APPLICATION OF MACHINE LEARNING IN ANALYZING SEVERITIES OF DOUBLE VEHICLE CRASHES IN BANGLADESH

SADIA MOHSIN

MASTER OF SCIENCE IN CIVIL ENGINEERING (TRANSPORTATION)

DEPARTMENT OF CIVIL ENGINEERING

BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

DHAKA-1000, BANGLADESH

March 2023

APPLICATION OF MACHINE LEARNING IN ANALYZING SEVERITIES OF DOUBLE VEHICLE CRASHES IN BANGLADESH

by

SADIA MOHSIN

Student ID No.: 1015042420 P

A thesis submitted to Department of Civil Engineering,

Bangladesh University of Engineering and Technology, Dhaka

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN CIVIL ENGINEERING

(TRANSPORTATION)

DEPARTMENT OF CIVIL ENGINEERING

BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY
DHAKA

March 2023

The thesis titled "Application of Machine Learning in Analyzing Severities of Double Vehicle Crashes in Bangladesh" submitted by Sadia Mohsin, Roll No.: 1015042420 P, Session: October 2015 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Science in Civil Engineering (Transportation) on the 20th March, 2023.

BOARD OF EXAMINERS

Dr. Md. Mizanur Rahman Professor Department of Civil Engineering BUET, Dhaka-1000	Chairman (Supervisor)
Dr. Abu Siddique Professor and Head Department of Civil Engineering BUET, Dhaka-1000	Member (Ex-Officio)
Dr. Md. Hadiuzzaman Professor Department of Civil Engineering BUET, Dhaka-1000	Member
Dr. Md. Asif Raihan Assistant Professor Accident Research Institute (ARI) BUET, Dhaka-1000	Member (External)

DECLARATION

It is hereby declared that, except for the contents where specific references have been made to the work of others, the research work presented in this thesis has been carried out by the author under the supervision of Dr. Md. Mizanur Rahman, Professor, Department of Civil Engineering, BUET. This thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

Sadia Mohsin

Student ID: 1015042420 P

Date: 20.03.2023

ACKNOWLEDGEMENT

All praises to Allah the most beneficial and merciful. By the grace and proper guidance of Almighty Allah, I have been able to complete the thesis as a requirement for the degree of Master of Science in Civil Engineering (Transportation).

I am grateful to all my family members for their unconditional support, love, and blessing. My parents are the one who comes next to the Almighty. Though words can never be enough to express my gratefulness to them, I just want to say that I am greatly indebted to them for bringing me up with love and encouragement to this stage and for everything they did for me. I would also like to thank my husband for all the inspiration and support throughout the research work. I am indebted to him for all the tireless motivation and support. I would also like to thank my in-laws for all the supports throughout the whole journey. For all these people the journey of my entire research career till today has been so easy and comfortable.

I would like to express my utmost gratitude to my supervisor, Dr. Md. Mizanur Rahman, Professor, Department of Civil Engineering, Bangladesh University of Engineering and Technology (BUET), for his earnest supervision, prudent guidance, inclusive and important comments, along with timely and efficient discussion sessions. It was my honor and pleasure to work under his constant guidance. His impeccable focus on perfection has always galvanized me to strive for better outcomes. I would also like to express gratitude to Dr. Md. Asif Raihan, Assistant Professor, ARI, BUET, for his continuous support. His expertise and valuable knowledge sharing throughout the research work have provided immense assistance to me. I would also like to express gratitude to B.M Tazbiul Hassan Anik and Fajle Rabbi Ashik, Research Assistant, ARI, BUET for their immense support in this research.

Next, I am thankful to all the honorable board members for their valuable time. My sincere gratitude to the teachers of Department of Civil Engineering, BUET for sharing their knowledge and experience and thus enlightening me.

Finally, I would like to thank all my colleagues and well-wishers for all their encouraging words and constant support.

Dedicated

to

My Parents

ABSTRACT

Road traffic crashes have become one of the leading causes of death worldwide. Bangladesh, a developing country, is rapidly becoming a major victim of road accidents. Due to traffic crashes, different types of injuries eventuate depending on the severity level of the crashes. Double vehicle crashes are the most critical type of road accidents that have the potential to cause serious injuries and fatalities. Unfortunately, Bangladesh is still in nascent stage in dealing with road accidents, especially for double vehicle crashes. A precise prediction of crash severity in road accidents significantly improves traffic safety. Therefore, there has recently been a tactical shift among safety researchers to apply machine learning (ML) algorithms to estimate crash severity due to their superior predictive ability. Although there have been an increasing number of applications of machine learning methods in crash severity research, however there is a limited applicability of these methods in estimating the severity of a double vehicle crashes. As a result, this study aimed to apply machine learning algorithms in predicting double vehicle crash severities in the context of Bangladesh.

The aim of this study is to compare the predicted performance of numerous machine learning and traditional statistical regression techniques in modeling double vehicle crash severities, as well as to identify the contributing components and how they impact crash severity prediction. Using Dhaka's most recent crash record collected from Accident Research Institute (ARI), BUET (2017-2020), this study employed classification and regression tree, support vector machine, random forest, adaptive boosting, logistic regression, and soft voting classifier-based hybrid models. This study compared the performance of logistic regression and other machine learning classifiers using the most commonly known evaluation criteria: Accuracy (ACC), Receiver Operating Characteristics (ROC) Curve, and Area Under the Curve (AUC) Value. The comparison of predictive performance revealed that the hybrid model, built on logistic regression, random forest, and adaptive boosting, outperforms other individual models with a subset of twenty explanatory variables and with an accuracy of 75% and an AUC score of 0.71. With the same subset of features, random forest performs better with an accuracy of 70% and an AUC score of 0.69 within the individual models. This study uses the SHAP (Shapley Additive Explanation) methodology to determine how well the features contribute to the severity prediction, thus finding influential factors. SHAP Global

Feature Importance represents the marginal contribution of each feature in the prediction. SHAP Local Explanation identifies how the contributing factors affect double vehicle crash severities. According to the SHAP (Shapley Additive Explanation) technique, the most significant elements of double vehicle crash severities are the day of the week, vehicle type, time of day, vehicle maneuver, road geometry and they have important contribution in predicting crash severities by an average of 10.2, 4.8, 4.6, 3.8 and 2.9 percentage points respectively. This means that the factor day of week alone contributed in predicting whether the double vehicle crash severity would be fatal or not by an average of 10.2 percentage points. In addition, vehicle type is another most critical variables in predicting double vehicle crash severities whether it would be fatal or not by an average of 4.8 percentage points.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML	:	Machine Learning
LR	:	Logistic Regression
RF	:	Random Forest
CART	:	Classification and Regression Tree
SVM	:	Support Vector Machine
ARI	:	Accident Research Institute
ACC	:	Accuracy
AUC	:	Area Under Curve
ROC	:	Receiver Operating Characteristics
ROS	:	Random Over Sampling
WHO	:	World Health Organization
SHAP	:	Shapley Additive Explanation
GNP	:	Gross National Product
GDP	:	Gross Domestic Product
SPFs	:	Safety Performance Functions
RTA	:	Road Traffic Accident
ARF	:	Accident Report Form
LSA	:	Local Sensitivity Analysis
SA	:	Sensitivity Analysis
ANN	:	Artificial Neural Networks
GBDT	:	Gradient Boosting Decision Tree
XGBoost	:	eXtreme Gradient Boosting
GEE	:	Generalized Estimation Equation
MAAP5	:	Microcomputer Accident Analysis Package 5
OOB	:	Out of Bag Data
TN	:	True Negative
TP	:	True Positive
FN	:	False Negative
FP	:	False Positive

Chapter 1

INTRODUCTION

1.1 Background

Road traffic fatality rates are higher in low- and middle-income countries (21.5 and 19.5 per 100,000 population, respectively) than in fast countries (10.3 per 1,00,000) (WHO 2018). Low and middle-income countries account for 93 percent of global road accident fatalities despite owning only 60 percent of the world's motor vehicles (WHO 2022). Global losses from traffic crashes are estimated to be $518 billion, costing nations around 1% and 3% of their gross national product (GNP) - greater than the total amount of development assistance received by these countries. While accidental fatality rate in many high-income nations have resolved or declined in recent decades, data indicate that the global epidemic of automobile accidents is still growing in most regions of the world. Road deaths are expected to rise to the fifth top cause of death by 2030, resulting in around 2.4 million fatalities per year unless immediate action is taken (WHO, 2009). The actual fatality rate is likely to be higher. Between 1982 and 2000, the number of accidents increased by 43%, while the death toll increased by around 400% (Louis Burger, 2005).

Bangladesh, a low-income Asian country, has the highest death rate in Asia, with 1020 persons killed each year per 100,000 motor vehicles (WHO 2018). According to official statistics, there are more than 60 fatalities in road accidents for every 10,000 cars in Bangladesh (WHO 2018). Every day, approximately eight people are killed in traffic accidents. Bangladesh's number of deaths from accidents is comparable to that of countries at war, like Sierra Leone and Liberia (Al-Mahmood, 2007). According to a report published by the Bangladesh Road Transport Authority (BRTA), there were a total of 4,625 road accidents in Bangladesh in 2021, resulting in 4,999 deaths and 7,460 injuries. Over the past two decades, the death toll has increased 3.5 times, reaching over 3,000 annually. However, some have speculated that the number is higher than 12,000 annually (due to non-reporting and misreporting) (UNESCAP, 2007). As the country's population, total road length, and modal share of road transport continue to grow, the number of fatalities from road accidents is expected to rise further.

Although Bangladesh's official traffic accident data paints a rosy picture of traffic safety, the reality is quite different. Traffic accidents have earned a permanent place in the print and annual deaths from road accidents could be 20,000, taking into account under reporting and definitional inconsistencies, whereas police reported statistics show that it is around 3,000 each year. Road accidents in Bangladesh cost the economy nearly 2% of its gross domestic product (GDP) (Hoque et al., 2008). As a result, measures to reduce accidents based on a thorough understanding of the underlying causes are of great interest to Bangladesh.

Road accidents are caused by several crash groupings, including single cars, two vehicles, and multi-vehicles. Among all subgroups, two-vehicle crashes are a critical type of road accidents that have the potential to cause serious injuries and fatalities. The severity of two-vehicle road accidents varies around the world, depending on a range of factors such as road infrastructure, driver behavior, the implementation of effective safety measures, and so on. Despite the use of safer vehicles, improved road design, and better enforcement of traffic laws, in developed countries, two-vehicle road accidents can still be severe and result in fatalities and serious injuries. For example, in the United States, two-vehicle collisions accounted for over 60% of all traffic fatalities in 2019, according to the National Highway Traffic Safety Administration (NHTSA). In Europe, approximately 40% of all road accidents involve two or more vehicles colliding with each other (Source: European Commission). In Australia, two-vehicle crashes account for approximately 62% of all fatal crashes. (Source: Australian Department of Infrastructure, Regional Development and Cities). Unfortunately, Bangladesh is still in nascent stage in dealing with road accidents, especially for two-vehicle crashes. No work on identification to identifying factors influencing two-vehicle crash outcomes and prediction to predicting two-vehicle crash outcomes has been done till now.

Researchers in transportation safety have been developing and implementing safety performance functions (SPFs) to achieve better traffic safety. In the past, research has shown that analyzing overall accidents without defining probable subgroups can miss connections between of subgroups and lead to inaccurate results when developing

SPFs (Geedipally, S., and D. Lord, 2010, Ma, J., and K. M. Kockelman, 2006, Geedipally, S. R., and D. Lord, 2010). Appropriately, the researchers attempted to develop the SPFs of multiple crashes simultaneously by dividing the total crashes into different categories based on injury severity, crash types, and the number of vehicles involved in a crash (Geedipally, S. R., and D. Lord, 2010, Martensen, H., and E. Dupont, 2013, Kitali, A. E., and P. E. T. Sando, 2017). Modeling collisions with possible clusters in crash data can aid in gaining a better understanding of the impact of multiple factors on every crash category, allowing for the development of effective protective measures.

Researchers frequently divide crash data into two groups when modeling crash frequency based on the total number of vehicles involved: single-vehicle crashes and two-plus vehicle crashes. (i.e multi-vehicle crashes) (Geedipally, S., and D. Lord, 2010, Geedipally, S. R., and D. Lord, 2010, Martensen, H., and E. Dupont, 2013, Chen, F., and S. Chen, 2011, Pasupathy et al., 2000, Ma, X. et al., 2016). Previous research has shown that crashes involving two or more vehicles differ significantly from those involving only one car, As a result, the two crash types must be modeled individually (Geedipally, S., and D. Lord, 2010, Geedipally, S. R., and D. Lord, 2010, Ma, X. et al., 2016, Qin, X. et al., 2004, Lord, D. et al., 2005 and Griffith, M. S., 1999). According to the findings of these studies, developing separate models for single-vehicle and two-plus vehicle accidents offers more accurate predictions than establishing models that combine the two crash categories (Geedipally, S. R., and D. Lord, 2010).

1.2 Present State of the Problem

Two vehicle collisions are the most dominant types of traffic accidents in Bangladesh, accounting for 35% of total accidents and 30% of total fatalities (Raihan et al. 2017). As a result, developing preventive mechanisms to minimize two vehicle crash fatalities is crucial. The estimation and use of disaggregate level crash severity models is a vital component of the preventative measure in identifying and obtaining a full understanding of the elements that lead to two vehicle crash severity. Furthermore, an independent two vehicle crash severity modeling is required because modeling aggregate accidents without specifying relevant subgroups may fail to find associations between subgroups, resulting in incorrect parameter estimations (Kitali et al. 2021). Remarkably it is essential to place emphasis on identifying the factors responsible for two vehicle crash severities.

1.3 Objectives of the Study

The purpose of the research was to investigate the prediction accuracy of different machine learning (ML) and statistical methods to predict two vehicle crash severity in a low-income country context, Bangladesh, specifically Dhaka city using Road Traffic Accident (RTA) data (2017- 2010) from ARI, BUET. The global objective of this study is employing machine learning algorithms to predict drivers' injury severities in two-vehicle crashes in Dhaka, Bangladesh. The specific research objectives are:

- to evaluate the potential of different ML models, both individual and hybrid models, and parametric regression model to predict the severity of a crash involving two vehicles.
- to identify the contributing factors and the ways in which they impact the prediction of crash severities in Dhaka, Bangladesh.
- to compare the hybrid model with individual classifiers to see if individual or hybrid model can forecast the severity of a two vehicle crash with greater accuracy.

1.4 Scope of the Research

This research is concerned with the crash severity prediction of two vehicles using Logistic Regression (LR) as a statistical method and some popular ML methods such as classification and regression tree (CART), support vector machine (SVM), random forest (RF), adaptive boosting and soft voting classifier- based hybrid model. The study shows how accident intensity is related to various factors associated of accident events, as well as which factors cause what type of accident severity. In-depth analyses of the study results needed to create defensive measures and strong policy decisions, however, were outside the scope of this thesis.

1.5 Thesis Outline

The thesis consists of six chapters.

Chapter 1 has explained the background, present state of the problem, purpose and objectives as well as the scope of the research.

Chapter 2 has been dedicated to review the relevant literature of two vehicle crashes in the context of this study.

Chapter 3 has illustrated the data description, data preparation, and the fundamentals of various machine learning, statistical and hybrid methods that have been applied in this thesis. These include Logistic Regression (LR), Classification and Regression Tree (CART), Support Vector Machine (SVM), Random Forest (RF), Adaptive Boosting (AdaBoost) and Classifier-specific Soft Voting as Hybrid model. The feature selection and model evaluation metrics have also described in this chapter. The descriptions are brief yet self-containing.

Chapter 4 has addressed the detailed analysis and interpretation of model results regarding two vehicle crash severity predictions.

Chapter 5 has presented the major findings of the thesis along with its limitations and future scopes.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Two vehicle collisions are the quite common types of traffic accidents currently in Bangladesh and developing preventive mechanisms to minimize two vehicle crash fatalities is crucial to improve road traffic safety. Statistical models have traditionally been the most commonly used methods for analyzing crash injury severity. For better prediction, besides traditional parametric statistical method, some popular machine learning techniques have been introduced in this study. This chapter commences by defining traditional parametric statistical method along with machine learning method. The limitations of statistical method, the advantages and shortcoming of machine learning method have also been incorporated in this chapter. It then clarifies the concept from the standpoint of transportation. The paper then summarizes the previous relevant literatures, conducting a thorough review of the objective and guidance in this evolving and absolutely vital research field.

2.2 Relevant Studies on Two vehicle Crashes

Numerous studies have been conducted to investigate the mechanism of a single vehicle collision, however just a few studies have been performed to evaluate two vehicle crashes (WHO 2022). Earlier researches in many nations have mostly concentrated on evaluating the causes of crashes involving two cars at signalized junctions, construction sites, urban and rural locations (Dancan et al., 1998; Chiou et al., 2020 and Yuan et al., 2022).

Yuan et al. (2022) in Pennsylvania, applied mixed logit models to discover the elements that determine injury severity in a two-vehicle incident, taking vehicle characteristics of the distinct crash roles into consideration. The result revealed that the type and movement of vehicles have a substantial impact on crash severity (WHO 2022).

Two-vehicle crashes have been studied extensively, with Champahom et al. (2020) and Wang & Abdel-Aty (2006) focusing on studying the mechanism of crashes at signalized intersections, in construction zones, in urban, and rural settings. The factors that contribute to the severity of crashes involving two vehicles at unsignalized intersections are poorly understood.

Two-vehicle collisions under similar traffic situations can vary greatly in severity due to the difference in the performance, type, weight, and vehicle movement (struck and striking vehicles), as determined by Yuan et al., (2017) and Lee & Li, (2014).

It has been found by Shao et al. (2020) that the severity of injuries sustained in truck-versus-car crashes differs significantly from those sustained in truck-only crashes. According to Abay et al. (2013), a front-facing vehicle poses a greater threat of injury to the driver than a frontal collision.

Lee and Li (2014) studied in Ontario, Canada, the severity of driver injuries in one and two-vehicle crashes and analyzes the impact of independent factors amongst various crash scenarios using heteroscedastic ordered logit (HOL) models. The study showed that, young car drivers have reverse impacts in car-to-car collisions, while side-impact collisions have distinct consequences in car-car and truck-truck collisions. They also found that car- heavy truck crashes are the mostly turn into fatal injury (Lee, C., and X. Li , 2014).

Zeng et al. (2016) studied the interaction influence on vehicle unit injury severity in two-vehicle crashes and found that, compared to cars, other types of vehicles were significantly more severe. Injury severity is lower for the driver of the vehicle himself, but to a greater extent for the driver of other vehicles (Yang et al., 2019).

Chiou et al. (2020) in Taiwan, considered the severity of the crash by two parties (referred to as the "responsible party" and the "non-responsible party") using the Generalized Estimation Equation (GEE) and the result indicated that the most significant variable that contributes to the severity of a crash is the vehicle type (motorcycle), followed by speed, angle, impact and alcohol consumption (Duncan et al., 1998).

Lombardi et al. (2017) researched age-related disparities in fatal accidents between two vehicles, and the findings revealed that older and younger drivers were more prone to engage in road accidents than average-aged drivers (Zeng et al., 2016).

Ji and Levinson (2020) in USA, analyzed some popular ML models and some ensemble techniques such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), Random Forest (RF), and AdaBoost (Adaptive Boosting) for predicting occupant injuries collision between two vehicles and result showed that, combining models can perform better than individual model (Ji, A., and D. Levinson, 2020).

Yang et al. (2019) in Japan, examined the critical variables that determine the severity of driver injuries in two vehicle crashes, passenger car and truck, using bivariate ordered probit model and found that, time of the day, locations, traffic conditions, types of collisions, and types of roads have different effects on the two vehicle crashes where the weather condition and age of drivers have relatable impacts for two types of crashes (Jamal et al., 2021).

Duncan et al. (1998) in USA, studied the effect of numerous factors on the injuries sustained by occupants in two vehicle incidents by using the ordered probit model and found that the factors responsible for injuries to passenger vehicle occupants in rear-end collisions on split roads are darkness, high speed differentials, high speed limits, grades, particularly if they are damp (Fan et al., 2019).

Sobhani et al. (2011) in Australia, aimed to measure the extent of injuries to people in road crashes involving two vehicles using a Log-Gamma regression model and result showed that, the interaction of the type of impact, presence of airbag, presence of seat belt, and age of the occupants are the triggering factors for the crash injury severity. It also showed that the severity of crash injury is higher for crashes where the airbags and seat belts are available for near side crashes rather than crashes on far side and front crashes (Liao et al., 2018).

Previous research has discovered that accident type is one of the determining elements of two vehicle crash severity (Yang et al., 2019; Duncan et al., 1998; Chiou et al., 2020; Lee, C., and X. Li, 2014 and Ji, A., and D. Levinson, 2020). Vehicle factors such as vehicle type and vehicle movement have also been demonstrated to have a substantial impact on the severity of a two vehicle crash (Yuan et al., 2022 and Zeng et al., 2016). The characteristics of the roadway and surrounding environment (for example, traffic conditions, road types, and lighting conditions) can explain the level of crash severity (Yang et al., 2019; Duncan et al., 1998). According to certain research, the temporal characteristics of accidents (i.e., day of week, specific time of day) and the level of alcohol intake of drivers are also connected with the level of severity in two vehicle crashes (Yang et al., 2019 and Chiou et al., 2020). However, these conclusions are derived from the setting of high-income countries, despite the fact that the hotspot of accidents is in low- and middle-income countries. Because of the differences in their contexts, further research should be conducted in low- and middle-income areas to build an effective and comprehensive preventive strategy.

