CRIME RATE ANALYSIS USING MACHINE LEARNING TECHNIQUE

BY

JERIN TASNIM
ID: 191-15-2772
AND

NADIA RAHMAN
ID: 191-15-2763

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mushfiqur Rahman
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Md. Sabab Zulfiker
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled **"CRIME RATE ANALYSIS USING MACHINE LEARNING TECHNIQUE"**, submitted by **Jerin Tasnim**, ID No: 191-15-2772 and **Nadia Rahman**, ID No: 191-15-2763 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *30/01/2023*.

BOARD OF EXAMINERS

Chairman

31/01/23

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Ms. Sharmin Akter
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mushfiqur Rahman, Senior Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

01.02.23

Mushfiqur Rahman
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Sabab Zulfiker
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Jerin Tasnim

Jerin Tasnim
ID: 191-15-2772
Department of CSE
Daffodil International University

Nadia Rahman

Nadia Rahman
ID: 191-15-2763
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Mushfiqur Rahman**, **Senior lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Machine learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan,** Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

A method for methodically detecting crime, analyzing crime patterns, and anticipating crime trends is crime analysis. The information gleaned from machine learning is of great use to police officers and can be applied to a large number of crime datasets. This issue could be resolved by utilizing a Random Forest in security analysis and law enforcement. Since the Random Forest algorithm has been cited as the most effective machine learning algorithm for predicting crime data, this work investigated the construction of a prototype model for crime prediction using the Random Forest algorithm. In addition to displaying criminal offense areas within a region, our algorithm is able to identify and forecast locations with a high likelihood of occurrence. The experimental results show that the Random Forest was able to correctly identify the unknown category in the crime data by 0.82, which is good enough to trust the system for predicting future crimes. This method's results can be used to raise awareness of risky areas and assist law enforcement in predicting future crimes in a particular area within a given time frame. Due to the expanding use of computerized and informational systems, data analysts of crime may be able to assist police departments in speeding up the process of solving crimes in our society. The machine learning system is simple to set up and works with the spatial plot of crime and criminal activities to improve the performance of our police and other government agencies. The Bangladesh police can reduce crime and solve cases as quickly as possible by implementing this developed system.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Everyone's quality of life and economic activity are impacted by crime, which is a widespread social issue. The application of various classification algorithms to real-time crime data and the comparison of their accuracy in predicting crimes will be the primary focus of this area of crime forecasting [1]. Crime is social oppression that is thought to be bad for citizens' well-being [2]. It is a prohibited act because it is deemed detrimental to public welfare [3]. This method of forecasting can be created using both qualitative and quantitative approaches. The long-term nature of criminal activity can be recognized using qualitative methods like environmental scanning and situation composition [4]. It is thought to be a crucial factor in deciding whether people move to a new town or not and which places to avoid when traveling. Police departments are looking for cutting-edge spatial data systems and cutting-edge data processing methods to boost crime analytics and community safety as crime rates rise. Despite the fact that crimes can happen everywhere and at any time, it’s extremely common for gangsters to take a shot at opportunities and openings they come across in places that are highly known to them. We will in general wish to increase people’s awareness of dangerous areas in any bound eras by providing a machine learning technique to discover and find the main criminal minded hotspots and comprehend the type, location, and time of any carried-out wrongdoings. Furthermore, having this type of information might push people to improve their living location choices in a particularly large city. Police authorities, on the other hand, will use this objective to expand the scope of potential crime and search for ways to prevent it before it occurs. It will promote the internal circulation of police powers at the most obviously wrongful locations for an indefinite period of time, allowing for conservative and highly productive use of police resources. We usually prepare to make our location more secure for the individuals who live there and furthermore, for other people who can go there, by having the majority of this information out there connected to wrongdoings. Criminals will decide to commit a

crime if they are aware that there is a system in place that can forecast their motives and actions even before they do them. It's a fantastic weapon for fighting and combating modern-day criminals, thanks to the data and information revolution. There are many tools available for machine learning, for this research Google Colab is chosen. It is an open source platform [10]. The objective of this research is to apply a suitable machine learning algorithm on crime data to predict the likelihood of a county having low, medium or high violent crimes.

1.2 Motivation

In recent years, every political party and government in the world that aims to reduce the rate of crime views security as a crucial aspect. Hoodlums and wrongdoing danger for the general public in all pieces of world for quite a while now since ascent of human civilization and precautionary measures are expected to eliminate violations from our local area. Our goal is to use crime and criminal activity prevention applications to keep the general public safe. Current police tactics involve searching the scene of the crime and the perpetrators, typically after the incident has occurred. However, we can use crime history data to create a usable pattern and use these patterns to predict crimes before they occur with the assistance of technological and informational advancements.

In the current situation, criminals and guilty parties are progressing automatically and effortlessly committing wrongdoing. The challenge posed by analysts and police organizations is that it is difficult to study a huge amount of information about wrongdoing and psychologically violent activities. As a result, law enforcement agencies need to be limited in their ability to develop new methods to prevent wrongdoing, catch criminals, and stay ahead of the criminals and the police. In order to comprehend this, appropriate procedure must choose to conduct a wrongdoing investigation based on the data uncovered. As the term "information handling" refers to the extraction or mining of data from enormous amounts of data, it is used in this context on a massive wrongdoing dataset, and information obtained from information preparation systems is useful and supports legal implementations.

The Naïve Bayes algorithm forecasts a test case just in time rather than building a model. It has a few advantages over other bunch algorithmic programs, like the fact that it uses machines with less machine unpredictability and is good for large group data sets. In that limit, it's been with progress utilized in different explanation and looks at limits, and also feature division, workstation vision, geo statistics, urbanology and agribusiness. It is frequently used as a preprocessing task for available elective calculations, such as evaluating a starting setup. Because the nature of wrongdoings varies greatly and wrongdoing data are typically packed with uncertain violations

1.3 Rationale of the Study

When conducting an investigation into wrongdoing, the appropriate information mining strategy should be chosen, and grouping is a related information mining strategy. A group of articles in particular means that questions in the same group are much more alike than those in elective groups and involve a variety of calculations that have a significant impact on participants conceptions of what constitutes a group and how to effectively recognize them. K indicates that this examination employs group approach information mining to remove supportive learning from the large dataset of wrongdoing. to translate the dataset that aids law enforcement in recognizing and breaking down examples of wrongdoing in order to anticipate wrongdoing and reduce the number of instances of wrongdoing. Additionally, this provides information to minimize future wrongdoing by anticipating it. During this investigation, the K mean group is supported by open-source information preparation tools that are diagnostic tools used to examine data. We choose Goggle Colab over the new open source information handling suites like KNIME, WEKA, R, ORANGE, and quick excavator. K means that the whole thing is done with the help of the fast mineworker device, which is a Java-based open-source science and data preparation PC code with a variety of flexible data handling support options. In a similar vein, a dataset containing records kept by the police in the United States from 1990 to 2011-12 is used for wrongdoing investigations. During this investigation, the wrongdoing committed by a human by killing another human is being investigated.

In areas where wrongdoing is risky, grouping calculations are being used to investigate and anticipate wrongdoing. There are a few grouping calculations that can be used to group the important data into the groups that are needed for the examination. The unpredictability of connections between these types of information has also made human science an excellent field for implementing and evaluating information handling strategies, which has resulted in the enormous volumes of wrongdoing data set. Human science could be defined as the logical examination of wrongdoing, criminal behavior, and its application. It is a method that anticipates locating criminal characteristics and wrongdoing. It is one of the most fundamentally important fields in which the learning mining approach will produce fundamentally important results. This method of foreseeing violations is based on the activity of recognizing characteristics of wrongdoing. The knowledge and data gathered through the process of handling information could be an extremely useful tool that could support and encourage legal requirements. This method of bunching takes a dataset and divides it up into groups that are further analyzed to find unsafe areas for wrongdoing. These groups show the areas of violations on the outside, overlaid on a guide for the requirements of the law. Along the edges of various subtle elements of wrongdoing characteristics like type and time, groups store wrongdoings. On the preface of their distinctive individuals, these bunches are arranged and arranged in a particular order. Thinly possessed groups are also prone to wrongdoing, but fewer individuals are noticed. With wrongdoing groups supported by wrongdoing hazardous areas, measures are taken to avoid, used, and upheld predictable.

1.4 Research Question

This report summarizes the research's objectives as follows:

1. Decide and inspect investigation and different review that address the long run of wrongdoing, with a focus on the initial twenty years of the twenty first 100 years.
2. Find the individuals and communities in Bangladesh and around the world who have developed a crime prediction system, examine their objectives and methods, the nature of their analysis, and the methods, instruments, and tools they employed for analytics.

3. The present analysis and forecasts are made by analyzing the strengths and weaknesses.
4. Analyze the organizations in North America, the United States, and Australia to compare Bangladesh's capabilities and practices to those of other nations.
5. Provide the Police Force with suggestions for additional research and measurements to be carried out in Bangladesh.
6. Utilize machine learning technique to develop a prediction model and anticipate criminal activity in advance.

