Crime analysis and prediction of Bangladesh using machine learning tools

by

Gautam Kumar Shom

Supervisor by

SARNALI BASAK
Assistant Professor
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

Jahangirnagar University
November, 2018

Declaration

The research work entitled ”Crime analysis and prediction of Bangladesh using machine learning tools” has been carried out in the Department of Computer Science and Engineering, Jahangirnagar University is original and conforms the regulations of this university. I understand the university’s policy on plagiarism and declare that no part of this thesis has been copied from other sources or been previously submitted elsewhere for the award of any degree.

Gautam Kumar Shom
ID: CSE201703071
Session: 2017-18

Counter Signed By

Sarnali Basak
Assistant Professor
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

Approval of Acceptance

This thesis report written by Gautam Kumar Shom and ID: CSE201703071 entitled ”Crime analysis and prediction of Bangladesh using machine learning tools” is submitted to the PMSCS Program in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The project is done under the supervision of Sarnali Basak.
I have examined this report and recommend it’s acceptance:

Sarnali Basak
Assistant Professor
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

Dr. Md. Ezharul Islam
Associate Professor and 2nd examiner
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

Prof. Md. Abul Kalam Azad
Professor
Coordinator, PMSCS program, 2018
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh

Contents

List of Figures

Crime analysis and prediction of Bangladesh using machine learning tools

Abstract

Though average crime rates in Bangladesh increasing for the last few years so analyzing and predicting crime rate has a grate importance in society.If we can analysis and predict crime then in the future there should be less crime from now. This thesis report represents crime analysis and prediction in Bangladesh crime rate based on machine learning algorithms. There is an open source crime statistics data from Bangladesh police and we are working on that data to analyze and predict. We normalize those data and add some more fields. After that we used different learning methods and compare each methods.

Acknowledgment

First of all thanks to my supervisor Sarnali Basak, Assistant Professor, CSE, JU who gave me so much support to do this work and also thanks to all of my respected teachers. She helped me a lot in every aspect of our work and guided me with proper directions whenever I sought one. I exposed to the real wonderful and work ”Crime analysis and prediction of Bangladesh using machine learning tools” by her careful supervision.

Date: 09 November, 2018
Jahangirnagar University
Savar, Dhaka, Bangladesh

Chapter 1

Introduction

Studies have shown that crime tends to be associated with slower economic growth at both the national level[8] and the local level, such as cities and metropolitan areas[5]. So it is important for us to focus on this section. Several studies in criminology and sociology have provided evidence of significant concentrations of crime at micro levels of geography, regardless of the specific unit of analysis defined[4] [13].

1.1 Overview

In this thesis we try to explain about crime analysis and prediction about Bangladesh crime. Bangladesh police open their statistical data to public, which is open source data. So we collect those data and normalize those data. After that we implement various learning methods to analyze those data and predict based on those normalize data.

To predict future crime accurately with a better performance, it is a challenging task because of the increasing numbers of crime in present days. We are interested in applying machine learning methods to datasets regarding crime (crime statistics in particular division range and metropolitan police range) and possible related factors (such as Dacoity, Robbery, Murder, Kidnapping etc.). Specifically, we are interested in investigating if it is possible to predict criminal events for a specific time and place in the future.

In this thesis. first and foremost, we are interested in seeing if we can analyze and predict the criminal incidents, perhaps for a specific type of crime, for a time range and geographic region. Second, we are interested in learning which features have the most predictive power with respect to crime.

1.2 Objectives

The main objective of this thesis is to analyze and predict Bangladesh crime, focusing crime prediction based on Bangladesh police statistical data. To predict future crime accurately with a better performance, it is a challenging task because of the increasing numbers of crime in present days. We are interested in applying machine learning methods to datasets regarding crime (crime statistics in particular division range and metropolitan police range) and possible related factors (such as Dacoity, Robbery, Murder, Kidnapping etc.). Specifically, we are interested in investigating if it is possible to predict criminal events for a specific time and place in the future.

In this thesis. first and foremost, we are interested in seeing if we can analyze and predict the criminal incidents, perhaps for a specific type of crime, for a time range and geographic region. Second, we are interested in learning which features have the most predictive power with respect to crime.

1.3 Summary

As a problem-driven project, for this project we will include implementing various supervised learning methods . The first challenge in this problem will be feature extraction. For analyze correlation between features we extract, as well as simple data exploration methodologies (such as analyzing crime data over-time, etc.) will be helpful for feature prediction.

For machine learning models, we will begin with simple DTC(Dicision Tree Classifier), Naive Bayes, Linear SVC(Support Vector Classifier) on the features we will extract. Ultimately we would like to try to implement a model for practice and to see if they lead to improved performance. Beyond that, we will need to explore various other regression models to determine what is appropriate for the data and context. If time allows, we would also like to explore other models typically good for time series data, such as Polynomial Kernel and K means.

In the end, we will likely need to develop a more sophisticated model that more fully captures the dynamics of crime. These types of modeling enhancements may lead to better predictive power[12] . We expect to spend a good amount of time tweaking different model ideas, and measuring their performance.

Chapter 2

System Modeling

2.1 Normalization

Normalization significantly reduces the training time in feed-forward neural networks[2] [7] . Here is the normalization formula we used in our data for normalize all those data.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{2.1}$$

where x=(x1,...,xn) and zi is our ith normalized data. We needed to normalize the inputs, otherwise the model will be not stable. In essence, normalization is done to have the same range of values for each of the inputs to the model. This can guarantee stable convergence of weight and biases. Normalization is the process of reducing measurements to a standard scale. We normalized our data from 0 to 1 which help us to get fine results.

