

# Prediction of Crime Occurrence in Bangladesh using Machine Learning Models

Faisal Tareque Shohan<sup>\*a</sup>, Abu Ubaida Akash<sup>\*a</sup>, Muhammad Ibrahim<sup>\*\*b</sup> and Mohammad Shafiul Alam<sup>a</sup>

<sup>a</sup>Dept. of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh

<sup>b</sup>Dept. of Computer Science and Engineering, Univeristy of Dhaka, Dhaka, Bangladesh

## ARTICLE INFO

### Keywords:

Crime prediction  
Machine learning  
Data science  
Data curing  
Feature engineering  
Decision tree  
Random forest  
AdaBoost  
XGBoost

## ABSTRACT

Crime is an unlawful act that carries legal repercussions. Bangladesh has a high crime rate due to poverty, population growth, and many other socio-economic issues. For law enforcement agencies, understanding crime patterns is essential for preventing future criminal activity. For this purpose, these agencies need structured crime database. This paper introduces a novel crime dataset that contains temporal, geographic, weather, and demographic data about 6574 crime incidents of Bangladesh. We manually gather crime news articles over a seven year time span from a daily newspaper archive. We extract basic features from these raw text. Using these basic features, we then consult standard service-providers of geo-location and weather data in order to garner these information related to the collected crime incidents. Furthermore, we collect demographic information from Bangladesh National Census data. All these information are combined that results in a standard machine learning dataset. Together, 36 features are engineered for the crime prediction task. Five supervised machine learning classification algorithms are then evaluated on this newly built dataset and satisfactory results are achieved. We also conduct exploratory analysis on various aspects the dataset. This dataset is expected to serve as the foundation for crime incidence prediction systems for Bangladesh and other countries. The findings of this study will help law enforcement agencies to forecast and contain crime as well as to ensure optimal resource allocation for crime patrol and prevention.

## 1. Introduction

Crime is a prevalent concern of any society. It has an impact on the quality of life and economic prosperity of a society. It is a critical factor in determining whether or not individuals should visit a city or country at a specific time or which areas they should avoid if they want to do so. It is also a key indicator of the development and social well-being of a country. Hence minimizing the crime activities has always been a priority of a government.

Law enforcement organizations leverage lawful means to contain the crime rate. Artificial intelligence, in particular, data science and machine learning disciplines have been offering tremendous benefits to almost every sector of a society including the law and order sector. The law enforcement agencies continue to seek advanced information systems that employ state-of-the-art machine learning techniques to better safeguard their communities as crime rates are on the rise across the world. In many regions of the world, the domain of analyzing and detecting crimes using machine learning is gaining intensive research both from academia and industry.


Although crimes can occur anytime anywhere, criminals usually operate in their comfort zones and strive to recur the crime under similar circumstances once they are successful (Tayebi et al., 2012). This implies that crime incidents often leave trails of consistent patterns. If this pattern is detected by the crime prevention authorities, then the crime may be preemptively stopped. Machine learning algorithms are being used to analyze and forecast crime, giving the security agencies a totally new viewpoint (ToppiReddy et al., 2018).

### 1.1. Motivation and Research Questions

As machine learning relies primarily on historical data to predict future outcomes, crime records from the past are necessary for analyzing and forecasting criminal incidents. These records must be in structured and standard form in

<sup>\*</sup>Equal Contribution

<sup>\*\*</sup>Corresponding author

 faisaltareque@hotmail.com (F.T. Shohan\*); akash.ubaida@gmail.com (A.U. Akash\*); ibrahim313@du.ac.bd (M. Ibrahim\*\*); shafiul.cse@aust.edu (M.S. Alam)

ORCID(s): 0000-0002-3872-5758 (F.T. Shohan\*); 0000-0003-3284-8535 (M. Ibrahim\*\*)

order to be used by machine learning algorithms. While in some countries this research is emerging (as detailed in Section 7), Bangladesh, despite having a sizable population of over 160 million, lacks any such study. To the best of our knowledge, no standard, structured, and day-to-day based crime data are available for Bangladesh. On Bangladesh Police's official website, some aggregated crime statistics<sup>1</sup> are available to the general public. However, only the annual counts of a few crime types are provided there. Using these data, a Github repository containing a few types of crimes are available<sup>2</sup>, but this dataset is neither systematically developed nor can it be considered extensive and standard. Also, the major factors that influence crime occurrence are not present in these data. In countries such as United States<sup>3</sup>, United Kingdom<sup>4</sup>, and Canada<sup>5</sup>, there are numerous widely used crime datasets available for researchers. Our research aims at minimizing this gap. We attempt to address the following research question:

*“In order to facilitate prediction of crime occurrence in Bangladesh, how can we develop a standard machine learning dataset so as to apply supervised machine learning algorithms on it?”*

The above question spawns a few sub-questions that include: (1) How to collect past crime data? (2) How to extract information of crime incidents? (3) How do climate and demographic information impact crime occurrence? (4) How to select or engineer meaningful features for crime prediction? (5) How to apply machine learning models to the crime data of Bangladesh?

## 1.2. Contribution

Below we list the major contributions made by this study:

- We introduce a crime dataset with spatio-temporal, weather, and demographic information covering six types of crime incidents of Bangladesh over a seven year time span.
- Several factors that influence the crime occurrence are systematically extracted from a daily newspaper, a weather database, and Bangladesh National Census report. Thus, in our dataset, the impact of place, time, weather, and demographic data on crime occurrence are duly reflected.
- We apply several feature engineering methods to further improve the quality of the data and to make it suitable for machine learning tasks.
- We perform exploratory data analysis of the newly developed dataset and elicit useful patterns.
- Several popular supervised machine learning algorithms, namely Random Forest, XGBoost, AdaBoost, Extra Tree, and Decision Tree are employed to forecast criminal incidents based on our dataset.
- SMOTE oversampling technique is utilized to minimize the data imbalance problem.
- We achieve satisfactory accuracy in the crime prediction task. The findings of this study is expected to assist the law enforcement agencies in prediction, prevention of crimes in Bangladesh and to help in better resource allocation for crime patrol.

The remaining portion of the paper is organized as follows: Section 2 describes the methodology of this study. Proposed dataset and its curating process is presented in detail in Section 3. Section 4 describes the data processing steps and feature engineering methods. Section 5 performs exploratory data analysis on the newly developed dataset. Section 6 discusses the experimental results after applying machine learning models. Section 7 relates this work to the existing research on crime forecasting. Section 8 brings the paper to a conclusion.

<sup>1</sup><https://dmp.gov.bd/crime-data/>

<sup>2</sup><https://github.com/mjtbasif/Bangladesh-Crime-Data>

<sup>3</sup><https://www.sanfranciscopolice.org/stay-safe/crime-data/crime-dashboard>, <https://home.chicagopolice.org/statistics-data/crime-statistics/>

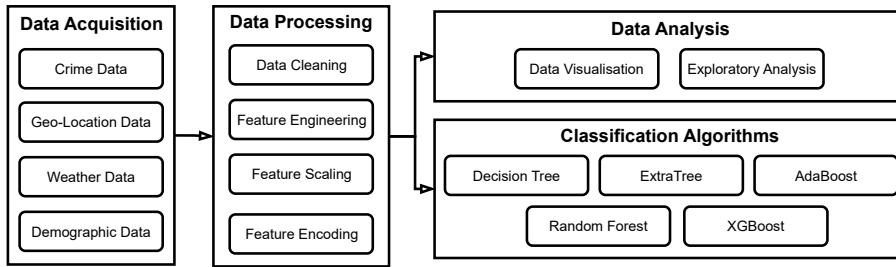
<sup>4</sup><https://data.police.uk/data/>

<sup>5</sup>[https://www.statcan.gc.ca/en/subjects-start/crime\\_and\\_justice](https://www.statcan.gc.ca/en/subjects-start/crime_and_justice)

## 2. Methodology

The workflow of this investigation can broadly be divided into four phases as depicted in Figure 1. Briefly, the phases are as follows:

1. The first phase is data acquisition. We collect, both manually and automatically, crime news articles of the years 2013 – 2019 from a popular daily newspaper of Bangladesh named “The Daily Star”<sup>6</sup>. We also garner geographic, weather, and demographic information related to the collected crime incidents from different sources since these factors affect the crime occurrences.
2. The second phase is data processing and feature extraction. Since the data contains raw text, cleaning this data is necessary. After cleaning the data using standard tools, we derive more crime features from the text. We also apply feature engineering techniques to further improve the quality of the dataset, thereby make the dataset ready for being fed into machine learning prediction algorithms. This way we develop the dataset which we call *CrimeDataBD* – the first-ever standard crime dataset of Bangladesh containing 6574 crime instances and 36 features.
3. The third phase involves analysis of the newly developed dataset using an exploratory approach. The goal here is to better understand the characteristics of the dataset before applying the machine learning models.
4. Finally, in the fourth phase we apply several machine learning classification algorithms on the dataset and analyze the results. We employ five supervised algorithms, namely, Decision Tree, Extra Tree, Random Forest, Adaboost, and XGBoost. The prediction accuracy produced by the classification algorithms are then analyzed and found to be satisfactory.



