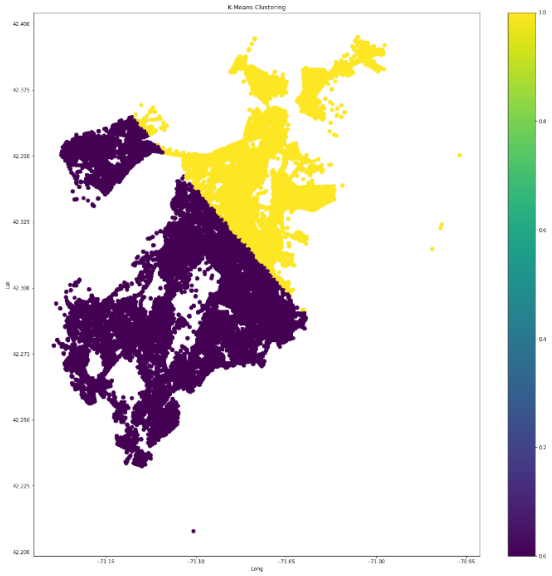


Figure 48: 2-Cluster

12.5.2 3-CLUSTER

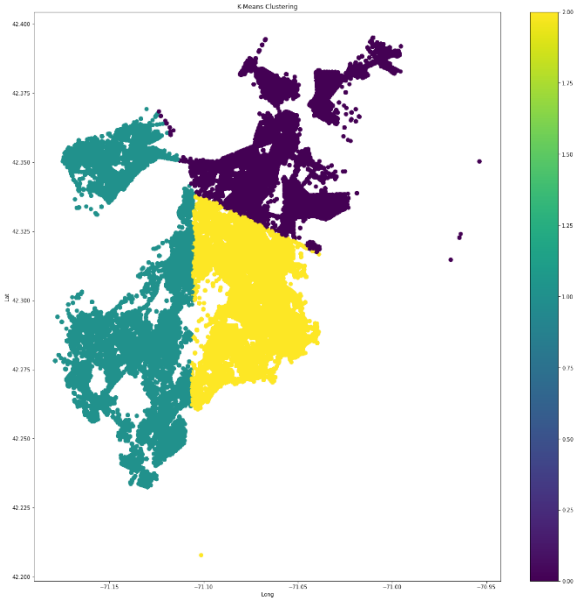


Figure 49: 3-Cluster

12.5.3 5-CLUSTER

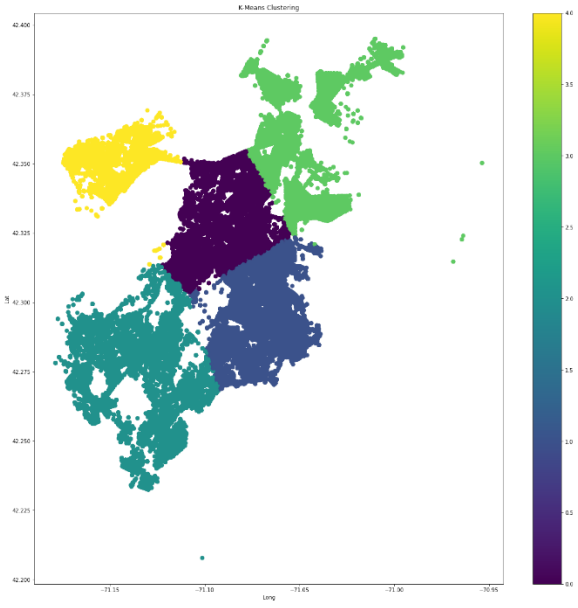


Figure 50: 5-Cluster

12.5.4 10-CLUSTER

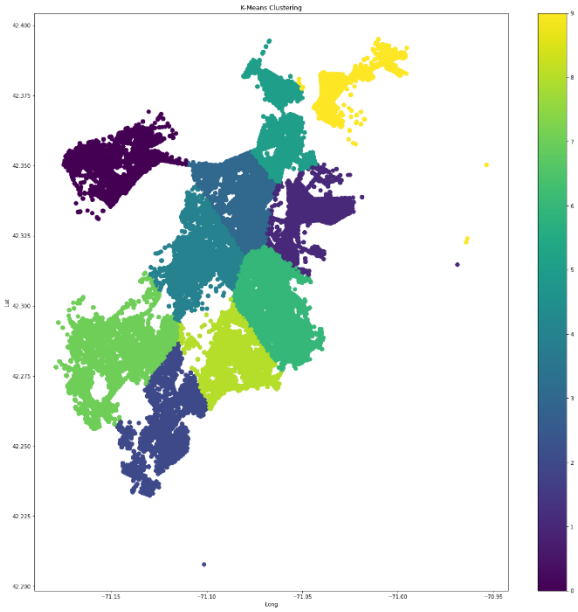


Figure 51: 10-Cluster

12.6 CLUSTERING WITH LOCATION AND OFFENSE CODE DATA

By calculating the data count, means, std, min, max with the Latitude and Longitude, a data format fig [52] is created by using that a visualization is generated for detailed understanding of the topic.

	OFFENSE_CODE	Long	Lat
count	439461.000000	439461.000000	439461.000000
mean	2304.839852	-71.083170	42.322190
std	1196.088318	0.030067	0.031968
min	111.000000	-71.178674	42.207760
25%	1001.000000	-71.097718	42.297555
50%	3001.000000	-71.077647	42.326146
75%	3201.000000	-71.062131	42.348577
max	99999.000000	-70.953726	42.395042

This 2 K-Means clustering is showing a plot for the max and the min value of crime scenario.

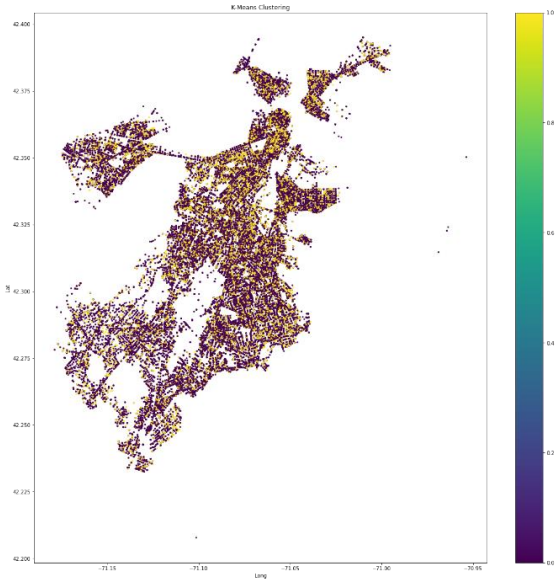


Figure 52: Clustering with Location & Offense Code's Data

12.8 CLUSTERING WITH LOCATION AND MONTHS

This plot is showing scatter data of crime occurrences on behalf of the month's data.