2.2 Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method. For classification and regression we had to used this learning method. It is possible to generate knowledge in the form of decision trees that is capable of solving difficult problems of practical significance[9]. Main objective of this learning method is to generate a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

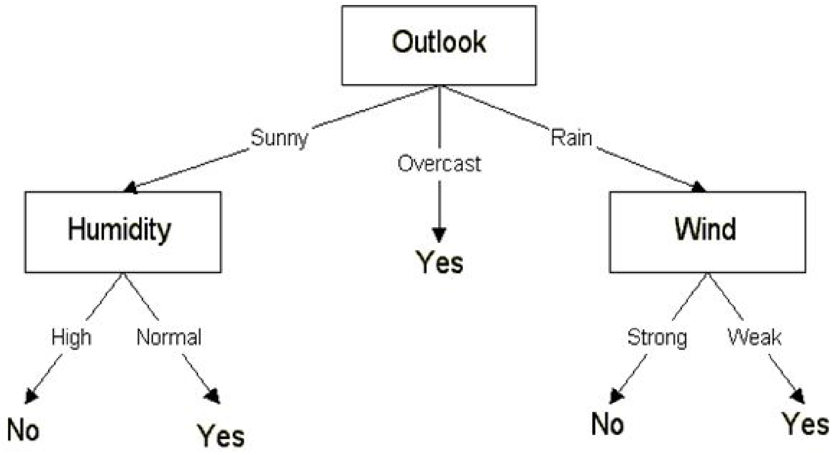


Figure 2.1: Simple weather decision tree

We used decision trees, the reasons are:

- Accomplish well results even if it's assumptions are somehow violated[11] .
- It's very easy to understand and to implement. Trees can be also visualized.
- Timing cost of using decision tree is fast because it is logarithmic in the number of data points used to train the tree.
- Multi-output problems can be handle properly by using this learning method.
- Numerical and categorical both type of data can be able to handle easily. Other learning methods are usually handle analyzing datasets that have only one type of variable.

2.3 Linear Classification

Commonly used approach in statistics for obtaining a linear classifier is logistic regression[15].There are two types of linear classification we used in our data.

1. Gaussian NB(Naive Bayes) to learn a Naive Bayes Classifier
2. Linear SVC(Support Vector Classification) to learn a Linear Support Vector Machine

2.3.1 Gaussian Naive Bayes

Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Naive Bayes is used because its simplicity and its computational efficiency make it an attractive choice[3]. It's competitive performance in classification is surprising, because the conditional independence assumption on which it is based, is rarely true in real world applications[14]. Gaussian Naive Bayes is algorithm for classification. Here is the formula for GaussianNB algorithm:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{2.2}$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

Main thing in Gaussian Naive Bayes algorithm is, it's a special type of NB algorithm. When the features have continuous values then Gaussian naive Bayes used[10] [6] . It's also assumed that all the features are following a Gaussian distribution i.e, normal distribution.

2.3.2 Linear Support Vector Classification(LinearSVC)

LinearSVC is much like SVC but there is a parameter in LinearSVC, which is kernel='linear'.Since it is implemented in terms of liblinear rather than libsvm, so it is more flexible in the area of loss functions and penalties. It's also scale better in large numbers. This classifier supports both dense and sparse input. We used LinearSVC rather than SVC because there are some differences between LinearSVC and SVC. Here are some differences:

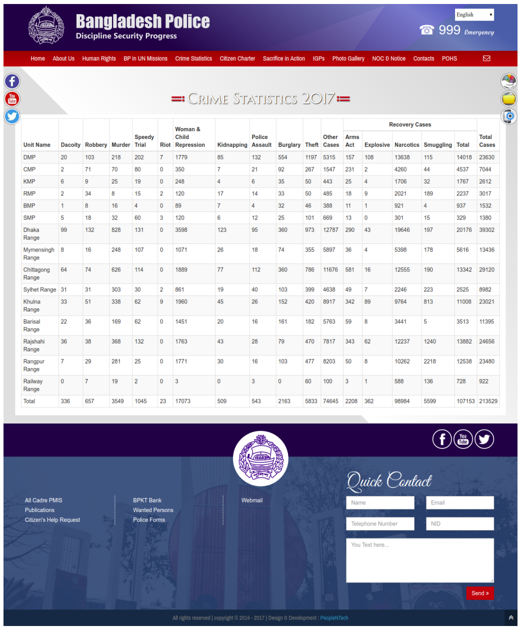
- SVC minimizes the regular hinge loss but LinearSVC minimizes the squared hinge loss.
- SVC uses the One-vs-One multi class reduction but LinearSVC uses the One-vs-All (also known as One-vs-Rest) multi class reduction.

Chapter 3

Design

3.1 Collecting statistic data

Bangladesh police make their data open to all in their website. Here is the crime data url[1] the reference section. From where we get our data. Here is a sample image of our data from where we collect:



Crime Statistics 2017

Unit Name	Dacoity	Robbery	Murder	Speedy Trial	Riot	Woman & Child Repression	Kidnapping	Police Assault	Burglary	Theft	Other Cases	Recovery Cases					Total Cases
												Arms Act	Explosive	Narcotics	Smuggling	Total	
DMP	20	103	218	202	7	1779	85	132	554	1197	5315	157	108	13638	115	14018	23630
CMP	2	71	70	80	0	350	7	21	92	267	1547	231	2	4260	44	4537	7044
KMP	6	9	25	19	0	248	4	6	35	50	443	25	4	1706	32	1767	2612
RMP	2	34	8	15	2	120	17	14	33	50	485	18	9	2021	189	2237	3017
BMP	1	8	16	4	0	89	7	4	32	46	388	11	1	921	4	937	1532
SMP	5	18	32	60	3	120	6	12	25	101	669	13	0	301	15	329	1380
Dhaka Range	99	132	828	131	0	3598	123	95	360	973	12787	290	43	19646	197	20176	39302
Mymensingh Range	8	16	248	107	0	1071	26	18	74	355	5897	36	4	5398	178	5616	13436
Chittagong Range	64	74	626	114	0	1889	77	112	360	786	11676	581	16	12555	190	13342	29120
Sylhet Range	31	31	303	30	2	861	19	40	103	399	4638	49	7	2246	223	2525	8982
Khulna Range	33	51	338	62	9	1960	45	26	152	420	8917	342	89	9764	813	11008	23021
Barisal Range	22	36	169	62	0	1451	20	16	161	182	5763	59	8	3441	5	3513	11395
Rajshahi Range	36	38	368	132	0	1763	43	28	79	470	7817	343	62	12237	1240	13882	24656
Rangpur Range	7	29	281	25	0	1771	30	16	103	477	8203	50	8	10262	2218	12538	23480
Railway Range	0	7	19	2	0	3	0	3	0	60	100	3	1	588	136	728	922
Total	336	657	3549	1045	23	17073	509	543	2163	5833	74645	2208	362	98984	5599	107153	213529