2.3 Statistical Methods in Crash Severity Modeling

The four main ways in which statistical methods aid classification are, : developing probability models for data and classes to identify probable classifications for a given set of data; creating tests of validity of specific classes produced by a classification scheme; contrasting the relative efficacy of various classification schemes; and increasing the search for ideal classifications by probability- based research techniques. Algorithms for standard hierarchical and splitting data are analyzed statistically (J.A. Hartigan, 2001). Statistical models have traditionally been the most frequently used techniques for analyzing crash injury severity. Several parametric statistical techniques have been employed in previous studies to model the severity of a two-vehicle incident in an attempt to uncover potential risk factors for death. The most popular modeling method among them is probably the ordered probit (OP) model (C. Lee and M. Abdel-Aty, 2005, N. Siddiqui et al., 2006, K. K. W. Yau, 2006, Y. Xie, 2009 and Wang et al., 2011) which may order the severity of an accident into categories (Yang et al., 2019, Duncan et al., 1998 and Chiou et al., 2020). Savolainen et al. (2011) provided a thorough literature review of approaches to assessing crash injuries. In the literature, multinomial logit (MNL) model (V. Shankar and F. Mannering, 1996, A. Khorashadi, 2005 and P. Savolainen and F. Mannering, 2007), the binary logit (BL) model (A. SAl-Ghamdi, 2002), Logistic Regression (LR) (J.A. Hartigan, 2001), Mixed Logit Model, and Binomial Regression Model are all frequently used methodologies (Kitali, 2021, Yuan et al., 2022, Lee, C., and X. Li, 2014 and Zeng et al., 2016).

Crash injury severity modeling has made extensive use of various types of statistical regression. Multi-level ordered logit models, binomial logistic models, and their variants are the most popular regression prediction techniques. Regression models like these can analytically shed light on the connections between various factors and provide a plausible theoretical interpretation (J. Tang et al., 2019). However, they require that you adhere to certain mathematical forms when relating your dependent and explanatory variables (Z. Li et al., 2012). Due to their inflexible premise, regression models have poorer predictive power than other algorithms (A. Iranitalab and A. Khattak, 2017). There are some caveats to using statistical models, despite the fact that their mathematical interpretation is sound and they help shed light on the part played by various predictor variables. To begin with, they are predicated on a set of presumptions (with regards to linear link functions and error distribution terms) and a predefined relationship between the variables, and violating any of these can lead to skewed model estimation (Ullah et al., 2021; Zahid, Chen, Jamal, Al-Ahmadi, et al., 2020; Zahid, Chen, Jamal, and Memon, 2020). Second, they have low reliability and poor prediction accuracy. In addition, class imbalance problems have been identified in statistical models that use past crash data (Vilaça et al., 2019; Wang et al., 2019; Elamrani Abou Elassad et al., 2020). Statistical procedures assume, by definition, that classrooms are evenly populated (Leevy et al., 2018). Hyden's safety pyramid describes the distribution of crash severity, which is a common example of a dataset with class imbalance (Laureshyn et al., 2010). The results have been shown to support the majority class, leading to biased predictions and even misleading conclusions if there is major class imbalance (rare events are under about 5%) in the datasets (Ferrari & Bacciu, 2021; King & Zeng, 2001). Analyzing the efficacy of various sampling strategies for addressing class imbalance using crash datasets is crucial for reducing the impact of class imbalance problems. Researchers in the past have turned to machine learning techniques for predicting crash injuries' severity (M. A. Abdel-Aty and H. T. Abdelwahab, 2004, L.-Y. Chang and H.-W. Wang, 2006, Y. Xie et al., 2007, J. de Oña, 2011, J. Abellán, 2013 and A. Iranitalab and A. Khattak, 2017) to get around the shortcomings of statistical models.

The output of statistical models is often straight forward formulas that illustrate the relationships between the dependent and explanatory variables. However, in spite of their solid conceptual foundation, statistical models have certain caveats. A linear function is used to establish a connection between the dependent variable and the explanatory factors in statistical modeling, for instance, and this technique needs assumptions about the data distribution. There is no guarantee that all of those premises hold true. Incorrect parameter estimates will be generated if such assumptions are broken (L. Mussone et al., 1999 and C. Chen et al., 2016). Model estimate may also be impacted by other factors, like multicollinearity (S. Washington et al., 2010), within-crash correlation (P. Savolainen and F. Mannering, 2007 and R. Paleti et al., 2010) and unobserved heterogeneity (P. Xu and H. Huang, 2015 and Z. Li et al., 2013). Complex frameworks are often needed to mitigate the destructive effects of such problems, making the corresponding statistical models challenging to solve. (P. T. Savolainen et al., 2011). Conventional parametric regression models, for example, require linear functions to link the response variable to determinants and rely on a specific distribution of crash data. There are instances when these assumptions can lead to incorrect estimations and biased model inferences (Jamal et al., 2021). These data-related shortcomings, which are a prevalent limitation in parametric regression models, can be eliminated by using ML algorithms, which have the ability to uncover key factors and enhance prediction accuracy (Raihan et al. 2017, Fan et al., 2019 and Liao et al., 2018).

The severity of a crash injury is often indicated by different classes such as damage to property, possible injuries, capacitating injury, incapacitating injury, fatality, and so on. Damage to property only (PDO)/no injuries, injuries, and deaths have all been utilized as injury severity groups in numerous researches (C. Ma et al., 2018 and Mesa-Arango et al., 2018). Because crash injury severity levels are discrete, discrete outcome models such as binary or multinomial logit/probit models have been extensively used (Azimi et al., 2020; Rifaat & Chin, 2007; Shankar & Mannering, 1996; Yu & Abdel-Aty, 2014a). To accommodate for variability and causality, as well as the ordinal character of within-crash correlation a number of complex models as Bayesian hierarchical (Huang et al., 2008; Li et al. 2018), ordered logit models (Azimi et al., 2020; C. Chen et al., 2016; Khattak et al., 1998; O'Donnell & Connor, 1996), bivariate/multivariate models (Aguero-Valverde & Jovanis, 2009; C. Lee & Abdel-Aty, 2008; Russo et al., 2014; Zeng et al., 2017), nested logit model (Osman et al., 2016; Shankar et al., 1996), random parameter model (Milton et al., 2008; J. Wang et al., 2020), Markov switching multinomial model (Malyshkina & Mannering, 2009; Xiong et al., 2014), and their mixed versions (Christoforou et al., 2010; Eluru & Bhat, 2007; Huang et al., 2011; Li et al. 2019), were reviewed.

2.4 Machine Learning (ML) Methods in Crash Severity Modeling

The term "machine learning" (ML) refers to an approach to "learning" by analyzing data. Discovering regularities in the information is a key part of this process. Predictions and classifications are aided greatly by the robust algorithms made available by Machine Learning (ML). The goal of ML models is to increase prediction precision via a non-parametric method (L. Wahab and H. Jiang, 2019). Researchers have paid close attention to machine learning (ML) methods over the past two decades due to their rapid development and accurate regression and classification performance. More and more studies have used ML techniques to examine crash severity. Unlike conventional statistical methods, which have rigid and well-defined functional forms, ML approaches are extremely adaptable, make few if any assumptions about the crash severity data, and can deal with missing values, noises, and outliers (Tang et al., 2019). The models used in machine learning make no assumptions about the connections between different variables. Some studies have found that machine learning techniques outperform statistical ones at producing fitting.

In order to generate predictions or choices without being explicitly taught to do so, machine learning algorithms develop a mathematical model based on sample data, often known as training data (Bishop, 2006). The challenge of concentrating on the most pertinent information in a potentially overwhelming amount of data has become more significant as machine learning attempts to tackle bigger, more complex tasks.

Machine learning-based models have appeared as valuable technology in road safety studies in recent years, overcoming the drawbacks of statistical methods due to rapid advances in soft computing methods. However, there are still data and methodology issues with ML that have yet to be resolved. To begin, traffic crash severity datasets are inherently imbalanced and, in some cases, under reported. Many studies have found that while ML methods frequently produce high overall prediction accuracy, they produce poor accuracy for severity categories with fewer observations, such as potentially deadly and serious accidents (Abdel-Aty & Abdelwahab, 2004; Chang &

Wang, 2006; Chene et al., 2016a; Chene et al., 2016b; Lie et al., 2012). Second, most ML approaches suffer from the "black-box" problem, in which it is unclear how to interpret the modeling results and derive the underlying relationships between independent/explanatory variables and crash severity outcomes. Sensitivity analysis (SA), and more specifically local sensitivity analysis (LSA), has been implemented to get around this issue. In order to reduce the 'black box' effect, researchers have developed techniques like sensitivity analysis (M. B. Anvari et al., 2017; R. Yu and M. Abdel-Aty, 2014; A. Das et al., 2009; J. Zhang et al., 2018; L. Jiang et al., 2019; X. Li et al., 2008; Li et al. 2012 and Y. Zhang & Xie, 2007). An application of sensitivity analysis is in extracting features and ranking the relative importance of different variables in relation to a given target variable. It's made it much easier to implement ML models into research on vehicular safety. To capture the joint effects of multiple risk factors, however, sensitivity analysis must make the potentially false assumptions of linearity, normality, and local variations. When applying ML approaches to crash severity analysis, additional data/methodology-related issues, such as model performance metrics, crash spatiotemporal correlations, causality, transferability, and heterogeneity, often arise.

To address the shortcomings of statistical approaches, various machine learning (ML) models are being investigated for modeling possibly nonlinear correlations between accident contributing elements and injury severity outcomes. (Abdel-Aty & Abdelwahab, 2004; Iranitalab & Khattak, 2017; Li et al. 2012; Pradhan & Sameen, 2020; Sameen & Pradhan, 2017; Sarkar et al., 2020; Tang et al., 2019) Machine learning models have the benefit of being more adaptable to processing outliers, noisy, or missing data, as well as being more flexible with no or few post assumptions for input variables.

When compared to statistical methods, ML methods are said to fit better. Machine learning models have been extensively used to predict the severity of traffic accidents. (M. Taamneh et al., 2016). Some widely used ML algorithms in crash severity modelling domain are: Artificial Neural Networks (ANN) (Abdelwahab & Abdel-Aty, 2001; Amiri et al., 2020; Zeng & Huang, 2014), Support Vector Machines (SVM) (Dong et al., 2015; Mokhtarimousavi et al., 2019; Zhibin Li et al. 2012), Decision Trees (DT) (Abellán et al., 2013; Oña et al., 2013; P. Lu et al., 2020), K-means

Clustering (KC) (Anderson, 2009; Fiorentini & Losa, 2020; Mauro et al., 2013), Random Forest (Iranitalab & Khattak, 2017; Mondal et al., 2020; J. Zhang et al., 2018), and Naïve Bayes (Arhin & Gatiba, 2020; Budiawan et al., 2019; C. Chen et al., 2016).

The comprehensive literature review shows that most of the previous studies have been done based on parametric regression whereas there are few studies which are based on Machine Learning (ML). To the best knowledge, Ji and Levinson (Ji, A., and D. Levinson , 2020) in USA, analyzed some popular ML algorithms for predicting occupant injuries collision between two vehicles.

2.5 Ensemble or Hybrid Methods in Crash Severity Modeling

In comparison to any of the individual classifiers, ensemble learning improves prediction accuracy by combining a number of weak classifiers.

Bagging (also known as random forests) and boosting are the two main ensemble learning techniques used in crash severity analysis (Wen et al., 2021).

Boosting and Bagging are two examples of ensemble methods, which are a type of cutting-edge learning strategy in which multiple learners are trained separately and then combined for application. It is common knowledge that multiple learners in an ensemble can improve accuracy significantly over a single learner, and ensemble methods have seen huge success in many practical uses (Zhi-Hua Zhou, Ensemble Methods Foundations and Algorithms, 2012). Liu, L. et al (2020) applied an ensemble model (CSSV-AGX) of AdaBoost (Adaptive Boosting), GBDT (Gradient Boosting Decision Tree), and XGBoost (eXtreme Gradient Boosting) based on Classifier-specific Soft Voting and Several major factors have been analyzed to determine connections between accident factors (like speed) and the distribution of various occupant accident severity levels. The intense Gradient Boosting (XGBoost) model was studied by Jamal, A. et al. (2021) to determine if it could be used for analyzing crash injuries more accurately than more conventional machine learning algorithms like logistic regression, random forest, and decision tree.

2.6 Summary

There are some significant differences of this study from previous studies. Firstly, this study has employed the Shapley Additive explanations (SHAP) approach, developed by Lundberg and Lee (Lundberg, S. M., and S.-I. Lee, 2017), to explain the model's output. An individual feature's contribution, based on its marginal contribution, to the predicted value can be calculated using SHAP (Parsa et al., 2019). Secondly, this study has used feature engineering technique to extract useful features. An essential and challenging problem in ML is deciding on the best subset of features to use. Choosing a decent collection of attributes reduces computational load while simultaneously improving accuracy. Noisy data can lead classifiers to form inaccurate connections, and redundant or linked features raise classification intricacy without delivering innovative information to the system (Moons et al., 2016 and Wang et al., 2017). In this study, RF model has been used to select the effective features. Thirdly, in this study, the Classifier-specific Soft Voting has been employed to integrate individual models in predicting the two-vehicle crash severity. Soft voting employs class-specific weights to boost combinatorial performance while reducing computing cost. On top of all that, it improves classifier weightings by taking into account both soft class probabilities (Cao et al., 2015). Fourthly, the majority of prior studies evaluated the influence of vehicle features on crash severity at the accident level, but they failed to consider the effect of diverse roles in an accident (Yuan et al., 2022). In this study, besides vehicle characteristics (vehicle type, vehicle maneuver), it has been also focused on driver characteristics (age, sobriety condition, seatbelt/helmet usage), roadway conditions/environment (road geometry, surface condition, light condition, junction, road class, traffic control, movement), crash characteristics (collision type) and temporal features (time, day of week) in predicting two vehicle crash severities. One of the benefits of using many explanatory factors is that it allows models overcome the biasness associated with the absence of potential independent variables. In this study different machine learning (ML) methods and traditional statistical regression method have been applied to identify the contributing factors responsible for such crashes and also compared the prediction accuracy of such models.

Chapter 3

DATA COLLECTON AND METHODOLOGY

3.1 Introduction

Several machine learning methods, as well as a statistical method, were used in this study and also employed a hybrid model and evaluate the best performing model in predicting two vehicle crash severities. This chapter gives a brief but comprehensive description of these methods, as well as their applicability. The chapter also discusses the data collection and preparation procedures for this study.

3.2 Methods and Work Flow of the Study

The following Figure 3.1 depicts the entire operational framework of this thesis.

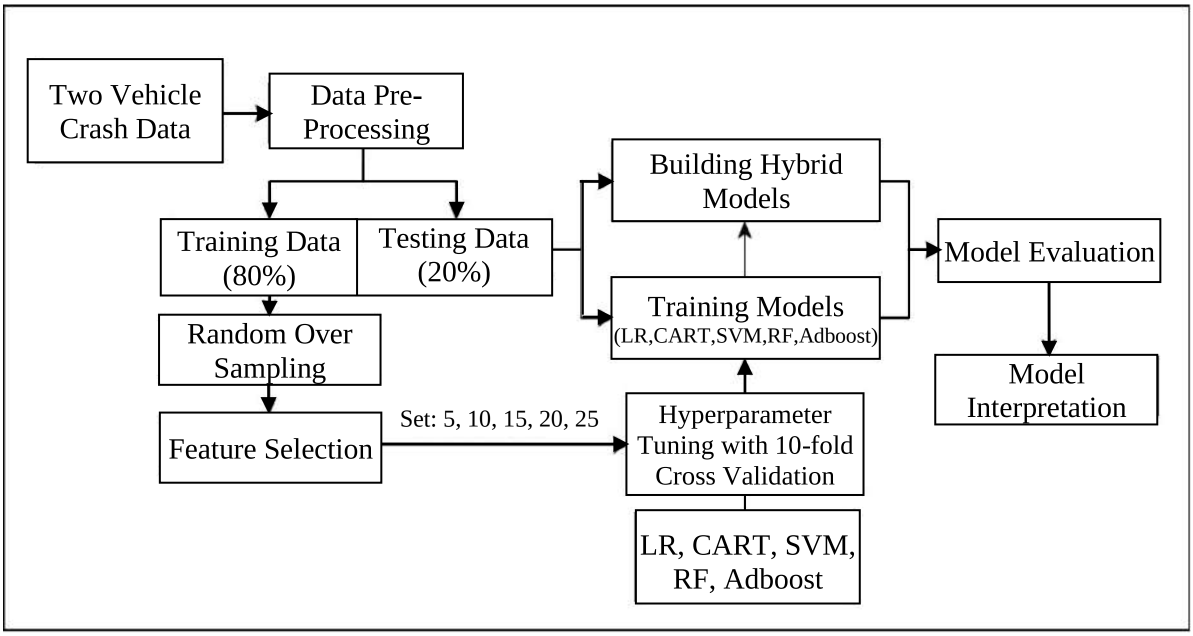


Figure 3.1: Work methodology of the study

The work flow of this study can be listed as followed:

(i) The filtered data of two vehicle crashes has been divided into training dataset (80%) and testing dataset (20%) and trained the models through a 10-fold cross-validation. To resolve data imbalance issue, method of random over-sampling (ROS) has been adopted on training dataset.

(ii) Important features have been selected using Random Forest (RF) and all the models (ML, Statistical method and all possible hybrid models) have been trained with all possible combinations of feature subset to assess the performance of different predictors. Hyper parameter tuning has been done with 10-fold cross validation.

(iii) The efficiency of the various models has been evaluated utilizing the most commonly known performance metrics and compared the prediction accuracy of different classifiers using Confusion Matrix.

(iv) An individual feature's contribution, based on its marginal contribution, to the predicted value has been calculated using SHAP.

Following sections describe the methods sequentially.

3.3 Data Description

This study has been utilized the recent crash data of ARI for two vehicle crash severity prediction analysis. The overall data processing and description have been mentioned in the following sections.

3.3.1 Data Overview

In Bangladesh, police are mainly concerned with the collection of crash data at field level in Accident Report Form (ARF) (Appendix-A). A guideline for filling the Accident Report Form (ARF) has been published by Accident Research Institute (ARI) of Bangladesh University of Engineering and Technology (BUET) (Appendix-B). The reports are then recorded in the Microcomputer Accident Analysis Package 5 (MAAP5) repository by the Accident Research Institute (ARI) of Bangladesh University of Engineering and Technology (BUET). In our study, crash data of 2017-2020 for Dhaka city have been collected from ARI, BUET. At time of collection, these data were available in a tabular format with each sample containing information about crash, roadway geometry, environment, vehicle, driver, passenger, and pedestrian characteristics.

3.3.2 Data Preparation

Since this study focuses on modeling two-vehicle crashes, the very first task of processing was filtering out crashes that involved only two vehicles. The Excel database extracted from MAAP5 database needed some further processing to be used for machine learning. At first, the Excel format was converted to CSV (comma delimited) format; so that it can be imported by the python software. Later, it was found that the computer that was designated for the modeling could not handle this huge database. Therefore, it became urgent to reduce the size of this accident database.

There were numerous crash data that were collected from Accident Research Institute (ARI), BUET (2017-2020). This study, which focuses on two-vehicle crashes, was reduced from 357 variables to 25 variables by eliminating the irrelevant variables. Index system has been adopted to eliminate single and multi-vehicle crash data. Eventually total 1494 crash data have been reduced to 692 crash data which was further reduced to 329 crash data for better accuracy.

Different independent variables like vehicle characteristics (vehicle type, vehicle maneuver), driver characteristics (age, sobriety condition, seatbelt/helmet usage), roadway conditions/environment (road geometry, surface condition, traffic control, movement, light condition, junction, road class,), crash characteristics (collision type) and temporal features (time, day of week) have been used in predicting two vehicle crash severities. Summary of the filtered data with information on crash, roadway geometry, environment, temporal, vehicle, and driver characteristics have been presented in Table 3.1. A total of 658 drivers were involved in 329 traffic crashes reported during the years 2017-2020 in Dhaka. The dataset then divided into 80% ‘Training’ dataset with 526 drivers and the 20% ‘Testing’ dataset with remaining 132 drivers.

Table 3.1: The Descriptive Statistics of Variables

Variable	Variable Description	Frequency	Ratio (%)
Target Variable			
Injury Severity	0 - Non-Fatal	468	71.1
	1 – Fatal	190	28.9
Explanatory Variables			
Crash Characteristics			
Collision Type	not a rear end collision	288	43.7
	a rear end collision	370	56.3
Roadway Characteristics			
Junction	no junction was present	326	49.5
	junction was present	332	50.5
Traffic Control	no traffic control system is present	142	21.5
	traffic control system is present	516	78.5
Movement	the road was one way	332	50.5
	the road was two-way	326	49.5
Surface Condition	the road surface was dry	632	96
	the road surface was not dry	26	4
Road Geometry	the road was not straight	66	10
	the road was straight	592	90
Road Class	the road flows within the city	340	51.7
	the road flows to outside of the city	318	48.3
Environment Characteristics			
Light Condition	dawn/dusk	108	16.4
	Daylight	312	47.4
	Dark	238	36.2
Temporal Characteristics			
Day of Week	Weekday	492	74.8
	Weekend	166	25.2
Time	during night hours	258	39.2
	during off-peak hours	168	25.5
	during peak hours	232	35.3

Vehicle Characteristics			
Vehicle Type	one of the involved vehicles was a bus	190	28.9
	one of the involved vehicles was a car	55	8.4
	one of the involved vehicles was a	162	24.6
	one of the involved vehicles was an NMV	62	9.4
	one of the involved vehicles was a pick-up	25	3.8
	one of the involved vehicles was a truck	73	11.1
	one of the involved vehicles was a van/SUV	23	3.5
	one of the involved vehicles was any other	68	10.3
Vehicle Maneuver	vehicle was going straight	435	66.1
	vehicle was not going straight	223	33.9
Driver Characteristics			
Driver Age	driver age <=30	231	35.1
	driver age >50	14	2.1
	driver age 31-40	295	44.8
	driver age 41-50	118	18
Sobriety Condition	driver was not suspected drunk	591	89.8
	driver was suspected drunk	67	10.2
Seatbelt/Helmet	driver/biker did not-worn seatbelt/helmet	533	81
	driver/biker worn seatbelt/helmet	125	19

Each feature has been categorized for better distribution and more accuracy. According to the severity level, the distribution of each individual category has been observed through further visualization.

