UNIVERSITY OF BAHRAIN

College of Science



Department of Post Graduate Studies

Machine Learning Approach to Traffic Accidents Analysis in the Kingdom of Bahrain.

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of M.Sc. in Big Data Science and Analytics

Prepared by

Salman Bader Abdulla Alhaddad

ID Number: 20081603

Supervised by

Dr. Ahmed Mohammed Zeki

Assistant Professor

University of Bahrain

Kingdom of Bahrain

March 2023

UNIVERSITY OF BAHRAIN

College of Science

Department of Post Graduate Studies

Machine Learning Approach to Traffic Accidents Analysis in the Kingdom of Bahrain.

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of M.Sc. in Big Data Science and Analytics

Prepared by

Salman Bader Abdulla Alhaddad

ID Number: 20081603

Supervised by

Dr. Ahmed Mohammed Zeki

Assistant Professor

University of Bahrain

Kingdom of Bahrain

March 2023

Defense Committee's Approval

University of Bahrain
Deanship of Graduate Studies
And scientific Research

جامعة البحرين
عمادة الدراسات العليا والبحث العلمي

Thesis Discussion Committee Decision

Student Name: **SALMAN BADER ABDULLA ALHADDAD**	Student ID: **20081603**
Department: Choose an item.	College: **Science**
Program: **M.Sc. in Big Data Science and Analytics**	
Thesis Title: Machine Learning Approach to Traffic Accidents Analysis in the Kingdom of Bahrain	
Supervisor Name: **Dr. Ahmed M. Zeki**	Student GPA: **3.45**

The Discussion Committee met on 29-May-23 In S41-1002. Upon discussing the student whose details are stated above in public, the discussion committee recommends the following:

The student is awarded a grade Choose an item. in the thesis subject. Type text here

☐ Without making any changes to the thesis.

☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

Name		Signature
1. Dr. Ahmed M. Zeki	(Advisor)	
2. Dr. Riadh Ksantini	(Internal Examiner)	
3. Prof. Hani Bani-Salameah	(External Examiner)	

Approved at the University Council 6th meeting, 2013
In accordance with Decision No. 736/2013 dated 14/3/2013
Modified by University Council Decision No. 2043/2013 dated 20/11/2013.

Abstract

The Road Traffic Accidents is one of the main causes that contribute to increase the death rate worldwide. So, many countries including the Kingdom of Bahrain are cooperating to reduce accidents as much as possible. In recent years, the machine learning approach has been used noticeably in many fields such as medicine, economics, education, biology, and transportation. In this study, many classifiers, namely KNN, SVM, DT, RF, and NB have been investigated for their high performance in previous studies to predict the severity of RTAs based on several factors such as accident date, weather, and road type by choosing appropriate parameters for each model.

Five real datasets were collected by the statistic section at the general directorate of traffic in the Kingdom of Bahrain. Each dataset refers to a specific year from 2018 to 2022. In addition, After the preparation process with data cleaning, data structuring, and feature engineering, the datasets under investigation combined into 5191 injuries caused by RTAs.

The comparison between mentioned classifiers is done by using several evaluation methods which are accuracy, F-measure, recall, precision, and AUC. Interestingly, the classes of the target attribute were imbalanced. For that, the researcher used the imbalanced dataset with over-sampling and under-sampling methods to evaluate the performance of models in these cases.

The results of this study were satisfactory. The best classifier with all evaluation methods was achieved with the RF algorithm of 85% for all criteria by using the over-sampling method. Meanwhile, the worst classifier is the NB when the dataset used was imbalanced with a precision of 63%, recall of 10%, F-measure of 10%, and accuracy of 10%. Also, through building all models the class fatal has the highest AUC compared to other classes. It is observed that the over-sampling method outperformed other methods with most models. In addition, the most important features that affect the RTAs are cause type and accident day of week, the least important features with significant differences are road condition and road surface. The results of the research are promising and are expected to be applied by the Bahraini General Directorate of Traffic.

Dedicate

I dedicate this work to my mother, who always sacrificed for me and encouraged me to do my best, and to my deceased father, who supported me in life and made knowledge a valuable acquisition for me to rely on. Also, I dedicate this thesis to my beloved wife, who shares all the moments with me to build my bright future, and to my children Bader and Farah. In addition, this thesis is dedicated to my brothers who always stand by me and support me, and to my family and friends.

Acknowledgements

Firstly, I would like to sincerely thank my thesis supervisor assistant professor Dr. Ahmed Zeki for his continuous support and guidance for me when needed. Also, I extend my deepest gratitude to Dr. Sawsan Hilal the coordinator of the program for providing the necessary facilities during the study period of the Big Data Science and Analysis Program. I extend my greetings to all the distinguished professors in this program for providing the necessary knowledge and skills as this helped me to complete this work.

My sincere thanks to the Statistics Department at the Bahraini General Directorate Traffic for facilitating obtaining the data used in this study, and especially for Mr. Ahmed Meshaal for his cooperation.

In the end, my sincere thanks go to my family, friends, and everyone who helped me and stood by me until the completion of my master's thesis.

Salman Bader Alhaddad

Table of Contents

Defense Committee's Approval.....A

Abstract.....B

Dedicate.....C

Acknowledgements.....D

Table of Contents.....E

List of Tables.....J

List of Figures.....L

List of Abbreviations.....N

List of Appendices.....P

Chapter One

Introduction

1.0. Introduction.....2

1.1. Research Problem.....6

1.2. Research Questions.....6

1.3. Research Objectives.....7

1.4. Research Significance.....7

1.5. Research Setting.....7

1.6. Research Organization.....8

Chapter Two

Conceptual Framework and Literature Review

2.0. Introduction 9

2.1. Machine Learning 9

2.2. Types of Machine Learning 10

2.2.1. K-Nearest Neighbor (KNN) 10

2.2.2. Support Vector Machine (SVM) 11

2.2.3. Naïve Bayes (NB) 13

2.2.4. Random Forest (RF) 14

2.2.5. Decision Tree (DT) 16

2.2.6. Multilayer Perceptron (MLP) 16

2.2.7. K-means Algorithm 17

2.2.8. Summary of the Common ML Algorithms 18

2.3. Types of Machine Learning Summary 19

2.4. Feature Selections 26

2.5. Contributing Factors to Road Traffic Accidents 26

2.5.1. The Human Factor 26

2.5.2. The Vehicle Factor 26

2.5.3. The Environment Factors 27

2.6. Previous Studies on RTA Using Machine Learning Models 27

2.6.1. RTAs in the Worldwide 27

2.6.2. RTAs In the Arab Region and the GCC 29

2.6.3. RTAs in the Kingdom of Bahrain 30

2.7. Previous Studies Summary 32

2.8. Summary of the Number of Algorithms Used 37
2.9. Conclusion 38

Chapter Three

Research Methodology

3.0. Introduction.....40
3.1. Research Methodology 40
3.2. Data Description 42
3.2.1. Dataset One Summary (2018) 42
3.2.2. Dataset Two Summary (2019).....43
3.2.3. Dataset Three Summary (2020).....44
3.2.4. Dataset Four Summary (2021).....44
3.2.5. Dataset Five Summary (2022) 45
3.3. Description Original Datasets.....45
3.4. Data Combining.....48
3.5. Data Understanding 48
3.6. Data Preparation.....53
3.6.1. Data Cleaning..... 53
3.6.2. Data Structuring..... 54
3.6.3. Feature Engineering..... 55
3.7. Dataset Description After Handling 55
3.8. Data Splitting 61
3.9. Resampling Methods 61
3.10. ML Models Used 61
3.11. Evaluation Methods 62

3.11.1. Accuracy 62
3.11.2. Recall 62
3.11.3. Precision 63
3.11.4. F-measure 63
3.11.5. Area Under the Curve (AUC) 63
3.12. Conclusion 63

Chapter Four

Results and Discussion

4.0. Introduction 66
4.1. Tools Used 66
4.2. K-Nearest Neighbor (KNN) Model 66
4.2.1. Best Value of K 67
4.2.2. KNN with Resampling Methods 67
4.3. Random Forest (RF) Model 70
4.3.1. Parameters of RF 70
4.3.2. RF with Resampling Methods 71
4.4. Naïve Bayes (NB) Model 73
4.4.1. Parameters of NB 74
4.4.2. NB with Resampling Methods 74
4.5. Support Vector Machine (SVM) Model 77
4.5.1. Parameters of SVM 77
4.5.2. SVM with Resampling Methods 78
4.6. Decision Tree (DT) Model 81

4.6.1. Parameters of DT 81

4.6.2. DT with Resampling Methods 82

4.7. Best Models Comparison 84

4.8. Feature Selections 87

4.9. Conclusion 89

Chapter Five

Conclusion and Recommendations

5.0. Introduction 90

5.1. Conclusion 90

5.2. Limitations 91

5.3. Recommendations 91

5.4. Future Work 92

References 93

Appendices 99

الملخص 100

List of Tables

Table (2.1): Summary of Common Machine Learning Algorithms in the Field of RTAs 19
Table (2.2): Summary for Studies About RTAs 32
Table (2.3): The Number of Times the Algorithms were Used in the Identified Previous Studies 37
Table (3.1): Details of Missing Values in Dataset One 43
Table (3.2): Details of Missing Values Related to Dataset Two 43
Table (3.3): Missing Values of Dataset Three 44
Table (3.4): Missing Values of Dataset Four 45
Table (3.5): Missing Values of Dataset Five 45
Table (3.6): Describe Datasets One and Two 46
Table (3.7): Describe Dataset Three, Four, and Five 47
Table (3.8): Details of Missing Values Relative to the Whole Datasets after Combining 54
Table (3.9): Description Dataset after Preparation Process 56
Table (4.1): The Performance of KNN Algorithm by using Different Resampling Methods 68
Table (4.2): The Performance of RF Algorithm by using Different Resampling Methods 71
Table (4.3): The Performance of NB Algorithm by using Different Resampling Methods 75
Table (4.4): The Performance of SVM Algorithm by using Different Resampling Methods 79
Table (4.5): The Performance of DT Algorithm by using Different Resampling Methods 82
Table (4.6): Comparison Between All Best Models of Applied Algorithms 85

Table (4.7): Comparison Between the Best Results of the Algorithm for the Values of the AUC.....86

List of Figures

Figure (1.1): The Main Causes of Death that Expected in 2030 3
Figure (2.1): Explain the Way to Find the Value of K in the Model of KNN 11
Figure (2.2): Hyperplane Concept Related to SVM Algorithm 12
Figure (2.3): An Illustrated Diagram of the Random Forest Algorithm ..15
Figure (3.1): Flowchart of the Research Methodology 41
Figure (3.2): Relation Between Months and the Number of Injuries.....49
Figure (3.3): The Percentage of Injuries for each Year From 2018 to 202249
Figure (3.4): The Number of Injuries During Day of Week 50
Figure (3.5): The Difference in Percentage of Road Traffic Injuries Between Gender of Male and Female 50
Figure (3.6): The Relation Between Number of Injuries and the Governorates at the Kingdom of Bahrain.....51
Figure (3.7): The Difference Between the Types of Surface Road.....52
Figure (3.8): The Variate Between Variables of Accident Junction 52
Figure (3.9): The Comparing Between Number of Injuries with Different Types of Visibility 53
Figure (3.10): The Relation Between Accident Hour and the Number of Injuries 59
Figure (3.11): The Age Attribute in The RTAs Dataset.....59
Figure (3.12): Road Speed Attribute of the Dataset Under Investigation 60
Figure (3.13): The Variables of Injury Type as Target Attribute in RTAs Dataset.....60
Figure (4.1): Confusion Matrix for the KNN Models of the RTAs Classes69
Figure (4.2): AUC for the Models of KNN Related to the RTAs Classes70
Figure (4.3): Confusion Matrix for the RF Models of the RTAs Classes 72
Figure (4.4): AUC for the Models of RF Related to the RTAs Classes...73

Figure (4.5): Confusion Matrix for the NB Models of the RTAs Classes76
Figure (4.6): AUC for the Models of NB Related to the RTAs Classes ..77
Figure (4.7): Confusion Matrix for the SVM Models of the RTAs Classes80
Figure (4.8): AUC Values for the SVM Models Related to the RTAs Classes.....81
Figure (4.9): Confusion Matrix for the DT Models of the RTAs Classes83
Figure (4.10): AUC Values for the DT Models Related to the RTAs Classes 84
Figure (4.11): Comparison Between Best Accuracies for all Models of Applied Algorithms 86
Figure (4.12): Most Important Features Effect on the Road Accidents in the Kingdom of Bahrain 88

List of Abbreviations

RTAs	Road Traffic Accidents
RoSPA	The Royal Society for the Prevention of Accidents
UN	United Nations
SDGs	Sustainable Development Goals
WHO	World Health Organization
GCC	Gulf Cooperation Council
GDT	General Directorate of Traffic
RNN	Recurrent Neural Network
MLP	Multilayer Perceptron
SVM	Support Vector Machine
DT	Decision Tree
ML	Machine Learning
KNN	K-Nearest Neighbor
RPBL	Random Parameter Binary Logit
NB	Naïve Bayes
AUC	Area Under the Curve
ROC	Receiver Operating Characteristics
RF	Random Forest
MSE	Mean Squared Error
RMSE	Root Mean Square Error
ANN	Artificial Neural Network
LR	Logistic Regression
NN	Neural Network
LCC	Latent Class Clustering
CRISP-DM	Cross Industry Standard Process for Data Mining
TNR	True Negative Rate

TPR	True Positive Rate
FNR	False negative Rate
FPR	False Positive Rate
LSTM	Long Short-Term Memory

List of Appendices

Appendix (1) An example of the thesis's codes.

Chapter One

Introduction to the Study

1.0. Introduction 2
1.1. Research Problem 6
1.2. Research Questions 6
1.3. Research Objectives 7
1.4. Research Significance 7
1.5. Research Setting 7
1.6. Research Organization 8

Chapter One

Introduction

1.0. Introduction

Because people and society are affected by accidents on the road, several entities and organizations have tried to create a definition that explains and describes the meaning of Road Traffic Accidents (RTAs). The Royal Society for the Prevention of Accidents (RoSPA) has defined RTAs as an accident is a random event involving more than one factor, and such an accident is always preceded by a situation in which at least one user fails to deal with the environment around the road. Also, Agbeboh and Osabuohien explained that RTA is an unplanned event that occurs in an auto crash that may cause injuries and loss of life and public or private property (Micheale, 2017). Moreover, RTA can be explained as an accident happen if at least one vehicle is involved in this accident on public road, and at least one person is injured or killed (Bokaba et al., 2022).

Around the world, the high rate of RTA is attributed to the increased use of land vehicles among urban residents (Touahmia, 2018). Therefore, it is necessary to know that RTAs are associated with high rates of death, serious injuries, and significant economic losses (Parvareh et al., 2018). In addition, many countries consider RTAs as public health issues that need to be studied to find out their causes and consequences to reduce it as possible (Hamadeh, & Ali, 2013). For this reason, in 2015 when the United Nations (UN) issued 17 goals and 169 targets known as the Sustainable Development Goals (SDGs), the target of 3.6 stated halve the number of injuries and deaths caused by road accidents by 2020. Furthermore, in

August of the year 2020, UN adopted Resolution A/RES/74/299, which considered the period between 2021 and 2030 to be the second decade of action for road safety to reduce injuries and deaths by at least 50 percent (ESCAP, 2020).

According to World Health Organization (WHO) that the RTAs are the main cause of death for people between the ages of 5 and 44 which effects on people's health and the economy (Woyessa et al., 2021). Figure (1.1) shows the main expected causes of death during 2030 according to the WHO. In 2004, RTAs were the eighth cause of all deaths globally; by 2030, RTAs are expected to rank fifth among the largest contributors to global deaths (Butt et al., 2020). Furthermore, every year about 1.3 million people die on the road and about 50 million more people are injured, and Africa records the highest annual death rate in the world because of road accidents (Abegaz, & Gebremedhin, 2019).

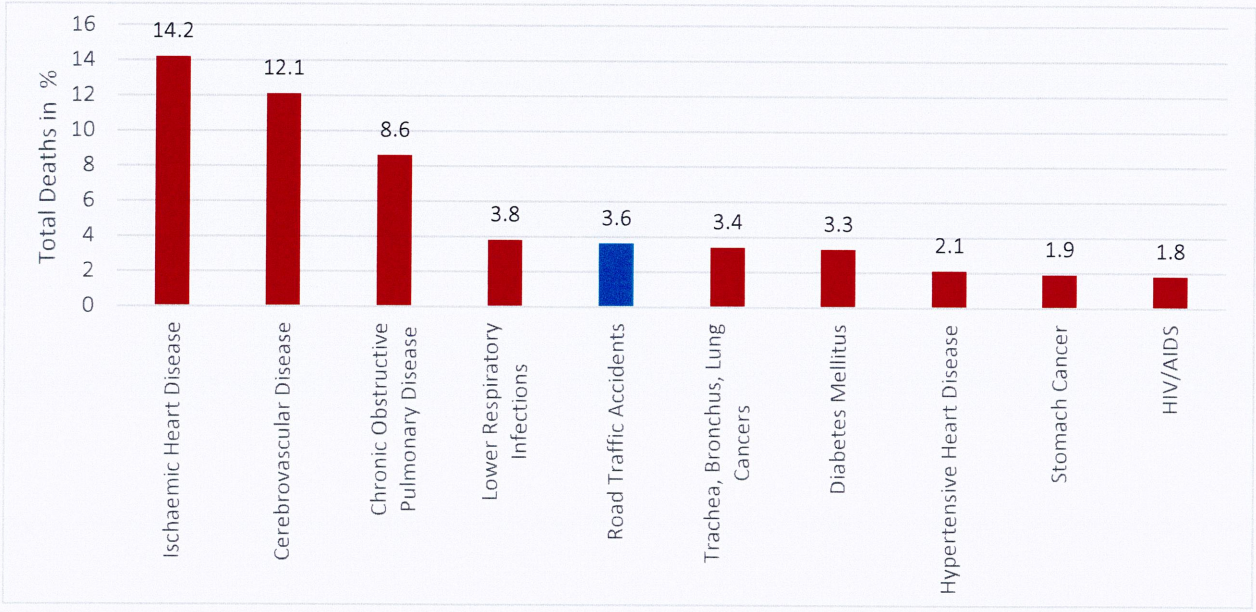


Figure (1.1): The Main Causes of Death that Expected in 2030 (Butt et al., 2020)

It is worth noting, the death rate due to RTAs in low-income countries is rising while it has been decreasing continuously in high-income countries. Moreover, around 90% of deaths resulting from RTAs occurred in low-income countries (Parvareh et al., 2018). In 2015, according to the report issued by the WHO, the number of deaths resulting from RTAs varies between countries of the world, low-income countries record 24.1 deaths per 100,000 people, while high-income countries recorded 9.2 deaths per 100,000 people (Bokaba et al., 2022).

But in the Middle East, high-income countries score differently due to traffic-related deaths rising extraordinarily. In this regard, the main factors for the high rate of road traffic deaths for the high-income countries of The Gulf Cooperation Council (GCC) countries are driving too fast, needing to use a seat belt, need paying attention, and improper vehicle maneuvering (Gasana et al., 2022). These reasons contribute to a significant increase in road and fatality rates compared to Western countries. Although, statistics covering 15 European countries have shown that the rate of road traffic deaths has decreased from 13.5 deaths per 100,000 people in the eighties to 5.5 deaths per 100,000 people today, while the death rate from road traffic accidents in the GCC region remained almost constant at 23 deaths per 100,000 people during the same period. (Butt et al., 2020).

In the Kingdom of Bahrain one of the GCC countries, RTAs are the eighth cause of death, and many strategies and legislation have been developed to ensure road safety including determining the speed limit for each road, the obligatory wearing of motorcycle helmets, wearing seat belts, and setting regulations for driving under the influence of alcohol. In general, public compliance with these laws is poor. Moreover, the two

leading causes of death due to RTA in Bahrain are driving too fast, and while intoxicated (Hamadeh, & Ali, 2013).

In the year 2020, the General Directorate of Traffic (GDT) in the Kingdom of Bahrain recorded 613 road injuries, including 53 deaths because of RTAs. The capital governorate recorded the largest number of deaths and injuries reached 21 and 222 respectively. In general, Bahrainis constituted 65.4% of the drivers who made the mistake, and 13.5% of injured pedestrians are under 10 years old. In addition, among the years 2011 and 2020, the years 2019 and 2020 recorded the lowest number of RTAs (General Directorate of Traffic, 2020).

Predicting RTAs can improve road safety for users. The process of collecting various data contributes to identifying the most important factors affecting the occurrence of traffic accidents on the road. Many studies discussed the topic of predicting the most important influencing factors that lead to RTAs. Some focused-on vehicle factors, others preferred to study human factors as the most important, while some studies considered roads and environmental factors. On the other hand, to predict risk, many researchers collect all these factors (Alrajhi, & Kamel, 2019; Bokaba et al., 2022; Yan, & Shen, 2022).

In recent years, the Machine Learning (ML) approach used in many fields related to traffic accidents to avoid it as possible; some focuses on predicting the injury severity of RTAs, finding the impact of drinking on injury severity, and identification of accident-prone areas on roads. All these are done by using different algorithms such as Recurrent Neural Network (RNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Decision Tree (DT) (Assi et al., 2020; Megnidio-Tchoukouegno, & Adedeji, 2023).

ML techniques are used in road accidents to decide the important influences linked to accident severity. Also, these techniques could show how each factor influences RTAs from each type of vehicle in proportion. In addition, ML techniques could be used to identify the most critical factors that might lead to accidents, detect accident characteristics, and the influence of vehicle users in accident injuries. In advance, RTAs patterns can be clarified using ML. However, ML techniques can handle both nominal and discrete variables with more than two levels and deal well with multi-collinear explanatory variables as well (Al-Moqri, et al., 2020).

1.1. Research Problem

The growth of population in the Kingdom of Bahrain is clearly noticeable, in 2012 the population was recorded as 1,208,964 people while in 2021 recorded as 1,504,365 people (Al-Ayam Newspaper, 2022). Also, the number of vehicles licensed that use the road in the Kingdom of Bahrain is steadily increasing. In 2016, the GDT registered 653,040 vehicles, while the number of vehicles increased in 2020 to reach 736,850 vehicles (General Directorate of Traffic, 2020). Therefore, the population is exposed to RTAs that affect the health and life of the population in the Kingdom of Bahrain, which leads to a negative impact on the economy of the country. For that, this study tries to use ML models based on high performance with several evaluation methods to analyze RTAs in the Kingdom of Bahrain.

1.2. Research Questions

In this study, the researcher aims to answer the following questions:

- What is the most effective ML algorithm that can classify the severity of RTAs in the Kingdom of Bahrain based on the highest criteria that could be achieved?

- Can ML algorithms be used to determine the factors leading to RTAs?
- What are the effects of resampling methods on the performance of the selected algorithms?

1.3. Research Objectives

The objectives of this research are:

- To classify the severity of injuries caused by traffic accidents using a ML approach.

- To determine the most important factors leading to accidents on the roads using a ML approach.

- To enrich the understanding of the research community for RTAs prediction in the Kingdom of Bahrain.

