# SAFE AREA PREDICTION: AN ANALYTICAL METHOD USING MACHINE LEARNING TECHNIQUES

## BY

**Partho Protim Das ID: 183-15-11849**

## AND

**Mahid Ahmed Tonmoy ID: 183-15-12041**

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

Supervised By

# Shah Md. Tanvir Siddiquee

Assistant Professor

Department of Computer Science & Engineering Daffodil International University

Co-Supervised By

# Mr. Md. Sadekur Rahman

Assistant Professor

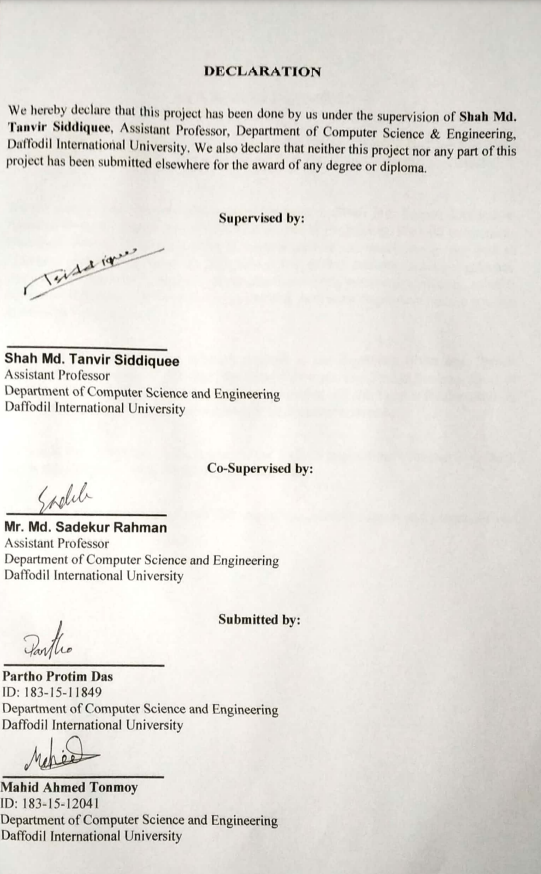
Department of Computer Science & Engineering Daffodil International University



# DAFFODIL INTERNATIONAL UNIVERSITY DHAKA, BANGLADESH

**January, 2023**

# 



# ACKNOWLEDGEMENT

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

We are grateful and wish our profound indebtedness to **Shah Md. Tanvir Siddiquee**, Assistant Professor, Department of Computer Science & Engineering, 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 drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to our Supervisor **Shah Md. Tanvir Siddiquee**, Co-Supervisor **Mr. Md. Sadekur Rahman**, and Touhid Bhuiyan, Head of Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

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

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

# ABSTRACT

A safe area is expected when people reside in or visit a location. The rise in crime rates contradicts this notion, which is why many research efforts are targeted at collecting information from criminal records to truly comprehend criminal behavior and, ultimately, prevent future crimes. In South Asia countries like Bangladesh cannot provide enough safety for the citizens in every situation and this causes several vexatious situations like- hijacking, theft, rape, and many other crimes. People visit new places or move to a new area to live in and always don't know enough about the safety of those places. Usually, it is seen that when people go somewhere at first, they check google maps for directions or traffic conditions. The method of Google Maps can also be applied in finding the safest route or area to pass through or live. Machine learning can play a vital role in predicting safe areas using past data for building this system. In this job, a prediction model is built which helps to forecast safe areas as though people can know about area safety to avoid unpleasant circumstances. Algorithms used for training the dataset are- K-nearest Neighbor (KNN), Random Forest (RF), Naïve Bayes (NB), XGBoost (XGB), and others. The results, looking at predictive performance and accuracy indicate that Random Forest can predict safe areas more reliably.

|  |  |
| --- | --- |
| **Table of Contents** |  |
| [APPROVAL](#_bookmark0) | [i](#_bookmark0) |
| [DECLARATION](#_bookmark1) | [ii](#_bookmark1) |
| [ACKNOWLEDGEMENT](#_bookmark2) | [iii](#_bookmark2) |
| [ABSTRACT](#_bookmark3) | [iv](#_bookmark3) |
| [LIST OF FIGURES](#_bookmark4) | [viii](#_bookmark4) |
| [LIST OF TABLES](#_bookmark5) | [ix](#_bookmark5) |
| CHAPTER 1 | 1 |
| INTRODUCTION | 1 |
| 1.1 Introduction | 1 |
| 1.2 Motivation | 2 |
| 1.3 Rationale of Study | 2 |
| 1.4 Research Questions | 3 |
| 1.5 Expected Outcome | 3 |
| 1.6 Report Layout | 3 |
| CHAPTER 2 | 5 |
| LITERATURE REVIEW | 5 |
| 2.1 **Error! Bookmark not defined.** |  |
| 2.2 Related Works | 5 |
| 2.3 Comparative Analysis | 7 |
| 2.4 Scope of the Problem | 7 |
| 2.5 **Error! Bookmark not defined.** |  |
| CHAPTER 3 | 8 |
| RESEARCH METHODOLOGY | 8 |
| 3.1 Introduction | 8 |
| 3.2 Research Subject | 8 |
| 3.3 Machine Learning Techniques | 8 |
| 3.3.1 Supervised Learning | 9 |
| 3.4 Classification Techniques | 9 |
| 3.4.1 Learning | 9 |
| 3.4.2 Classification | 10 |
| 3.5 Algorithm Details | 10 |
| 3.5.1 K-Nearest Neighbor | 10 |
| 3.5.2 Random Forest | 10 |
| 3.5.3 Naïve Bayes | 11 |

* + 1. [eXtreme Gradient Boosting (XGBoost) 11](#_TOC_250034)
    2. [Stochastic Gradient Descent 12](#_TOC_250033)
    3. [Discriminant Analysis 13](#_TOC_250032)
  1. [Proposed Method 13](#_TOC_250031)
     1. [Data Collection 14](#_TOC_250030)
     2. [Dataset 14](#_TOC_250029)
     3. [Data Pre-processing 14](#_TOC_250028)
     4. [Data Normalization 15](#_TOC_250027)
     5. [Data Splitting 15](#_TOC_250026)
     6. [Algorithm Implicate 15](#_TOC_250025)
     7. [Model Analysis 15](#_TOC_250024)
     8. [Extract Appropriate Algorithm 16](#_TOC_250023)
  2. [System Architecture](#_TOC_250022)

16

CHAPTER 4 17

EXPERIMENTAL RESULTS AND DISCUSSION 17

* 1. [Introduction 17](#_TOC_250021)
  2. [Experimental Results 17](#_TOC_250020)
     1. [Data Acquisition 17](#_TOC_250019)
     2. [Data Utilization 19](#_TOC_250018)
     3. [Feature Importance 19](#_TOC_250017)
  3. [Result and Discussion 20](#_TOC_250016)
     1. [Confusion Matrix 20](#_TOC_250015)
     2. Classification report 22
  4. Result Analysis 22

28

[CHAPTER 5 30](#_TOC_250014)

[IMPACT ON SOCIETY & SUSTAINABILITY 30](#_TOC_250013)

* 1. [Introduction 30](#_TOC_250012)
  2. [Impact on Society 30](#_TOC_250011)
  3. [Impact on Environment 30](#_TOC_250010)
  4. [Ethical Aspects 30](#_TOC_250009)
  5. [Sustainability Plan 31](#_TOC_250008)

[CHAPTER 6 32](#_TOC_250007)

[FUTURE SCOPE AND CONCLUSION 32](#_TOC_250006)

* 1. [Introduction 32](#_TOC_250005)
  2. [Future Scope of this Study 32](#_TOC_250004)
  3. [Recommendations 32](#_TOC_250003)
  4. [Conclusion 33](#_TOC_250002)

[REFERENCES 34](#_TOC_250001)

[PLAGIARISM REPORT 36](#_TOC_250000)

# LIST OF FIGURES

## Figures Page No.

Figure 3.1: Proposed Method to predict Safe Area 14

Figure 3.2: System Architecture 16

Figure 4.1: Safe Area Result Frequency 18

Figure 4.2: Accuracy Chart 23

Figure 4.3: Jaccard Similarity Chart 24

Figure 4.4: Cross-validated Chart 25

Figure 4.5: Misclassification Chart 27

Figure 4.6: AUC Chart 28

Figure 4.7: ROC Curve 29

# LIST OF TABLES

|  |  |
| --- | --- |
| **Table** | **Page No.** |
| Table 4.1: Safe Area Result Frequency | 18 |
| Table 4.2: Data Acquisition | 18 |
| Table 4.3: Data Unit name | 19 |
| Table 4.4: Data Description | 19 |
| Table 4.5: Confusion Matrix | 20 |
| Table 4.6: Accuracy Score | 23 |
| Table 4.7: Jaccard Score | 24 |
| Table 4.8: Cross Validated | 25 |
| Table 4.9: Standard Deviation | 26 |
| Table 4.10: Misclassification Score | 27 |
| Table 4.11: AUC Score | 28 |

**CHAPTER 1 INTRODUCTION**

# Introduction

Safety and Security are one of the major concerns when we are continuing to live in intensity and tangle. Our modern world has unbound popularity and inadequate security and so people face unpleasant situations every day. Citizens of every country including Bangladesh expect their safety to lead their life in a worry-free way. Government is responsible for keeping the country crime-free and controlling the crime rate. Analyzing crime information to implement future crime prevention decisions is a challenge for the government because of the amount of criminal information generated by law enforcement agencies each year.