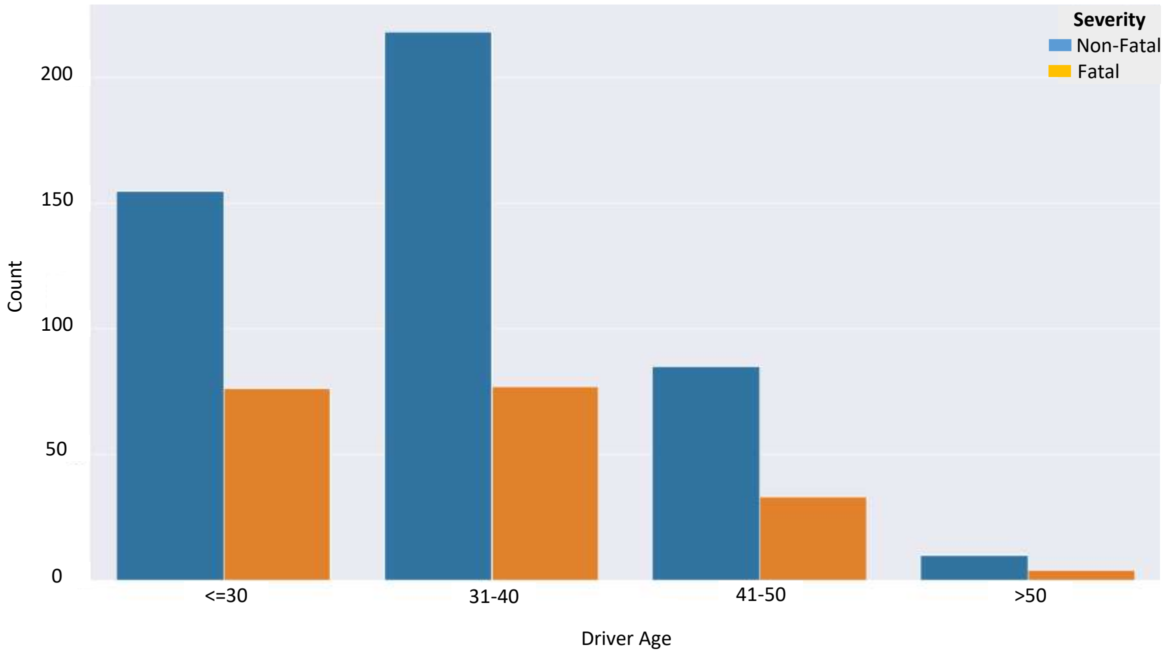


Figure 3.2: Distribution Graph of Driver Age

After reviewing previous studies, in this study, driver's age has been categorized in four categories "age<=30", "age=31-40", "age=41-50" and "age>50". From Figure 3.2, it has been found that total counts of two vehicle crashes were highest for 31-40 aged drivers. Crashes has found lowest for drivers' aged greater than 50.

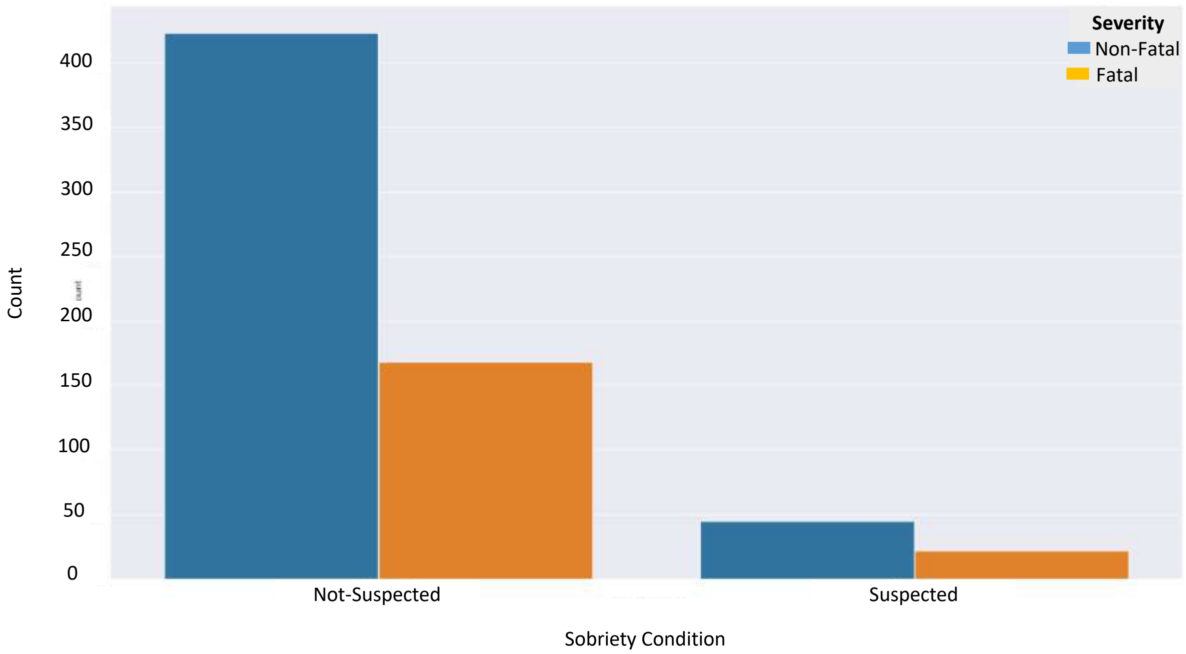


Figure 3.3: Distribution Graph of Sobriety Condition

Driver's sobriety condition has been categorized in two categories "suspected" and "non-suspected". From Figure 3.3, it can be identified that the total counts of two vehicle crashes were high for sobriety condition "not suspected". For "non-suspected" category the rate of fatality has found about one-third compared with the non-fatality rate. Whereas, the fatality rate was almost half of the non-fatality rate in case of "suspected" condition.

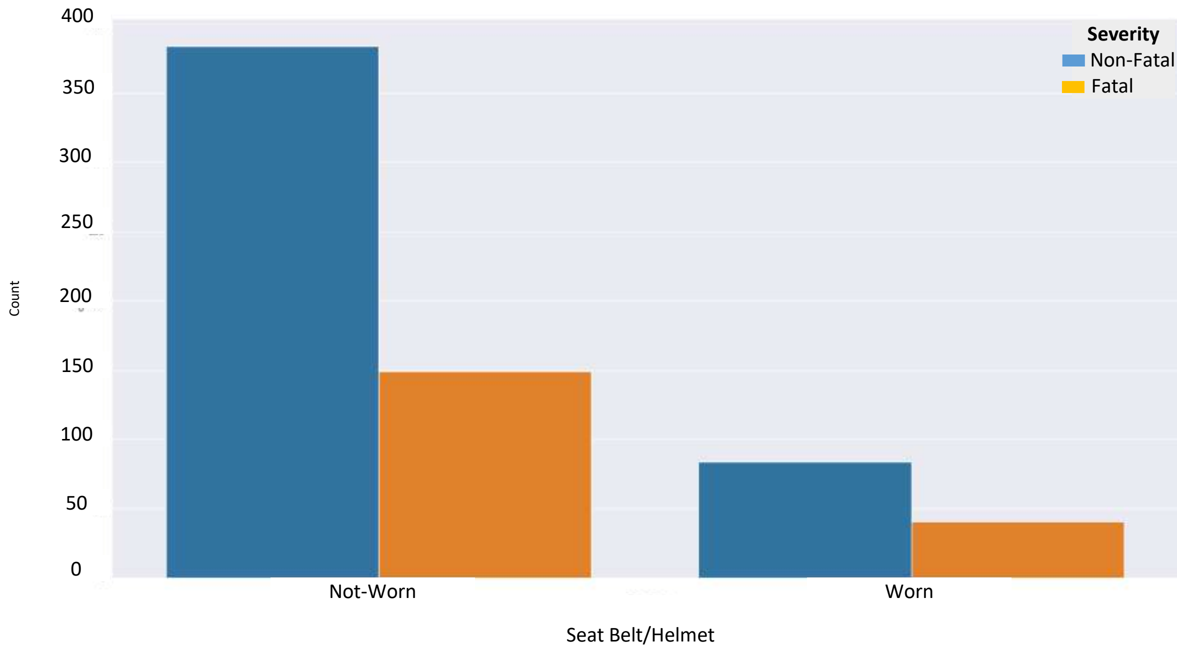


Figure 3.4: Distribution Graph of Seatbelt/Helmet Use

Driver/Biker's seatbelt/helmet use condition has been categorized in two categories "worn" and "not-worn". From Figure 3.4, it can be identified that the total counts of two vehicle crashes were high for the drivers/bikers who did not wear seatbelt/helmet.

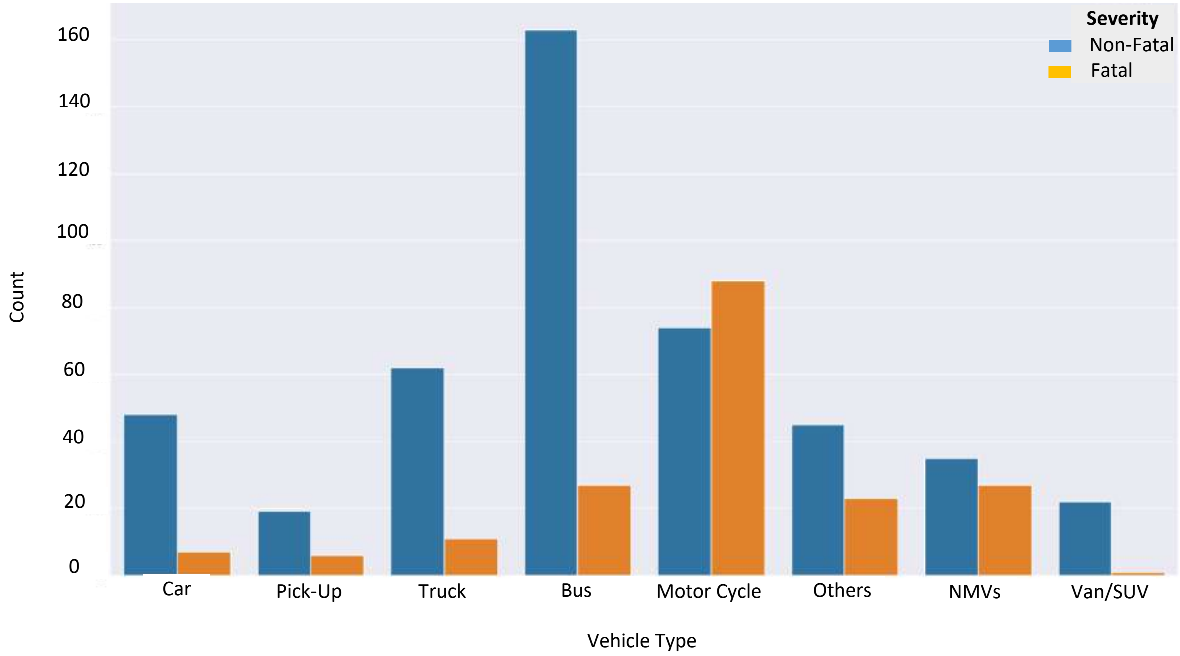


Figure 3.5: Distribution Graph of Vehicle Type

The variable "vehicle type" has been categorized in eight categories bus, car, motorcycle, NMV, pick-up, truck, van/SUV and others. From Figure 3.5, it has been found that the total counts of two vehicle crashes were high if one of the vehicles involved in the crash was bus whereas, the rate of fatality has found maximum when one of the vehicles involved in two vehicle crashes was motorcycle.

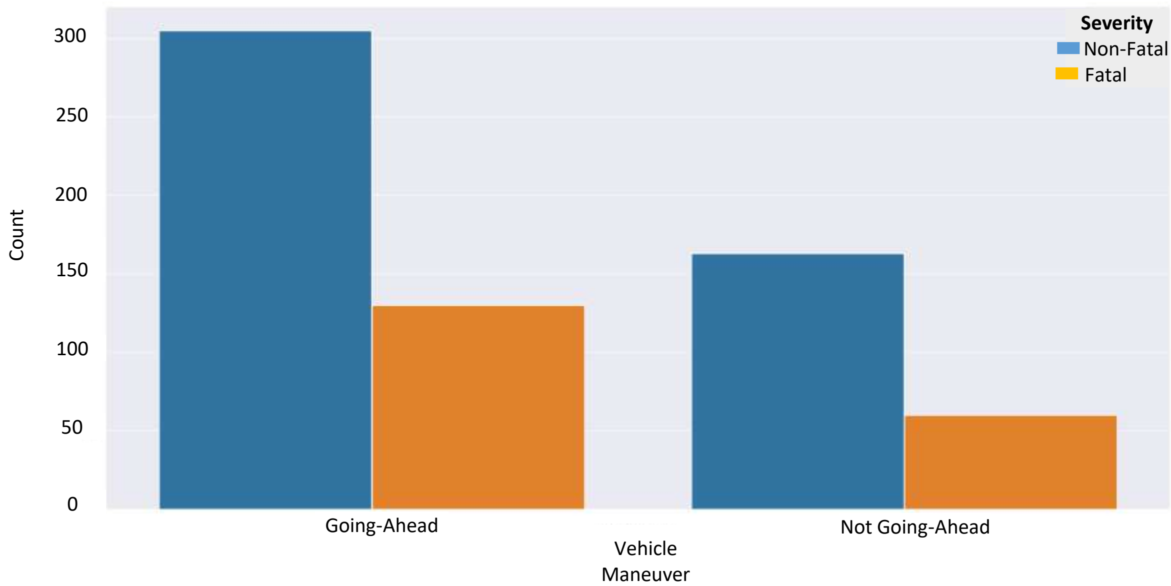


Figure 3.6: Distribution Graph of Vehicle Maneuver

The variable "vehicle maneuver" has been categorized in two categories "going ahead" and "not going ahead". From Figure 3.6, it can be identified that the total counts of two vehicle crashes were high for "going ahead" maneuver.

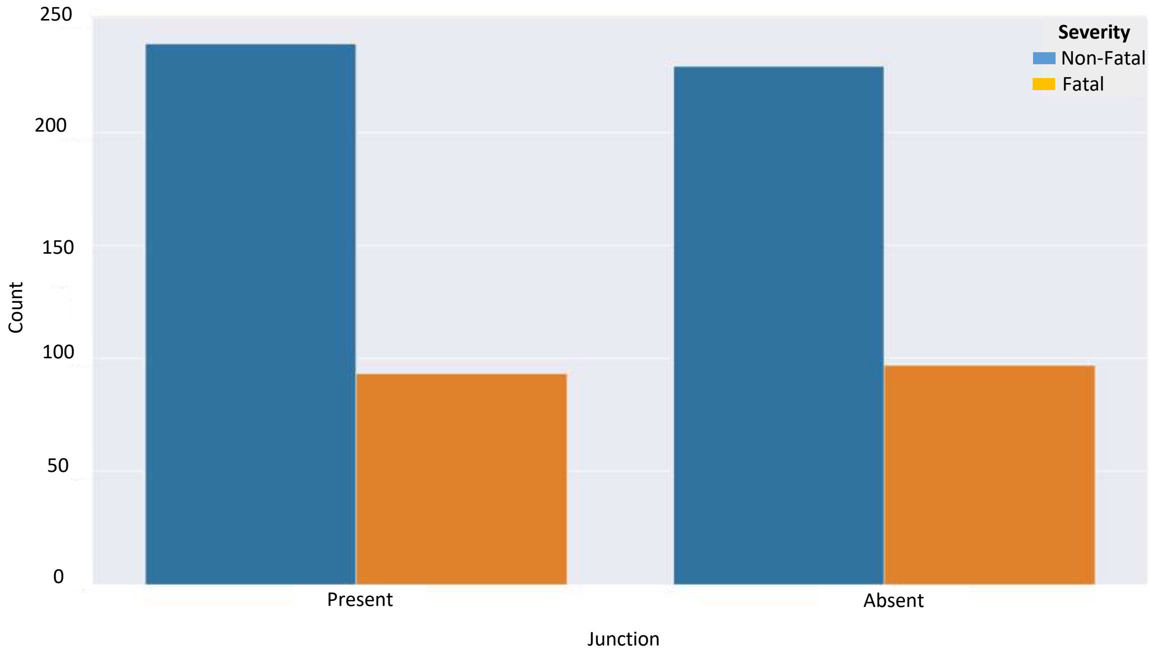


Figure 3.7: Distribution Graph of Junction

The roadway characteristics "junction" has been categorized in two categories "present" and "absent". From Figure 3.7, it can be identified the total counts of two vehicle crashes were nearly similar for both the conditions.

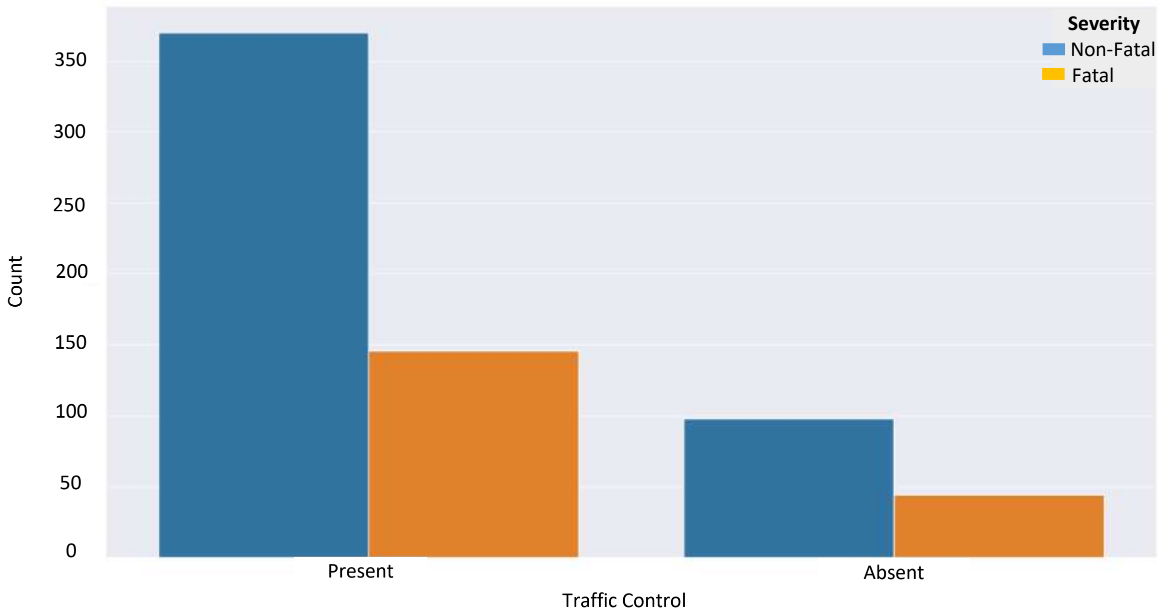


Figure 3.8: Distribution Graph of Traffic Control

The roadway characteristics “traffic control” has been categorized in two categories “present” and “absent”. From Figure 3.8, it can be identified the total counts of two vehicle crashes were high where traffic control was present.

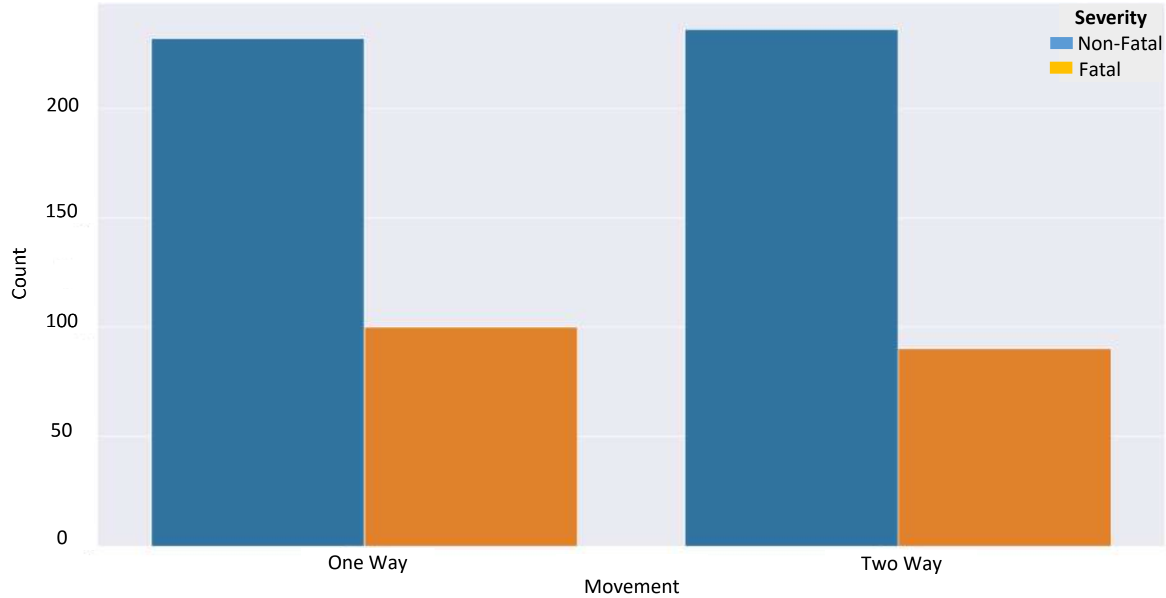


Figure 3.9: Distribution Graph of Movement

The roadway characteristics “traffic movement” has been categorized in two categories “one-way” and “two-way”. From Figure 3.9, it can be identified that the total counts of two vehicle crashes were almost similar for both one-way and two-way movements.

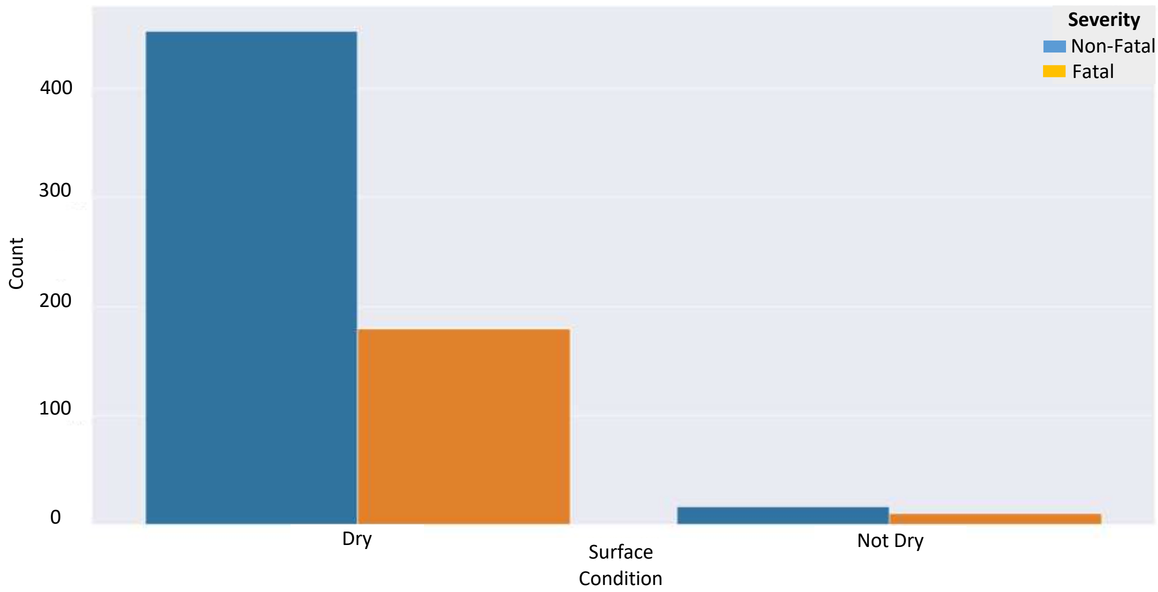


Figure 3.10: Distribution Graph of Surface Condition

The roadway characteristics “surface condition” has been categorized in two categories “dry” and “not dry”. From Figure 3.10, it can be identified that the total counts of two vehicle crashes were much higher when the surface was dry.