**Figure 1:** Methodology of this research.

The following sections describe these four phases in more details.

## 3. Data Acquisition

This section describes how we collect the crime data and prepare it for the next stages. The overall process of data acquisition is administered as follows: First, we select a good number of crime news articles from the newspaper (this process will be elaborated shortly) and categorize the articles into six type of crimes. Second, for each of the crime incidents of all six categories, we extract various information about the crime such as date, place, time, victim’s information, criminal’s information etc. This comprises the basic dataset. Third, among the information extracted in the previous step, we choose the features (such as crime place, time etc.) that are able to assist in predicting future occurrence of crime incidents. Here we also derive some other features which are not present in the news articles such as weather information of the crime occurrence time and demographic information of the area of the crime. Thus, we get a dataset that contains various useful information for predicting future crime occurrences.

Regarding the first step mentioned above, the following steps are taken in order to fetch the crime incidents from the daily newspaper “The Daily Star” and to extract features from the articles:

1. We read the Front, Back, City, and Country sections of the newspaper archive covering the years 2018 and 2019. After skimming through around twenty eight thousand news articles, two thousand potential crime news articles

<sup>6</sup><https://www.thedailystar.net/>

are shortlisted and divided into six crime categories which are murder, rape, assault, robbery, kidnapping, and body-found.

2. We then extract some keywords for each category of crime from the 2000 articles. We do this by first tokenizing (i.e., breaking down into words aka tokens) the headlines of these articles. We then select the tokens having the most frequency of occurrence in a category. We exclude some irrelevant keywords using manual judgement.
3. Our next step is to automate the process of fetching even more crime news articles. We do this by using a Web Crawler tool<sup>7</sup>. We crawl, without mentioning any criteria, around sixty thousand news links and headlines from the aforesaid four sections of the newspaper archive for the duration of the years 2013 – 2017. Definitely all articles of this huge collection are not crime news, so we then use the keywords collection (developed in the previous step) to fetch the crime news among these 60,000 articles. Here we employ another tool called FuzzyWuzzy<sup>8</sup>. This way we retain around 4,700 crime news articles among the 60,000 ones. Added to these are the previously collected 2000 articles. Some duplicate articles are manually discarded. Thus the final number of crime news articles stands at 6574. At this stage, the collection is ready for feature selection and extraction.

In the remainder of this Section we further elaborate the above-mentioned steps.

### 3.1. Crime News Source Selection

Since every newspaper covers crime news, it is a great source for obtaining information about past criminal activities. The Daily Star is selected as the source for crime news since it is Bangladesh's one of the most widely read daily English-language newspapers. To avoid news duplication, we work with only one newspaper as different newspapers cover the same crime incident.

### 3.2. Basic Data Acquisition

Among the published crime news, the most prevalent types of crime news are found to be murder, rape, assault, robbery, kidnapping, and corruption. The frequency of corruption-related crimes is lower than that of other types of crime. In contrast, the number of reports containing discovery of an unknown person's body is found to be notably high. Even though body-found news is classified as murder news, we notice that this category of report usually lacks a significant amount of the information that are found in murder news. Therefore, we classify body-found as a separate type of crime.

The selected news may include articles about murder, rape, assault, robbery, kidnapping, or body-found. In addition, if an incident involving multiple crimes is reported in the news, each sort of crime is identified. War crimes, accidents, deaths or injuries resulting from landslides, road collisions, gunfights, police shootings, human trafficking etc. are disregarded.

News about crime incidents are usually found on Front, Back, Country, and City sections in the Archive portal. At the time of data collection, the Archive portal<sup>9</sup> of The Daily Star was publicly accessible. Unfortunately, the Archive portal is not accessible anymore. However, all the collected news are still accessible through the direct links (which is provided in the dataset).<sup>10</sup>

#### 3.2.1. News Link Collection using Manual Process

Since The Daily Star does not offer a news API, at first we resort to a manual process for acquiring data. We begin our news link collection with the news published during 1st January, 2019 – 31st December, 2019 found in the Archive portal. After reading the headline and skimming through the content of the report, we determine if the news is about crime and to what category it belongs. Every news article on the Front Page, Back Page, City, and Country is reviewed to identify crime-related news. After reviewing around 28000 archived news articles, 2000 crime-related news links are selected.

Sometimes it is possible to misinterpret a non-criminal news as a criminal one if only headline is read. For instance, the headline “Two Workers Killed” suggests a crime has happened. However, a thorough reading of the news reveals that it, in fact, reports an accident.

<sup>7</sup>A web crawler is a tool that automatically fetch URL links from Internet. We choose *BeautifulSoup4* as the web crawler.

<sup>8</sup><https://pypi.org/project/fuzzywuzzy/>

<sup>9</sup><https://archive.thedailystar.net/>

<sup>10</sup>Like this one: <https://www.thedailystar.net/city/news/man-stabbed-dead-badda-1781593>



### 3.2.2. News Link Collection using Web Crawler

Since The Daily Star lacks a news API, we use Web Crawling as the next stage. To automatically gather more news links, we construct a web crawler using a tool called *BeautifulSoup*<sup>11</sup>. Using this tool we traverse all articles published from January 1, 2013 to December 31, 2017 from the Front Page, Back Page, City, and Country sections from the archive portal of The Daily Star and gather all headlines and links. This way approximately 60,000 news links are collected.

### 3.2.3. Filtering Crawled News Collection to Identify Crime News

Obviously not all of the 60,000 crawled articles report crime news. Now we need a method to automatically select crime news from this vast collection. While manually gathering news links, we analyze a headline to see whether there are any significant cues within the headlines about crime news. For instance, the words ‘harass’, ‘attacked’, ‘assault’, ‘torturing’, ‘stabbed’, ‘beat’, ‘brutally’, ‘shoot acid’, ‘molested’, ‘burn’, ‘shot’, ‘stabbing’, ‘chained’, ‘tied’, ‘injury’, ‘harassed’, ‘abuse’, ‘brutalised’, ‘forcibly’, ‘cuts’ etc. indicate the assault crime. Other keywords are discovered by analyzing headlines from the previously (manually) collected 2000 crime news. Here we automate the process of calculating frequencies of all words of all the 2000 headlines, and then manually select a group of keywords from each crime category. These keywords are given in Table 1. Using the FuzzyWazzy<sup>12</sup> tool for string matching, we then utilize these keywords to filter the crime news from the 60000 news links. More intervention from us was necessary to identify and remove news about accidents, natural disasters, war crimes, human trafficking, and news from outside Bangladesh. This way 4700 news articles are selected as crime news from the 60000 ones.

**Table 1**  
Keywords in each Crime Category

Category	Keywords
Murder	murder, murdered, kill, killed, homicide, slaying, manslaughter, shoot, dead, assassinate, stabbed, suffocated, poisoned
Rape	raping, rape, raped, gang-raped, rapes, rapist, gang-rape
Assault	harass, attacked, assault, attack, torturing, tortur, stabbed, beat, brutally, unconscious, forced, boiling, shoot, acid, thrown, molested, assaulted, burn, assaults, shot, stabbing, tied, chained, brutality, sexually, injury, harassed, abuse, brutalised, assailant, brutalise, attacks, forcibly, cuts, stalking, sexual, molestation, shave, throwing, cruelty, caned, wrath, abusing, burnt, hack, molest, mercilessly, resists, stab
Robbery	mugger, mugged, robber, loot, snatch, robbed, robbery, looted
Kidnap	abduction, abduct, abducted, abducting, kidnap, rescued, missing, traceless
Body Found	found body, body found, bodies found, body recovered, bodies recovered, found dead, found murdered, found hanging, found bodies, murdered found, dead found, recovered body, recovered bodies, hanging body, hanging bodies, bullet-hit body, bullet-hit bodies, decomposed body, decomposed bodies

**Table 2**  
Distribution of crime instances in each crime category of the dataset.

Murder	Rape	Assault	Robbery	Kidnap	Body Found
1518	1193	1097	598	651	1517

<sup>11</sup><https://pypi.org/project/beautifulsoup4/>

<sup>12</sup><https://pypi.org/project/fuzzywuzzy/>

Note that a few of these news were found to be duplicate news, we manually identified and eventually discarded them. Finally, 6574 crime incidents are recorded in our dataset that occurred during the time span of 2013 – 2019. The data collected for each category of crime are listed in Table 2.

### 3.3. Basic Feature Preparation for Machine Learning

Now that we have gathered the crime news reports in natural language format, it is time to extract useful information from these so that a machine learning model can utilize this dataset to predict future crime occurrences. Machine learning algorithms need some features that can signal about the possibility of future crimes. In this subsection we elaborate this feature preparation process.