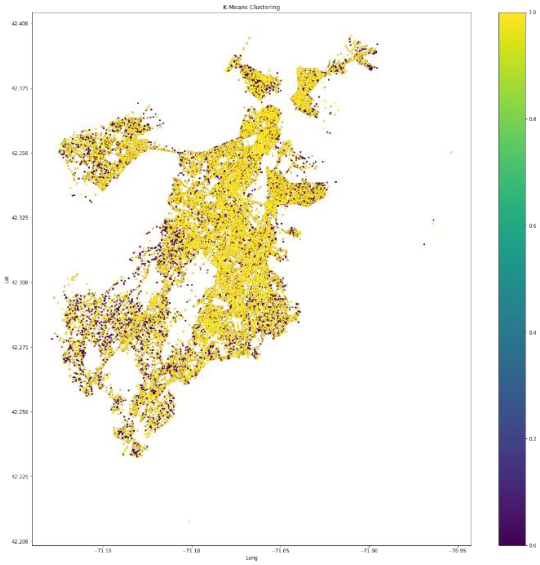


Figure 53: Clustering with Location & Month's Data

12.9 Dataset Shape

The total set of data and column number

```
###################### Shape ######################
(465592, 17)
```

Figure 54: Data Shape

12.10 DATA TYPES

Data types mentioned in the columns are shown in the [fig55].

```
###################### Types ######################
INCIDENT_NUMBER         object
OFFENSE_CODE             int64
OFFENSE_CODE_GROUP      object
OFFENSE_DESCRIPTION     object
DISTRICT                object
REPORTING_AREA          object
SHOOTING                object
OCCURRED_ON_DATE        object
YEAR                     int64
MONTH                    int64
DAY_OF_WEEK             object
HOUR                     int64
UCR_PART                object
STREET                  object
Lat                    float64
Long                   float64
Location                object
```

Figure 55: Types of Data

12.11 DATASET HEAD

Headcount details of the dataset is shown here in the figure 56.

```
#################### Head ####################
  INCIDENT_NUMBER  OFFENSE_CODE OFFENSE_CODE_GROUP    OFFENSE_DESCRIPTION  \
0      I192074488           619            Larceny     LARCENY ALL OTHERS
1      I192073511          1107              Fraud  FRAUD - IMPERSONATION
2      I192073187          2629         Harassment             HARASSMENT

  DISTRICT REPORTING_AREA SHOOTING     OCCURRED_ON_DATE  YEAR  MONTH  \
0      C11            351      NaN  2017-08-23 00:00:00  2017      8
1       B3            435      NaN  2017-02-21 00:01:00  2017      2
2       C6            231      NaN  2017-07-13 00:00:00  2017      7

  DAY_OF_WEEK  HOUR  UCR_PART        STREET        Lat       Long  \
0   Wednesday     0  Part One      ADAMS ST  42.300605 -71.059230
1     Tuesday     0  Part Two  ARMANDINE ST  42.284315 -71.074108
2    Thursday     0  Part Two    E FIFTH ST  42.333989 -71.032606

                    Location
0  (42.30060526, -71.05923027)
1  (42.28431486, -71.07410838)
2  (42.33398889, -71.03260574)
```

Figure 56: Dataset Head

12.12 DATASET TAIL

Tail count with its data from the dataset is shown in the figure 57.

```
#################### Tail ####################
       INCIDENT_NUMBER  OFFENSE_CODE OFFENSE_CODE_GROUP  \
465589   I080542626-00          1848     Drug Violation
465590   I030217815-08          3125    Warrant Arrests
465591   I030217815-08           111           Homicide

                                OFFENSE_DESCRIPTION DISTRICT  \
465589  DRUGS - POSS CLASS B - INTENT TO MFR DIST DISP       A1
465590                                WARRANT ARREST      E18
465591            MURDER, NON-NEGLIGIENT MANSLAUGHTER      E18

       REPORTING_AREA SHOOTING     OCCURRED_ON_DATE  YEAR  MONTH DAY_OF_WEEK  \
465589            111      NaN  2015-08-12 12:00:00  2015      8   Wednesday
465590            520      NaN  2015-07-09 13:38:00  2015      7    Thursday
465591            520      NaN  2015-07-09 13:38:00  2015      7    Thursday

        HOUR    UCR_PART       STREET        Lat       Long  \
465589    12    Part Two  BOYLSTON ST  42.352312 -71.063705
465590    13  Part Three     RIVER ST  42.255926 -71.123172
465591    13    Part One     RIVER ST  42.255926 -71.123172

                       Location
465589  (42.35231190, -71.06370510)
465590  (42.25592648, -71.12317207)
465591  (42.25592648, -71.12317207)
```

Figure 57: Dataset tail

12.13 NULL DATA

All the null data fields in the dataset are shown in figure 58.

```
###################### NA ######################
INCIDENT_NUMBER             0
OFFENSE_CODE                0
OFFENSE_CODE_GROUP     112339
OFFENSE_DESCRIPTION         0
DISTRICT                 2238
REPORTING_AREA              0
SHOOTING               351798
OCCURRED_ON_DATE            0
YEAR                        0
MONTH                       0
DAY_OF_WEEK                 0
HOUR                        0
UCR_PART               112436
STREET                  11207
Lat                     22530
Long                    22530
Location                    0
```

Figure 58: NaN Fields

12.14 QUANTILES

Here is a visualization of quantiles as given in the figure 59.

```
#################### Quantiles ####################
                    0.00         0.05         0.50         0.95         0.99  \
OFFENSE_CODE   111.000000   522.000000  3005.000000  3831.000000  3831.000000
YEAR          2015.000000  2015.000000  2017.000000  2020.000000  2021.000000
MONTH            1.000000     1.000000     7.000000    12.000000    12.000000
HOUR             0.000000     0.000000    14.000000    22.000000    23.000000
Lat             -1.000000    42.265685    42.325538    42.371422    42.382801
Long          -71.178674   -71.146980   -71.077370   -71.039291   -71.004155

                    1.00
OFFENSE_CODE  9.999900e+04
YEAR          2.021000e+03
MONTH         1.200000e+01
HOUR          2.300000e+01
Lat           4.239504e+01
Long          5.249691e-08
```

Figure 59: Quantiles

12.15 MISSING DATA

Here is a plot of missing data in the dataset where all the white space in the plot is representing data in the column which is missing.