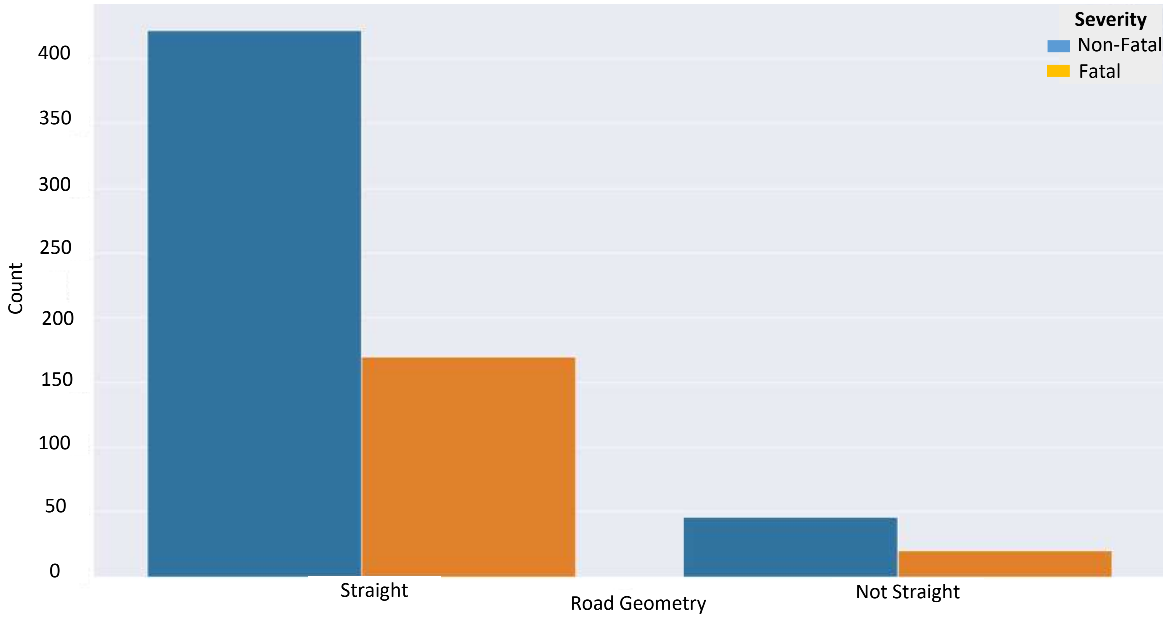


Figure 3.11: Distribution Graph of Road Geometry

The roadway characteristics “road geometry” has been categorized in two categories “straight” and “not straight”. From Figure 3.11, it can be identified that the total counts of two vehicle crashes were much higher when the road geometry was straight.

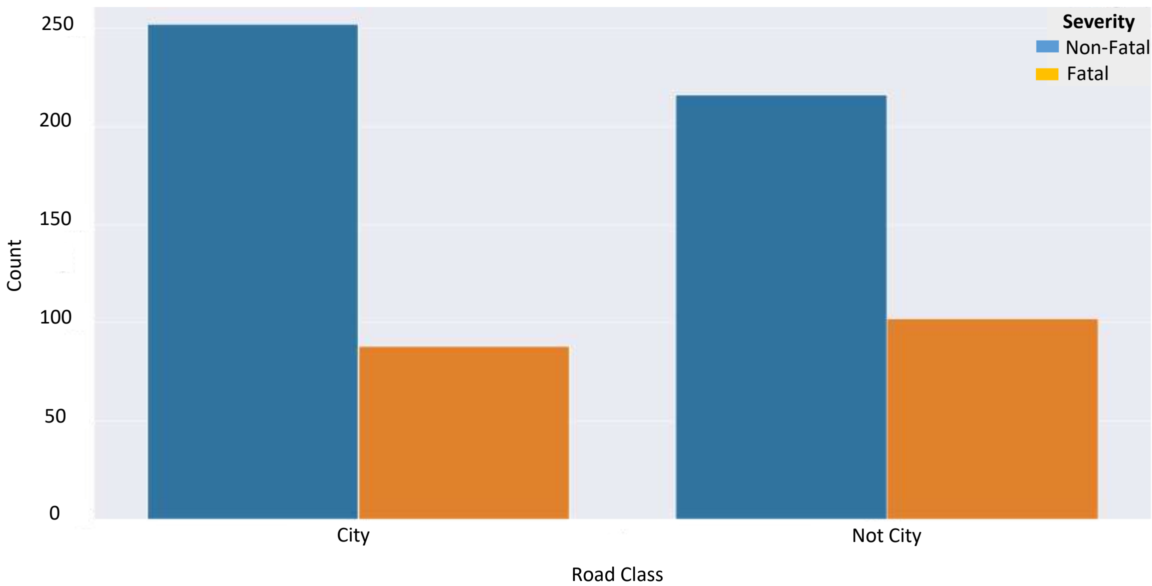


Figure 3.12: Distribution Graph of Road Class

The roadway characteristics “road class” has been categorized in two categories “city” and “not city”. From Figure 3.12, it can be identified that the total counts of two vehicle crashes were nearly similar for both road class categories.

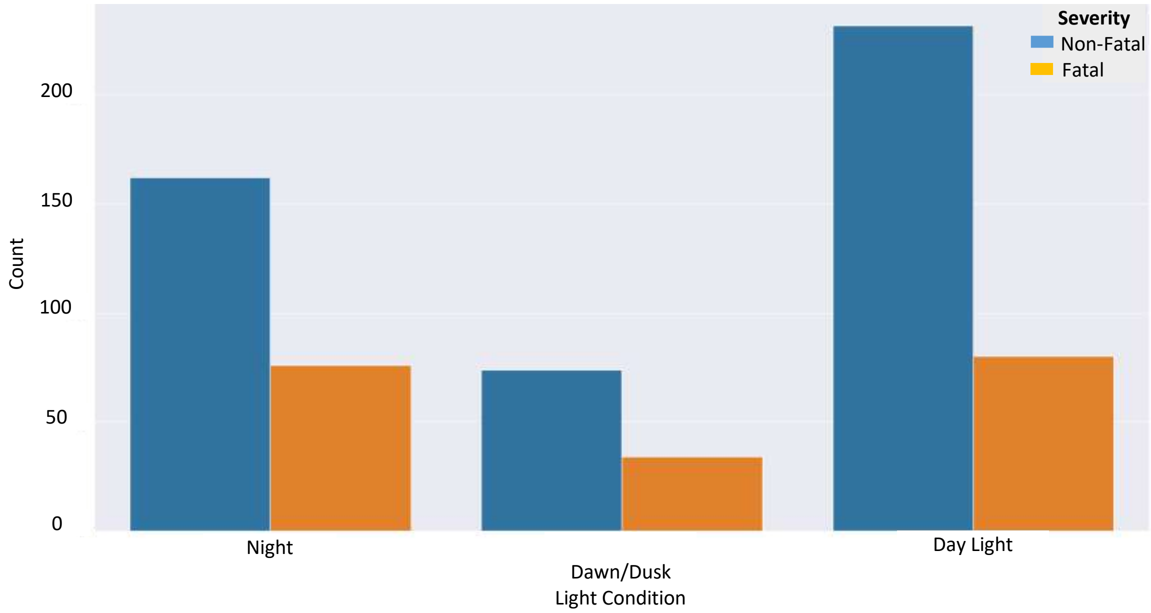


Figure 3.13: Distribution Graph of Light Condition

The environmental characteristics “light condition” has been categorized in three categories “night”, “dawn/dusk” and “daylight”. From Figure 3.13, it can be identified that the total counts of two vehicle crashes were slightly high in daylight condition than in night condition. Number of crashes have found low in dawn/dusk condition.

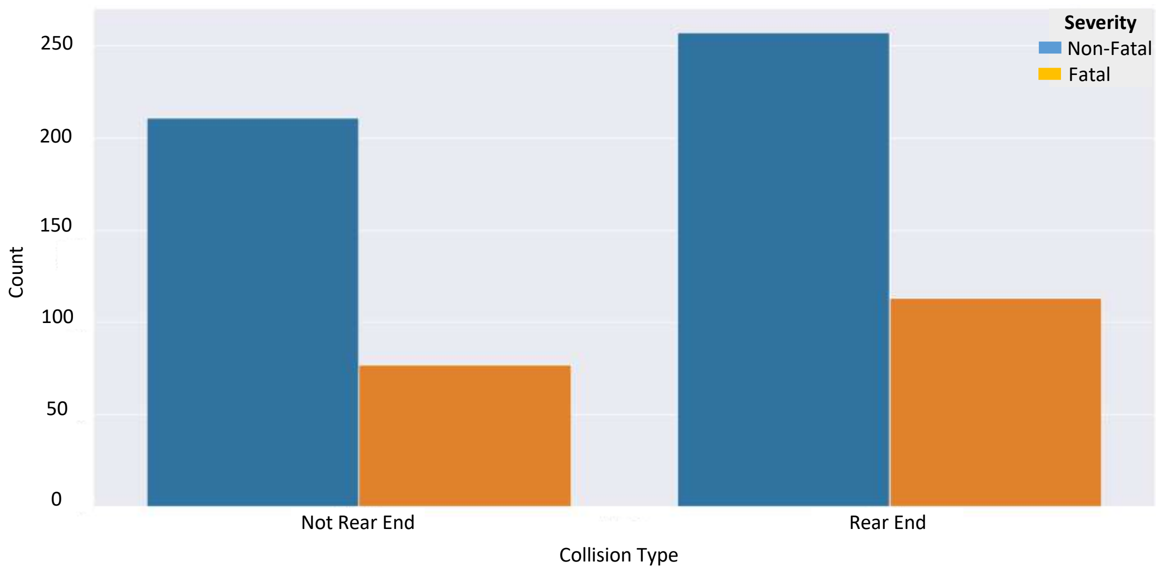


Figure 3.14: Distribution Graph of Collision Type

The crash characteristics “collision type” has been categorized in two categories “not rear end” and “rear end”. From Figure 3.14, it can be identified that the total counts of two vehicle crashes were high for rear end collision.

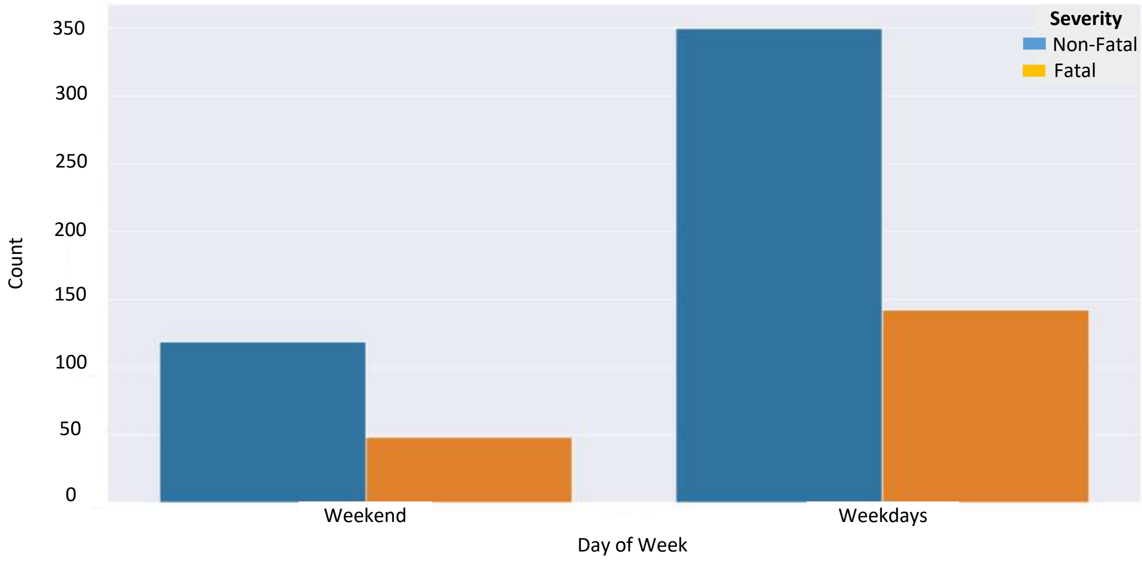


Figure 3.15: Distribution Graph of Day of Week

The temporal characteristics “day of week” has been categorized in two categories “weekday” and “weekend”. From Figure 3.15, it can be identified that the total counts of two vehicle crashes were high while the accidents took place on a weekday.

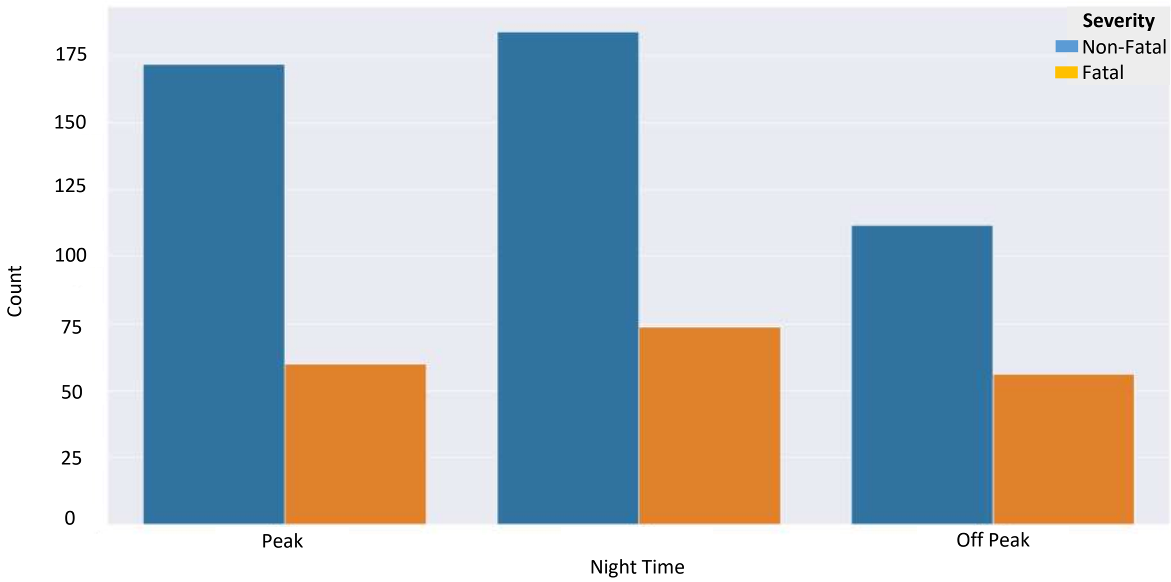


Figure 3.16: Distribution Graph of Time

The temporal characteristics “time” has been categorized in three categories “peak hours”, “off-peak hours” and “night hours”. From Figure 3.16, it can be identified that the total counts of two vehicle crashes were high while the accidents took place at night time and low at off-peak hours.

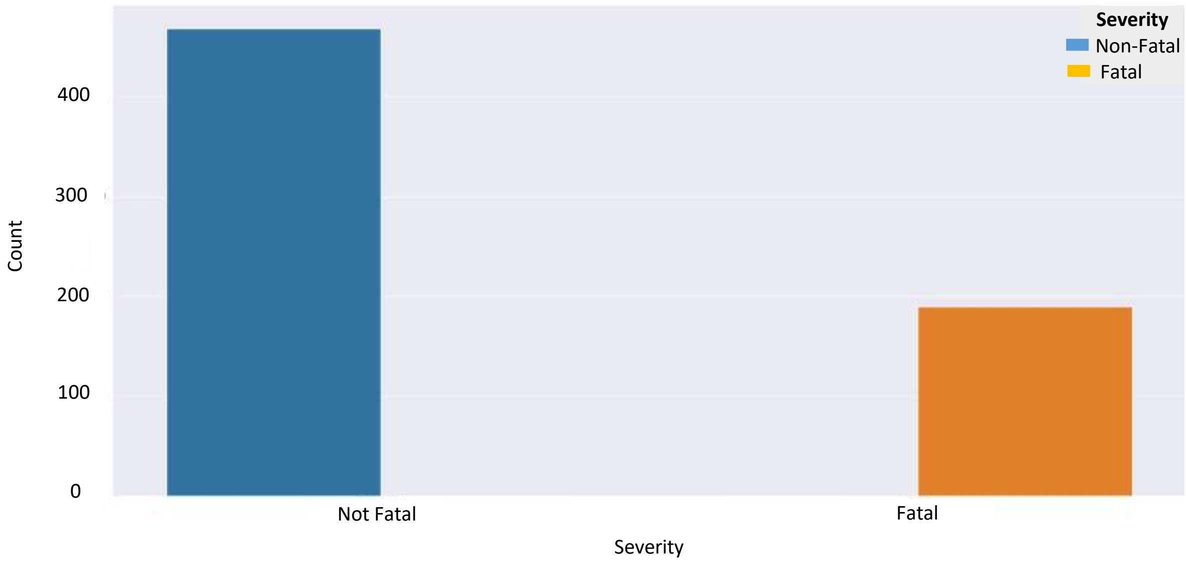


Figure 3.17: Distribution Graph of Severity

Among four injury severity categories such as fatal, grievous, simple injury and non-fatal, this study has used two injury severity categories: fatal and non-fatal. However, a major issue of data imbalance was observed in the dataset. As shown in Figure 3.17, a sum of 468 drivers' involvement resulted in non-fatal (NF), 190 drivers' involvement resulted in fatal (F) (Non-Fatal: 71%, Fatal: 29%). Further re-sampling has been done for balancing the data.

Before re-sampling in this study, as all the variables were categorical, hot coding or dummy variables have been created. In machine learning, categorical variables (such as geometric characteristics) can only range over a series of fixed values. Generally, a feature of *k* possible values needs to be encoded as a set of *k* derived dummy variables so that all the categories within the feature can be represented. For each sample, there is only one setting for each of the derived dummy variables that has a value of 1. For the remaining parts of the dummy variables, they are equal to 0. According to the thumb rule, for each feature among all dummy variables one dummy has to be discarded randomly.