#### 3.3.1. Information Extraction from News Reports

We carefully inspect every category of crime news to find out what information about the crime is described in the news. Most common information found in a crime news article are:

- |   |                                    |
|---|------------------------------------|
| i) News date                              | vii) Victim age                    |
| ii) Crime approach                        | viii) Victim's address             |
| iii) Relation between victim and criminal | ix) Criminal's age                 |
| iv) Incident place                        | x) Motive behind the crime         |
| v) Incident time                          | xi) If criminal is arrested or not |
| vi) Victim profession                     |                                    |

In addition, news articles may include information such as the victim's or offender's political and religious involvement. However, not all crime-related news articles contain all these fields. Therefore, we manually extract these information from all the news articles. For each type of crime, the following information are extracted:

**Murder:** news date, incident date, if arrested, part of the day of incident, incident place, murder approach, murder weapon, motive, victim age, victim gender, victim profession, victim religion, victim address, criminal age, criminal gender, criminal profession, criminal religion, relation between victim and criminal.

**Rape:** news date, incident date, if arrested, part of the day of incident, incident place, no of victims, victim age, victim gender, victim profession, victim religion, victim address, criminal age, criminal gender, criminal profession, criminal religion, relation between victim and criminal.

**Assault:** news date, incident date, if arrested, part of the day of incident, incident place, motive, victim age, victim gender, victim profession, victim religion, victim address, criminal age, criminal gender, criminal profession, criminal religion, no of criminal, relation between victim and criminal.

**Robbery:** news date, incident date, if arrested, part of the day of incident, robbed area, incident place, victim state, victims injured, criminal age, criminal gender, criminal religion, no of criminal.

**Kidnap:** news date, abduction date, rescue date, if arrested, part of the day of incident, incident place, rescued place, victim age, victim gender, victim profession, victim religion, victim address, victim injured, criminal age, criminal gender, criminal profession, criminal religion, no of criminal, relation between victim and criminal.

**Body Found:** news date, incident date, part of the day of incident, incident place, victim age, victim gender, victim profession, victim religion, victim address, body state.

#### 3.3.2. Feature Identification for Machine Learning

While the dataset built so far is informative, it is not yet ready for machine learning tasks. One of the goals of the study is to forecast criminal activity. Before predicting a crime incident, it is necessary to identify the underlying pattern. If we carefully analyze the information fields mentioned in the previous discussion, we see that incident date, part of the day of incident, and incident location may hint about the pattern of the crime. Fortunately, these three pieces of information are present in all crime news. Algorithms may be able to predict crime incidents with the help of the useful patterns hidden in these spatio-temporal data. So we select these three information as features for machine learning. Table 3 lists these features.

### 3.4. Features Derived from Basic Ones

While the above three features, i.e., incident date, incident place, and part of the day, may predict the chance of occurring a crime at a particular place in a particular timeframe, it is, from the machine learning perspective, better to have more information that may affect the crime occurrence. Therefore, using these basic features we derive three

**Table 3**

Selected features for crime prediction from each crime incident.

Feature	Type
Incident date	Date
Incident place	Categorical
Part of the day of the incident	Categorical

**Table 4**

Geo-Location Features Derived from Incident Place (cf. Table 3).

Feature	Type	Derived From
Latitude	Numerical	Incident Place
Longitude	Numerical	Incident Place

additional types of features, namely, geo-location features, weather features, and demographic features. Below we describe each of these.

### 3.4.1. Geo-location Data Acquisition

We obtain the geographic coordinates of a given location using the Map-box<sup>13</sup> Geocoding API from the Search service. Later, we add latitude and longitude for addresses.

For some addresses, Mapbox does not provide a response in terms of latitude and longitude. For example, even though the official name of a district is “Chapainawabganj”, Geocoding services refer to it as “Nawabganj”. Another example is, the district airport is still known as “Jessore Airport”, even though the official name has been changed to “Jashore”<sup>14</sup>. This kind of inconsistency led to some oddities which we had to manually address.

Google Map<sup>15</sup> is used to search for incident areas where no response or an incorrect response is received from Mapbox. A Google Map search gives us the latitude and longitude. Table 4 lists the two weather features.

### 3.4.2. Weather Data Acquisition

It is reasonable to assume that weather and criminal activity have some sort of association. Moreover, Heilmann and Kahn (Heilmann and Kahn, 2019) show that on days with maximum daily temperatures being above 85 degrees Fahrenheit (29.4 degrees Celsius) the overall crime rises by 2.2 percent, and violent crime rises by 5.7 percent. Among various weather information, we select the ones that have a direct affect on crime such as visibility, cloudiness, temperature etc.

The Weatherstack<sup>16</sup> API delivers accurate weather data – both real-time and historical – for a particular time and place.<sup>17</sup> We collect the following information: temperature, weather type (encoded as a numerical code), precipitation, humidity, visibility, cloud coverage, heat, and season. Table 5 shows the full list of the weather features.

### 3.4.3. Demographic Data Acquisition

Demographic data refers to information that is socioeconomic in nature and represents the characteristics of a geographic location. These data include population, household size, education, literacy rate etc. Demographic factors are crucial in understanding the crime rates of a place. The crime rate of an area are influenced by them. For example, the higher the literacy rate, the lower the possibility of crime; the higher the number of playgrounds, the the lower the possibility of crime; the lower the number of police stations, the higher the possibility of crime etc.

The National Census of Bangladesh was conducted in 2011 by the Bangladesh Bureau of Statistics<sup>18</sup> and the reports are made publicly available<sup>19</sup>. We manually extract these census data on population size, households, sex and age distribution, marital status, economically active population, literacy and educational attainment, religion, etc.

<sup>13</sup><https://www.mapbox.com/>

<sup>14</sup><https://www.dhakatribune.com/bangladesh/2018/04/02/english-spellings-chittagong-comilla-barisal-jessore-bogra-chang>

<sup>15</sup><https://www.google.com.bd/maps>

<sup>16</sup><https://weatherstack.com/>

<sup>17</sup>However, access to the historical weather API is not included in the free version of Weatherstack, and so we purchased the paid version.

<sup>18</sup><http://www.bbs.gov.bd/>

<sup>19</sup><http://www.bbs.gov.bd/site/page/2888a55d-d686-4736-bad0-54b70462afda/->

**Table 5**

Weather features collected from WeatherStack API (cf. Tables 3 and 4).

Feature	Type	Derived From
Max Temp	Numerical	Latitude, Longitude, Incident Date
Avg Temp	Numerical	Latitude, Longitude, Incident Date
Min Temp	Numerical	Latitude, Longitude, Incident Date
Weather Code	Categorical	Latitude, Longitude, Incident Date
Precipitation	Numerical	Latitude, Longitude, Incident Date
Humidity	Numerical	Latitude, Longitude, Incident Date
Visibility	Numerical	Latitude, Longitude, Incident Date
Cloudcover	Numerical	Latitude, Longitude, Incident Date
Heatindex	Numerical	Latitude, Longitude, Incident Date
Season	Categorical	Latitude, Longitude, Incident Date

for all of Bangladesh's districts and Upazilas<sup>20</sup>, as well as for the major cities. Table 6 shows the full list of features extracted from Bangladesh National Census report.

At this stage, we have 31 features in our dataset which are mentioned in Tables 3, 4, 5, and 6.

**Table 6**

Demographic features collected from Bangladesh Census Data, 2011 (cf. Table 3).

Feature	Type	Derived From
Household Number	Numerical	Incident Place
Male Population	Numerical	Incident Place
Female Population	Numerical	Incident Place
Total Population	Numerical	Incident Place
Gender Ratio	Numerical	Incident Place
Avg House Size	Numerical	Incident Place
Population Density	Numerical	Incident Place
Literacy Rate	Numerical	Incident Place
Religious Institution	Numerical	Incident Place
Playground	Numerical	Incident Place
Park	Numerical	Incident Place
Police Station	Numerical	Incident Place
Cyber Cafe	Numerical	Incident Place
School	Numerical	Incident Place
College	Numerical	Incident Place
Cinema	Numerical	Incident Place

## 4. Data Processing and Feature Enhancement

In machine learning community, it is believed that the real predictive power of these astonishing mathematical techniques oftentimes lies in the quality and quantity of the data they work on. As we build our dataset from real-world text data, we need to apply a variety of methods to clean the dataset from noise and to standardize the dataset so it becomes ready-to-use for machine learning.

### 4.1. Processing and Transforming Crime Data

Real world data is usually inconsistent, incomplete, and prone to a variety of errors. This situation aggravates when the data is text of natural language. In our case, since we develop the dataset from scratch, i.e., from a daily newspaper, emphasize on data cleaning and feature engineering is imperative to make it eligible for machine learning model fitting.

<sup>20</sup>In Bangladesh, several Upazilas comprise a district, and several Unions comprise an Upazila.

In this sub-section we describe how we process and cure the three basic features, namely, incident place, incident date, and part of the day of the incident.

#### 4.1.1. Incident Place

Incident locations are presented in news articles in a variety of formats. In some cases, only the district of crime is mentioned, while in some other cases, both Upazila and the district information are available, and in some other cases, only the Union information is available. In our dataset, for each crime incident we keep all three fields but leave blank if any field is missing.

The spellings of the Union, Upazila, and District names are collected from Wikipedia<sup>21</sup> and Bangladesh government websites<sup>22</sup>. Some of these spellings, however, are found to be inconsistent with the official names.

