MOTORBIKE ACCIDENT SEVERITY PREDICTION USING MACHINE LEARNING ALGORITHMS

BY

DEWAN FUAD HASSAN OVI
ID: 181-15-2067
AND

SANJANA SUROVI MEEM
ID: 181-15-2057

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

Supervised By

Md. SABAB ZULFIKER
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised By

Md. MAHFUJUR RAHMAN
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

The research paper " Motorbike Accident Severity Prediction Using Machine Learning Algorithms" submitted by Dewan Fuad Hassan Ovi (ID: 181-15-2067) and Sanjana Surovi Meem (ID: 181-15-2057) to Daffodil International University's Department of Computer Science and Engineering has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13 January 2022.

Board of Examiners

Dr. Touhid Bhuiyan **Chairman**
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Tsiddique

Shah Md. Tanvir Siddique **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Reduan

Md. Reduanul Haque **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Farid

Dr. Dewan Md. Farid **External Examiner**
Associate Professor
Department of Computer Science & Engineering
United International University

DECLARATION

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

Supervised by:

____

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

Co-Supervised by:

____

Md. Mahfujur Rahman
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

Dewan Fuad

____

Dewan Fuad Hassan Ovi
ID: 181-15-2067
Department of CSE
Daffodil International University

Sanjana surovi meem

____

Sanjana Surovi Meem
ID: 181-15-2057
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First and foremost, we express our heartfelt gratitude to Almighty Allah for His glorious blessings, which have enabled us to successfully finish the final year project/internship.

Md. Sabab Zulfiker, Lecturer in the Department of Computer Science and Engineering at Daffodil International University in Dhaka, is deserving of our thanks and respect. Our supervisor must have considerable knowledge of and a strong interest in the subject of "Machine Learning" in order to perform this task. This study was accomplished thanks to his never-ending patience, intellectual direction, frequent and energetic monitoring, constructive criticism, helpful recommendations, and reading numerous poor copies and rectifying them at all stages.

We would like to convey our heartfelt gratitude to Professor Dr. Touhid Bhuiyan, Head, Department of CSE, as well as other academic members and personnel from Daffodil International University's CSE department, for their invaluable assistance in completing our project.

We would like to express our gratitude to all of our Daffodil International University classmates who participated in this conversation as part of their course work.

Finally, we must recognize and acknowledge our parents' unwavering support and patience.

ABSTRACT

Motorbike accident is most dangerous among all accident on road. Many people dies because of bike accident. Motorbike accident damages both public and private property. A victim's family have to spent a big amount of money for the treatment of the biker. Sometimes their family falls into financial crisis to arrange money for treatment of the victim. Bikers need to be aware about accident and should avoid the facts that causes accident. In this study different machine learning algorithm has been employed to predict the severity of the motorbike accident. Then we collect data depending on those criteria, such as speeding, overtaking, turning, bike fitness issues, speed-breakers without signs, unsafe lane changes, talking with a passenger, and highways without road dividers, among others. We only collect information from bikers who have been in an accident. We processed all of the data once it was collected and developed a processed dataset. On our processed dataset, we used machine learning techniques. Machine learning has been employed in various prediction and detection systems since their inception. Random Forest, Multilayer Perception (MLP), Decision Tree, Logistic Regression, k-Nearest Neighbors (KNN), AdaBoost, GNB, SVM with RBF Kernel, Linear SVC and Gradient Boosting are just a few of the techniques we utilize. MLP offered the greatest results in terms of accuracy. It showed an accuracy of 83.10%. And again MLP performs better in terms of sensitivity, specificity, F1-Score and precision.

TABLE OF CONTENTS

CONTENTS **PAGE NO:**

LIST OF FIGURES

LIST OF TABLES

TABLES **PAGE NO:**

CHAPTER 1

INTRODUCTION

1.1 Introduction

Road accident is one of the most life-threatening hazards that humans face. Numerically motorbike accident occurs more than other accidents and it's more dangerous than all type of accidents. Bikers fall into accidents because of unawareness, over speeding, not obeying the traffic rules, signals, road sign etc. Among all of these causes over speeding and unawareness are two main reasons for motorbike accident. Because motorbike is the first choice of today’s young generation. When they ride a bike they try to get the top speed. They drive bikes with high-speed as a result they fall into accidents. Sometimes they take their friends as pillions in their bike and talk to each other when riding bikes and fall into accidents. Motorbike accident also hampers to state/government properties. Motorbike accident causes huge loss to family of the victims, public and government properties. Sometime victims’ families spent almost whole of their property for treatment of the victim. Making people aware about motorbike accident is one of the best solution to reduce motorbike accident. In this study, machine learning algorithms has been used to predict motorbike accident severity.

1.2 Motivation

Motorbike accident is very dangerous. A huge number of deaths occurs because of bike accidents. Most of the time victim have to face fatal severity. Accident cause severe financial problem to the victims family as they have to arrange money for treatment of victims. The family of the victims fall into financial crises. It also destroys the public and private properties. So, we have to aware people about this and motorbike riders should ride motorbike safely. We can reduce the occurrences of motorbike accidents and their fatality rates if we can predict the possibility of motorbike accidents and their severity earlier.

1.3 Rationale of study

As previously stated, there has been less major study done in the past with Motorbike accident severity prediction in Bangladesh. That is why working with motorbike accident and machine learning approaches is appealing to us.

Machine learning is a branch of artificial intelligence that employs a number of optimization, probabilistic and statistical approaches to allow computers to "learn" from previous examples and detect difficult-to-find patterns in big, noisy or complex data sets. Machine learning algorithms are utilized in a variety of applications, including detection and classification. Cancer prediction, a holistic study of software defect prediction, dermatological illness diagnosis, and other applications of machine learning are examples of using machine learning algorithms. Machine learning is increasingly being used to perform several forms of detection and risk prediction. In more complicated issues, machine learning approaches may be useful as a supplement to regression results. We decided to use machine learning for our prediction work because of its broad scope.

1.4 Expected outcome

Bike accident is one of the most life threatening accident on the road. Many family fall into economic crisis to arrange money for treatment of the victim. Bike accident also make cost to government as it hampers government properties. By this study people will come to know about the reasons that causes bike accident. Biker will be aware and they will be well prepared before they bike ride. They will avoid the reasons that causes bike accident. So, bike accident and its' severity will be reduced.

People have to go one place to another place for works. Sometimes they have to take ride on bikes as pillion. By this research they will deeply know about bike accident. If they afraid that the bike is going to be fall into accident then they can take proper action about it. They can aware that biker about this. Many of our relatives, friends, neighbors use bikes in their daily life. We can aware them about the reason of bike accident and severity behind every reason. So that they will be aware about bike accident and we can avoid those reasons for which bike accident occurs. Parents of young bikers can be also consulted. So that their parents can take care of them. Because young generation are in risk of bike accident as they like to drive rush.

1.5 Layout of the report

The following is a list of the contents of this research paper:

- ✓ The first chapter discusses the research's objective, justification for the study and predicted outcomes.

- ✓ The second chapter discusses related works, a study summary, the problem's scope, and difficulties.

- ✓ The workflow of this study, the data gathering technique, statistical analysis, and future implementation are all covered in Chapter 3.

- ✓ Chapter four discusses experimental evaluation and other related debates, as well as the numerical and graphical results of study.

- ✓ The significance of this research on society is discussed in Chapter 5.

- ✓ The sixth chapter offers a summary of this study.

CHAPTER 2

STUDY OF THE BACKGROUND

2.1 Introduction

Relevant works, a study synopsis, the breadth of the topic, and problems will be examined in this part. Several related works, research papers, classifiers, primary methodologies and accuracies that are relevant to our study is outlined in the related work section. A summary of selected relevant studies has been produced. The breadth of the problem section outlines how the work paradigm can contribute to the problem. Finally, in the difficulties section, we discuss the hurdles and dangers we encountered while conducting this research.

2.2 Related Works

This research paper's literature review portion will show recent related works on prediction done by some researchers. We watched and studied their work to learn more about the techniques and strategies they employ.

For the prediction of occupational accidents, Sarkar, et al. [1] employed Genetic Algorithms, Artificial Neural Networks, and Particle Swarm Optimization. They looked at categorical data as well as unstructured text. With a score of 90.67 percent, PSO-SVM has the best accuracy.

Iranitalab, et al. [2] employed machine learning to forecast the severity of a crash. To anticipate the severity of the collision, they applied four distinct classification algorithms. In their research, they used Nearest Neighbors Classification, Multinomial Logit, Random Forest, and Support Vector Machine.

For real-time highway accident prediction, Shuming, et al. [4] applied the k-Nearest Neighbor approach. For the first time, the K-nearest neighbor algorithm was used to forecast a real-time highway collision.

Rujun, et al. [5] investigated traffic accident prediction. The RBF neural network was used to create this model. They used the RBF neural network model to extrapolate and anticipate the number of deaths and economic losses from 2000 to 2006.

Machine learning techniques were utilized to analyze road accidents by Abraham, et al. [6]. They used four machine learning approaches to estimate the severity of traffic accidents. They used a concurrent hybrid model, SVM, Decision Tree and hybrid learning methodologies. The hybrid decision tree-neural network technique outperformed the separate approaches among the machine learning paradigms studied.

Satu, et al. [7] used different decision tree induction techniques to examine 892 traffic incidents on Bangladesh's N5 National Highway in an attempt to determine traffic accident patterns. They also drew up restrictions for trees for the purpose of reducing the number of accidents on road.

Kumar, et al. [8] utilized two data mining approaches, association rule mining and K-means clustering to identify the most accident-prone areas in India, as well as the primary characteristics connected to those incidents.

Bülbül, et al. [9] used machine learning techniques to investigate the current state of road accident occurrence in Istanbul. The authors used CART algorithm to calculate the risk of accident and got an accuracy of more than 81.5 percent.

Elahi, et al. [10] proposed a strategy for detecting the possibility of traffic accidents using vision-based techniques in the setting of Bangladesh. They learned from roadside video data and reached an accuracy rate of 85 percent in specific conditions.

Nandurge et al. [11] used two data mining techniques, association rule mining and K-means clustering to identify the primary factors linked to traffic accidents.

Esmaeili, et al. [12] employed logistic regression to determine the impact of road faults on accident severity depending on the vehicle state after an accident.

Beshah, et al. [13] used decision trees, KNN and Naïve Bayes data mining approaches to uncover links between recorded road features and accident intensity in Ethiopia in their study. They also devised a set of guidelines to improve Ethiopian road safety based on the foundational principles.

A machine learning-based accident prediction model was proposed by Mohanta, et al. [14]. In this model, they applied a variety of machine learning classifiers. For forecasting accident severity, they used Logistic Regression, Artificial Neural Networks, Decision Trees, K-Nearest Neighbors, and Random Forest.

In their work, Ramani, et al. [15] used Random Tree, Decision Tree algorithms, J48, C4.5, and Decision Stump to analyze a database of fatal incidents that happened in the United Kingdom in 2010. For performance comparison, they employed K-folds Cross-Validation techniques to quantify the unbiased estimate of the four prediction models.

Theofilatos et al. [16] compared machine learning and deep learning algorithms to forecast real-time crashes in their paper. Random forest, k-nearest neighbor, Naïve Bayes, support vector machine, decision tree, shallow neural network, and deep neural network were among the models investigated. The Naïve Bayes model was shown to have a high accuracy in this investigation.