The accuracy of classification algorithm is seriously compromised when built on imbalanced data (Yuan et al., 2022). Since, this study builds on maximizing prediction capability of classifiers, re-sampling was a big necessity before diving further into modeling details. Sampling strategies aim to resolve data imbalance issue by balancing class distribution in the dataset, either by eliminating some data from the majority class (under-sampling) or adding some artificially generated data to the minority class (over-sampling) (Elamrani et al., 2020). In this study, after straining different re-sampling strategies the method of random over-sampling (ROS) has been adopted instead of under-sampling because the latter often results in the loss of important information (Ma et al., 2022). So, soon after the crash data was reduced to 526 samples by removing outliers, irrelevant levels, and missing data, the revised data was sliced into training (80%) and testing (20%) sets, and ROS was applied on the training data (Non-Fatal: 376, Fatal: 150) making the ratio 1:1 (Non-Fatal: 376, Fatal: 376) where the total data then became increase of 752 in terms of total driver count. After ROS, the further distribution of each individual category in terms of the severity level can be observed as follows:

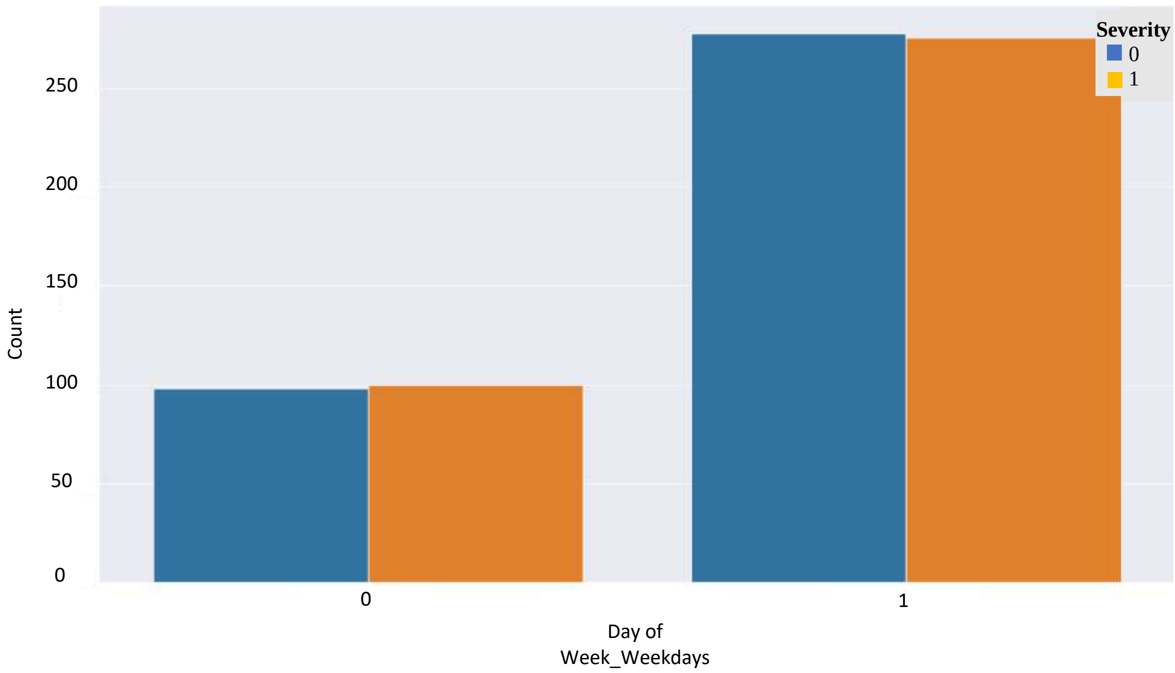


Figure 3.18: Distribution Graph of Day of Week

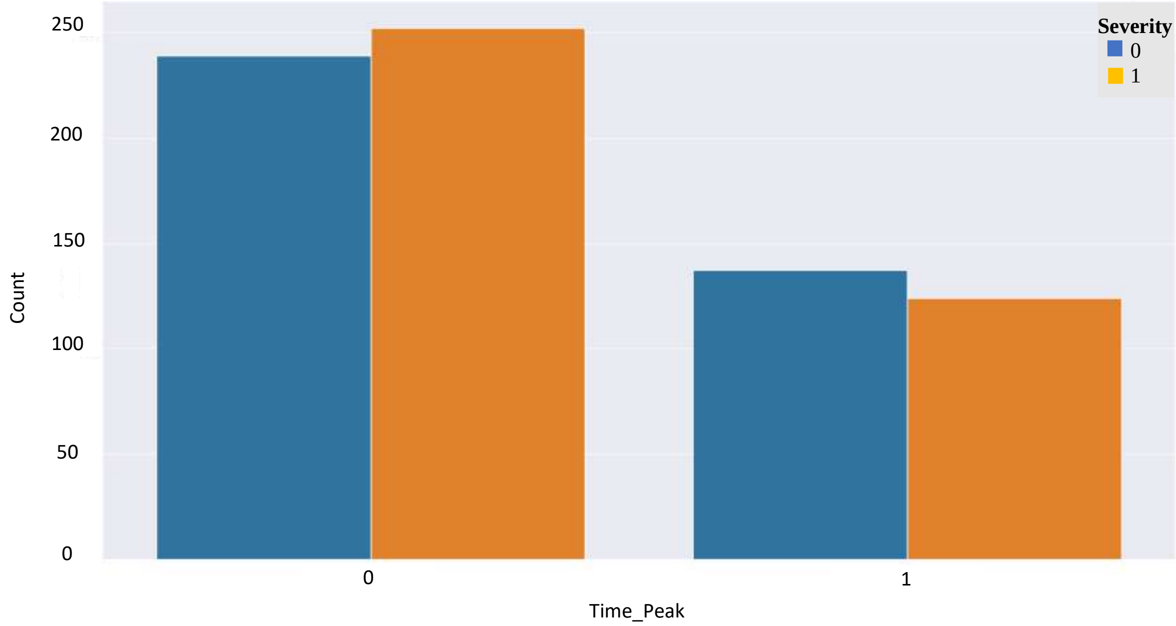


Figure 3.19: Distribution Graph of Peak-Time

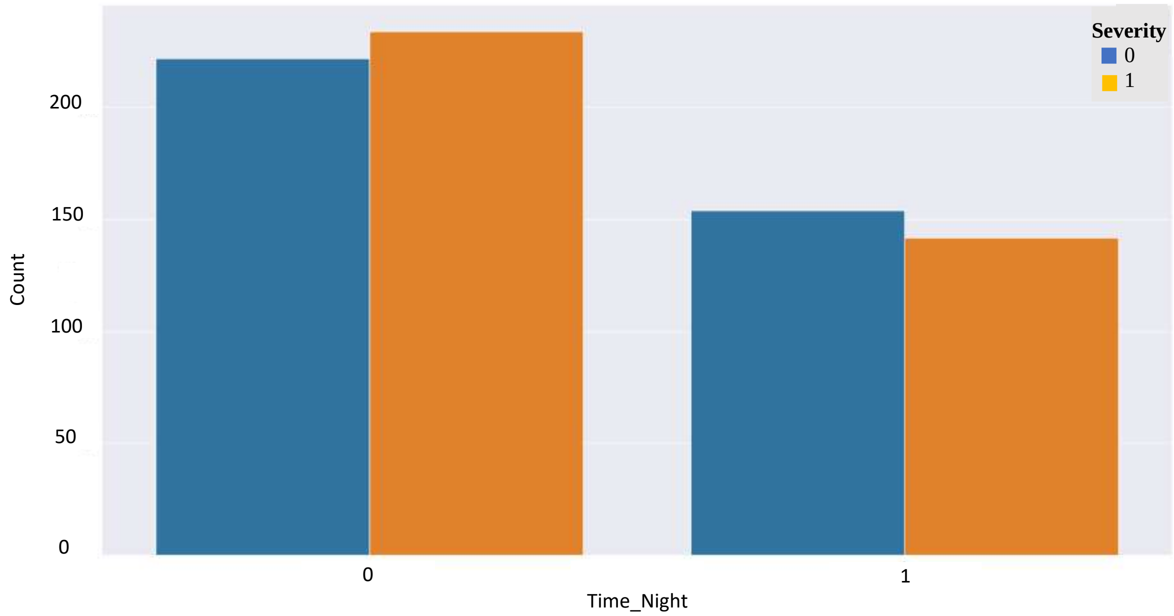


Figure 3.20: Distribution Graph of Night-Time

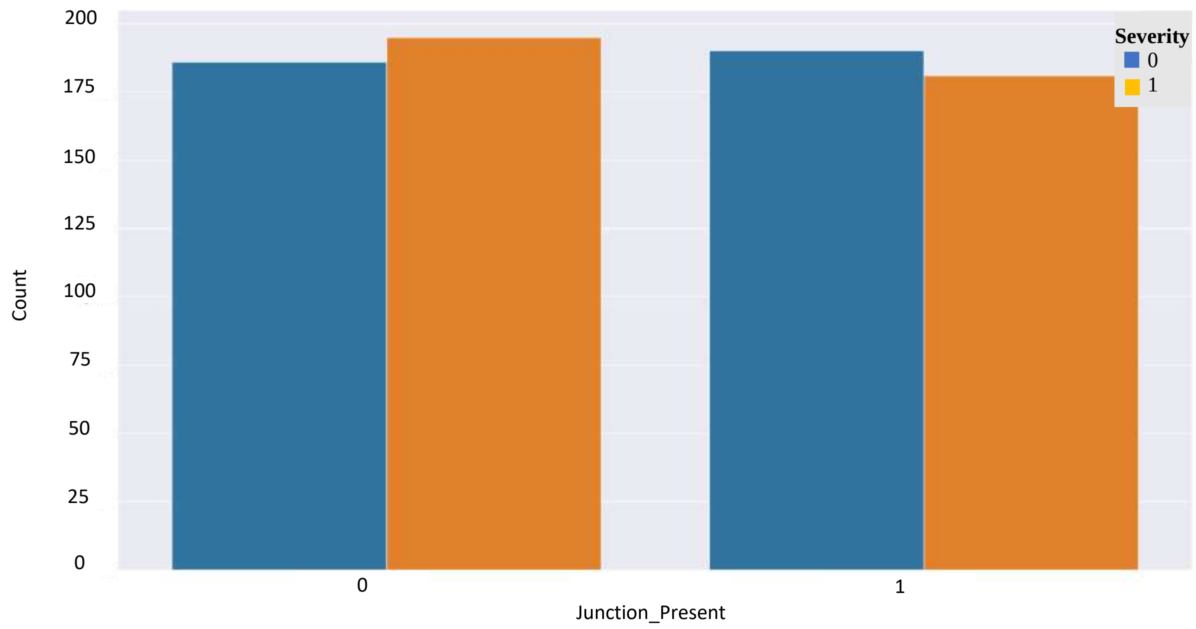


Figure 3.21: Distribution Graph of Junction

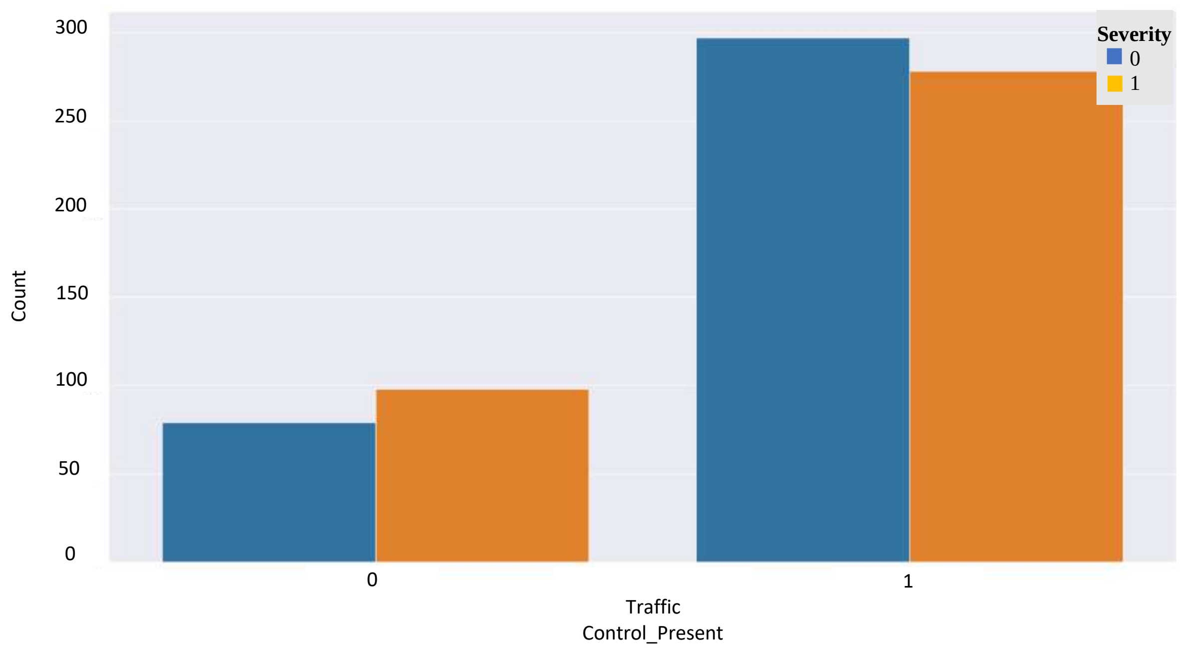


Figure 3.22: Distribution Graph of Traffic Control

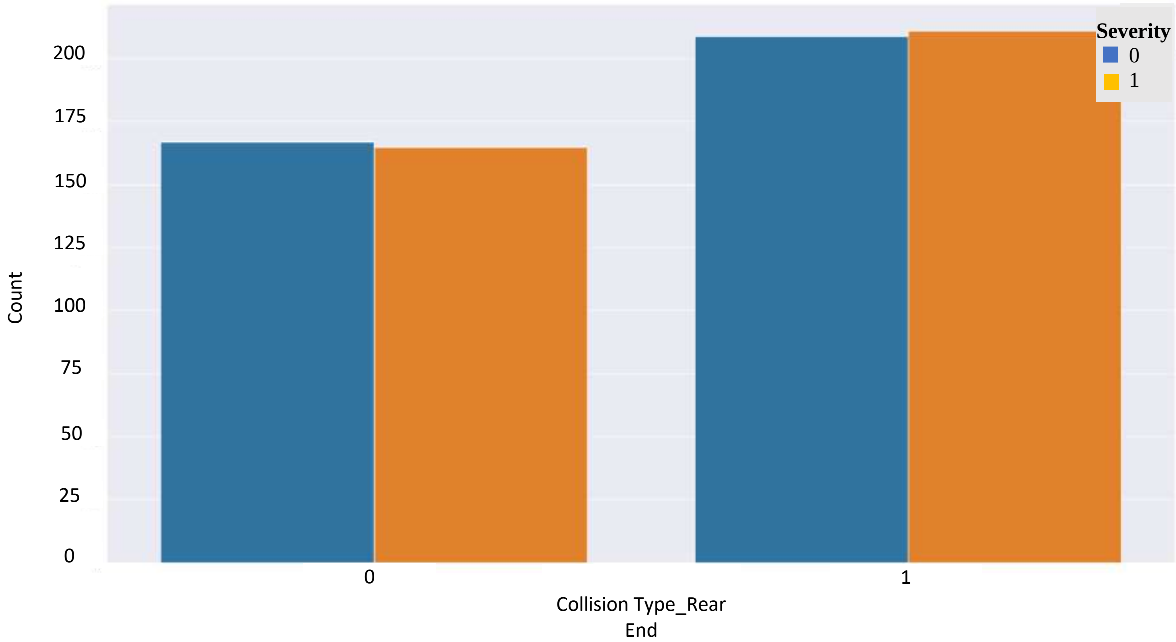


Figure 3.23: Distribution Graph of Collision Type

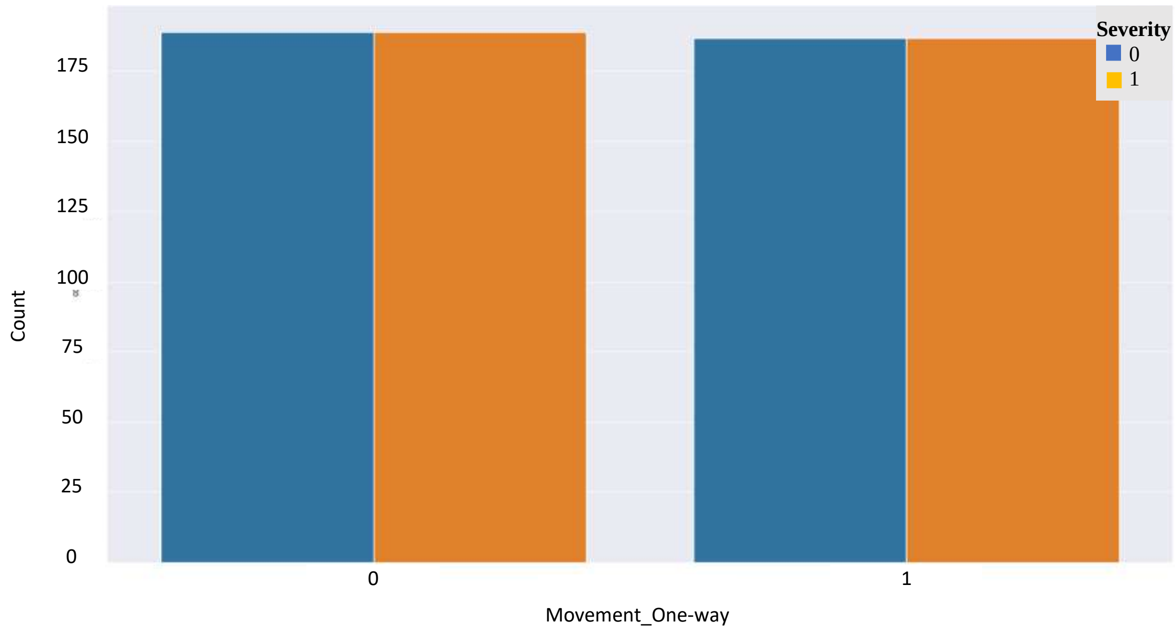


Figure 3.24: Distribution Graph of Movement

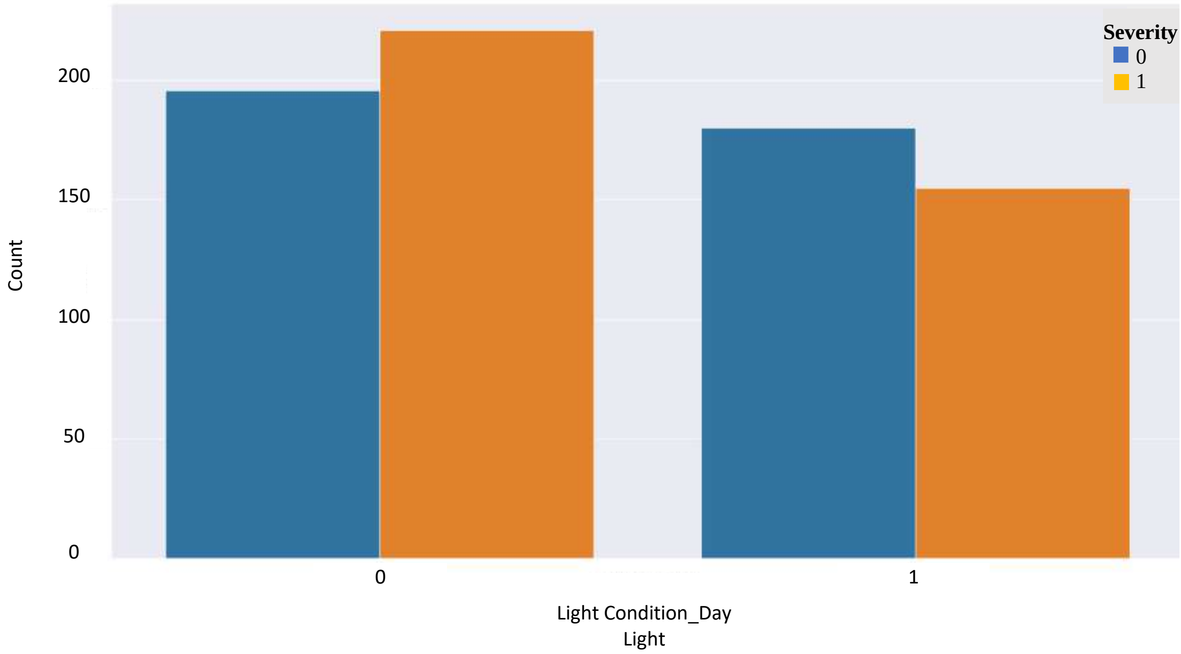


Figure 3.25: Distribution Graph of Light Condition (Daylight)

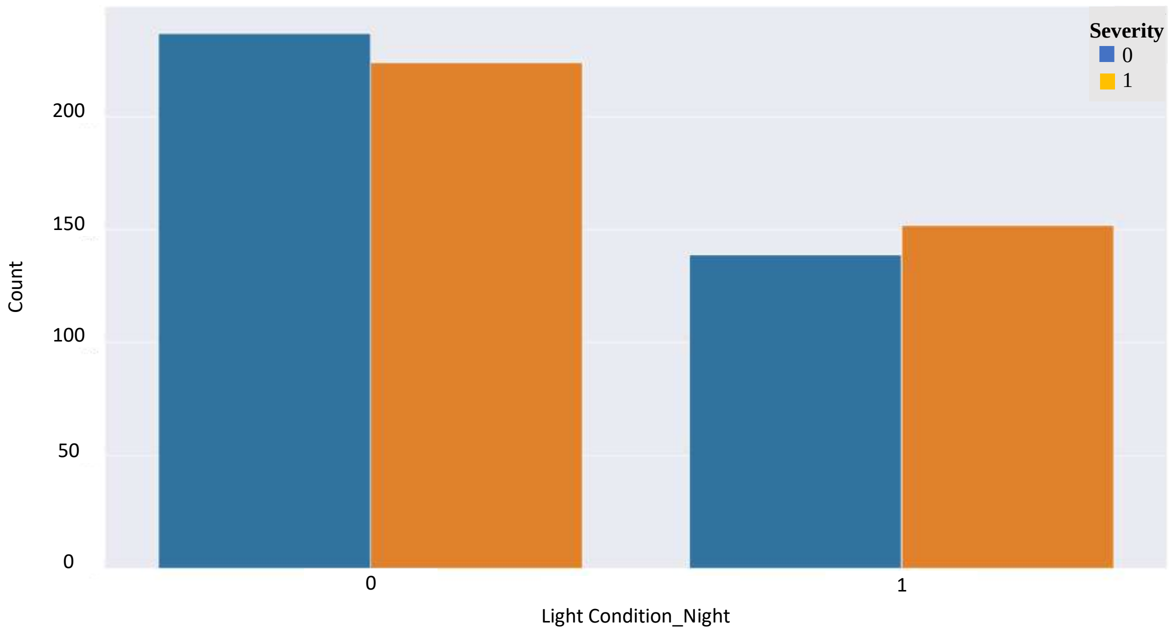


Figure 3.26: Distribution Graph of Light Condition (Night)