Since manually correcting these misspellings of place names is inefficient and time-consuming, we opt for string matching and replacement tools. The FuzzyWuzzy<sup>23</sup> string matching module is used to correct all misspellings present in the news articles.

The majority of the crime occurrences include Upazila and District information, but do not include Union or a specific location name. As a result, having only the Upazila and District names for the incident location seems to be a reasonable choice. However, some of the news articles do not mention the Upazila or Union names. Instead, they use other location information like village, bus station, university, town, school, police station, hospital, market etc. We manually search all of these place names in Google search engine and extract the corresponding Upazila or municipal area. We then substitute those names with corresponding Upazila or municipal area names. The process, we note, was quite laborious.

#### 4.1.2. Incident Date

Some of the news dates are written in various formats like “March 29”, “June 27, 2018”, “April 2nd”. But the majority of incident dates found in news articles are mentioned in implicit form such as “last Friday”, “yesterday”, “the day before”, “before an event” and so on. In order to determine the actual incident date, we manually adjust these phrases with the news publication dates. All these extracted dates are then manually formatted in "Day-Month-Year" format and included in the dataset.

#### 4.1.3. Part of the Day of the Incident

The time when the crime was committed is referred to as the incident time. Very few news articles state the exact time of the incident. The time of the crime are instead recorded as parts of the day as it is oftentimes difficult to find out the exact time of the crime occurrence. As a result, when categorizing incident time into different parts of the day, we consider the slots/parts of a day shown in Table 7 as the crime occurrence time.

### 4.2. Feature Engineering

Table 8 shows the features we derive from the features of Table 3. Incident place is split into three features, which are: incident division, incident district, and incident place. Additionally, incident dates has been transformed into date-time objects using the Python's Datetime<sup>24</sup> module. Then, incident month, incident week, and incident weekday are extracted from these *datetime* objects.

**Feature Scaling:** Feature scaling is a technique for normalizing the numerical valued features in a fixed range. Without scaling features, the learning algorithm may be biased toward features with greater magnitude values. As a result, feature scaling brings all features into the same range, and thus the learning model makes good use of all features. For numeric features, we apply a well-known scaling technique called min-max normalization. This technique brings every numerical attribute in a defined range. The most common ranges used are [0, 1] and [-1, 1]. The equation for range [-1, 1] is as follows:  $x_i = 2 * \frac{x_i - \min(x)}{\max(x) - \min(x)} - 1$ , where  $x$  is the feature at hand and  $x_i$  is the value of  $x$  at  $i$ th datapoint.

**Feature Encoding:** Machine learning algorithms can only operate on numerical values. Hence it is necessary to convert the categorical features into numeric ones. There are various methods for encoding categorical features. We

<sup>21</sup><https://en.wikipedia.org/wiki/Upazila>

<sup>22</sup><http://ddm.portal.gov.bd>

<sup>23</sup><https://pypi.org/project/fuzzywuzzy/>

<sup>24</sup><https://docs.python.org/3/library/datetime.html>

**Table 7**

Deciding the slot/part of the day

Time	Part of the Day
6 am - 11:59 am	Morning
12 pm - 3:59 pm	Noon
4 pm - 5:59 pm	Afternoon
6 pm - 7:59 pm	Evening
8 pm - 5:59 am	Night

**Table 8**

Derived Features from three core features of Table 3.

Feature	Type	Derived From
Incident Month	Categorical	Incident Date
Incident Week	Categorical	Incident Date
Incident Weekday	Categorical	Incident Date
Weekend	Categorical	Incident Date
Part of the Day	Categorical	Part of the Day
Incident Place	Categorical	Incident Place
Incident District	Categorical	Incident Place
Incident Division	Categorical	Incident Place

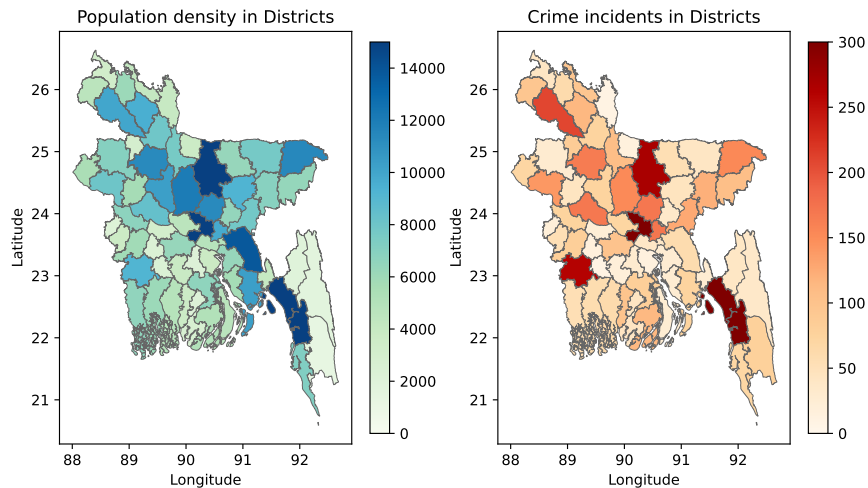
use the label encoding method which turns categorical data into machine-readable numeric form by assigning a unique number (beginning with 0) to each class of a particular feature.

Different types of information are contained in geo-location data, temporal data, weather data, and demographic data. All of these features are combined in the dataset, thereby resulting in a strong set of features for the prediction task. Table 9 shows all these features along with some of their properties. Thus the final dataset contains 6674 instances of crimes and 36 features along with a label field indicating the type of the crime.

## 5. Data Analysis

Now that the dataset is fully developed, we perform some exploratory data analysis to better understand the data.

The heatmap shown in Figure 2 depicts the relationship between population density and crime rate. The general trend is that the crime is more prevalent in the more populous districts.

**Figure 2:** Population density and crime relation in Districts.



**Table 9**

Complete list of features (cf. Tables 4, 5, 6, and 8) in the final dataset and some of their properties.

No.	ID	Feature	Type	Unique	Range	Median	Information Source
1	C1	Incident Month	Categorical	12	-	-	Crime News
2	C2	Incident Week	Categorical	53	-	-	Crime News
3	C3	Incident Weekday	Categorical	7	-	-	Crime News
4	C4	Weekend	Categorical	2	-	-	Crime News
5	C5	Part of the Day	Categorical	5	-	-	Crime News
6	C6	Latitude	Numerical	-	20.87 - 26.48	23.86	Crime News
7	C7	Longitude	Numerical	-	88.14 - 92.44	90.27	Crime News
8	C8	Incident Place	Categorical	660	-	-	Crime News
9	C9	Incident District	Categorical	64	-	-	Crime News
10	C10	Incident Division	Categorical	8	-	-	Crime News
11	W1	Max Temp	Numerical	-	17 - 45	33	Weather API
12	W2	Avg Temp	Numerical	-	16 - 40	30	Weather API
13	W3	Min Temp	Numerical	-	8 - 32	25	Weather API
14	W4	Weather Code	Categorical	21	-	-	Weather API
15	W5	Precipitation	Numerical	-	0 - 204.6	0.8	Weather API
16	W6	Humidity	Numerical	-	13 - 97	68	Weather API
17	W7	Visibility	Numerical	-	4 - 10	10	Weather API
18	W8	Cloudcover	Numerical	-	0- 100	23	Weather API
19	W9	Heatindex	Numerical	-	15 - 43	33	Weather API
20	W10	Season	Categorical	3	-	-	Weather API
21	D1	Household Number	Numerical	-	4872 - 2030000	83300	BD Census
22	D2	Male Population	Numerical	-	11300 - 4930000	187000	BD Census
23	D3	Female Population	Numerical	-	11200 - 3970000	188000	BD Census
24	D4	Total Population	Numerical	-	22900 - 8910000	377000	BD Census
25	D5	Gender Ratio	Numerical	-	84 - 203	101	BD Census
26	D6	Avg House Size	Numerical	-	3.58 - 8.42	4.44	BD Census
27	D7	Population Density	Numerical	-	23 - 30600	1239	BD Census
28	D8	Literacy Rate	Numerical	-	26.7 - 74.6	53.9	BD Census
29	D9	Religious Institution	Numerical	-	0 - 4289	818	BD Census
30	D10	Playground	Numerical	-	0 - 253	25	BD Census
31	D11	Park	Numerical	-	0 - 17	1	BD Census
32	D12	Police Station	Numerical	-	0 - 74	3	BD Census
33	D13	Cyber Cafe	Numerical	-	0 - 478	1	BD Census
34	D14	School	Numerical	-	1 - 242	45	BD Census
35	D15	College	Numerical	-	0 - 64	8	BD Census
36	D16	Cinema	Numerical	-	0 - 40	2	BD Census

Figure 3 illustrates crime incidents in various divisions. It is unsurprising to notice that the most of the recorded crime incidents occurred in the Dhaka division. In addition to having the largest population, Dhaka district, the capital of Bangladesh, is located within this division. We also see that Chattogram has the second highest number of crime incidents, which is perceivable given that it is the second largest city after Dhaka and has the busiest seaport on the Bay of Bengal. It is, however, unusual, that despite having a small population, Rangpur has a relatively high rate of crime. This is perhaps due to the fact that the law enforcement infrastructure is poor there, and also that the economic condition of that division is not that good.