Abdel-Aty, et al. [17] studied angle collisions at unsignalized junctions using machine learning approaches. MARS and Negative Binominal were employed in this investigation. MARS was integrated with the 'random forest' machine learning approach. In their investigation, they advocated MARS as an effective approach for prediction.

Using machine learning, Labib, et al. [18] examined traffic accidents in Bangladesh and predicted accident severity. In their research, they employed kNN, Decision Tree, Naïve Bayes, and AdaBoost. AdaBoost outperformed the others in this trial.

Matías, et al. [19] suggested a machine learning framework for workplace accident investigation. They employed classification trees, SVM, and Extreme Learning Machines as machine learning approaches.

Wahab, et al. [20] offered a comparison of machine learning-based methods for predicting the severity of motorcycle crashes. In Ghana, there is a dearth of research on this topic. The data was divided into four groups based on injury severity: dead, hospitalized, wounded, and damage-only. Random Forest, Instance-Based Learning with Parameter k, and Decision Tree are the machine learning approaches they employed. Random Forest produces the greatest results in this investigation.

Zhang, et al. [21] examined the efficacy of several machine learning and statistical approaches in predicting collision injury severity. In this work, they used four machine learning algorithms. Random Forest, Decision Tree, k-Nearest Neighbors, and SVM are the four approaches. In this investigation, Random Forest proved to be more accurate.

A model for accident prediction at a Korean container port was proposed by Jae Hun Kim [22]. The model is based on machine learning. By analyzing the outcomes of many models in terms of accuracy, precision, recall, and F1 score, the best model was found. In terms of all performance indicators, the gradient boosting model with a 6-hour interval had the best results.

Machine learning approaches were used by Komol, et al. [23] to assess the severity of crashes involving vulnerable road users. The study's crash data comes from Queensland, Australia, and spans the years 2013 to 2019. They compared SVM, kNN, and Random Forest machine learning algorithms. Random Forest had the highest accuracy in this investigation.

R.E. AlMamlook, et al. [24] examined machine learning techniques to predict the severity of traffic accidents. This work provides models for selecting a collection of influential elements and developing a model for categorizing the severity of injuries. Various machine learning approaches were used in this model, including Naïve Bayes, Logistic Regression, AdaBoost, and Random Forest. Among all the approaches, Random Forest had the best accuracy (75.5%).

2.3 Exploration and Overview

Currently, new technologies that combine artificial intelligence, machine learning and deep learning are being investigated for use in any type of prediction and detection model. Various machine-learning methods have lately been used to prediction and detection. For any detection model MLP, kNN, CNN, SVM, Logistic Regression and other methods are common. According to prior research, the popularity and effectiveness of the kNN, SVM, Random Forest, MLP, GNB and Decision Tree algorithms for detection or prediction models are highly uses. In this study, Random Forest, MLP, Decision Tree, Logistic Regression, kNN, AdaBoost, SVM, Linear SVC, GNB, and Gradient Boosting algorithms has been utilized to predict the severity of motorbike accident in Bangladesh, with the highest accuracy 83.10%.

2.4 Scope of the problem

This research focuses on developing a model through data analysis and machine-learning methods. The severity of a motorcycle collision can be predicted using our proposed model. This prognosis will have a profound social impact. Biker, particularly the younger generation, can avoid the causes of bike accidents. This model will be useful to traffic cops in a variety of situations. Motorbike accident is far more dangerous than any other road collision. As a result, this model will assist bikers, ordinary people, and mindful people in avoiding bike accidents.

This concept would also assist the government and riders' families to save money, as bike accidents cause damage to government and public property, as well as costing their families money for treatment of the victim. Occasionally, the bikers are the family's only source of income. If he has an accident, their family will be in financial trouble. This concept will be beneficial to bikers in preventing bike accidents and resolving these issues. Machine learning and artificial intelligence have recently been applied for various detection and prediction, with promising results. As a result, we decided to construct motorbike accident severity prediction model using machine learning.

2.5 Difficulties

Some difficulties had been faced while conducting the research. Gathering data was very difficult and challenging for us. The whole research had been done in Covid-19 pandemic. Almost the entire year 2020-2021 was spent under lockdown. School, Colleges, Universities, offices, almost all the organizations were remain closed during lockdown. People couldn’t stay out from home except emergency need. Vehicles didn’t seen on the road during lockdown. So, it was very difficult to collect data. As the medicine shops were remain open in the lockdown, Some copy of data collecting questions had been printed and taken it to some medicine shop. The medicine shop owners and their stuff allowed us to collect data by standing in their shop. We stood in those medicine shops and collected data. Sometimes govt. of Bangladesh overhauled the lockdown. At that time, those printed copies was taken to some departmental store, mobile shops etc. Store owners and stuffs also allowed us to collect data by standing in their shop. We stood in those shops and collected data. Some data has also been collected from the road.. We talked to some bikers and collected data. A few data was collected from online. An online from had been created and spread it to some bikers group. A little number of data was collected from newspaper, TV news and Facebook. Some of our relatives also faced bike accident. Some data was collected from them. Covid-19 pandemic interrupt us to collect data. Otherwise more data could be collected.

CHAPTER 3

THE METHODOLOGY OF RESEARCH

3.1 Introduction

The goal of this study is to develop a model that can predict the severity of a motorcycle collision. The Prediction Model is built using data from people's daily lives as well as other connected data. A variety of machine-learning methods has been used to develop this model. In this study, the following algorithms have been employed: Logistic Regression, kNN, SVM with RBF kernel, Decision Tree, AdaBoost, Random Forest, GNB, MLP, Gradient Boosting, Linear SVC. A total of twenty significant criteria has been used that were all linked to motorbike accident. We looked into some of the factors that contributed to the outcome. Prior to implementation, data has been processed. Then, the processed data has been fed to the machine learning algorithms. Different performance metrics had been computed for each of classifiers and compared them. It was discovered that MLP has the highest accuracy and is the most appropriate for our suggested model.

3.2 Process of gathering data

The data set consists of a large number of related and necessary factors that causes motorbike accident. Firstly we tried to collect data from our relatives and neighborhood. Some printed copy of questions has been taken to some departmental store, pharmacy, mobile shop etc. Those shopkeeper and their stuffs allowed us to collect data from their customers by standing in their shops. A little number of data collected from newspapers, online news portal and Facebook. And few data collected through online form. We made a google from and spread it to many bikers group. Data has also been collected from the bikers on road. We talked with them and they filled the printed form. All of our data has been collected from Savar, Dhamrai, Manikganj and online google form. Total 567 data has been collected.

Table 3.1 shows the factors that have been considered for predicting the severity of motorbike accident.

Variable Name	Variable Type	Variable Description	Possible values
OS	Predictor	Whether the biker faced accident for over speeding or not.	Yes (1), No (0)
OT	Predictor	Whether the biker faced accident for overtaking or not.	Yes (1), No (0)
Turn	Predictor	Whether the biker faced accident for turning or not.	Yes (1), No (0)
BFP	Predictor	Whether the biker faced accident for bike fitness problem or not.	Yes (1), No (0)
SBWS	Predictor	Whether the biker faced accident for a speed-breaker without any sign or not.	Yes (1), No (0)
ULC	Predictor	Whether the biker faced accident for unsafe lane changing or not.	Yes (1), No (0)
BRC	Predictor	Whether the biker faced accident for bad road condition or not.	Yes (1), No (0)
DIA	Predictor	Whether the biker faced accident for driver improper action or not.	Yes (1), No (0)
TWP	Predictor	Whether the biker faced accident for talking with pillion or not.	Yes (1), No (0)
NRD	Predictor	Whether the biker faced accident for no road divider on highway or not.	Yes (1), No (0)
UOB	Predictor	Whether the biker faced accident for unawareness of biker or not.	Yes (1), No (0)
UOAV	Predictor	Whether the biker faced accident for unawareness of another vehicle or not.	Yes (1), No (0)
IB	Predictor	Whether the biker faced accident as he is an inexperienced biker or not.	Yes (1), No (0)
TWI	Predictor	Whether the biker faced accident for three wheeler involvement or not.	Yes (1), No (0)

PDI	Predictor	Whether the biker faced accident for pedestrian involvement or not.	Yes (1), No (0)
SUS	Predictor	Whether the biker faced accident for sudden stop or not.	Yes (1), No (0)
RC	Predictor	The condition of the road surface.	Wet (1), Dry (0)
HELM	Predictor	Whether the biker wearing helmet or not at the time of accident.	Yes (1), No (0)
NOPR	Predictor	Number of rider on the bike at the time of accident including biker.	1, 2, 3
Severity	Target	Accident severity	Fatal (1), Mild (0)

Table 3.1: Variables that are used to predicting the severity of motorbike Accident

Severity has been divided into two part, one is mild and another is Fatal. Table 3.2 shows their detailed information.

Mild	**Fatal**
➢ Simple tissue damage	➢ Fracture ➢ Hospitalized ➢ Paralyzed ➢ Fall into coma

Table 3.2: Detailed information of the severity of accidents.

3.3 Analyzing Data

Figure 3.1 shows number of mild severity caused for different factors.

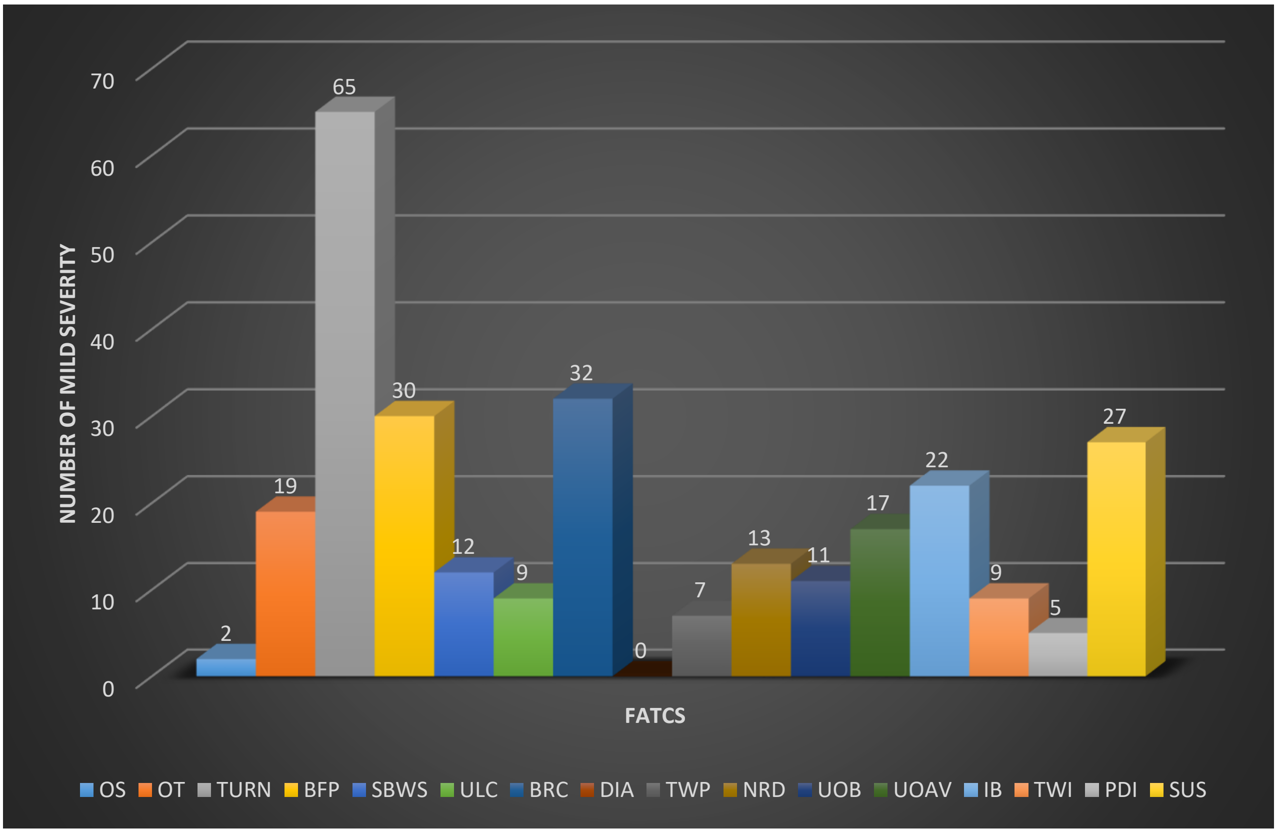


Fig 3.1: Number of mild severity caused for different factors.