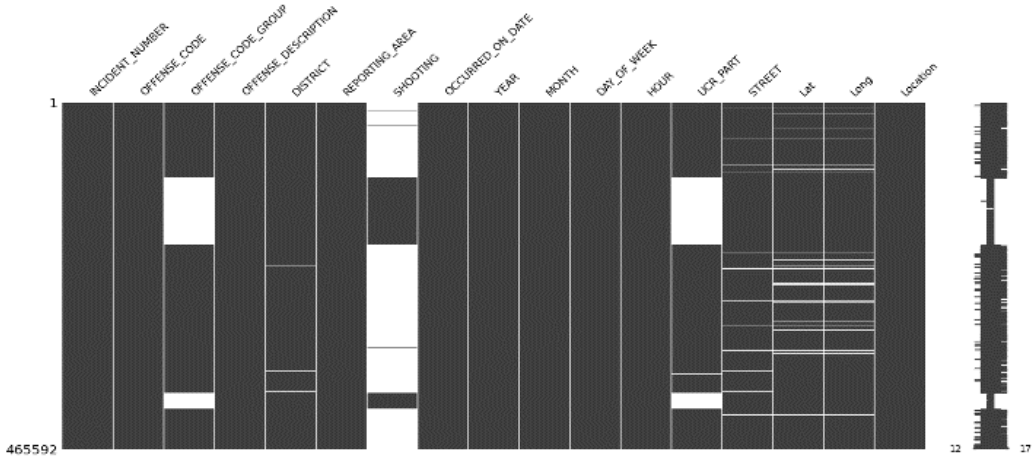


Figure 60: Missing Data

12.16 DATA COUNT

The plot in the fig [61] is showing the count of data according to its type available in the 17 columns in the dataset.

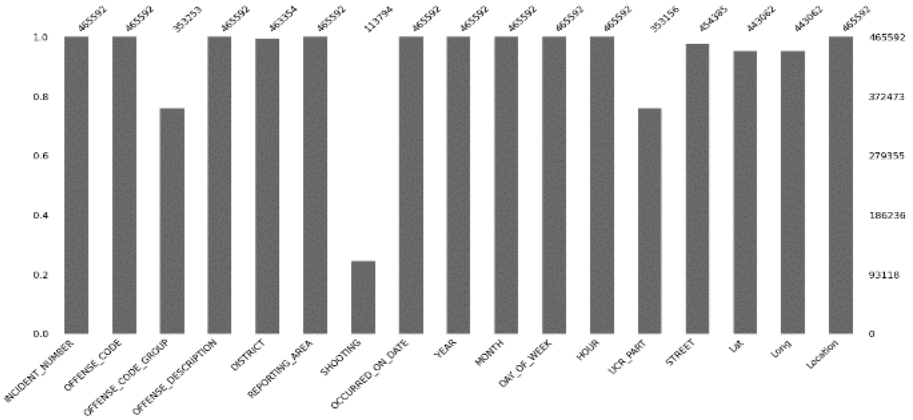


Figure 61: Data Count

12.17 DATA MODIFICATION

In terms of making the work process easier, the dataset is modified into a simple way which is easier to understand for both humans and machines. [Table 5]

	INCIDENT_NUMBER	Group	Description	District	REPORTING_AREA	Date	Year	Month	Day	Hour
0	I192074488	Larceny	LARCENY ALL OTHERS	C11	351	2017-08-23 0:00:00	2017	8	Wednesday	0
1	I192073511	Fraud	FRAUD - IMPERSONATION	B3	435	2017-02-21 0:01:00	2017	2	Tuesday	0
2	I192073187	Harassment	HARASSMENT	C6	231	2017-07-13 0:00:00	2017	7	Thursday	0
3	I192072907	Simple Assault	ASSAULT SIMPLE - BATTERY	D4	171	2017-09-01 16:32:00	2017	9	Friday	16
4	I192072900	Fraud	FRAUD - FALSE PRETENSE / SCHEME	D4	151	2017-09-01 0:00:00	2017	9	Friday	0

UCR_PART	Street	Lat	Long	dayofweek	quarter	dayofyear	dayofmonth	weekofyear
Part One	ADAMS ST	42.3000605	-71.059923	2	3	235	23	3
Part Two	ARMANDINE ST	42.284315	-71.074108	1	1	52	21	8
Part Two	E FIFTH ST	42.333989	-71.032606	3	3	194	13	28
Part Two	HARRISON AVE	42.335119	-71.074917	4	3	244	1	35
Part Two	WARREN AVE	42.345163	-71.071291	4	3	244	1	35

Table 5: Modified Data

12.18 UNIQUE DATA

This table is showing the unique data that the dataset holds into it in types.

INCIDENT_NUMBER	423519
Group	67
Description	292
District	13
REPORTING_AREA	881
Date	346928
Year	7
Month	12
Day	7
Hour	24
UCR_PART	4
Street	12027
Lat	33018
Long	33018
Location	33034
dayofweek	7
quarter	4
dayofyear	366
dayofmonth	31
weekofyear	53

Figure 62: Unique Data

12.19 MONTHLY DISTRIBUTION OF CRIMES

This plot is showing the crimes in monthly orientation. Here it is visible that June has a higher crime rate of 10% and October, November and December have lower rate of crime which is 7%.

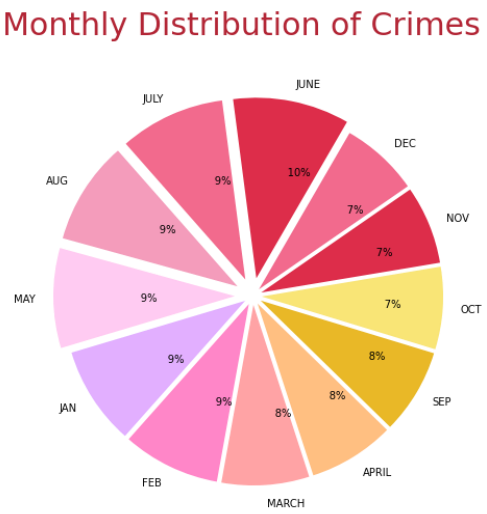


Figure 63: Month Wise Data

12.20 YEARLY DISTRIBUTION OF CRIME

In this diagram, the crime is distributed in yearly orientation where 2016 has a 21.1% crime rate which is higher and 2020 has a lower crime rate of 5.1%. 2021 is not over yet; for that cannot be taken as the lower rate of crime.

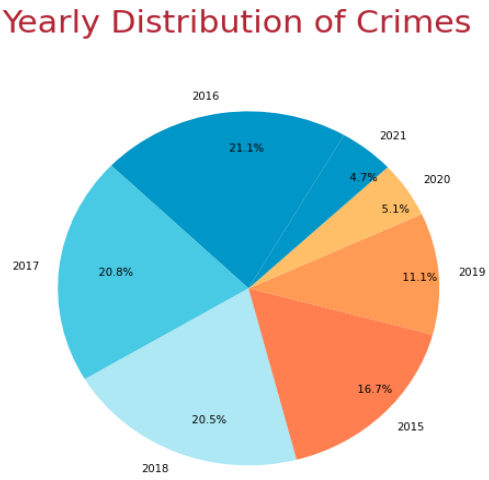


Figure 64: Year Wise Data

12.21 CATEGORY WISE CRIME COUNT

This plot describes the crime of motor vehicle accident response, larceny, medical assistant group, investigate person and external crime in the district orientation.