Crime rate in different seasons can be visualized in Figure 4. Broadly, the three distinct seasons of Bangladesh are: the hot pre-monsoon summer season which lasts from March through May, the wet monsoon rainy season which lasts from June through October, and the cold dry winter season which lasts from November through February. The Figure shows that the crime occurrence reaches its peak during the rainy season. The murder rate is found to be higher in the summer and rainy season.

Figure 5 shows another significant statistic which is the number of crimes of various categories reported in each month of the year. We see that the number of murders committed each month has a seasonal pattern: the majority of

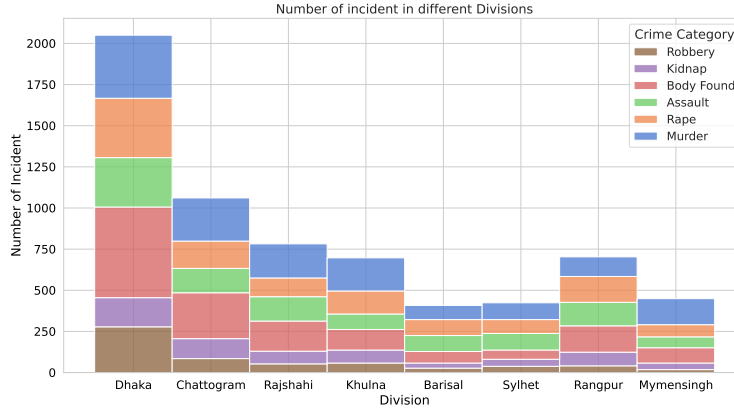


Figure 3: Crime incidents occurring in various Divisions.

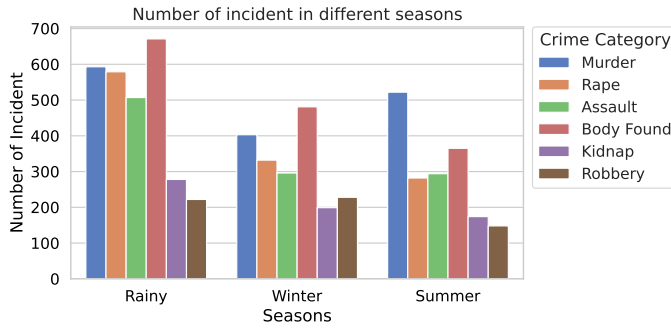


Figure 4: Crime incidents occurring in various seasons.

murders were reported in the first six months of the year. On the other hand, from January to May, there was a gradual decrease in the number of robberies. November and December broadly experience low crime rates.

As reflected in Figure 6, the impact of weather on crime rate is crucial. Crime rates are noticeably higher in hot weather when the average temperature ranges from 28 to 33 degrees Celsius. Also, during the rainy season with a high humidity level, there can be seen an increase in crime incidents.

Figure 7(a) demonstrates the trend of the relation between incident time and crime category. Incident time is expressed here as a part of the day. We see that the majority of murders and rapes occurred at night. The absence of daylight implies that criminals are less exposed which is conducive to criminal activities. The fact that most bodies are discovered in the morning is understandable since most murders that take place at night are found in the morning.

Figure 7(b) illustrates the frequency of crime incidents during a week. In general, more crimes are recorded during the workdays. From Friday to Sunday, fewer murders occurred as compared to the period of Monday to Wednesday. A trend can be seen that most of the crimes occurred on Tuesdays.

Figure 8 demonstrates the occurrence of crime incidents with population density and literacy rate. We see that there is no clear pattern found in these figures.

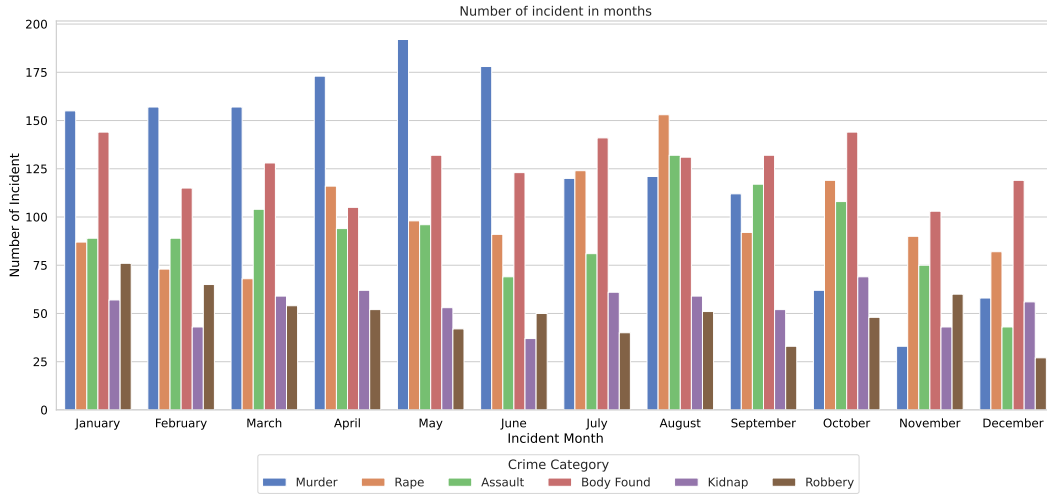
## 6. Machine Learning Model Fitting and Result Analysis

In this section, We first apply several machine learning models on the original dataset. We then deal with the imbalanced nature of the dataset and conduct experiments on a different setting of the dataset.

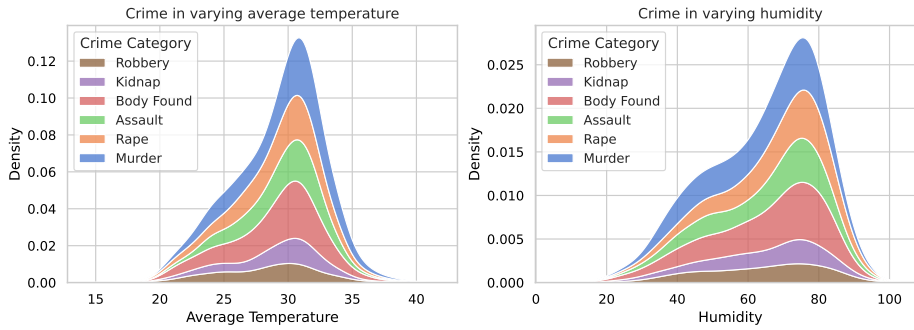
### 6.1. Experimental Setup

We employ five supervised machine learning algorithms that are well-known for their high accuracy, namely Decision Tree, Extra Tree, Random Forest, Adaboost, and Extreme Gradient Boost (XGBoost).

## Prediction of Crime Occurrence in Bangladesh using Machine Learning Models



**Figure 5:** Crime incidents occurring in various months.



**Figure 6:** Crime incidents occurring in varying average temperature and humidity.

The Decision Tree is a classic classification algorithm coined in 1980s where data are split based on a specific feature (Ibrahim, 2022). Extra Tree (Geurts et al., 2006) classifier is an ensemble learning technique that aggregates the classification results of multiple de-correlated decision trees. Random Forest (Breiman, 2001) is another aggregate classifier that contains a number of decision trees and aggregates their predictions in order to reduce high variance of individual trees. Adaptive Boosting or AdaBoost (Schapire, 1999) is a technique that combines multiple weak classifiers to collectively yield a strong classifier. Extreme Gradient Boosting or XGBoost (Chen et al., 2015) is a distributed gradient-boosted decision tree-based ensemble algorithm. Since ensemble models are popular due to their effectiveness, we employ four of them.

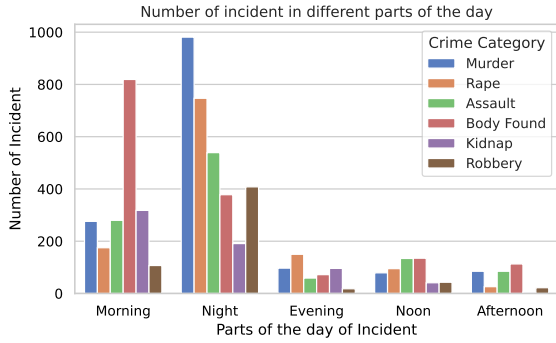
We use Decision Tree, Extra Tree, Random Forest, and AdaBoost implementations from ScikitLearn<sup>25</sup> library package. For XGBoost, we use another library<sup>26</sup>. For Decision Tree, the maximum depth of a tree is set to 20. For Extra Tree, the number of trees is set to 100. For Random Forest, the number of trees is set to 2000, and gini co-efficient is used for splitting. For AdaBoost, the number of trees is set to 50 and learning rate is set to 1.0. For XGBoost, the number of trees, maximum depth, and the learning rate are set to 100, 10, and 0.3 respectively. All other hyperparameters are set to the default settings of the libraries. All the experiments are performed on Google Colab<sup>27</sup> environment.