Criminal acts directly affect our socio-economic development and the quality of our lifestyle. Crimes like – murder, rape, vehicle crime, theft, volume crime, drug dealing, and other major crimes are a threat to citizens. Armed forces are not always there when anyone faces threatening situations. These types of circumstances sometimes take lives or make a lifetime spot of people. And it is a common problem in Bangladesh.

In this era of the modern world, we depend on technologies for good deeds. We use Google Maps to choose a road, to decide which road is shorter, to see traffic conditions, etc. Similar technology can also be built with the help of machine learning, which can analyze previous crime data and predict safe areas. Crime data analysis cannot be done one by one because the amount of criminal information increases every day and a human can’t manage that. For this case, machine learning can be used for analyzing data and predicting safe areas.

Bangladesh is a developing country and it is getting a digital country as well. All over the country, people are now used to technologies and modern gadgets. People of Bangladesh like to apply digital methods in every sector of their life. But there is no such technology in Bangladesh that can show the safety of such areas so that people can be conscious of dangerous places or avoid those. So, developing a technology that can forecast safe areas for the citizen’s safety might be helpful for the people of Bangladesh.

In this paper, we see the use of technology to find a safe place for Bangladesh people so that no one faces any crime scene or gets a victim. We use past crime rates to predict low crime areas. We can mark future crime locations based on past crime information and it helps to

find safe areas. Armed force (Police) will update crime information daily and the system will automatically predict the location-wise crime possibility. General people will see the safety of the area on the system’s map and decide on their own.

For this work, data from the Bangladesh Police website from 2010 to 2021 is used. Classification techniques are used for the prediction such as (K-nearest Neighbor (KNN), Random Forest (RF), Naïve Bayes (NB), XGBoost (XGB), Stochastic Gradient Decent (SGD), Discriminate Analysis (DA)) to detect hot-spots from this dataset. The dataset is analyzed to find the crime rate in past areas under particular thanas.

The rest paper is organized as the followings –Sect. 1 introduces the whole paper, Sect. 2 the earlier work on safe area prediction is described in the literature review section., Sect. 3 This section presented data processing, data management, safe area forecasting, and the many methods that we utilized in the crime data evaluation, as well as the general paper information in the Methodology section. Sect. 4 proposes the model for this work. Sect.5 gives the dataset source description and dataset analysis. Sect. 6 demonstrates the result and discussion, Sect. 7 At the end of this chapter, acknowledgments, and references are mentioned, as well as the overall conclusion of the book.

# Motivation

There are several reasons why criminal behavior occurs. For instance, crime rates are typically greater in places of extreme poverty because people may resort to extreme methods when they feel they lack the resources to survive. A criminal's desire for vengeance, the rush of committing a crime, an intellectual disorder, or addiction are possible additional drivers. Criminal justice professionals need to learn about different types of crimes, the reasons why some people choose to live this way, and how to identify criminals, catch them, and bring them to justice.

# Rationale of Study

Every country has a crime problem, and people from all over the world want to see their families in a secure environment. Armed forces can't always protect us; therefore, this model was created so that individuals may utilize it for their benefit, avoid places with a high crime rate, and locate the safest place to live or travel quietly. A certain authority will utilize this model, and the police will supply the data. The general public will be able to determine which areas have a high or low crime rate. The primary goals of this study are to facilitate people's

lives and ensure their safety. Since there are so many international tourists to Bangladesh, who will profit alongside the locals, it is very vital.

# Research Questions

We encounter some of these issues when writing our papers.

* + - Is it possible to compile genuine data for crime analysis?
    - Can we effectively employ data mining techniques?
    - Is it possible to preprocess the raw data before using it with machine learning techniques?
    - Which crimes have increased in frequency in recent years?
    - Can the police and the public benefit from it?

# Expected Outcome

We want to identify where crimes will occur. We can forecast the area's crime rate based on a variety of crimes, including smuggling, murder, kidnapping, and dacoity. As a consequence, if someone wants to go to or stay the number of crimes and the year, which will help us identify the safest or riskiest areas. We utilize an in a new location, they may avoid these dangerous locations and learn about the places with the greatest crime rates.

# Report Layout

Six distinct chapters are included in this report to make the analysis report more concise and effective for readers or researchers.

Chapter 1 marks an essential beginning of this study. This is a brief that describes COVID-19 in depth. The purpose of the research, the justification for the study, relevant research questions, anticipated findings, and general management information are all explained in this chapter.

Chapter 2 offers a thorough overview of the study's historical context. Based on the findings of this research study, such as automated categorization systems and related activities. The scope of the problem statement, comparative analysis, and perceived obstacles are also covered in this chapter.

Chapter 3 covers the research project's approach, recommended system, and system architecture. The algorithmic details of each developed method are described in depth, from mathematical conception to current state.

Chapter 4 showed the full results of each stage's outcome analysis. The best algorithm, Jaccard score, cross-validation score, confusion matrix, and classification report are included in the report's conclusion. The standard deviation, misclassification, mean absolute error and mean squared error are all covered in the latter section of this chapter.

Chapter 5 is to emphasize the research's impact on society, it explored ethical issues, which are crucial for any influential research endeavor. A consideration of the sustainability of the research project finishes the chapter.

Chapter 6 simply outlined the growth of this research study as the future scope of this research activity. The useful conclusion in this chapter brings the entire study report to a close by succinctly summarizing the key findings.

# Introduction

**CHAPTER 2 LITERATURE REVIEW**

Up until this point in this chapter, crime has been a significant problem. As a result, there has been a great deal of loss and historical conversion in the backdrop of this situation. To save lives, a lot of work has been done to find a safe spot. The context and history of this issue are discussed in this chapter. This chapter has covered some relevant work on this subject. The extent to which the current work was impacted was shown in a comparative study that was delivered at the end.

# Related Works

## S. Sathyadevan et al. 2014[1]

A collaboration between computer science and criminal justice is being created to build a data mining, real-time, and location data technique that will aid in the faster resolution of crimes. Their technology can identify places with a high risk of crime and display crime hotspots. Crime data analysts can assist law enforcement personnel in speeding up the process of investigating crimes, thanks to the rising use of digital systems. They can uncover heretofore unrecognized, relevant information from various sources using this technique.

## Sohrab Hossain et al. 2020[2]

They devise a model for detecting clusters and examine previously recorded criminal activity to uncover hidden trends. If law enforcement authorities have previous knowledge of criminal activity and information on crime hotspots, they can efficiently deal with offenses. They used a supervised learning algorithm to predict criminal activity and their work was on the data on San Francisco criminal activity.

## Ying-Lung Lin et al. 2018[4]

By using the Google Places API to theft data, their research integrates the idea of a criminal environment into grid-based crime prediction modeling and develops a variety of spatial- temporal characteristics based on 84 categories of geographic information. It is a Taoyuan City, Taiwan-based work and based on theft data only. According to their work, the best model was the Deep Neural Network.

## Shanjana A.S et al. 2021[7]

They developed a method that can analyze, identify, and forecast the likelihood of certain crimes in a given area. Using a variety of data mining approaches, this study discusses several forms of criminal analysis and crime prediction. They collected their data from various blogs, news, and websites. Their main target was to detect ‘hot-spot’ but for a given specific region. Their research focused on developing predicting models for monthly crime frequency by crime type.

## Nikhil Dubey et al. 2014[8]

This study looks at a variety of computational approaches for predicting future crime. They also included a comparison of Data Mining Techniques for detecting and predicting future crime. A review of several sorts of data mining approaches for predicting crime hotspots is offered, which can aid in the allocation of police resources to create a safe environment.

## Hitesh Kumar et al. 2018[10]

For determining the crime distribution across a region, several visualization approaches and machine learning algorithms are used in this study. This article offers a crime prediction system based on online mapping and visualization.

## Tahani Almanie et al. 2015[11]

Using a database of known crime statistics, they hope to discover spatiotemporal criminal hotspots. They'll aim to pinpoint the most likely crime hotspots and when they're most likely to occur. They will also forecast what sort of crime will occur next in a certain area at a specific time.

## Ginger Saltos et al. 2017[12]

They investigate models for identifying the frequency of various sorts of crimes by LSOA code (Lower Layer Super Output Areas — a UK police administrative system of areas) and the frequency of anti-social behavior offenses in this research.

## Shiju Sathyadevan et al. 2014[14]

Their technology can forecast crime-prone locations and show crime hotspots. Crime data analysts can assist law enforcement personnel in speeding up the process of investigating crimes as computerized systems become more prevalent.