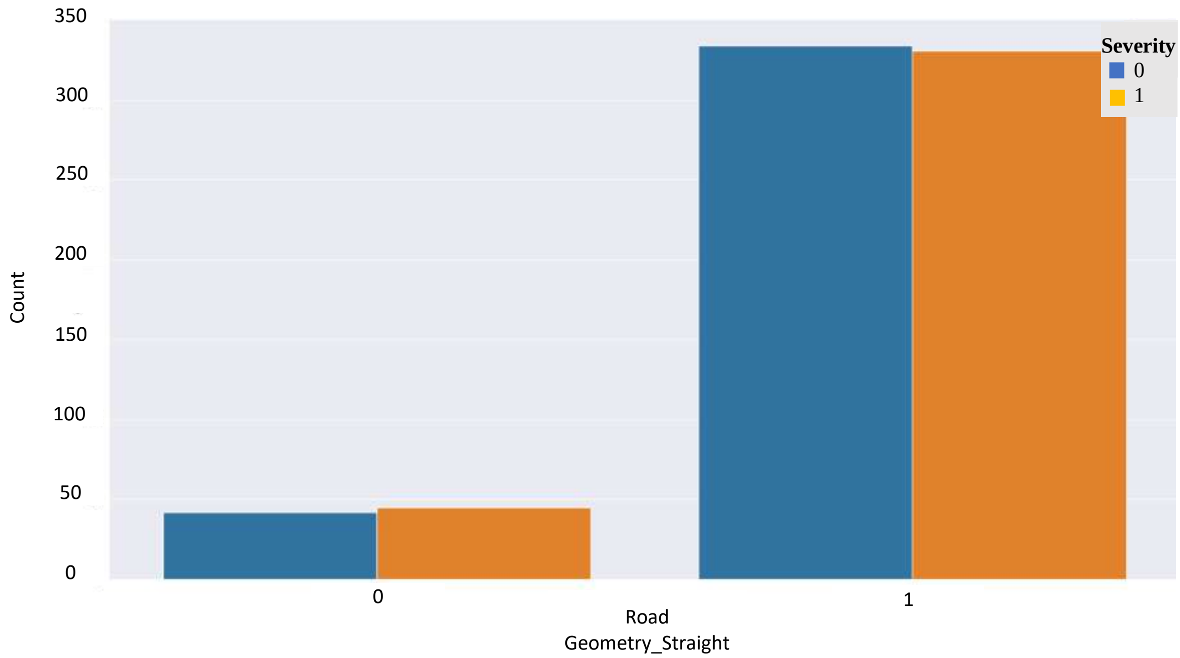


Figure 3.27: Distribution Graph of Road-Geometry

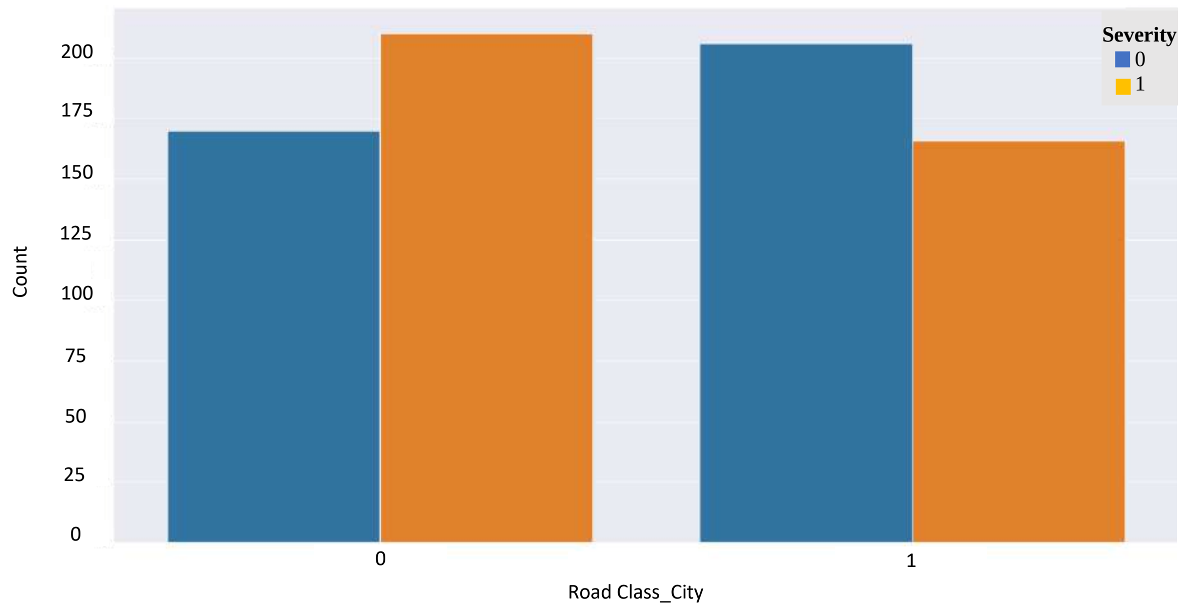


Figure 3.28: Distribution Graph of Road Class

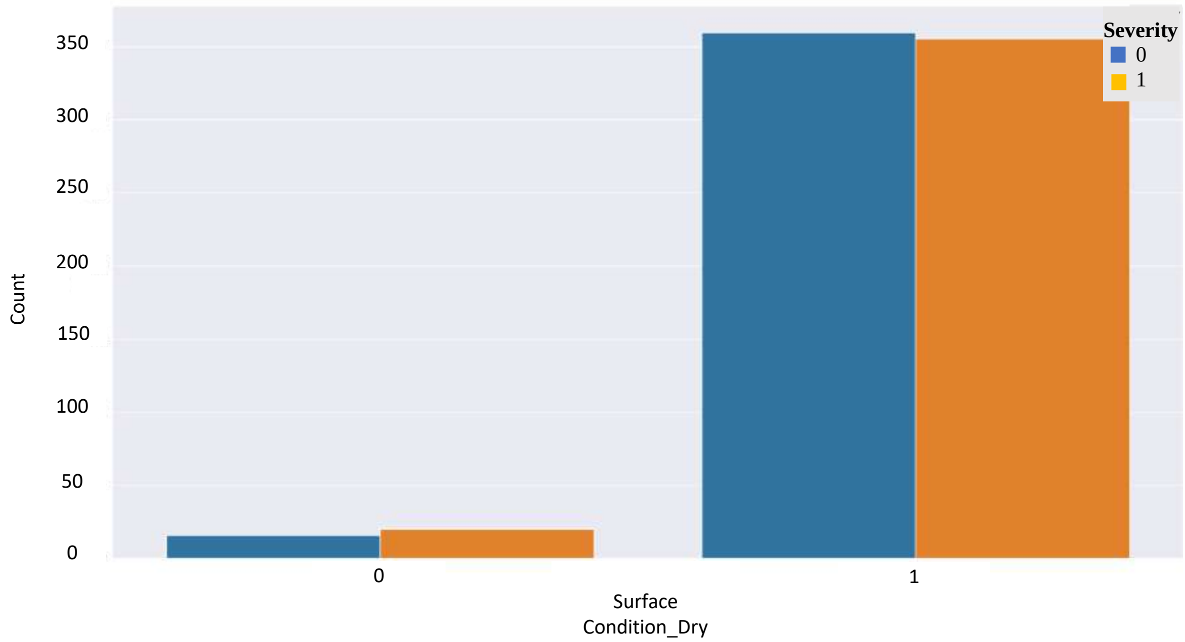


Figure 3.29: Distribution Graph of Surface Condition

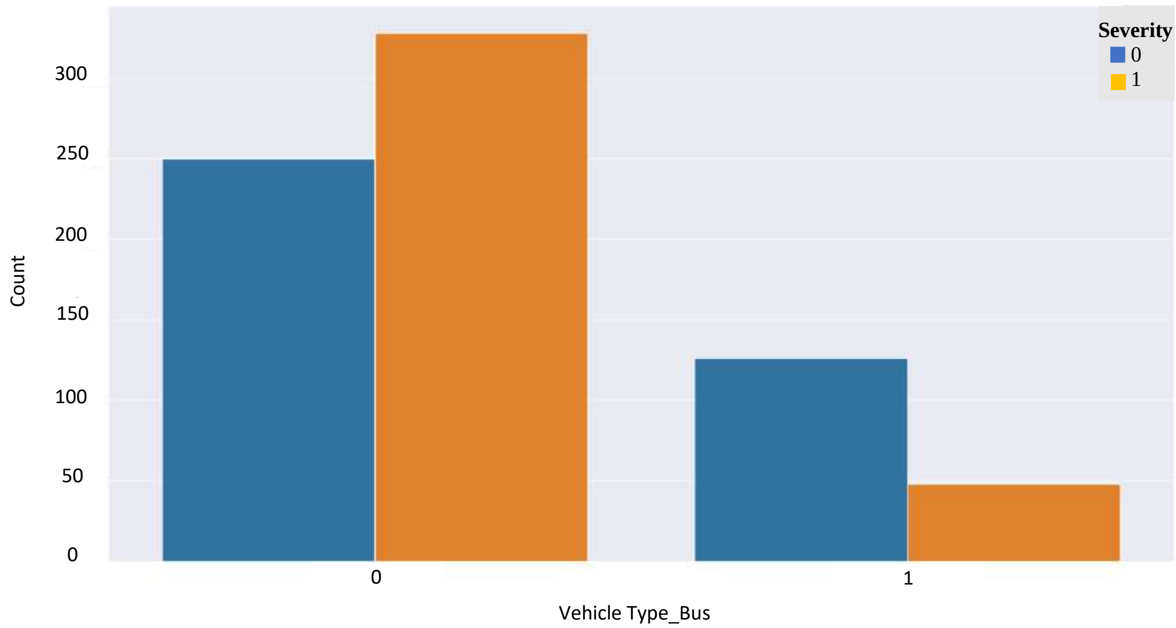


Figure 3.30: Distribution Graph of Vehicle Type (Bus)

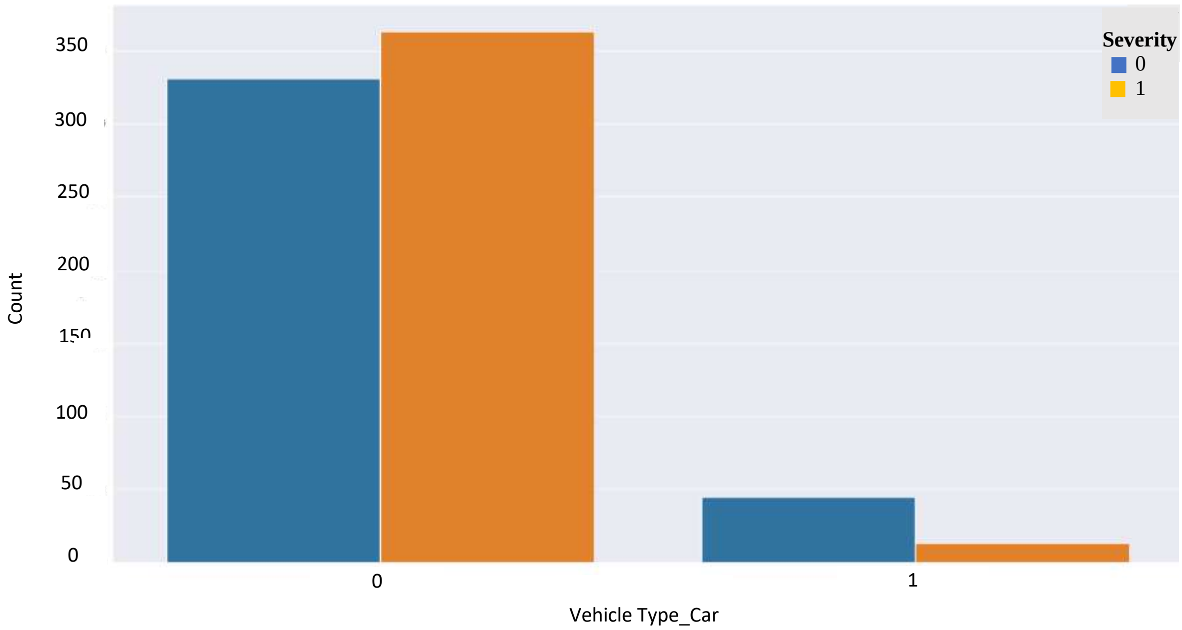


Figure 3.31: Distribution Graph of Vehicle Type (Car)

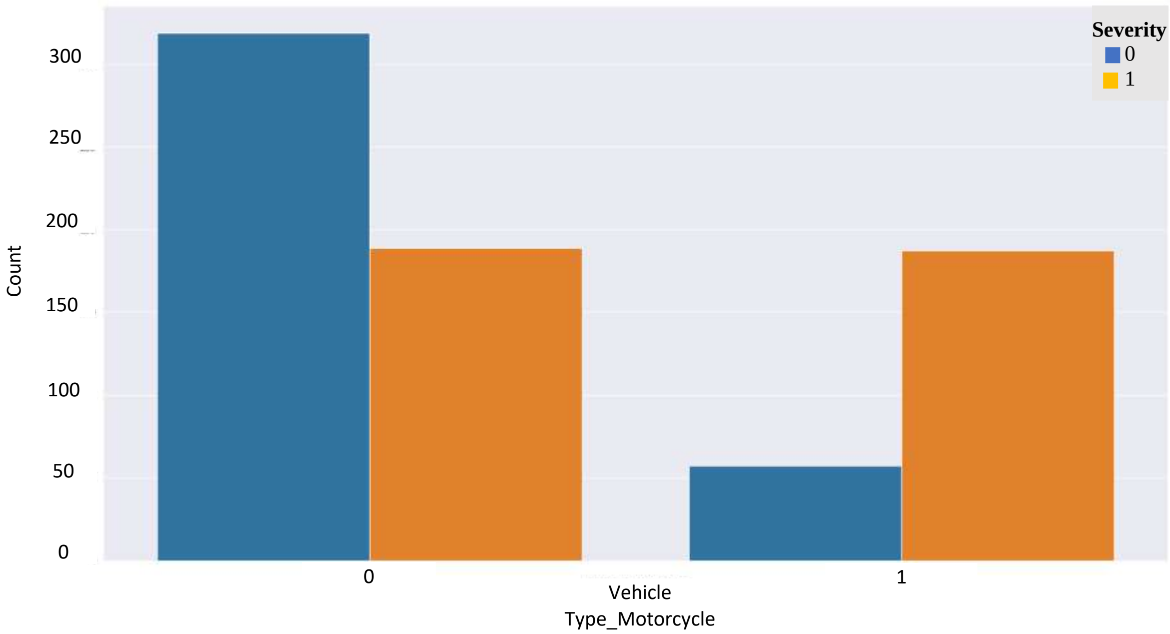


Figure 3.32: Distribution Graph of Vehicle Type (Motorcycle)

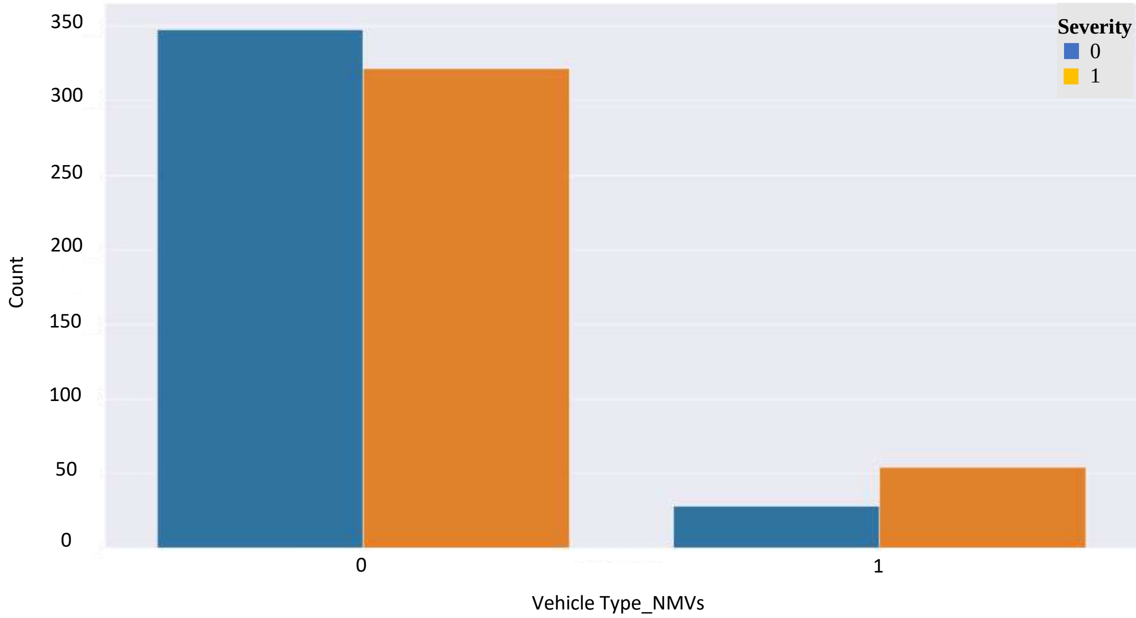


Figure 3.33: Distribution Graph of Vehicle Type (NMVs)

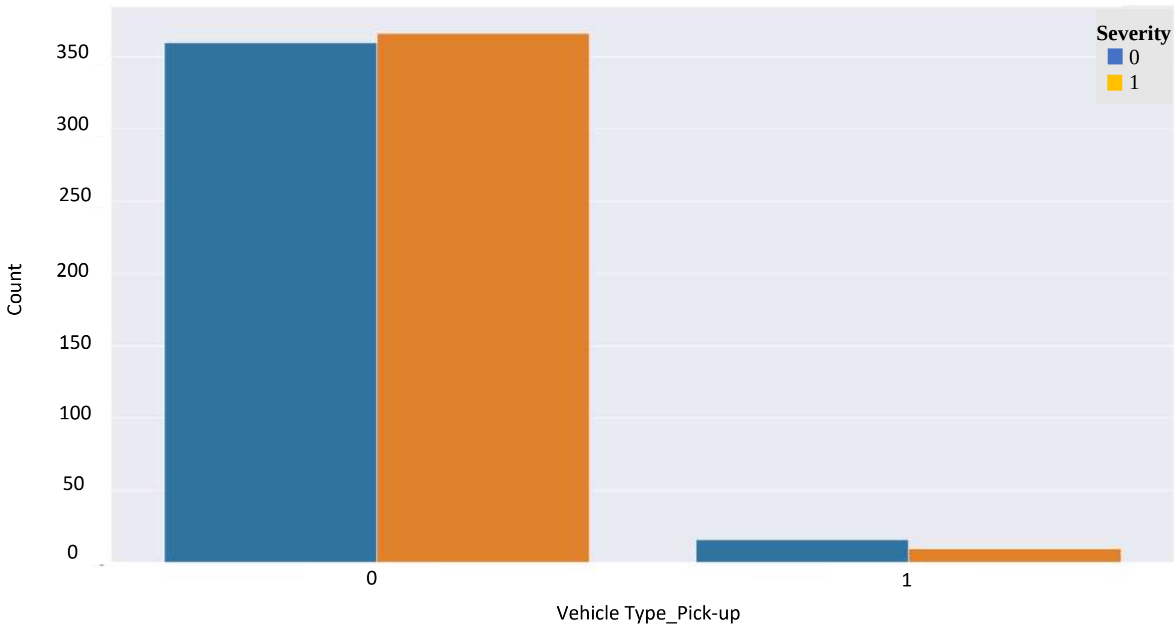


Figure 3.34: Distribution Graph of Vehicle Type (Pick-Up)

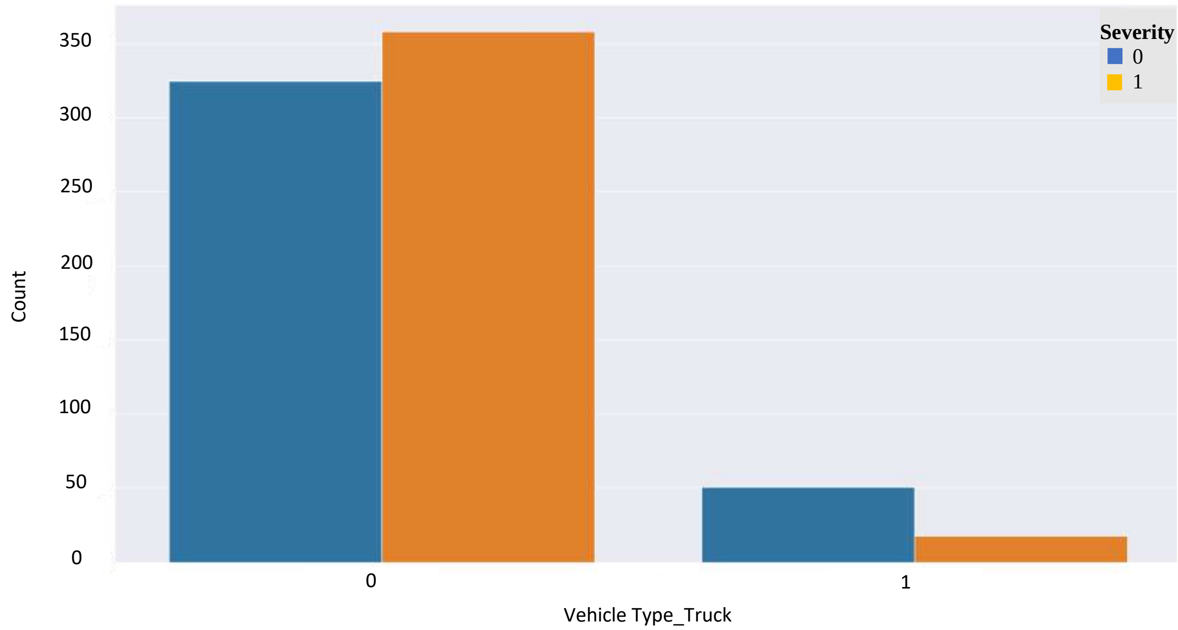


Figure 3.35: Distribution Graph of Vehicle Type (Truck)

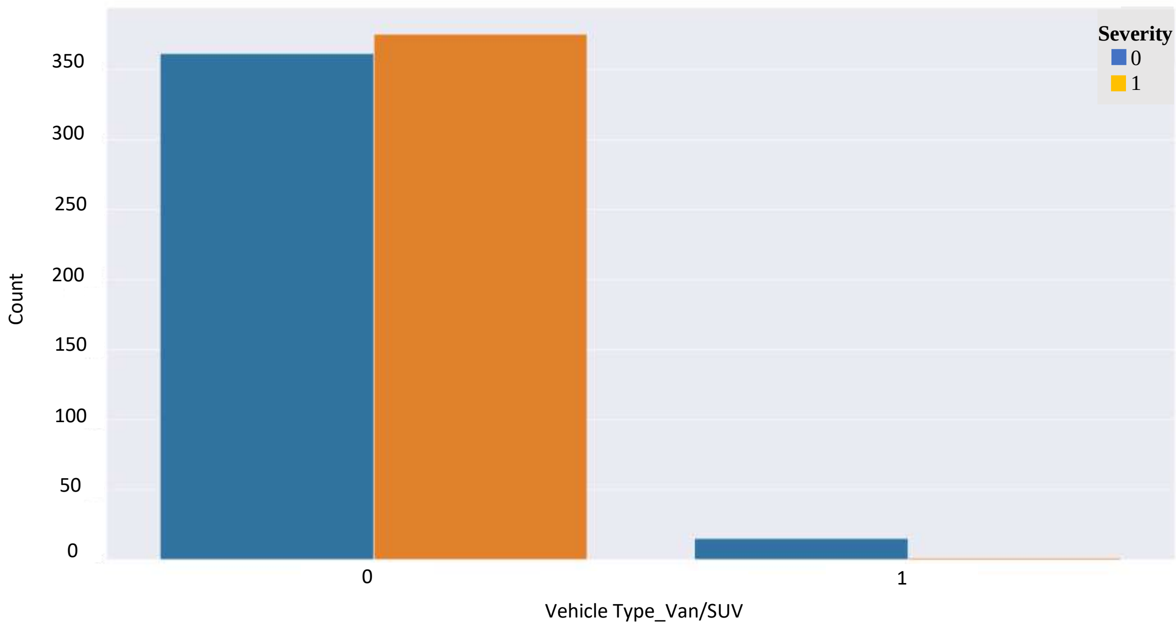


Figure 3.36: Distribution Graph of Vehicle Type (Van/SUV)