1.4. Research Significance

This study aims to contribute to avoiding or reducing traffic accidents. Also, explains to Bahrain residents what are the most common factors that lead to RTAs. Determine the algorithm capable of predicting the severity of RTAs. Since no previous study in the Kingdom of Bahrain used the ML approach to analyze the RTAs, this is the first study to use this approach to analyze traffic accidents on roads in the Kingdom of Bahrain. Therefore, this study contributes to supporting Resolution A/RES/74/299 issued by the UN in working for road safety to reduce injuries and deaths in the period 2021 and 2030.

1.5. Research Setting

In this study, the data under investigation were collected from the GDT in the Kingdom of Bahrain. These data described the features of

accidents that happen in different places of the kingdom of Bahrain such as level of accidents, causes, locations, Governorates, and several other features. The dataset within this research discussed the RTAs with a time range from January 2018 until June 2022.

1.6. Research Organization

This thesis is organized into five chapters. The first one provides an introduction related to RTAs. The conceptual framework and literature review presents in chapter two, while chapter three covers the research methodology, the results and discussion found in chapter 4, and the conclusion with recommendations in chapter 5.

Chapter Two

Conceptual Framework and Literature Review

2.0. Introduction 9

2.1. Machine Learning 9

2.2. Types of Machine Learning 10

2.3. Types of Machine Learning Summary 19

2.4. Feature Selections 26

2.5. Contributing Factors to Road Traffic Accidents 26

2.6. Previous Studies on RTA Using Machine Learning Models. 27

2.7. Previous Studies Summary 32

2.8. Summary of the Number of Algorithms Used 37

2.9. Conclusion 38

Chapter Two

Conceptual Framework and Literature Review

2.0. Introduction

This chapter explains the meaning of ML and illustrates several algorithms for supervised and unsupervised learning. Also, it discusses several types of ML algorithms, especially those used frequently in previous studies that focused on datasets in the field of RTAs.

The rest of this chapter is organized as follows: section 2.4. presents feature selection. Section 2.5. introduces the three most important factors affecting RTAs: human, vehicle, and environmental. Finally, section 6 discusses the essential findings of previous studies by focusing on evaluation methods such as accuracy, sensitivity, and recall.

2.1. Machine Learning

ML is defined as a study field that helps computers to get the ability to learn without the need for explicit programming and it is divided into supervised learning which depends on labeled data, unsupervised learning doesn't need to label data, and semi-supervised learning is the combination between supervised and unsupervised ML (Osisanwo, et al., 2017; Mahesh, 2020).

Many programmers and mathematicians try to design ML algorithms to solve problems including big data in various fields around the world such as Medicine, education, sports, transportation, and economics (Colliot, 2023; Javeed, et al., 2023). This study will focus on supervised learning in the field of RTAs.

2.2. Types of Machine Learning

ML algorithm techniques are applied in many fields. According to previous studies, it is observed that the present time ML algorithms are increasingly being used to determine the causes of traffic accidents as well as to detect the severity of RTAs.

During this part, the researcher explains the most used ML algorithms in the field of RTAs based on previous studies and illustrates the findings of several papers related to these algorithms.

2.2.1. K-Nearest Neighbor (KNN)

One type of Supervised ML method is the K-Nearest Neighbor (KNN) algorithm, which is used as a classification method. This algorithm includes one important parameter known as K which refers to the number of nearest points. The idea of the KNN algorithm is to use a training dataset to find the nearest data neighbors. Finding the distances between the training points and the test point helps rank the closest data points. After determining the k points closest to the data, it implements the rule of majority voting to decide which class appeared more than others. The final classification of the query is the most visible class (Fiorentini & Losa, 2020; Uddin, et al., 2022). Mailagaha & Luukka (2022) explain that one of the ways to find the distance between two points is using the formula of Minkowski distance which is defined as follows:

$$d(x_i, x_j) = \left(\sum_{t=1}^{m} \left|x_t^i - x_t^j\right|^p\right)^{1/p} \quad (1)$$

An example of a point that can be classified based on most voting of the nearest neighbors provides in Figure (2.1).

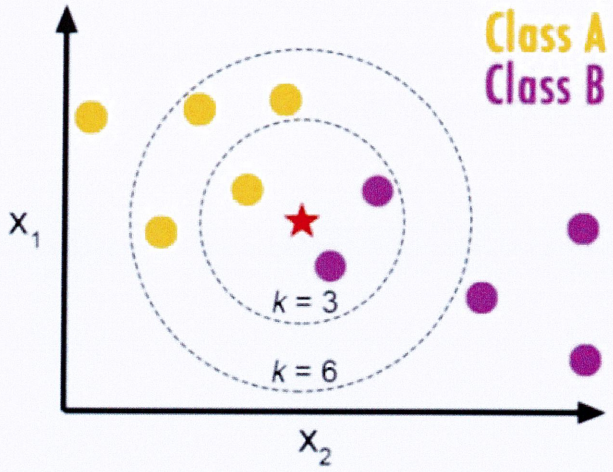


Figure (2.1): Explain the Way to Find the Value of K in the Model of KNN (Prihatmono, et al., 2022)

Labib, et al. (2019), illustrate two experiments by using four then two target classes to analyze RTAs in Bangladesh by using a ML approach. In the mentioned study, by focusing on the algorithm of KNN the results appeared an improvement with all evaluation methods. The precision score increased from 68% to 70%, the accuracy increased from 67% to 69%, and the F1-score increased from 67% to 69%. In addition, the researchers found when the surface type of road is Sealed the probability is high of an accident.

Zhang, et al. (2018) used ML and statistical methods about accident vehicles focusing on Florida city to predict accuracy and find the important variables. The results show the ML algorithm methods score higher accuracy than statistical methods. The highest accuracy was achieved with KNN and Random Forest (RF) of 52.9% and 53.9% respectively.

2.2.2. Support Vector Machine (SVM)

Liao, et al. (2018) explain that SVM is a type of supervised ML algorithm, and it is based on the theory of statistical learning. This algorithm was created by H. Drucker and C. Cortes. The characteristics of this algorithm are the ability to learn the model by using small samples and

it contains good generalization performance. SVM was used firstly in solving problems related to linear discrete data of binary classification. Figure (2.2) shows the principle of finding the satisfied hyperplane required for a specific dataset, and the maximum margin obtained between two points. To solve cases related to nonlinear problems, should map the input variables into the high dimensional and then find the suitable hyperplane. Choosing the satisfied kernel functions to create a linear function from a nonlinear one, and the following functions related to linear and nonlinear classification respectively:

$$f(x) = sgn\{w^*.x + b^*\} = sgn\{\sum_{k=1}^{m} a_k^* y_k (x_k.x) + b^*\} \quad (2)$$

$$f(x) = sgn\{\sum_{k=1}^{m} a_k^* y_k K(x, x_k) + b^*\} \quad (3)$$

Where a_k^*, b^*use to choose the optimal hyperplane. $(x_k.x)$ is dot related to two vectors. And y_k refers to classes of the target.

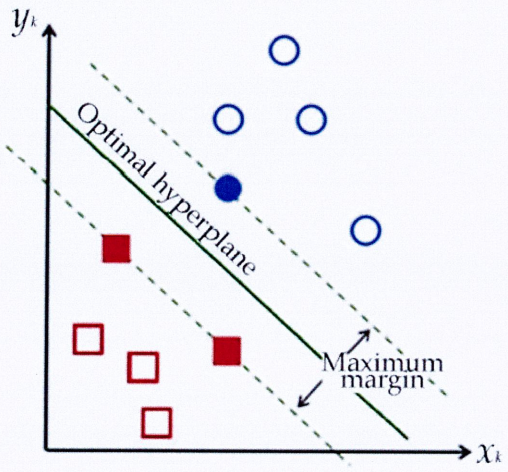


Figure (2.2): Hyperplane Concept Related to SVM Algorithm (Liao, et al., 2018)

Rahim & Hassan (2021) used the algorithms of CART and MARS to decide the most important variable which is the number of vehicles, followed by Driver airbag non-deployed and Collison manner (Sideswipe same direction) respectively. In addition, the algorithms used to predict the severity of RTAs are SVM and deep learning models. When the SVM is

fitted on the CART variable the value achieved on recall and precision is 0.17 and 0.09 respectively.

Hosseinzadeh, et al. (2021) study large truck accidents to predict the factors that influence accident severity. The algorithms used are SVM and Random Parameter Binary Logit (RPBL). The most important factor found in both models is a deviation to the left. This factor improves the probability of fatal accidents by 0.122 in single vehicles and by 0.245 in multivehicle when the truck is not at fault. In addition, Rollover depends as one of the significant factors in SVM with the category of single-vehicle truck accidents but is not significant in the algorithm of RPBL.

2.2.3. Naïve Bayes (NB)

Naïve Bayes (NB) is a supervised algorithm based on the idea of the Bayesian theorem with robust independence assumptions. This model provides an efficient way to deal with any number of attributes because it is based on the probability theorem. NB works as a practical learning algorithm and provides prior knowledge of observed data. It works well with the high dimensionality of the dataset (Bahiru, et al., 2018). The formula related to the NB algorithm is:

$$P(c/x) = \frac{P(x/c)\,P(c)}{P(x)} \tag{4}$$

Where $P(c/x)$ is the posterior probability, $P(x/c)$ is the likelihood, $P(c)$ is class prior probability, and $P(x)$ is predictor prior probability.

Bahiru, et al. (2018) aim in their study to predict the severity of RTA and identify the most important variables that affect accidents through using the NB algorithm and three types of DT: CART, ID3, and J48. Also, by applying techniques of feature selection like gain ratio and information gain, the most important feature was Weather condition then followed by Accident time and Number of lanes. This study used 10-fold cross-

validation to evaluate the performance of prediction for algorithms under investigation. The results of accuracy achieved 96.30%, 96.22%, 94.62%, and 94.09% for ID3, NB, CART, and J43 respectively, and the best Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC) were achieved with the model of NB.

The study of Jeong, et al. (2018) aims to classify the severity of injury in RTA. The results found the NB achieved an accuracy of 84.54% by using five target classes with no bagging method and no treatment used. the accuracy of NB increased to 84.98% by reducing the number of classes to three. In addition, the value of accuracy with NB keeps increasing by reducing the number of classes into two.

2.2.4. Random Forest (RF)

The main idea of the Random Forest (RF) algorithm depends on reproducing a subset of data and selecting the input variables randomly. Moreover, RFs are a set of decision trees where each tree is based on values of a random vector taken individually. Also, using the best subset of predictors selected randomly helps in node segmentation. This algorithm has achieved very well in the part of overfitting compared with many other classifiers. In this method, it is important to determine the parameters like the number of variables selected randomly and the number of trees. In addition, to increase the correlation between the trees, a higher value of the layers is chosen to divide the tree nodes to provide identical results during voting. (Chen, Chen, 2020; Ijaz, et al., 2021). The Instructure of the RF algorithm is illustrated in Figure (2.3), and the formula related is defined as:

$$p = 1 - \sum_{i=1}^{N} P_i^2 \qquad (5)$$

Where p refers to dataset and N refers to number of classes.

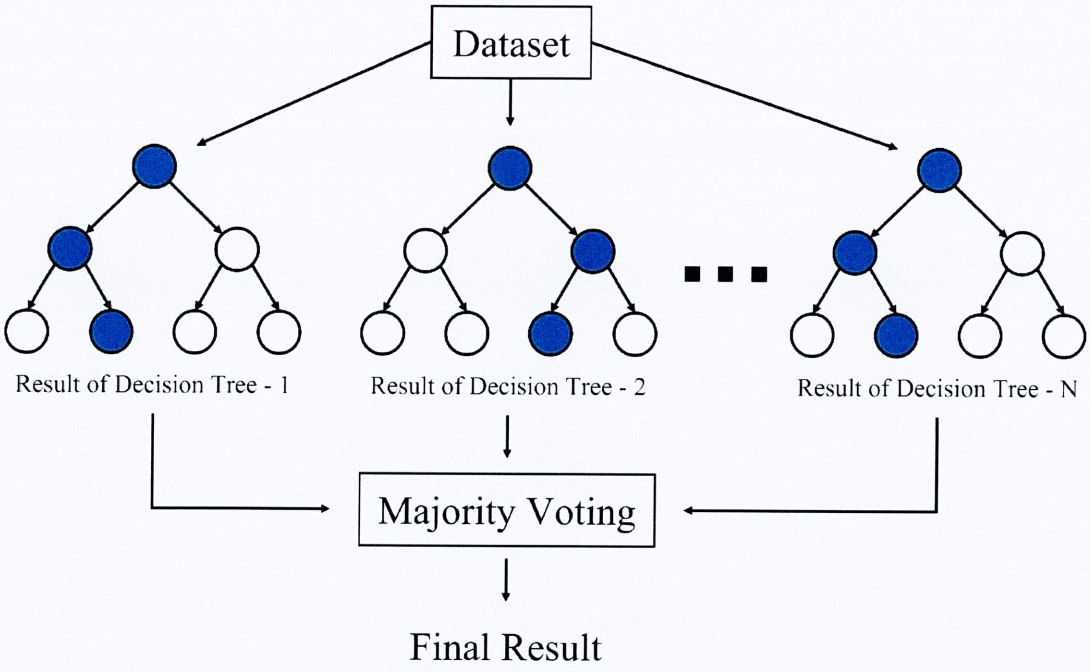


Figure (2.3): An Illustrated Diagram of the Random Forest Algorithm

Lee, et al. (2019) identify the optimal number of trees by calculating the out-of-bag estimate of the error rate of each model. Then divided 500 trees into samples of five sizes, N= {100, 200, 300, 400, 500}. The closest values to observed values were achieved by the RF model. The lowest Mean Squared Error (MSE) and Root Mean Square Error (RMSE) were yielded with RF then followed by DT and Artificial Neural Network (ANN) respectively. By using RF, the important variables were Rainfall Intensity, Curve Length, and Driver Gender respectively.

Chakraborty, et al. (2021) propose to provide causal inference and injury severity classification for highway traffic crashes in the state of Texas. The results found the most influential factors affecting accident severity are speed limit, weather conditions, and traffic volume. Also, the DT algorithm creates the highest performance for Property Damage Only, and the RF algorithm makes the greatest performance for Possible and Non-sever Injury.

2.2.5. Decision Tree (DT)

DT is a supervised algorithm used for classification problems. The main idea of this algorithm is to predict the results of the variable by deducing decision rules created from attributes of the dataset. Creating the construction of the algorithm of DT needs a root node that is based on the best feature, then generating the sub-nodes depending on the quality at each node. By continuing this process recursively for each node, the class is defined which is known as the leaf (Azhar, et al. 2022; Labib, et al., 2019).

Ospina-Mateus, et al. (2021) aim to analyze motorcyclist accidents to increase safety on the road. Decision rules are used to define patterns in the established conditional variables such as if the age of pedestrians is between 40 and 79 then the fatality is more probable. Also, if the day is Monday the motorcycles are more probability to be involved in RTA. On the other hand, the algorithms of SVM, CART, C4.5, and Random Tree achieved the same highest accuracy of 74%, and the highest recall of 74% was achieved with SVM, Cart, and C4.5.

Sripuram, et al., (2022) created models to predict traffic accidents and identify patterns associated with them to develop safety measures. The results of DT model scored a recall of 77.78%, an F1-measure of 87.5%, and an accuracy of 77.78% achieving a value higher than Logistic Regression (LR) but lower than RF.

2.2.6. Multilayer Perceptron (MLP)

MLP is a supervised algorithm and is considered a type of neural network which needs to add many hidden layers. Also, for computing, the total weight of biases and input needs to use linear or nonlinear activation functions. MLP includes the iterative procedure and adjusting the weights to minimize the error to achieve well prediction (Al-Moqri, et al., 2020).

Cigdem & Ozden (2018) compare ML algorithms to estimate the results of the RTA that happen in Adna of Turkey. The researchers changed the size of the target classes because there are just a few records of nonfatal classes. The results found the performance of the MLP algorithm achieved better results in many evaluation methods by changing the size of the target class of nonfatal from threefold into one-fold where the AUC increased from 0.839 to 0.872 and False Positive Value from 0.521 to 0.797. In addition, by using the size of one-fold the results found the value of AUC achieved the highest score of 0.925 followed by the KNN of 0.902.

Hussain, et al., (2019) proposed to predict the classes of injuries that are caused by RTA. MLP performed better than other algorithms in all other evaluation methods. The highest accuracy for MLP is 85.33% followed by BayesNet and J48 of 80% and 78.67% respectively.

2.2.7. K-means Algorithm

ML algorithm of k-means is unsupervised where there is no importance to knowing the dependent variables. The k-means algorithm is used in various fields such as data analysis, information retrieval, and image processing. Also, this algorithm is used to solve the problem of unobserved heterogeneity with features. In addition, this algorithm needs M points, N dimension, and K clusters after finding the distance from the cluster data point to the nearest cluster. The main objective of the K-means algorithm is reducing the Euclidean distance between points from the centroid. This helps to reduce variance and increase the similarity of inter-cluster (Yassin, 2020). The squared error is shown in the next formula:

$$f(x) = \sum_{n=1}^{k} \sum_{n=1}^{n} \left| X_i - C_j \right|^2 \quad (6)$$

Where n is the case number, C_j is the centroids number, X is data points, and k is the number of clusters.

Yassin (2020) mentions there are fourteen variables, and the target includes three classes: fatal, serious injury, and light injury. Also, the author divided the dataset into two parts, the first part 70% to train the model and the second part 30% to evaluate the model. This research used K-means to create a new feature to improve the performance of classification, which effect to make dramatically improved evaluation methods. Elbow and other techniques help to choose the appropriate value of K to be 3. In addition, RF achieved the highest values in all evaluation metrics before and after adding the new feature followed by LR.

2.2.8. Summary of the Common ML Algorithms

The previous works include several insights that could be highlighted in this section. For example, the researchers used supervised and unsupervised machine learning models that were created based on the target of the study and the type of features of the dataset such as SVM, KNN, and K-means. Most of the used models fall under classification methods that are applied to predict the severity of RTAs. Although the data used in the mentioned studies differed, there are many models that could achieve good performance such as RF, NB, and DT.

2.3. Types of Machine Learning Summary

This part provides a summary of the common ML algorithms in the RTAs.

Table (2.1): Summary of Common Machine Learning Algorithms in the Field of RTAs

Authors	Aim	Dataset Description	Methods	Criteria	results	Positives	Comments / Improvement need
(Lee, et al., 2019)	Build a predictive model of traffic accidents and identify the important features to reduce the frequency of accidents.	9- years duration from Seoul in Korea. 518 instances. Target Classes are: critical incident and non-critical incident.	RF DT ANN	the out-of-bag estimate of error rate, MSE, RMSE	**MSE** RF: 0.047 ANN: 0.102 DT: 0.079 **RMSE** RF: 0.217 ANN: 0.319 DT: 0.281	• Well description for variables of the dataset under investigation.	• The number of accidents is small to evaluate the models. • It is recommended to add more evaluation methods.
(Chakraborty, et al., 2021)	Classify highway traffic crashes.	State of Texas, USA. Period of 6-years from 2014 to 2019 The number of accidents is 156,166. Three classes:	DT RF XGBoost Deep Neural Network	Confusion matrix	The most important factor is the speed limit. Light conditions have the lowest effect on accidents.	• Discussed the important factors affecting the RTAs in detail.	• Recommended using more critical methods to evaluate the algorithms.

		severe injury and fatal, non-severe and possible injury, and property damage only.					
(Bahiru, et al., 2018)	Use different classification methods to find the more influential factors and predict the target of RTAs to prevent new accidents.	2-years duration from 2015 to 2016 The dataset was collected from the UK. 3,050 records Classes of target are fatal, serious, and slight.	NB ID3 DT (CART, and J43)	Accuracy ROC AUC	The highest accuracy was achieved with J48 of 96.30% and NB achieved an accuracy of 94.6%. Correctly classified instances for J48 are 2936 and incorrectly are 114 instances. NB scores highest ROC and AUC.	• Used several algorithms and evaluation methods to compare results. • Explain each algorithm clearly in subsections.	• To analyze attributes, it is recommended to explain variables in a simplified table instead of using a screenshot taken from the tool Weka. • There is no mention of the number of training and testing datasets. • There is a difference between illustrated attributes and analyzed attributes. the first case includes a Seat belt but changed in the second case to the number of lanes.

(Rahim & Hassan, 2021)	This research aims to propose a deep-learning algorithm to predict the target classes of RTA.	5-years duration from 2014 to 2018 Louisiana, United State of America. 10,048 accidents The target includes three levels of classes: PDO, fatalities, and injuries.	DT (CART, and MARS) SVM Deep learning.	Recall Precision F1-sccore	The deep learning algorithm performed better than SVM in the process of classifying injury and fatal accidents. Better recall is achieved with Mars model.	• Used algorithms of CART and MARS to find the significant variables.	• Need to use more algorithms to compare the results between them.
(Ospina-Mateus, et al., 2021)	Analysis the road traffic accidents of motorcyclists to increase road safety.	The dataset covered six years duration from 2013 to 2018. Bogotá, Colombia 34,232 motorcyclists' accidents The variables of target attributes are only damages, accident with injuries, and accident with fatalities.	SVM KNN NB NN RF Random Tree (RT) DT (CART, ID3, and C4.5)	Recall Precision Accuracy	The NB achieved the highest precision of 75% and the lowest accuracy of 62%. On the secondary roads, if the man is the motorcyclist that may cause two victims females aged between 20 and 39.	• Discuss the rules deeply in a proper way.	• Recommended to explain briefly the algorithms used in the study.

(Labib, et al., 2019)	To determine the factors that affect accidents that happened on the roads, and to analyze RTAs by using the ML approach.	15-years duration from 2001 to 2015 Bangladesh 43,089 records. The classes of the target are Fatal, Grievous, Simple Injury, and Motor Collision	DT KNN NB AdaBoost	Precision Accuracy F1 Score	Based on evaluation methods, the AdaBoost model achieved the best performance.	• Used two types of experiments by reducing the number of classes from 4 to 2. • Discuss the significant factors that affect the RTAs in a proper way.	• There is a slight increase in the value of accuracy for all types of algorithms used.
(Hosseinzadeh, et al., 2021)	Predict the factors that affect the accident injury severity caused by large trucks.	5-years duration between 2011 and 2014. Iran 8,552 truck-involved accidents. The output data is classified into non-fatal crashes and fatal crashes.	Random Parameter Binary Logit (RPBL) (SVM).	ROC AUC	The most significant factor found is Deviation to the left. Track driver is fault: AUC RPBL = 0.71 AUC SVM = 0.76 Truck driver is not fault: AUC RPBL = 0.67 AUC SVM = 0.71	• Clearly discussed the important factors that affect the severity of accidents.	• There are only 8 attributes that affect the predicting model. • Recommended adding more than two evaluation methods like accuracy and precision.