# Comparative Analysis

This work is unique from previous works in that it has some spectacular effects and a unique approach. In the majority of situations, crime analysis is done to determine the rate of crime increase. Other studies that used foreign data included the City of Chicago. However, no study is constructed utilizing the Bangladesh crime dataset that is utilized in this paper. In particular, forecasting safe areas using data from Bangladesh is quite unusual. This model's accuracy percentage is substantially greater than prior studies, which is only approximately 98 percent. This work is more effective than previous ones because of all these features and capacities.

# Scope of the Problem

Bangladesh is the eighth most populous nation on the planet. Considering this enormous population, the nation lacks the necessary military might to defend every citizen. If we look at the crime rate over the last few years, we may get an idea of the situation in Bangladesh. There are crime hotspots in almost every neighborhood. For jobs or travel, a lot of people travel to different locations every day. They are unfamiliar with these locations; hence they frequently fall prey to traps and suffer severe repercussions. Crime is a problem that no one can predict when it will occur or where. This model was created to prevent situations of this nature. This model will aid individuals in selecting the safest route, neighborhood, or place to travel through or reside in.

# Challenges

Machine learning takes a lot of data to foresee the expected. One significant problem was gathering statistics for Bangladesh. Criminal data is known to have significant missing values. The task of handling these null data is challenging. This could be a drawn-out process. The dataset contained many variables, however, not all of them were necessary to forecast a safe location. The selection of useful qualities from the raw dataset got more challenging because data visualization was needed. The development of an algorithm that was suitable for the situation was challenging as well. The dataset was trained using several different algorithms, and the researchers chose the techniques that accurately identified safe areas.

# Introduction

**CHAPTER 3 RESEARCH METHODOLOGY**

When beginning a research project, the research approach should be taken into account. First, a limitation should be discovered, and then strategies must be devised to tackle that problem. This chapter explains the methodology. The constructed algorithms were then given, along with a thorough explanation of the strategy and a schematic illustration for better comprehension. A system architecture was additionally displayed for improved visualization.

# Research Subject

The mission of preventing crime is essential since it is one of our society's most unsettling and threatening features. A methodical approach to identifying and investigating trends and qualities in crime is through crime-based area prediction. For authorities in government, predicting crime hotspots is not an easy task. A systematic method for identifying and analyzing patterns and trends in crime, crime area prediction identifies the impacted regions that have a high likelihood of experiencing crime. We evaluate several categorization methods to determine which is most effective at identifying crime "hotspots" in certain geographic areas. Finally, we recommend an adequate forecasting approach to achieve the most secure results. Our research has produced a model that can accurately forecast where crimes will occur by utilizing implicit and explicit geographical and temporal information.

# Machine Learning Techniques

To categorize items, classification and supervised machine learning can be combined. A machine learning system is self-contained and capable of continuously integrating data for decision-making. This can be accomplished by drawing on prior knowledge, making analytical observations, and using other techniques. There are several machine learning techniques accessible.

# Supervised Learning

Algorithms for supervised machine learning are commonly employed in this study. To make predictions, supervised machine learning techniques use labeled samples of the past. In order to make predictions about the output values, the learning approach builds an inferred function from an analysis of a dataset that has been thoroughly trained. The system can handle an unlimited amount of training after that. The learning algorithm may identify errors and modify the model as needed when comparing its output to what was intended. The safe region is classified and predicted early on in this thesis using a variety of supervised learning approaches.

# Classification Techniques

Data classification is a subset of data analysis that involves creating models that define pertinent data classes. The most popular and efficient machine learning technology is this one. These models, sometimes referred to as classifiers, are capable of categorical target class prediction under supervised learning. The forecasts are discrete and unordered. Multiclass classification is one of the many classification techniques used in supervised machine learning. There might be a huge number of class labels in some circumstances. A multi- classification system has more classes than just two. In multi-class categorization, only one target label is assigned to each sample. Classify, for instance, a collection of images of leaves that may be aspen, maple, or lupin. A leaf can only be an aspen leaf or a maple leaf at once. The categorization learning technique may be used on labeled data. There are two sorts of data used in classification learning. The first category is known as training data, while the second is known as test data. Training data are used to create the model, while test data are used to evaluate the model. The classifying procedure involves two phases.

# Learning

During the learning phase, a classifier is created using an appropriate approach and training data, which is then assessed against the actual world. When a classification algorithm and training data are coupled, a classifier is simply a set of rules that may be applied to a variety of different events in appropriate circumstances.

# Classification

The classifier or model created during the prediction phase as a consequence of the learning phase can be used to select which class of unknowable data will be forecasted. The test data are utilized in this section to evaluate the accuracy of a model's predictions.

# Algorithm Details

This research effort uses six of the best-supervised machine learning algorithms. An algorithm is, generally speaking, a structured collection of instructions that tell computer software how to transform a set of input data into useful data. Facts are statistics, and useful knowledge is any knowledge that is significant to people, technology, or systems. A similar methodology is used by machine learning algorithms, which involves a procedure and some mathematics. The mathematical transformation used by each machine learning algorithm varies generally. On the other hand, the whole system architecture of this research study includes the most important machine learning algorithms as well as pertinent algorithmic operations.

# K-Nearest Neighbor

By implementing a modest supervised machine learning methodology, the k-nearest neighbors' method (KNN) can be used to perform both classification and regression issues. KNN is easy to comprehend and put into practice. The fundamental theorem of KNN is the Distance measure. Because the dataset is divided into two classes, KNN is used in this classification. The exact formula for the K-Nearest Neighbor algorithm is generated from equation (i).

𝐷(𝑥 , 𝑥 ) = √(𝑥

2

− 𝑥 ) + (𝑦

− 𝑦

2

… … … (𝑖)

𝑖 𝑗

𝑖 𝑗

𝑖 𝑗)

# Random Forest

Random forests, also known as random decision-making forests, are a supervised learning approach that uses numerous decision-making trees to categorize, retrograde, and perform

other tasks. For classification problems, the random forest output is the category picked by the preponderance of trees. For regression problems, the mean or average prediction of the different trees is presented.

∑𝐵(𝑓 (𝑥′) − 𝑓)2

𝜎 = √ 𝑏 𝑏 … … … (𝑖𝑖)

𝐵 − 1

The experimental value, B, is an uncharged variable in formula (ii). Hundreds to thousands of trees are commonly used, based on the scale and nature of the workshop. The average validation loss for each sample of training X', and just the trees with no X' in their random subset, can help determine an ideal number of B trees by validation set or analysis of the out- of-bag error. The training and validation mistakes start to degrade after a few trees are fitted.

# Naïve Bayes

Gaussian Naive Bayes is a variant of Naive Bayes that allows for Gaussian normal distribution and continuous data. Naive Bayes is a set of related supervised learning algorithms for classification algorithms based on the Bayes theorem. The categorizing method is simple but effective. When working with continuous data, it's acceptable to assume that the values for each class follow a normal (or Gaussian) distribution. The complete formulation approach for the Gaussian Naive Bayes algorithm is derived in equation (iii).

1

P(𝑥 |𝑦) ==

(𝑥𝑖 − 𝜇𝑦

exp (− ) … … … (𝑖𝑖𝑖)

𝑖

√2πσy

𝑦

𝑦

2

2𝜋𝜎2

# eXtreme Gradient Boosting (XGBoost)

XGBoost is a decision-making Machine Learning approach that uses a gradient boosting architecture. It comprises a sequence of categorized and reverted (also known as CART) trees as ensemble learning, then improves tree performances by constructing a collection of trees that reduces a normalized goal function to the smallest possible value. Concepts such as split- wisdom discovery in each tree, memory compatible approximate techniques for identifying splits, and appropriate computation of gradient boosting techniques were used to create an approach with a significant computational performance and strong forecasting ability. The XGBoost model may be written as follows for the set of data

𝐷 = {(𝑥𝑖, 𝑦𝑖)} (𝑥𝑖 ∈ 𝑅𝑚, 𝑦𝑖 ∈ 𝑅, 𝑖 = 1,2, … 𝑛) with n m aspects:

𝑘

𝑦𝑖 = ∑ 𝑓𝑘(𝑥𝑖), 𝑓𝑘 ∈ 𝐹 (𝑖 = 1,2 … , 𝑛) … … … (𝑖𝑣)

𝑘=1

Here 𝐹 = {𝑓 (𝑥) = 𝑤𝑞(𝑥)}(𝑞: 𝑅𝑚 → {1,2, … 𝑇}, 𝑤 ∈ 𝑅𝑇) exemplifies the collection of CART decision dendrogram, q denotes the tree structure of the sample map to the intermediate node, T refers to the collection of the intermediate node, and w indicates the real score of the intermediate nodes in the CART decision tree structure collection.

# Stochastic Gradient Descent

The Stochastic Gradient Descent (SGD) approach is a straightforward yet effective way to learn linear and regressive classifiers using convex loss functions, such as (linear) Support Vector Machines and Logistic Regression. Even though SGD has been around for a long time in the field of machine learning, it has only recently gained popularity in the context of large- scale learning. A basic stochastic gradient descent learning process is provided by the class SGD Classifier, which supports a range of classification loss functions and penalties.