Here we can see the number of mild severity caused for different factors. Here we can see that for over-speeding number of mild severity is 2, for overtaking number of mild severity is 19, for turning number of mild severity is 65, for bike fitness problem number of mild severity is 30, for speed-breaker without any sign number of mild severity is 12, for unsafe lane change number of mild severity is 9, for bad road condition number of mild severity is 32, for driver improper action number of mild severity is 0, for talking with pillion number of mild severity is 7, for no road divider on highway number of mild severity is 13, for unawareness of biker number of mild severity is 11, for unawareness of another vehicle number of mild severity is 17, for inexperienced biker number of mild severity is 22, for three wheeler involvement number of mild severity is 9, for pedestrian involvement number of mild severity is 5, for sudden stop number of mild severity is 27. So, highest number of mild severity happened for turning with number 65 and no mild severity happened for driver improper action.

Figure 3.2 shows number of fatal severity caused for each fact.

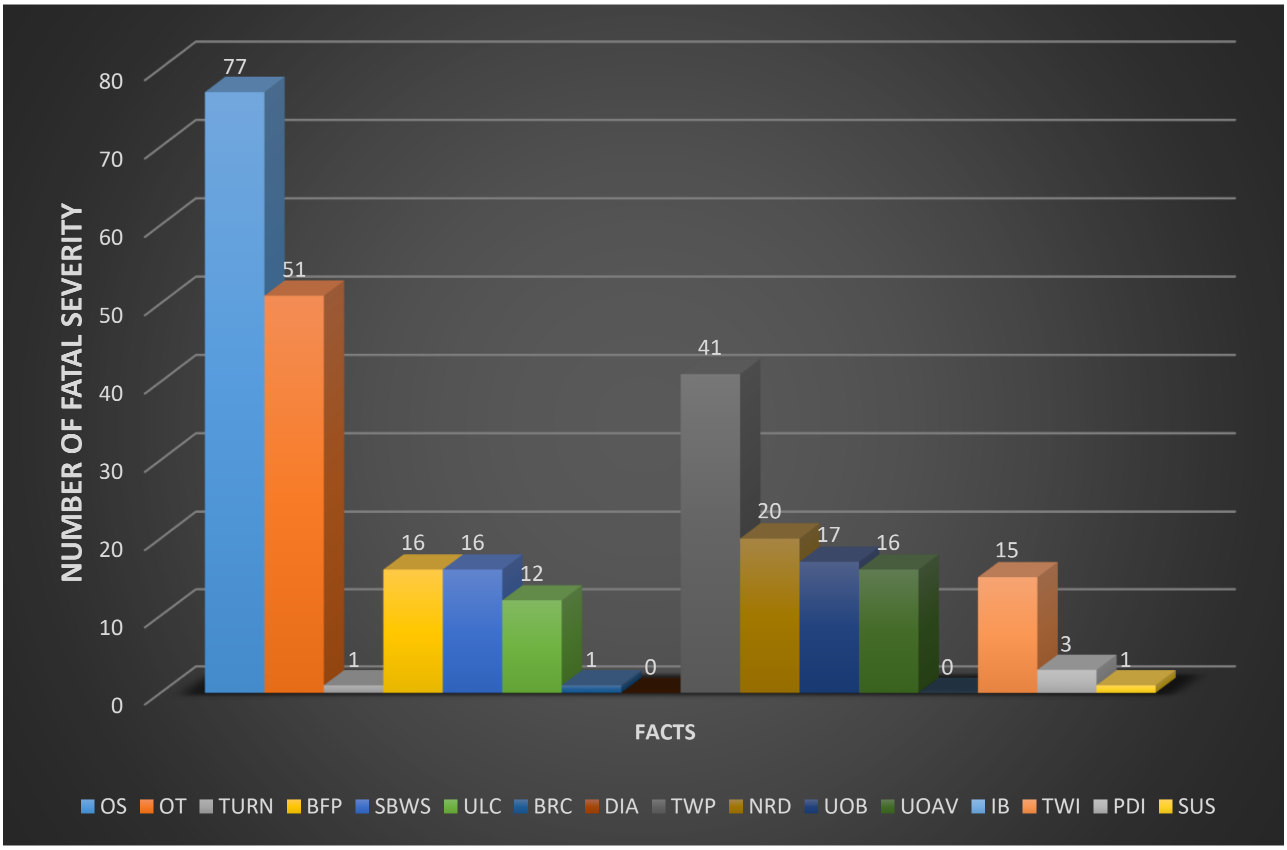


Fig 3.2: Number of fatal severity caused for each fact

Here we can see the number of fatal severity caused for different factors. Here we can see that for over-speeding number of fatal severity is 77, for overtaking number of fatal severity is 51, for turning number of fatal severity is 1, for bike fitness problem number of fatal severity is 16, for speed-breaker without any sign number of fatal severity is 16, for unsafe lane change number of fatal severity is 12, for bad road condition number of fatal severity is 1, for driver improper action number of fatal severity is 0, for talking with pillion number of fatal severity is 41, for no road divider on highway number of fatal severity is 20, for unawareness of biker number of fatal severity is 17, for unawareness of another vehicle number of fatal severity is 16, for inexperienced biker number of fatal severity is 0, for three wheeler involvement number of fatal severity is 15, for pedestrian involvement number of fatal severity is 3, for sudden stop number of fatal severity is 1. So, highest number of fatal severity happened for over-speeding with number 77 and no fatal severity happened for driver improper action and inexperienced biker.

Figure 3.3 shows number of mild and fatal injuries by road condition.

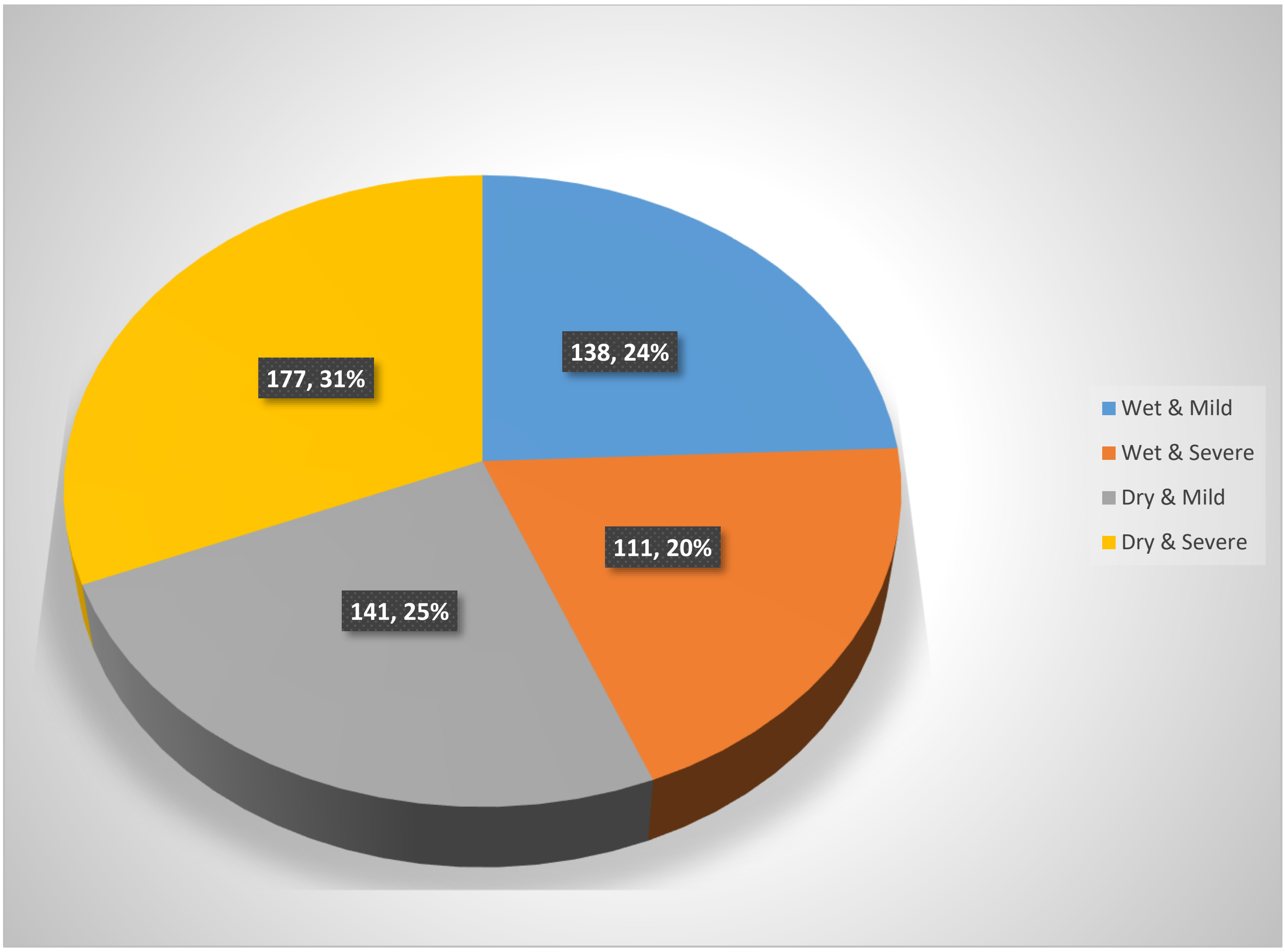


Fig 3.3: Number of mild and fatal injuries by road condition.

In the pie chart we can see the number and percentage of mild and fatal injuries by road condition. For dry road number of mild severity is 141 which is 25% of total severity and number of fatal severity is 177 and it is 31% of total severity. We can also see that for wet road condition mild severity is 24% with number 138 and fatal severity is 20% with number 111.

Figure 3.4 shows number of fatal and mild injuries by helmet.