Following the standards of machine learning, we split the dataset into two groups: train and test with percentages 90% and 10% respectively. The training set is used for learning (and validation) purpose, and the test set is used for evaluation of the metrics.

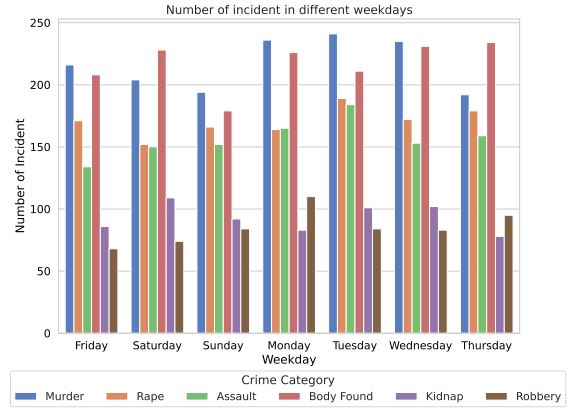
<sup>25</sup><https://scikit-learn.org/stable/>

<sup>26</sup><https://xgboost.readthedocs.io/en/stable/index.html>

<sup>27</sup><https://colab.research.google.com/>

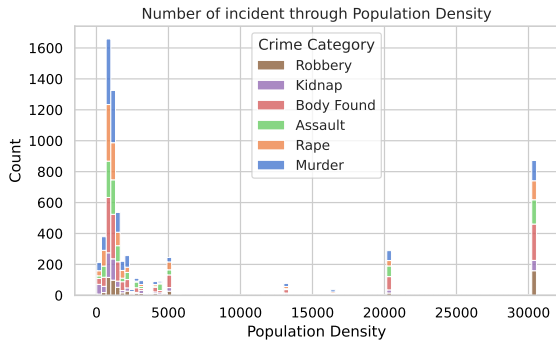


(a) Crime incidents in different part of the day.

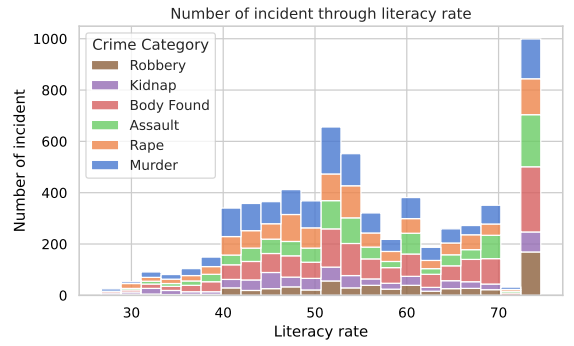


(b) Crime incidents in different weekdays.

Figure 7: Crime incidents in different part of the day and weekdays.



(a) Crime incidents with density of population.



(b) Crime incidents with literacy rate.

Figure 8: Crime incidents with density of population and literacy rate.

To assess a classifier's performance, we compute four widely used metrics, namely, accuracy, precision, recall, and F1 score. ScikitLearn Metrics and scoring library<sup>28</sup> are used to compute these values.

## 6.2. Experiment on Original Dataset

Table 10 shows performance of the classifiers in terms of evaluation metrics. Random Forest's performance tops the list, which is followed by that of Extra Tree and XGBoost. To experimentally stress the need for deploying classifier algorithms, we include the performance of random guessing (the last row of the Table) which is found to be quite below all the classifiers. We consider the accuracy satisfactory given that the dataset is developed from a real-world scenario with a lot of missing values and with the presence of noise, and that the task of predicting crime is inherently difficult (We discuss the challenges of predicting crime in Section 7.2).

Table 11 shows the performance of all five algorithms for every type of crime. We see that in general, the performance of body-found crime is the highest, and performance of kidnapping and robbery crimes are the lowest across all algorithms. Notably, kidnapping and robbery are the two minority classes in terms of the number of instances in the dataset. This raises a question: can we improve the performance by somehow making the dataset more balanced? We try to answer this question next.

<sup>28</sup>[https://scikit-learn.org/stable/modules/model\\_evaluation.html](https://scikit-learn.org/stable/modules/model_evaluation.html)

**Table 10**

Five classifiers' performance on original dataset. Results of random guessing is also included.

Classifier	Accuracy	Precision	Recall	F1
Random Forest	0.44	0.44	0.44	0.43
XGBoost	0.41	0.42	0.41	0.41
Decision Tree	0.32	0.32	0.32	0.32
Ada Boost	0.32	0.31	0.32	0.30
Extra Tree	0.42	0.42	0.42	0.42
Random guess	0.14	0.14	0.14	0.14

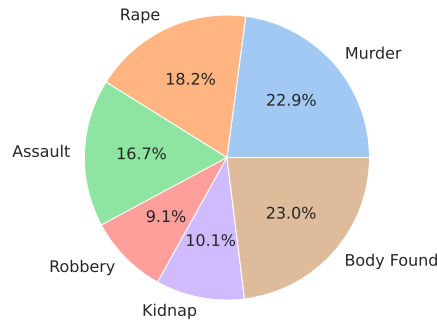
**Table 11**

Evaluation of all five algorithms for each of the six classes with original dataset.

Class	Accuracy					F1-Score				
	Random Forest	XG Boost	Decision Tree	Ada Boost	Extra Tree	Random Forest	XG Boost	Decision Tree	Ada Boost	Extra Tree
Assault	0.38	0.35	0.27	0.13	0.37	0.38	0.34	0.26	0.16	0.37
Body-Found	0.52	0.54	0.37	0.50	0.54	0.47	0.50	0.36	0.42	0.49
Kidnap	0.21	0.26	0.22	0.06	0.25	0.29	0.35	0.22	0.11	0.32
Murder	0.48	0.47	0.35	0.51	0.48	0.45	0.44	0.35	0.41	0.45
Rape	0.39	0.41	0.34	0.25	0.40	0.39	0.40	0.35	0.26	0.40
Robbery	0.25	0.25	0.27	0.16	0.28	0.33	0.32	0.27	0.20	0.36

### 6.3. Experiment on Oversampled Dataset

Figure 9 depicts the distribution of classes in the dataset. We see that the distribution is somewhat imbalanced: robbery and kidnapping have a much lower percentage of data than other crime categories. The imbalanced nature of a dataset poses a challenge for the learning algorithms as the algorithm struggles to learn the discriminating traits of the minority classes.

**Figure 9:** Class distribution of the dataset.

In order to address the challenges posed by an imbalanced dataset, one of the methods that is most frequently chosen by the researchers is resampling the data (Ibrahim, 2020b). There are primarily two types of methods: undersampling and oversampling. Oversampling techniques are typically preferred over undersampling techniques because the latter group of methods removes some instances from the data that may contain crucial information. Synthetic Minority Oversampling (SMOTE) (Chawla et al., 2002) is a popular method for oversampling in which artificial samples of the minority classes are systematically generated. We apply this method to our dataset to convert it into a balanced one. For implementation, Imbalanced Learn library<sup>29</sup> is used. Table 12 shows the count of datapoints in each class before and after applying SMOTE oversampling.

<sup>29</sup><https://imbalanced-learn.org/stable/>

**Table 12**

Datapoints in each class before and after oversampling.

Dataset	Assault	Body Found	Kidnap	Murder	Rape	Robbery	Total
Original	1097	1517	651	1518	1193	598	6574
Oversampled by SMOTE	1518	1518	1518	1518	1518	1518	9108

Table 13 (left side) show performance of five classifiers after applying SMOTE oversampling technique on the dataset. We see that performance of all algorithms have increased after this modification. This time XGBoost and Extra Tree top the list with Random Forest being the second.

**Table 13**

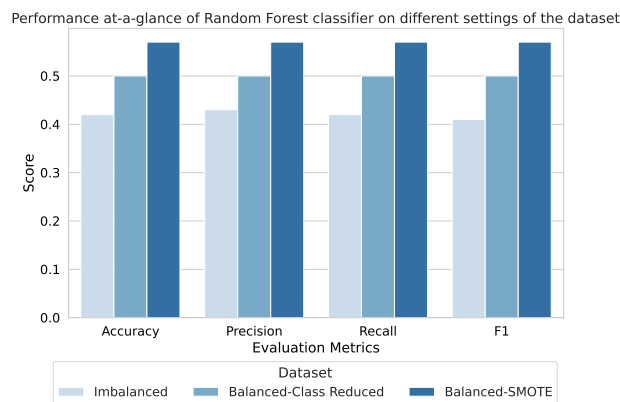
Classifiers' performance on oversampled and class-reduced data.