The hinge loss, which is identical to a linear SVM, is used to train an SGDClassifier. The SGD approach’s mathematical fundamentals are detailed here. Here are several instructive instances, such as (𝑥1, 𝑦1) − − − (𝑥𝑛, 𝑦𝑛) 𝑤ℎ𝑒𝑟𝑒 𝑥𝑖 ∈ 𝑅𝑚 𝑎𝑛𝑑 𝑦𝑖 ∈ 𝑅(𝑦𝑖 ∈ −1,1 for classification). This system must learn a linear scoring function 𝑓(𝑥) = 𝑤𝑇𝑥 + 𝑏 with model parameter 𝑤 ∈ 𝑅𝑚 and intercept 𝑏 ∈ 𝑅 . The system must consider the sign-off while producing binary classification predictions 𝑓(𝑥). The normalized training error given by the equation is used to discover model parameters (v).

𝑛

1

E(𝑥, 𝑏) = 𝑛 ∑ 𝐿(𝑦𝑖, 𝑓(𝑥𝑖) )+ 𝖺 𝑅(𝑤) … … … (𝑣)

𝑖=1

where L is a loss function that examines model fit and R is a normalization term that prevents the model from becoming complicated; α > 0 is a non-negative linear combination that determines the normalization strength.

# Discriminant Analysis

To estimate possibilities, the Linear Discriminant Analysis model employs Bayes' Theorem. They generate forecasts based on the likelihood that a current input dataset will fall into each of the classes. The resulting category is the one with the greatest chance, and the Linear Discriminant Analysis makes a prediction based on it. The probability is calculated using Bayes' Theorem, which predicts the likelihood of the output variable provided the entry. They also utilize the likelihood of every category and the possibility of each category's data:

[(𝑃𝑙𝑘 × 𝑓𝑘(𝑥))]

𝑃(𝑌 = 𝑥|𝑋 = 𝑥) = … … … (𝑣𝑖) [(𝑠𝑢𝑚 (𝑃𝑙𝐼 × 𝑓𝑙(𝑥)))]

Where x denotes the entry.

k denotes the output category.

𝑃𝑙𝑘

= 𝑁𝑘, which is the basic probability of each class in the training phase.

𝑛

In Bayes' Theorem, it's also known as a probability value.

𝑓𝑘(𝑥) is the possibility of 𝑥 relating to class k as assessed by 𝑓𝑘(𝑥).

The 𝑓(𝑥) is visualized by using the Gaussian Distribution function before being fed through into the formula above, yielding the following formulae:

𝑚𝑒𝑎𝑛

𝑚𝑒𝑎𝑛2

𝐷𝑘(𝑥) = 𝑥 ∗ (

∑ 2 ) − ((2 ∗ ∑ 2)) + 𝑙𝑛(𝑃𝑙𝑘) … … … (𝑣𝑖𝑖)

The classifier equation Dk(x) is known as the dependent variable for class k given entry 𝑥, mean, 2, and 𝑃𝑙𝑘 are all determined from either the data and the class are determined as having an enormous number, which will be regarded in the final classifying.

# Proposed Method

From data collection to result, all are included in the proposed method shown in Figure 3.1. Collection data from Bangladesh Police website and then turn it into usable for applying machine learning models. Find predicting results by the help of machine learning models. System architecture is wider representation of the proposed system.

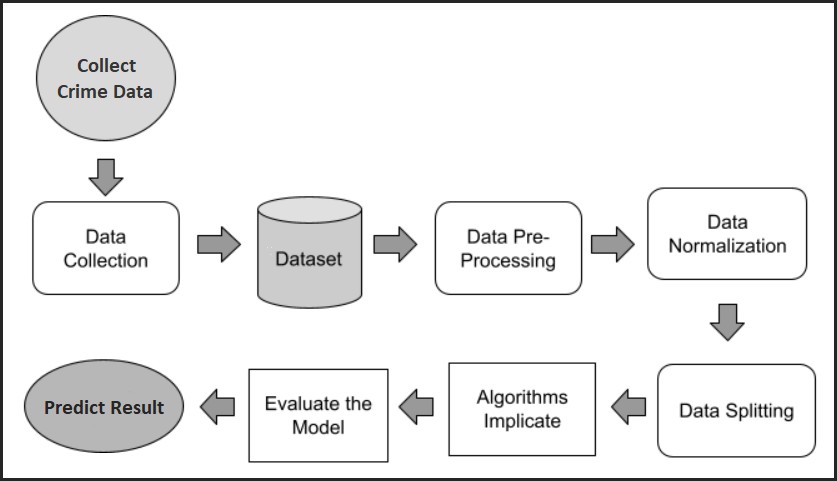


Figure 3.1: Proposed Method to Predict Safe Area

# Data Collection

To predict safe area and save lives, the system required real-life data. Since very few securities sector shares their sensitive data, it was tough to find the needed data. Then the team found the dataset Bangladesh Police official website. From 2010 to 2021, crime values from eight different area is included in this dataset. The first phase in this study is to gather all of the necessary data from diverse sources. Following that, a single CSV file containing all of the data was created for analysis and interpretation.

# Dataset

After consolidating all of the collected data into a digital Comma Separated Value file, the frequency of rows remained perfect for integrating different machine learning algorithms. Machine learning algorithms need a lot of data to foresee anything. The original data, which has 216 rows and 18 columns, contains some inconsistencies.

# Data Pre-processing

The actual world is messy, as we all know. As a result, secondary datasets are much more problematic because real-world datasets are more likely to contain missing, noisy, redundant, and inconsistent data. Therefore, we first had to look at the dataset and understand its

characteristics in order to prepare it for modeling. Data pre-processing is what is referred to as this.

# Data Normalization

The process of converting numerical values to a scale-based representation while preserving value variation is known as normalization. Since the majority of the characteristics are binary, a min-max scaler was employed. The predictor variables (X) were then normalized for increased accuracy.

# Data Splitting

Even before the machine learning approach is applied, the data set has to be divided into training and testing portions. 20% of the data were used to test the model, while the remaining 80% were used to train it. The training portion of the dataset may be used to create the specific model, which can then be used to predict anything. The testing portion can then be used to determine how accurately the data is predicted.

# Algorithm Implicate

Twelve alternative algorithms have been utilized to determine the best accuracy and select the best strategy. The six techniques are K-nearest Neighbor, Random Forest, Naive Bayes, XGBoost, Stochastic Gradient Descent, Perceptron, and Discriminant Analysis. All of these techniques led to various analytical discoveries.

# Model Analysis

After analyzing the Confusion Matrix, Accuracy Score, Jaccard Score, Cross Validated Score, AUC Score, Mean Absolute Error, and Mean Squared Error for each algorithm, the results were translated into tables. The confusion matrix provides a summary of how accurately the data is anticipated. The Accuracy Score, Jaccard Score, Cross Validated Score, and AUC Score all show the percentage of projected data that is accurate. Mean Absolute Error and Mean Squared Error are used to calculate the error rate of the algorithms.

# Extract Appropriate Algorithm

The ideal approach was identified by assessing and examining all of the crucial data from the tables. The extracted technique has the highest accuracy rate and the lowest error rate in the supplied dataset. First and foremost, the dataset must be used effectively by a suitable algorithm. It is preferable in this situation to utilize a range of algorithms as models before choosing the best one. In this study, a variety of analytical criteria, such as accuracy score, Jaccard score, cross-validation score, AUC score, and others, were employed to determine the most efficient approach. The XGBoost Classifier has been shown to be the best method that is suitable for the study's crime dataset. In all of the aforementioned metrics, it received the top ratings. After selecting the proper algorithm, the process advances to the next phase.

# System Architecture

A system architecture formation is required to assume the entire project system in order to improve the practical training in machine learning method and web implementation. Figure

3.2 displays the fundamental system architecture, which is a more comprehensive depiction of the suggested system.