Figure 3.1: Bangladesh police open data from where we collected our data

3.2 Information of statistic data

So figure 3.1 shows that Bangladesh police provides us all data that we need. Here they provide crime statistics by four main sections and the sections are:

1. Yearly (2010 to 2018)
2. Every months of 2018(till august 18)
3. Monthly human trafficking cases of year 2018(till august 18)
4. Comparative crime statistics(2002 to 2015)

And for yearly (2010 to 2018) and every months of 2018(till august 18) section there are about 15 names of crime. Which are:

- Dacoity
- Robbery
- Murder
- Speedy Trial
- Riot
- Woman and Child Repression
- Kidnapping
- Police Assault
- Burglary
- Theft
- Arms Act
- Explosive
- Narcotics
- Smuggling
- Other Cases

Also in yearly (2010 to 2018) and every months of 2018(till august 18) section there are about 15 ranges in whole Bangladesh where this crime occurred. Which are:

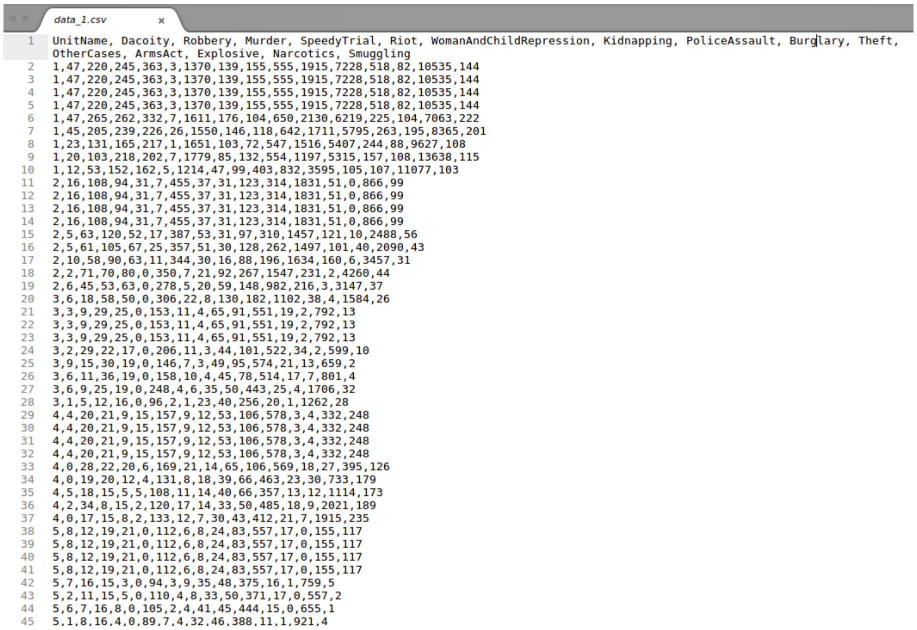
- DMP(Dhaka Metropolitan Police)
- CMP(Chittagong Metropolitan Police)
- KMP(Khulna Metropolitan Police)
- RMP(Rajshahi Metropolitan Police)
- BMP(Barisal Metropolitan Police)
- SMP(Sylhet Metropolitan Police)
- Dhaka Range
- Mymensingh Range
- Chittagong Range
- Sylhet Range
- Khulna Range
- Barisal Range
- Rajshahi Range
- Rangpur Range
- Railway Range

We will convert every statistics in many division and apply different supervised algorithm into that for predicting and analyzing all types of crime in Bangladesh.

So we will convert our data into CSV(Comma-Separated Values) format and make it in our way for apply supervised learning methods in python.