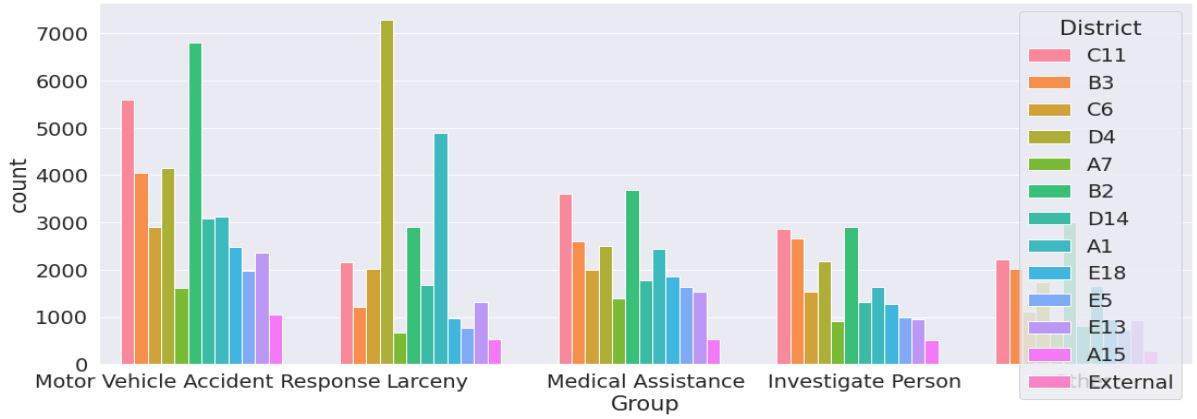


Figure 65: Category Wise Crime Count

12.22 DISTRICT WISE MONTHLY CRIME

This plot represents the crime in monthly orientation but in box diagram format.

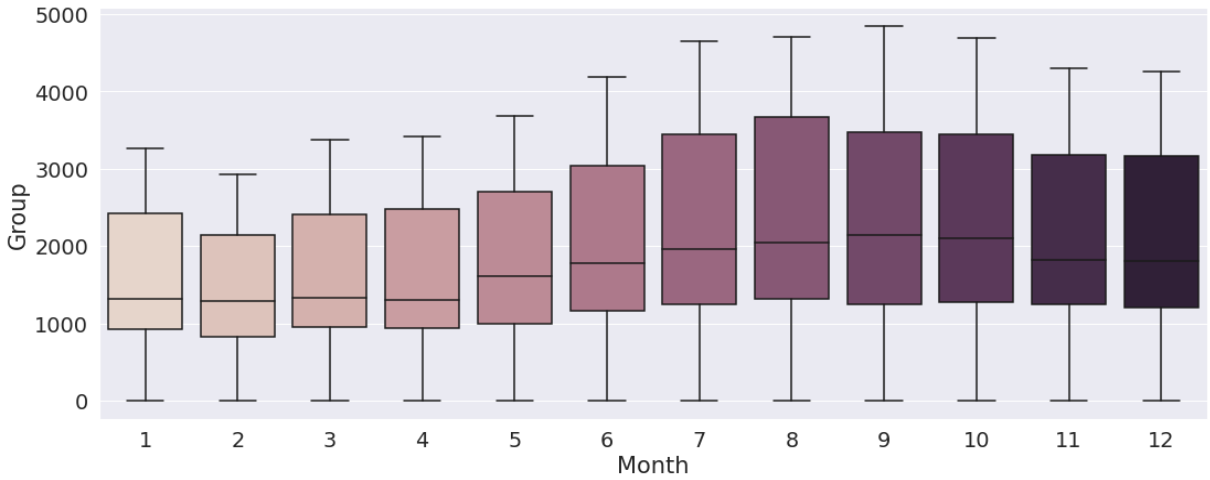


Figure 66: District Wise data

12.23 OVERALL CRIME COUNT

The chart shows the overall crime count of all the crime categories complying with the number.

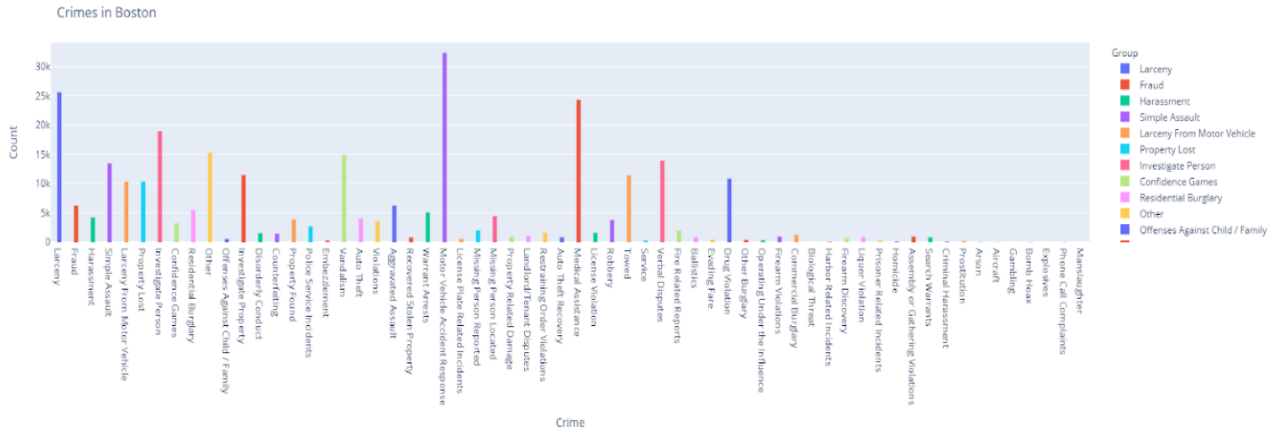


Figure 67: Overall Crime Count

12.24 UCR AND DISTRICT DISTRIBUTION

Crime data diagram in UCR and district orientation which shows what crime occurred in which district and what is its UCR count.

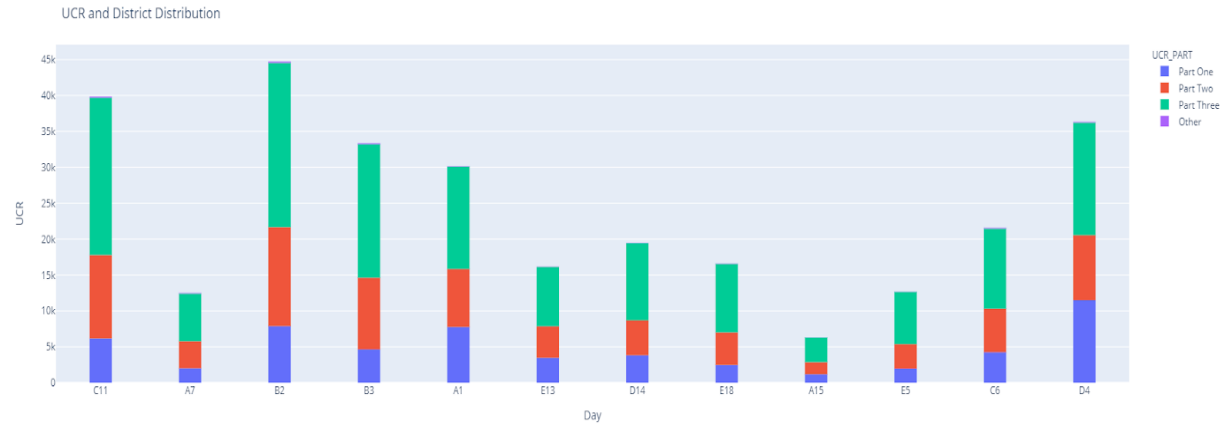


Figure 68: UCR & District Data

12.25 DAY AND DISTRICT DISTRIBUTION

This part is showing the number of crime occurrences in each day of a week district wise.