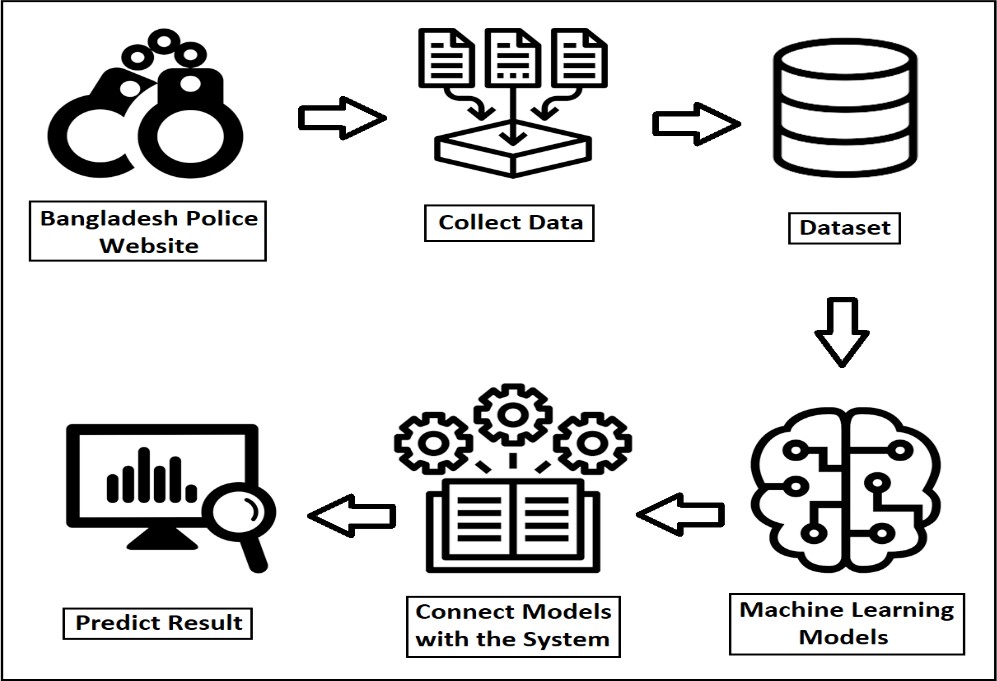


Figure 3.2: System Architecture

# CHAPTER 4 EXPERIMENTAL RESULTS AND DISCUSSION

# Introduction

For every study and project kind, the outcome is crucial. Consequently, the output is part of a project. The table includes all of the outcomes from this chapter. The dataset for crime rate is thoroughly explained in this chapter, along with data collection, data use, and feature importance. Then a confusion matrix table was made to show the outcomes of different algorithms. A classification report was used to help show precision, recall, and the F1-Score in a table. The accuracy, Jaccard score, cross-validation score, AUC score, and ROC curve were all displayed in different plots. The information has been arranged in a table for ease of understanding. In certain cases, the standard deviation is displayed as a table. The solution was then shown as a table. A screenshot of the output display was provided in this chapter.

# Experimental Results

The best approach to forecast a safe location was discovered after the machine learning model development was successfully completed and each algorithm showed its distinct accuracy and scores. As a result, the analytical element of the experimental results section allows for the examination of all possible scores for all algorithmic applications and processes.

# Data Acquisition

The dataset, which consisted primarily of actual crime data from the Bangladesh Police website, was used to train the algorithm. The website included annual records of numerous crime kinds, such as overall murder rates or dacoity, among others. Approximately 216 crime data were used to train the algorithm. There are two numerical values for each class: "1" for "high crime rate" and "0" for "low crime rate." If the result indicates a high crime rate, the location is not safe, and if the result indicates a low crime rate, the region is safe. According to Table 4.1 and Figure 4.1, 48.6 percent of the dataset's area has been designated as dangerous.

Table 4.1: Safe Area Result Frequency

|  |  |
| --- | --- |
| Safe | Unsafe |
| 51.4% | 48.6% |



**Dataset**

**48.60%**

**51.40%**

Safe

Unsafe

Figure 4.1: Safe Area Result Frequency

The dataset contains a variety of information that may be used to estimate Safe Area and obtain the result, including fundamental details like unit name and year as well as criminal information like dacoity, robbery, murder, and swift trial. The utilized dataset has 18 important components or characteristics, all of which are numerical variables. Table 4.2 contains a quick summary of the dataset.

Table 4.2: Data Acquisition

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Name | Unit/Values | Feature Name | Unit/Values |
| Unit Name | Nominal | Police Assault | Numerical |
| Year | Numerical | Burglary | Numerical |
| Dacoity | Numerical | Theft | Numerical |
| Robbery | Numerical | Other Cases | Numerical |
| Murder | Numerical | Arms Act | Numerical |
| Speedy Trial | Numerical | Explosive | Numerical |
| Riot | Numerical | Narcotics | Numerical |
| Woman & Child  Repression | Numerical | Smuggling | Numerical |
| Kidnapping | Numerical | Crime Rate | Numerical |

# Data Utilization

It could be simpler to manage the data in a computer system if each category (nominal) variable was coded separately. 1 and 0 were used to categorize the high/low answers, with 1 being high and 0 denoting low. Integer values are assigned to all other characteristics. The data descriptions for the 18 attributes were extracted in order to have a better knowledge of the dataset. The numbers Count, Mean, Standard Deviation, Min, 25%, 50%, 75%, and Max generated from the dataset are displayed in Table 4.3.

Table 4.3: Data Description

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | Mean | STD | Min | 25% | 50% | 75% | Max |
| Unit\_Name | 216 | 8.5 | 5.200179 | 0 | 4 | 8.5 | 13 | 17 |
| Year | 216 | 2015.5 | 3.460071 | 2010 | 2012.75 | 2015.5 | 2018.25 | 2021 |
| Dacoity | 216 | 26.86111 | 40.56702 | 0 | 0 | 6.5 | 38 | 184 |
| Robbery | 216 | 47.51389 | 62.43963 | 0 | 3 | 23 | 63 | 294 |
| Murder | 216 | 203.963 | 293.9737 | 0 | 15 | 37 | 278 | 1395 |
| Speedy\_Trial | 216 | 78.36111 | 112.6729 | 0 | 2 | 22.5 | 110.5 | 563 |
| Riot | 216 | 4.578704 | 8.214752 | 0 | 0 | 0 | 7 | 56 |
| W\_&\_C\_Repression | 216 | 989.463 | 1339.007 | 0 | 8 | 170 | 1640.5 | 5115 |
| Kidnapping | 216 | 38.37963 | 50.75435 | 0 | 1 | 12 | 56.25 | 204 |
| Police\_Assault | 216 | 34.60185 | 50.90994 | 0 | 1 | 14 | 41 | 336 |
| Burglary | 216 | 136.6944 | 183.9549 | 0 | 2 | 52.5 | 161.5 | 686 |
| Theft | 216 | 381.1759 | 529.0699 | 0 | 30 | 128 | 485.75 | 2240 |
| Other\_Cases | 216 | 4339.574 | 5705.168 | 0 | 129.5 | 761.5 | 7003.5 | 22429 |
| Arms\_Act | 216 | 95.23611 | 141.9115 | 0 | 3 | 22.5 | 152 | 723 |
| Explosive | 216 | 27.78241 | 53.18108 | 0 | 0 | 4 | 27 | 387 |
| Narcotics | 216 | 2699.56 | 4038.545 | 0 | 161.75 | 775.5 | 3445 | 22682 |
| Smuggling | 216 | 301.0556 | 553.0123 | 0 | 2 | 52.5 | 204.25 | 2509 |
| Crime\_Rate | 216 | 0.486111 | 0.500968 | 0 | 0 | 0 | 1 | 1 |

# Feature Importance

"Feature significance" refers to techniques that evaluate input characteristics according to how well they can predict a target variable. The term "feature importance" refers to a group

of techniques for calculating scores for the features used as inputs to a predictive model, indicating the relative weighting of each piece of data when producing a forecast. The feature significance score may be used to build a prediction model as well as get insight into the dataset and model.

# Result and Discussion

The values of High Crime and Low Crime are rendered positive and negative, respectively, in this study. To display intended results and assess the effectiveness of machine learning algorithms, the confusion matrix has been utilized. Table-4.5 displays the template for the confusion matrix for various algorithm types.

# Confusion Matrix

To verify the implementation phase's findings, a confusion matrix must be created. A N x N matrix called a confusion matrix is used with N target classes to assess how well a classification model is working. The accuracy of the machine learning model is evaluated by the matrix by comparing the actual target values to the anticipated ones. This shows how well the algorithmic model is working and what mistakes it is committing. Binary classification will allow for the determination of Precision, Recall, and Accuracy utilizing some mathematical calculations. Additionally, these average values need to be examined using a micro or macro average for multi-class categorization. It's important to comprehend the four components that make up the calculation of various measurement metrics before we proceed. The four forms of false positives and false negatives are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (FN). Table 4.4 depicts the Confusion Matrix's real evolution, which comprises True Positive, False Positive, False Negative, and True Negative.

Table 4.4: Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | Actual Values | |
| Predicted Values | True Positive (TP) | False Positive (FP) |
| False Negative (FN) | True Negative (TN) |

## True Positive (TP)

Tuples that have been accurately classified by the classifier are considered optimistic.

## True Negative (TN)

Positive tuples that the classifier misclassified are referred to as negative tuples.

## False Positive (FP)

Our focus today is on these negatively labeled tuples that the classifier misclassified as positive. The symbol FP can be used to represent this kind of relationship.

## False Negative (FN)

The classifier incorrectly identified these positive tuples as negative. It is denoted by the letter FN.

## Precision

One metric for determining how accurate something is may be its precision. It is also known as the percentage of retrieved events that are actually relevant. Equation provides the mathematical method for measuring accuracy (viii).

## Recall

𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 =

𝑇𝑃

𝑇𝑃 + 𝐹𝑃

… … … (𝑣𝑖𝑖𝑖)

The quantity of optimistic tuples that are identified as such in machine learning is a measure of completeness. A relevant instance is defined as the portion of relevant instances that have been discovered relative to the entire number of relevant cases. Equation provides the mathematical method for calculating Recall (ix).