3.3 Convert data to CSV file

Since those statistical data is open source for all so we collect all data and manually saved in a CSV(Comma-Separated Values) file. We added another filed in the name of crime which is UnitName. UnitName column contains all 15 range name in number which are (1 to 15 respectively DMP to Railway range). Here is a sample image of our CSV file:



```
UnitName, Dacoity, Robbery, Murder, SpeedyTrial, Riot, WomanAndChildRepression, Kidnapping, PoliceAssault, Burglary, Theft,
OtherCases, ArmsAct, Explosive, Narcotics, Smuggling
1,47,220,245,363,3,1370,139,155,555,1915,7228,518,82,10535,144
1,47,220,245,363,3,1370,139,155,555,1915,7228,518,82,10535,144
1,47,220,245,363,3,1370,139,155,555,1915,7228,518,82,10535,144
1,47,220,245,363,3,1370,139,155,555,1915,7228,518,82,10535,144
1,47,265,262,332,7,1611,176,104,650,2130,6219,225,104,7063,222
1,45,205,239,226,26,1550,146,118,642,1711,5795,263,195,8365,201
1,23,131,165,217,1,1651,103,72,547,1516,5407,244,88,9627,108
1,20,103,218,202,7,1779,85,132,554,1197,5315,157,108,13638,115
1,12,53,152,162,5,1214,47,99,403,832,3595,105,107,11077,103
2,16,108,94,31,7,455,37,31,123,314,1831,51,0,866,99
2,16,108,94,31,7,455,37,31,123,314,1831,51,0,866,99
2,16,108,94,31,7,455,37,31,123,314,1831,51,0,866,99
2,16,108,94,31,7,455,37,31,123,314,1831,51,0,866,99
2,5,63,120,52,17,387,53,31,97,310,1457,121,10,2488,56
2,5,61,105,67,25,357,51,30,128,262,1497,101,40,2090,43
2,10,58,90,63,11,344,30,16,88,196,1634,160,6,3457,31
2,2,71,70,80,0,350,7,21,92,267,1547,231,2,4260,44
2,6,45,53,63,0,278,5,20,59,148,982,216,3,3147,37
3,6,18,58,50,0,306,22,8,130,182,1102,38,4,1584,26
3,3,9,29,25,0,153,11,4,65,91,551,19,2,792,13
3,3,9,29,25,0,153,11,4,65,91,551,19,2,792,13
3,3,9,29,25,0,153,11,4,65,91,551,19,2,792,13
3,2,29,22,17,0,206,11,3,44,101,522,34,2,599,10
3,9,15,30,19,0,146,7,3,49,95,574,21,13,659,2
3,6,11,36,19,0,158,10,4,45,78,514,17,7,801,4
3,6,9,25,19,0,248,4,6,35,50,443,25,4,1706,32
3,1,5,12,16,0,96,2,1,23,40,256,20,1,1262,28
4,4,20,21,9,15,157,9,12,53,106,578,3,4,332,248
4,4,20,21,9,15,157,9,12,53,106,578,3,4,332,248
4,4,20,21,9,15,157,9,12,53,106,578,3,4,332,248
4,4,20,21,9,15,157,9,12,53,106,578,3,4,332,248
4,0,28,22,20,6,169,21,14,65,106,569,18,27,395,126
4,0,19,20,12,4,131,8,18,39,66,463,23,30,733,179
4,5,18,15,5,5,108,11,14,40,66,357,13,12,1114,173
4,2,34,8,15,2,120,17,14,33,50,485,18,9,2021,189
4,0,17,15,8,2,133,12,7,30,43,412,21,7,1915,235
5,8,12,19,21,0,112,6,8,24,83,557,17,0,155,117
5,8,12,19,21,0,112,6,8,24,83,557,17,0,155,117
5,8,12,19,21,0,112,6,8,24,83,557,17,0,155,117
5,8,12,19,21,0,112,6,8,24,83,557,17,0,155,117
5,7,16,15,3,0,94,3,9,35,48,375,16,1,759,5
5,2,11,15,5,0,110,4,8,33,50,371,17,0,557,2
5,6,7,16,8,0,105,2,4,41,45,444,15,0,655,1
5,1,8,16,4,0,89,7,4,32,46,388,11,1,921,4
```

Figure 3.2: CSV format of our data

3.4 Normalize CSV file

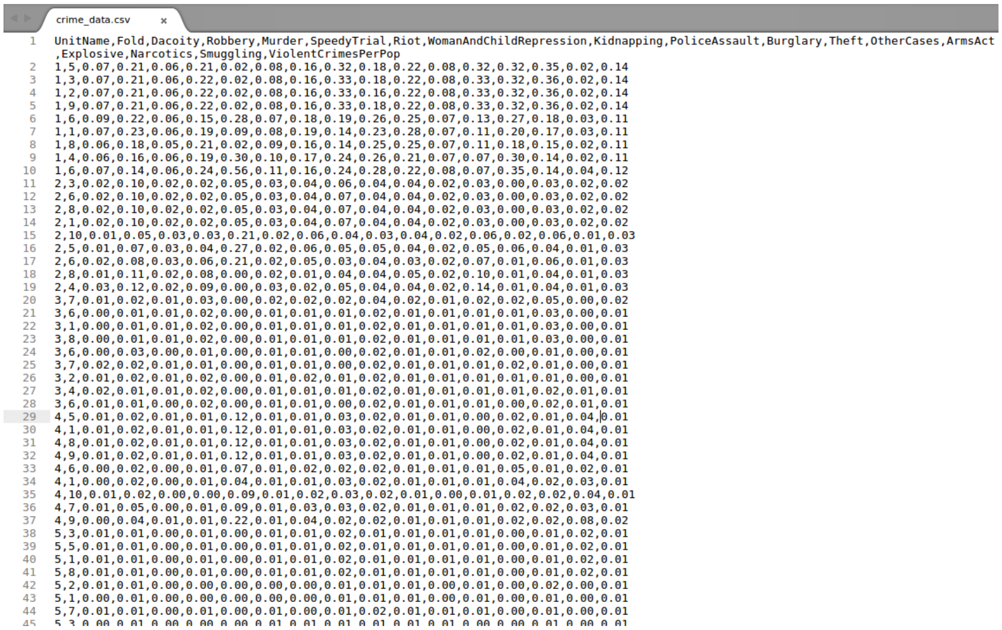
After we collect all data then we normalize all data. We normalize those data because it is easier for learning method to handle and also we discussed about why we need normalization in section 2.1. Here is function code in python we used for normalize those data.

```
def normalize(lst):
    s = sum(lst)
    value = map(lambda x: float(x)/s, lst)
    value = [ '%.2f' % elem for elem in value ]
    return value
```

In this normalize process we also normalize a new column which is ViolentCrimesPerPop(Violent Crimes Per Population). We normalize every column that has been given and for ViolentCrimesPerPop we took Total Cases field. Then calculate every data and write on that CSV file.