(Cigdem & Ozden, 2018)	Using ML methods to determine the influences of weather and other factors in traffic accidents.	11-years duration from 2005 to 2015 25,015 records Adana, Turkey The classes of the target are fatal and nonfatal.	KNN NB MLP DT SVM LR	Confusion matrix Sensitivity Specificity Accuracy ROC AUC	By using the size of three-fold for nonfatal the highest accuracy was achieved with DT. By using the size of one-fold for nonfatal the highest AUC was achieved with MLP.	• Used different types of evaluation methods. • Several algorithms were used to compare the results between them.	• Need to explain all abbreviations.
(Zhang, et al., 2018)	Comparison between the performance of statistical and ML methods in the prediction accuracy and significant variables.	3-years duration. Florida, United State of America (USA). 5,538 records. Five classes of target: No injury, invisible injury, No-capacitating injury, Incapacitating injury, and Fatal.	KNN DT RF SVM	Accuracy	KNN and RF achieved the highest accuracy.	• Use various methods to compare the performance.	• Need to mention the years of collecting the data. • Recommended to use of more evaluation methods.

(Jeong, et al., 2018)	Train the models to predict the injury severity caused by RTAs.	2-years to collect the dataset from 2016 to 2017. Michigan, United Stat of America. 297,113 accidents. Classes of the target: Fatal injury, Incapacitating injury, non-incapacitating Injury, Possible injury, and No injury.	LR DT Neural Network (NN) Gradient boosting model NB	Accuracy	The highest accuracy was achieved by using bagging with the model of DT.	• The summary of variables is presented clearly. • The accuracy raised from less than 80% to more than 90%.	• Recommended using more critical methods to create a clear evaluation of the performance.
(Sripuram, et al., 2022)	Designing algorithms to predict traffic accidents to develop safety measures on the road.	37 attributes	DT RF LR	Accuracy	The accuracy: DT: 77.87% RF: 86.86% LR:67.48%	• Used several methods for prediction.	• Need to mention more information about the dataset. • It is proper to prepare the results in the table rather than taking screenshots from the software tools.

(Yassin, 2020)	Using clustering and classification models to predict RTA severity.	Addis Ababa, Ethiopia. The dataset includes 5,000 records collected from 2011-2018. Severity classes are Fatal, Severe Injury, and Light Injury.	K-means LR RF SVM KNN ANN	Accuracy Recall Specificity Precision F1 Score	RF achieved the highest value in all evaluation metrics before and after adding a new feature followed by DT. The accuracy for models with new feature is: LR: 99.13 RF: 99.86 SVM: 73.13 KNN: 68.58	• Used supervised and unsupervised algorithms. • There is a noticeable improvement in all evaluation metrics.	• Need to mention the contributions of this research. • It is recommended to compare with other studies.
(Hussain, et al., 2019)	Use a different type of ML algorithms to predict the classes.	Data collected from January 2014 to June 2016 Kano, Nigeria. 150 instances	MLP DT (J48) Bayes Network	Accuracy Specificity Recall Precision	The accuracy: MLP: 85.33% J48: 78.66% BayesNet: 80.66% The specificity: MLP: 0.473 J48: 0.624 BayesNet: 0.787	• Many classifiers are used to compare results. • Clearly explain each classifier.	• The number of records is very low to use in evaluation and predicting the model. • Recommended to illustrate the summary description of the dataset.

2.4. Feature Selections

In the study related to Sarkar & Sarkar (2020), they decided the best features by using several feature ranking techniques such as Information gain, Symmetric uncertainty, and Gain ratio. All variables were tested using the ensemble ranking method to develop the model based on their high prediction. The researchers found that between twenty-four input variables, the significant features selected for the prediction of accident severity are eight features they are Annual Average Daily Traffic, Road Width, Percentage of Left-Turn Vehicles, Driver's Age, Percentage of Right-Turn Vehicles, Type of Vehicles Involved, Spot Speed of Major Road, and Cause of Accidents.

2.5. Contributing Factors to Road Traffic Accidents

There are many previous studies that discuss the contributory factors that caused RTAs. This part explains human factors, environmental factors, and vehicle factors due to their impact on the severity of RTAs.

2.5.1. The Human Factor

More than fifty percent of accidents on the road happen due to human factors. Some examples of human factors are drunk driving, speeding, and fatigue driving (Zhang, et al., 2019).

Lee, et al. (2019) divided the contributory factors into four parts: road alignment factors, road environment factors, weather factors, and human factors including driver age and driver gender. Also, Bokaba, et al. (2022) used features of gender and rider age as human factors that affect the RTAs.

2.5.2. The Vehicle Factor

Although vehicle factors are important in the field of RTAs, there are few studies using datasets that include them. Liao, et al. (2018) used

vehicle weight rating and vehicle type of the frontal vehicle as vehicle factors. In addition, the factors related to vehicles used in the study by Santos, et al. (2021) are heavy vehicles, light vehicles, and motorcycles.

2.5.3. The Environment Factors

Most of previous studies used dataset involved road factors. Bokaba, et al. (2022) mention features related to road factors they are Wet Road, No. of Travel Lanes, No. of Vehicles Involved, and Roadway Name. Kumar, et al. (2018) use attributes of Lighting Condition, Roadway Feature, Road Type, and Surrounding Area as environment factors.

2.6. Previous Studies on RTA Using Machine Learning Models.

This section illustrates many types of algorithms used in previous studies to compare their best results by using different critical methods.

2.6.1. RTAs in the Worldwide

In the research of Bokaba, et al. (2022), the dataset under investigation covered four years of duration related to RTAs. There are eight features selected after removing others due to insufficient entries. The researchers compared the performance of the methods by using default models and after optimizing the same models. The RF achieved the best performance in all evaluation models for default and optimizing. After optimizing for accuracy and RMSE the RF shows values of 97% and 0.12 respectively.

Lu, et al. (2022) integrate path analysis with ML to identify behavioral pathways with bicycle-motor accidents in North Carolina. The results find the highest accuracy of behavioral models with Neural Network (NN) of 64%, while the severity models show the best accuracy of 49% for RF and NN.

Mokhtarimousavi, et al., (2020) determine the factors affecting traffic accidents for pedestrians during the days of the week. The results found that most accidents of Sever Injury happened at weekends 46.8%. The best AUC, F1-Score, Specificity, Sensitivity, and Accuracy were achieved with the algorithm of (WOA-ANN).

In the research of Ijaz, et al. (2021), the researcher uses a dataset covering three years from 2017 to 2019. The study focused on the severity of injuries due to motorized three-wheeled Rickshaws. Jungle forest achieved a high accuracy of 83.7% outperforming the RF and DT. By using the tuning parameter, the number of decision DAGs, the maximum width of decision DAGs, and the number of optimization steps were 16, 56, and 6,581 respectively. By using DT, the results indicate that the most important factors affecting the injury severity are weather, type of crash, rider's age, speed limit, and lighting condition. Moreover, regarding the drivers' factors, if the gender was male and the age between 20 -30 years ago, then speeding is associated positively with injury severity.

Kumar, et al. (2018) used Latent Class Clustering (LCC), Balanced Iterative Reducing and Clustering Using Hierarchies, and K-modes as clustering techniques on data related to traffic accidents to find that LCC creates better clusters than others. After that, the researchers used three classification algorithms SVM, RF, and NB to classify the data into the severity of RTA. Finally, the result found that the RF performed the highest accuracy then followed by NB, and SVM achieved 81%, 79%, and 72% respectively.

Santos, et al. (2021) choose clustering as an unsupervised ML algorithm to discover patterns through datasets. The data is separated into two clusters, the first one includes 2,0732 and the second cluster includes 7,371 instances. After that, the researchers used the R package of

(randomForest) and then found the influential feature was MoterCycle then followed by TypeAcid, HitAndRun, and AirTemp respectively. Moreover, with the classification algorithm, the dataset was divided into a 70% training set and a 30% test set where RF and LR achieved the highest accuracy of 73%.

2.6.2. RTAs In the Arab Region and the GCC.

Al-Moqri, et al. (2020) used several algorithms to predict the severity of road traffic crashes in Yemen. The classifiers used during this study: NB, J48, RF, SVM, and Multilayer Perceptron (MLP). Between these algorithms, the RF scored the highest accuracy, recall, precision, and F-measure of 94.84%, 94.8%, 94.9%, and 94.8% respectively.

AlKheder, et al., (2020) aim to understand the correlation between input and output variables. The algorithms of DT, SVM, and Bayesian network are used to compare the performance in predicting the selected variables. The results found the highest probability of fatal and severe injuries was for pedestrians. Also, speed limits affect the level of injuries for pedestrians, and when the gender of the driver was male the fatalities frequent more than females.

Jamal, et al., (2021) compared the performance between the XGBoost model and traditional algorithms for predicting the severity of RTA in the city of Riyadh, Kingdom of Saudi Arabia. The most important features were collision type, weather status, and road surface. The results found the performance of XGBoost outperformed other algorithms in the evaluation methods of accuracy, recall, precision, and F-measure.

Taamneh, et al., (2017) established classifiers to predict the level of injury in Abu Dhabi. The results exhibited less accuracy with the NB algorithm. The fatal severity was associated with nationality, gender,

accident year, and age as the most significant factors. Also, the accidents are frequently more with male drivers than female drivers.

Alrajhi & Kamel (2019) aim to forecast future injuries, fatalities, and accidents in the Kingdom of Saudi Arabian. The researcher used only one algorithm, Long Short-Term Memory (LSTM), in predicting. The results found the number of accidents in 2013 was around 35000. While the algorithm of LSTM predicts to be less than 25000 accidents in 2030.

2.6.3. RTAs in the Kingdom of Bahrain

There are no previous studies discussing the RTAs using the ML approach in the Kingdom of Bahrain. For that, this is the first study to try to analyze the accidents on the road of the Kingdom of Bahrain using techniques related to ML.

However, the Bahraini General Traffic Department elaborated on the most important factors affecting traffic accidents. In 2017, the GDT find drivers are more likely to be involved in traffic accidents when they are between the ages of 20 and 34 and if the gender is male. Pedestrians between the ages of 5 and 9 are more probability to be involved in traffic accidents. Also, the highest number of traffic violations was issued for speeding and crossing a red light (General Directorate of Traffic, 2017).

Results of RTAs in 2018 found the Capital Governorate recorded the largest number of traffic accidents of various types of slight, serious, and fatal, and the most traffic accidents were recorded on Wednesday (General Directorate of Traffic, 2018). In 2019, October was the month with the highest number of traffic accidents recorded, and the main cause of traffic accident injuries is crossing a red light. Vehicles over 8 years old are more likely to be involved in traffic accidents (General Directorate of Traffic, 2019).

In 2021, the Capital Governorate recorded the highest number of accidents on the road. Twenty deaths were recorded throughout the mentioned year due to RTAs, and the main cause of traffic accidents is over speed. In addition, the two cities with the most traffic accidents were Manama and Riffa, and the largest number of traffic accidents was recorded in November (General Directorate of Traffic, 2021).

2.7. Previous Studies Summary

RTAs studies have been summarized in this section.

Table (2.2): Summary for Studies About RTAs

Authors	Aim	Dataset Description	Methods	Criteria	results	Positives	Comments / Improvement need
(Bokaba, et al., 2022)	Comparing the prediction models to assist policymakers and transport authorities.	Dataset collects from the Gauteng Department of Community Safety.	NB LR KNN AdaBoost RF SVM	Accuracy Precision Recall ROC RMSE	RF has the best value in accuracy of 97% and all other evaluation methods.	• Use six methods of algorithm which is a good number to compare between their performance.	• Didn't mention the dataset refers to which years. • Need to add summary statistics of variables.
(Lu, et al., 2022)	Determining the behavioral paths in bicycle and car collisions.	Dataset covered 3-years from 2017-2019 North Carolina, USA 9,296 accidents Three levels of injury: Minor Injury, Serious Injury, and Killed.	RF NB SVM AdaBoost Neural Network (NN)	Accuracy F1-sccore AUC	**Accuracy of Behavioral Models** RF: 55% SVM: 62% NN: 64% AdaBoost: 55% NB: 54% **Accuracy of Severity Models** RF: 49%	• Variables are described clearly on a nice table. • Several methods of algorithms used.	• prefer to use more evaluation methods.

					SVM: 46% NN: 49% AdaBoost: 47% NB: 44%		
(Ijaz, et al., 2021)	Assess the risk features related to motorized three-wheeled Rickshaws (3W-Mr) accident severity.	3- years duration (2017 - 2019) from Rawalpindi in Pakistan. 2,734 instances. Target Classes are: No injury, Minor Injury, Severe Injury, and Fatal injury.	Decision Jungle (DJ) RF DT	Overall accuracy. Macro-average recall. Macro-average accuracy. Geometric mean. AUC	DJ algorithm achieved higher accuracy of 83.7%. The performance of AUC: DJ: 0.956 DT: 0.829 RF: 0.874	• The first research uses ML with motorized (3W-Mr) accident severity. • Add new columns about weather and climate to create the dataset.	• Should explain the pre-processing in more detail.
(Kumar, et al., 2018)	Compare between three clustering technique LCC, k-modes, and BIRCH on data related to traffic accidents to decide which one is the best for	Data from Muzzafarnagar in India. 2,300 accidents records. Target Classes are: Killed or Severe Slight Injury	K-modes clustering. LCC Clustering. Clustering using (Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH)	To choose the appropriate number of clusters: AIC BIC CAIC For classification methods: Accuracy	The elbow plot appears the appropriate number of clusters is 2. To compute different clusters the k-modes take less time. The LCC clustering method	• Used supervised and unsupervised methods.	• Literature review contains many studies established before 2008. • Need to mention the duration of the dataset. • The authors use only accuracy as one evaluation method for classification algorithms.

	classification purposes.		RF NB SVM		achieved the highest accuracy among all classification methods.		
(Santos, et al., 2021)	Developing a ML algorithm to identify factors that affect accidents to use to classify the severity of RTAs.	Setúbal, Portugal. This data includes 28,102 records. from 2016 to 2019. The classes were No Victims or Victims.	DT (C5.0) LR RF	Accuracy AUC Score Precision Sensitivity False positive rate Specificity	Generally, RF achieved the best results. DT achieved lowest accuracy of 65%.	• The author used cluster and classification methods.	• The data description could be illustrated in the table rather than explained in points. • It would be better to name the variables in each attribute. • Prefer to mention the reason behind choosing the LR, DT, and RF.
(Al-Moqri, et al., 2020)	Use ML algorithms to predict the severity of RTAs.	Dataset collected from August to October in 2015 in Yamen. 156 recorded The classes of injury severity: Severe, Serious, and Minor.	NB DT (J48) RF SVM MLP	Precision Recall Accuracy F-measure Confusion matrix Kappa statistics	The best accuracy was achieved with the RF of 94.84%. The method of SMOTE is used to balance the classes of output.	• Use 6 types of evaluation methods to decide the best performance of algorithms.	• The low number of observations used which that affects the performance of algorithms.
(AlKheder, et al., 2020)	Study the potential factors that contribute to	Abu Dhabi, The United Arab Emirates.	DT Bayesian network SVM	Accuracy.	The most important factor for SVM is the injured seat of	• explain in detail the factors that lead to different classes of RTA.	• There is no clear comparison between algorithms.

	determining the level of injuries resulting from RTA.	5,740 cases related to RTA collected in the period from 2008 to 2013.			position, the Bayesian network is Road Classification, and DT is the injured seat of position.		• Need to use more evaluation methods.
(Mokhtarimousavi, et al., 2020)	Investigate the factors that affect the severity of RTA of pedestrians on weekdays.	California, United State of America. 5-years duration in the period of 2010-2014 10,146 crashes Classes are: Severe Injury. Minor Injury. No injury.	ANN Random Parameter Ordered Probit. Random Parameter Ordered Logit. ANN Whale Optimization Algorithm	Accuracy Sensitivity Specificity F-Score AUC	Highest overall accuracy and specificity achieved with WOA-ANN.	• Used various algorithms and evaluation methods to find the results.	• Need to present a summary for data description.
(Jamal, et al., 2021)	Compare XGBoost with LR, RF, and DT algorithms for predicting the injury severity of RTA.	Riyadh, Kingdom of Saudi Arabia. 3-years duration from 2017 to 2019. 13,546 accidents.	XGBoost LR RF DT	Average Accuracy Cohen's Kappa Precision Recall F-measure	Accuracy: XGBoost: 95% LR: 92% RF: 89% DT: 72% Recall: XGBoost: 70% LR: 69% RF: 60%	• Several algorithms were used to classify the severity of RTA. • The result of accuracy reaches 95% which is a very good result.	• Recommended to find the results of evaluation methods after selecting the best features.

		Injury severity is property damage. Only, Injury, and Fatal.			DT: 48%		
(Taamneh, et al., 2017)	Identify the important factors of accident severity.	The number of accidents is 5,973 covers period of 6-years from 2008 to 2013 in Abu Dhabi Four level of classification: minor, moderate, severe, and death.	DT (J48) Rule Induction (PART) NB MLP	Accuracy Recall Confusion matrix	Accidents related to males more than females. The number of traffic accidents is gradually decreasing over the years.	• The result of accuracy for most injury severity increased with significant change by using Resampled training set technique.	• Prefer to use figures to compare the algorithms used.
(Alrajhi & Kamel, 2019)	Predict fatalities, injuries, and accidents.	Dataset covered period between 1996 and 2013. Kingdom of Saudi Arabia (KSA)	Long Short-Term Memory.	Mean absolute error. Mean Squared Error. Root Mean Squared Error.	The highest number of accidents was recorded in the year 2012.	• Using time series data with LSTM.	• Need to mention the number of records and attributes during the dataset. • Used only one algorithm of LSTM. • Record the results even if they were not high. • Important to provide a summary description of the dataset.

Previous studies can support the work of this study in many ideas. For example, it is found that many researchers combine several factors to create the dataset such as road, driver, and weather factors. The target feature of the datasets that are used contains two, three, or four classes. Various previous studies aim to compare the performance of classifiers by building several algorithms such as DT, SVM, and KNN. Also, it is observed some researchers find patterns and conditions that affect the RTAs. And the best performance in many studies was achieved by using the RF algorithm, but in some studies, the DT achieved the worst results.

2.8. Summary of the Number of Algorithms Used

This section presents the most used algorithms in previous studies that were discussed in this research.

Table (2.3) shows that the algorithm used frequently in the previous studies which is the DT, then followed by RF, SVM, NB, and LR respectively.

Table (2.3): The Number of Times the Algorithms were Used in the Identified Previous Studies

Algorithms	The Number of Repetitions
NB	10
LR	7
KNN	6
AdaBoost	3
RF	13
SVM	11
XGBoost	3
DT *	18
Bayesian Network	2
ANN	6
MLP	4
K-means	2
Others	13

2.9. Conclusion

This chapter begins with the concepts of ML with supervised, unsupervised, and semi-supervised learning, then explains the idea for several types of algorithms. Also, it is observed that the ML algorithms are used widely with the dataset related to RTA. Many researchers divided the attributes of the RTA dataset into contributing factors, which the most of them human factors, vehicle factors, and environmental factors.

Many studies mentioned how to use ML algorithms to predict the severity of RTA, but it is rarely used in the GCC and there are no previous studies that discussed this technique in the Kingdom of Bahrain. The most frequently used algorithms in previous studies mentioned were DT, SVM, and NB respectively.

After this review, it is now clear that the DT algorithm is used to select the most important features affecting RTAs. In addition, it is preferring to use many evaluation methods such as recall, sensitivity, accuracy, and MSE to compare the performance between different types of algorithms.

Chapter Three

Research Methodology

3.0. Introduction 40

3.1. Research Methodology 40

3.2. Data Description 42

3.3. Description Original Datasets 45

3.4. Data Combining 48

3.5. Data Understanding 48

3.6. Data Preparation 53

3.7. Dataset Description After Handling 55

3.8. Data Splitting 61

3.9. Resampling Methods 61

3.10. ML Models Used 61

3.11. Evaluation Methods 62

3.12. Conclusion 63

Chapter Three

Research Methodology

3.0. Introduction

The research methodology applied in this study is described in detail in this chapter. The preparation and handling of the dataset are first conducted to create an appropriate dataset to classify injuries resulting from RTAs. In addition, in this chapter, different algorithms and evaluation metrics are discussed.

3.1. Research Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is followed in this research. It includes five stages namely: business understanding, data understanding, data preparation, modeling, and evaluation and comparison as shown in Figure (3.1).

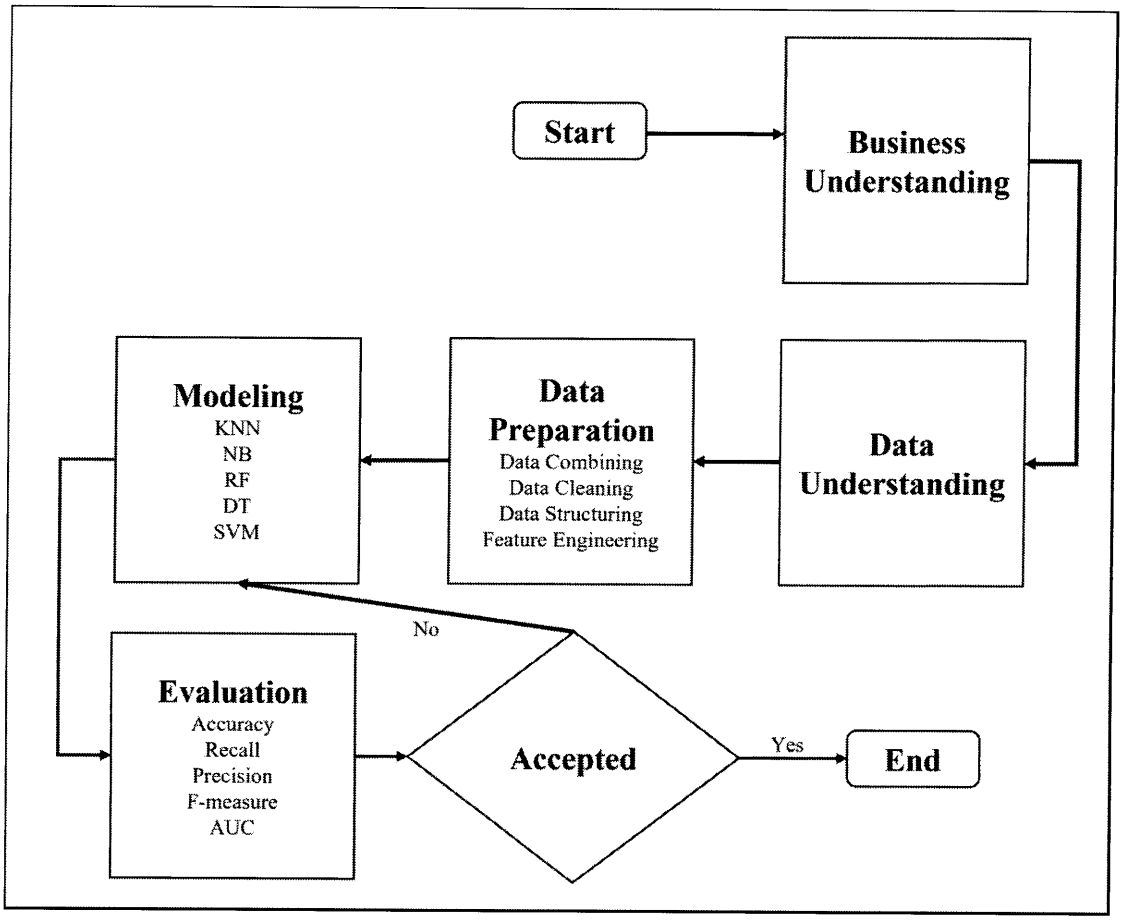


Figure (3.1): Flowchart of the Research Methodology

The stage of the literature review presents the procedures and results of previous studies that discussed the problem of RTAs. Different types of ML algorithms and methods usually used to maximize performance were explained in chapter two.