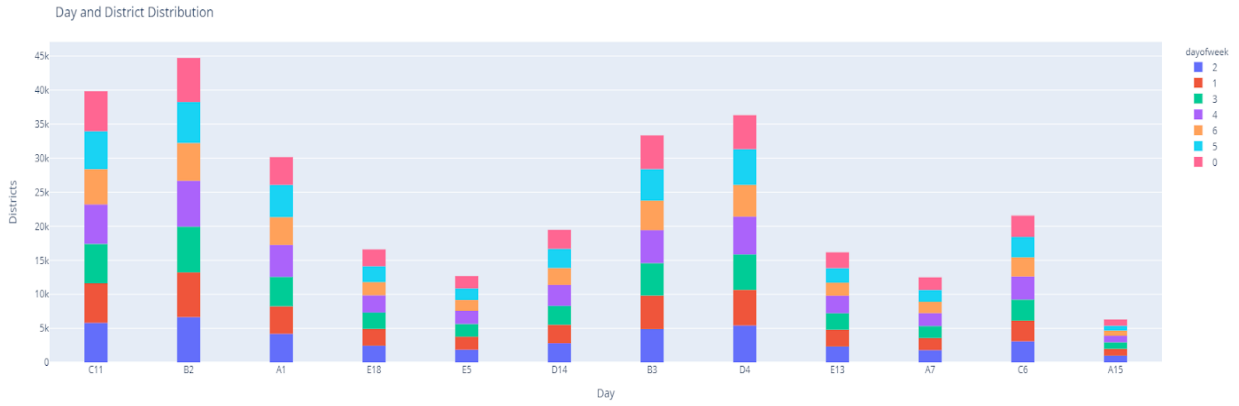


Figure 69: Day & District Data

12.26 CRIME DETECTION RESULT FORMAT

This plot will help to check the previous data for crime records by crime group, district, reporting area, time, and location for the crime in a table format. [Table 6]

	Group	District	REPORTING_AREA	Year	Month	dayofweek	Hour	Street	Lat	Long
0	0	0	0	2017	8	0	0	0	42.300605	-71.059230
1	1	1	1	2017	2	1	0	1	42.284315	-71.074108
2	2	2	2	2017	7	2	0	2	42.333989	-71.032606
3	3	3	3	2017	9	3	16	3	42.335119	-71.074917
4	1	3	4	2017	9	3	0	4	42.345163	-71.071291

Table 6: Data Format

12.27 CRIME HEAT MAP TIME TO TIME VIEW OF BOSTON

Here the heat map of crime data for motor vehicle response is shown from June 2015 to May 2021 from the dataset. If anyone follows carefully it will be visible the crime rate and its change rate from the diagrams [fig 70].

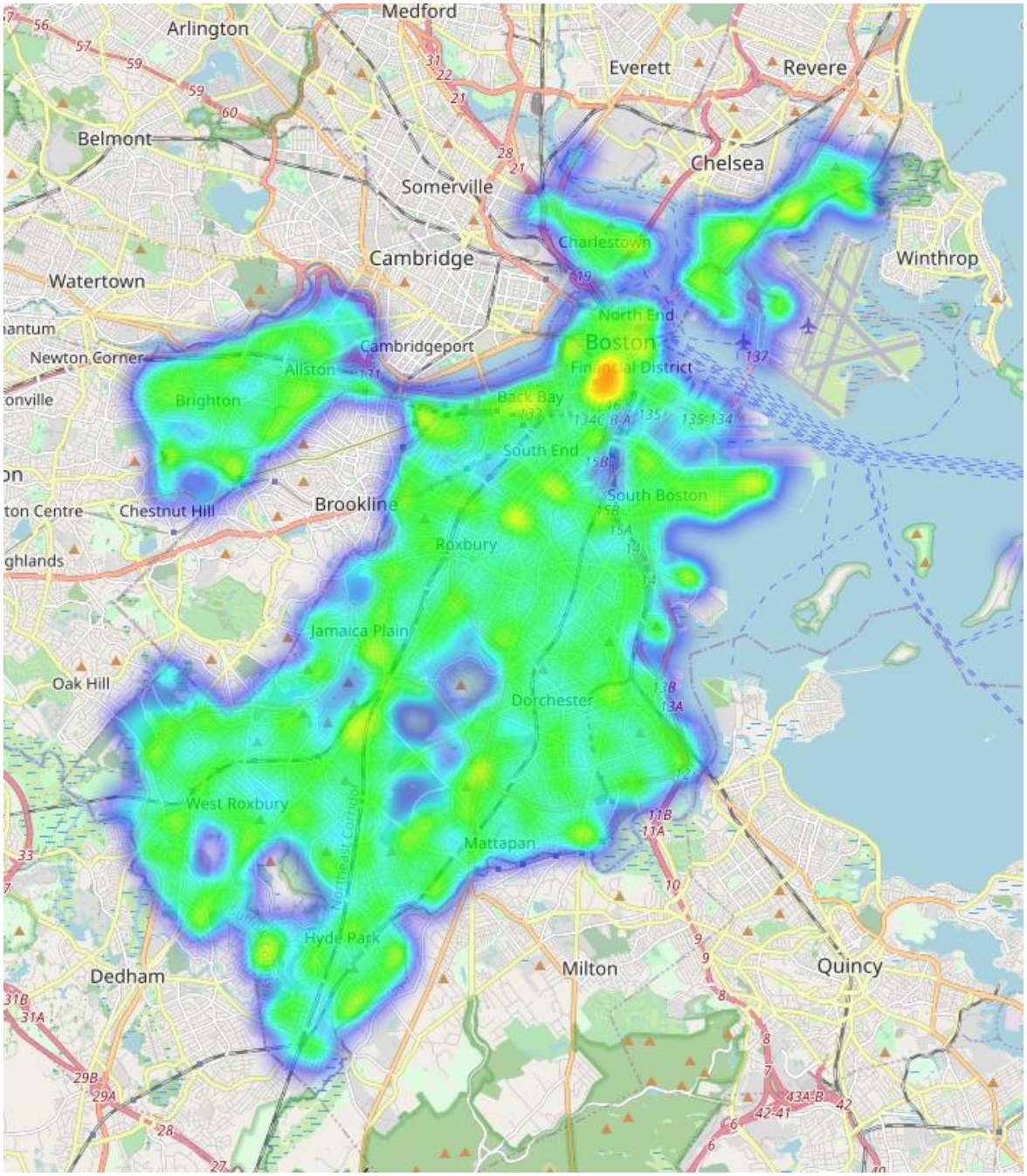


Figure 70: Motor Vehicle Response Heat Map

Chapter 13

RESULT ANALYSIS

13.1 CRIME WISE ACCURACY

By analyzing the accuracy data of the calculation, it is visible that for predicting data crime-wise, LSTM Architecture is giving the best accuracy of 28%. [Table 7]