1.5 Expected output

The rate of wrongdoing is rising steadily and impressively. The problem with wrongdoing is that it can't be predicted because it doesn't happen in a planned or irregular way. In addition, cutting-edge innovations and approaches encourage criminals to effortlessly accomplish their goals. Regarding the Crime Records Bureau, crimes like robbery, riot, and other legal offenses are reduced while crimes like woman harassment, murder, and assault are increased. We will anticipate the location where chance wrongdoing predominates, despite the fact that it is intolerable to foresee who could also be the victims of wrongdoing. Although the anticipated outcomes cannot be guaranteed to be 100% accurate, they do demonstrate that our application reduces the rate to an accurate level by providing security in areas where wrongdoing is likely. Law enforcement agencies can use the framework, and they will be ready to oil the areas where they see the most wrongdoing more often to stop a wrongdoing that was predicted using our framework.

Using our framework, the areas that are prone to wrongdoing and where the wrongdoing event is anticipated can be depicted graphically using a warmth outline to show movement level. Typically, low action is represented by lighter shading, while high action is represented by darker shading.

1.6 Project Management and Finance

There are a couple restrictions through this project that we hope to get past in the future. We employed feature selection approaches and fewer classifiers that we would like to work in the future.

Finance: Our research project finance by Daffodil International University.

1.7 Report Layout

There are five chapters in total in this report. In the first, we have described the project's purpose and goal, as well as its summary and introduction. We have provided crime data related to incidents to validate our decision to select this project and provided a comprehensive explanation of why we chose it.

In chapter two, we talked about the background of this research, previous related works in the same field, and how it was put into practice successfully. Additionally, we have described the difficulties encountered during the research and how they were overcome.

Chapter three provides an in-depth explanation of our research's methodology, with illustrations and diagrams illustrating the various methods we employed.

The discussions and findings of our research are presented in Chapter 4. In conclusion, the findings of our study and recommendations for future enhancements are presented in chapter 5.

CHAPTER 2

Background study

2.1 Preliminaries

In the study and analysis of criminology, machine learning can be divided into two main areas: crime control and crime concealment. While criminal suppression attempts to capture a criminal by utilizing his or her history that has been recorded in machine learning, crime control typically makes use of knowledge that is derived from the analyzed data in order to control and prevent the occurrence of crime. The management, fusion, and translation of information from various sources made use of data combination. The main reason was to get rid of the confusion caused by contradictory reports and backgrounds that were cluttered or loud. As a result, large databases were searched for patterns and connections using machine learning. Methods for locating and preventing crime are connected to a variety of applications, including cross-border security, internet security, and violations at home. Utilizing the recurrence event of incidents, a strategy to use computer log records as history data to examine a few connections was proposed. After that, they looked at the result and created performance-enhancing profiles.

2.2 Related Works

In the study of real crime data and execution of predictive methods in this dataset, crime analysts can predict crime zones [1]. This particular paper use for this dataset, they built some regression models and achieved accuracy for Linear Regression is 0.83, Polynomial Regression is 0.95 and for Random Forest Regression is 0.90. Therefore,they observed that polynomial and random forest regression is better than linear regression in predicting crime trends and patterns [2]. The methodology is divided into two parts, these are dataset description and model selection. By using this methodology, the author can predict crime trends. A comparable concept we're using in our investigation is to find unusual cases from known data and datasets. [3] From this

paper, we can learn about different approaches on crime predictions. Here are all the approaches related to research on predicting crimes.[4] Their main goals were to model and create wrongdoing-based events, investigate the use of wrongdoing-based events in improving order and grouping, develop an intelligent recovery framework for news related to wrongdoing, and visualize news related to violations in a powerful and understandable way, combine them into a useful, strong, and trustworthy frame- work, and evaluate the framework's ease of use and execution, and the study will contribute to a deeper understanding of the facts associated with wrongdoing [5]. Examines the use of group inquiry in the provided space, with a focus on inconsistency finding during review. The goal of his investigation is to see how bunching technology can be used to automate misrepresentation filtering and identify through the issues associated with submitting extortion during a review. He used group investigation to help evaluates focus their efforts when evaluating their group life insurance claims, which is suspicious. Introducing integration methods, it is not an easy task to find the right structure for a particular database. Many forecasting ideas have been proposed to address the problem of predicting machine learning in India. [6]. This study provides guidelines for finding the pattern of crime, crime reporting and criminal justice proposal procedures. The feasibility of using machine learning techniques to build such platforms has been explored earlier. [7]. The primary goal of this research is to demonstrate how law enforcement agencies may utilize machine learning to detect, predict, and solve crimes at a significantly faster pace, therefore lowering the crime rate. [8] The machine learning techniques are used to investigate the city's crime information from the Police Department by using a bunching/group-based model to predict wrongdoing trends. The consequences of this machine learning might be effectively used to reduce and even prevent misbehavior in the future. [9]. There is a strong body of evidence to support the theory that crime is predictable (in the statistical sense) mainly because criminals tend to operate in their comfort zone [10]. The re- search focused on developing a wrongdoing examination tool for the Canadian scenario using a unique information mining method that may assist police divisions in conducting viable and successful wrongdoing investigations. [11]. As crime prediction, the authors suggested using technology to obviate criminal activities, and they have gathered a huge amount of data to research this

issue. [12]. For this paper we find datasets accuracy are Support Vector machines - 0.573, Random Forest- 0.6723 and Gradient Boosting Machines- 0.6423. Second dataset accuracy are Support Vector machines - 0.7939, Random Forest- 0.6579, Gradient Boosting Machines- 0.6167 and Neural Network 0.7402. [13] The analysis of the crime data will help decision making agencies to take precautionary steps to control the crime rate over demographic places. [14] Different approaches connected with statistical learning techniques result in predictive database models. That will help us to take necessary action against crime. [15]

2.3 Comparative Analysis and Summary

This section provides an overview of the research conducted and the findings that serve as a model for predicting wrongdoing for the underlying decades of the 21st century. Additionally, a specific breakdown of the ways in which wrongdoing's scope and character will shift has been provided. Crime is expected to decrease, both quantitatively and subjectively. While this outline examines how future wrongdoing rates may differ from those of the past, the majority of its focus is on how the concept of wrongdoing is expected to evolve in the future. This section begins by summarizing the most important variables, specifically statistical, comprehensive financial, and innovative variables, that research indicates will influence the scope and nature of wrongdoing in the future. Following that, a brief summary of research that has predicted specific future wrongdoing rates is presented. A survey of expectations associated with the concept of wrongdoing follows this, with a focus on future patterns and anticipated improvements that will recognize wrongdoing in the twenty-first century. A survey of expectations associated with the concept of wrongdoing follows this, with a focus on future patterns and anticipated improvements that will recognize wrongdoing in the twenty-first century. This study looks at products and services that will be targeted by criminals in the future, specific types of new wrongdoings that are expected to grow in the future, the extent to which wrongdoings will affect society in the future, and a profile of exploited people and future criminals.

Despite the fact that there is no specific or authoritative explanation for why the rate of wrongdoing has decreased over the past few years or why it began to significantly increase in circumstances such as the 1960s, there are a variety of factors that influence the type of wrongdoing.

However, the two most important factors that have had a significant impact on criminal activity rates in the past and are expected to have a significant impact on future rates are large-scale financial variables such as the quality of the economy, consumer spending, unemployment rates, and the number of men in the criminally inclined age group.

Technology:

Innovation is likely to have a significant impact on the nature and scope of both current and future wrongdoing, in addition to statistical and large-scale financial factors, which are essential for determining the reasons for wrongdoing in the future. The effect of development on the inevitable destiny of bad behavior and criminal activities can be depicted into three general characterizations:

1. Misrepresentation, burglary, tax evasion, and falsification are just some of the common crimes that will continue to be encouraged by technological advancements.

2. Technology itself will be the target of criminal acts and activities (such as the theft of telecommunications services and the spread of infections).

3. To stop or change criminal assault, new technology will be used.

The development of products like personal computers is one of the primary reasons for the rise in fraud in recent years. The openness of criminals has increased as a result of these innovations' lower costs and accessibility to businesses. As a result, a greater number of people, novices and experts alike, now have the opportunity to engage in a variety of misrepresentation-related offenses that used to be the domain of exceptionally skilled forgers or falsifiers.

2.4 Scope of the Problem

Using data analysis, machine learning algorithms, and ensemble techniques, this project is contributing to the development of a model that can predict crime location. We can use this model to identify the location where most of the crime happen, which will help society deal with those problems.

As a result, we considered developing a model that could anticipate crime locations.

2.5 Challenges

The fact that crime and crime rate forecasting is a novel, high-tech innovation that is still in the early stages of development is the most tedious aspect. It is currently in use in China, the United States, the UK, and some parts of Europe. However, the outcomes have not always been positive. There are several purposes for it. Since algorithms have not yet matured and developed, they are susceptible to errors. If the data pertaining to crime is inconsistent, it is impossible to accurately predict crime patterns. In the end, this results in inaccurate predictions and forecasts of crime. At the moment, experts believe that predictive crime prevention is not yet universally accepted. It is only a matter of time before data science and artificial intelligence become an essential component of general law enforcement given that they are already utilized in the defense of numerous nations. There are numerous interdependencies in crime prediction and forecasting at this time.

CHAPTER 3

Research methodology

3.1 Research Subject and Instrumentation

Google Colab: Colab is essentially a totally mist based, free Jupyter sketchbook atmosphere. Most significantly, Colab doesn't essential to be established, and the sketchpads you generate can have multiple team associates editing that one at once, alike to how you run documents in Google Docs dictionary. The fact that Colab cares the most broadly castoff machine learning collections and that they are modest to weight onto your notebook is its greatest advantage.
It's a held Jupyter notebook with a countless allowed kind that delivers free admission to Google dispensation incomes like GPUs and TPUs and necessitates no arrangement.