The business understanding section gives an overview related to the severity of the RTAs dataset, explains the quality of the dataset, and describes where the dataset of this study was collected from to accomplish the research goals.

The data preparation aims to prepare the final shape of the dataset to be used in the modeling stage. To achieve this phase, it is necessary to

make several operations such as removing values, adding new attributes, combining datasets, and converting the values to compatible types.

The modeling stage focuses on building models and then assessing them after selecting the algorithms that will be used. Also, during this stage, the process of splitting data into the training set and testing set takes place.

In the evaluation phase, the best model that meets the dataset of road accidents can be decided. This stage also includes procedures such as reviewing the work and summarizing findings.

3.2. Data Description

The dataset under investigation used in this study was obtained from the Bahraini General Directorate of Traffic (GDT) and collected by Development & IT Division at the Statistics Section. This dataset includes RTAs that happened in different places around the Kingdom of Bahrain showing the level of severity of injuries of three levels: slight, serious, and fatal based on several factors recorded during the accident such as gender, nationality, governorate, road condition, day of the accident, and weather. These datasets exist in five separate excel files because each one contains details about all traffic injuries that happen in a specific year from 2018 until June 2022.

3.2.1. Dataset One Summary (2018)

This dataset refers to accidents collected in the year 2018 with 1,550 casualties and included 21 different factors. The whole dataset is complete except attributes of gender, age, road surface, road condition, governorate, site, and accident junction missed several values. Table (3.1) shows the missing values for these columns, for instance age attribute has 20 missing values, while road condition attribute has only one missing value. In

addition, Table (3.6) provides the needed information about the feature name, feature description, and data type to understand dataset one better.

Table (3.1): Details of Missing Values in Dataset One

Feature Name	No. of Missing Values	Percentage of Missing Values (%)
Gender	6	0.4
Age	20	1.3
Road Surface	2	0.1
Road Condition	1	0.1
Governorate	1	0.1
Site	1	0.1
Accident Junction	7	0.5

3.2.2. Dataset Two Summary (2019)

Dataset two covers all traffic accidents that occurred in the year 2019. This dataset contains 1,363 accidents with 21 factors that were recorded during the time of the accident. Five attributes have some missing values presented with their percentages in Table (3.2). The highest number of missing values existed in the attribute Age of 5.5%, while the Road Surface has the lowest value of 0.2%. Also, Table (3.6) presents all attributes' name of dataset two with their descriptions and types.

Table (3.2): Details of Missing Values Related to Dataset Two

Feature Name	No. of Missing Values	Percentage of Missing Values (%)
Age	75	5.5
Road Surface	3	0.2
Road Condition	4	0.3
Site	4	0.3
Accident Junction	7	0.5

3.2.3. Dataset Three Summary (2020)

The RTAs recorded in dataset three occurred in the year 2020. This dataset includes 927 injuries from road accidents and 23 attributes with two additional features more than previous datasets. Missing values details about dataset three are represented in Table (3.3). This dataset shows the attributes: Age, Road Type, and Weather have only 2 missing values while Accident Junction missed 13 values. In addition, information about the description and names of features with their data type is shown in Table (3.7).

Table (3.3): Missing Values of Dataset Three

Feature Name	No. of Missing Values	Percentage of Missing Values (%)
Age	2	0.2
Road Type	2	0.2
Site	4	0.4
Accident Junction	13	1.4
Weather	2	0.2
Governorate	1	0.1

3.2.4. Dataset Four Summary (2021)

Dataset four contains the data about the year 2021 and includes 23 attributes and 958 injuries related to road accidents. Table (3.4) shows attributes that missed some values with their number and percentage to the whole data. The Road Condition attribute has the lowest value of missing data of 0.1%, while the attribute of Site has the highest value of missing data of 0.7%. More details can be found in Table (3.7) such as data type of features.

Table (3.4): Missing Values of Dataset Four

Feature Name	No. of Missing Values	Percentage of Missing Values (%)
Age	3	0.3
Road Type	3	0.3
Road Surface	2	0.2
Road Condition	1	0.1
Site	7	0.7
Accident Junction	5	0.5

3.2.5. Dataset Five Summary (2022)

Dataset five covered six months of accidents in the year 2022 from January to June. This set of data includes 585 accidents and 23 attributes. Table (3.5) shows the percentage and number of missing values. The highest attributes that have missing values are road type and accident junction followed by site, age, and visibility. The name of features of dataset 5 is presented in Table (3.7).

Table (3.5): Missing Values of Dataset Five

Feature Name	No. of Missing Values	Percentage of Missing Values (%)
Age	5	0.9
Road Type	7	1.2
Site	6	1
Accident Junction	7	1.2
Visibility	1	0.2

3.3. Description of Original Datasets

The first type of datasets covers datasets one and two for the years 2018 and 2019 as shown in Table (3.6).

Table (3.6): Describe Datasets One and Two

NO.	Feature Name	Feature Description	Data Type
1	Report Number	Unique number for each accident	Int64
2	Accident Date	The date on which the accident occurred	Datetime64
3	Accident Month	The month of the accident	Int64
4	Accident Year	The year of the accident	Int64
5	Accident Time	The time of accidents with hour and minutes.	Object
6	Accident DOW	The day of the week in which the accident occurred	Object
7	Location	The area and/or the name of the street	Object
8	Gender	The sex of the victim	Object
9	Age	The victim's age	Int64
10	Nationality	The country to which the victim belongs	Object
11	Road Type	Speed limits on the road	Object
12	Road Surface	The material from which the road is made	Object
13	Road Condition	The condition of road during accident.	Object
14	Site	The type of junction at the place where the accident occurred	Object
15	Accident Junction	Junction control during accidents	Object
16	Weather	The weather condition type	Object
17	Visibility	The condition of light	Object
18	Road Impact	Impact of the road on accident	Object
19	Person Involved	The type of victim	Object
20	Cause Type	Main cause of road accident	Object
21	Injury Type	The injury level of accident	Object

The second type of dataset describes datasets three, four, and five from 2020 to June 2022. This sort of dataset contains 23 attributes, and its type is differentiated from one attribute to another such as Int64, Datetime64, and Object. Table (3.7) presents the name of the attribute then describes it briefly and mentions its data type. During this period of the dataset, there are two additional attributes of Governorate and Region.

Table (3.7): Describe Dataset Three, Four, and Five

NO.	Feature Name	Feature Description	Data Type
1	Report Number	The unique number of accidents	Int64
2	Accident Month	The month on which the accident occurred	Int64
3	Accident Date	The date of the accident	Datetime64
4	Accident Year	The year of the accident	Int64
5	Accident DOW	The weak of the accident	Object
6	Accident Time	The time of accidents with hour and minutes.	Object
7	Location	The area and/or the name of the street	Object
8	Age	The victim's age	Int64
9	Gender	The sex of the victim	Object
10	Nationality	The country to which the victim belongs	Object
11	Governorate	The governorate of the accident	Object
12	Region English	The name of the area where the accident occurred	Object
13	Road Surface	The material from which the road is made	Object
14	Road Type	Speed limits on the road	Object
15	Road Condition	The condition of road during accident.	Object
16	Accident Junction	Junction control during accidents	Object
17	Site	The type of junction at the place where the accident occurred	Object
18	Weather	The weather condition type	Object
19	Road Impact	Impact of the road on accident	Object
20	Visibility	The condition of light	Object
21	Person Involved	The type of victim	Object
22	Injury Type	The injury level of accident	Object
23	Cause Type	Main cause of road accident	Object

3.4. Data Combining

To combine all five datasets together, the researcher adds one new attribute of the governorate to the first type of dataset and deleted the attribute of the Region from the second type of dataset which contributes to creating the same attributes in both types of datasets. To complete variables of the Governorate, the researcher depends on the attribute of location and then refers to approved zoning maps created by the Urban Planning and Development Authority in the Kingdom of Bahrain. Also, the positions of the features were arranged to be the same as each other in all datasets.

3.5. Data Understanding

This dataset contains information about causalities of RTAs that are organized with different features such as Report Number, Road Type, Weather, Visibility, and Injury Type. Some accidents may have the same Report Number due to involving more than one person in the same accident. Also, the date and age with gender are recorded. The months written as a number vary from 1 to 12. The attribute of Cause Type mentions the main reasons for the accidents with 50 different variables. In addition, the injury type attribute used in the classification process into fatal, light, and serious.

The attribute of months reveals that the highest number of road injuries during all years that covered by the dataset under investigation occurred in January and includes about 600 injuries as shown in Figure (3.2), and June has the lowest number of injuries then followed by August and September.

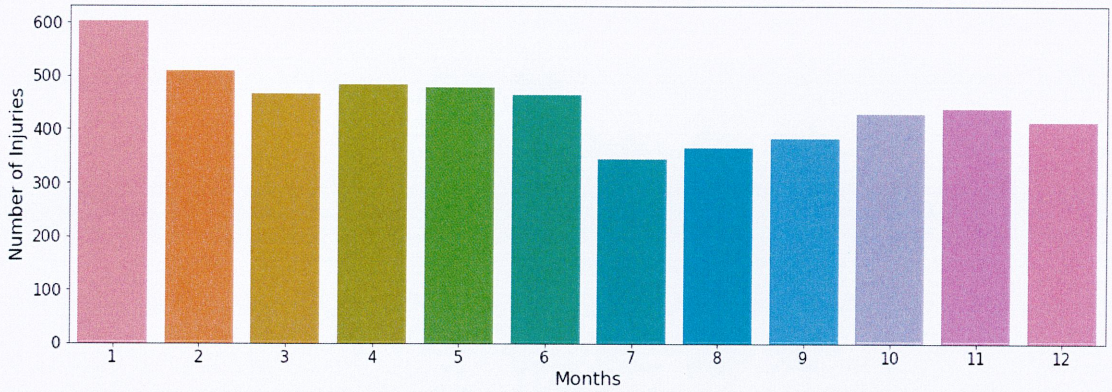


Figure (3.2): Relation Between Months and the Number of Injuries

Figure (3.3) shows that the highest number of road injuries happen in the year 2018 achieved 28.8% then followed by years 2019, 2020, 2021, and 2022 achieved 25.3%, 17.2%, 17.8%, and 10.9% respectively. by taking into consideration the year 2022 covered only injuries until June. It is observed that by increasing the years the number of injuries is reduced.

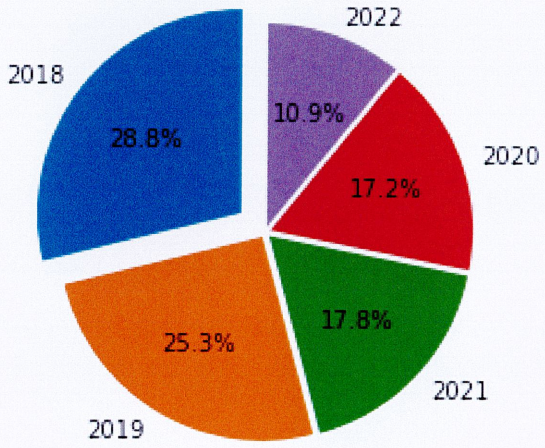


Figure (3.3): The Percentage of Injuries for each Year From 2018 to 2022

Figure (3.4) reveals there is no significant difference by focusing on the relation between the number of injuries and the day of the week.

However, the highest number of injuries happens on Thursday, and it is observed the day Tuesday has the lowest number of injuries at all.

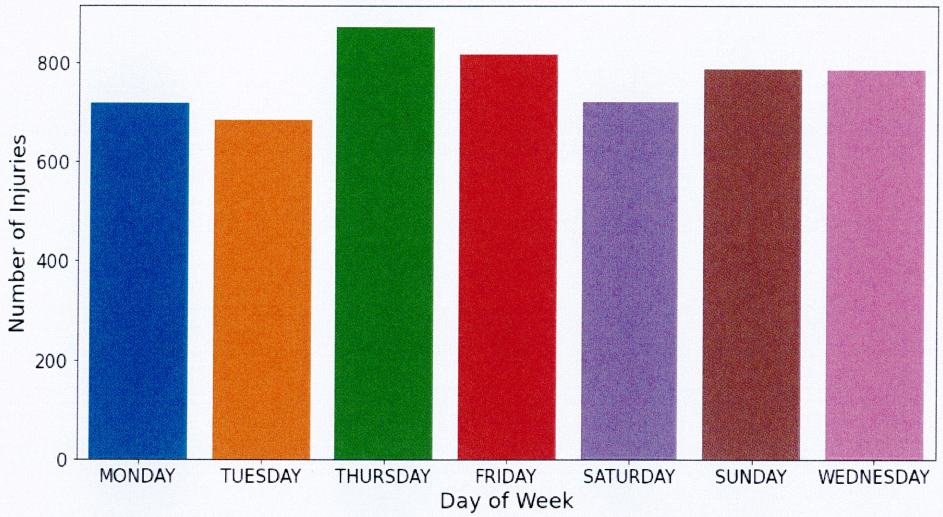


Figure (3.4): The Number of Injuries During Day of Week

The gender attribute implies that most road injuries involved males more than females with a significant difference. As shown in Figure (3.5), the percentage of injuries caused by RTAs among males reaches 71.1%, while it reaches 28.9% for injuries among females.

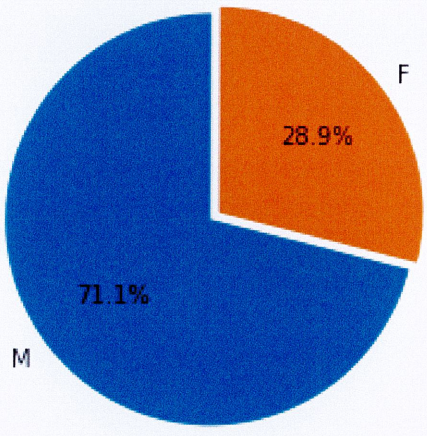


Figure (3.5): The Difference in Percentage of Road Traffic Injuries Between Gender of Male and Female

Figure (3.6) represent the percentage of injuries that occurred in the Kingdom of Bahrain in four governorates. The Capital governorate has the most injuries then followed by the Northern governorate, Southern governorate, and Moharraq governorate with 37.1%, 28.2%, 19.3%, and 15.5% respectively.

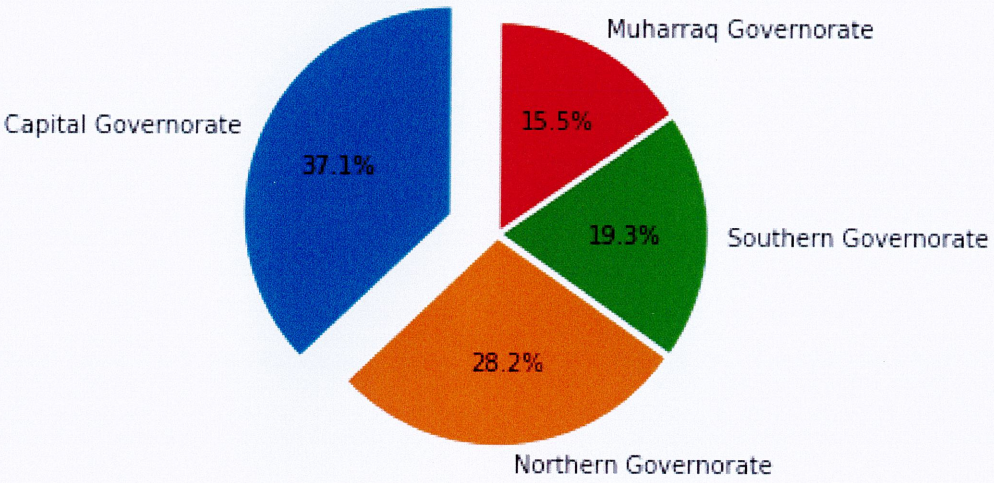


Figure (3.6): The Relation Between Number of Injuries and the Governorates at the Kingdom of Bahrain

By focusing on the type of road surface, it is observed there is big difference between the variables within this attribute. When the type of surface is concrete or asphalt the number of injuries is about 5300, while the number of injuries reaches about 45 when the type is earth or sand as presented in Figure (3.7).

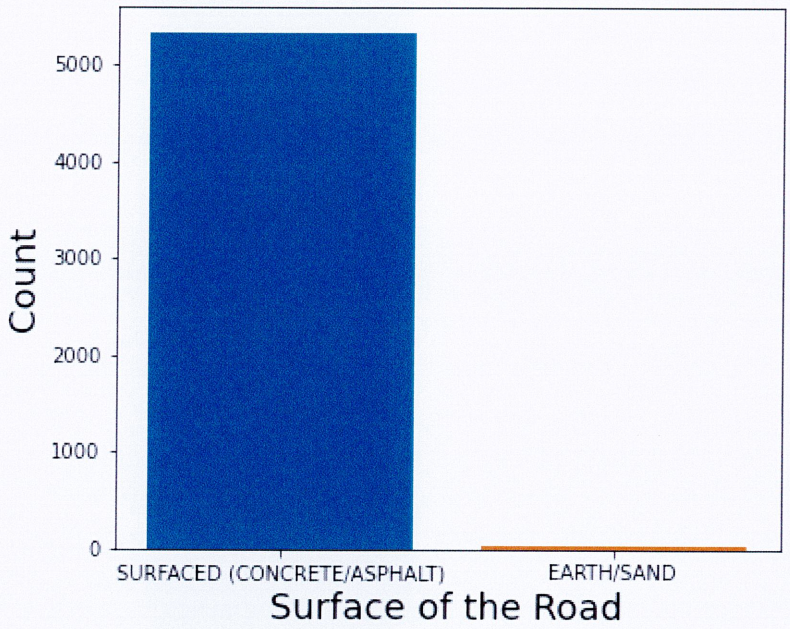


Figure (3.7): The Difference Between the Types of Surface Road

The most important reason for the accident standing on the attribute of accident junction happened when the junction is uncontrolled. In addition, when there is a policeman at the junction the number of accidents were the lowest. Figure (3.8) shows the reasons for accidents at the junctions.

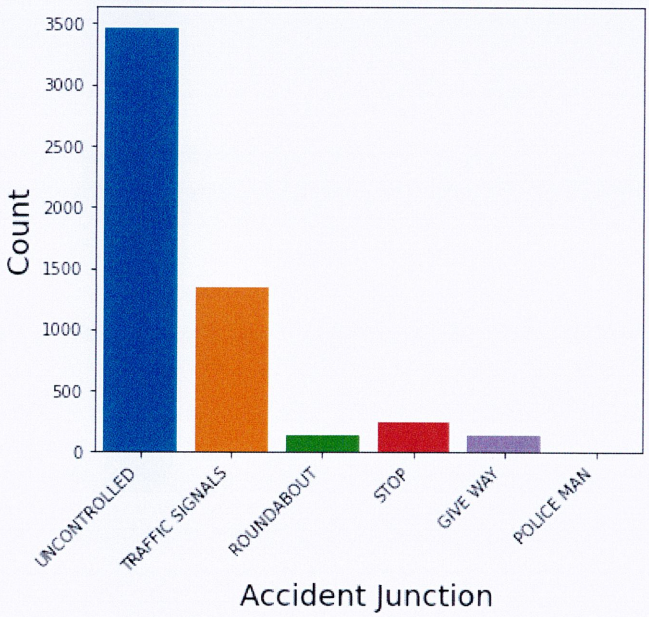


Figure (3.8): The Variate Between Variables of Accident Junction

The visibility attribute represents that most of the casualties around 2500 injuries are caused during daylight, around 2150 injuries during night-lit of streetlights, and the rest of the injuries occurred during the different degrees of visibilities, as displayed in Figure (3.9).

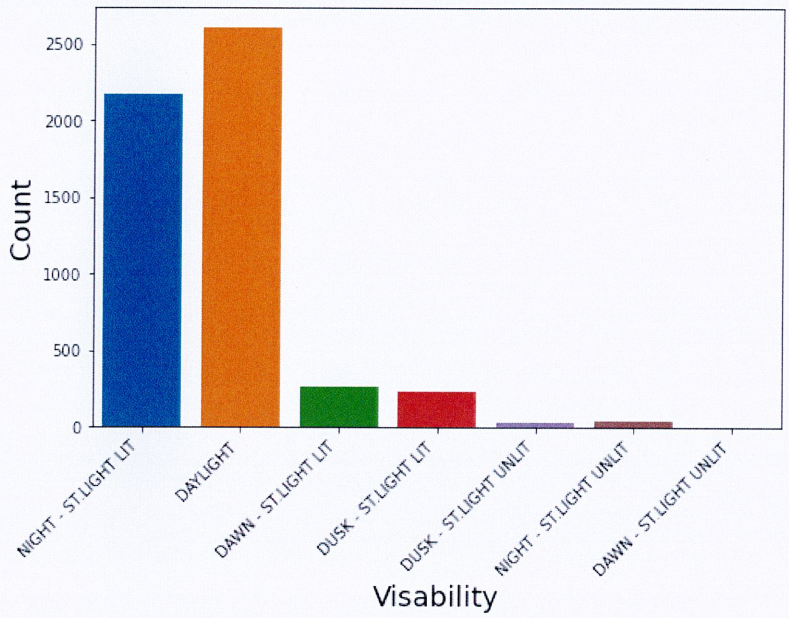


Figure (3.9): The Comparing Between Number of Injuries with Different Types of Visibility

3.6. Data Preparation

This section presents procedures that are implemented to prepare the dataset to be appropriate to use in building the models. The data cleaning was treated firstly then data structuring and finally with the feature engineering process.

3.6.1. Data Cleaning

In the data cleaning process, the attribute of the report number was deleted for its unique values for each injury involved in the same accident. The Accident Time attribute is deleted because there is a new attribute created known as Accident Hour. Since the features of day, month, and year were recorded in the data frame there is no need to keep the attribute of Accident Date.

The Location attribute is deleted for the great diversity of variables. Also, many variables in this attribute refer to the road's name and were confused with the position of the accident due to many roads across several governorates. Moreover, some values of the road's name are repeated many times in different ways with the same meaning such as (Sheikh Khalifa Street) and (Sheikh Khalifa).

Related to missing values in the dataset, the researcher dropped all rows have any missing values. Since the highest attribute has null values is Age of 2% there is no effect on the composition of the whole dataset by dropping these rows. Table (3.8) explains all attributes of the dataset containing null values with their percentages and numbers. Attribute Age has the highest number of missing values reach to 105, while the lowest attribute Visibility has only one value as missing.

Table (3.8): Details of Missing Values Relative to the Whole Datasets after Combining

Feature Name	NO. of Missing Values	Percentage of Missing Values
Age	105	2
Gender	6	0.11
Governorate	2	0.03
Road Surface	7	0.13
Road Type	12	0.22
Road Condition	6	0.11
Accident Junction	39	0.72
Site	22	0.41
Weather	2	0.03
Visibility	1	0.02

3.6.2. Data Structuring

This part explains how the attributes in the dataset are divided into categories rather than keep it in values with wide variance. For example, the month attribute with values varying from 1 to 12 has been changed to

have four labels as January-March, April-June, July-September, and October-December. The variables in the accident hour attribute converted into four categories instead of keeping them in 24 hours to be in ranges of 06:00-11:59, 12:00-17:59, 18:00-23:59, and 00:00-05:59. Regarding the age feature, there are five groups created they were < 20 years, from 20 to 29, from 30 to 39, from 40 to 49, and ≥ 50 years. The other data structuring appears in Table (3.9).