Classifier	Oversampled dataset				Class-reduced dataset			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Random Forest	0.57	0.57	0.57	0.57	0.50	0.50	0.50	0.50
XGBoost	0.59	0.59	0.59	0.59	0.49	0.49	0.49	0.49
Decision Tree	0.46	0.46	0.46	0.46	0.41	0.42	0.41	0.41
Ada Boost	0.35	0.34	0.35	0.34	0.46	0.45	0.46	0.44
Extra Tree	0.59	0.59	0.59	0.59	0.49	0.49	0.49	0.48

#### 6.4. Experiment on Class-Reduced Dataset

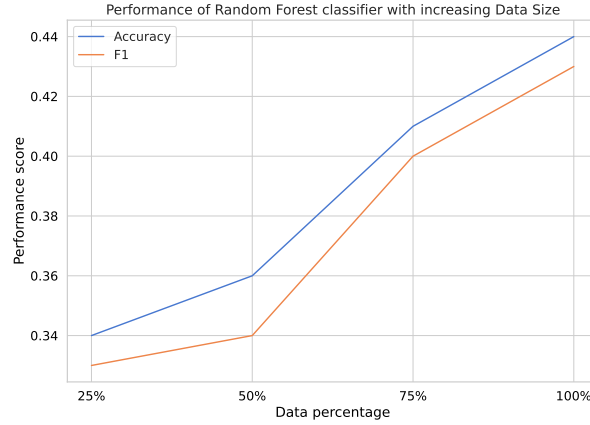
To further investigate the imbalanced data problem, in this experiment we exclude two minority classes altogether – Robbery and Kidnapping – to examine whether their exclusion improves the overall performance. Table 13 (right side) shows the results with four classes in the dataset. We see that performance in general have increased for all algorithms as compared to the original dataset (cf. Table 10). Adaboost algorithm notices this increase more than others: the balanced dataset contributes to the highest performance gain for it. This time Random Forest tops the list with XGBoost and Extra Tree being second.

To summarize the findings of these three experiments: when algorithms are trained on more evenly distributed data, we have noticed a gradually increasing performance in general. Figure 10 illustrates this finding.

**Figure 10:** Performance of Random Forest on different settings of the dataset.

The next question we deal with is: if we increase the dataset size, can we expect a better performance? To investigate this, we conduct a pilot experiment with Random Forest where we start with a random 25% of the dataset, and we then gradually increase the dataset size up to 100%. Figure 11 shows the plot. From this experiment, we can say that having a larger dataset is likely to yield a better performance, which is as per the conventional knowledge of machine learning.





**Figure 11:** Performance of Random Forest with varying dataset size.

In our experiments discussed above, we have found that Random Forest is the most robust algorithm in the sense that it has been among the top few performs across all the various settings. This is perhaps due to the fact that the dataset, being built from a real-world scenario, is inherently noisy; and it is known from the literature (Ibrahim, 2020a) that Random Forest is relatively robust to various inconsistencies of the dataset.

### 6.5. Experiment on Feature Importance

The term *feature importance* refers to methods that assign a score to each input feature in a given model indicating the importance of a particular feature on prediction. A higher score indicates that the particular feature has a greater impact on the model's prediction of the target variable.

Among many available feature importance calculation methods (Allvi et al., 2020), we use the one provided by, or embedded in, the Random Forest classifier. In this method, first the dataset is fit with a Random Forest model. Then, the importance score of a feature is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree. Figure 12 presents the feature importance scores of all 36 attributes. We see that the impact of weather data and spatial-temporal crime features on crime prediction is relatively higher. Feature importance of incident week and part of the day are very significant. As was noticed in Figure 7(a), a high number of crimes are committed at night. Weather information such as temperature, humidity, and cloud-cover are also found to be of high feature importance which is in accordance with an existing study by Heilmann and Kahn mentioned earlier (Heilmann and Kahn, 2019). Also, in Figure 6 we have already demonstrated that the average temperature and humidity have high impact on criminal activities.

## 7. Related Works and Discussion

In this section, we discuss the relevant existing works and then compare them with our research.

### 7.1. Comparison with Existing Works

The study of eliciting hidden patterns from crime data using machine learning techniques has been gaining popularity among the researchers since the last few years. Below we discuss some relevant works in this field.

Buczak and Gifford (2010) investigate the use of fuzzy association rule mining to find community crime patterns. They use the Communities and Crime Dataset from the UCI Machine Learning Repository that includes 2215 crime instances and 128 attributes. Almanie et al. (2015) concentrate on identifying temporal and spatial criminal hotspots using a collection of real-world crime statistics of Denver and Los Angeles. Their method is designed to focus on three key aspects of crime data: the type of a crime, when it occurs, and where it occurs. The authors elicit intriguing patterns for crime hotspots using an apriori algorithm. The paper also demonstrates how the decision tree and naive bayes classifiers can be used to predict potential crime types. Nguyen et al. (2017) use Portland Police Bureau's data for crime forecasting. The authors merge the data with demographic information obtained from various public sources. The

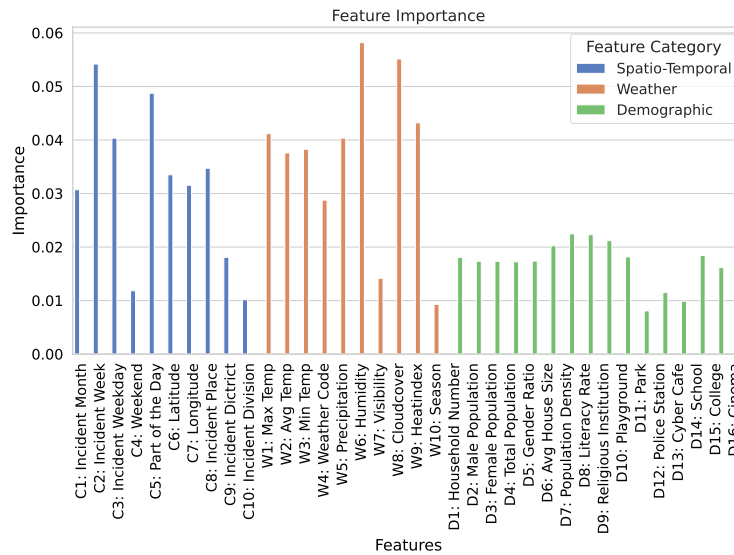


Figure 12: Feature Importance

dataset is then supplemented with additional census data such as economic and ethnic characteristics. The entire dataset is used to predict the type of crime in a specific location over time using a variety of machine learning algorithms. The authors assert that the major variables necessary for crime prediction are location and time. Bogomolov et al. (2014) predict the crime-proneness of a particular area of London city using data collected from mobile phones of users and also using demographic data.

The K-means clustering-based technique is used by Agarwal et al. (2013) and Tayal et al. (2015) to analyze yearly crime occurrence patterns. The dataset used for the former work is crimes recorded by the police in England and Wales from 1990 to 2011. The latter work utilizes the data provided by the National Crime Records Bureau and Committee to Protect Journalists. Varan (2007) also examines K-means clustering with a few modifications to assist identification of crime patterns. A semi-supervised learning technique is applied to discover knowledge from real crime records collected from a local sheriff's office. Sivaranjani et al. (2016) and Pednekar et al. (2018) use some clustering algorithms such as K-means, agglomerative, and DBSCAN to create criminal clusters. The information from the three resulting clusters is then used to predict the class of the crime. The authors obtain data from India's National Crime Records Bureau (NCRB) across six cities from 2000 to 2014 with 1760 incidents and 9 attributes. ToppiReddy et al. (2018) employ Naive Bayes and KNN classifiers to determine the type of crime that is likely to occur using location and day information. Their data comes from the official website of the United Kingdom Police Department that contains a total of 11 attributes, of which they use crime type, location, date, latitude, and longitude data.

We now discuss some works that apply machine learning and data mining techniques on Bangladesh crime data.

In order to forecast crime trends in Bangladesh, Awal et al. (2016) investigate a linear regression model on aggregate data from Bangladesh Police sources. Crime forecasting for robbery, murder, women and children Repression, kidnapping, and other crimes in the various regions of Bangladesh is attempted. Their findings indicate that the majority of crimes are on the rise as the population increases. Parvez et al. (2016) propose a spatio-temporal street crime prediction model that exploits street crime data of Dhaka City. Their dataset is obtained from Dhaka Metropolitan Police (DMP), which consists of the records of crimes from June 2013 to May 2014 but only in aggregate form. Rahman et al. (2021), Islam et al. (2022), and Biswas and Basak (2019) use Bangladesh Police Statistics (cited earlier) as the source for their datasets. This dataset contains crime data in a given division over the course of a year, i.e., in aggregate form. It lacks detailed spatio-temporal information. In addition, these authors do not consider demographics in their studies, which is an important feature to predict crime.

From the above discussion we see that although there are some existing works that employ machine learning techniques on Bangladesh crime data, the dataset is neither systematically developed nor comprehensive. Their dataset is in aggregate form, and moreover, lacks sophisticated and effective features such as weather and demographic features. Not only that, the experiments used in these works are not extensive and comprehensive in terms of effective machine

learning algorithms. Our work is relatively more comprehensive than these works. Moreover, this is, to the best of our knowledge, the first-ever developed standard dataset for Bangladesh. We have included not only information from various reliable sources that affect crime occurrence, but also leveraged feature engineering techniques to further improve the quality of the data. Demographic and weather features are duly included in our dataset. Moreover, the size of our dataset can be considered large enough for being suitable for state-of-the-art machine learning algorithms, as already demonstrated in our experimental studies section.