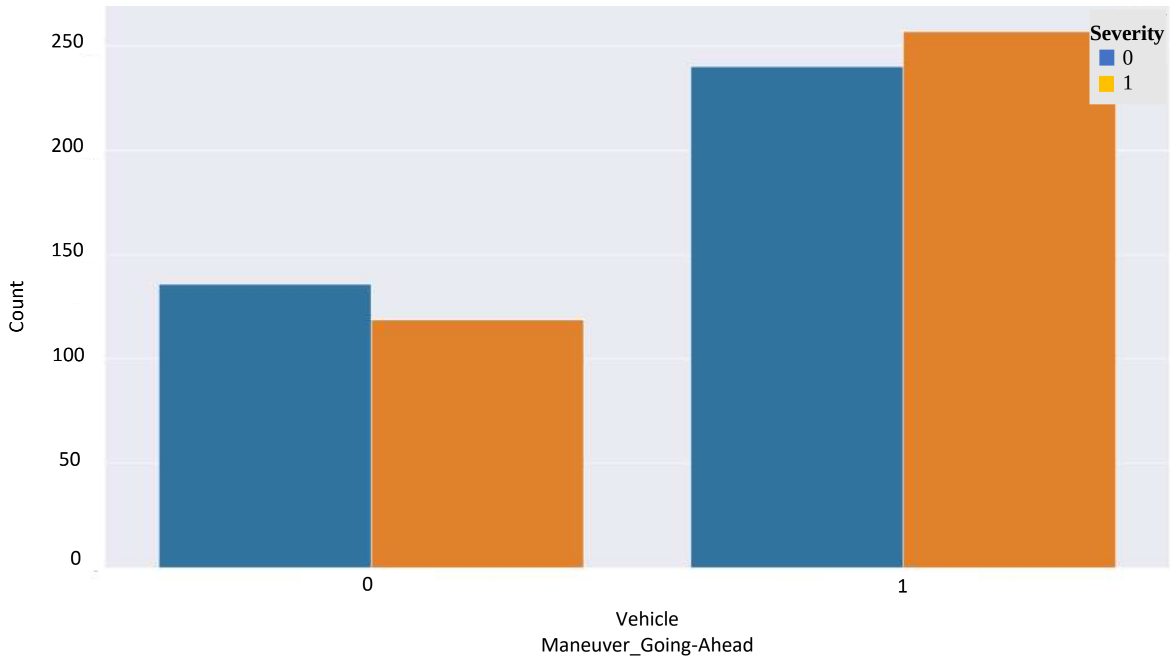


Figure 3.37: Distribution Graph of Vehicle Maneuver

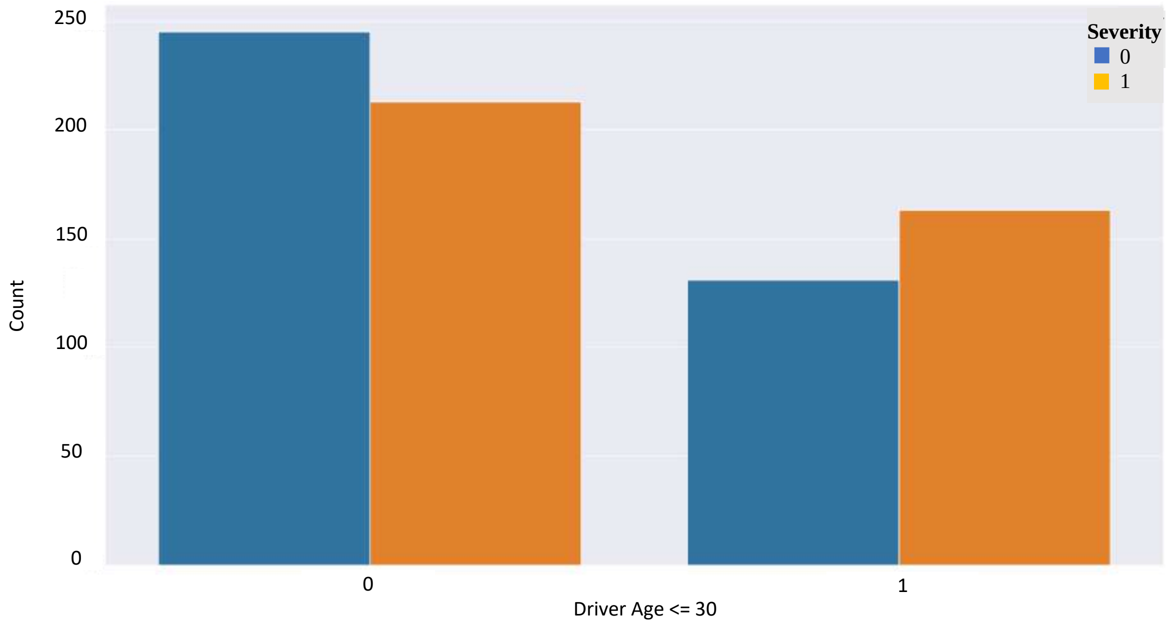


Figure 3.38: Distribution Graph of Driver Age <= 30

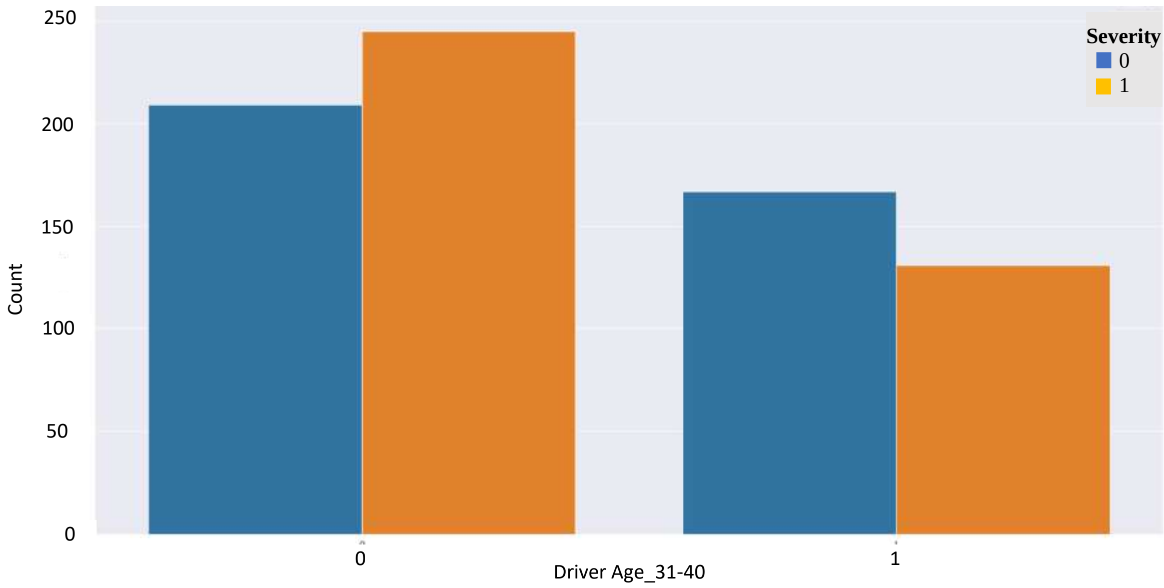


Figure 3.39: Distribution Graph of Driver Age (31-40)

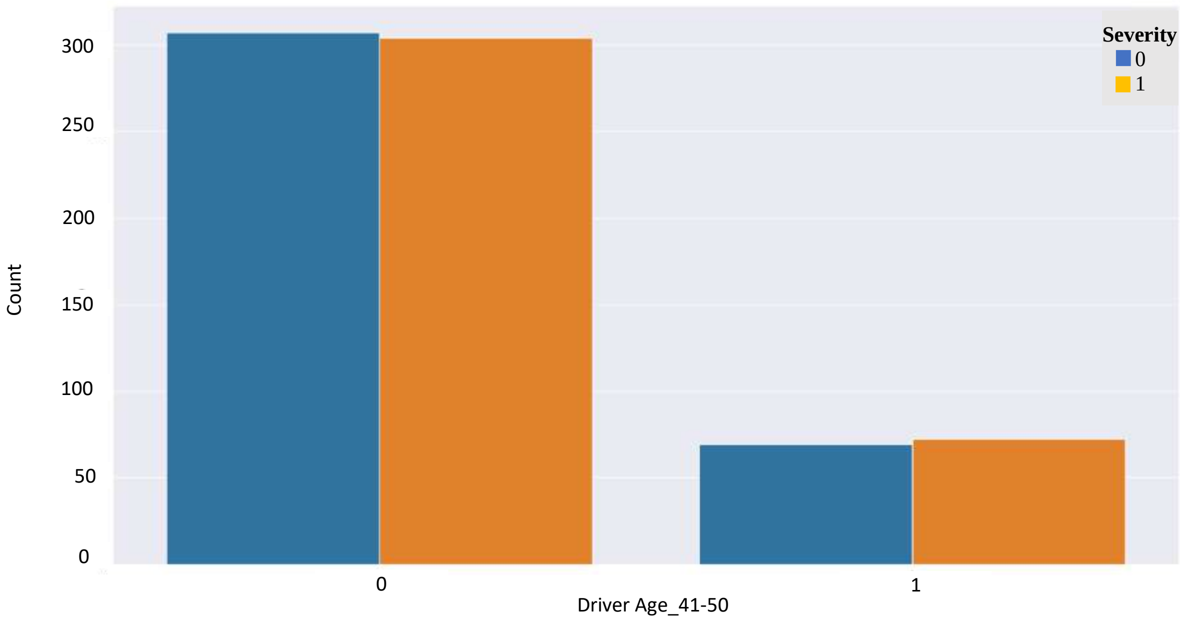


Figure 3.40: Distribution Graph of Driver Age (41-50)

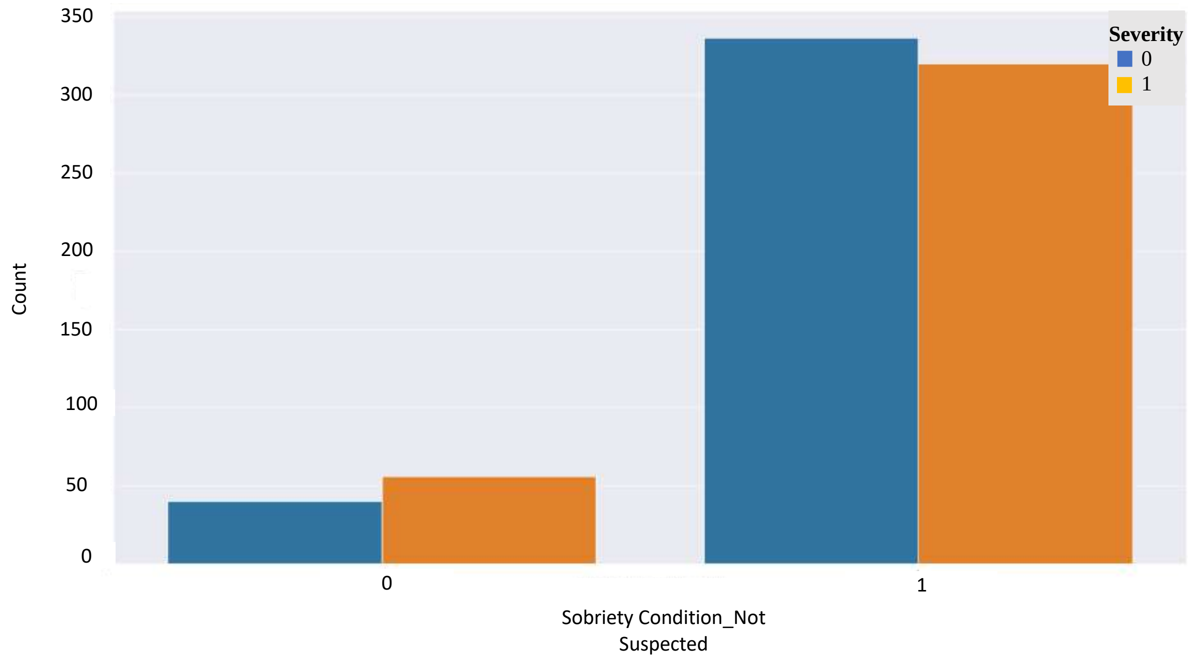


Figure 3.41: Distribution Graph of Sobriety Condition

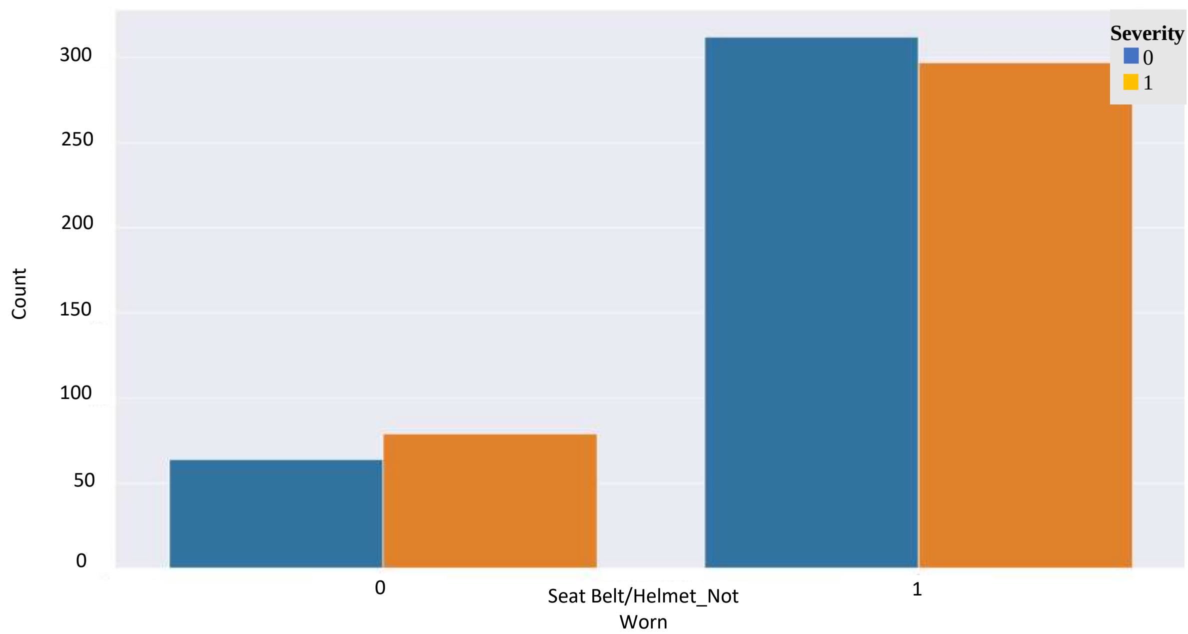


Figure 3.42: Distribution Graph of Seatbelt/Helmet Use

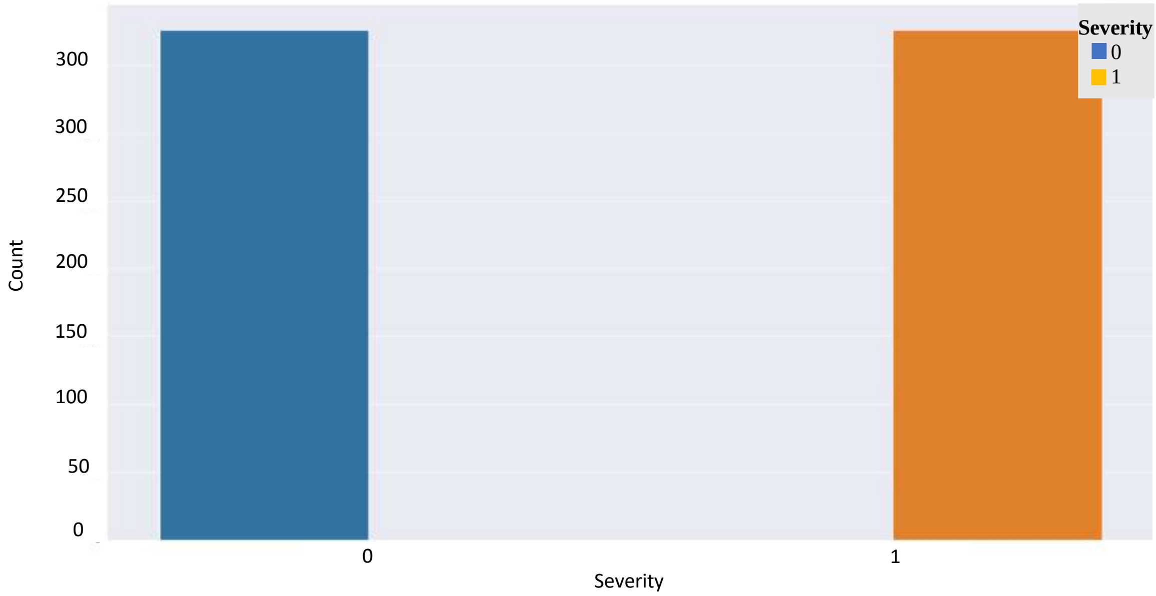


Figure 3.43: Distribution Graph of Severity

In the above Figure 3.18 to Figure 3.43, it can be seen that after creating dummy variables and after random over sampling (ROS), the ratio of fatal and non-fatal have balanced.

In this study, the dataset has been spitted into the 'Training' set (80%) and the 'Testing' set (20%) and trained the models through a n-fold cross-validation by the 'Training' set (here n = 10). That means, during the training process, the data have split into n subsets (n-1 for training and 1 for validation), and the hyper parameters have been tuned by repeating the procedure n times. The dedicated 'Testing' set have then tested the trained model used in this study.

3.4 Logistic Regression (LR)

LR is a regression analysis suitable to perform when the dependent variable is dichotomous (binary). It is a widely used tool for predictive analysis. The correlation between a binary dependent variable and one or more nominal, ordinal, interval, or proportional independent variables is explained using LR. In the LR model, independent variables are used to anticipate the probability that the response variable will obtain on a particular value (Abdelwahab, H. T., and M. A. Abdel-Aty, 2001). LR models are sometimes difficult to interpret; The Statistical Intelligence tool makes it easy to perform analysis and then interpret the results in simple language. Since it is possible for linear regression to produce probabilities larger than one or less than zero, it cannot be used for analysis of crash severity classification. That's why you should employ logistic regression instead (LR). Also, unlike the linear regression model, the assumptions made by logistic regression can be tested. Instead of producing discrete classes, LR generates probabilities on a scale from one to zero (Jamal et al., 2021)**.**

The LR model determines the relationship between the target class y = (y_1, y_n) given p = (p_1, p_n) and set of j predictors X = (x_1, ..., x_j). The strategy makes an effort to model the connection f between a set of independent variables x and a set of class variables y. The dependent/target variable was designed to have two possible outcomes: {y_1 = non fatal damage; y_2 = fatal}, which can be coded as {y_1 = 0; y_2 = 1}. The LR modeling function characterizes the connection between the set of independent or predictor variables and the probability of a specific class, such as y = 1. The equations (3.1 and 3.2) below illustrate a common form of the LR model:

$$P(y = 1|x) = \frac{1}{1+e^{(-z)}} = \frac{e^{(z)}}{1+e^{(z)}} \in [0,1] \qquad (3.1)$$

$$z = \beta_\circ + \beta_1 x_1 + \ldots \ldots \ldots + \beta_n x_n = x\beta \qquad (3.2)$$

Where xβ represents the sigmoid S-shaped function. A fatality, injury, or property loss was recorded in the data set when the probability was greater than 0.5. Many variables, such as the number of iterations, epsilon, learning rate strategy, step size, and regularization, were considered in logistic regression. In addition, both the regularization and the learning rate strategy were assumed to be constants.

This study used scikit-learn package of Python program to conduct the activities related to LR.

3.5 Classification and Regression Trees (CART)

CART is a type of classification algorithm established by Breiman et al. (2001) that builds a decision tree on the basis of Gini's impurity index. As mentioned by Wen et al. (2021), the categorical and continuous variable types are both acceptable for use as inputs and outputs in CART models. They also stated that, CART utilizes a repetitive binary splitting strategy. In this strategy, the training dataset is offered to the root node initially, and then it is split into two inner nodes. Then the splitting is repeatedly employed to each inner node. Certain requirements govern the split strategy ensure that an internal node's outputs are as uniform as possible. Here the procedure is continued until no further division is possible (Wen et al., 2021 and Chong et al., 2005). Finally, every leaf node in the tree indicates a distinct crash severity.

Let there be x_i in the learning datasets for i = 1 to M. After splitting, let t_p be the parent node and t_l, t_r be the left and right child nodes. The splitting rule in CART aims to split the data into two portions with the greatest possible homogeneity. The algorithm determines This study will demonstrate the Gini splitting rule for separating nodes and cross-validation for pruning trees, even though there are many other algorithms that can do the same thing. The splitting value xi^R in such a way that xi^R maximizes homogeneity of the child nodes for all splitting values of all variables. This is determined by developing an impurity function $I(t)$. The concept emphasizes that xi^R will maximize the difference in impurity between the parent and child nodes, as shown in Equation. 3.3 (Hossain, 2011):

$$\text{arg max } [\triangle I(t) = I(tp) - Pl{*}I(tl) - Pr{*}I(tr)] \tag{3.3}$$

where P_l and P_r represent the percentages of left and right node information. To determine the correct value of xi^R, several algorithms exist for defining the impurity functions that meet the conditions of the equation 3.3.

Anyway, it has been established that the algorithm has no bearing on the final tree. In this analysis, the Gini index is used to determine how to divide up nodes. The Gini index will be in the range (-1/K) to (1-1/K) if the outcome variable has K categories. The minimum value is seen for pure nodes (those that only contain data from one class), while the highest value is reached for nodes with an even distribution of outcome classes. At any given node t, the Gini index is defined as (Hossain, 2011):

$$I(t) = \sum_{j \neq 1} p(j|t)\, p(l|t) = \sum_j p(j|t)(1 - p(p(j|t)) = \sum_j p(j|t)^2 = 1 - \sum_j p(j|t)^2 \tag{3.4}$$

where j and l are the outcome variable categories and is $p(j|t)$ the proportion of outcome class j in node t. Now, by plugging Equation 3.4 into Equation 3.3, the change in impurity can be calculated. By minimizing [$Pl*I(tl) + Pr*I(tr)$], the change in impurity can be. The tree is grown to the maximum depth using this splitting algorithm by recursive splitting until each node contains a pure class. Following that, the tree is pruned based on a trade-off between the tree's complexity and the miscalculation error. It is accomplished by minimizing the cost-complexity (cp) function, a compound function, as shown in Equation 3.5.

$$\min R\alpha(T) = R(T) + \alpha(T') \tag{3.5}$$

where $R(T)$ denotes the misclassification error of tree T; T' denotes the total number of terminal nodes in tree T and $\alpha(T')$ denotes the complexity. The cross-validation method computes the value of by repeatedly using a portion of the data as a learning sample to build the tree and the remaining portion to test classification accuracy (Hossain, 2011).