3.5 After Normalize data

Since we normalize all data and added some fields so here is our final data we worked on:



```
UnitName,Fold,Dacoity,Robbery,Murder,SpeedyTrial,Riot,WomanAndChildRepression,Kidnapping,PoliceAssault,Burglary,Theft,OtherCases,ArmsAct,Explosive,Narcotics,Smuggling,ViolentCrimesPerPop
1,5,0.07,0.21,0.06,0.21,0.02,0.08,0.16,0.32,0.18,0.22,0.08,0.32,0.32,0.35,0.02,0.14
1,3,0.07,0.21,0.06,0.22,0.02,0.08,0.16,0.33,0.18,0.22,0.08,0.33,0.32,0.36,0.02,0.14
1,2,0.07,0.21,0.06,0.22,0.02,0.08,0.16,0.33,0.16,0.22,0.08,0.33,0.32,0.36,0.02,0.14
1,9,0.07,0.21,0.06,0.22,0.02,0.08,0.16,0.33,0.18,0.22,0.08,0.33,0.32,0.36,0.02,0.14
1,6,0.09,0.22,0.06,0.15,0.28,0.07,0.18,0.19,0.26,0.25,0.07,0.13,0.27,0.18,0.03,0.11
1,1,0.07,0.23,0.06,0.19,0.09,0.08,0.19,0.14,0.23,0.28,0.07,0.11,0.20,0.17,0.03,0.11
1,8,0.06,0.18,0.05,0.21,0.02,0.09,0.16,0.14,0.25,0.25,0.07,0.11,0.18,0.15,0.02,0.11
1,4,0.06,0.16,0.06,0.19,0.30,0.10,0.17,0.24,0.26,0.21,0.07,0.07,0.30,0.14,0.02,0.11
1,6,0.07,0.14,0.06,0.24,0.56,0.11,0.16,0.24,0.28,0.22,0.08,0.07,0.35,0.14,0.04,0.12
2,3,0.02,0.10,0.02,0.02,0.05,0.03,0.04,0.06,0.04,0.04,0.02,0.03,0.00,0.03,0.02,0.02
2,6,0.02,0.10,0.02,0.02,0.05,0.03,0.04,0.07,0.04,0.04,0.02,0.03,0.00,0.03,0.02,0.02
2,8,0.02,0.10,0.02,0.02,0.05,0.03,0.04,0.07,0.04,0.04,0.02,0.03,0.00,0.03,0.02,0.02
2,1,0.02,0.10,0.02,0.02,0.05,0.03,0.04,0.07,0.04,0.04,0.02,0.03,0.00,0.03,0.02,0.02
2,10,0.01,0.05,0.03,0.03,0.21,0.02,0.06,0.04,0.03,0.04,0.02,0.06,0.02,0.06,0.01,0.03
2,5,0.01,0.07,0.03,0.04,0.27,0.02,0.06,0.05,0.05,0.04,0.02,0.05,0.06,0.04,0.01,0.03
2,6,0.02,0.08,0.03,0.06,0.21,0.02,0.05,0.03,0.04,0.03,0.02,0.07,0.01,0.06,0.01,0.03
2,8,0.01,0.11,0.02,0.08,0.00,0.02,0.01,0.04,0.04,0.05,0.02,0.10,0.01,0.04,0.01,0.03
2,4,0.03,0.12,0.02,0.09,0.00,0.03,0.02,0.05,0.04,0.04,0.02,0.14,0.01,0.04,0.01,0.03
3,7,0.01,0.02,0.01,0.03,0.00,0.02,0.02,0.02,0.04,0.02,0.01,0.02,0.02,0.05,0.00,0.02
3,6,0.00,0.01,0.01,0.02,0.00,0.01,0.01,0.01,0.02,0.01,0.01,0.01,0.01,0.03,0.00,0.01
3,1,0.00,0.01,0.01,0.02,0.00,0.01,0.01,0.01,0.02,0.01,0.01,0.01,0.01,0.03,0.00,0.01
3,8,0.00,0.01,0.01,0.02,0.00,0.01,0.01,0.01,0.02,0.01,0.01,0.01,0.01,0.03,0.00,0.01
3,6,0.00,0.03,0.00,0.01,0.00,0.01,0.01,0.00,0.02,0.01,0.01,0.02,0.00,0.01,0.00,0.01
3,7,0.02,0.02,0.01,0.01,0.00,0.01,0.01,0.00,0.02,0.01,0.01,0.01,0.02,0.01,0.00,0.01
3,2,0.01,0.02,0.01,0.02,0.00,0.01,0.02,0.01,0.02,0.01,0.01,0.01,0.01,0.01,0.00,0.01
3,4,0.02,0.01,0.01,0.02,0.00,0.01,0.01,0.01,0.02,0.01,0.01,0.01,0.01,0.02,0.01,0.01
3,6,0.01,0.01,0.00,0.02,0.00,0.01,0.01,0.00,0.02,0.01,0.01,0.01,0.00,0.02,0.01,0.01
4,5,0.01,0.02,0.01,0.01,0.12,0.01,0.01,0.03,0.02,0.01,0.01,0.00,0.02,0.01,0.04,0.01
4,1,0.01,0.02,0.01,0.01,0.12,0.01,0.01,0.03,0.02,0.01,0.01,0.00,0.02,0.01,0.04,0.01
4,8,0.01,0.02,0.01,0.01,0.12,0.01,0.01,0.03,0.02,0.01,0.01,0.00,0.02,0.01,0.04,0.01
4,9,0.01,0.02,0.01,0.01,0.12,0.01,0.01,0.03,0.02,0.01,0.01,0.00,0.02,0.01,0.04,0.01
4,6,0.00,0.02,0.00,0.01,0.07,0.01,0.02,0.02,0.02,0.01,0.01,0.01,0.05,0.01,0.02,0.01
4,1,0.00,0.02,0.00,0.01,0.04,0.01,0.01,0.03,0.02,0.01,0.01,0.01,0.04,0.02,0.03,0.01
4,10,0.01,0.02,0.00,0.00,0.09,0.01,0.02,0.03,0.02,0.01,0.00,0.01,0.02,0.02,0.04,0.01
4,7,0.01,0.05,0.00,0.01,0.09,0.01,0.03,0.03,0.02,0.01,0.01,0.01,0.02,0.02,0.03,0.01
4,9,0.00,0.04,0.01,0.01,0.22,0.01,0.04,0.02,0.02,0.01,0.01,0.01,0.02,0.02,0.08,0.02
5,3,0.01,0.01,0.00,0.01,0.00,0.01,0.01,0.02,0.01,0.01,0.01,0.01,0.00,0.01,0.02,0.01
5,5,0.01,0.01,0.00,0.01,0.00,0.01,0.01,0.02,0.01,0.01,0.01,0.01,0.00,0.01,0.02,0.01
5,1,0.01,0.01,0.00,0.01,0.00,0.01,0.01,0.02,0.01,0.01,0.01,0.01,0.00,0.01,0.02,0.01
5,8,0.01,0.01,0.00,0.01,0.00,0.01,0.01,0.02,0.01,0.01,0.01,0.01,0.00,0.01,0.02,0.01
5,2,0.01,0.01,0.00,0.00,0.00,0.00,0.00,0.01,0.01,0.01,0.00,0.01,0.00,0.02,0.00,0.01
5,1,0.00,0.01,0.00,0.00,0.00,0.00,0.00,0.01,0.01,0.01,0.00,0.01,0.00,0.01,0.00,0.01
5,7,0.01,0.01,0.00,0.01,0.00,0.01,0.00,0.01,0.02,0.01,0.01,0.01,0.00,0.01,0.00,0.01
5,3,0.00,0.01,0.00,0.00,0.00,0.01,0.01,0.01,0.01,0.01,0.01,0.00,0.00,0.01,0.00,0.01
```