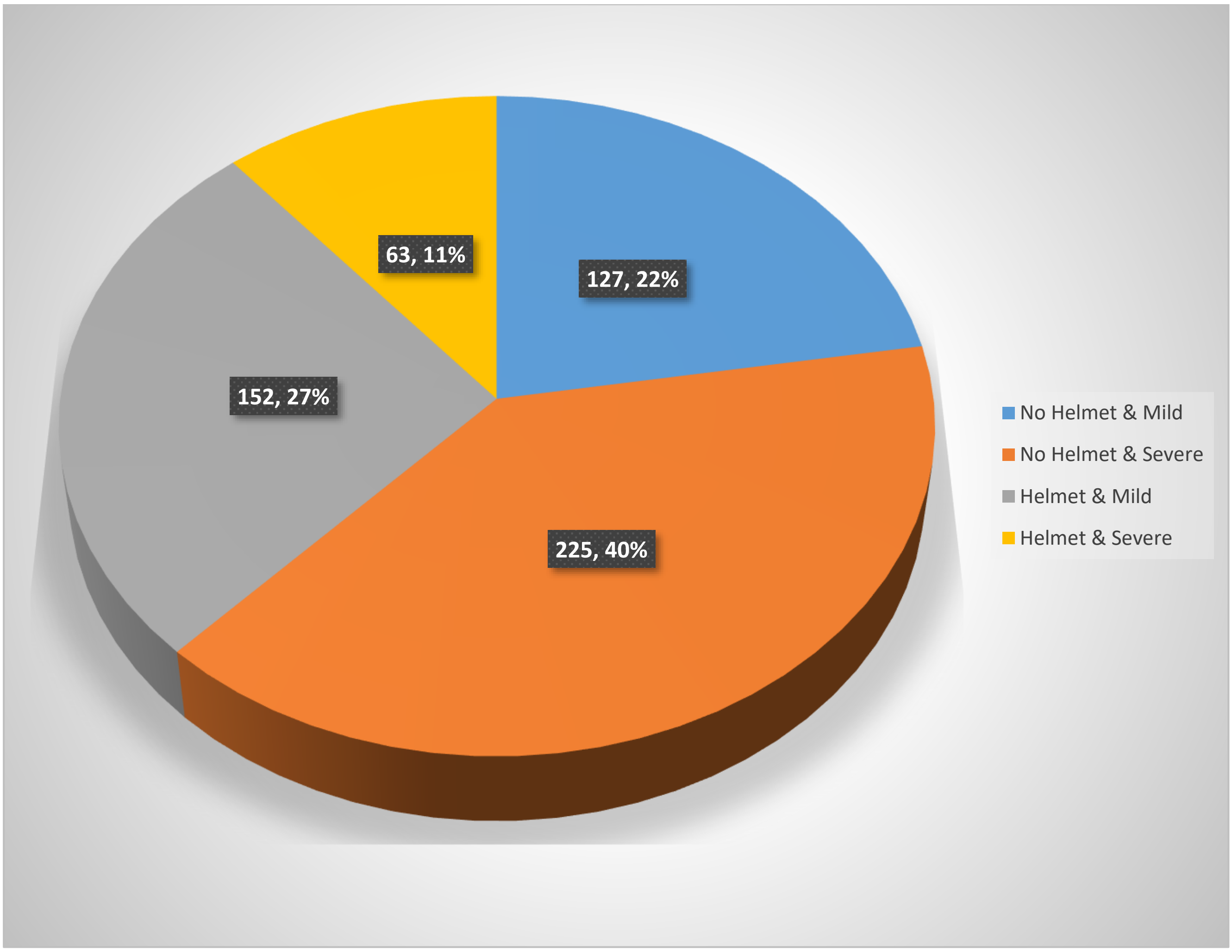


Fig 3.4: Number of fatal and mild injuries by helmet.

In the pie chart we can see the number and percentage of mild and fatal injuries by helmet. For helmet number of mild severity is 152 which is 27% of total severity and number of fatal severity is 63 and it is 11% of total severity. We can also see that for no helmet mild severity is 22% with number 127 and fatal severity is 40% with number 225.

3.4 Proposed Methodology

Fig 3.5 shows the step by step processes for implementing the study.

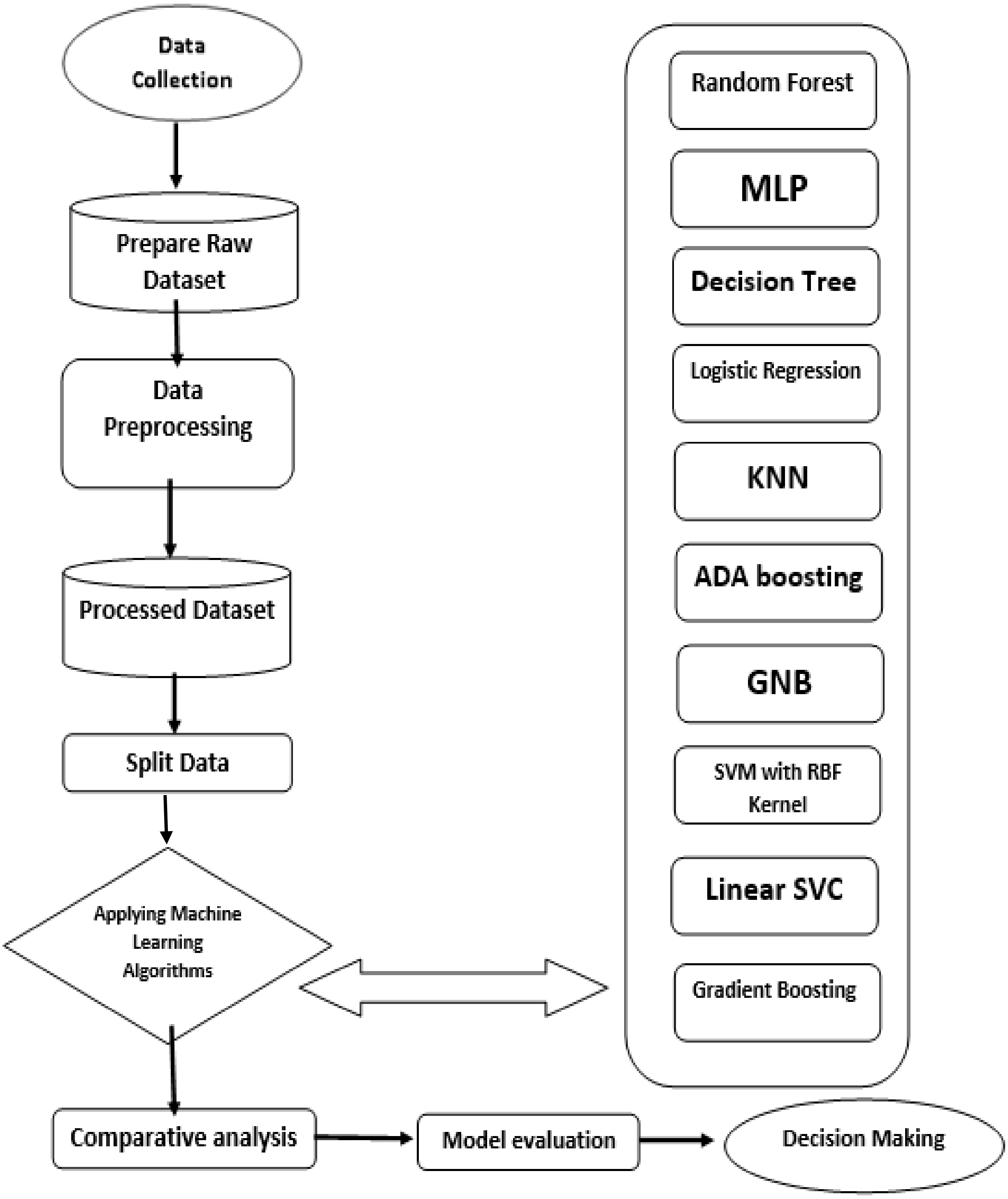


Fig 3.5: Steps of our proposed methodology.

First the data has been collected. We have talked about our data collection process before in this paper. The data has been collected physically and using online google form. Then raw dataset has been prepared using Microsoft excel. All the data inserted into a excel file.

After that processing of the data has been started. Some categorical data, missing data, text data and numerical data has been detected after collecting data. Then we decide that by processing the data, we can make it acceptable for algorithms. Data processing is the ability to convert data into a usable format once it has been collected. Data or information is processed in a specified format to facilitate result.

Figure 3.2 depicts our data preparation procedure.

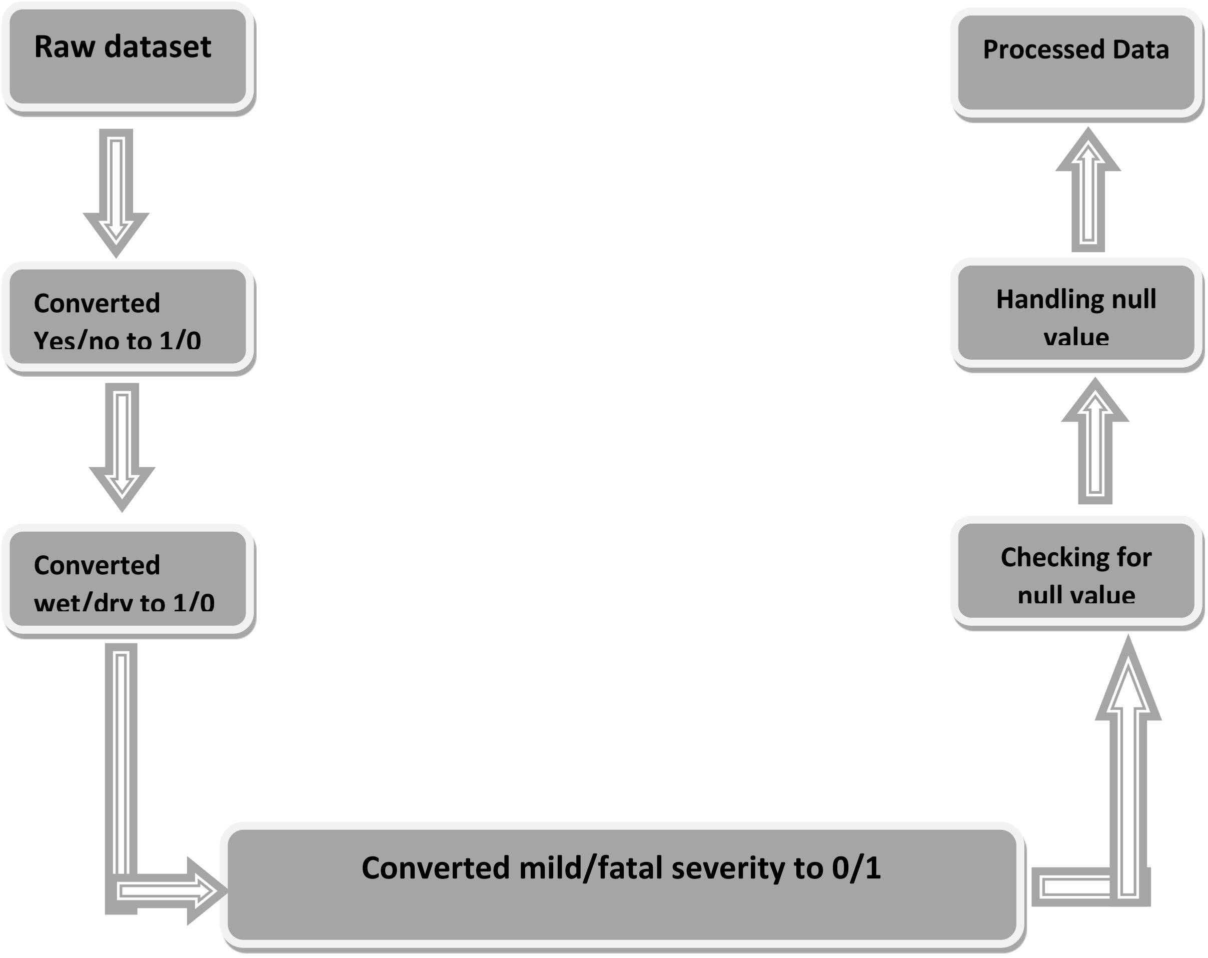


Fig 3.6: Data preparation procedure

First, the raw dataset was taken. Then all the yes and no was converted to 1 and 0. After that wet and dry was converted to 1 and 0. Severity has been divided into two option. One is mild severity and another is fatal severity. So, fatal and mild was transformed to 1 and 0 as well. Then all null values was checked. When any null value was found then it was handeled. And finally got the processed data.