𝑅𝑒𝑐𝑎𝑙𝑙 = 𝑆𝑒𝑛𝑠𝑖𝑡𝑖𝑣𝑖𝑡𝑦 =

𝑇𝑃

𝑇𝑃 + 𝐹𝑁

… … … (𝑖𝑥)

## F1-Measure

The visual mean, sometimes referred to as the F measure, is a measurement that assesses the precision and recall of a test. Equation provides the mathematical method for calculating the F1-Measure (x).

𝐹 =

2 × 𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 × 𝑅𝑒𝑐𝑎𝑙𝑙

𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 + 𝑅𝑒𝑐𝑎𝑙𝑙

… … … (𝑥)

## Accuracy

The accuracy of a classifier on a test set is measured by the percentage of test set tuples that the classifier correctly identified. Equation analysis may make accuracy evaluation simpler (xi).

𝐴𝑐𝑐𝑢𝑟𝑎𝑐𝑦 =

𝑇𝑃 + 𝑇𝑁

𝑇𝑃 + 𝐹𝑁 + 𝐹𝑃 + 𝑇𝑁

… … … (𝑥𝑖)

# Classification report

A classification report is a statistic that is used in machine learning to evaluate the effectiveness of the system. It is used to assess a trained classification model's accuracy, recall, F1 Score, and support. For a classification-based machine learning model, it is a common technique. Accuracy, recall, F1 score, and support for the model are all displayed. It gives a more realistic picture of the trained model's overall performance. Classification outcomes from machine learning models must be assessed using all metrics provided in the research project.

# Result Analysis

After calculating all practical components, such as Precision, Recall, F1-Measure, Accuracy, and so on, it is crucial to review the findings analysis. The best algorithm among all algorithms and the algorithm that does this when compared to others will be identified in this study.

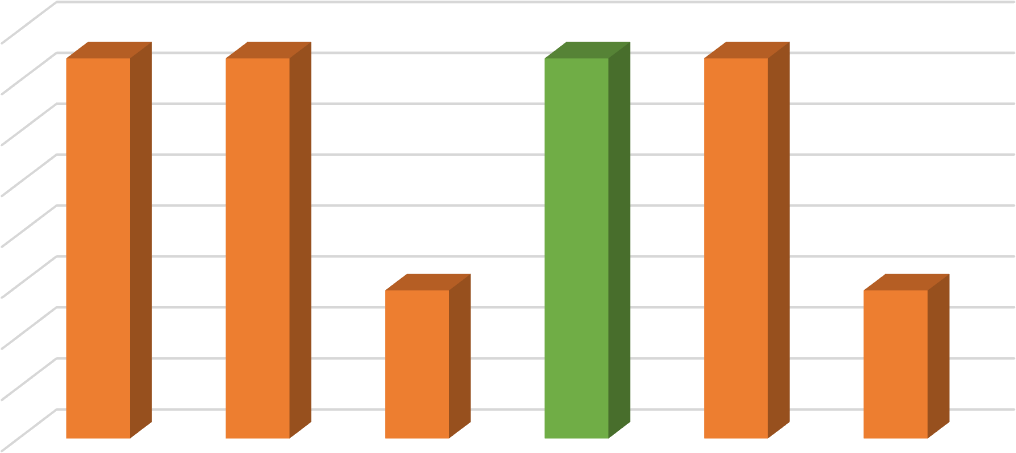
# Accuracy

The best performance of an algorithm is determined by its accuracy. How well it performs will depend on the information that are given to it. A probabilistic approach can be used to

evaluate accuracy performance. Extreme Gradient Boosting is the most accurate approach compared to the others, whereas Gaussian Naive Bayes is the least accurate. One of the most potent and scalable gradient boosting machine implementations, the eXtreme Gradient Boosting technique has proven that it can outperform boosted tree methods in terms of processing power. It was developed with the primary goal of enhancing model performance and speeding up computer data analysis. Figure 4.2 and Table 4.5 display the accuracy graph and percentages for each approach that was used to make predictions in this model.

Table 4.5: Accuracy Score

|  |  |
| --- | --- |
| **Algorithm Name** | **Accuracy Score (%)** |
| KNN | 97.73 |
| Random Forest | 97.73 |
| Naive Bayes | 95.45 |
| XGBoost | 97.73 |
| Stochastic Gradient Descent | 97.73 |
| Discriminant Analysis | 95.45 |



Accuracy Chart

98

97.73

97.73

97.73

97.73

97.5

97

96.5

96

95.45

95.45

95.5

95

94.5

94

K-Nearest

Neighbor

Random

Forest

Naïve Bayes XGBoost

Classifier

Stochastic Discriminant

Gradient Analysis

Decent

Figure 4.2: Accuracy Chart

# Jaccard Score

A statistic used to assess the variety and similarity of sample sets is the Jaccard score. In terms of the Intersection to Union ratio, they are equivalent. You may compare the similarity of two finite sample sets statistically by using the Jaccard coefficient, which is calculated as intersection size divided by union size. Equation (xii), Figure 4.3, and Table 4.6 all display the accuracy chart and percentages for the various forecasting techniques utilized in this model.

J (A, B) =

|𝐴 ∩ 𝐵|

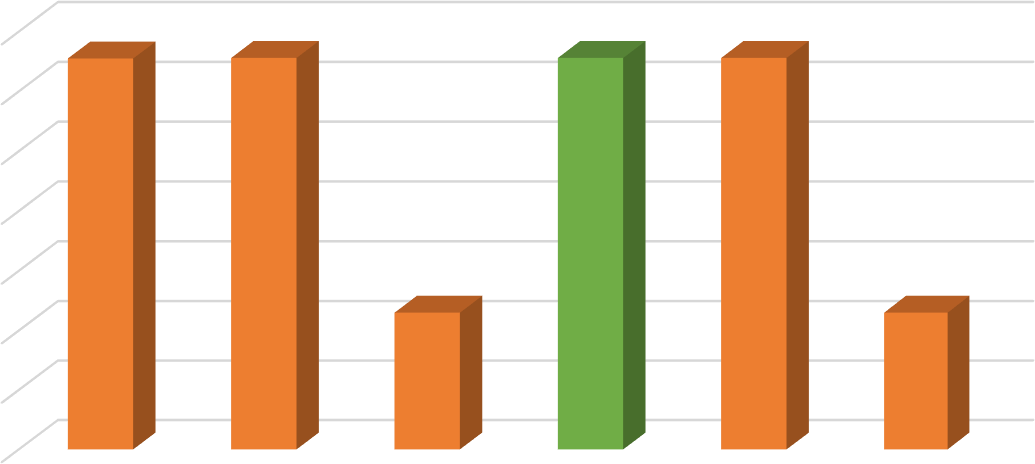
|𝐴 𝖴 𝐵| =

|𝐴 ∩ 𝐵|

|𝐴| + |𝐵| − |𝐴 ∩ 𝐵| … … … (𝑥𝑖𝑖)

Table 4.6: Jaccard Score

|  |  |
| --- | --- |
| **Algorithm Name** | **Jaccard Score (%)** |
| KNN | 95.55 |
| Random Forest | 95.56 |
| Naive Bayes | 91.30 |
| XGBoost | 95.56 |
| Stochastic Gradient Descent | 95.56 |
| Discriminate Analysis | 91.30 |



Jaccard Similarity Chart

96

95.55

95.56

95.56

95.56

95

94

93

92

91.3

91.3

91

90

89

K-Nearest

Neighbor

Random

Forest

Naïve Bayes

XGBoost

Classifier

Stochastic Discriminant

Gradient Analysis Decent

Figure 4.3: Jaccard Similarity Chart

# Cross Validated Score

The statistical technique of cross-validation is used to assess machine learning models. The data are first separated into k folds and then shuffled to begin the cross-validation process.

After fitting the data with k models, 1

𝑘

samples of the data are evaluated. The results k−1

k

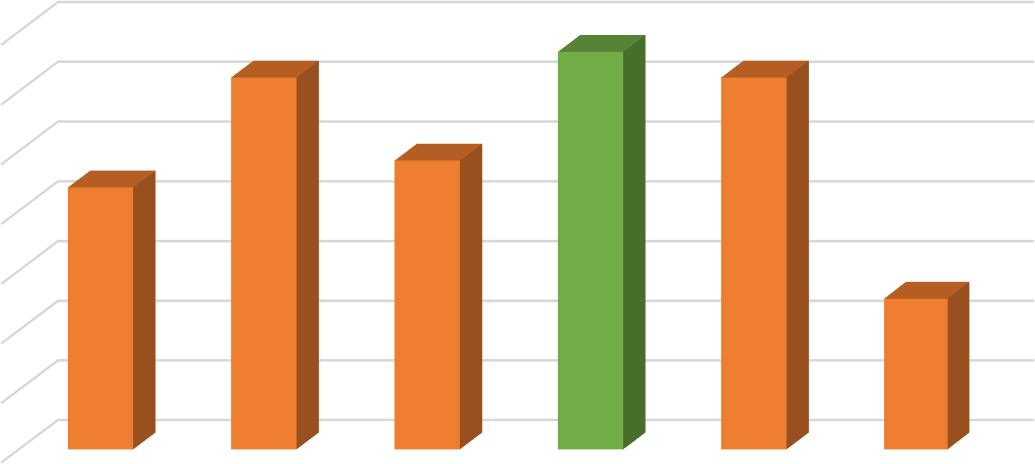
(

) of

each evaluation are averaged to provide the final score, and the resultant model is then fitted to the whole dataset in order to be used. Figure 4.4 and Table 4.7 display the cross-verified score chart and the proportion of each estimation technique employed in this model.