Figure 3.3: Final csv data screenshot

Chapter 4

Experimental Result

4.1 System Information

Processor	: Intel(R) Core(TM) i3-5005U CPU @ 2.00GHz
RAM	: 3.9 GB
Operating System	: Ubuntu 16.04.5 LTS

4.2 Project Information

Project name	: Crime analysis and prediction of Bangladesh using machine learning tools
Programming Language	: Python
Python IDE	: Jupyter Notebook
Libraries and packages	: pandas, numpy, sklearn, matplotlib, IPython, pydotplus, time, math.
Algorithms	: Decision Trees, LinearSVC, Linear Classification, Regression

4.3 Overall Approach

In this section we will discuss about our whole work we had done. Here is simple flowchart diagram we made for visualize our work:

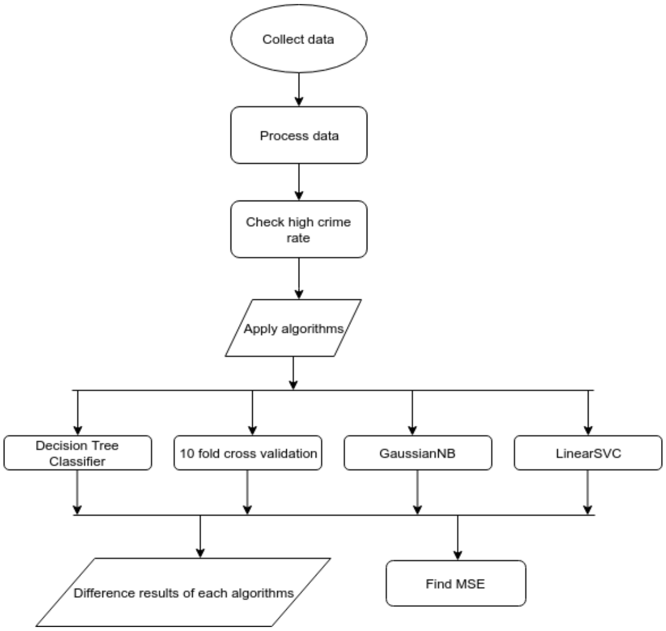


Figure 4.1: Overall approach of our work

In figure 4.1, we can see that first we collected data then process it as well. Then we checked high crime rate. Checked high crime rate means that if overall crime rate is positive or negative. After check high crime rate we applied around 4 algorithms to find which gives us better results in our data. We also make differences between our algorithms we used and also checked which one in better. We also calculate MSE(Minimum Square Error) of our results. We will fully discuss about this in other sections.

4.4 Decision tree results

In this decision tree section we checked and applied many methods. First of all we checked if overall high crime rate is positive or negative. And also got the results of the percentages of positive and negative results. Then we will also discuss about decision tree classifier, find out the depth and accuracy of the DTC, main features for classification, top main features for classification, 10-fold cross-validation in the subsections of decision tree results section.