Then we split our data into test and train dataset and applied ten algorithms. We applied : kNN, logistic regression, SVM with RBF kernel, GNB, random forest, decision tree, multilayer perception (MLP), Linear SVC, gradient boosting and ADA boosting classifier. The theoretical description of the used classifiers are given below:

A random forest generates a large number of de-correlated trees for prediction. It minimizes tree correlation by including randomness in the tree-growing process. The randomization method is split-variable. Each tree split in a random forest has a smaller feature search space.

Multilayer perception is abbreviated as MLP. MLP is made up of a variety of multilayer neurons. The 1st layer is the input layer, the 2nd is the concealed layer, and the 3rd layer is the output layer. It accepts data from the input layer and sends it to the output layer for processing.

A decision tree is a model that is built on trees. Using splitting criteria, it divides the features into smaller sections with similar response values. The tree diagram is created using the divide-and-conquer approach. A limited amount of pre-processing is required for decision trees, and they can readily handle category features without it.

The logistic function employed in logistic regression is known as a sigmoid function. The real numbers are positioned between 0 and 1 on an S-shaped curve

The supervised machine learning algorithm K-nearest neighbors (kNN) is a simple one. The kNN algorithm can be used to address classification and regression problems. For classifying the unseen test data,t he training observation is memorized by the kNN algorithm.The kNN algorithm finds items that are similar in a nearby area.

Yoav Freund and Robert Schapire proposed ADA boosting, also known as adaptive boosting. It creates a classifier by combining many low-performing classifiers. In each cycle, it sets the weight of classifiers and trains the data.

GNB is a Naïve Bayes variation that allows continuous data and follows the Gaussian normal distribution. The Bayes theorem is the foundation for the Naïve Bayes family of supervised machine learning classification methods. It's a straightforward categorization method with a lot of punch. They're useful when the inputs have a lot of dimensions. Complex classification problems can also be solved with the Naïve Bayes Classifier.

The Support Vector Machine is a supervised machine learning technique. This can also be utilized to solve classification and regression issues. The values of the characteristics are given in the particular coordinate once the data items are arranged in n-dimensional space. It is termed hyperplane because it produces the most homogeneous points in each part. In sklearn's SVM classification algorithm, the default kernel is RBF.

The Linear Support Vector Classifier (SVC) classifies data using a linear kernel function and works well with big datasets. The Linear SVC model has more parameters than the SVC model, such as loss function and penalty normalization ('L1' or 'L2'). The kernel method cannot be changed because linear SVC is based on the kernel linear approach.

Gradient boosting is a machine learning method for solving classification and regression predictive modeling problems. Gradient boosting is also referred to as stochastic gradient boosting (a subset of gradient boosting), gradient tree boosting and gradient boosting machines (abbreviated as GBM).

After that, the performance of the algorithms has been analyzed comparatively. Based on the performance metrics the final decision has been taken.

3.5 Requirements for implementation

Information has been collected via online forms and handwritten forms. Microsoft Excel has been used to construct data sets. "Jupyter notebook" has been utilized to implement the algorithms.

Microsoft Excel is a spreadsheet program that allows users to format, calculate and organize data using formulae. This program is part of the Microsoft Office suite and can work with other Office programs.

Jupyter is a free, open-source, interactive web platform that allows academics to combine software code, computational output, explanatory prose, and multimedia materials in one document. Despite the fact that computational notebooks have been around for decades, Jupyter has grown in popularity in recent years. According to co-founder Fernando Pérez, the notebook's rapid adoption was aided by an enthusiastic community of user–developers and a redesigned architecture that allows it to speak dozens of programming languages — a fact reflected in its name, which was inspired by the programming languages Julia (Ju), Python (Py), and R.

CHAPTER 4

RESULTS OF EXPERIMENTS AND REVIEW

4.1 Introduction

The dataset and dataset processing techniques were covered in the preceding section. Some algorithms use the processed data, and the algorithm's results will be explained in this section. kNN, logistic regression, SVM with RBF kernel, GNB, decision tree, multilayer perception (MLP), random forest, gradient boosting, Linear SVC and ADA boosting have been used and the results are assessed to see which approach delivers the best accuracy:

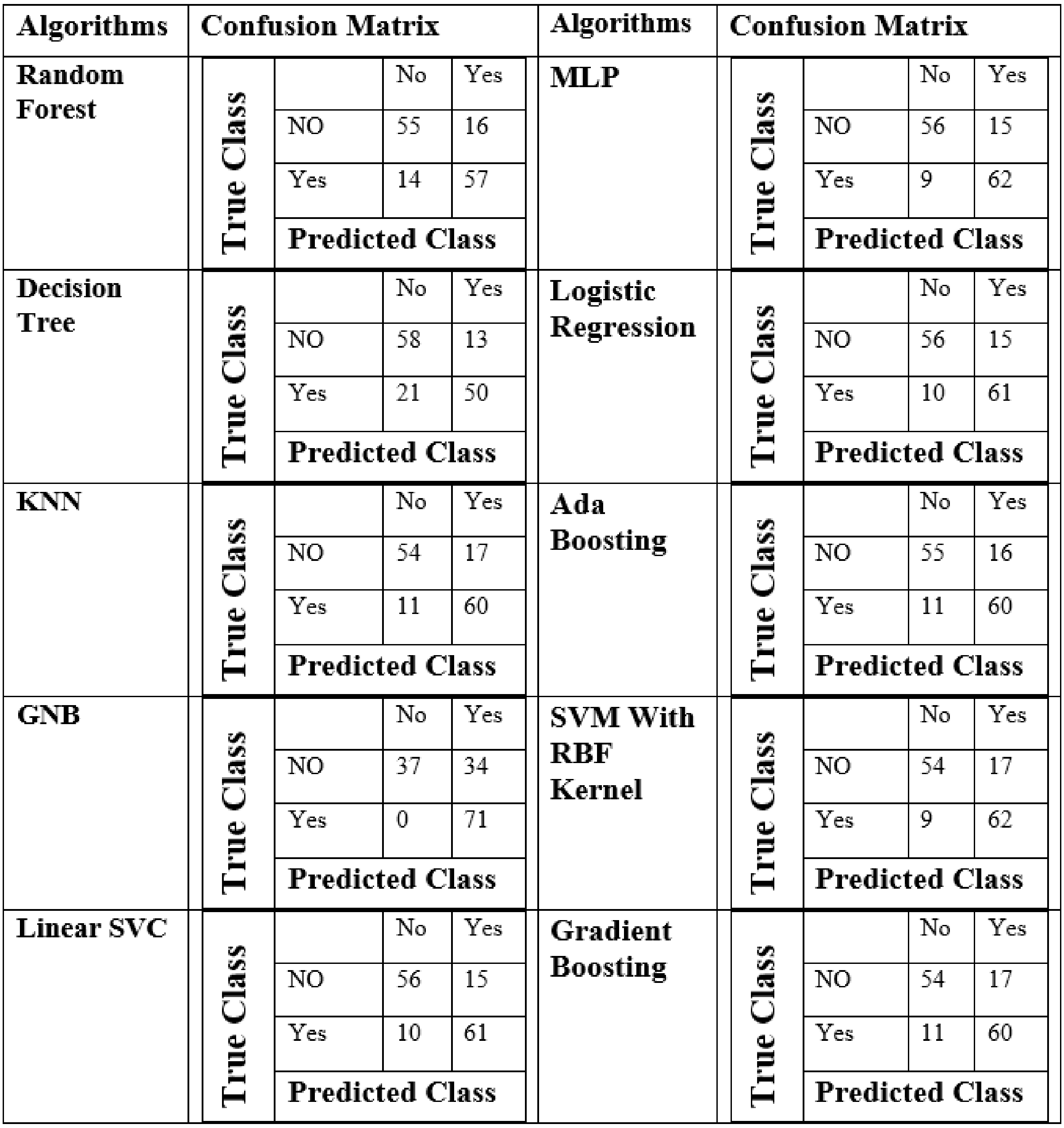
4.2 Analyzing the Results of Experiments

Ten machine learning algorithms has been employed and compared these algorithms by assessing their accuracy.

4.2.1 Descriptive Analysis

The confusion matrix is one of the most essential performance assessment approaches for machine learning classification. It will apply the classification models to the test data and produce a tabular output of true positive, false positive, true negative and false negative values. The Confusion Matrix is essential for assessing the performance of any classifier.

The confusion matrix of all methods utilized in our model is shown in Table 4.1. In the following table, the model evaluation of each classifier is described with a value.



Algorithms	Confusion Matrix				Algorithms	Confusion Matrix			
Random Forest	True Class		No	Yes	MLP	True Class		No	Yes
		NO	55	16			NO	56	15
		Yes	14	57			Yes	9	62
		Predicted Class					Predicted Class		
Decision Tree	True Class		No	Yes	Logistic Regression	True Class		No	Yes
		NO	58	13			NO	56	15
		Yes	21	50			Yes	10	61
		Predicted Class					Predicted Class		
KNN	True Class		No	Yes	Ada Boosting	True Class		No	Yes
		NO	54	17			NO	55	16
		Yes	11	60			Yes	11	60
		Predicted Class					Predicted Class		
GNB	True Class		No	Yes	SVM With RBF Kernel	True Class		No	Yes
		NO	37	34			NO	54	17
		Yes	0	71			Yes	9	62
		Predicted Class					Predicted Class		
Linear SVC	True Class		No	Yes	Gradient Boosting	True Class		No	Yes
		NO	56	15			NO	54	17
		Yes	10	61			Yes	11	60
		Predicted Class					Predicted Class		

Table 4.1: All classifier’s confusion matrix.

We also calculated specificity, sensitivity, precision, f-score, confusion matrix and roc curve. Any model selection requires evaluation of that model. Certain classifiers must be measured when it comes to model evolution. For better measurement, classifications are assessed using the test data set.

The percentage of correct predictions for the test data is known as accuracy. It's simple to figure out by dividing the number of correct guesses by the total number of predictions.