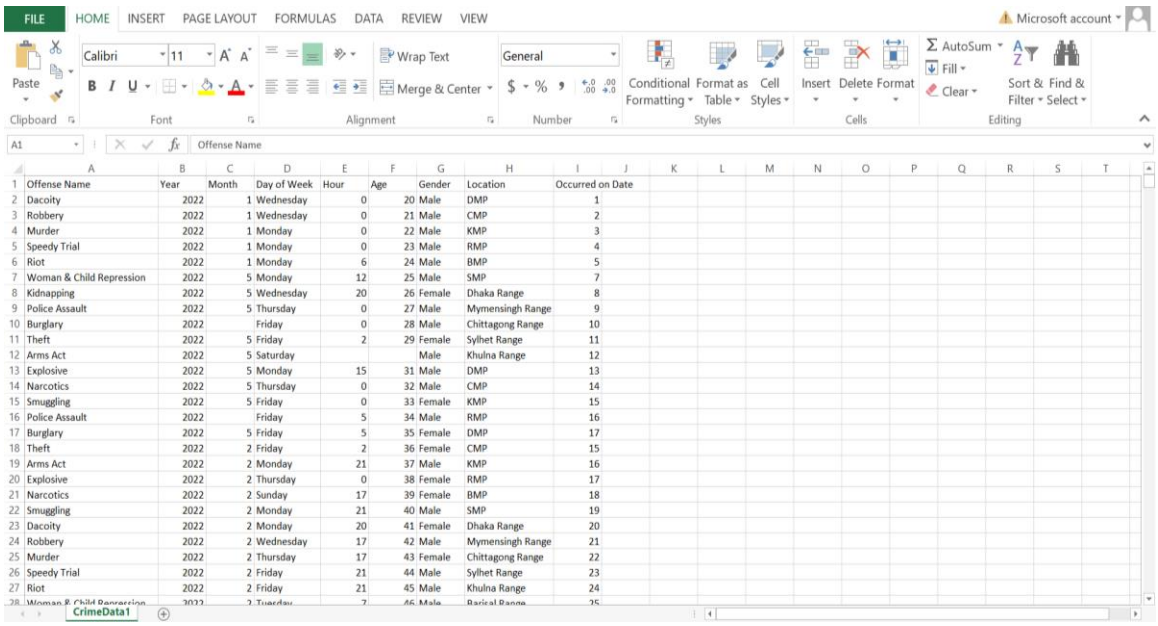
Kaggle: Kaggle is a Jupyter Notebooks environment that doesn't need to be set up and can be changed. You can get free GPUs and a huge repository of data and code that has been published by the community. Kaggle is an online community for enthusiasts of machine learning and data science.

Microsoft Excel: The spreadsheet program Microsoft Excel is a member of the Office product group for business applications. Microsoft Excel can be used to format, organize, and calculate a spreadsheet. One of Excel's primary functions is data entry. Accounting and data management Excel data types are the four distinct categories of values in Microsoft Excel. The four types of data are text, numbers, logical, and error data. With each type, you can play a variety of roles, so knowing which ones to use and when is very important.

3.2 Data Collection Procedure/Dataset Utilized

We are collecting the dataset from various sites. Bangladesh Police Crime report is downloaded in CSV format and use for research purpose. Total of 1400 individual data is used to run the project.

In figure-3.1, it shows the dataset in CSV format-



	Offense Name	Year	Month	Day of Week	Hour	Age	Gender	Location	Occurred on Date
2	Dacoity	2022	1	Wednesday	0	20	Male	DMP	1
3	Robbery	2022	1	Wednesday	0	21	Male	CMP	2
4	Murder	2022	1	Monday	0	22	Male	KMP	3
5	Speedy Trial	2022	1	Monday	0	23	Male	RMP	4
6	Riot	2022	1	Monday	6	24	Male	BMP	5
7	Woman & Child Repression	2022	5	Monday	12	25	Male	SMP	7
8	Kidnapping	2022	5	Wednesday	20	26	Female	Dhaka Range	8
9	Police Assault	2022	5	Thursday	0	27	Male	Mymensingh Range	9
10	Burglary	2022		Friday	0	28	Male	Chittagong Range	10
11	Theft	2022	5	Friday	2	29	Female	Sylhet Range	11
12	Arms Act	2022	5	Saturday			Male	Khulna Range	12
13	Explosive	2022	5	Monday	15	31	Male	DMP	13
14	Narcotics	2022	5	Thursday	0	32	Male	CMP	14
15	Smuggling	2022	5	Friday	0	33	Female	KMP	15
16	Police Assault	2022		Friday	5	34	Male	RMP	16
17	Burglary	2022	5	Friday	5	35	Female	DMP	17
18	Theft	2022	2	Friday	2	36	Female	CMP	15
19	Arms Act	2022	2	Monday	21	37	Male	KMP	16
20	Explosive	2022	2	Thursday	0	38	Female	RMP	17
21	Narcotics	2022	2	Sunday	17	39	Female	BMP	18
22	Smuggling	2022	2	Monday	21	40	Male	SMP	19
23	Dacoity	2022	2	Monday	20	41	Female	Dhaka Range	20
24	Robbery	2022	2	Wednesday	17	42	Male	Mymensingh Range	21
25	Murder	2022	2	Thursday	17	43	Female	Chittagong Range	22
26	Speedy Trial	2022	2	Friday	21	44	Male	Sylhet Range	23
27	Riot	2022	2	Friday	21	45	Male	Khulna Range	24

Figure 3.1: Dataset in CSV format

3.3 Statistical Analysis

- Two types of data present in dataset such as Categorial and Numerical.
- Categorical data are Location, Gender and Day of Week.
- Numerical data are Month, Year, Age, Hour and Occurred on Date
- Microsoft excel used to save Dataset in extension is CSV.

3.4 Proposed Methodology/Applied

Here are some steps to analyze the crime rate:

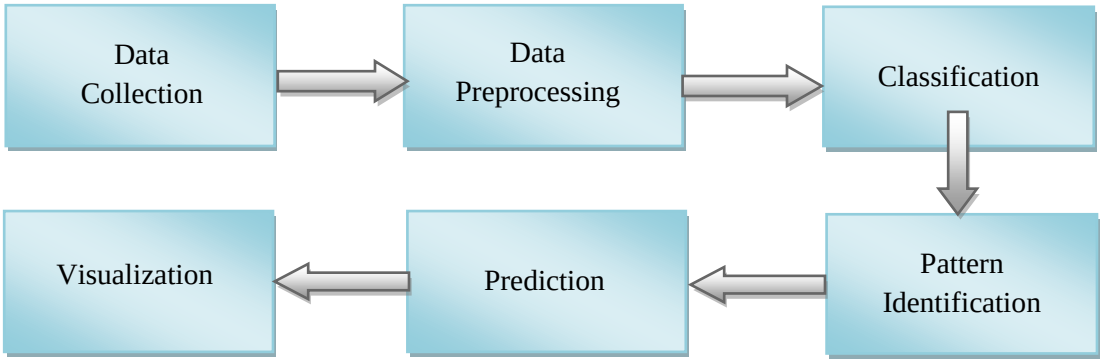


Figure 3.2: Stages in Crime Rate Analysis

Data Preprocessing:

Data preprocessing edits, rewrites, and organizes data for analysis. Classification often ignores some facts. This study's data cleaning and preparation are described below:

Handling Null Values: There are several null values in the columns. The values are replaced using mean-mode method. The arithmetic mean is obtained by adding the numbers and dividing the result by the total number of numbers in the list. The term "average" is typically used to refer to this. Mode is the value that appears most often in a list.

Replaced categorical values with numerical: Replaced the categorical values of the features to numerical value to train the data into the model. The converted values are: "DayofWeek":{'Friday':0,'Saturday':1,'Sunday':2,'Monday':3,'Tuesday':4,'Wednesday':5,' Thursday':6},'Gender':{'Male':1,'Female':0}.

Synthetic Minority Oversampling Technique (SMOTE): SMOTE is carried out to address the loan dataset's data imbalance issue. Synthetic Minority Oversampling Technique, or SMOTE for short, is a potent remedy for data imbalances. SMOTE, an algorithm for data augmentation, generates fictitious data points based on the real ones. SMOTE can be thought of as an enhanced form of oversampling or as a particular

method for enhancing data. With SMOTE, we produce artificial data points that are only slightly different from the actual data points rather than producing duplicate data points.