4.4.1 High crime rate

After we have our clean data we first and foremost find out if overall high crime rate is positive or negative. We used this function to check which is true if the crime rate ViolentCrimesPerPop is greater than 0.1. We discussed about ViolentCrimesPerPop in the section 3.4. So here if ViolentCrimesPerPop is greater than 0.1 then it will be high crime. Here is the python function we made:

```
def setHighCrime(df):
    if df['ViolentCrimesPerPop'] > 0.1:
        return True
    else:
        return False
```

So the result was like:

```
Percentage Positive Instance = 31.1111111111
Percentage Negative Instance = 68.8888888889
```

4.4.2 Decision Tree Classifier(DTC)

We used decision tree classifier to learn a decision tree to predict highCrime on the entire dataset. So we got the training accuracy, precision, and recall for this tree. And here is the first result for DTC learning algorithm.

```
Training Accuracy = 1.0
Precision = 1.0
Recall = 1.0
```

The scores values shows overfitting so we can define max depth to avoid the complexity of the tree and to reach a point from where there is a decrease in the cross validation performance. So we calculated the depth of the DTC. Here is the result:

```
Depth: 1 Accuracy: 0.954
Depth: 2 Accuracy: 0.962
Depth: 3 Accuracy: 0.954
Depth: 4 Accuracy: 0.954
```

After that we can see that the point up to which the performance is increasing is the depth 4. We can specify the depth and witness the results. Then we used DTC again so that there won't be any overfitting. So the final result for DTC learning algorithm is:

Accuracy = 0.9925925925925926
Precision = 0.9767441860465116
Recall = 0.9767441860465116

4.4.3 Main features used for classification

In the DTC classification there are some major features by which every data got separate classes. Here is our classification graph result we got. So as we can see by this Figure 4.2 the top main feature is Kidnapping because it is the split point of the tree.

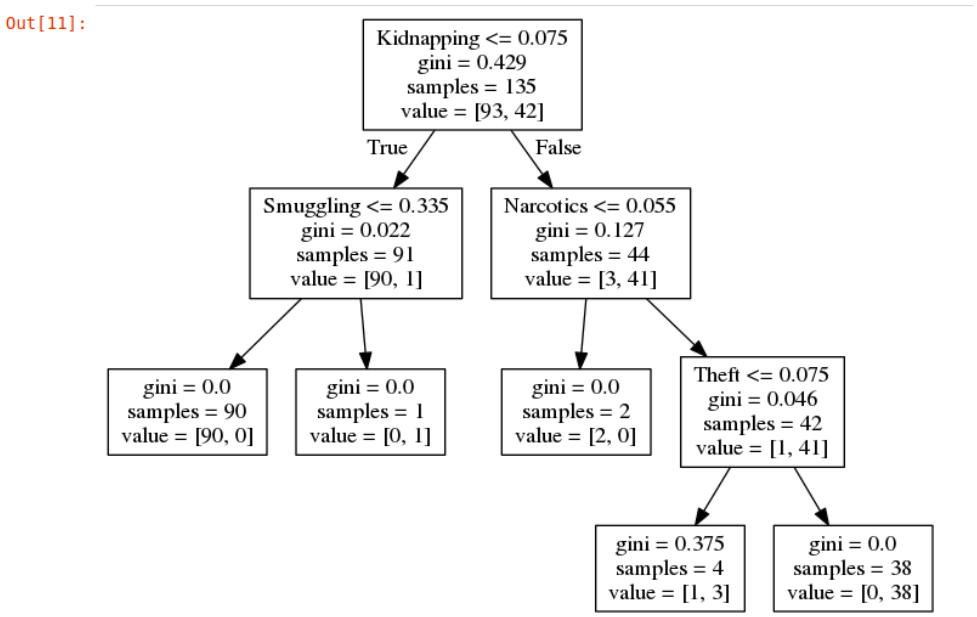


Figure 4.2: Main features used for classification

4.4.4 10-fold cross-validation using DTC

We also applied cross-validation to do 10-fold cross-validation to estimate the out-of-training accuracy of decision tree learning for this task. And here is our result after applied 10-fold cross-validation learning method.

Accuracy = 0.9461538461538461
Precision = 0.925
Recall = 0.9466666666666667

4.5 Linear Classification

In this linear classification section we also checked and applied many methods. We discussed about linear classification in section 2.3. First of all we applied 10-fold cross-validation, 10 most predictive features, LinearSVC learning methods to see their accuracy, prediction and also recall.

4.5.1 10-fold cross-validation using Using GaussianNB

Using GaussianNB(Gaussian Naive Bayes) to learn a Naive Bayes classifier to predict highCrime. For this we train a 10-fold cross-validation and here is the result:

Accuracy = 0.9384615384615385
Precision = 0.9071428571428571
Recall = 0.975

And here is the 10 most predictive features for GaussianNB classifier.

- Robbery
- Murder
- Theft
- Burglary
- SpeedyTrial
- Smuggling
- Narcotics

- Kidnapping
- ArmsAct
- Explosive

4.5.2 LinearSVC

We used LinearSVC(Linear Support Vector Classification) to learn a linear Support Vector Machine model to predict highCrime. In this learning method results are here:

Accuracy = 0.9615384615384615
Precision = 0.9371428571428572
Recall = 0.975

4.6 Difference between learning methods

In this section we will discuss about which learning methods gives us better results and difference between them. In the above sections we saw the results of accuracy, precision and recall of each learning methods. So in here we will see the visualization of learning methods differences. First we compare between DTC vs GaussianNB and DTC vs LinerSVC learning methods results.

4.6.1 10-fold cross validation vs GaussianNB

In the section of 4.4.4 and 4.5.1 we discussed about results of 10-fold cross validation using DTC and GaussianNB respectively. So here is graphical results of them:

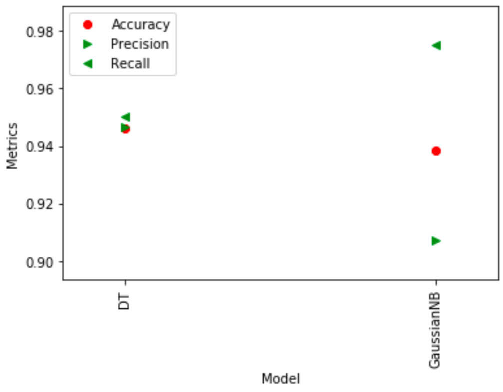


Figure 4.3: 10-fold using DTC vs GaussianNB learning method

Here we can see that 10-fold cross validation and GaussianNB learning method result figure. In this we can see that accuracy for both learning methods are more likely same but 10-fold cross validation is better in precision where GaussianNB is better in recall.