## 7.2. Discussion

While machine learning astonishes the world with high accuracy in predicting future events in various sectors of human society including healthcare (Callahan and Shah, 2017), agriculture (Ahmed et al., 2023), and economy (Muhammad et al., 2022), predicting events of social well-being (such as crime prediction) has not been an easy task (Kleinberg et al., 2018). The reasons are multifarious. First, social incidents heavily depend on human psychology and state of mind, which is highly unstable. Second, political and social upheavals may change the factors that influence a social incident, thereby making an otherwise predictable incident (from historical data) unpredictable. Third, the prepared dataset may be biased towards a particular group of people due to lack of sufficient and representative data (Angwin et al., 2016). All these and other reasons contribute to the inherent hardness of crime prediction task.

The dataset preparation process for real-time crime scenarios needs to make a number of difficult choices such as the news source selection criteria, the categorization of crime incidents, the type of information to be collected, the process of extracting direct features and deriving indirect ones (Einav and Levin, 2014), the choice of appropriate automated tools etc. While the practices we have followed in this study are standard and effective, there are more avenues to explore in almost each of these processes.

Although law enforcement agencies of Bangladesh currently, like any other country, use human common sense to judge the crime-proneness of a particular time and date (for example, crimes are more likely to occur at night as compared to daytime), taking an informed and data-driven decision will definitely boost their success rate in crime prevention. As such, these agencies can reap benefit from studies of this type in a number of ways. First, they get a crime database which, even if no prediction algorithms are used, is inherently valuable for manual data analysis (like the exploratory approach we adopted in Section 5). Second, they may deploy a real-time prediction system using this dataset and machine learning algorithms in their regional offices and thereby administering their crime patrols more effectively. Third, unlike their current practice of recording crime data in an unplanned and unstructured way (as done in Bangladesh Police website cited earlier) they will be inspired to record the crime incidents in a structured and comprehensive manner following the guidelines of this study.

While the dataset is prepared from Bangladesh crime incidents, the findings of this study may easily be extended to other countries, especially to the countries with whom Bangladesh has resemblance in terms of demography, social structure, and weather. Furthermore, using Transfer Learning setting (Pan and Yang, 2010), this dataset may also be used for predicting crime of another country having a smaller dataset.

## 7.3. Limitations and Possible Extensions

This research, being the first of this kind for Bangladesh, spawns a number of avenues for further investigation. We may benefit from access to more comprehensive and accurate data sources, such as social media, IoT devices, and sensor networks. Tuning of hyper-parameters of the algorithms may also yield better performance. Selection of the best evaluation metrics for this type of social incidence prediction problem also needs further attention from the research community. Combination of various types of machine learning techniques such as unsupervised learning, deep learning and reinforcement learning may lead to improved crime prediction models. Including information on other types of crimes such as theft, smuggling, narcotics, cyber crime etc. will increase the usability of such dataset.

## 8. Conclusion

The present study is an endeavor to extend the benefit of machine learning discipline to social well-being. The investigation is inspired by the absence of standard, systematic, sizable, and comprehensive crime datasets. By collecting, analyzing, and engineering historical crime occurrence data from real-life crime scenarios and incorporating weather and demographic data from reliable sources, we show that it is indeed possible to get crime forecasting – with a reasonable confidence – about future crime occurrence at a particular place and time. Exploratory data analysis is performed on the newly developed dataset. State-of-the-art supervised machine learning algorithms are then applied

with different settings of the data, and satisfactory accuracy in prediction task is achieved. This research will help the law enforcement agencies predict and prevent crimes that may occur in a specific geographical area. The proposed dataset will serve as a foundation for developing a systematic crime database of a country. This information may assist law enforcement agencies in better planning and preparing for optimal resource utilization and management.

## References

- Agarwal, J., R. Nagpal, and R. Sehgal (2013, 12). Crime analysis using k-means clustering. *Intl. Journal of Computer Applications* 83, 1–4.
- Ahmed, S. I., M. Ibrahim, M. Nadim, M. M. Rahman, M. M. Shejunti, T. Jabid, and M. S. Ali (2023). Mangoleafbd: A comprehensive image dataset to classify diseased and healthy mango leaves. *Data in Brief*, 108941.
- Allvi, M. W., M. Hasan, L. Rayan, M. Shahabuddin, M. M. Khan, and M. Ibrahim (2020). Feature selection for learning-to-rank using simulated annealing. *International Journal of Advanced Computer Science and Applications* 11(3), 699–705.
- Almanie, T., R. Mirza, and E. Lor (2015, 08). Crime prediction based on crime types and using spatial and temporal criminal hotspots. *International Journal of Data Mining & Knowledge Management Process* 5.
- Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications.
- Awal, M., J. Rabbi, S. Hossain, and M. Hashem (2016, 05). Using linear regression to forecast future trends in crime of bangladesh. pp. 114–118.
- Biswas, A. and S. Basak (2019, 09). Forecasting the trends and patterns of crime in bangladesh using machine learning model. pp. 114–118.
- Bogomolov, A., B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. S. Pentland (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. *Proceedings of the 16th International Conference on Multimodal Interaction*.
- Breiman, L. (2001). Random forests. *Machine learning* 45, 5–32.
- Buczak, A. L. and C. M. Gifford (2010). Fuzzy association rule mining for community crime pattern discovery. In *ACM SIGKDD workshop on intelligence and security informatics*, pp. 1–10.
- Callahan, A. and N. H. Shah (2017). Machine learning in healthcare. In *Key Advances in Clinical Informatics*, pp. 279–291. Elsevier.
- Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16, 321–357.
- Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2* 1(4), 1–4.
- Einav, L. and J. Levin (2014). The data revolution and economic analysis. *Innovation Policy and the Economy* 14(1), 1–24.
- Geurts, P., D. Ernst, and L. Wehenkel (2006). Extremely randomized trees. *Machine learning* 63, 3–42.
- Heilmann, K. and M. E. Kahn (2019, June). The urban crime and heat gradient in high and low poverty areas. Working Paper 25961, National Bureau of Economic Research.
- Ibrahim, M. (2020a). An empirical comparison of random forest-based and other learning-to-rank algorithms. *Pattern Analysis and Applications* 23(3), 1133–1155.
- Ibrahim, M. (2020b). Sampling non-relevant documents of training sets for learning-to-rank algorithms. *Int. Journal of Machine Learning and Computing* 10(2), 406–415.
- Ibrahim, M. (2022). Evolution of random forest from decision tree and bagging: A bias-variance perspective. *Dhaka University Journal of Applied Science and Engineering* 7(1), 66–71.
- Islam, S. S., M. S. Haque, M. S. U. Miah, T. B. Sarwar, and A. Bhowmik (2022). A trend analysis of crimes in bangladesh. ICCA '22, New York, NY, USA, pp. 501–508. Association for Computing Machinery.
- Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The quarterly journal of economics* 133(1), 237–293.
- Muhammad, T., A. B. Aftab, M. Ahsan, M. M. Muhu, M. Ibrahim, S. I. Khan, M. S. Alam, et al. (2022). Transformer-based deep learning model for stock price prediction: A case study on bangladesh stock market. *arXiv preprint arXiv:2208.08300*.
- Nguyen, T., A. Hatua, and A. Sung (2017, 01). Building a learning machine classifier with inadequate data for crime prediction. *Journal of Advances in Information Technology*, 141–147.
- Pan, S. and Q. Yang (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 1345–1359.
- Parvez, M. R., T. Mosharraf, and M. E. Ali (2016). A novel approach to identify spatio-temporal crime pattern in dhaka city. Association for Computing Machinery.
- Pednekar, V., T. N. Mahale, P. Gadhave, and A. Gore (2018). Crime rate prediction using knn. *International Journal on Recent and Innovation Trends in Computing and Communication* 6.
- Rahman, P., A. I. Hoque, M. F. Ahmed, I. Kashem, A. Alam, and N. Hossain (2021, 08). Bangladesh crime reports analysis and prediction. pp. 453–458.
- Schapire, R. E. (1999). A brief introduction to boosting. In *Ijcai*, Volume 99, pp. 1401–1406. Citeseer.
- Sivaranjani, S., S. Sivakumari, and M. Aasha (2016). Crime prediction and forecasting in tamilnadu using clustering approaches. In *2016 International Conference on Emerging Technological Trends (ICETT)*, pp. 1–6.
- Tayal, D. K., A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi (2015, Feb). Crime detection and criminal identification in india using data mining techniques. *AI & SOCIETY* 30(1), 117–127.
- Tayebi, M. A., R. Frank, and U. Glässer (2012). Understanding the link between social and spatial distance in the crime world. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pp. 550–553.
- ToppiReddy, H. K. R., B. Saini, and G. Mahajan (2018). Crime prediction & monitoring framework based on spatial analysis. *Procedia computer science* 132, 696–705.
- Varan, S. (2007, 01). Crime pattern detection using data mining. pp. 41 – 44.