3.6.3. Feature Engineering

In the stage of feature engineering, there were several attributes that were handled to build new features. The Accident Day attribute was extracted from the attribute of Accident date to have separate features of day, month, and year. Also, the Accident Hour attribute was generated from the Accident Time attribute. Furthermore, the Road Speed attribute was added from the Road Type attribute to get a proper understanding of the dataset with building the accurate model.

3.7. Dataset Description After Handling

After completing the preparation processes, the final shape of the accident dataset covering the period from 2018 to June 2022 obtained from GDT in the Kingdom of Bahrain includes 5191 injuries caused due to RTAs and nineteen attributes. The target classes include three values they were slight, serious, and fatal.

The description of the dataset that was prepared to build models to classify the severity of RTAs is illustrated in Table (3.9). The Accident Day feature has been divided into three variables with no significant difference between them. However, the first ten days of the month have more accidents than other days of 33.7%. On the other hand, it is observed on the Accident Month attribute that the first three months from January to March have the

most injuries on the road score of 29.3%, while the months from July to September have the lowest number of injuries reach of 20.2%.

Table (3.9): Description Dataset after Preparation Process

No.	Attribute Name	Variables	Code	Number of Injuries			Total	Percentage %
				Slight	Serious	Fatal		
1	Accident Day	From 1 to 10	0	1009	660	80	1749	33.7
		From 11 to 20	1	1021	621	74	1716	33.1
		From 21 to 31	2	1003	637	86	1726	33.2
2	Accident Month	April - June	0	802	526	53	1381	26.6
		January - March	1	884	560	79	1523	29.3
		July - September	2	630	368	49	1047	20.2
		October - December	3	717	464	59	1240	23.9
3	Accident Year	2018	-	1000	471	45	1516	29.2
		2019	-	762	443	59	1264	24.3
		2020	-	479	373	55	907	17.5
		2021	-	480	399	60	939	18.1
		2022	-	312	232	21	565	10.9
4	Accident Hour	00:00-05:59	0	465	300	68	833	16.0
		06:00-11:59	1	652	388	46	1086	20.9
		12:00-17:59	2	946	616	44	1606	30.9
		18:00-23:59	3	970	614	82	1666	32.1
5	Accident Day of Week	Friday	0	453	292	45	790	15.2
		Monday	1	419	238	32	689	13.3
		Saturday	2	382	263	38	683	13.2
		Sunday	3	433	299	30	762	14.7
		Thursday	4	496	315	32	843	16.2
		Tuesday	5	410	236	25	670	12.9
		Wednesday	6	440	275	314	754	14.5
6	Governorate	Capital Governorate	0	1117	705	94	1916	36.9
		Muharraq Governorate	1	498	279	31	808	15.6
		Northern Governorate	2	868	527	72	1467	28.3
		Southern Governorate	3	550	407	43	1000	19.3
7	Gender	Female	0	1074	405	31	1510	29.1

		Male	1	1959	1513	209	3681	70.9
8	Age	< 20 years	0	530	307	24	861	16.6
		From 20 to 29	1	780	454	55	1289	24.8
		From 30 to 39	2	827	573	78	1478	28.5
		From 40 to 49	3	464	284	40	788	15.2
		≥ 50 years	4	432	300	43	775	14.9
9	Nationality	Bangladesh	0	250	304	46	600	11.6
		Egypt	1	38	27	2	67	1.3
		India	2	255	199	32	486	9.4
		Kingdom of Bahrain	3	2065	1068	111	3244	62.5
		Other	4	177	107	21	305	5.9
		Pakistan	5	93	124	17	234	4.5
		Philippines	6	53	38	5	96	1.8
		Saudi Arabia	7	52	35	5	92	1.8
		Yemen	8	50	16	1	67	1.3
10	Road Speed	100 KPH	0	797	411	96	1304	25.1
		50 KPH or LESS	1	1036	760	55	1851	35.7
		60-80 KPH	2	1200	747	89	2036	39.2
11	Road Surface	EARTH/ SAND	0	11	14	2	28	0.5
		Surfaced (Concrete/ Asphalt)	1	3022	1904	238	5164	99.5
12	Road Condition	Dry	0	2961	1866	239	5066	97.6
		Wet	1	72	52	1	125	2.4
13	Site	Crossroad	0	608	323	30	961	18.5
		Section	1	1684	1036	137	2857	55.0
		Sliproads	2	104	105	29	238	4.6
		T-Junction	3	637	454	44	1135	21.9
14	Accident Junction	Give Way	0	84	54	4	142	2.7
		Roundabout	1	86	45	1	132	2.5
		Stop	2	127	104	8	239	4.6
		Traffic Signal	3	812	443	51	1306	25.2
		Uncontrolled	4	1924	1272	176	3372	65.0
15	Weather	DRY & SUNNY	0	1139	687	66	1892	36.4
		Dust or Sundstrom	1	30	18	1	49	0.9
		Fog	2	5	8	1	14	0.3

		Glaring Sunshine	3	161	110	8	279	5.4
		High Humidity	4	11	10	1	22	0.4
		Others	5	1658	1060	162	2880	55.5
		Rain	6	29	25	1	55	1.1
16	visibility	Day	0	1675	1028	1676	2807	54.1
		Night	1	1358	891	135	2384	45.9
17	Road Impact	Impact	0	834	480	100	1414	27.2
		No Impact	1	2199	1438	140	3777	72.8
18	Person Involved	Driver	0	1309	978	120	2407	46.4
		Passenger	1	1093	403	50	1546	29.8
		Pedestrians	2	631	537	70	1238	23.8
19	Cause Type	Changing Lanes / Turning Carelessly	0	651	473	51	1175	22.6
		Collision With Pedestrian While Driving Carelessly (Driver)	1	431	314	42	787	15.2
		Drunk / Drugged	2	124	116	27	267	5.1
		Falling to Comply Traffic Sign	3	567	308	26	901	17.4
		Following Too Closely Behind Vehicle	4	499	172	13	684	13.2
		Ill / Disability	5	66	29	4	99	1.9
		Losing Control Before Accident	6	134	125	13	272	5.2
		Other Errors of Judgment	7	307	152	12	471	9.1
		Speed Too Great for Prevailing Cond	8	170	118	39	327	6.3
		Walking / Running on Road	9	84	111	13	208	4

The relation between injuries and hours was exhibited in Figure (3.10). There are 1666 injuries recorded in the period of 18:00-23:59 and there are 1606 injuries in 12:00-17:59, but there are only 833 injuries through the time of 00:00-05:59.

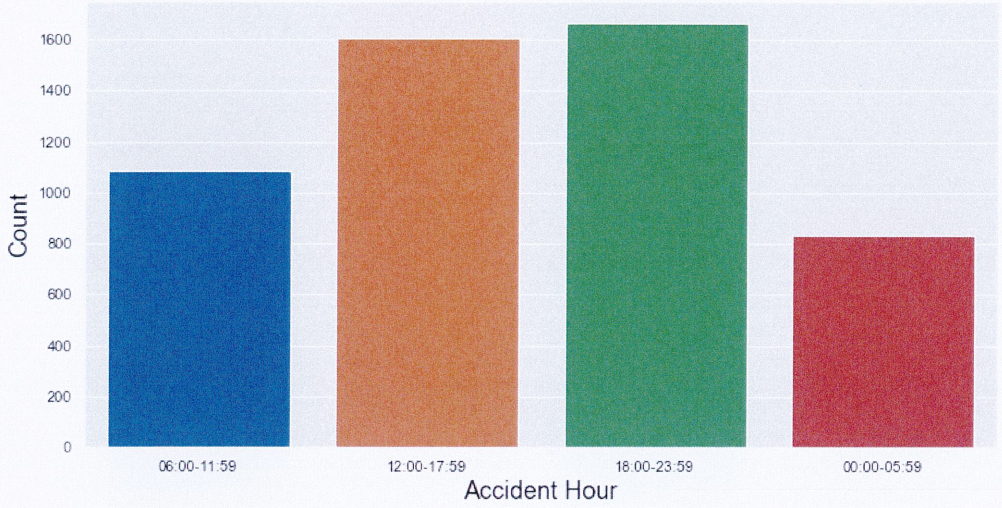


Figure (3.10): The Relation Between Accident Hour and the Number of Injuries

The age from 30 to 39 has the highest number of injuries on the road of 28.5% followed by the age from 20 to 29 is 24.8% and the lowest number of injuries was recorded when the age equal or more than fifty years old of 14.9% as shown in Figure (3.11).

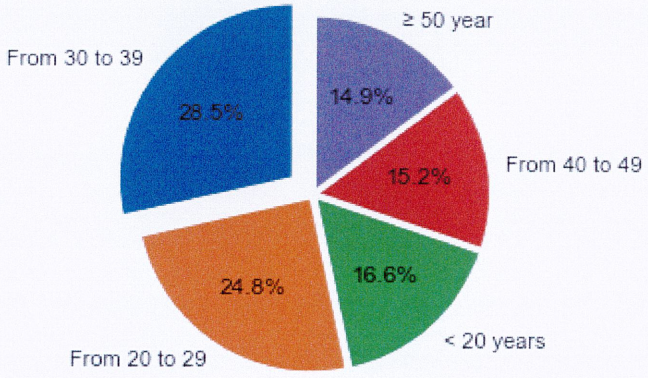


Figure (3.11): The Age Attribute in The RTAs Dataset

Figure (3.12) shows that the highest performance of injuries occurred when the road speed is between 60 and 80 kilometers per hour and the lowest number of injuries were found on roads with a speed of 100 kilometers per hour.

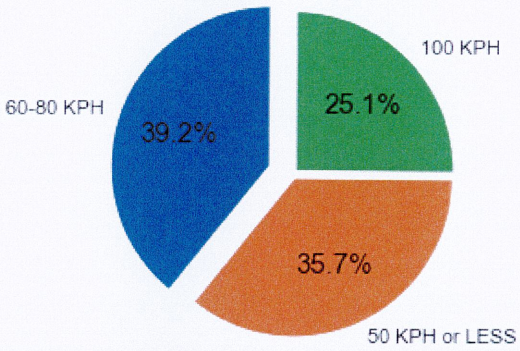


Figure (3.12): Road Speed Attribute of the Dataset Under Investigation

Figure (3.13) presents an overview of the classes of the target attribute where there is a big difference between the variables. It is observed that the injuries number of serious, slight, and fatal reaches 3033, 1918, and 240 respectively.

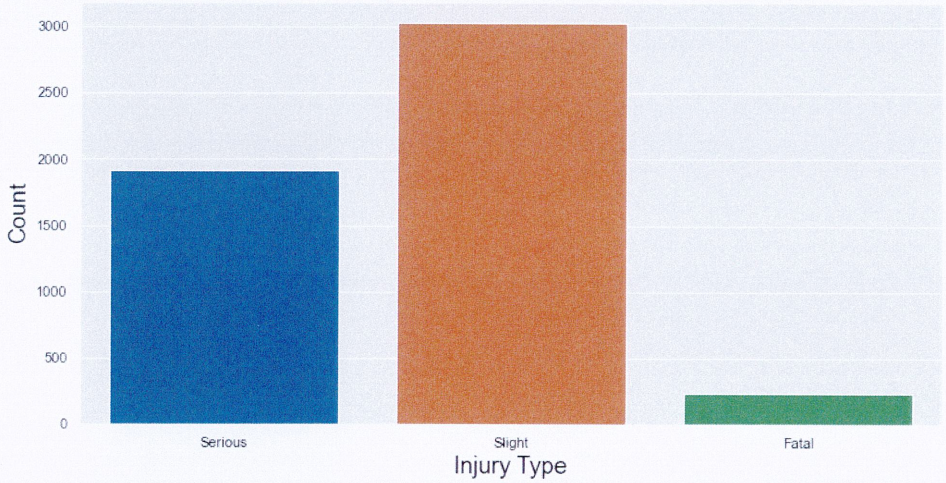


Figure (3.13): The Variables of Injury Type as Target Attribute in RTAs Dataset

3.8. Data Splitting

The dataset of RTAs was divided into two parts, the first one of 70% as a training set and the second one of 30% as a testing set based on previous studies. The training dataset includes 3633 instances used to build the models, while the testing dataset includes 1558 instances used to evaluate the models. In addition, the random seed was used to ensure the results are reproducible.

3.9. Resampling Methods

By focusing on the dataset under investigation, it is observed that the target classes were imbalanced. The ratio of target classes was 36.9% for the Serious, 58.4% for the Slight, and 4.6% for Fatal. After balancing data using over-sampling methods, the variables became equal for all classes due to raising minority percentages to the same level as majority percentages, and using the under-sampling method makes the ratio of classes the same, because the majority percentage is decreased to equal the minority.

3.10. ML Models Used

This study used the algorithms RF, KNN, SVM, NB, and DT to classify the severity of RTAs. RF and DT algorithms were used for their high performance in previous studies. Furthermore, the main reason for choosing KNN and NB was simplicity. Moreover, KNN is a nonparametric algorithm, so it does not require any assumptions to implement it. Additionally, SVM is used for its ability to manage both non-linear and linear datasets. However, all these algorithms are compatible to solve classification problems as this research discusses.

3.11. Evaluation Methods

According to this study, four metrics are used to determine the performance of classification algorithms. These metrics are F-measure, accuracy, recall, and precision. To deal with these criteria it is important to calculate the True Negative Rate (TNR), True Positive Rate (TPR), False negative Rate (FNR), and False Positive Rate (FPR). Furthermore, these terminologies are important to build the Confusion Matrix. The formulas and concepts related to the metrics under investigation are presented in section 3.11.1. Interested readers can refer to these two references of Eibe, et al. (2017) and Giuseppe (2018) to get more information about critical methods.

3.11.1. Accuracy

ML accuracy refers to how many correct classifications a trained model achieves (7). In accuracy, the degree of closeness is measured to the true value.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{7}$$

3.11.2. Recall

The ratio between the number of positive samples correctly predicted as positive and all number of positive samples used to calculate Recall (8). Recall finding the ability of the algorithm to discover positive samples.

$$\text{Recall} = \frac{TP}{(FN+TP)} \tag{8}$$

3.11.3. Precision

Precision is the proportion of correct positive samples to the whole number of positive samples either incorrectly or correctly (9). Precision presents the reliability as positive in the model.

$$Precision = \frac{TP}{(\mathrm{FP+TP})} \tag{9}$$

3.11.4. F-measure

To calculate F-measure need to find the average of Recall and Precision (10). It considered FNs and FPs into account.

$$\mathrm{F-measure} = \frac{2(\mathrm{recall}*\mathrm{precision})}{(\mathrm{recall+precision})} \tag{10}$$

3.11.5. Area Under the Curve (AUC)

The AUC is an evaluation method used to measure classification accuracies. By using AUC, the overall performance is calculated by taking all thresholds of classifiers, where the AUC equal one is the best classifier.

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{11}$$

3.12. Conclusion

This chapter describes the methodology of the research that covers the steps needed to prepare a dataset to find the best algorithm that can classify the severity of injuries of RTAs.

The first step of this chapter explains the ways to deal with five datasets obtained from the statistic section of the GDT in the Kingdom of Bahrain. These datasets cover the period from 2018 to June 2022. There was an explanation provided for each set of data containing a description of the number of attributes, rows, and the percentage of the missing values.

The next stage presents the visual description for the dataset after combining the five datasets in one data frame to understand the nature of attributes and discover the most important variables that affect road injuries.

In the following step, to create the final form of the dataset implemented the several stages of preparation include explaining the reason behind deleting unnecessary attributes, describing the changes that happen for variables of attributes, and presenting the new features created to obtain the finalized dataset.

The next step was applied to give a deep description of the dataset related to the research problem after handling and preparing the final shape of the dataset by using tables, bar charts, and pie charts to provide information about the variables for each attribute.

The final stage mentioned algorithms of KNN, RF, DT, NB, and SVM that wer used in the next chapter, and explained the critical methods of Recall, Precision, F-measure, Accuracy, and AUC to evaluate the model

Chapter Four

Results and Discussion

4.0. Introduction 66

4.1. Tools Used 66

4.2. K-Nearest Neighbor (KNN) Model 66

4.3. Random Forest (RF) Model 70

4.4. Naïve Bayes (NB) Model 73

4.5. Support Vector Machine (SVM) Model 77

4.6. Decision Tree (DT) Model 81

4.7. Best Models Comparison 84

4.8. Feature Selections 87

4.9. Conclusion 89

Chapter Four

Results and Discussion

4.0. Introduction

This section presents the results for NB, KNN, SVM, DT, and RF algorithms using several evaluation methods such as accuracy and recall, which helps identify the best classifier appropriate for the dataset under investigation, and to predict the accident's severity whether it is slight, serious, or fatal. Additionally, this part demonstrates the difference between imbalanced data and resampling methods. Also, it identifies the most important features that affect road accidents.

4.1. Tools Used

In this study, the researcher used the test dataset to build the models using Python 3.8.8 in Jupyter Notebook with all CPUs in most operations. Scikit-learn is used to compute the criteria for KNN, NB, SVM, DT, and RF algorithms. Moreover, Matplotlib is utilized in the graphical and visualization of the results obtained in this research.

4.2. K-Nearest Neighbor (KNN) Model

One of the algorithms used in this study to analyze the injuries by road traffic in the Kingdom of Bahrain is the KNN algorithm. This part provides the results of critical methods for the mentioned model that was found by the researcher with different resampling methods, then gives an overview to compare them. Also, a discussion of the way to get the best results by using this algorithm is discussed in this section considering that the type of distance used between points is Minkowski.

4.2.1. Best Value of K

To choose the best value of K in the KNN algorithm, the researcher used two methods. The first one considered the value of K equals 39 which is found by calculating the square root for the number of total casualties that exist in this dataset. The second one is done by using the method of trial and error until finding the high results of evaluation methods.

4.2.2. KNN with Resampling Methods

Table (4.1) compares the results of the KNN algorithm by using two types of resampling methods: under-sampling, and over-sampling with imbalanced data, by using different values of K.

The result related to the KNN classifier achieved the best values of metrics for imbalanced data when K = 39, for under-sampling when K =25, and for over-sampling when K = 3. The researcher found the over-sampling method when K=3 achieved the highest results for all metrics precision, recall, F-measure, and accuracy of 74%, 75%, 74%, and 75% respectively, and observed significant differences by comparing to other methods. In contrast, the lowest values were found with the under-sampling method when K=3 for accuracy, F-measure, recall, and precision of 42%, 41%, 42%, and 42% respectively.

comparing this study and the study done by (Labib, et al., 2019) which approximately used the same factors, it is found there is no significant difference between the best results obtained through building the KNN model; this may be referred to the fact that both datasets concern road accidents.

Table (4.1): The Performance of KNN Algorithm by using Different Resampling Methods

K-Nearest Neighbors (KNN) Algorithm					
Resampling Methods		Metrics (%)			
		Precision	Recall	F-Measure	Accuracy
Imbalanced Data	K=39	55	61	56	61
	K=3	57	57	57	57
	K=25	52	58	54	58
Under-sampling	K=15	43	42	42	42
	K=3	42	42	41	42
	K=25	43	43	42	43
Over-sampling	K=55	52	52	51	52
	K=3	74	75	74	75
	K=25	55	56	54	56

Figure (4.1) illustrated the confusion matrices related to the KNN algorithms with the best value of K for each type of resampling method to explain the behavior of this model with the target classes of the dataset under study.

The KNN algorithm by using imbalanced data with K=39 appears high predictions for the slight class, poorly for the serious class, and there is no correct prediction for the fatal class as presented in shape (1). Shape (2) presents the predictions of the KNN classifier by using the under-sampling method when K=3, the worst prediction was achieved with the slight class, but the prediction of the serious class was better than the previous one, and the best prediction is done for fatal class. Meanwhile, the over-sampling method when K=3 obviously has the highest prediction of fatal class, and regarding classes of serious and slight the classification results are approximately the same each other.

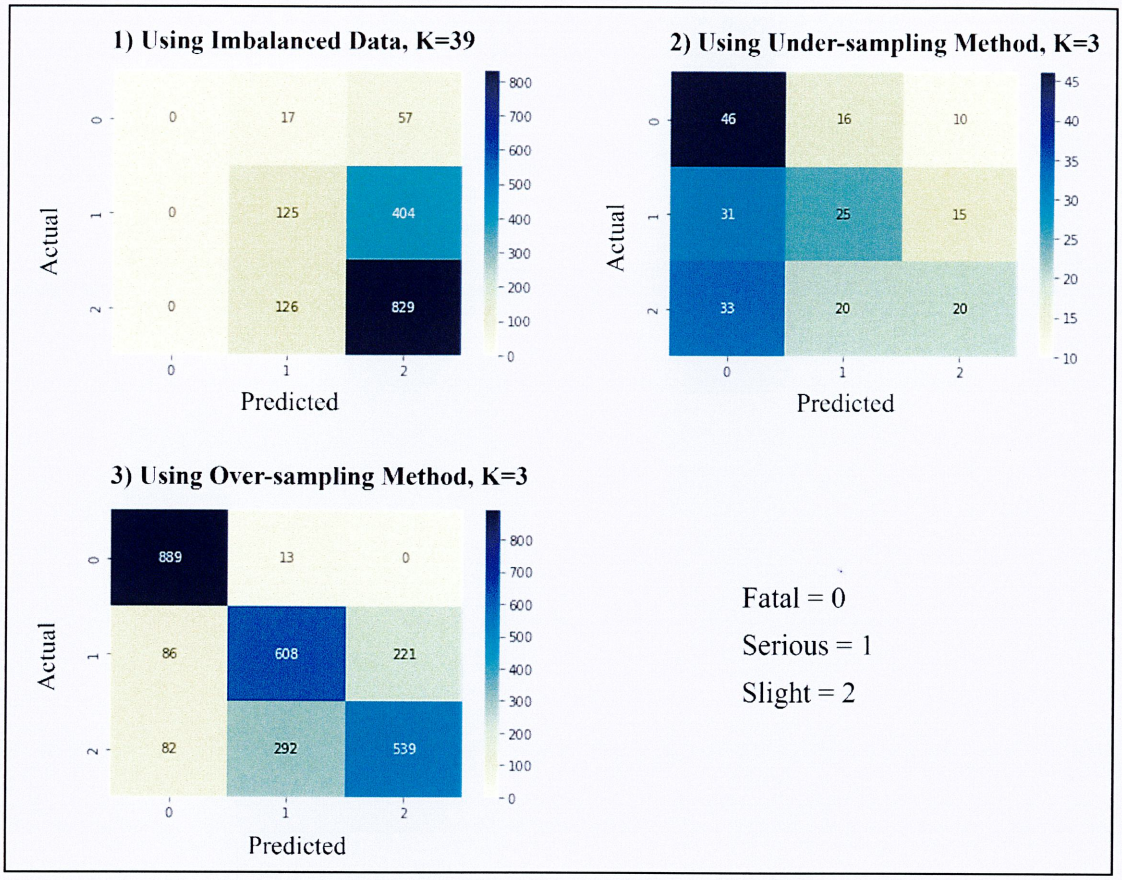


Figure (4.1): Confusion Matrix for the KNN Models of the RTAs Classes

Figure (4.2) presents the results and graphs of the AUC related to the KNN algorithms by using the best value of K for each resampling method. It is clearly noticeable that the class of fatal scored the highest value of AUC with imbalanced data, Over-sampling, and Under-sampling methods of 62%, 98%, and 65% respectively. Meanwhile, the class of slight scored the worst results of AUC with the Over-sampling of 82%, and the Under-sampling of 57%. Moreover, the highest values of AUC for all classes were achieved with significant differences by using the Over-sampling method when K equals 3.