4.6.2 10-fold cross validation vs LinearSVC

In the section of 4.4.4 and 4.5.2 we discussed about results of 10-fold cross validation using DTC and LinearSVC respectively. So here is graphical results of them: Here we can see that 10-fold cross validation and LinearSVC

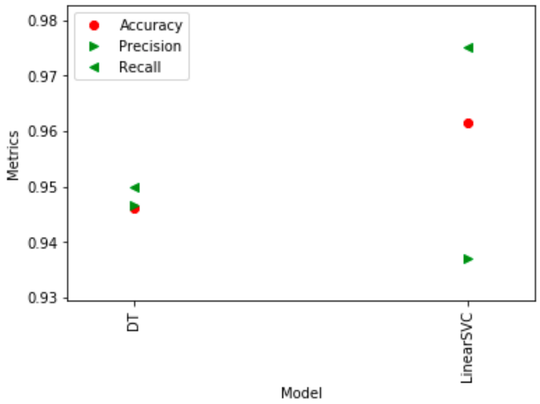


Figure 4.4: 10-fold using DTC vs LinearSVC learning method

learning method result figure. In this we can see that accuracy for LinerSVC is gives much better result than 10-fold and it is also better in precision where 10-fold cross validation is better in recall.

4.6.3 Learning method results

In all learning methods there are accuracy, recall and precision. So figure 4.5 shows the results of DTC, 10-fold cross validation, GaussianNB and LinearSVC. Here we can also see that DTC gives better accuraacy than other learning methods. And for recall DTC and GaussianNB gives same better results than other two learning methods. DTC and LinearSVC gives better results for precision. So we can conclude that DTC gives promising results in every fields.

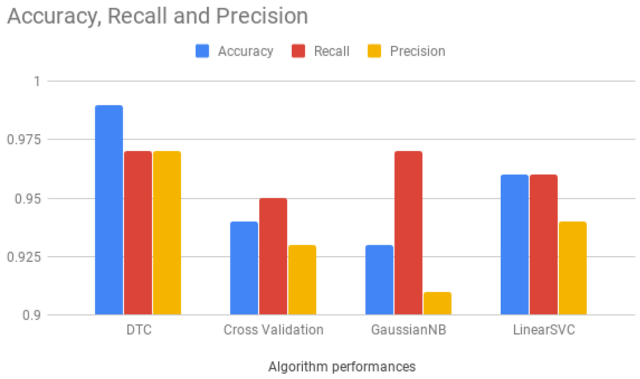


Figure 4.5: All learning method results

4.7 Mean Squared Error (MSE)

We used linear regression to learn a linear model directly predicting the crime rate on ViolentCrimesPerPop field. So by using 10-fold cross-validation, there is an estimated meansquared- error (MSE) of the model and here is the result of that model:
Estimated meansquared- error (MSE) of the model is', 3.5104200570706014e-05

	Linear regression	Ridge regression	Polynomial regression
10-fold	3.5104e-05	0.0002	0.03431
Training set	1.4526e-05	0.00	0.16

Here we used Linear regression, Ridge regression and Polynomial regression to find out the MSE of 10-fold cross validation and training set.

4.8 Computational time

In this section we discussed about computation time taken by all four learning methods. Here is table which shows us time taken of each learning methods.

	DTC	10-fold CV	GaussianNB	LinearSVC
Time taken	0.009	0.135	0.120	0.139

In this table we can see that DTC(Decision Tree Classifier) takes less time than other learning methods.

Chapter 5

Conclusion

Summary of this report is that how our model analysis and predict crime rate and also the differences between learning methods. We had to go through many things to do these things. Analyzing and predicting crime is an important and also interesting fact to work.

We introduced various machine learning algorithm(mainly supervised learning) used in crime prediction, these technique are capable for moreover analyzing and predict the future crime but they also have some disadvantage. In the future work we will try to overcome disadvantages of these technique. There has been some work done already on this topic, but not in Bangladesh crime statistics. So we took the chance and made some model to check the results.

In the end, we develop a more sophisticated model that more fully captures the dynamics of crime. These types of modeling enhancements may lead to better predictive power. We spend a good amount of time tweaking different model ideas, and measuring their performance. We have applied various models including: Decision Trees, Gaussian NB, Regression, Linear SVM. We also performed the 10 fold cross validation. The results are different and we have plotted results based on the metrics for the different models. It can further be tested on many models to identify the best that can be used to predict the crime rate.

5.1 Future work

Since we worked on various models for analyzing the data but here are some points for future work based on this topic:

- Using Polynomial Kernel and K means learning methods
- More details work on Regression
- Using Gradient Boosting Classifier learning methods
- Random Forests Classifier learning methods

Reference

[1] Bangaldeh police crime statistical data (www.police.gov.bd) , 2018. [Online; accessed 15-October-2018].

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Paul N Bennett. Assessing the calibration of naive bayes posterior estimates. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2000.

[4] Patricia L Brantingham and Paul J Brantingham. A theoretical model of crime hot spot generation. *Studies on Crime & Crime Prevention*, 1999.

[5] Julie Berry Cullen and Steven D Levitt. Crime, urban flight, and the consequences for cities, 1999.

[6] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2-3):103–130, 1997.

[7] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[8] Halvor Mehlum, Karl Moene, and Ragnar Torvik. Crime induced poverty traps. *Journal of Development Economics*, 77(2):325–340, 2005.

[9] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[10] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

[11] Lior Rokach and Oded Maimon. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487, 2005.

[12] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544, 2001.

[13] David Weisburd and Lorraine Green. Defining the street-level drug market. 1994.

[14] Harry Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.

[15] Tong Zhang and Frank J Oles. Text categorization based on regularized linear classification methods. *Information retrieval*, 4(1):5–31, 2001.