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \times 100\%$$

The true positive rate is called sensitivity. Sensitivity is defined as the ratio of accurately diagnosed positive tuples to the total number of positive tuples.

$$Sensitivity = \frac{TP}{\mathrm{TP+FN}} \times 100\%$$

The true negative rate is known as specificity. That is, specificity is the ratio of correctly diagnosed negative tuples to the total number of negative tuples.

$$Specificity \ = \frac{TN}{\mathrm{TN+FP}} \times 100\%$$

The term "precision" refers to the measuring of exactness. The ratio between the genuine positive value and the expected positive value.

$$Precision = \frac{TP}{\mathrm{TP+FP}} \times 100\%$$

The harmonic mean of recall and precision is calculated with the F1 score. In order to calculate, it takes into account both false positive and false negative numbers.

$$F1\ Score = \frac{\mathrm{2TP}}{\mathrm{2TP+FP+FN}} \times 100\%$$

The performance of each algorithm is described in Table 4.2. On the basis of these algorithm performances and their correctness, method will be picked which is best for our model. It is clear that MLP performs the best based on this accuracy. MLP performs better in terms of sensitivity, specificity, F1-Score and precision. Taking everything into analyzing, this algorithm can be used to achieve the optimum model performance.

Algorithms	**Accuracy**	**Precision**	**F1 Score**	**Sensitivity**	**Specificity**
Random Forest	78.87 %	78.08%	79.17%	80.28%	77.46%
MLP	83.10 %	80.52%	83.78%	87.32%	78.87%
Decision Tree	76.05 %	79.37%	74.62%	70.42%	81.69%
Logistic Regression	82.39 %	80.26 %	82.99%	85.91%	78.87%
KNN	80.28 %	77.92 %	81.08%	84.51%	76.06%
ADA Boosting	80.99 %	78.94 %	81.63%	84.51%	77.46%
GNB	76.05 %	67.62 %	80.68%	100%	52.11%
SVM with RBF Kernel	81.69 %	78.48 %	82.67%	87.32%	76.06%
Linear SVC	82.39 %	80.26 %	82.99%	85.92%	78.87%
Gradient Boosting	80.28 %	77.92 %	81.08%	84.51%	76.06%

Table 4.2: Classifier performance evaluation

From the table we can see that MLP shows highest accuracy 83.10%. In Sensitivity GNB Shows the highest value 100%, in Specificity Decision Tree shows highest value 81.69%, in Precision MLP shows the highest value 80.52% and finally in F1 Score MLP shows the highest value 83.78%. So we can say that MLP is better for our study.

Here in the graph we can see that

On processed datasets with a total of 20 features, we ran ten machine-learning algorithms. Random Forest appears to have achieved 78.87% accuracy, MLP 83.10% accuracy, decision tree 76.05 % accuracy, logistic regression 82.39% accuracy, kNN 80.28% accuracy, ADA

Boosting 80.99% accuracy, GNB 76.05% accuracy, SVM with RBF kernel 81.69% accuracy, Linear SVC 82.39 accuracy and gradient boosting 80.28% accuracy.

The accuracy of ten algorithms is depicted in Figure 4.1.

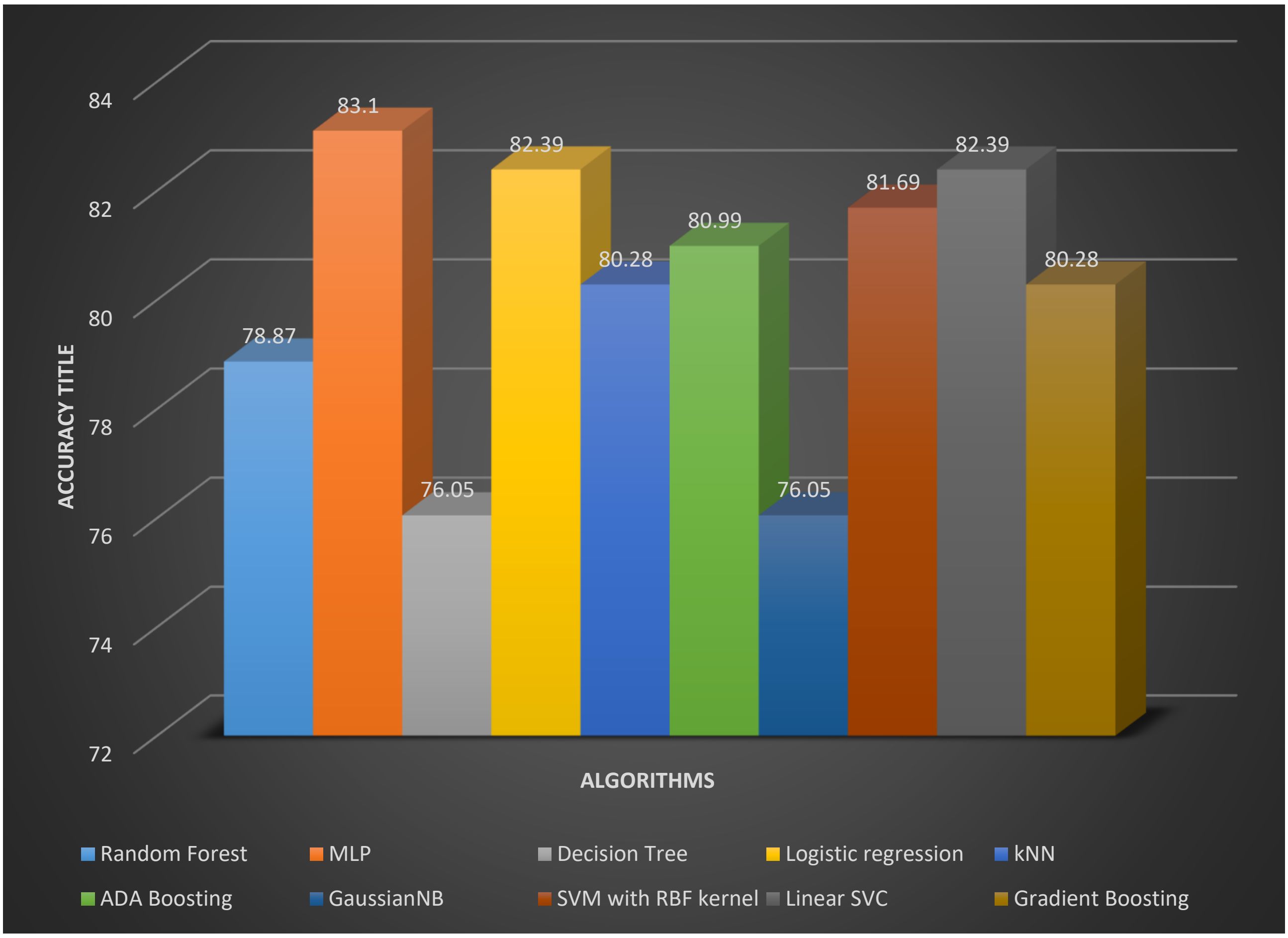


Fig 4.1: Accuracy of ten algorithms

Receiver operating characteristics (roc) curves are a great way to compare categorization models visually. A ROC curve is made up of a false-positive and a true-positive rate. The diagonal line reflects a random estimate. Fig (4.2 – 4.11) shows the ROC-Curves of the each proposed algorithms.

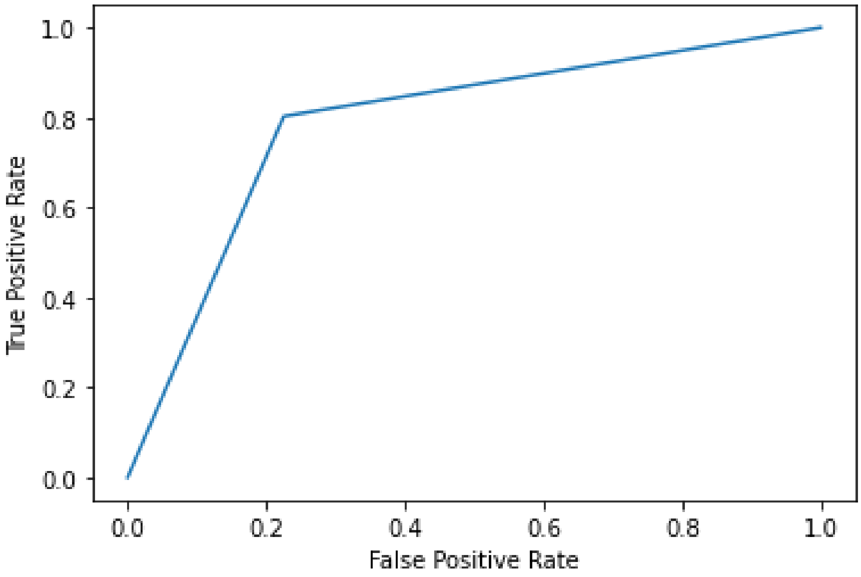


Fig 4.2: ROC-Curve (Random Forest)

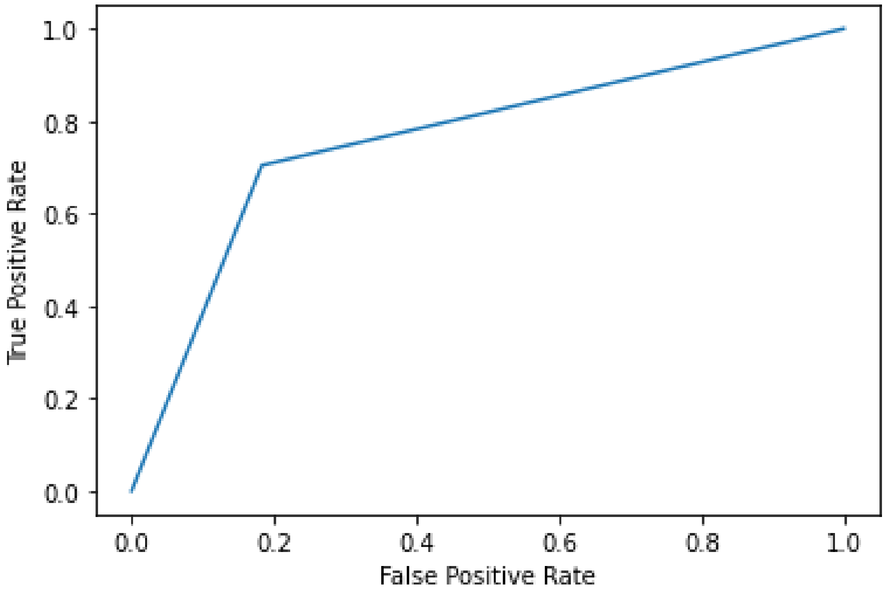


Fig 4.3: ROC-Curve (Decision Tree)

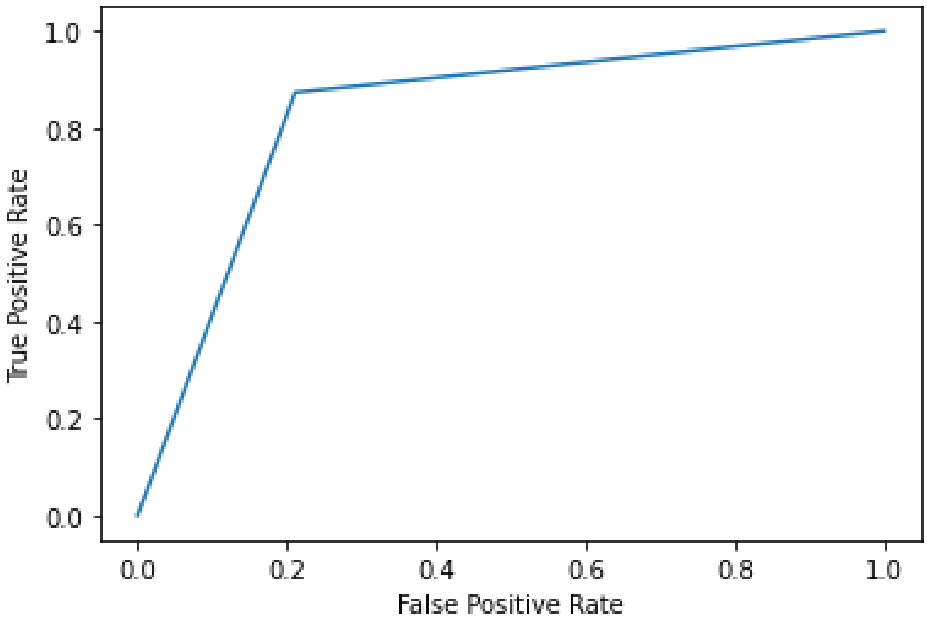


Fig 4.4: ROC Curve (MLP)