Implementation:

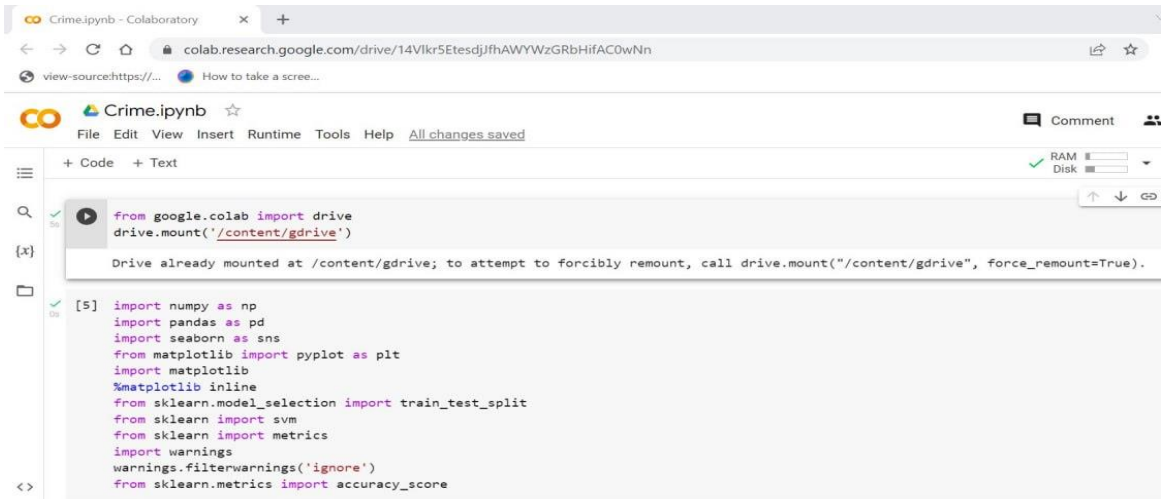


Figure 3.3: Mount Drive

	Offense Name	Year	Month	Day of Week	Hour	Age	Gender	Location	Occurred on Date
0	Dacoity	2022	1.0	Wednesday	0.0	20.0	Male	DMP	1
1	Robbery	2022	1.0	Wednesday	0.0	21.0	Male	CMP	2
2	Murder	2022	1.0	Monday	0.0	22.0	Male	KMP	3
3	Speedy Trial	2022	1.0	Monday	0.0	23.0	Male	RMP	4
4	Riot	2022	1.0	Monday	6.0	24.0	Male	BMP	5

Figure 3.4: Dataset Description

```
Offense Name          0
Year                  0
Month                 3
Day of Week           0
Hour                  1
Age                   1
Gender                0
Location              0
Occurred on Date      0
dtype: int64
```

Figure 3.5: Null value

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1400 entries, 0 to 1399
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Offense Name        1400 non-null   object
 1   Year                1400 non-null   int64
 2   Month               1397 non-null   float64
 3   Day of Week         1400 non-null   object
 4   Hour                1399 non-null   float64
 5   Age                 1399 non-null   float64
 6   Gender              1400 non-null   object
 7   Location            1400 non-null   object
 8   Occurred on Date    1400 non-null   int64
dtypes: float64(3), int64(2), object(4)
memory usage: 98.6+ KB
```

Figure 3.6: Information

	Year	Month	Hour	Age	Occurred on Date
count	1400.0	1397.000000	1399.000000	1399.000000	1400.000000
mean	2022.0	6.454545	13.907076	35.922802	14.359286
std	0.0	3.839954	7.816014	10.413403	7.549800
min	2022.0	1.000000	0.000000	20.000000	1.000000
25%	2022.0	2.000000	9.000000	27.000000	8.750000
50%	2022.0	7.000000	17.000000	36.000000	14.000000
75%	2022.0	11.000000	20.000000	42.000000	20.000000
max	2022.0	12.000000	23.000000	65.000000	31.000000

Figure 3.7: Statistical measure

```
[12]  # filling missing values for numerical terms - mean
      dt['Month'] = dt['Month'].fillna(dt['Month'].mean())
      dt['Hour'] = dt['Hour'].fillna(dt['Hour'].mean())
      dt['Age'] = dt['Age'].fillna(dt['Age'].mean())

[13]  dt.isnull().sum()