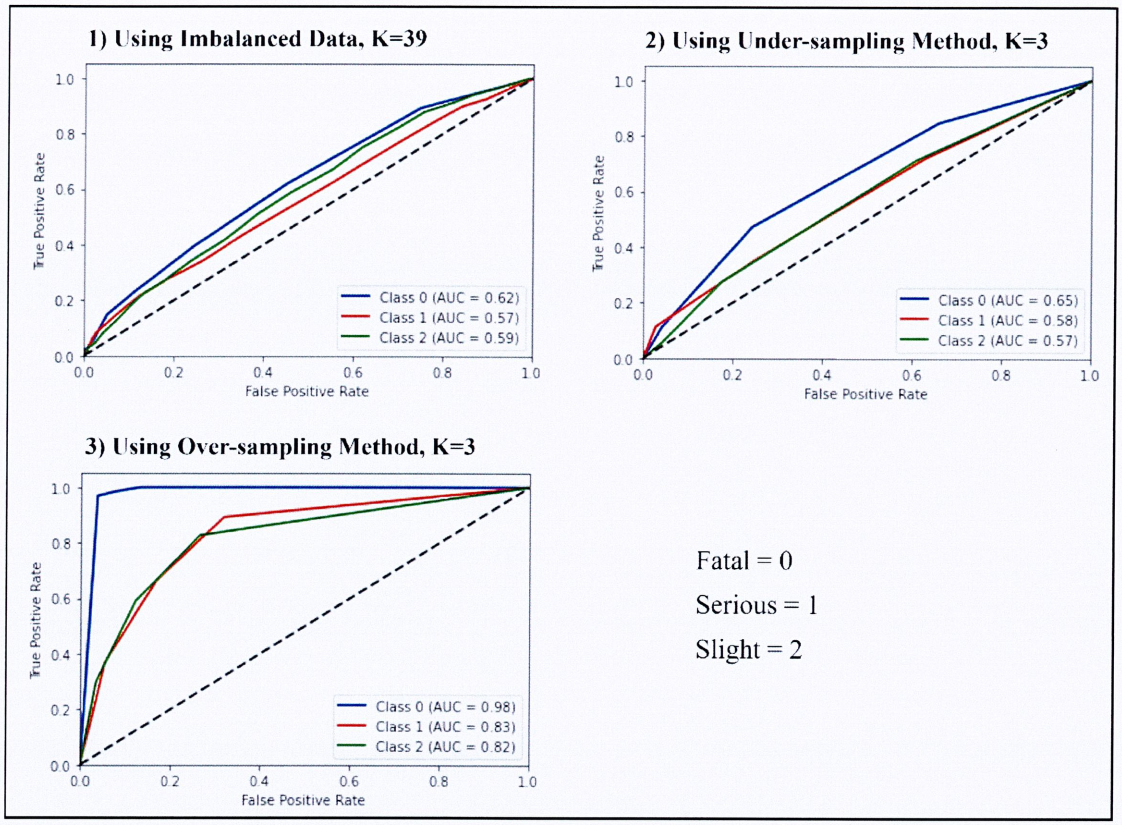


Figure (4.2): AUC for the Models of KNN Related to the RTAs Classes

4.3. Random Forest (RF) Model

This part illustrates all results found with different types of evaluation methods related to the algorithm of RF and then compares them. Also, a discussion of the ways of choosing the optimal parameters and deciding the best type of resampling method takes place in this section.

4.3.1. Parameters of RF

The two most important parameters that are considered through working on the RF algorithm are the number of estimators (trees) and the number of features. The estimator number used is 150 based on its best results by using trial and error method. Related to feature numbers, it's always permuted randomly in each split. However, since this algorithm needs time in the running process to build the model, the researcher used all processes that exist in the machine.

4.3.2. RF with Resampling Methods

For models of RF, the results of evaluation methods were presented in Table (4.2) with different types of resampling methods. The best value of accuracy for the RF algorithm was achieved with the over-sampling method of 85%. The researcher found the over-sampling method achieved the highest result for all metrics with a significant difference. In contrast, 50% was the worst accuracy that RF gave with the under-sampling method. The lowest values of the RF algorithm for all metrics were achieved with the method of under-sampling. This insufficient learning happens for the decreased number of instances resulting from the method of under-sampling.

Furthermore, it is observed by using the under-sampling method the precision was reduced by 8% compared to imbalanced data. The value of the metric recall with the over-sampling method achieved a noticeable increase of 24% from 61% to 85%. Also, the other metrics of precision, F-measure, and accuracy result in higher values in the method of over-sampling compared to the original dataset. It is clearly noticed that the criteria result of RF models can be changed by removing and adding data. In other words, by using the over-sampling method the results increased and by using the under-sampling method the results decreased.

Table (4.2): The Performance of RF Algorithm by using Different Resampling Methods

Random Forest (RF) Algorithm				
Resampling Methods	**Metrics (%)**			
	Precision	**Recall**	**F-Measure**	**Accuracy**
Imbalanced Data	58	61	59	61
Under-sampling	50	50	50	50
Over-sampling	85	85	85	85

The confusion matrices obtained within building the RF models by using several resampling methods are shown in Figure (4.3). By focusing on the main diagonal in shape (1), it is observed that the best class predicted correctly is slight and the worst one is fatal when the dataset used is imbalanced. The predictions for class fatal outperform the rest in the under-sampling method as presented in shape (2). Moreover, the best confusion matrix in terms of comparison between actual classes and predicted classes is the one used with the over-sampling method of the RF model as shown in shape (3).

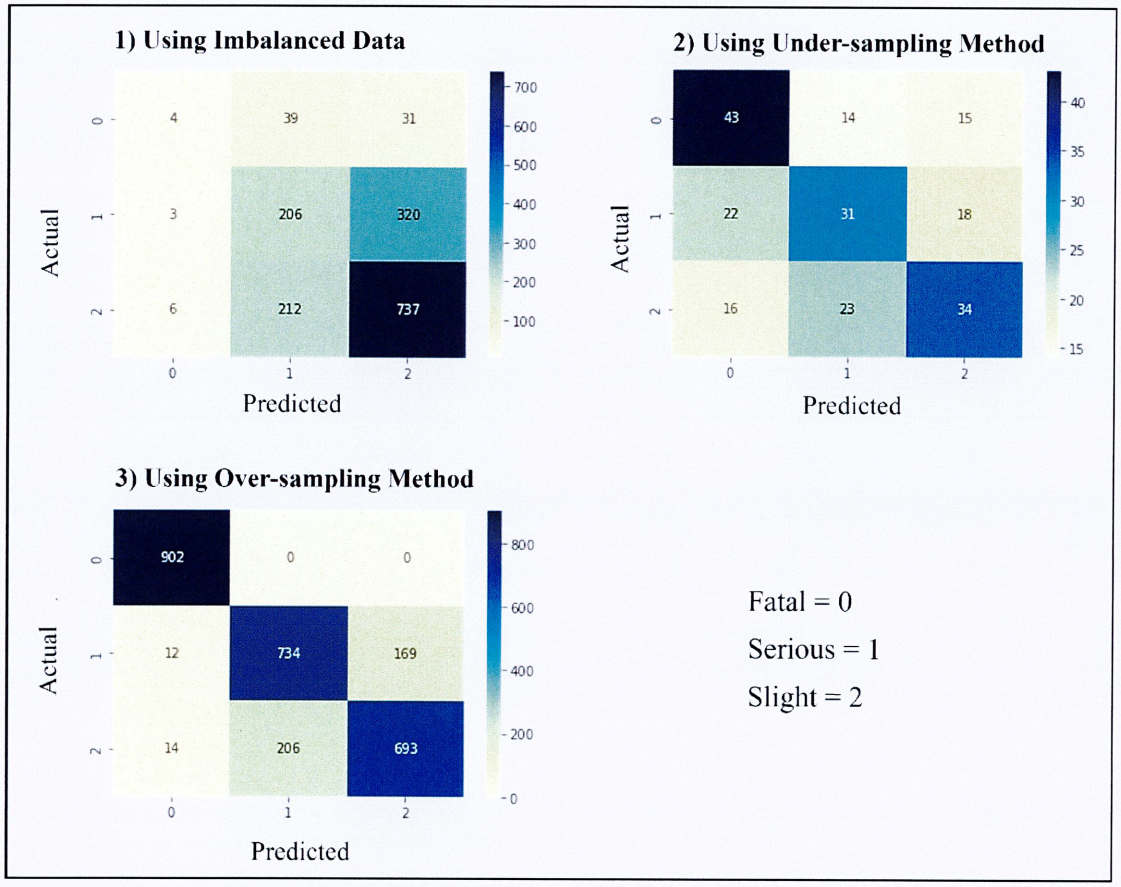


Figure (4.3): Confusion Matrix for the RF Models of the RTAs Classes

Figure (4.4) illustrates the values of AUCs related to each class of RTAs dataset by building the RF models through using several resampling methods. The best value of AUC was achieved for all resampling methods with the fatal class and the highest AUC for this class scored with the over-sampling method of 100%. The lowest results of AUC were achieved by using the under-sampling method, and the highest results were achieved significantly with the over-sampling method. Moreover, the lowest value of AUC scored in the serious class for all resampling methods. In addition, it is noticeable the results of AUC for all classes are above 50%.

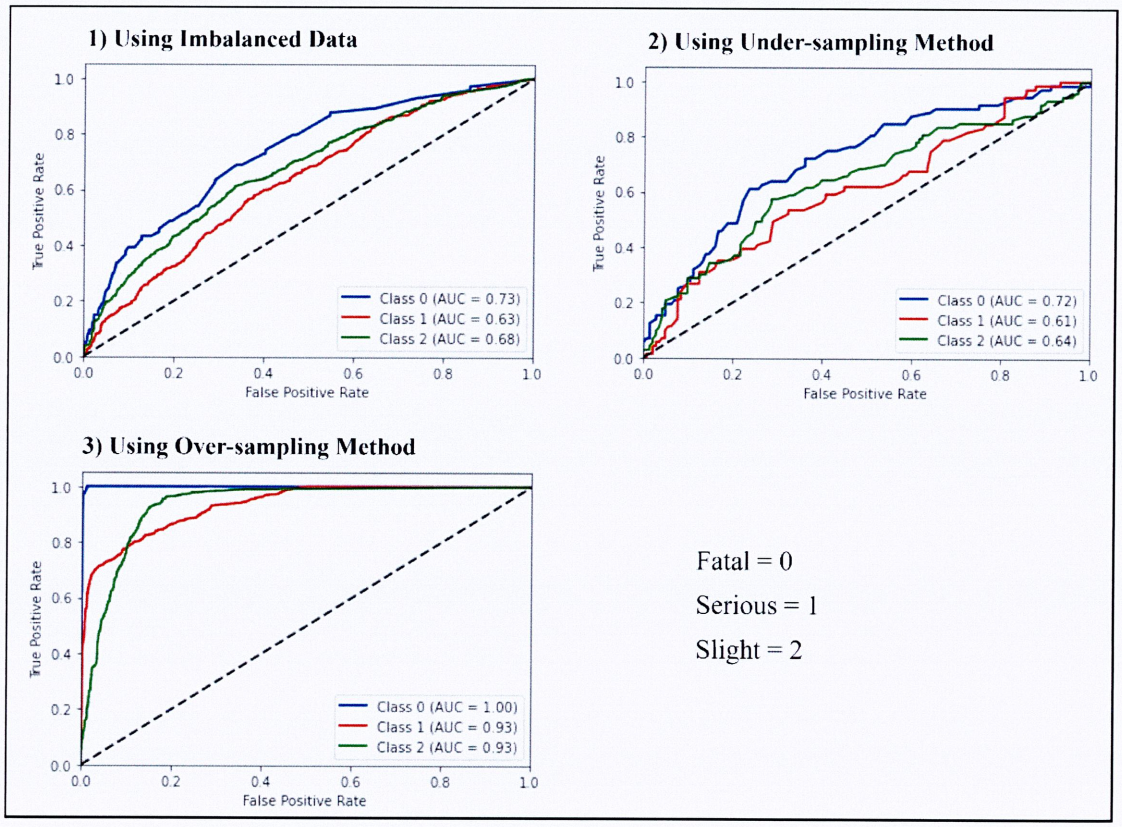


Figure (4.4): AUC for the Models of RF Related to the RTAs Classes

4.4. Naïve Bayes (NB) Model

Within this part, the study discussed the results gathered from metrics by using the NB algorithm with different types of resampling methods. Moreover, it indicates whether any resampling methods could improve the results.

4.4.1. Parameters of NB

NB is one of the algorithms that is considered a straightforward classifier due to its no need to change a lot of parameters to get the results of the evaluation methods. In addition, one of the important advantages of this algorithm is that it didn't spend so much time in the classification process. In this section, all parameters used with the model of NB were default.

4.4.2. NB with Resampling Methods

Table (4.3) shows the three types of resampling methods used to find the results for evaluation methods of the NB algorithm. Firstly, by using the dataset without any resampling method the precision scored 63% and 10% for other metrics. Secondly, by using the under-sampling method the precision decreased to 35% and the left metrics increased according to imbalanced data. Thirdly, regarding the over-sampling method, the other metrics improved again except the precision which decreased compared to imbalanced data.

The worst value of accuracy with a significant difference was achieved when there was no change in the dataset of 10%. There is a slight difference in accuracy between over-sampling and under-sampling, the first one has 45% and the second one has 40%. The worst result for F-measure was with imbalanced data, while there is a slight difference of 7% between under-sampling and over-sampling methods. The worst value of 10% was achieved for metrics of recall when the data is imbalanced, meanwhile, the other resampling methods approximately achieved the same results. In addition, it's observed that the values of evaluation methods improve by using a balanced dataset, this is consistent with the findings of (Cigdom & Ozden, 2018).

Generally, there is a significant difference between the results of resampling methods and the imbalanced dataset. For that, it is not recommended to use the NB algorithm for this type of dataset that is related to RTAs without using any kind of resampling method due to the unsatisfactory results.

Table (4.3): The Performance of NB Algorithm by using Different Resampling Methods

Naïve Bayes (NB) Algorithm				
Resampling Methods	**Metrics (%)**			
	Precision	**Recall**	**F-Measure**	**Accuracy**
Imbalanced Data	63	10	10	10
Under-sampling	35	40	32	40
Over-sampling	47	45	39	45

The confusion matrix results of the NB algorithms are shown in Figure (4.5). The poor prediction appeared in shape (1), and there are 879 instances of slight class classified incorrectly as fatal when the researcher used a dataset without change known as the imbalanced dataset. On the other hand, over-sampling and under-sampling methods appear more improvement to classify the classes correctly, especially with fatal and slight classes. Also, it is observed that there are a lot of serious classes classified incorrectly as fatal with the last two resembling methods.

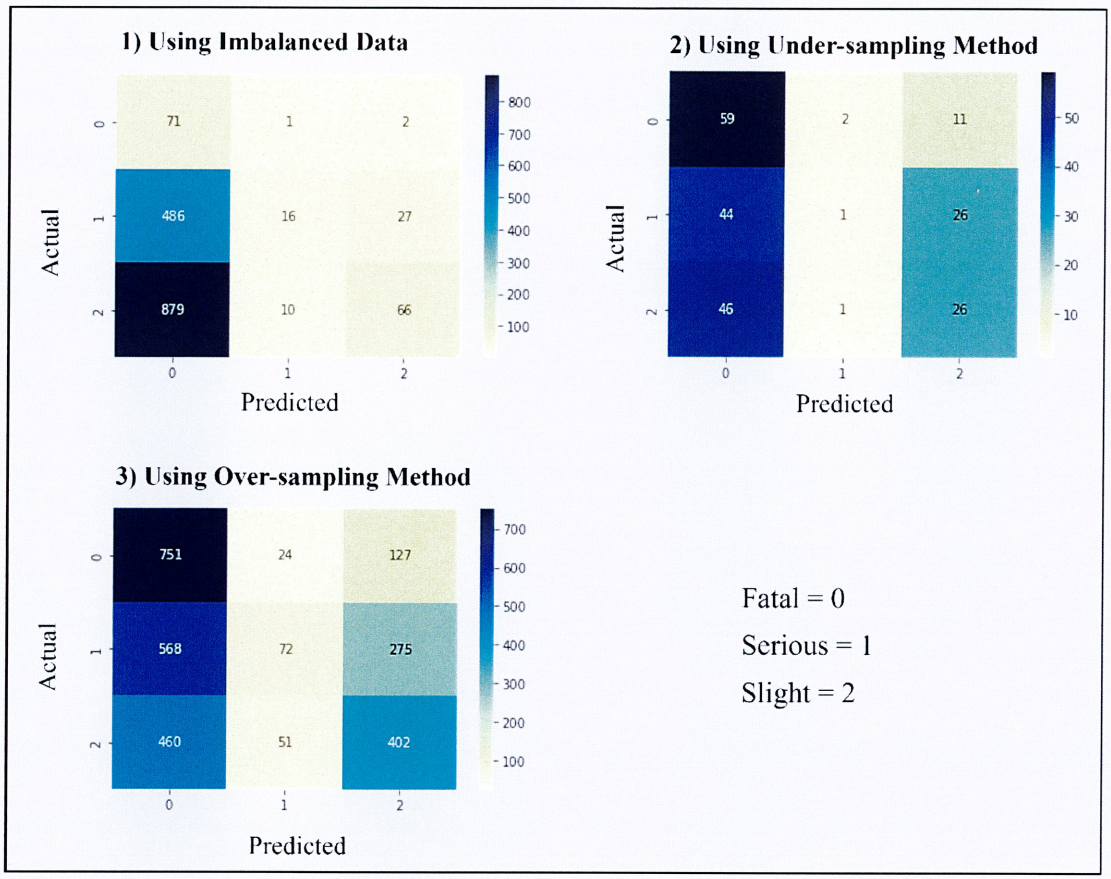


Figure (4.5): Confusion Matrix for the NB Models of the RTAs Classes

Figure (4.6) shows the area under the curve for the classes target of the dataset under investigation by using three types of resampling methods associated with the NB algorithm. There is no significant difference between the results of AUC for resampling methods used, but the highest values of AUC found with the over-sampling method are 71% for the fatal class, 60% for the serious class, and 66% for the slight class. The contrast, the lowest results of AUC were obtained through building the RF algorithm while using imbalanced data, which classes fatal, serious, and slight scored percentages of 65%, 46%, and 60% respectively.

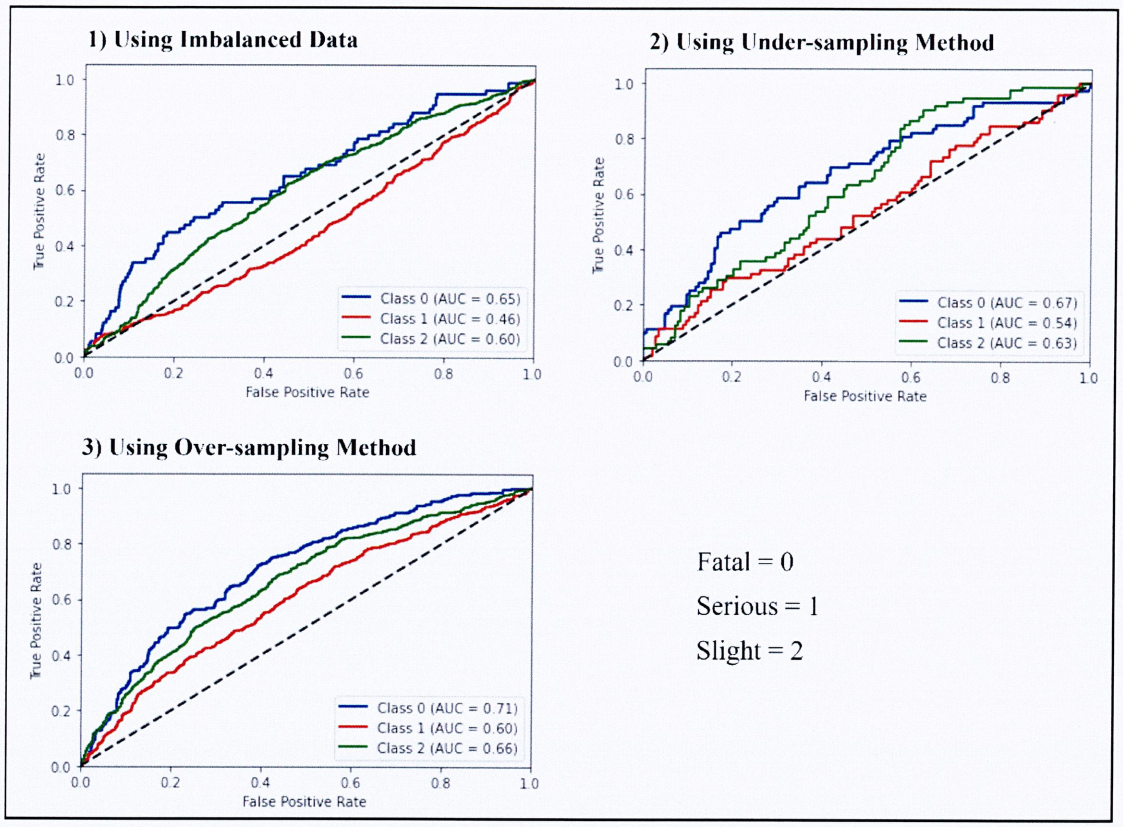


Figure (4.6): AUC for the Models of NB Related to the RTAs Classes

4.5. Support Vector Machine (SVM) Model

Through this section, the researcher presents the results of running the model for SVM algorithms. These results cover evaluation methods of precision, F-measure, accuracy, and recall.

4.5.1. Parameters of SVM

There are several parameters that could be used with the algorithm of SVM such as Regularization and Gamma but there is no significant change occurred in the results of evaluation methods. In contrast, by changing the Kernel function from RBF into linear the metric of precision increased from 38% to 60%. For that, the important parameter with the algorithm of SVM by using the dataset under investigation is Kernel. In addition, the polynomial degree used during building the SVM model is three.

4.5.2. SVM with Resampling Methods

Table (4.4) shows the results of different types of metrics with resampling methods for the SVM algorithm by fixing the kernel to be linear and other parameters used as default. The best result for the precision of 60% was achieved when there was no change in the dataset, while the worst result of 38% for the under-sampling method which means there is a significant difference between them. The highest value for recall 61% scored with imbalanced data. Also, superior results were achieved by keeping the dataset without change for the metrics F-measure and accuracy of 47% and 61% respectively.