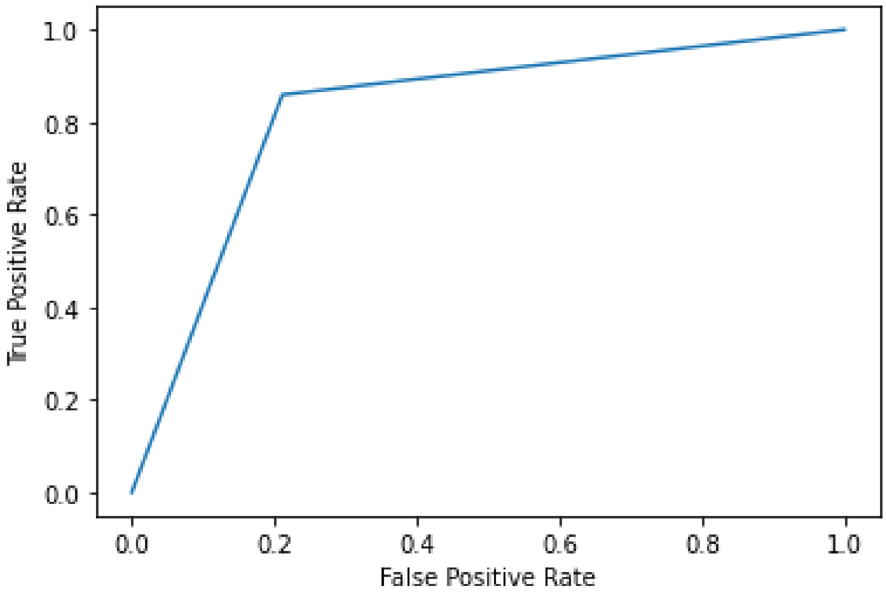


Fig 4.5: ROC-Curve (Logistic Regression)

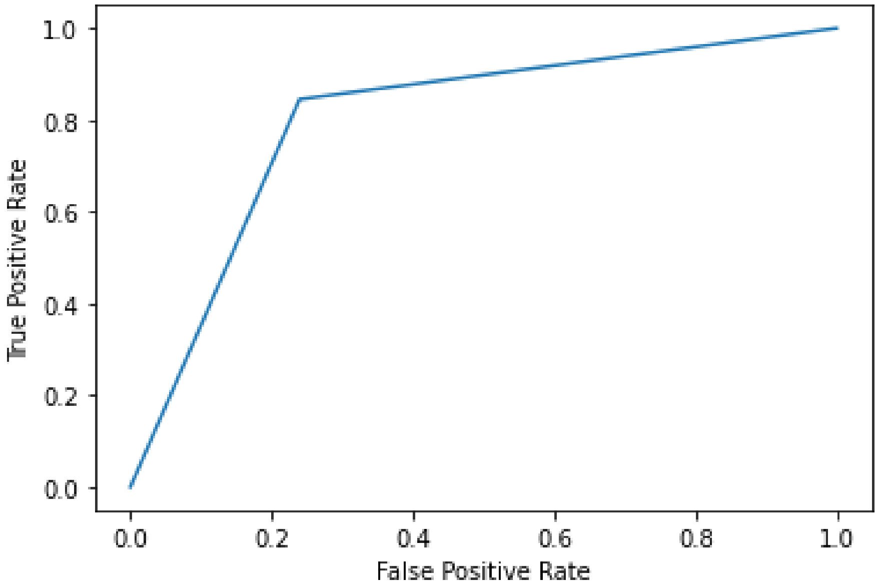


Fig 4.6: ROC-Curve (KNN)

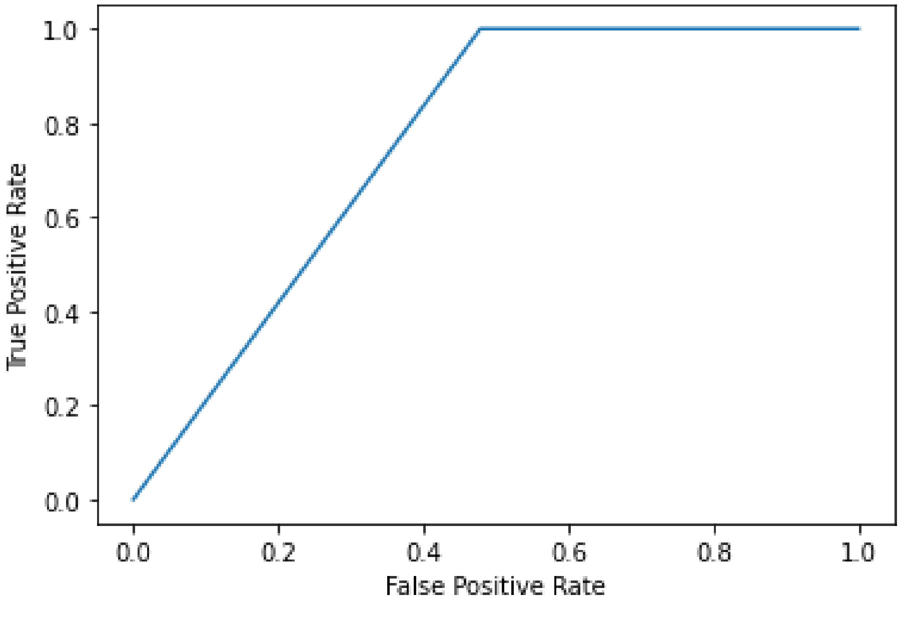


Fig 4.7: ROC-Curve (GNB)

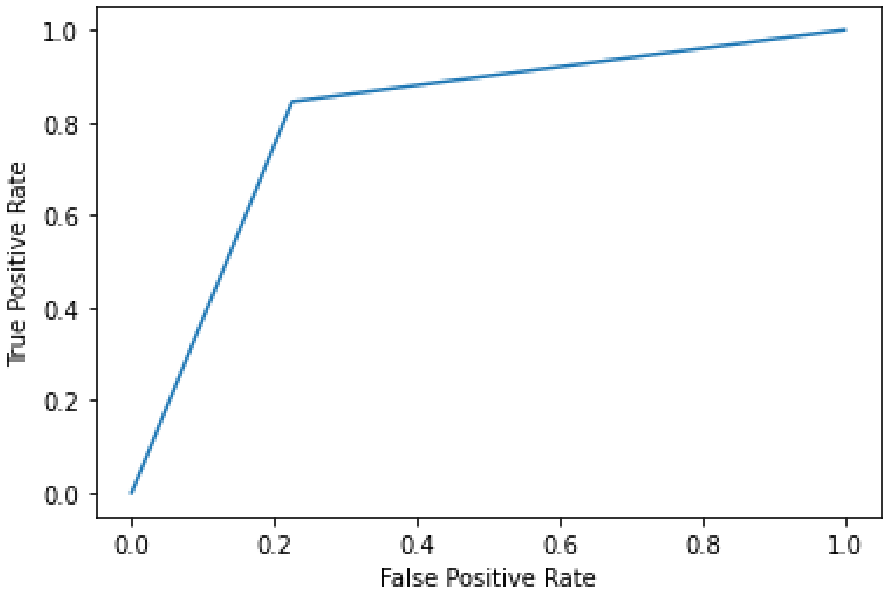


Fig 4.8: ROC-Curve (ADA Boosting)

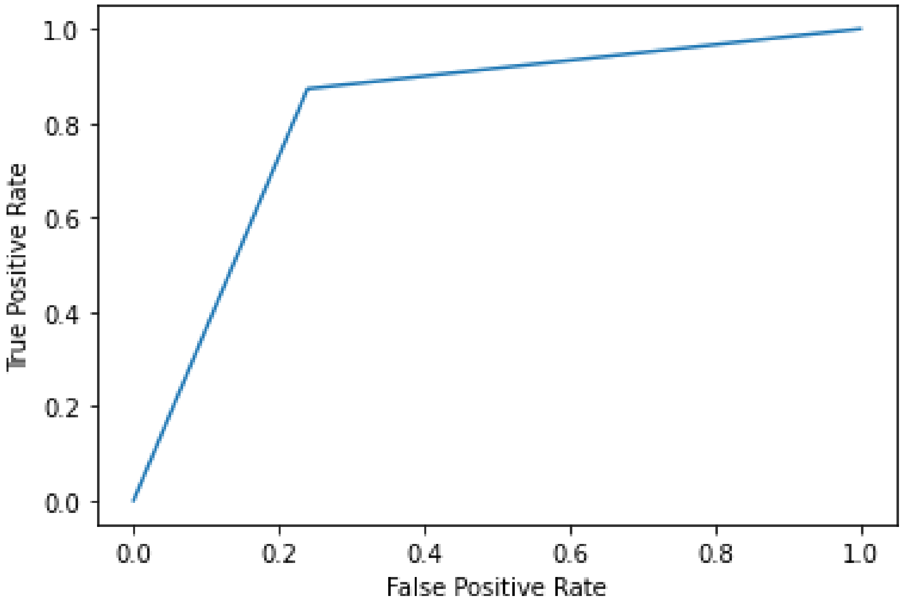


Fig 4.9: ROC Curve (SVM with RBF Kernel)

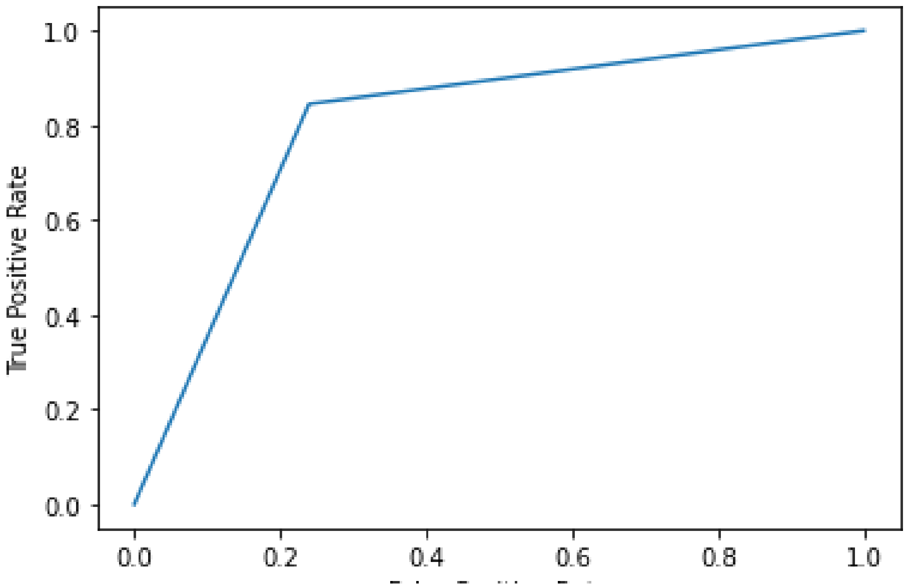


Fig 4.10: ROC Curve (Gradient Boosting)