Table 4.7: Cross Validated

|  |  |
| --- | --- |
| **Algorithm Name** | **Cross Validated Score (%)** |
| KNN | 95.39 |
| Random Forest | 97.23 |
| Naive Bayes | 95.84 |
| XGBoost | 97.66 |
| Stochastic Gradient Descent | 97.23 |
| Discriminate Analysis | 93.53 |



Cross-validatied Chart

97.66

98

97.23

97.23

97

95.84

96

95.39

95

94

93.53

93

92

91

K-Nearest

Neighbor

Random

Forest

Naïve Bayes

XGBoost

Classifier

Stochastic Discriminant

Gradient Analysis Decent

Figure 4.4: Cross-validated Chart

# Standard Deviation

The information from this study's data may also be used to assess the standard deviation. A dataset's standard deviation shows how far from the mean it deviates. The square root of the

variance of each data point is used to determine the standard deviation. The standard deviation rises when the data points deviate farther from the mean. The standard deviation for the top six methods is shown in Table 4.8.

Table 4.8: Standard Deviation

|  |  |
| --- | --- |
| **Algorithm Name** | **Standard Deviation** |
| KNN | -- |
| Random Forest | 0.07 |
| Naive Bayes | -- |
| XGBoost | 0.17 |
| Stochastic Gradient Descent | -- |
| Discriminate Analysis | 17.48 |

# Misclassification and Error

Determining an algorithm's effectiveness makes detecting any flaws difficult. The accuracy of a machine-learning model takes into account absolute error and mean square error after misclassification. Misclassification may happen if an incorrect attribute is selected. Misclassification occurs when all classes, groups, or categories of a variable have the same error rate. The amount of measurement inaccuracy is described by the absolute error. The term "Mean Absolute Error" refers to the average of all measurement-related absolute errors (MAE). Equation represents the Mean Absolute Error formula (xiii).

𝑛

1

𝑀𝐴𝐸 =

𝑛 ∑|𝑥𝑖 − 𝑥| … … … (𝑥𝑖𝑖𝑖)

𝑖=1

How well a regression line matches a group of points is determined by its Mean Squared Error (MSE). Equation provides the formalized computation for Mean Squared Error (xiv).

𝑛

1

𝑀𝑆𝐸 =

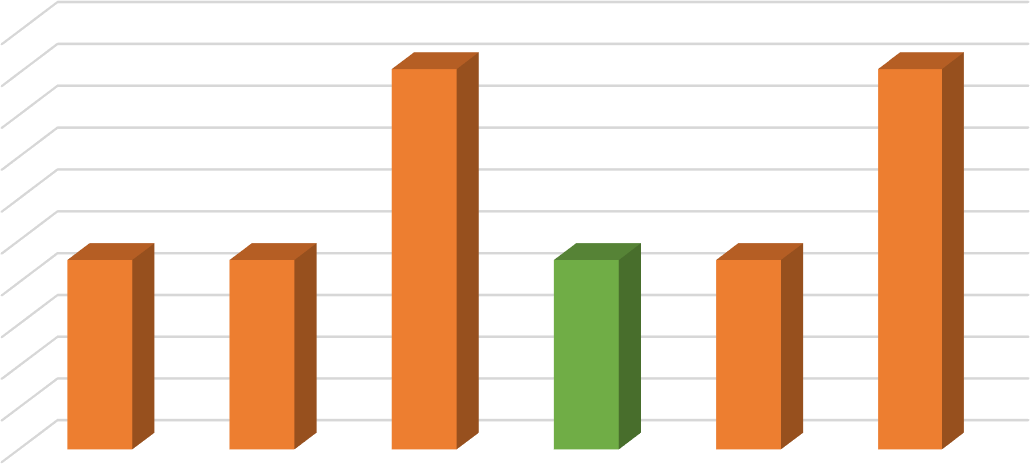
𝑛 ∑|𝑦𝑖 − 𝑦| … … … (𝑥𝑖𝑣)

𝑖=1

Table 4.9 and Figure 4.5 display the misclassification, mean absolute error, and mean square error in the algorithms.

Table 4.9: Misclassification Score

|  |  |
| --- | --- |
| **Algorithm Name** | **Misclassification (%)** |
| KNN | 2.27 |
| Random Forest | 2.27 |
| Naive Bayes | 4.55 |
| XGBoost | 2.27 |
| Stochastic Gradient Descent | 2.27 |
| Discriminate Analysis | 4.55 |



Misclassification Chart

5

4.5

4

3.5

3

2.5

2

1.5

1

0.5

0

4.55

4.55

2.27

2.27

2.27

2.27

K-Nearest Random Naïve Bayes XGBoost

Neighbor Forest Classifier

Stochastic

Gradient

Decent

Discriminant

Analysis

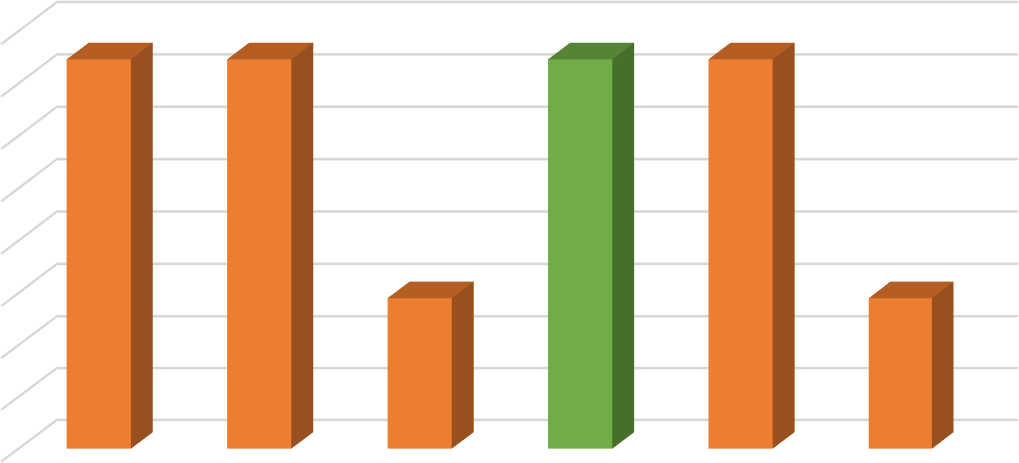
Figure 4.5: Misclassification Chart

# AUC Score

A system's potential categorization levels may be established using this performance metric in conjunction with machine learning technologies, which may be helpful in a variety of circumstances. The AUC may be calculated by contrasting a percentile of random positive instances where the model performs noticeably better to a percentile of random negative examples where the model performs noticeably worse. It is possible for this number to take on up to four distinct values, with one being the most likely. The potential values range from 0 to 1, with 0 being the lowest. Models that completely mis predict the future have an accuracy of zero, whereas models that completely correctly predict the future have an accuracy of one, as shown in Figure 4.6 and Table 4.10.

Table 4.10: AUC Score

|  |  |
| --- | --- |
| Algorithm Name | AUC Score (%) |
| KNN | 97.73 |
| Random Forest | 97.73 |
| Naive Bayes | 95.45 |
| XGBoost Classifier | 97.73 |
| Stochastic Gradient Decent | 97.73 |
| Discriminate Analysis | 95.45 |



AUC Chart

98

97.73

97.73

97.73

97.73

97.5

97

96.5

96

95.45

95.45

95.5

95

94.5

94

K-Nearest

Neighbor

Random

Forest

Naïve Bayes XGBoost

Classifier

Stochastic Discriminant

Gradient Analysis Decent

Figure 4.6: AUC Chart

# ROC Curve

ROC analysis is a crucial technique for assessing the efficacy of diagnostic tests and, more broadly, the precision of a statistical model that divides people into ill and healthy groups. One of the finest uses for ROC curve analysis is as a straightforward graphical tool to demonstrate the precision of a medical diagnostic test. The key curve score is displayed in Figure 4.7. Six machine learning methods were used to the dataset during the modeling stage to test their capacity to determine Safe Area. K-nearest Neighbor (KNN), Random Forest (RF), Naïve Bayes (NB), XGBoost (XGB), Stochastic Gradient Decent (SGD) and Discriminate Analysis (DA) are the methods used in this work.

The model predicted 0.98 auROC (area under the receiver operating characteristic curve) for the future test set, as shown in Figure 4.7.