Algorithm Name	Max F1	Min F1	Accuracy
DecisionTreeClassifier	48	10	22%
BernoulliNB	26	0	16%
ExtraTreeClassifier	47	9	21%
KNeighborsClassifier	37	7	22%
GaussianNB	26	0	15%
RandomForestClassifier	-	-	24%
LGBMClassifier	27	0	19%
LSTM Architecture	-	-	28%

Table 7: Crime Wise Result

13.2 DISTRICT WISE ACCURACY

In district-wise location prediction, the LGBMClassifier is giving the best accuracy and it is 19%.

[Table 8]

Algorithm Name	Max F1	Min F1	Accuracy
DecisionTreeClassifier	21	1	15%
BernoulliNB	27	0	16%
ExtraTreeClassifier	21	1	15%
KNeighborsClassifier	21	1	15%
GaussianNB	26	0	16%
RandomForestClassifier	21	1	15%
LGBMClassifier	27	0	19%
LSTM Architecture	-	-	-

Table 8: District Wise Result

13.3 UCR WISE ACCURACY

For predicting the UCR wise crime occurrence time, LGBMClassifier is giving the highest accuracy of 51%. [Table 9]

Algorithm Name	Max F1	Min F1	Accuracy
DecisionTreeClassifier	45	30	45%
BernoulliNB	66	0	49%
ExtraTreeClassifier	54	28	44%
KNeighborsClassifier	54	29	44%
GaussianNB	65	0	48%
RandomForestClassifier	63	28	50%
LGBMClassifier	66	17	51%
LSTM Architecture	-	-	-

Table 9: UCR Wise Result

13.4 FINALIZING ALGORITHM

By analyzing and calculating the accuracy rate, it is giving us a clear result that to predict data for crime detection by crime area, type and time; the LGBMClassifier, LSTM Architecture and LGBMClassifier are needed to be chosen for the best solution.

Chapter 14

OBSTACLES DURING THE MACHINE LEARNING

While doing the test cases, a human has to face some back-logs which are the main findings as a problem in the report writing.

14.1 LACK OF HIGH CONFIGURATION COMPUTER

To go well with the accuracy test, a minimum of 4.2 GHz processor and 32 GB of RAM is needed. In lack of this configuration, the calculation and learning process may get hung and sometimes it crashes.

14.2 TIME SHORTAGE AND COMPLEXITY

Machine training needs a large amount of time. Due to shortage of time, all the algorithms may not be able to be tested and where as it took so much time to calculate which may not be possible in a short period of time.

Chapter 15

CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

15.1 IMPLEMENTATION OF NLP

Though it was one of the prime goals of the research work, the NLP has not been implemented yet. We have a plan to take data from the daily newspaper every day. Every day an AI bot will read all daily newspapers and categorize all the news. Then its main job will be taking the crime

data and its associates. NLP enables more natural conversations, more efficient operations, reduced costs, higher customer satisfaction, and improved analysis. Creating an NLP Chabot or adding NLP capabilities to your existing Chabot is easier than ever before, with the advantages vastly outweighing the associated costs. NLP executed methods can be extremely helpful in a law requirement climate, particularly when unstructured and a lot of information should be prepared by criminal specialists.

As of now generally utilized procedures like OCR[6] and machine interpretations can be effectively utilized in criminal examinations. For instance, OCR is utilized in misrepresentation examinations to consequently change unstructured solicitations and other monetary papers into accessible and totaled spreadsheets. In the past, harder-to-execute strategies like the programmed outline of writings, data extraction, substance extraction, and relationship ex-foothold are currently coming into reach of law implementation and knowledge organizations. This is masculine so due to the decrease in cost per preparing unit and the way that these strategies need a lot of handling ability to have the option to utilize adequately.

15.2 IMPLEMENTATION OF CRIME CATEGORIZING BY LETHALITY OR SEVERITY

In the future, we have to categorize the crime depending on time, place, and type, which helps to assume how dangerous the crime is. To solve something, it is very important to understand the thing properly. We think that after categorizing the crime it is easier to reduce the crime rate.

15.3 ANALYZE THE CRIME REASON AND REQUISITION

Different crimes have different reasons. The main reason for crime is a family environment. An unhealthy family environment leads teenagers to commit crimes. Also, if the parents did not give their time to their children, children may not have much chance to talk with their parents so they mix with a bad company that pushes them to make many wrong decisions. Another reason is social and school life. An excessive amount of pressure will lead to a depressing life that is the reason to do the crime. To reduce the crime rate we need to be the focal point of youngsters' lives at home and at school. At home, guardians need to focus closer on the things their youngsters are doing and how they are feeling as they grow up and experience numerous new changes. In addition, youngsters should be urged to examine and realize what they are acceptable at so they feel sure about their fates.

15.4 MINING DATA BY PUBLIC OPINION

In the future, data might be collected by open form which can be used for data calculation and mining which can add more accuracy in the result. The main focus of this will be generating a report which will show the result of Bangladeshi people’s thought regarding the crime scenario.

15.5 CONCLUSION

To conclude the report, authors are hopeful about the future plans of the project and also very positive about the works done here. By simplifying the problem, the studied and detailed information on the present crime scenario in Bangladesh is “Automated crime news segmentation, knowledge-based creation, and public opinion mining for social awareness.” For a more precise understanding of why, how, where, and when something occurs. This report is necessary for

authorities such as police and other associates to grasp this crime situation, as well as a description of the criminal scenario and relevant action based on the case, in order to reduce crime casualties and eliminate crime scenarios in a consistent manner. People will also receive a valid decision from the government's court division.

Criminality has been described as a major problem in Bangladesh, and it is an evident aspect of contemporary societal structures. The cost of crime varies from a direct cost to a cost that is incurred later. As a social issue, crime has the potential to limit people's possibilities within the network and to instill widespread fear, which can trump national security, work, the cost of essential goods, impoverishment, and other concerns. Crime victims may suffer long-term mental harm. Outside speculation will be stifled by the fear rate, which will result in the mockery of neighborhoods or even entire sections of town. To combat crime, the desire to have a safe and secure environment necessitates a variety of open techniques and administrative mediations. Whatever the case may be, the crime's difficulty has thus far persisted. With the help of this report anyone will able to find and use increasingly skilled techniques to resolving cloud security challenges for a better and more secure online knowledge as well as observe a less real life crime scenario which can take our surrounding to a better position.