      Offense Name        0
      Year                0
      Month               0
      Day of Week         0
      Hour                0
      Age                 0
      Gender              0
      Location            0
      Occurred on Date    0
      dtype: int64
```

Figure 3.8: Null value replacement

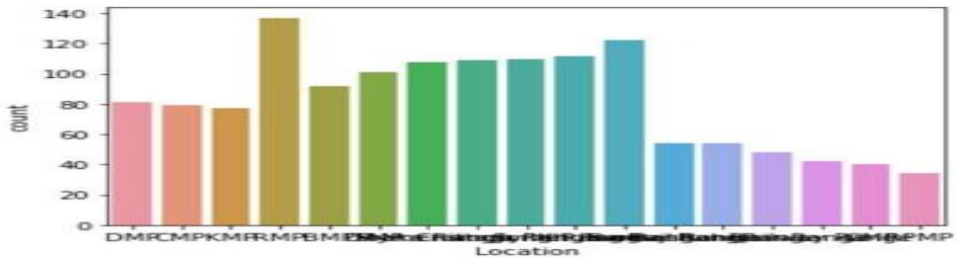


Figure 3.9: Before location visualization

Figure 3.10: After process location visualization

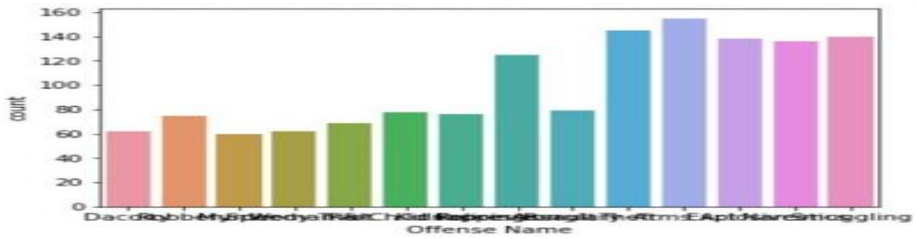


Figure 3.11: Before offense name visualization

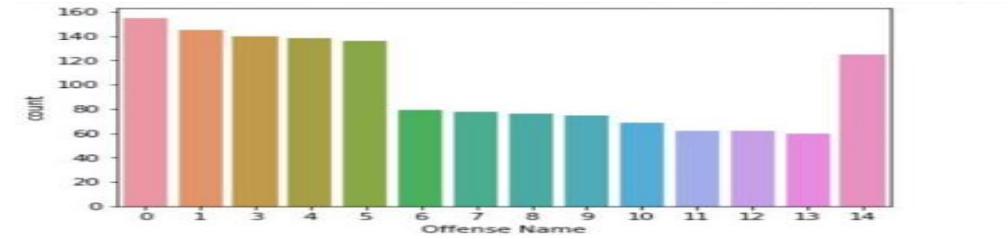


Figure 3.12: After offense name visualization

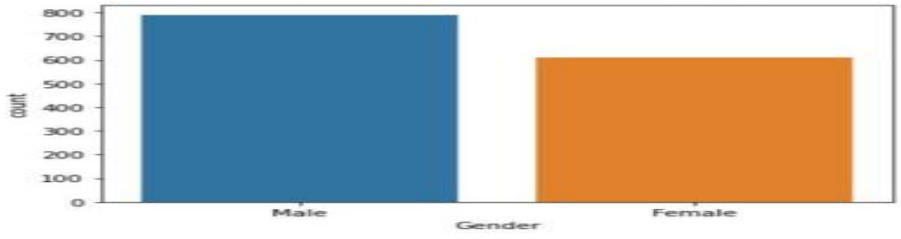


Figure 3.13: Before gender visualization

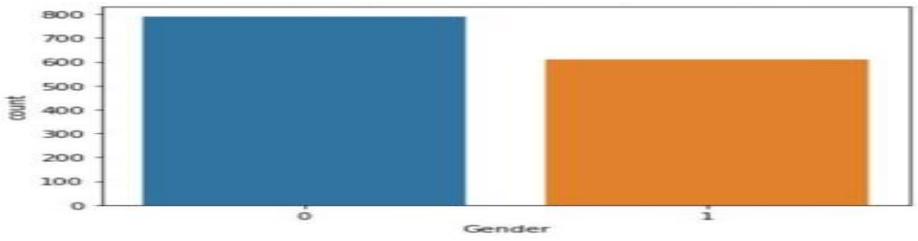


Figure 3.14: After gender visualization

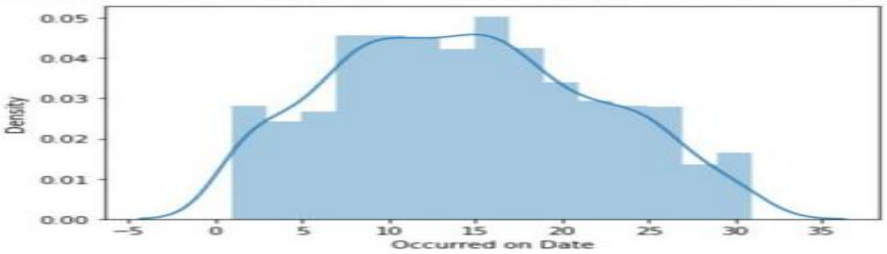


Figure 3.15: Before occurred on date visualization

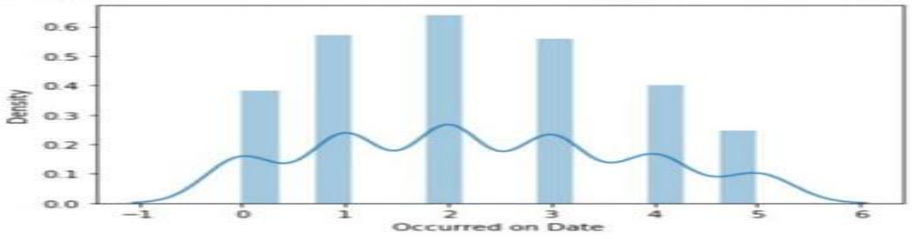


Figure 3.16: After occurred on date visualization

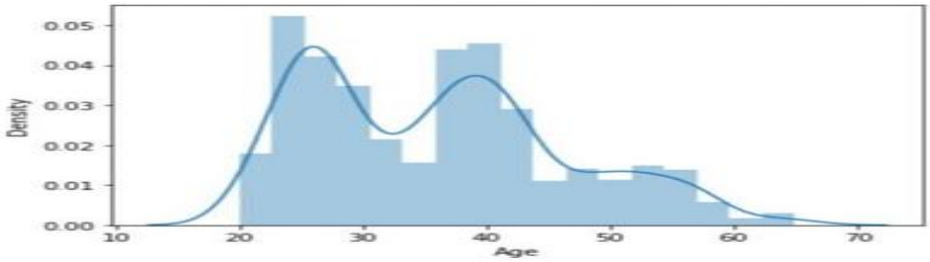


Figure 3.17: Before age visualization

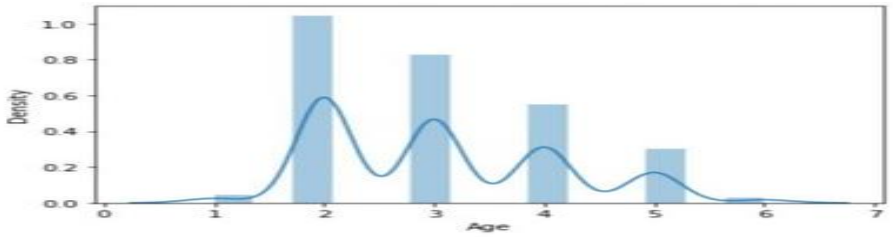


Figure 3.18: After age visualization

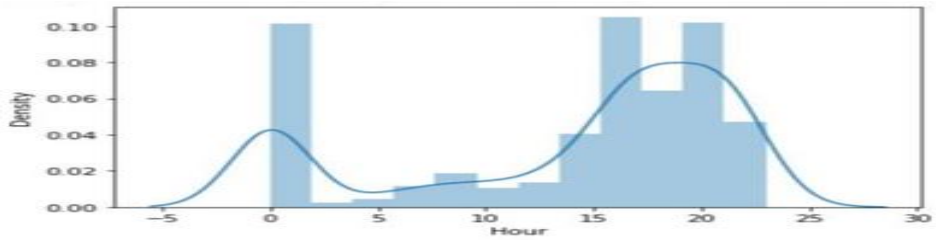


Figure 3.19: Before hour visualization

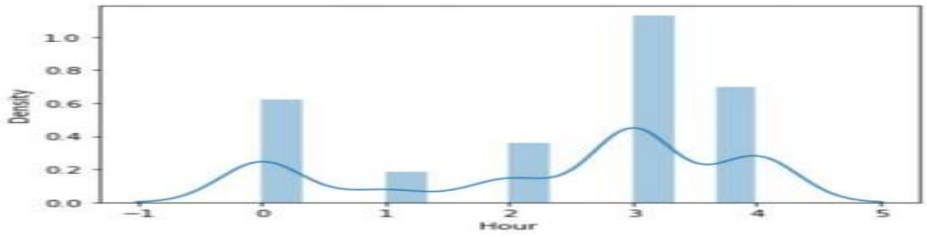


Figure 3.20: After hour visualization

Figure 3.21: Divide data

Figure 3.22: Train test split

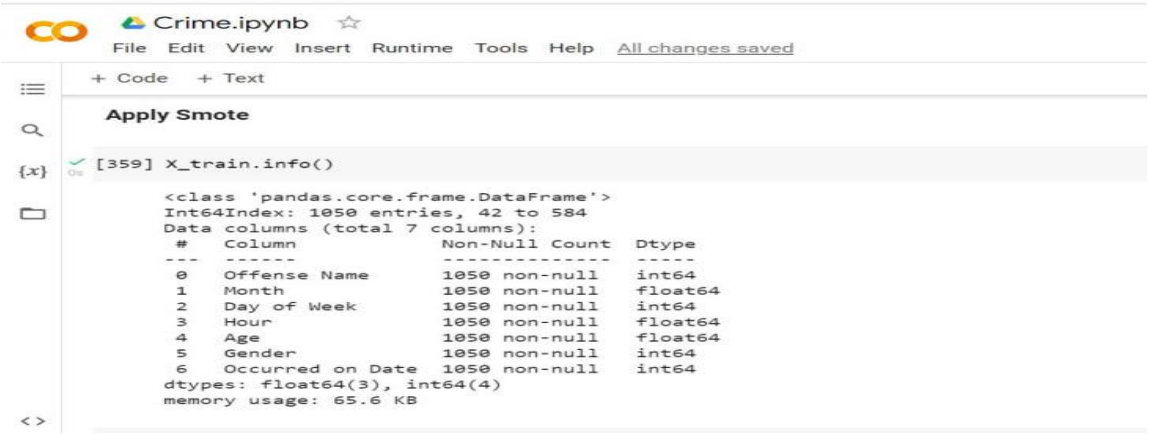


Figure 3.23: Apply smote

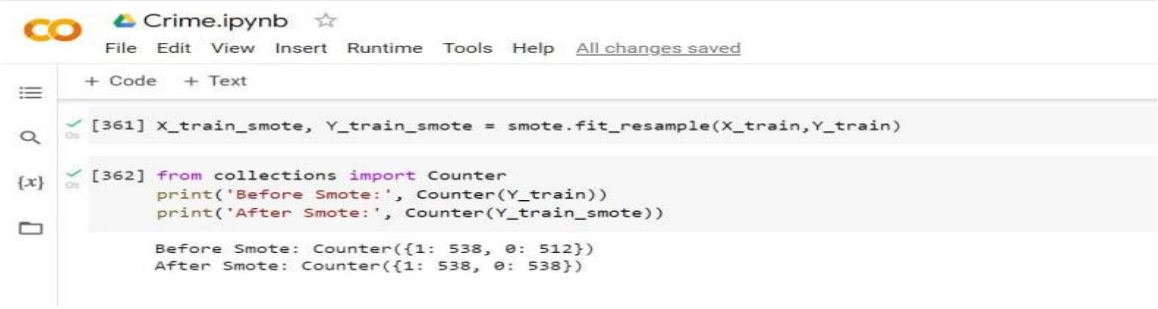


Figure 3.24: Apply smote 2

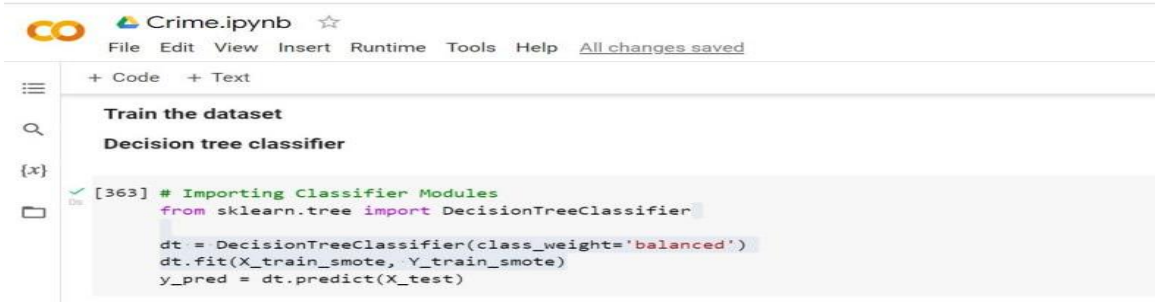


Figure 3.25: Train dataset with Decision Tree

Figure 3.26: Train dataset with Random Forest

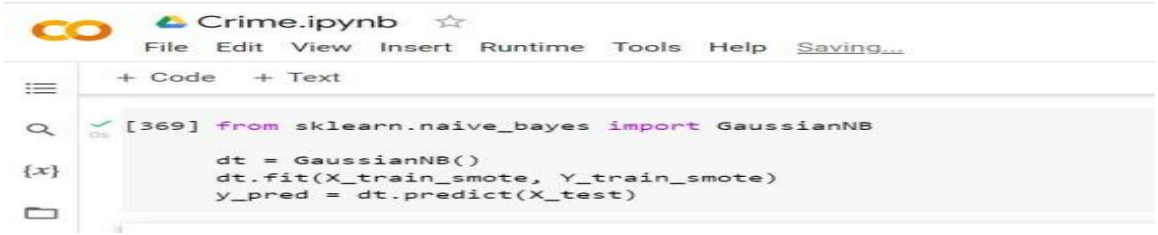


Figure 3.27: Train dataset with NaiveBayes

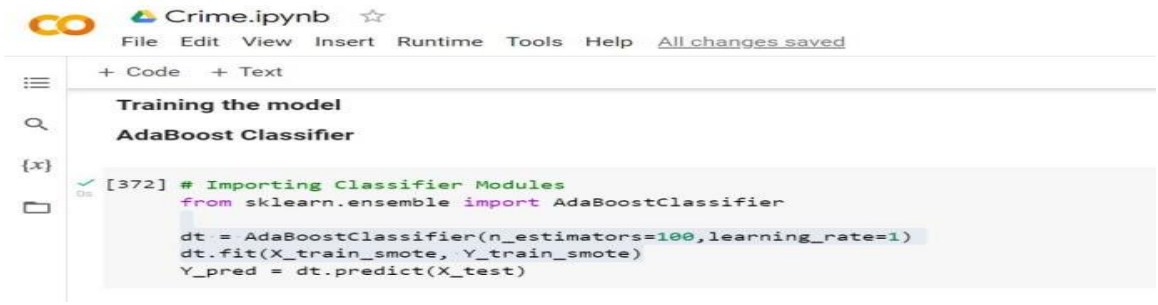


Figure 3.28: Train dataset with AdaBoost

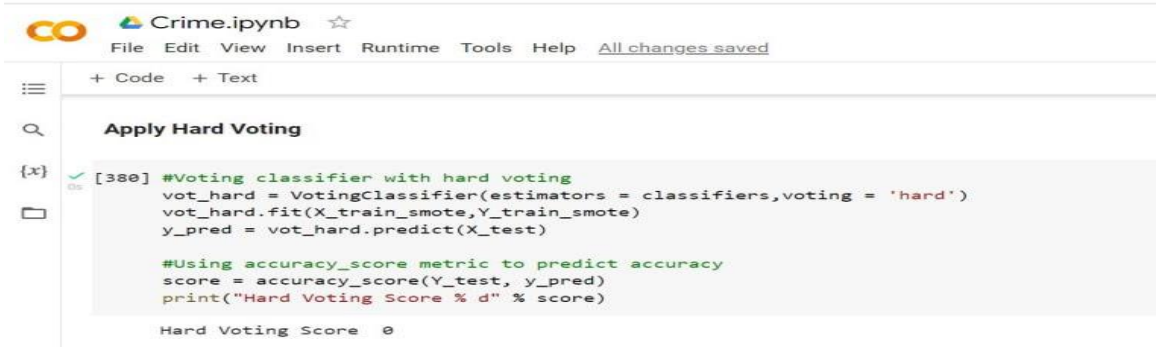


Figure 3.29: Apply hard voting

Classification:

The algorithm that we have included in our research paper are:

Decision Tree: We are making use of a technique known as a decision tree in order to produce the pattern. A decision tree is used as a predictive model in the decision tree algorithm. A root, an internal node, and an external node make up this method. The dataset, which is its own distinct set, is primarily characterized by the root. There is a distinct distinction between each internal node. The leaf hub gives the class name, which is the result. The Decision Tree is structured like a tree, with each node representing a test on a value of an attribute. Classification models are predicted by the leaves, which indicate classes or distributions of classes. Classes are denoted by feature combinations shown in the trees. Classification rules for the utilized dataset can be quickly generated using the tree structure, which has a lot of potential.

Naive Bayes: Naive Bayes classifiers are a group of basic probabilistic classifiers that are based on the Bayes theorem and machine learning feature independence assumptions. Problem cases are given class labels by the method, which are represented as vectors of feature values and selected from a finite set, using classifier models. The following are the dimensions of the large cube that makes up a Naive Bayes model: The value of the input field for discrete fields or the range of values for continuous fields is referred to as the name of the input field. Continuous fields are subdivided into discrete bins by the Naive Bayes algorithm. Target field value the frequency with which an input field value and a target field value occur together is tracked by a Naive Bayes model.

Random Forest Classifier (RF): A huge number of decision trees are created and combined to form a "forest" using the complex and flexible supervised machine learning method known as Random Forest. Problems involving classification and regression can be resolved with this.