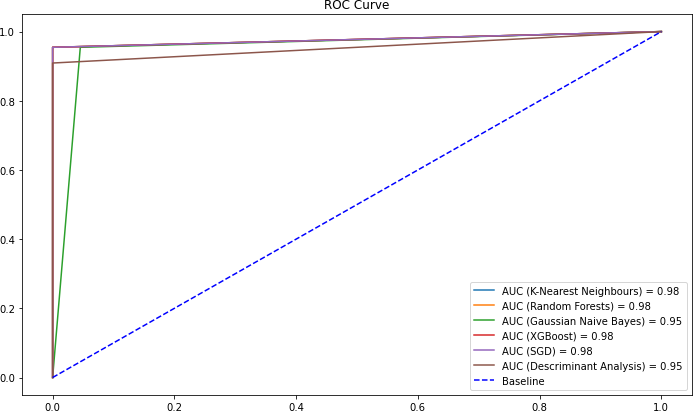


Figure 4.7: ROC Curve

# CHAPTER 5

# IMPACT ON SOCIETY & SUSTAINABILITY

# Introduction

Once a project is finished, its impact on society must be investigated and evaluated. This chapter's three parts discuss the effects of the Prediction of Safe Area project. This project will have a tremendous impact on humanity, as discussed in the section on how it will affect society or the state. Then there are ethical issues to think about. To understand how this project may assist patients and the medical industry of a country, the ethical component of the study was carefully evaluated. The project's long-term viability was also taken into account. It is discussed how this concept may develop in the future and benefit more people.

# Impact on Society

Police in Bangladesh gather crime statistics from eight different cities before informing the public about hotspots for specific types of crime. If information is shared on the internet, anyone can also view it on their own. According to our statistics, residents in the regions with the greatest crime rates will be able to learn about the types of crimes that are most common there and, as a consequence, will be ready and vigilant. Based on anticipated knowledge, newcomers to Bangladesh will avoid such locations and use caution if they must go there.

# Impact on Environment

The most criminally damaged location can have a significant influence on the local economy, according to Area Prediction Based on Year and Crime Types inside Bangladesh using Machine Learning. Depending on the topic, if there is an economic crisis, the local population will go, ruining the civilization and having a negative effect on the environment.

# Ethical Aspects

Regarding our project's ethical components, there are two effects.

* + 1. **Positive outcome:** The government will take different measures to suppress crime in such regions as a result of increased public awareness of crime. The

government will combat these offenses in the proper way. The government will take action to stop such crimes from happening again. It will be feasible to gradually reduce these crimes if we take action against them and if we receive administrative support.

* + 1. **Negative outcome:** The area's reputation will be tarnished, the economy will suffer, and the people's way of life will be upended as a result of these crimes. Relationships with regions where these crimes are more common will deteriorate, which will damage national unity. The mobility of females on the streets may be significantly impacted by these crimes.

# Sustainability Plan

Initial updates to crime-based software will occur periodically. Second, software will be created and an application for crime zone forecasting will be maintained for a while. People will be able to utilize the program for free, making it simple for them to use and obtain crime updates. In the future, if the software is produced by an organization, that organization must update the program on a regular basis in order to assist the general public in receiving accurate information. Long-term viability of the sustainability plan is assured.

# CHAPTER 6

# FUTURE SCOPE AND CONCLUSION

# Introduction

This chapter covered the potential scope. How might this project help the company grow more effectively in the future? What techniques may be used to create a better machine in the future? There is a neat ending to this chapter, which is provided at the end of the chapter. At the end of this chapter, a list of references is given.

# Future Scope of this Study

As a future extension of our work, this prediction system can be used widely and wisely. By the help of this system, investigator or security persons can markdown unsafe areas and can take necessary steps. Also, web application and android application can be developed using these models. Citizen of Bangladesh can use the applications in their daily life as now-a-days people use other online applications. This application can be maintained directly from police station as this application need daily update and crime data. By using this application users can get to know about searched area’s safety where they live in or pass by regularly or new places so that they can take precaution in future or can avoid the area. These models can be updated and further enhanced.

# Recommendations

The goal of this work was to determine whether machine learning techniques might be utilized to forecast crime areas. This essay provides information on crime area, including the percentage of each type of location that is more vulnerable to crime and the percentage of each specific type of crime that occurs there. This approach can identify any unexpected stages depending on the typical level of each diagnostic component. Dacoity, Robbery, Murder, Kidnapping, Explosives, Smuggling and other types of crimes are primary features of this work. In some situations, data might not be updated or wrong data uploaded to the dataset and in that situation the system might give incorrect forecast. Most of the high crime rate areas have murder cases. Also kidnapping, narcotics, smuggling, robbery are commonly noticed in high crime rate areas. Individuals should try to avoid these kinds of activities. If

anyone see any type of criminal activity happening, they should call the police without making any delay. Following government rules in one of the most important responsibilities for people of a nation.

# Conclusion

We utilized machine learning prediction model for predicting Safe Area from Bangladesh Crime Data throughout this work. Data pre-processing, outliers, missing values, data reduction, feature selection, and modeling were all used to create the prediction system. As prediction model, we employed six machine learning methods. Some graphs and plots are used - ROC Curve Plot, Pie Chart to visualize the ratio of dataset target class, Bar Chart to visualize six machine learning model's multiple validation scores and other tables. One of the built models helped us to achieve very high prediction accuracy and the model can predict safe area 97.73% accurately and it was further double checked. Additionally, there is a problem with our inability to foresee when a crime will occur. Moment is a crucial factor in crime; we must foresee not only the crime's location but also the appropriate time.

# REFERENCES

1. “S. Sathyadevan, M. S. Devan, S. Surya Gangadharan, "Crime analysis and prediction using data mining", *First International Conference on Networks & Soft Computing, (ICNSC2014),* 2014”.
2. “Sohrab Hossain, Ahmed Abtahee, Imran Kashem, Mohammed Moshiul Hoque, Iqbal H. Sarker, "Crime Prediction Using Spatio-Temporal Data", *Communications in Computer and Information Science book series, (CCIS, volume 1235),* 2020”.
3. “Olta Llaha, "Crime Analysis and Prediction using Machine Learning", *43rd International Convention on Information, Communication and Electronic Technology (MIPRO),* 2020”*.*
4. “Ying-Lung Lin, Meng-Feng Yen, Liang-Chih Yu, "Grid-Based Crime Prediction Using Geographical Features", *ISPRS Int. J. Geo-Inf. 2018, 7(8), 298,* 2018”.
5. “P. B. Fajemiseye, C. Ugwu, L. N. Onyejegbu, "An Intelligent Crime Prediction System Using Artificial Neural Network and Geographical Information System", *International Journal of Scientific & Engineering Research Volume 11, Issue 9,* September-2020”.
6. “Prajakta Yerpude, "Predictive Modelling of Crime Data Set Using Data Mining", *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.4,* July-2017”.
7. “Shanjana A.S, Dr.R.Porkodi, "Crime Analysis and Prediction Using Data Mining: A Review", *IJCRT, Volume 9, Issue 2,* 2 February-2021”.
8. “Nikhil Dubey, S. K. Chaturvedi, "A Survey Paper on Crime Prediction Technique Using Data Mining", *Nikhil Dubey et al Int. Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 4, Issue 3(Version 1), March 2014, pp.396-400*, 2014”.
9. “Vikas Grover, Richard Adderley, Max Bramer, "Review of Current Crime Prediction Techniques", *International Conference on Innovative Techniques and Applications of Artificial Intelligence, SGAI 2006: Applications and Innovations in Intelligent Systems XIV, pp 233-237,* 2006”.
10. “Hitesh Kumar, Bhavn Saini, Ginika Mahajan, "Crime Prediction & Monitoring Framework Based on Spatial Analysis", *International Conference on Computational Intelligence and Data Science, Volume 132, Pages 696-705,* 2018”.
11. “Tahani Almanie, Rsha Mirza, Elizabeth Lor, "Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots", *International Journal of Data Mining &*

*Knowledge Management Process (IJDKP) Vol.5, No.4,* July 2015”.

1. “Ginger Saltos, Mihaela Cocea, "An Exploration of Crime Prediction Using Data Mining on Open Data", *International Journal of Information Technology & Decision Making, Vol. 16, No. 05, pp. 1155-1181*, 2017”.
2. “A. Malathi, Dr. S. Santhosh Baboo, "Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters", *Global Journal of Computer Science and Technology, Volume 11 Issue 11Version 1.0,* July-2011”.
3. “Shiju Sathyadevan, Devan M. S., Surya S Gangadharan, Crime Analysis and Prediction Using Data Mining", *Conference: First International Conference on Networks & Soft Computing (ICNSC),* 2014”.
4. “Chung-Hsien Yu, Max W. Ward, Melissa Morabito, Wei Ding, "Crime Forecasting Using Data Mining Techniques", *IEEE 11th International Conference on Data Mining Workshops*, 2011”.
5. “Devendra Kumar Tayal, Arti Jain, Surbhi Arora, Surbhi Agarwal, Tushar Gupta, Nikhil Tyagi, "Crime detection and criminal identification in India using data mining techniques", *Journal of Knowledge, Culture and Communication, AI & SOCIETY volume 30, pages117– 127*, 2015”.
6. “Suhong Kim, Param Joshi, Parminder Singh Kalsi, Pooya Taheri, "Crime Analysis Through Machine Learning", *IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON),* 2019”.

# PLAGIARISM REPORT