BIBLIOGRAPHY

1. Editor in Chief, 13 Major Pros and Cons of the Uniform Crime Report, ConnectUS, April 2019, Accessed 06 Apr 2020 [online] Available at https://connectusfund.org/13-major-pros-and-cons-of-the-uniform-crime-report.
2. Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J."A Novel Connectionist System for Improved Unconstrained Handwriting Recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence. 31 (5): 855–868, April 2009.
3. Collobert, Ronan; Weston, Jason. “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning”. NEC Labs America, 4 Independence Way, Princeton, NJ 08540 USA pp. 160–167, January 2008.
4. Bergstra, James; Bengio, Yoshua (July 2012)."Random Search for Hyper-Parameter Optimization" (PDF). Montreal, QC, H3C 3J7, Canada. 13: 281–305.
5. Song, Ruihua; Microsoft Research (September 2007). "Joint Optimization of Wrapper Generation and Template Detection" (PDF). The 13th International Conference on Knowledge Discovery and Data Mining.
6. Lukasz Kurgan and Petr Musilek. “A survey of Knowledge Discovery and Data Mining process models” The Knowledge Engineering Review. Volume 21 Issue 1, pp 1–24, March 2006.
7. Harrendorf, S Heiskane, M and Maldy S “International Statistics on Crime and justice.” HEUNI Publication Series No. 64, June 2010.
8. Bangladesh Police, Bangladesh Police Disciple Security Process, Bangladesh Police

a. January 2018 Accessed on 06 Apr 2020 [online] Available: https://www.police.gov.bd/en/crime_statistic/year/2019.

9. R. Chandramouli, Emerging Social Media Threats: Technology and Policy Perspectives, Researchgate,December 2017, Accessed on : 06 Apr 2020 [online] Available : https://www.researchgate.net/publication/252027118_Emerging_social_media_threats_Technology_and_policy_perspectives
10. Portland State University, Criminal Justice Policy Research Institute: Portland Crime Data, September 2012, Accessed on 04 Apr, 2020 [online] Available: www.pdx.edu/crime-data.
11. US Department of Justice, Federal Bureau of Investigation. "Uniform Crime Report Handbook", Federal Bureau of Investigation 1000 Custer Hollow Road Clarksburg, August 2004.
12. Plecas, R. G. (2017). *Introduction to Criminal Investigation: Processes, Practices and Thinking.* Darryl Plecas, University of the Fraser Valley: BCcampus.
13. Akash Desai, G. (2017). Criminal Investigation Tracker with Suspect Prediction. *IJESC*.
14. Nitin Sakhare, S. J. (2014). Criminal Identification System Based On Data Mining. *Recent Trends in Engineering & Technology* (p. 1). 2014: Elsevier Publication.
15. *Data Preprocessing in Data Mining - GeeksforGeeks*. (n.d.). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/
16. Raheel, S. (2018, November 9). *Choosing the right encoding method- Label vs OneHot Encoder*. Retrieved from Towards Data Science: https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b

17. *What is One Hot Encoding? Why and when do you have to use it?* (n.d.). Retrieved from Hackernoon: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f

18. Bronshtein, A. (2017). Train/Test Split and Cross Validation in Python. Understanding Machine Learning.

19. *Dataaspirant.* (2017). Retrieved from Dataaspirant: https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

20. Pupale, R. (2018). *Support Vector Machines(SVM) — An Overview*. Retrieved from Towards data science: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

21. *Classification Report - yellowbrick 0.9.1 documentation*. (2016). Retrieved from Yellow Brick: https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html

22. Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., & Ramanujam, K. (1997, March). Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery*, 121-125.

23. Brian, S. (2012, January 25). The Problem of Shot Selection in Basketball. *PLoS One*.

24. Choudhury, R. D., & Bhargava, P. (2007). Use of Artificial Neural Networks for Predicting the Outcome. *International Journal of Sports Science and Engineering, 1*(2), pp. 87-96.

25. Duckworth, F., & Lewis, T. (1999). *Your Comprehensive Guide to the Duckworth/Lewis Method for Resetting Targets in One-day Cricket.* University of the West of England.

26. Plecas, R. G. (2017). *Introduction to Criminal Investigation: Processes, Practices and Thinking.* Darryl Plecas, University of the Fraser Valley: BCcampus.

27. Sankaranarayanan, V. V., Sattar, J., & Lakshmanan, L. S. (2014). Auto-play: A data mining approach to ODI cricket simulation and prediction. *SIAM International Conference on Data Mining*, (p. 1064).

28. Tulabandhula, T., & Rudin, C. (2014). Tire Changes, Fresh Air, and Yellow Flags: Challenges in Predictive Analytics for Professional Racing. *Big data*.

29. *What is Data Preprocessing? - Definition from Techopedia*. (n.d.). Retrieved from Techopedia: https://www.techopedia.com/definition/14650/data-preprocessing

30. National Institute of Justice. (July 2001) Electronic Crime Scene Investigation. A Guide for First Responders. Available from: http://www.ncjrs.org/pdffiles1/nij/187736.pdf

31. Techniques of Crime Scene Investigation (8th edition, 2012) by Barry A. J. Fisher, Barry A. J. Fisher, David R. Fisher, David R. Fisher. Available from: https://www.nfstc.org/

32. *Which algorithms are best for machine learning? (n.d)* Retrieved from Builtin: https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies

33. Batchu V, Aravindhar D J et al., (2011). A classification based dependent approach for suppressing data, IJCA Proceedings on Wireless Information Networks & Business Information System (WINBIS 2012), Foundation of Computer Science (FCS).

34. Cios J, Pedrycz W et al., (1998). Data Mining in Knowledge Discovery, Academic Publishers.

35. Geenen P. L, van der Gaag L C et al., (2011). Constructing naive Bayesian classifiers for veterinary medicine: A case study in the clinical diagnosis of classical swine fever, Research in Veterinary Science, vol 91(1), 64–70.
36. Hamou A, Simmons A et al., (2011). Cluster analysis of MR imaging in Alzheimer's disease using decision tree reinement. International Journal of Artificial Intelligence, vol 6(S11), 90–99.
37. Han J, and Kamber M (2006). Data mining: concepts and techniques, Morgan Kaufmann Publishers, San Francisco, CA. Figure 2. Percentage comparison of correctly and incorrectly classified instances.Rizwan Iqbal et al 4225 www.indjst.org | Vol 6 (3) | March 2013 Indian Journal of Science and Technology | Print ISSN: 0974-6846 | Online ISSN: 0974-5645
38. Kochar B, and Chhillar R (2012). An Effective Data Warehousing System for RFID using Novel Data Cleaning, Data Transformation and Loading Techniques. Arab Journal of Information Technology, vol 9(3), 208–216
39. Kováč S (2012). Suitability analysis of data mining tools and methods, Bachelor's Thesis.
40. Kumar V, and Rathee N (2011). Knowledge discovery from database using an integration of clustering and classification, International Journal of Advanced Computer Science and Applications, vol 2(3), 29–32.
41. Li G, and Wang Y (2012). A privacy-preserving classification method based on singular value decomposition, Arab Journal of Information Technology, vol 9(6), 529–534.
42. Ngai E W T, Xiu L et al., (2009). Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, vol 36(2), 2592–2602.