In addition, the metric of F-measure achieved the lowest value compared to other metrics with under-sampling method, this is incompatible with the results found in some previous studies with the same field of dataset such as the study (Yassin, 2020). However, the best result found between the resampling method is imbalanced data. Meanwhile, the lowest values were found for the method of under-sampling.

Whatever, there are noticeable changes for resampling methods with all metrics except F-measure has no change between imbalanced data and over-sampling. Accordingly, the under-sampling method didn't achieve the highest score in any evaluation method due to the SVM algorithm being unsuitable when the dataset is small.

One possible reason for the drop in the performance of SVM after the use of over-sampling and under-sampling is related to overfitting, which the resampling methods could affect the overfitting to be increased. Also, building the SVM by using the under-sampling method could affect the loss of main information because some instances are removed. On the other hand, using the over-sampling method led to adding noise to the original dataset because some instances were duplicated.

Table (4.4): The Performance of SVM Algorithm by using Different Resampling Methods

Support Vector Machine (SVM) Algorithm				
Resampling Methods	**Metrics (%)**			
	Precision	**Recall**	**F-Measure**	**Accuracy**
Imbalanced Data	60	61	47	61
Under-sampling	38	38	38	38
Over-sampling	48	47	47	47

Figure (4.7) summarizes in the confusion matrix the predictions of SVM algorithms by classifying the variables of the target with three resampling methods. The confusion matrix of imbalanced data for the SVM algorithm classified the slight class more correctly than other classes and failed to classify any class of fatal correctly. The Confusion matrix of under-sampling methods appeared good predictions for the class of fatal, and the classes of slight and serious predicted the same number of instances. According to the over-sampling method, the confusion matrix presents that the highest class classified correctly is fatal, meanwhile, the lowest one is slight. Also, there are 240 instances classified wrongly as fatal rather than serious.

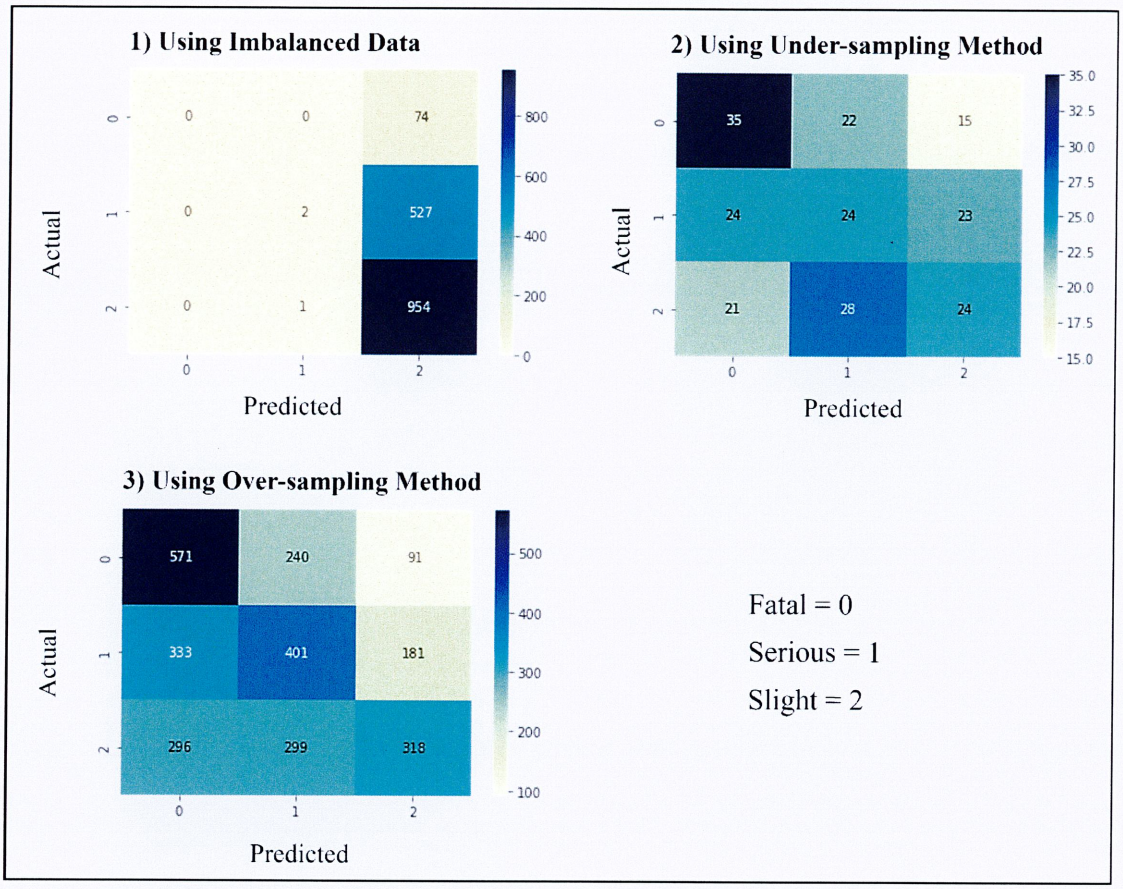


Figure (4.7): Confusion Matrix for the SVM Models of the RTAs Classes

Figure (4.8) shows the values of AUC for the target classes of RTAs dataset through building the SVM models by using different types of resampling method. The best results of the AUC were achieved with the over-sampling method, meanwhile the worst results were found by building the SVM model without making any change to the dataset.

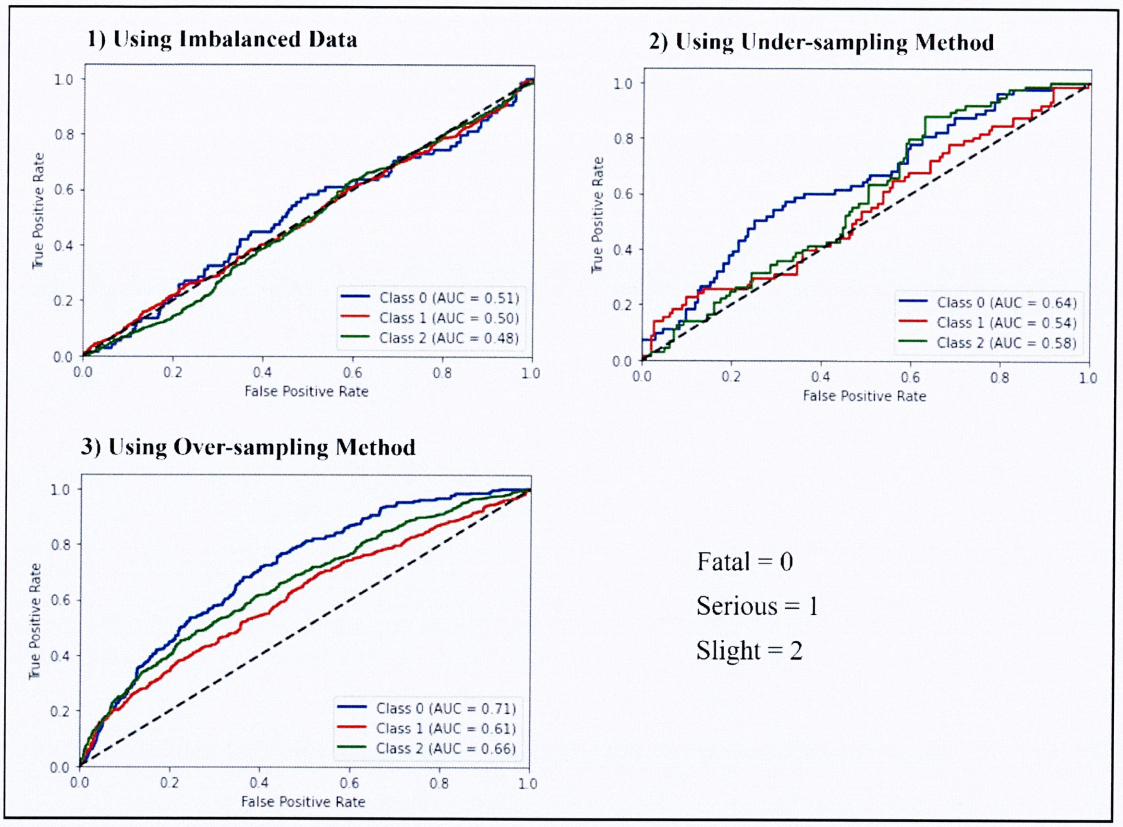


Figure (4.8): AUC Values for the SVM Models Related to the RTAs Classes

4.6. Decision Tree (DT) Model

During this section, the researcher finds the best results of evaluation methods by trying different methods of resampling related to the dataset of road accidents in the Kingdom of Bahrain from 2018 to 2022. These experiments lead to an understanding of how the DT algorithm can deal with imbalanced data, over-sampling, and under-sampling methods.

4.6.1. Parameters of DT

By using different types of functions to test the quality of a split, the researcher used the default setting due to there being no significant difference between Gini, Entropy, and Log Loss. Accordingly, since there are no interesting differences in the results of metrics by changing the other types of parameters such as the minimum samples in each node and the maximum depth of the tree, the researcher keeps using the default parameters.

4.6.2. DT with Resampling Methods

Table (4.5) summarizes the findings of this study by building the default model of the DT algorithm. The accuracy of the model by using imbalanced data was 53% same as recall, and precision was 54% same as F-measure. Compared to the imbalanced data model, the over-sampling method achieved better values of 81%, 81%, 80%, and 81% for precision, recall, F-measure, and accuracy respectively. As for the under-sampling method, the results were significantly lower than the previous method, where the F-measure, recall, and accuracy scored 41%. In general, the best result achieved with the over-sampling method for all metrics, while the worst value achieved for all metrics was the under-sampling method. The results can be interpreted as dataset type related.

Table (4.5): The Performance of DT Algorithm by using Different Resampling Methods

Decision Tree (DT) Algorithm				
Resampling Methods	**Metrics (%)**			
	Precision	**Recall**	**F-Measure**	**Accuracy**
Imbalanced Data	54	53	54	53
Under-sampling	40	41	41	41
Over-sampling	81	81	80	81

Figure (4.9) presents three confusion matrices for resampling methods used in the DT algorithm to describe the difference between them in the prediction of classifying classes. Regarding imbalanced data, the most class classified correctly more than others was slight followed by serious then fatal. Related to the under-sampling method, the lowest class predicted correctly was serious, meanwhile, the best class predicted in the right way was fatal. In addition, related to the over-sampling method the classifying process for the slight class was the worst prediction, and the higher prediction results achieved is for the class of fatal as shown in shape (3).

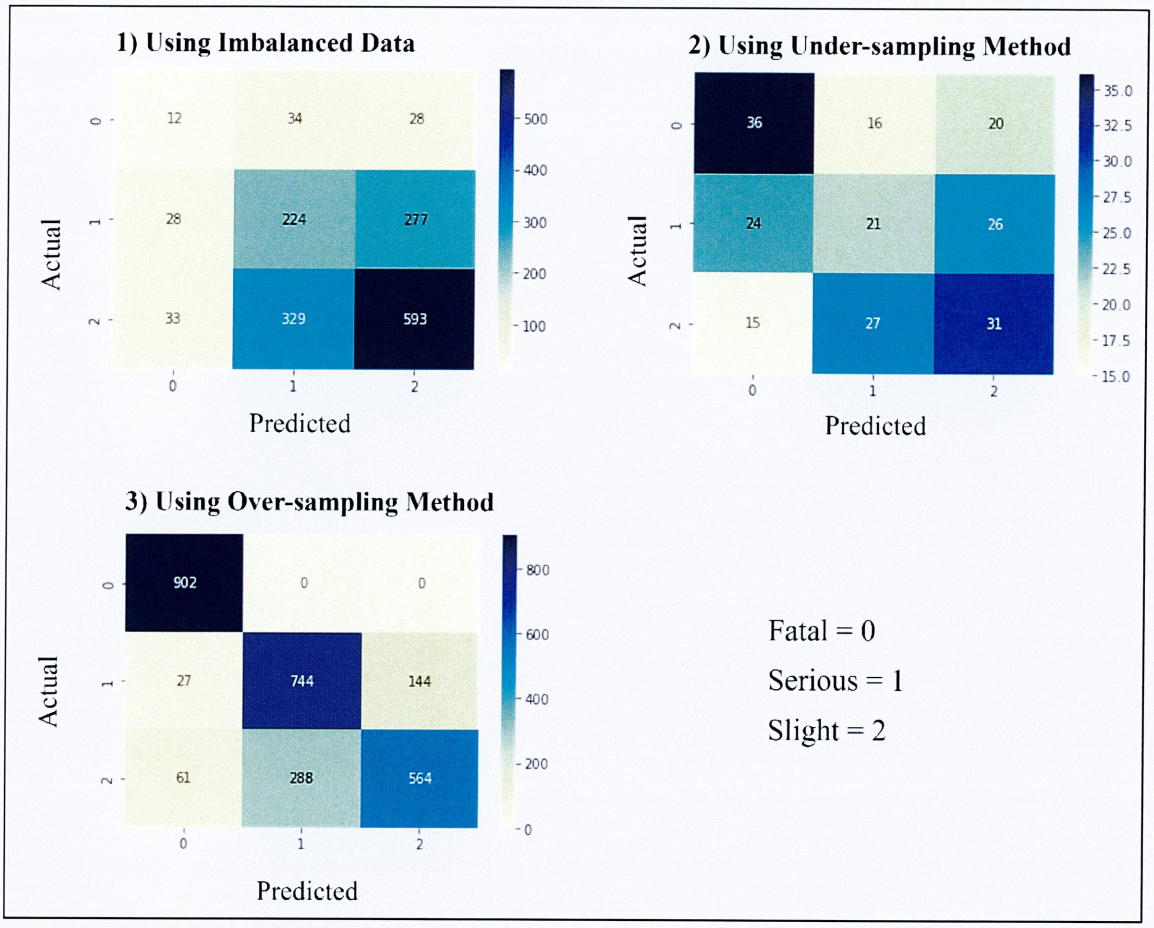


Figure (4.9): Confusion Matrix for the DT Models of the RTAs Classes

Figure (4.10) presents the values of AUC for target classes after building the DT model by using resampling methods. The best AUC values were achieved with the over-sampling method, which the class values of fatal, serious, and slight reached of 98%, 83%, and 77% respectively. And generally, the under-sampling method scored the lowest values. Moreover, the fatal class has the highest rate for AUC in all resampling methods.

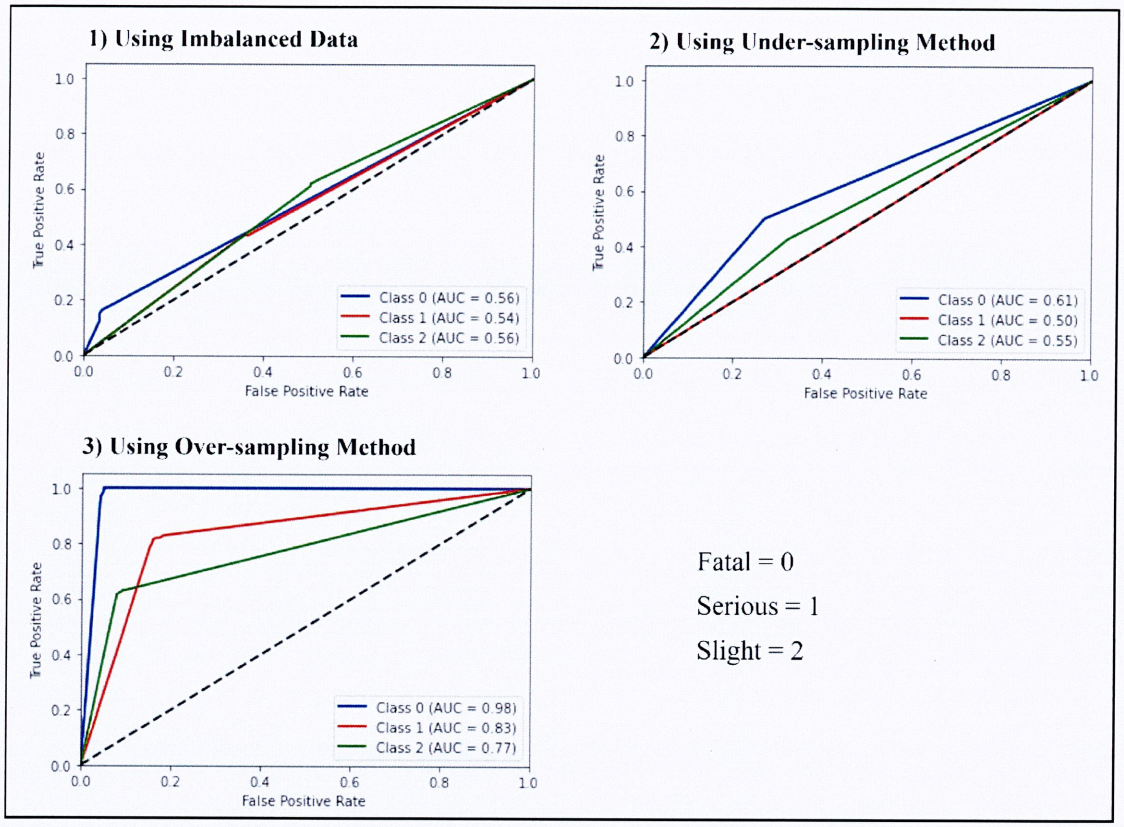


Figure (4.10): AUC Values for the DT Models Related to the RTAs Classes

4.7. Best Models Comparison

Table (4.6) presents the best results between the five types of models after trying different resampling methods and finding the appropriate parameters. It is observed, the resampling method that repeated more with different models used in this study is over-sampling to score the highest values with different evaluation methods used, the imbalanced data employed only one time with the SVM algorithm to be considered as the best method can achieve the highest results for metrics, while the under-sampling method didn't achieve the highest results in any type of algorithms that were worked on.

Regarding the results, the greatest values of 85% were achieved with RF for each one of the evaluation methods. The worst values scored for all

criteria between all best models was NB of 47%, 45%, 39%, and 45% for precision, recall, F-measure, and accuracy respectively.

Generally, Table (4.6) shows significant differences in the results between the best model of RF and the worst model of NB. For example, by focusing on metric recall the researcher found the difference reached 40%, and the NB algorithm was less than the RF algorithm by 46% with the metric of F-measure.

Table (4.6): Comparison Between All Best Models of Applied Algorithms

Comparing Between Best Models			
Model	**Metrics (%)**		
	Precision	**Recall**	**F-Measure**
KNN (Over-sampling, K=3)	74	75	74
RF (Over-sampling)	85	85	85
NB (Over-sampling)	47	45	39
SVM (Imbalanced Data)	60	61	47
DT (Over-sampling)	81	81	80

Figure (4.11) shows a comparison between the accuracies for the best results of all models implemented in the study. In terms of comparison, the RF model outperforms the rest, although the accuracy of the DT and RF models is nearly the same. In addition, the lowest result of accuracy achieved with the model of NB was 45%.

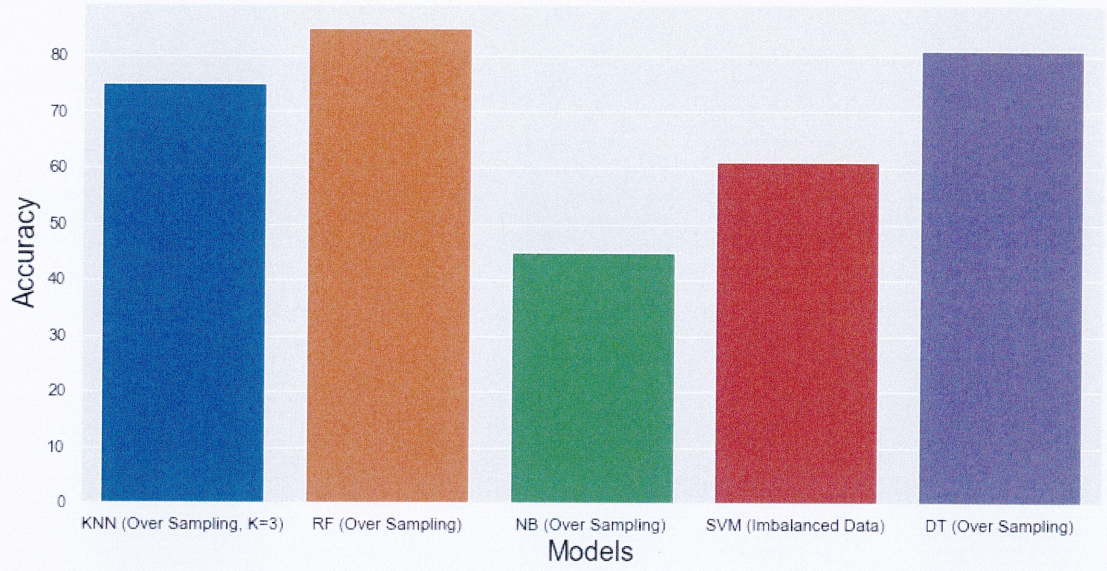


Figure (4.11): Comparison Between Best Accuracies for all Models of Applied Algorithms

Table (4.7) shows the values of the AUC results for the best models. The resampling method of over-sampling achieved the highest AUC with all models. The AUC values with the RF algorithm of fatal, serious, and slight are 100%, 93%, and 93 respectively. Also, it is observed that there is a significant difference between the RF algorithm and the NB algorithm reaching around 30% in all classes.

Table (4.7): Comparison Between the Best Results of the Algorithm for the Values of the AUC

Model	Results of the AUC (%)		
	Fatal (0)	Serious (1)	Slight (2)
KNN (Over-sampling, K=3)	98	83	82
RF (Over-sampling)	100	93	93
NB (Over-sampling)	71	60	66
SVM (Over-sampling)	71	61	66
DT (Over-sampling)	98	83	77

By that, the model of RF can represent the dataset under investigation and predict the targets more effectively than other algorithms that are used, followed by the DT algorithm. Meanwhile, the worst model to be used with this dataset is the algorithm of NB. However, these results agree with the research of (Ospina-Mateus, et al., 2021) which used the same algorithms and the same target classes with different datasets, where the NB algorithm achieved the lowest accuracy, and the RF algorithm scored the highest results.

Generally, focusing on the values achieved by evaluation methods, the researcher found that KNN has provided better results than SVM for several reasons. One possible reason is that KNN can handle noisy data better than SVM. Also, the KNN can work better with small and medium-sized datasets. On the other hand, SVM performance is affected by the high number of features.

4.8. Feature Selections

Since the model of RF achieved the best results than other models that were used in this study, the random forest feature importance for each variable was investigated within this part. Also, this section helps to solve one of the problems under investigation in this research to decide the important features that affect the causalities that occurred on the roads in the Kingdom of Bahrain.

Figure (4.12) shows the features with their importance; taking into consideration, a higher value indicates how the feature has a stronger impact on the prediction model. The most important feature among the 19 mentioned features is Cause Type then followed by Accident Day of Week, Age, Accident Year, and Accident Month respectively. Meanwhile, the least important features are Road Surface, Road Condition, Visibility,

Gender, and Road Impact respectively. However, it is obvious from the scale that most features depend as important because their rates are more than 0.5.