AdaBoost Classifier: A meta-estimator known as an AdaBoost classifier begins by applying a classifier primarily to the initial dataset. After that, multiple copies of the classification model are applied to the same dataset. The weights of instances that were incorrectly classified are changed so that subsequent classifiers can concentrate more on challenging cases. The notable helping calculation AdaBoost expects to join numerous feeble classifiers into a solitary solid classifier. An object's class might not be accurately predicted by a single classifier; However, we can group a number of weak classifiers together to create a strong algorithm and have each weak classifier gradually learn from the incorrectly classified items of the others. This classifier can be used with any of our standard classifiers, such as Logistic Regression, which is frequently used as the default, and Decision Trees, among others.

Pattern Identification:

The third step is the example differentiating evidence stage, in which we must detect wrongdoing possibilities and designs. We use Apriori calculations to find wrongdoing designs that occur from time to time. Apriori may be used to choose affiliation criteria based on database patterns that are both conventional and generic. The wrongdoing design for a given location is the result of this step. We consider the features of that location, such as climate characteristics, zone affectability, proximity to criminal gatherings, prominent occasion, and so on. When another case occurs after receiving a general wrongdoing design for a location, and if it follows the same wrongdoing design, we may say that the region has a probability for wrongdoing occurrence. Data on designs is used by authorized authorities to successfully encourage assets. They can prevent misbehavior by hiring protection in criminally inclined areas, resolving criminal cautions/CCTV, and so on.

Prediction:

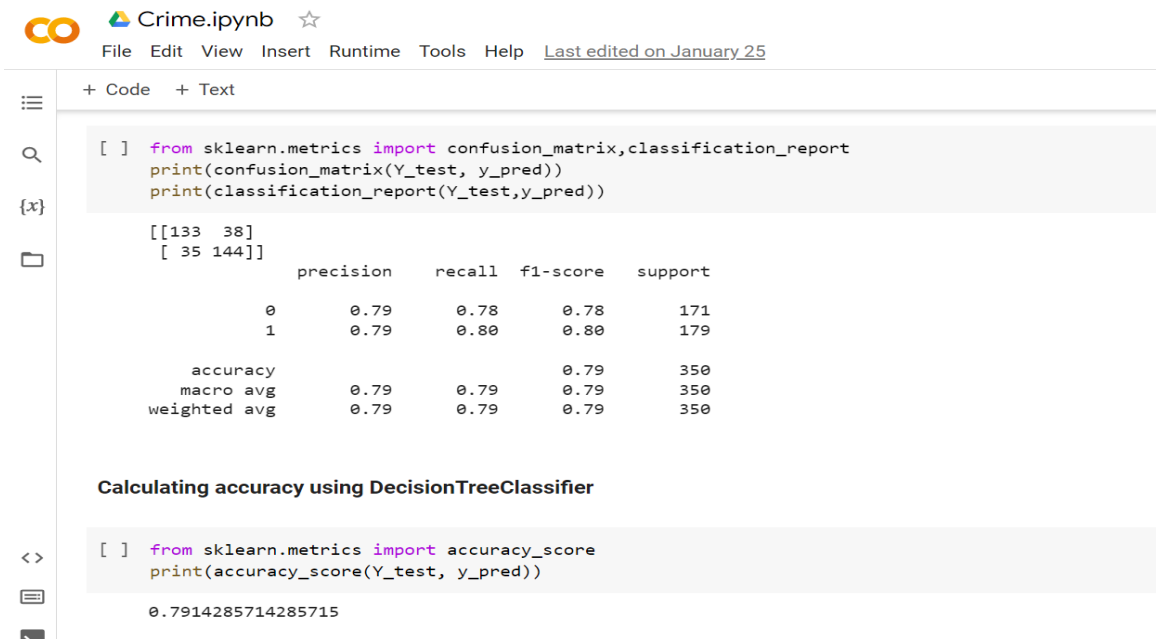


Figure 3.30: Decision tree performance

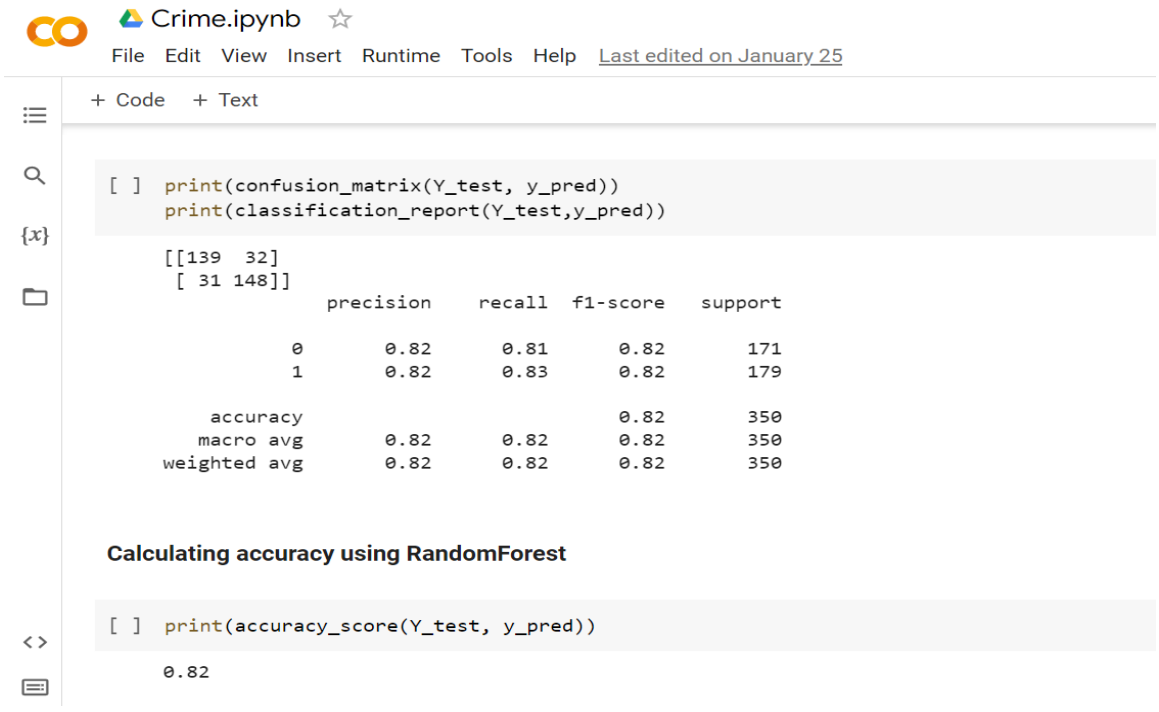


Figure 3.31: Random Forest performance

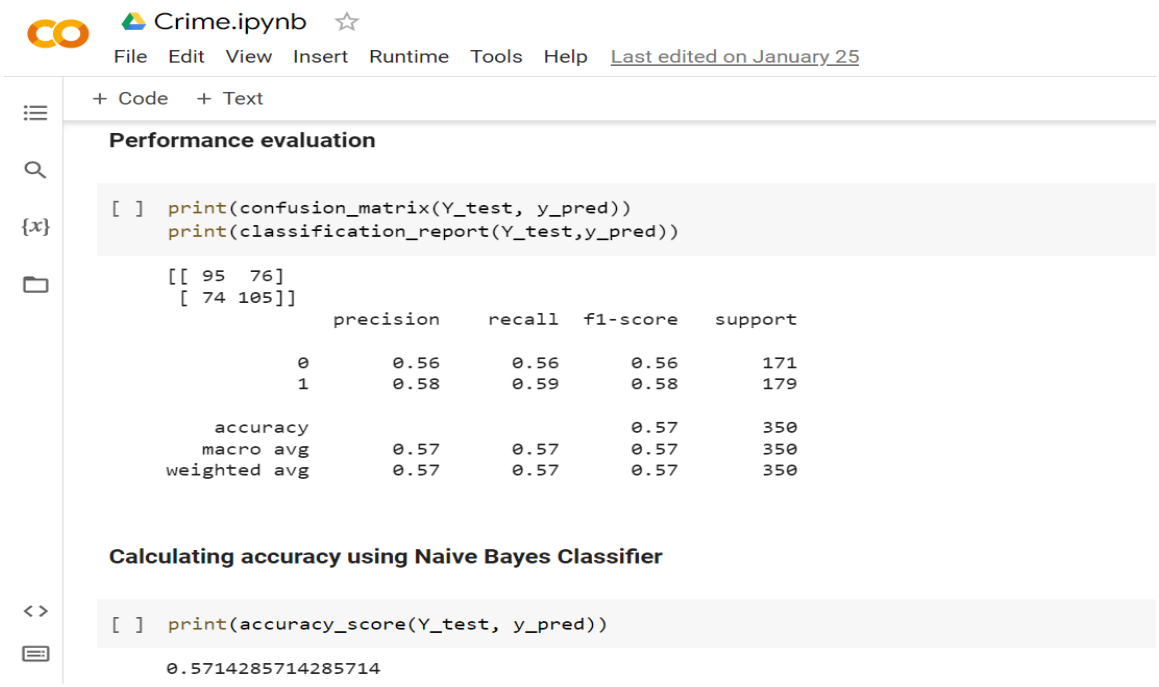


Figure 3.32: NaiveBayes performance

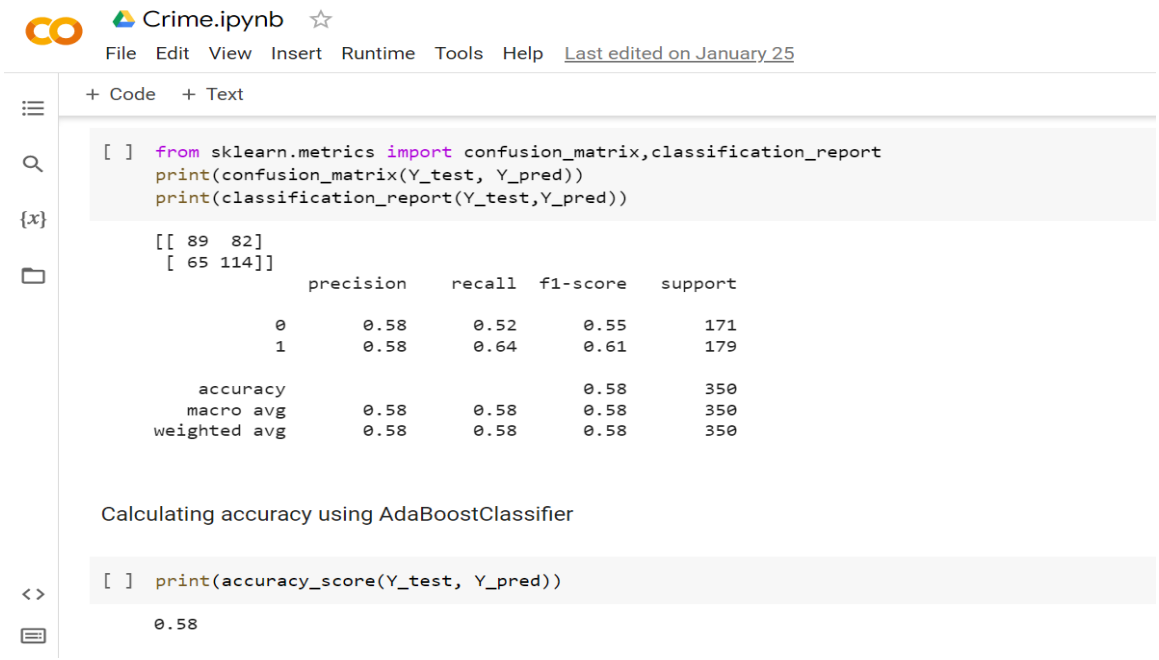


Figure 3.33: AdaBoost performance

Crime.ipynb ☆

File Edit View Insert Runtime Tools Help Last edited on January 25

+ Code + Text