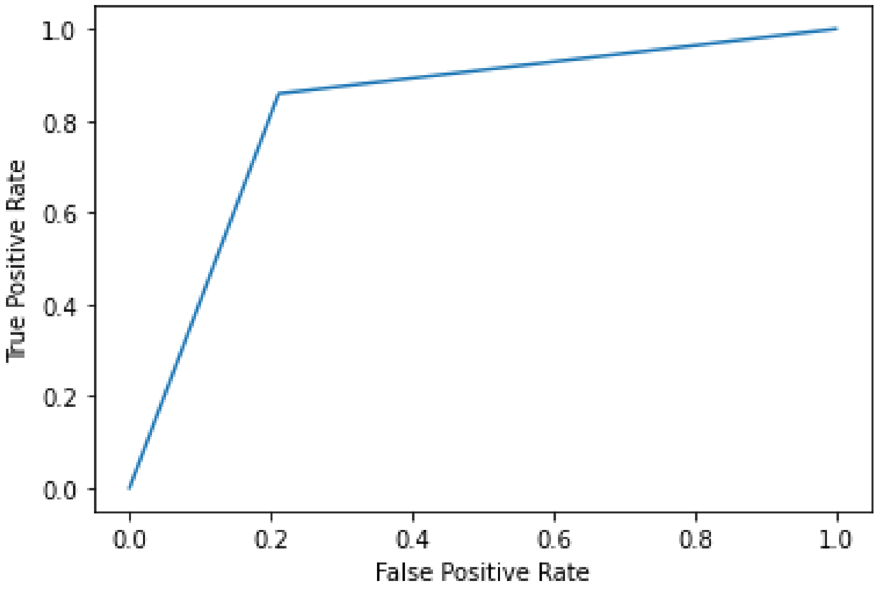


Fig 4.11: ROC Curve (Linear SVC)

4.3 Final Decision

From the above discussion we came to know that MLP shows highest accuracy 83.10%. In Sensitivity GNB Shows the highest value 100%, in Specificity Decision Tree shows highest value 81.69%, in Precision MLP shows the highest value 80.52% and finally in F1 Score MLP shows the highest value 83.78%. So we can say that MLP is better for our study. With specificity, sensitivity, F1-Score, precision and accuracy we can observe that MLP produce the best results. Finally, we discover that our Motorbike accident severity prediction model performs best when we approach MLP.

CHAPTER 5

SOCIETAL, ENVIRONMENTAL IMPACT AND SUSTAINABLE PLANNING

5.1 Societal Impact

Motorbike accident is one of the life-threating accident on the road. Bike accident is harmful for family, society and public, government property. For bike accident many public property damages. There are some reasons or facts which biker does to show their style and skill. They think that those thing show his personality and skills of drive a bike. For example: Biker thing over-speeding shows their skill and style. But most of the fatal severity happens for over-speeding. If there is one biker in the society with this type of thoughts, other bikers will be influenced by him. We think our model will be helpful for all the bikers. They can easily understood about the facts of accident. So, that they will try to avoid those facts. Parents, friends, relatives could consult them about bike accidents. So, motorbike accident severity prediction model will be helpful for society.

5.2 Aspects of morality

This accident severity prediction model is not unethical or infringes on human rights in any manner. There will be no privacy issues because the model does not collect any personal information, such as name, identification, or location. This model does not infringe on a person's right to enjoy or use, but it does play a part in raising awareness. The model for predicting the severity of motorcycle accidents was designed with all types of restrictions in mind, as well as privacy and confidentiality concerns. As a result, the model of motorbike accident severity prediction may be managed without difficulty utilizing machine-learning technology.

5.3 Sustainable Planning

The community, financial, and organizational aspects of the sustainability strategy. The Sustainability Plan provides us with a realistic picture of how a project will run and what the project's future plans will be. The goal of our model cohort is to figure out how serious a motorcycle accident is. This model must be tailored to make it simple for people to adapt, and it is critical to remember that using this model does not imply that people are inferior.

CHAPTER 6

SUMMARY,CONCLUSION, RECOMMENDATION AND FUTURE RESEARCH IMPLICATION

6.1 Summary of research

Data gathering, data preparation, methodology implementation, and experimental evaluation are all elements of our project. We gathered the required information using a Google form and a printed form. We only collected information from those who had been in a bike accident. Following data gathering, we process the data and use Jupyter Notebook to work on data processing and implementation. After preprocessing, we execute ten machine-learning algorithms, including random forest, MLP, decision tree, logistic regression, kNN, ADA Boosting, GNB, linear SVC, SVM with RBF Kernel and Gradient boosting, and evaluate their accuracy. It is clear that the MLP technique performs the best.

6.2 Limitations and Conclusions

Machine learning procedures are used in our research to anticipate motorbike accident severity. In our work and model, we have various restrictions and flaws. The dataset has been used was not particularly vast; a larger and more diverse data set would have been preferable. People from various professions, districts, and social groups were unable to collect data due to certain limitations. For data processing, a variety of advanced approaches may be utilized, and the model might be presented in a beautiful way using multiple variants in algorithm application. It is possible to determine the severity of a motorcycle collision using our proposed approach.

6.3 **Future Research Implication**

Technology and contemporary science have made our lives easier and faster in recent years. In the future, we aim to implement our model in a web application, Android application or software to continue the use of the internet and information technology in our country. We will be able to improve the accuracy of our model in the future by employing a larger dataset. Furthermore, the model's software can be made accessible to the public by constructing user-friendly GUIs. The model can be made more effective in the future by implementing new algorithms, adding more features and introducing alternative parameters A robust database can be built in the future by gathering data from various types of people according to the district.

APPENDICES

Abbreviation

MLP= Multilayer perception

kNN= k-nearest neighbors

SVM= Support Vector Machine

SVC= Support Vector classifier

RBF= Radial basis function

GNB= GNB

ROC= Receiver Operating Characteristic

Appendix: Research Reflections

When we began our research effort, we had no prior knowledge of artificial intelligence or machine learning detection and prediction. Our supervisor was quite pleasant and genuine. He provided us with invaluable advice and was quite helpful. We learnt numerous new techniques, new knowledge, how to employ algorithms, and how to work with various methodologies during the course of our research. The Jupyter notebook and Python programming language were both new to me. Working with them presented some challenges at first, but as we became more familiar with Jupyter notebook and Python, we were able to overcome them.

Finally, we acquired bravery and were encouraged to do more in the future as a result of our studies.

REFERENCES

1. Sarkar, S., Vinay, S., Raj, R., Maiti, J. and Mitra, P., 2019. Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research*, *106*, pp.210-224.

2. Iranitalab, A. and Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, pp.27-36.

3. Lingtao, W., Manjuan, Y., Chengcheng, T., Daoyue, P., Guohua, S., Tiejun, Z. and Zhanyang, G., 2010, November. Research on influence extention of two-lane highway intersections based on traffic accident database. In *2010 International Conference on Optoelectronics and Image Processing* (Vol. 2, pp. 244-246). IEEE.

4. Lv, Y., Tang, S. and Zhao, H., 2009, April. Real-time highway traffic accident prediction based on the k-nearest neighbor method. In *2009 international conference on measuring technology and mechatronics automation* (Vol. 3, pp. 547-550). IEEE.

5. Yu, R. and Liu, X., 2010, December. Study on traffic accidents prediction model based on RBF neural network. In *2010 2nd International Conference on Information Engineering and Computer Science* (pp. 1-4). IEEE.

6. Chong, M., Abraham, A. and Paprzycki, M., 2005. Traffic accident analysis using machine learning paradigms. *Informatica*, *29*(1).

7. Satu, M.S., Ahamed, S., Hossain, F., Akter, T. and Farid, D.M., 2017, December. Mining traffic accident data of N5 national highway in Bangladesh employing decision trees. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 722-725). IEEE.

8. Kumar, S. and Toshniwal, D., 2016. A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, *24*(1), pp.62-72.

9. Bülbül, H.İ., Kaya, T. and Tulgar, Y., 2016, December. Analysis for status of the road accident occurance and determination of the risk of accident by machine learning in istanbul. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 426-430). IEEE.

10. Elahi, M.M.L., Yasir, R., Syrus, M.A., Nine, M.S.Z., Hossain, I. and Ahmed, N., 2014, May. Computer vision based road traffic accident and anomaly detection in the context of Bangladesh. In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)* (pp. 1-6). IEEE.

11. Nandurge, P.A. and Dharwadkar, N.V., 2017, February. Analyzing road accident data using machine learning paradigms. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 604-610). IEEE.

12. Esmaeili, A., Khalili, M. and Pakgohar, A., 2012, September. Determining the road defects impact on accident severity; based on vehicle situation after accident, an approach of logistic regression. In *2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)* (pp. 1-4). IEEE.

13. Beshah, T. and Hill, S., 2010, March. Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia. In *2010 AAAI Spring symposium series*.

14. Mohanta, B.K., Jena, D., Mohapatra, N., Ramasubbareddy, S. and Rawal, B.S., 2021. Machine learning based accident prediction in secure iot enable transportation system. *Journal of Intelligent & Fuzzy Systems*, (Preprint), pp.1-13.

15. Ramani RG, Shanthi S (2012) Classifier prediction evaluation in modeling road traffic accident data. In: 2012 IEEE international conference on computational intelligence and computing research (ICCIC), pp 1–4

16. Theofilatos, A., Chen, C. and Antoniou, C., 2019. Comparing machine learning and deep learning methods for real-time crash prediction. *Transportation research record*, *2673*(8), pp.169-178.

17. Abdel-Aty, M. and Haleem, K., 2011. Analyzing angle crashes at unsignalized intersections using machine learning techniques. *Accident Analysis & Prevention*, *43*(1), pp.461-470.

18. Labib, M.F., Rifat, A.S., Hossain, M.M., Das, A.K. and Nawrine, F., 2019, June. Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1-5). IEEE.

19. Matías, J.M., Rivas, T., Martín, J.E. and Taboada, J., 2008. A machine learning methodology for the analysis of workplace accidents. *International Journal of Computer Mathematics*, *85*(3-4), pp.559-578.

20. Wahab, L. and Jiang, H., 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PLoS one*, *14*(4), p.e0214966.

21. Zhang, J., Li, Z., Pu, Z. and Xu, C., 2018. Comparing prediction performance for crash injury severity among various machine learning and statistical methods. *IEEE Access*, *6*, pp.60079-60087.

22. Kim, J.H., Kim, J., Lee, G. and Park, J., 2021. Machine Learning-Based Models for Accident Prediction at a Korean Container Port. *Sustainability, 13*(16), p.9137.

23. Komol, M.M.R., Hasan, M.M., Elhenawy, M., Yasmin, S., Masoud, M. and Rakotonirainy, A., 2021. Crash severity analysis of vulnerable road users using machine learning. *PLoS one*, *16*(8), p.e0255828.

24. AlMamlook, R.E., Kwayu, K.M., Alkasisbeh, M.R. and Frefer, A.A., 2019, April. Comparison of machine learning algorithms for predicting traffic accident severity. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (pp. 272-276). IEEE.