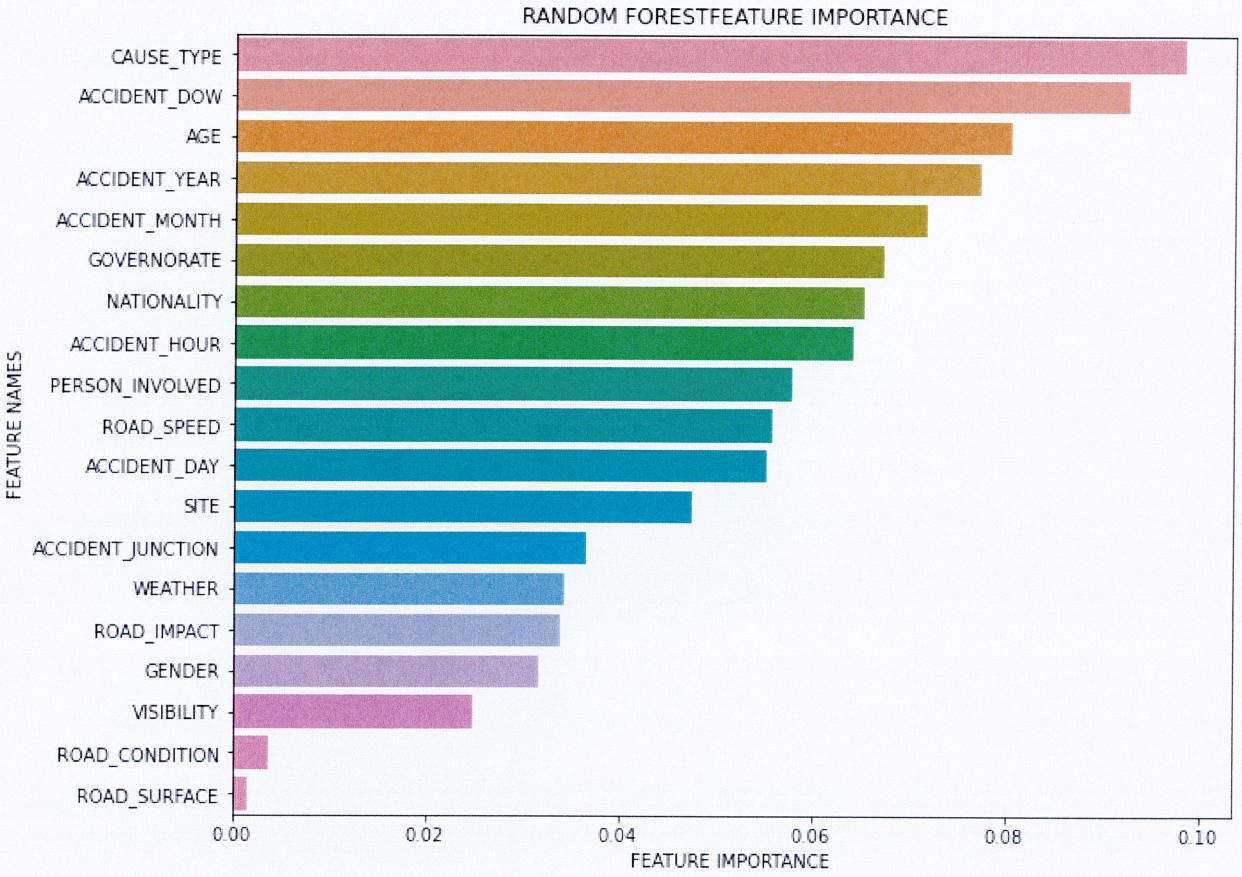


Figure (4.12): Most Important Features Effect on the Road Accidents in the Kingdom of Bahrain

For that, the researcher concludes that the most important feature effect on road accidents in the kingdom of Bahrain by using the method of RF feature selection is Cause Type which includes ten different variables such as Drunk, Ill, Walking on the Road, and Losing Control Before Accident. On the other hand, the least two important features contributing to causing accidents on the roads in the Kingdom of Bahrain were Road Surface and Road Condition; one of the reasons to make these two features are unimportant is the huge difference between variables number. For the Road Surface feature, the variable of Concrete/Asphalt is repeated of

99.5% while the other variable of Earth/Sand was 0.5%. For the second Feature of Road Condition, the variable of Dry recorded around 97.6% while the other variable of Wet was only 2.4%.

This result is compatible with the study of (Chakraborty, et al., 2021) where Visibility is one of the lower important variables that affect the RTAs. Meanwhile, the difference is with the variable of Road Surface, which is considered in this study as one of the lower important variables but is the highest important variable in the mentioned study.

4.9. Conclusion

This chapter discussed the results for NB, KNN, RF, DT, and SVM algorithms by changing the important parameters that affect to increase values of critical methods. Choosing the best classifier is done by focusing on the highest scores of the criteria. In this research, the evaluation methods used were recall, precision, accuracy, F-measure, and AUC.

After experiments, the results appeared that the RF algorithm outperformers other types of algorithms that used in this study, it was 85% for evaluation methods of accuracy, precision, recall, and F-measure by using the over-sampling method, and the AUC values for fatal, slight, and serious classes achieved of 100%, 93%, and 93%. On the other hand, the worst algorithm that achieved the lowest results was the NB by using the original dataset without any resampling methods, where the values for accuracy, F-measure, recall, and precision of 10%, 10%, 10%, and 63% respectively.

In addition, the best feature found that can affect the classifying of the classes of road injuries is Cause Type and Accident Day of Week, meanwhile the worst features are Road Surface and Road Condition.

Chapter Five

Conclusion and Recommendations

5.0. Introduction 90

5.1. Conclusion 90

5.2. Limitations 91

5.3. Recommendations 91

5.4. Future Work 92

Chapter Five

Conclusion and Recommendation

5.0. Introduction

This chapter provides an overview of the research, including the results achieved using several algorithms and evaluation methods, followed by the most important research limitations. section 3 considers some recommendations, and section 4 highlights the main future works.

5.1. Conclusion

This study aims to classify the injuries caused by RTAs by using the dataset related to accidents that occurred in all governorates of the kingdom of Bahrain during the periods from January 2018 to June 2022. This dataset exists in five excel files, which each file covers one specific year. The datasets mentioned are provided by the statistic section of GDT, and the number of injuries recorded reached 5383.

Moreover, ML algorithms used through this research were KNN, DT, RF, NB, and SVM. The evaluation methods used to decide the behavior of each algorithm with the dataset under investigation were accuracy, precision, recall, F-measure, and AUC. Also, the resampling methods used were over-sampling and under-sampling with imbalanced data.

Achieved the important parameters, the results found the highest accuracy achieved of 85% with the model of RF by using an over-sampling method, and the other critical methods of precision, recall, and F-measure achieved the same score of 85%. Meanwhile, the lowest accuracy, recall, and F-measure of 10% were achieved with the NB model by using an

imbalanced dataset, and the worst precision of 35% was found with NB algorithm when the method used is under-sampling. Generally, the best model that can predict the classes of road accidents is RF and the worst model is NB.

By using the random forest feature importance method, the most important features were effect on the injuries of RTAs that happen in the Kingdom of Bahrain were Cause Type, Accident Day of Week, Age, and Accident Year respectively. In contrast, the very weak features that affect the class of injuries were Road Surface, and Road condition.

There are many contributions for this research in the local and global arena. Firstly, it contributes to identifying the most important factors affecting traffic accident injuries. Also, this research is the first of its kind in the Kingdom of Bahrain that used the ML approach analyzing the injuries related to RTAs by using the dataset provided from GDT. In addition, this research compared the results of used models for SVM, RF, DT, NB, and KNN to know the best one between them appropriate to investigated dataset.

5.2. Limitations

One of the limitations of this study related to the size of the dataset which reached around five thousand records only and that may affect the accuracy in the process of training the models. Also, there are no previous studies that used the ML approach to discuss RTA severity in the Kingdom of Bahrain. In addition, the time available for completing this study is limited.

5.3. Recommendations

After completing this study, the researcher recommends the need to encourage the establishment of studies related to the analysis of road

accidents using ML approaches in the Kingdom of Bahrain. The important factors that cause traffic accident injuries must be considered during the construction of new roads in the Kingdom of Bahrain. Moreover, it is recommended to start using the RF classifier in predicting the classes of road injuries for its high results in evaluation methods.

5.4. Future Work

Since a variety of features may make a noticeable impact on road accidents, studying the impact of these features will be part of future work. Furthermore, other research can use additional classifiers such as CNN and ANN, and more resampling methods to enhance the results.

In addition, using a larger dataset will be a good idea to try to improve the performance of the models, and adding evaluation methods such as Kappa statistics will be considered in future work.

References

Micheale, K. G. (2017). Road traffic accident: Human security perspective. *International Journal of Peace and Development Studies*, *8*(2), 15-24.

Bokaba, T., Doorsamy, W., & Paul, B. S. (2022). Comparative study of machine learning classifiers for modelling road traffic accidents. *Applied Sciences*, *12*(2), 828.

Touahmia, M. (2018). Identification of risk factors influencing road traffic accidents. *Engineering, Technology & Applied Science Research*, *8*(1), 2417-2421.

Parvareh, M., Karimi, A., Rezaei, S., Woldemichael, A., Nili, S., Nouri, B., & Nasab, N. E. (2018). Assessment and prediction of road accident injuries trend using time-series models in Kurdistan. *Burns & Trauma*, *6*.

Hamadeh, R. R., & Ali, N. M. A. (2013). Fatalities from road traffic accidents among the young in Bahrain. *East Mediterr Health J.*, *19*(10), 854-60.

ESCAP, U. (2020). Road safety: saving lives beyond 2020 in the Asia-Pacific region.

Woyessa, A. H., Heyi, W. D., Ture, N. H., & Moti, B. K. (2021). Patterns of road traffic accident, nature of related injuries, and post-crash outcome determinants in western Ethiopia-a hospital based study. *African Journal of Emergency Medicine*, *11*(1), 123-131.

Butt, F. M., Ashiq, M., Rehman, S. U., Minhas, K. S., & Khan, M. A. (2020). Bibliometric analysis of road traffic injuries research in the Gulf Cooperation Council region. *F1000Research*, *9*.

Abegaz, T., & Gebremedhin, S. (2019). Magnitude of road traffic accident related injuries and fatalities in Ethiopia. *PloS One*, *14*(1), e0202240.

Gasana, J., Albahar, S., Alkhalidi, M., Al-Mekhled, Q., El Reda, D., & Al-Sharbati, M. (2022). Risky roads in Kuwait: an uneven toll on migrant workers. *International Journal of Environmental Research and Public Health*, *19*(15), 9726.

General Directorate of Traffic. (2020). Traffic Statistics. Statistic Section, General Directorate of Traffic, Kingdom of Bahrain.

Alrajhi, M., & Kamel, M. (2019). A Deep-Learning Model for Predicting and Visualizing the Risk of Road Traffic Accidents in Saudi Arabia: A Tutorial Approach. *International Journal of Advanced Computer Science and Applications*, *10*(11).

Bokaba, T., Doorsamy, W., & Paul, B. S. (2022). A Comparative Study of Ensemble Models for Predicting Road Traffic Congestion. *Applied Sciences*, *12*(3), 1337.

Yan, M., & Shen, Y. (2022). Traffic accident severity prediction based on random forest. *Sustainability*, *14*(3), 1729.

Al-Moqri, T., Haijun, X., Namahoro, J. P., Alfalahi, E. N., & Alwesabi, I. (2020). Exploiting Machine Learning Algorithms for Predicting Crash Injury Severity in Yemen: Hospital Case Study. *Appl. Comput. Math*, *9*(5), 155-164.

Assi, K., Rahman, S. M., Mansoor, U., & Ratrout, N. (2020). Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. *International Journal of Environmental Research and Public Health*, *17*(15), 5497.

Megnidio-Tchoukouegno, M., & Adedeji, J. A. (2023). Machine Learning for Road Traffic Accident Improvement and Environmental Resource Management in the Transportation Sector. *Sustainability*, *15*(3), 2014.

مليون ونصف عدد سكان البحرين . (2022, Feb 26). Al-Ayam Newspaper, 12012.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet]*, *9*, 381-386.

Colliot, O. (2023). A non-technical introduction to machine learning.

Javeed, A., Dallora, A. L., Berglund, J. S., Ali, A., Ali, L., & Anderberg, P. (2023). Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. *Journal of Medical Systems*, *47*(1), 1-25.

Fiorentini, N., & Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures*, *5*(7), 61.

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 1-11.

Mailagaha Kumbure, M., & Luukka, P. (2022). A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance. *Granular Computing*, *7*(3), 657-671.

Prihatmono, M. W., Arni, S., Iin, J. N., & Moeis, D. (2022). Application of the KNN Algorithm for Predicting Data Card Sales at PT. XL Axiata Makassar. In Conference Series (Vol. 4, pp. 59-64).

Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K., & 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.

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

Liao, Y., Zhang, J., Wang, S., Li, S., & Han, J. (2018). Study on crash injury severity prediction of autonomous vehicles for different emergency decisions based on support vector machine model. *Electronics*, *7*(12), 381.

Rahim, M. A., & Hassan, H. M. (2021). A deep learning based traffic crash severity prediction framework. *Accident Analysis & Prevention*, *154*, 106090.

Hosseinzadeh, A., Moeinaddini, A., & Ghasemzadeh, A. (2021). Investigating factors affecting severity of large truck-involved crashes: Comparison of the SVM and random parameter logit model. *Journal of Safety Research*, *77*, 151-160.

Bahiru, T. K., Singh, D. K., & Tessfaw, E. A. (2018, April). Comparative study on data mining classification algorithms for predicting road traffic accident severity. In *2018 Second International Conference on*

Inventive Communication and Computational Technologies (ICICCT) (pp. 1655-1660). IEEE.

Jeong, H., Jang, Y., Bowman, P. J., & Masoud, N. (2018). Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. *Accident Analysis & Prevention*, *120*, 250-261.

Chen, M. M., & Chen, M. C. (2020). Modeling road accident severity with comparisons of logistic regression, decision tree and random forest. *Information*, *11*(5), 270.

Ijaz, M., Zahid, M., & Jamal, A. (2021). A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. *Accident Analysis & Prevention*, *154*, 106094.

Lee, J., Yoon, T., Kwon, S., & Lee, J. (2019). Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. *Applied Sciences*, *10*(1), 129.

Chakraborty, M., Gates, T., & Sinha, S. (2021). Causal Analysis and Classification of Traffic Crash Injury Severity Using Machine Learning Algorithms. *arXiv preprint arXiv:2112.03407*.

Azhar, A., Ariff, N. M., Bakar, M. A. A., & Roslan, A. (2022). Classification of driver injury severity for accidents involving heavy vehicles with decision tree and random forest. *Sustainability*, *14*(7), 4101.

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051-10072.

Sripuram, A., Jayasree, T., Vaishnavi, T., Reddy, V. R., & Mtech, K. P. C. (2022). ROAD ACCIDENT ANALYSIS USING MACHINE LEARNING. International Research Journal of Modernization in Engineering Technology and Science, 4(3), 1364-1371.

Cigdem, A., & Ozden, C. (2018). Predicting the severity of motor vehicle accident injuries in Adana-Turkey using machine learning methods and detailed meteorological data. *International Journal of Intelligent Systems and Applications in Engineering*, *6*(1), 72-79.

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H. M., Almoshaogeh, M., Farooq, D., & Ahmad, M. (2021). Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *International Journal of Injury Control and Safety Promotion*, *28*(4), 408-427.

Taamneh, M., Alkheder, S., & Taamneh, S. (2017). Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *Journal of Transportation Safety & Security*, *9*(2), 146-166.

Alrajhi, M., & Kamel, M. (2019). A Deep-Learning Model for Predicting and Visualizing the Risk of Road Traffic Accidents in Saudi Arabia: A Tutorial Approach. *International Journal of Advanced Computer Science and Applications*, *10*(11).

General Directorate of Traffic. (2017). Traffic Statistics. Statistic Section, General Directorate of Traffic, Kingdom of Bahrain.

General Directorate of Traffic. (2018). Traffic Statistics. Statistic Section, General Directorate of Traffic, Kingdom of Bahrain.

General Directorate of Traffic. (2019). Traffic Statistics. Statistic Section, General Directorate of Traffic, Kingdom of Bahrain.

General Directorate of Traffic. (2021). Traffic Statistics. Statistic Section, General Directorate of Traffic, Kingdom of Bahrain.

Eibe, F., Ian, W., and Mark, H., Data mining: practical machine learning tools and techniques. Amsterdam: Morgan Kaufmann, 2017.

Giuseppe, B., Machine learning algorithms: popular algorithms for data science and machine learning. Birmingham, U.K.: Packt Publishing Ltd, 2018.

Appendices

The following figures show examples of the codes that were used in the research.

```
def plot_feature_importance(importance,names,model_type):
  #Create arrays from feature importance and feature names
  feature_importance = np.array(importance)
  feature_names = np.array(names)

  #Create a DataFrame using a Dictionary
  data={'feature_names':feature_names,'feature_importance':feature_importance}
  fi_df = pd.DataFrame(data)

  #Sort the DataFrame in order decreasing feature importance
  fi_df.sort_values(by=['feature_importance'], ascending=False,inplace=True)

  #Define size of bar plot
  plt.figure(figsize=(10,8))
  #Plot Searborn bar chart
  sns.barplot(x=fi_df['feature_importance'], y=fi_df['feature_names'])
  #Add chart labels
  plt.title(model_type + 'FEATURE IMPORTANCE')
  plt.xlabel('FEATURE IMPORTANCE')
  plt.ylabel('FEATURE NAMES')
```

A1: Codes that are used to find the feature importance.

```
# Predict probabilities on test set
# Just change the type of classifier
y_scores = clf.predict_proba(X_test)

# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(3):
     # Create binary classification labels (positive class: i, negative class: not i)
    y_true_i = (y_test == i).astype(int)
    y_scores_i = y_scores[:, i]
    # Compute ROC curve and ROC area
    fpr[i], tpr[i], _ = roc_curve(y_true_i, y_scores_i)
    roc_auc[i] = auc(fpr[i], tpr[i])

# Plot ROC curve for each class
plt.figure()
plt.plot(fpr[0], tpr[0], color='blue', lw=2, label='Class 0 (AUC = %0.2f)' % roc_auc[0])
plt.plot(fpr[1], tpr[1], color='red', lw=2, label='Class 1 (AUC = %0.2f)' % roc_auc[1])
plt.plot(fpr[2], tpr[2], color='green', lw=2, label='Class 2 (AUC = %0.2f)' % roc_auc[2])
plt.plot([0, 1], [0, 1], color='black', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend(loc="lower right")
plt.show()
```

A2: Codes are used to find the AUC for algorithms.

```
# To decided the Independent variables
X = oversampled_df.iloc[:,0:19]

# To decided the Dependent variable
Y = oversampled_df.iloc[:,19]

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler() # to make dataset between -1 and 1
scaler.fit(X)

 #Splitting dataset into 70% training and 30% test
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=1)

# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
model = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = model.predict(X_test)

# To see the results of evaluation methods and confusin matrics
print(classification_report(y_test,y_pred))
print(confusion_matrix(y_test, y_pred))


mat = confusion_matrix(y_test, y_pred)
p = sbn.heatmap(pd.DataFrame(mat), annot=True, cmap="YlGnBu" ,fmt='g')
```

A2: Codes are used to build the DT algorithm with Over-sampling method.

الملخص

الحوادث المرورية على الطريق تساهم في رفع معدل الوفيات حول العالم. ولهذا السبب فإن الكثير من الدول وبما فيها مملكة البحرين وضعت على عاتقها هدف تقليل الحوادث المرورية قدر الإمكان. ومن الجدير بالذكر إنه في السنوات الأخيرة تم ملاحظة تطبيق مناهج التعلم الآلي بصورة متزايدة في العديد من المجالات مثل الصحة والتعليم والإقتصاد وعلوم الأحياء والمواصلات. في هذه الدراسة تم استخدام عدة خوارزميات هي: أقرب الجيران، والبايزي الساذج، وآلة المتجهات الداعمة، والغابة العشوائية، وشجرة القرار للتنبؤ بشدة الحوادث المرورية على الطريق في مملكة البحرين، وتم ذلك بالاعتماد على عدة سمات مثل تاريخ الحادث، الطقس، ونوع الطريق، وقد تم اختيار المعاملات المناسبة لكل خوارزمية.

في هذه الاطروحة تم استخدام خمسة مجموعات من البيانات قد تم الحصول عليها من قسم الإحصاء بالإدارة العامة للمرور لمملكة البحرين. وكل مجموعة من البيانات تعود إلى سنة محددة في الفترة بين العام 2018م وحتى العام 2022م. وبعد تحضير هذه البيانات وتنقيحها تم الحصول على 5191 من الإصابات بعد تجميع كل البيانات في قائمة واحدة.

للمقارنة بين الخوارزميات المذكورة تم استخدام عدة أدوات للتقييم وهي: الدقة، والقياس، والإستدعاء، والضبط، والمساحة تحت المنحنى. ومن المثير للإهتمام إن المتغير التابع في هذه الدراسة يحتوي على أصناف غير متوازنة، لذلك عمد الباحث لاستخدام البيانات الأصلية الغير متوازنة بالإضافة إلى طريقة أخذ العينات الناقصة وطريقة أخذ العينات الإضافية من أجل تقييم الخوارزميات وفقا لهذه الحالات.

النتائج التي حققتها هذه الدراسة مرضية جدا، فقد تم التوصل إلى أن أفضل مصنف هو الغابة العشوائية بسبب تحقيق نتائج عالية تصل إلى 85% في جميع أدوات التقييم المستخدمة عبر استخدام طريقة العينات الإضافية. بينما أسوء مصنف تم الحصول عليه هو خوارزمية البايزي الساذج حيث تم استخدام البيانات غير المتوازنة، وكانت نسبة الضبط 63% بينما حصدت جميع أدوات التقييم الأخرى على نسبة 10%. ومن خلال بناء النماذج الخوارزمية تبين أن فئة الموت تمتلك أكبر مساحة تحت المنحنى مقارنة بباقي الفئات. وبصورة عامة تم ملاحظة أن نتائج طريقة البيانات الإضافية تكون أفضل من غيرها. علاوة على ذلك فإن أكثر عاملين يؤثران على الحوادث المرورية على الطريق هما سبب الحادث ويوم وقوع الحادث، بينما العوامل الأقل أهمية وباختلاف ملحوظ هما حالة الطريق وسطح الطريق. النتائج التي حققتها هذه الدراسة واعدة ومن المتوقع أن تستفيد منها الإدارة العامة للمرور في مملكة البحرين.

جامعة البحرين

كلية العلوم



قسم برامج الدراسات العليا

استخدام نهج التعلم الآلي لتحليل الحوادث المرورية في مملكة البحرين

إطروحة مقدمة كجزء من متطلبات الحصول على درجة الماجستير

في علوم وتحليل البيانات الضخمه

إعـــداد

سلمان بدر عبدالله الحداد

الرقم الجامعي: ٢٠٠٨١٦٠٣

إشـــراف

د. أحمد محمد زكي

أستاذ مساعد

جامعة البحرين

مـملكة البـــحرين

مارس ٢٠٢٣م