```
[ ] from sklearn.model_selection import cross_val_score
    c = []
    c.append(cross_val_score(clf1,X_train,Y_train,scoring = 'accuracy',cv = 10).mean())
    c.append(cross_val_score(clf2,X_train,Y_train,scoring = 'accuracy',cv = 10).mean())
    c.append(cross_val_score(clf2,X_train,Y_train,scoring = 'accuracy',cv = 10).mean())
    c.append(cross_val_score(clf2,X_train,Y_train,scoring = 'accuracy',cv = 10).mean())
    print(c)

    [0.7771428571428572, 0.8314285714285713, 0.8419047619047619, 0.8342857142857143]
```

Figure 3.34: Voting performance

Visualization:

Figure 3.35: Visualization

3.5 Implementation Requirements

We have explored previous crime history of Bangladesh police. We are evaluating each and every theoretical and statistical idea and method associated with this research project.

Software/Hardware:

- Operating System
- Hard disk (minimum 500GB)
- RAM (minimum 4GB)

Developing tools:

- Colab environment
- Google drive
- Good internet connection
- Any browser (Mozilla Firefox or Google Chrome)

CHAPTER 4

Experimental results and discussion

4.1 Introduction:

At first, we collected our dataset from Kaggle which have 1400 rows and 9 columns. Secondly, we preprocessed the dataset and the dataset is divided into two parts such as test data and train data. After that we trained the machine using the train data. The test data is then taken for analysis.

4.2 Experimental Results & Analysis

TP: A result any place the model accurately estimates the ideal example is dedicated to as a Genuine positive.

TN: An objective for which the model precisely predicts the negative lesson is similar to a genuine positive.

FP: When the model predicts the positive class incorrectly, this results in a false positive.

FN: A consequence for which classical predictions of the negative class are inaccurate is an untruthful undesirable

Precision: Precision is the ratio of correct optimistic to total true positive and false positive. Precision checks to determine the sample's number of false positives. The model was 100% precise if there were no false positives (FPs). When there are more FPs in the mix, the precision will look worse.

$$\text{Precision} = TP/(TP+FP)$$

Recall: Recall proceeds in a different manner. Instead of focusing on the number of false positives the model predicted, recall looks at the number of incorrect rejections that were included in the prediction process. The recall rate is penalized for each period in which there is a forecast false negative. Due to the penalties for precision and recall, the

equations themselves are opposites. Precision and recall are the yin and yang of evaluating the confusion matrix.

$$Recall = TP/ (TP+FN)$$

F-measure: Individual once precision and review are together 1 fixes the F1 Score end up being 1. Individually, recall and exactness play a significant role in the F1 score increase. The F1 score, which is the vocal mean of recall and precision, is a useful metric in addition to accuracy.

$$F\text{-score} = 2*((Precision*Recall) / (Precision + Recall))$$

Accuracy: By dividing the total number of explanations in the dataset by the total number of correct guesses, the exactness is strong. The highest level of precision is 1.0, with a range of 0.0 to 1.0. Additionally, it can be decisive by dividing by the ERR.

$$Accuracy = (TN +TP)/(TN+FP+FN+TP)$$

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Decision tree	79.14	79	80	80
NaiveBayes	57.14	58	59	58
Random Forest	82	82	83	82
AdaBoost	58	58	64	61

Table 4.1: Confusion matrix of the machine learning models using features of the Feature importance selection

4.3 Descriptive Analysis

A choice tree is created using descriptive analysis from a class agreement that marked the preparing test using this tree. In essence, it is a mental model that uses parallel standard to calculate class esteem. This tree makes choices;

1. At a hub which variable portion want to participate.

2. Choice when to stop or when to split.

3. Assigning terminal hub.

4.4 Discussion

After using voting classifier on the machine learning algorithms, we got the best accuracy which was 82.00% with the help of Random Forest Classifier. The result is accurately and precisely calculated by this technology. It accurately predicts the crime location.

CHAPTER 5

Impact on society, environment and sustainability

5.1 Impact on Society

One potential impact is that a crime location prediction system can help to increase the efficiency of the system and also increase the accuracy of the process. By performing certain aspects of the process, a system can help the system to decrease the time and resources required to evaluate crime location applications, which can benefit us.

On the other hand, there are also potential concerns about the use of a crime location prediction system, such as the possibility of errors or biases in the data used to train the system, which could lead to incorrect crime location prediction decisions.

Overall, the impact of a crime location prediction system on society will depend on how it is designed and implemented, and it is important to carefully consider the potential impacts and take steps to mitigate any negative effects.

5.2 Impact on Environment

The impact of a crime location prediction system on the environment will depend on the specific methods used to develop and implement the system. In general, the use of electronic or digital technologies, such as machine learning algorithms and computer systems, for the development and operation of a crime location prediction system could potentially have a negative impact on the environment due to the energy and resources required to power and maintain these systems.

However, there are also ways in which a crime location prediction system could potentially have a positive impact on the environment. For example, if the system is able to streamline and automate the process, it could potentially reduce the need for certain types of paper-based documentation and communication, which could help to reduce the environmental impact of paper production and disposal.

5.3 Ethical Aspects

There are several ethical aspects to consider in the development and use of a crime location prediction system.

One ethical concern is the potential for bias in the system. If the data we have used to train the machine is unrepresentative, it could lead to incorrect or unfair crime location decisions. For example, if the data we have used to train the system is predominantly from a certain demographic group, the system may be more likely to identify the exact location crime correctly.

Overall, it is important to carefully consider the ethical implications of a crime location prediction system and to take steps to ensure that it is fair, unbiased, and transparent.

5.4 Sustainability Plan

Here is a potential sustainability plan for a loan prediction system:

- **Continuous evaluation:** The prediction of crime location system should be continuously evaluated for its performance and accuracy. This can be done through regularly scheduled evaluations or by using a monitoring system that tracks the system's performance in real-time.
- **Regular updates:** The system should be regularly updated with new data and algorithms to improve its performance and accuracy. This can be done through updates to the training data or by incorporating new machine learning algorithms.

CHAPTER 6

Summary, conclusion, recommendation and implication for future research

6.1 Summary of the Study:

We tested the accuracy of arrangement and forecast using a variety of test informational collections in this investigation. The Random Forest, which demonstrated an accuracy of over 82%, serves as the foundation for the characterization. We developed a model and prepared various wrongdoing dataset by making use of this arrangement calculation. We incorporated better outcomes-indicating test data into the model for testing. The a priori calculation in our framework provides continuous examples of a place based on factors or characteristics of that place. The example is used to create a choice tree model. We made various outlines and charts and found interesting bits of knowledge that showed the benchmark to appreciate different metropolitan networks' infringement datasets. We connected Decision Tree and Naive Bayesian classifiers from that point on to assist and anticipate future violations in a given objective area within a given time. In one city, we achieved 51% forecast precision and 54% expectation exactness. Finally, we examined the focus by combining our findings from the wrongdoing dataset for urban areas with its statistics and land data. We intended to learn more about the discoveries made by our model and to investigate and identify network security-relevant components.

6.2 Conclusions:

A specific nation's wrongdoing hazardous districts are predicted by our structured framework. If we consider a specific state or region within a given time frame, it will be more precise. This study focuses on wrongdoing examination by performing bunching calculations on a particular wrongdoing dataset with the help of a quick digger, google colab.

Through preparation on these successive examples, we create an information show associated with each region. Since examples of wrongdoing tend to change over time, they cannot remain static. By preparing, we mean that we are supporting the framework

based on particular data and information sources. By evaluating the wrongdoing patterns, the framework then incorporates the example change in wrongdoing. Factors that contribute to wrongdoing change over time as well. We must identify new wrongdoing factors that motivate individuals to commit wrongdoing by going through the wrongdoing information. We need to discover more wrongdoing characteristics of locations rather than changing a few traits in order to predict signs of improvement. We have already figured out how to prepare our framework by using a few properties, but we want to add more models in the future to improve precision.

6.3 Implication for Further Study:

As a future increase of this endeavor, we plan to apply more portrayal models to extend assumption precision and to work on the overall execution of the structure. In addition, we intend to examine a wide range of urban areas and socioeconomic topographical data for wrongdoing. In addition, in addition to their statistical datasets, we need to consider various violations datasets from new urban communities.

For wrongdoing designs, visual and natural criminal and knowledge examination procedures can be created. We are able to carry out a variety of procedures for information mining because we have tried and connected various bunching strategies for investigating wrongdoing. Additionally, we are able to conduct research on a variety of datasets, including business datasets, destitution datasets, help viability datasets, and so forth.

Despite this another thought called Criminal profiling which assists the bad behavior trained professionals and police with driving to keep the traits of wrongdoers in a particular city or organization. It is a very precise tool for identifying the characteristics or interests of perpetrators or parties to crimes. It is a behavior and insight tool designed to assist law enforcement officers in precisely identifying and profiling the characteristics of obscure criminals. The primary goal of criminal profiling is to:

• To provide law enforcement with a criminal's mental and social assessment;

• To evaluate items discovered in the criminal's possession.

In order to accomplish this, we must dissect the criminal's background and records in order to gather the most relevant data. This way, wrongdoing records are used to gather the most extreme points of interest for each guilty party. When crimes like murder occur in a specific location, we can learn about the criminal's subtleties and method of operation from FIR reports. After acquiring these nitty-gritty details, we can consider those responsible for these practices. Therefore, it is difficult to filter through every wrongdoing record after a single wrongdoing event; instead, we will prepare to use perception systems to communicate with the wrongdoer's subtle elements in a manner that is humanly justifiable.

References

[1] R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. S. Panahy, and N. Khanah- madliravi, “An experimental study of classification algorithms for crime prediction,” Indian Journal of Science and Technology, vol. 6, no. 3, pp. 4219–4225, 2013.

[2] A. A. Biswas and S. Basak, “Forecasting the trends and patterns of crime in bangladesh using machine learning model,” in 2019 2nd Inter- national Conference on Intelligent Communication and Computational Techniques (ICCT), pp. 114–118, IEEE, 2019.

[3] M. A. Awal, J. Rabbi, S. I. Hossain, and M. Hashem, “Using linear regression to forecast future trends in crime of bangladesh,” in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 333–338, IEEE, 2016.

[4] D. Varshitha, K. Vidyashree, P. Aishwarya, T. Janya, D. G. KR, and R. Sahana, “Paper on different approaches for crime prediction system,” International Journal of Engineering Research & Technology (IJERT), vol. 5, p. 20, 2017.

[5] N. Ivan, E. Ahishakiye, E. O. Omulo, and D. Taremwa, “Crime prediction using decision tree (j48) classification algorithm.,” 2017.

[6] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim, and G. Sinha, “An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach,” IEEE Access, vol. 9, pp. 67488–67500, 2021.

[7] A. Sangani, C. Sampat, and V. Pinjarkar, “Crime prediction and anal- ysis,” in 2nd International Conference on Advances in Science & Technology (ICAST), 2019.

[8] V. Jain, Y. Sharma, A. Bhatia, and V. Arora, “Crime prediction using k- means algorithm,” GRD Journals-Global Research and Development Journal for Engineering, vol. 2, no. 5, 2017.

[9] A. Malathi and S. S. Baboo, “Evolving data mining algorithms on the prevailing crime trend–an intelligent crime prediction model,” Int J Sci Eng Res, vol. 2, no. 6, 2011.

[10] J. Azeez and D. J. Aravindhar, “Hybrid approach to crime prediction using deep learning,” in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1701–1710, IEEE, 2015.

[11] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, “Crime analysis through machine learning,” in 2018 IEEE 9th Annual Information Technol- ogy, Electronics and Mobile Communication Conference (IEMCON), pp. 415–420, IEEE, 2018.

[12] A. Kumar, A. Verma, G. Shinde, Y. Sukhdeve, and N. Lal, “Crime prediction using k-nearest neighboring algorithm,” in 2020 International Conference on Emerging Trends in Information Technology and Engi- neering (ic-ETITE), pp. 1–4, IEEE, 2020.

[13] T. T. Nguyen, A. Hatua, and A. H. Sung, “Building a learning ma- chine classifier with inadequate data for crime prediction,” Journal of Advances in Information Technology Vol, vol. 8, no. 2, 2017.

[14] S. Agarwal, L. Yadav, and M. K. Thakur, “Crime prediction based on statistical models,” in 2018 Eleventh International Conference on Contemporary Computing (IC3), pp. 1–3, IEEE, 2018.

[15] Z. M. Wawrzyniak, S. Jankowski, E. Szczechla, Z. Szyma n´ski, R. Pytlak, P. Michalak, and G. Borowik, “Data-driven models in machine learning for crime prediction,” in 2018 26th International Conference on Systems Engineering (ICSEng), pp. 1–8, IEEE, 2018.

CRIME RATE ANALYSIS USING MACHINE LEARNING TECHNIQUE

ORIGINALITY REPORT

23%	18%	3%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	12%
2	Submitted to Daffodil International University Student Paper	2%
3	www.ijcit.com Internet Source	1%
4	Submitted to Colorado Technical University Online Student Paper	1%
5	blogs.bmc.com Internet Source	1%
6	Submitted to Middle East College of Information Technology Student Paper	1%
7	Ghulab Nabi Ahmad, Shafiullah, Hira Fatima, Mohamed Abbas, Obaidur Rahman, Imdadullah, Mohammed S. Alqahtani. "Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by	<1%