43. Santhi P, and Bhaskaran V. M (2010). Performance of clustering algorithms in healthcare database, International Journal for Advances in Computer Science, vol 2(1), 26–31.
44. Selvaraj S, and Natarajan J (2011). Microarray data analysis and mining tools, Bioinformation, vol 6(3), 95–99. 13. UCI Machine Learning Repository (2012). Available from: http://archive.ics.uci.edu/ml/datasets.html
45. Wahbeh A H, Al-Radaideh Q A, et al., (2011). A comparison study between data mining tools over some classification methods, International Journal of Advanced Computer Science and Applications, Special Issue, 18–26.
46. Witten I, Frank E et al. (2011), Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. 17. Xu Y, Dong Z. Y et al., (2011) A decision tree-based on-line preventive control strategy for power system transient instability prevention, International Journal of Systems Science.
47. Amarnathan, L.C. (2003) Technological Advancement: Implications for Crime, The Indian Police Journal, April June.
48. Abraham, T. and de Vel, O. (2006) Investigative profiling with computer forensic log data and association rules," in Proceedings of the IEEE International Conference on Data Mining (ICDM'02), Pp. 11 – 18.
49. Brown, D.E. (1998) The regional crime analysis program (RECAP): A framework for mining data to catch criminals," in Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, Pp. 2848-2853.
50. Corcoran J.J., Wilson I.D. AND Ware J.A. (2003) Predicting the geo-temporal variations of crime and disorder, International Journal of Forecasting, Vol. 19, Pp.623–634.

51. David, G. (2006) Globalization and International Security: Have the Rules of the Game Changed?, Annual meeting of the International Studies Association, California, USA, http://www.allacademic.com/meta/ p98627_index.html.
52. de Bruin, J.S. , Cocx, T.K. , Kosters, W.A. , Laros, J. and Kok, J.N. (2006) Data mining approaches to criminal career analysis," in Proceedings of the Sixth International Conference on Data Mining (ICDM'06), Pp. 171-177.
53. Hauck, R.V.Atabakhsh, H., Ongvasith, P., Gupta, H. and Chen, H. (2002) Using Coplink to Analyze Criminal-Justice Data, Computer, Volume 35 Issue 3, Pp. 30-37.
54. Keyvanpour, M.R., Javideh, M. and Ebrahimi, M.R. (2010) Detecting and investigating crime by means of data mining: a general crime matching framework, Procedia Computer Science, World Conference on Information Technology, Elsvier B.V., Vol. 3, Pp. 872-830.
55. Nath, S. (2007) Crime data mining, Advances and innovations in systems, K. Elleithy (ed.), Computing Sciences and Software Engineering, Pp. 405-409.
56. Senator, T.E., Goldberg, H.G., Wooton, J., Cottini, M.A., Khan, A.F.U., Klinger, C.D., Llamas, W.M., Marrone, M.P. and Wong, R.W.H. (1995) The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions, AI Magazine, Vol.16, No. 4, Pp. 21-39
57. Yehya Abouelnaga. San Francisco crime classification. arXiv preprint arXiv:1607.03626, 2016. 13Figure 10: More Balanced data set Figure 11: Recall for imbalanced dataset
58. Tahani Almanie, Rsha Mirza, and Elizabeth Lor. Crime prediction based on crime types and using spatial and temporal criminal hotspots. arXiv preprint arXiv:1508.02050, 2015.

59. Exegetic Andrew B. Collier. Making Sense of Logarithmic Loss. http://www.exegetic.biz/blog/2015/12/making-sense-logarithmic-loss/, 2015.
60. Shen Ting Ang, Weichen Wang, and Silvia Chyou. San Francisco crime classification. University of California San Diego, 2015.
61. J. Bruin. Ucla: Multinomial logistic regression @ONLINE, February 2011.
62. City and County of San Francisco. Police Department Incidents. https://data.sfgov.org/PublicSafety/Police-Department-Incidents/tmnf-yvry/, 2017. 14Figure 12: Recall for a more balanced dataset
63. DataSF. Open government. https://www.data.gov/open-gov/. Accessed 2018-04-12.
64. Emre Eftelioglu, Shashi Shekhar, and Xun Tang. Crime hotspot detection: A computational perspective. In Data Mining Trends and Applications in Criminal Science and Investigations, pages 82–111. IGI Global, 2016.
65. Debopriya Ghosh, Soon Chun, Basit Shafiq, and Nabil R Adam. Big data-based smart city platform: Real-time crime analysis. In Proceedings of the 17th International Digital Government Research Conference on Digital Government Research, pages 58–66. ACM, 2016.
66. Jelle J Goeman and Saskia le Cessie. A goodness-of-fit test for multinomial logistic regression. Biometrics, 62(4):980–985, 2006.
67. Jacob Hochstetler, Lauren Hochstetler, and Song Fu. An optimal police patrol planning strategy for smart city safety. In 2016 IEEE 18th International Conference on HPCC/SmartCity/DSS, pages 1256– 1263. IEEE, 2016.
68. Dennis Hsu, Melody Moh, and Teng-Sheng Moh. Mining frequency of drug side effects over a large twitter dataset using apache spark. In Proceedings of the 2017 IEEE/ACM

International Conference on Advances in Social Networks Analysis and Mining 2017, pages 915–924. ACM, 2017.

69. Dan Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2009.
70. Brian Kolo. Binary and Multiclass Classification. Lulu. com, 2011.
71. Gabriela Hernandez Larios. Case study report: San Francisco crime classification, 2016.
72. Andy Liaw, Matthew Weiner, et al. Classification and regression by randomforest. R news, 2(3):18–22, 2002.
73. Shannon J Lining, Martin A Andresen, and Paul J Brantingham. Crime seasonality: Examining the temporal fluctuations of property crime in cities with varying climates. International journal of offender therapy and comparative criminology, 61(16):1866–1891, 2017.
74. Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. Astrophysics and space science, 39(2):447–462, 1976. 15
75. Paolo Neirotti, Alberto De Marco, Anna Corinna Cagliano, Giulio Mangano, and Francesco Scorrano. Current trends in smart city initiatives: Some stylised facts. Cities, 38:25–36, 2014.
76. Trung T Nguyen, Amartya Hatua, and Andrew H Sung. Building a learning machine classifier with inadequate data for crime prediction. Journal of Advances in Information Technology Vol, 8(2), 2017.
77. Philip H Swain and Hans Hauska. The decision tree classifier: Design and potential. IEEE Transactions on Geoscience Electronics, 15(3):142–147, 1977.

78. Luca Venturini and Elena Baralis. A spectral analysis of crimes in San Francisco. In Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics, page 4. ACM, 2016.
79. Xiaoxu Wu. An informative and predictive analysis of the San Francisco police department crime data, Master Thesis, 2016.