There are many algorithms for determining the value of, but they all produce the same tree in the end. For your convenience, another approach is outlined below. For the sake of argument, let's say the complexity parameter starts at 0. Now, we need to calculate the value of a function defined as tree costs plus the complexity parameter increased by the tree size for every tree (including the first, which has only the root node). You can make the root node the largest tree by continuously increasing the complexity parameter till the the function's value for the largest tree surpasses the the function's value for a smaller sized tree to become the new largest tree. It will be obvious to those versed in numerical analysis that this algorithm makes use of a penalty function. Expenses, which tend to reduce with tree size, are combined with tree size, which also tends to increase linearly, to form the function. To a certain point, larger trees incur a greater penalty for their complexity as the complexity parameter is increased. However, there is a point at which the additional complexity of the largest tree no longer justifies the additional cost of the smaller tree (Hill et al., 2006). The largest tree in a sequence generated by this algorithm exhibits several interesting properties.

There is a nesting relationship between trees that have been successively pruned because each larger tree contains every one of the nodes of the following smaller tree. When trying to move from one tree to the following smaller tree in the sequence, many nodes are frequently pruned, but very few nodes are pruned as the root node is approached. Due to the absence of a smaller-cost alternative, the sequence of the biggest trees is optimally pruned. Evidence and/or explanations of these properties can be found in Breiman et al. (1984).

The completed tree serves as a useful visual representation of the problem space and can also be used to infer additional information. Each data point can be run down the tree according to the splitting criteria, and the class of the data will become the dominant class of the node at which it lands. The scikit-learn package of the Python programming language was used to perform the CART-related tasks in this study.

3.6 Support Vector Machine (SVM)

SVM was developed by Vladimir Vapnik with colleagues in 1992 (Boser et al., 1992). It is a ML approach entrenched on statistical learning advanced theory of C. Cortes and H. Drucker (Shafizadeh et al., 2017) and the structural risk minimization principle (Cortes et al., 1995). SVM was initially applied to the binary classification problem of linear discrete data (Liao et al., 2018). It is also an algorithm for predicting and classifying linear and non-linear data (Farhat et al., 2020). SVM is able to handle complicated nonlinear classification issues by mapping the original data through some kernel methods in higher dimensional space where the input is nonlinear but the output relation can be linearized (Wen et al., 2021). In the n dimensional space, SVM seeks for the beneficial solution $n-1$ dimensional hyperplane to divide changed data into various groups, where the distance between the hyperplane and the nearest data points is optimized (Wen et al., 2021). The hyperplane is known as the maximum margin hyperplane, and the linear classifier is also recognized as the maximum margin classifier. The goal of SVM is to find the maximum margin of the hyperplane. The training examples closest to the hyperplane with the largest margin are called support vectors as shown in Figure 3.43.

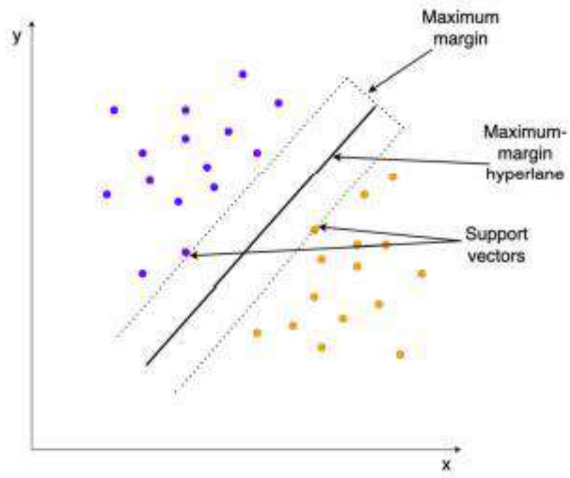


Figure 3.44: Maximum-margin Hyperplane Support Vector Machine

For this reason, SVM was first implemented for the binary classification problem of linear discrete data. Figure 3.44 depicts the basic idea, which is to locate an optimal hyper plane that satisfies the data classification demands and achieves the highest margin among two sample points while making sure classification accuracy. (Liao Y et al 2018).

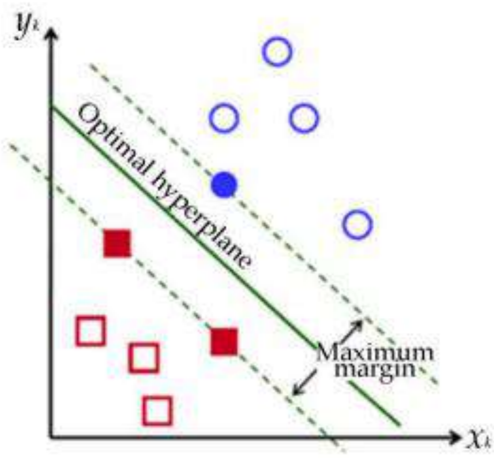


Figure 3.45: Concept of Optimal Hyperplane

The following is a brief mathematical description of the SVM algorithm (Equation 3.6). Assume a training set Q = $\{xi, yi\}_{i=1}^{N}$ with input vector xi = $\{xi^1, \dots, xi^n\}^T \in$ R and target labels yi ∈ (-1, +1), according to Vapnic Formula, satisfies the following conditions:

$$\begin{cases} W^T \emptyset(xi) + b \geq +1, if yi = +1 \\ W^T \emptyset(xi) + b \geq +1, if yi = -1 \end{cases} \tag{3.6}$$

Which is equivalent to:

$$yi\ [W^T \emptyset(xi) + b] \geq 1,\ i = 1 \tag{3.7}$$

Where the weight vector (maximum margin) and *b* is the bias (Equation 3.7).

In the case of linear classification, suppose the training sample is SV =

$\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, x ∈ Rd, yk ∈{-1,1}, $k = 1, 2, \ldots, m$, among which x_k is theinput variable, y_k represent the crash injury severity, *m* is the number of training samples, and Rd is a *d*-dimensional real number space.

SVM linear classification denotes the existence of a hyperplane ω.x + b = 0 that can correctly classify instances. Classify all samples, where ω is a weight vector that can be adjusted and b is the bias. The hyperplane must satisfy.

$$y_k(\omega \cdot x_k + b) \geq 1, k = 1,2, \ldots \ldots \ldots, m \tag{3.8}$$

Calculate the classification interval as shown in Equation 3.9,

$$\min_{\{x_k|y_k=1\}} \frac{\omega \cdot x_k = b}{||\omega||} - \min_{\{x_k|y_k=-1\}} \frac{\omega \cdot x_k = b}{||\omega||} = \frac{2}{||\omega||} \tag{3.9}$$

When the classification interval is maximized, that is, when the ||ω|| is minimized, then the optimal hyperplane problem can be written as finding the minimum function that satisfies the constraint of Equation (3.8).

The study used scikit-learn package of Python program to model SVM.

3.7 Random Forest

Random Forest (RF) was constructed by Breiman (2001) based on bagging method is a popular group learning method that involves various decision tree (DT) models with various attributes and integrates their model results to improve predictive accuracy.

According to Wen et al., (2021), in RF, a set of DT classifiers is trained, and each classifier is formed using samples obtained through bagging. For each DT classifier, only one casually selected subset of independent variables is used to separate the nodes. Each trained DT classifier votes for the severity results based on the input and ultimate classification is decided by majority votes (Wen et al., 2021). Before starting the search for optimal features and split points, the RF method requires the completion of two procedures. To begin, a predetermined number from the set of training data is selected at random A random subset of the growing trees is then selected each time by the RF. Over fitting can be reduced in RF based on two procedures. The model performance result in RF is achieved by combining respective results of all learners (Cai et al., 2022).

The prime steps of the RF algorithm are (Hossain, 2011):

(i) Let's say we have data set L with M predictors and N records, and we want to use a random forest (RF) with a total of B CART trees. Here we'll refer to Lb as the b-th bootstrap sample produced by selecting randomly n samples and replacing them with samples from L. Out of bag data (OOB) refers to the information that was left over after drawing the b-th bootstrap sample.

(ii) For the b-th tree Tb, instead of growing a CART tree with M predictors, m predictors are chosen randomly from M predictor space (M > m) at each node, and the best separator among m is used to split the node at each level, producing two maximum pure nodes.

(iii) Trying to predict from new data: run the newly collected data through every individual (here B number of trees) tree and the new data's class is the class of the leaf in every tree where it ended up. The final class of the data is determined by collecting the presumptions of the B trees. In the case of classification trees, it is achieved through majority voting.

(iv) Estimating OOB error rate: At each bootstrap iteration, the $L\text{-}L_b$ datasets are used to calculate the misclassification rate r_b of tree T_b (this misclassification rate r_b is used for calculating the variable importance as well). This is achieved by reducing the $L\text{-}L_b$ dataset to T_b grown in steps (ii). The majority vote determines the class of every data point (can be weighted). This majority vote is only needed to calculate the OOB error rate (not for variable importance). In other words, at the end, the r_b of all B trees is accumulated to measure the OOB error rate.

(v) Variable importance: Variable importance is a concept that differs from typical statistical approaches in RF. In this case, it is determined by permuting the values of every variable (one at a time) and then determining the new error rate. As any error in calculating its value has a significant impact on RF classification performance, the permuted variable with the maximum error rate is considered as the most critical variable. As a result, the values of the j-th predictor of M predictors in $L\text{-}L_b$ are permuted, and the data set is used to measure the misclassification rate rj_b. $|r_b - rj_b|$ represents the variable importance V_j of the j-th variable in the b-th tree. The technique is repeated for B trees, with the final variable importance calculated by averaging the V_j for every variable (j = 1 to M).

The study employed scikit-learn package of Python program to model random forest.

3.8 Adaptive Boosting (AdaBoost)

The AdaBoost was first proposed by Freund and Schapire (1997). Weak learners' errors are taken into account in this iterative algorithm. The distribution of the sample set is modified in each iteration depending on whether the pattern is accurately classified or not. In addition, it is essential to weight and integrate weak learners because the basic concept of AdaBoost is to teach many weak learners to construct a strong learner (Liu et al., 2020). Unlike RF, AdaBoost performs successive predictor learning and revises the weights on each analysis based on the error. At first, all findings are consistently weighted. Then, throughout iterative training, the learner's poorly approximated findings will be given more weight. Thus, the algorithm can successively adjust and minimize the deviation (Cai et al., 2022).

AdaBoost fits a series of learners to slightly modified versions of the original data at each boosting iteration. Through a series of iterations, the weights of correctly classified samples are decreased while the weights of incorrectly classified samples are increased.

The hypothesis of AdaBoost is shown in Equation 3.10,

$$h_f(x) = \begin{cases} 1 & if \\ 0 & otherwise \end{cases} \sum_{t=1}^{T} (\log \frac{1}{\beta_1}) h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \log \frac{1}{\beta_1} \quad (3.10)$$

The hypothesis h_f combines the outputs of the T weak hypotheses using a weighted majority vote.

3.9 Hybrid Model

Incorporating various algorithms in an ensemble or hybrid model can frequently deliver improved predictive abilities (Pradhan, B., and M. Ibrahim Sameen). So, in this study, besides training the LR and machine learning models, voting classifier method was implemented to integrate the individual models to develop a hybrid model. The aim to see if the hybrid model can better predict two-vehicle crash severity in Dhaka.

There are four distinct voting methods, including majority voting, simple voting, weighted voting, and soft voting (Zhou, Z.-H, 2012). Majority voting is a voting mechanism in which the output class label receives more than fifty percent of classifier votes for a class label. Simple voting, also known as proportional voting, determines the winner by casting the most votes. In addition, weighted voting is well suited for addressing the unequal performance classifiers. It gives more power for stronger classifiers when voting. Weighted voting can outperform both individual classifiers and majority voting when given rational weight tasks. Soft voting is commonly used for individuals who generate class probability outputs (Liu et al., 2020 and Zhou, Z.-H, 2012). Because the chosen base classifiers, LR, CART, SVM, RF, and Adaboost, generate class probability to determine the final class label, classifier-specific weight based soft voting may be a better method for obtaining the multi-label classifier (Zhou, Z.-H, 2012).

Here, to simply introduce the Classifier-specific Soft Voting we define that the individual classifier h_k outputs a l- dimensional vector $(h_k\ 1(x),\ldots,h_k\ l(x))T$ for the instance x_j, where $h_k\ j(x_j) \in [0,1]$ can be regarded as an estimate of the posterior probability $P\ (c_j \backslash x_j)$.

Each classifier is given a unique weight, and the combined result for class c_j is calculated and expressed as Equation (3.11) as follows:

$$H^j\ (x_j) = \sum_{k=1}^{T} \omega_k\ h_k\, j(x_j) \qquad (3.11)$$

Where ω_k is the weight assigned to the classifier h_k

Then, the calculation of output class label expressed as Equation 3.12 as follows:

$$\hat{y}_j = \arg\ {}^{\max}_{\mathrm{c}}\ [H^j\ (c_0|x_j),\ H^j(c_1|x_j),\ H^j\ (c_2|x_j) \qquad (3.12)$$

In this study, numerous classifier-specific soft voting models (hybrid models) were developed with all the possible combinations of the individual models: of LR, CART, SVM, RF, and Adaboost. The best performing hybrid model with the combination of LR, RF and Adaptive Boosting has presented in this study.

3.10 Feature Selection

As a data preprocessing strategy, feature selection is been shown to be effective and efficient in preparing data (particularly high-dimensional data) for various machine-learning problems. Building simpler and more understandable models, bettering machine learning performance, and getting ready clean, understandable data are all goals of feature selection. The curse of dimensionality occurs when machine-learning algorithms are applied to high-dimensional data. It refers to the phenomenon in which data becomes sparser in high-dimensional space, which has a negative impact on algorithms designed for low-dimensional space (Hastie et al. 2005). Furthermore, with a high number of features, learning models tend to overfit, that can lead to performance degradation on unseen data. High-dimensional data can significantly increase memory storage needs and computational costs for data analytics. Dimensionality reduction is one of the most effective tools for addressing the aforementioned problems. It consists primarily of two parts: feature extraction and feature selection. Feature extraction maps the original, high-dimensional features to a new, low-dimensional feature space. Typically, the newly created feature space is a linear or nonlinear combination of the actual features. In contrast, feature selection directly selects a subset of pertinent features for model construction (Guyon and Elisseeff 2003; Liu and Motoda 2007). Both feature extraction and feature selection have the benefits of enhancing learning performance, growing computational efficiency, reducing memory requirements, and constructing more accurate generalization models.

Feature selection preserves the physical significance of the original features and improves the readability and interpretability of models. In many applications, such as text mining and genetic analysis, feature selection is therefore commonly preferred. Despite the fact that feature dimensionality is often not that high, feature extraction/selection still plays a crucial role in certain circumstances, like improving learning performance, preventing overfitting, and reducing computational costs. There are numerous unrelated, redundant, and noisy features in real-world data.

Eliminating these features through feature selection lessens storage and computational costs without causing significant information loss or learning performance degradation.

The technique for selecting features not only saves estimation expenses but also performs adequately (Li et al., 2020). So, all the models (LR, CART, SVM, RF, Adaboost, and all possible hybrid models) were trained with all possible combinations of features (i.e., set of first one feature, set of first two features, etc.). However, before creating feature groups, at first the features were ranked based on their importance. In this study, prior to using Random Forest (RF) as classifiers, we used RF to rank features to conduct feature selection.

3.11 Model Evaluation Metrics

There are considerable methods of performance evaluation for ML algorithms. In this study, the most commonly known performance metrics were utilized to test the efficiency of the various techniques. Confusion matrix can be used to evaluate classification method performance. Confusion matrix contains a comparison between the outcomes of the system's classification and the results that should have been achieved (Prasetyo, E., 2012).

Based on Table 2, this study compared the performance of LR and different ML classifiers using the following evaluation criteria: Accuracy (ACC), Receiver Operating Characteristics (ROC) Curve, and Area Under the Curve (AUC) Value. For classification problems, the confusion matrix is made up of four possible scenarios, i.e., true (TP) positive rate which indicates the positive data entered into the system is detected correctly by the system, true negative (TN) rate indicates negative data entered into the system is detected incorrectly by the system, false positive (FP) rate indicates the negative data entered into the system is detected correctly by the system, and false negatives (FN) rate indicates positive data entered into the system is detected incorrectly by the system, that are shown in Table 2.

Table 3-1: Confusion Marix for evaluating model's performance

	Predicted Fatal Injury	Predicted Non-Fatal Injury
Actual Fatal Injury	True Positive (TP)	False Negative (FN)
Actual Non-Fatal Injury	False Positive (FP)	True Negative (TN)

Accuracy, precision, and recall can be calculated using the True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) values. Accuracy values describe the precision with which a system can classify data. In other words, the accuracy value is a comparison among correctly classified data and total data.

3.11.1 Accuracy (ACC)

Accuracy values describe the precision with which a system can classify data. In other words, the accuracy value is a comparison between correctly classified data and total data. The "error rate" is the proportion of misclassified samples in relation to total samples (Zhang et al., 2022). According to the same author, if there are m misclassified samples among the total samples n, the error rate is $E = \frac{m}{n}$. In a similar fashion, ACC can be expressed as Equation 3.13,

$$ACC = \frac{TP+TN}{TP+FN+TN+FP} \qquad (3.13)$$

3.11.2 Receiver Operating Characteristic (ROC) Curve and Area under the Curve (AUC)

The ROC curve is utilized to assess the performance of a classifier by plotting Sensitivity versus Specificity (Zhang et al., 2022). For binary classification problems, the AUC (Equation 3.16) is used. It identifies the two-dimensional region under the entire receiver operating characteristic curve (Zhang et al., 2022).

Specificity and sensitivity (also referred to as Recall) are two metrics described in the following section. As defined in Equation 3.14, the proportion of correctly predicted negative samples within all predicted negative class samples is referred to as specificity. Sensitivity is described as the proportion of correctly predicted positive samples among all real positive class samples, as shown in Equation 3.15.

$$Specificity = \frac{TN}{TN + FP} \tag{3.14}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.15}$$

$$AUC = [x = Sensitivity, y = 1 - Specificity] \tag{3.16}$$

Many studies have used sensitivity analysis (Tang et al., 2019 and Chen et al., 2016) to uncover the connections in between independent and dependent variables. The values of sensitivity, specificity, and the area under the ROC possibility curve (AUC) can ascertain how effectively and identifiably the models predict positive and negative classes (Ji, A., and D. Levinson, 2020).

3.12 Shapley Additive Explanations (SHAP)

Despite the fact that machine learning is designed to produce extremely precise estimates, it has proven to be challenging to assess the effect of explanatory variables on the output. This study examines the interpretability of tree-based ensemble models in order to better identify road safety solutions. The SHAP method developed by Lundberg and Lee (2017) is utilized to characterize the significance of the factors and to determine how these factors have an impact. To illustrate the prediction model's output, SHAP makes use of game theory. The Shapley value (Equation 3.17) can be determined using the following formula:

$$\varphi_t = \sum_{S \subseteq F-i} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{SU\{i\}}(x_{SU\{i\}}) - f_S(x_S)] \tag{3.17}$$

where $|F|$ is the total number of explanatory variables, S resembles any subset of explanatory variables that doesn't contain the i^{th} variable and $|S|$ is the size of that subset. $f_{SU\{i\}}(x_{S\cup\{i\}})$ indicates model trained with i, and $f_S(x_S)$ is model trained without i.

3.13 Summary

This chapter discussed the methodologies and the data collection processes that were adopted to achieve the objectives of this research. The chapter discussed the data collection and preparation procedures for this study. Different machine learning methods as well as a statistical method were discussed along with their applicability in this chapter. In this chapter, the most commonly known performance metrics were also discussed.

Chapter 4

MODEL DEVELOPMENT AND INTERPRETATION OF RESULTS

4.1 Application of RF for Feature Engineering

Feature Selection is a fundamental concept in machine learning which has a significant impact on the model's performance. Important for classification, feature selection eliminates irrelevant method to enhance model performance, make the model simpler to comprehend and decrease its running time. Feature Selection (variable elimination) facilitates data comprehension, reduces computation requirements, mitigates the curse of dimensionality, and enhances predictor performance. The objective of feature selection is to select a subset of variables from the input that can effectively characterize the input data while minimizing the effects of noise or irrelevant variables and still producing accurate predictions (Guyon and Elisseeff 2003).

In terms of feature selection, extensive experiments were used to propose a minimum redundancy–maximum relevance (MRMR) method for selecting the key features, which significantly improved class predictions (Ding and Peng, 2008). The random forest (RF henceforth), that is an ensemble learning algorithm based on decision trees, has just been widely utilized in a variety of fields and offers excellent predictive ability. Furthermore, the model is even more rigorous than other well-known models. Feature Selection based on the Random Forest (FSRF henceforth) can assess the significance of the features and select a subset of the most significant ones with good interpretability. The FSRF was utilized to extract the global, local, and evolutionary characteristics from protein data (Pan and Shen, 2009). To select the effective features, the RF model has been implemented. The current popular classification models LR, SVM, CART, AdaBoost, and RF were used to evaluate the efficacy of the feature selection of various feature combinations. Moreover, the performance of classifiers with various combinations of features has been compared using a variety of metrics.

To achieve specific goals or develop a model with excellent predictive performance, optimum features must be chosen from raw data. In this study, the RF model has been used to discover effective features since it can assess feature relevance and identify a group of relevant indicators with improved interpretability (Li et al., 2020). It is clear from Figure 4.1 that RF extracted various types of features from the raw feature set, including vehicle related factors (vehicle type, vehicle maneuver), driver characteristics (age, sobriety condition, seatbelt/helmet usage), roadway and environment conditions (road geometry, surface condition, junction, road class, traffic control, movement, light condition), crash characteristics (collision type), and temporal features (time, day of week).

4.2 Feature Selection Using RF

After extracting useful features using RF, twenty-five features have been selected for better predictive performance of the different classifiers as shown in Figure 4.1. In case of vehicle characteristics vehicle type and vehicle maneuver have been used as independent variables.

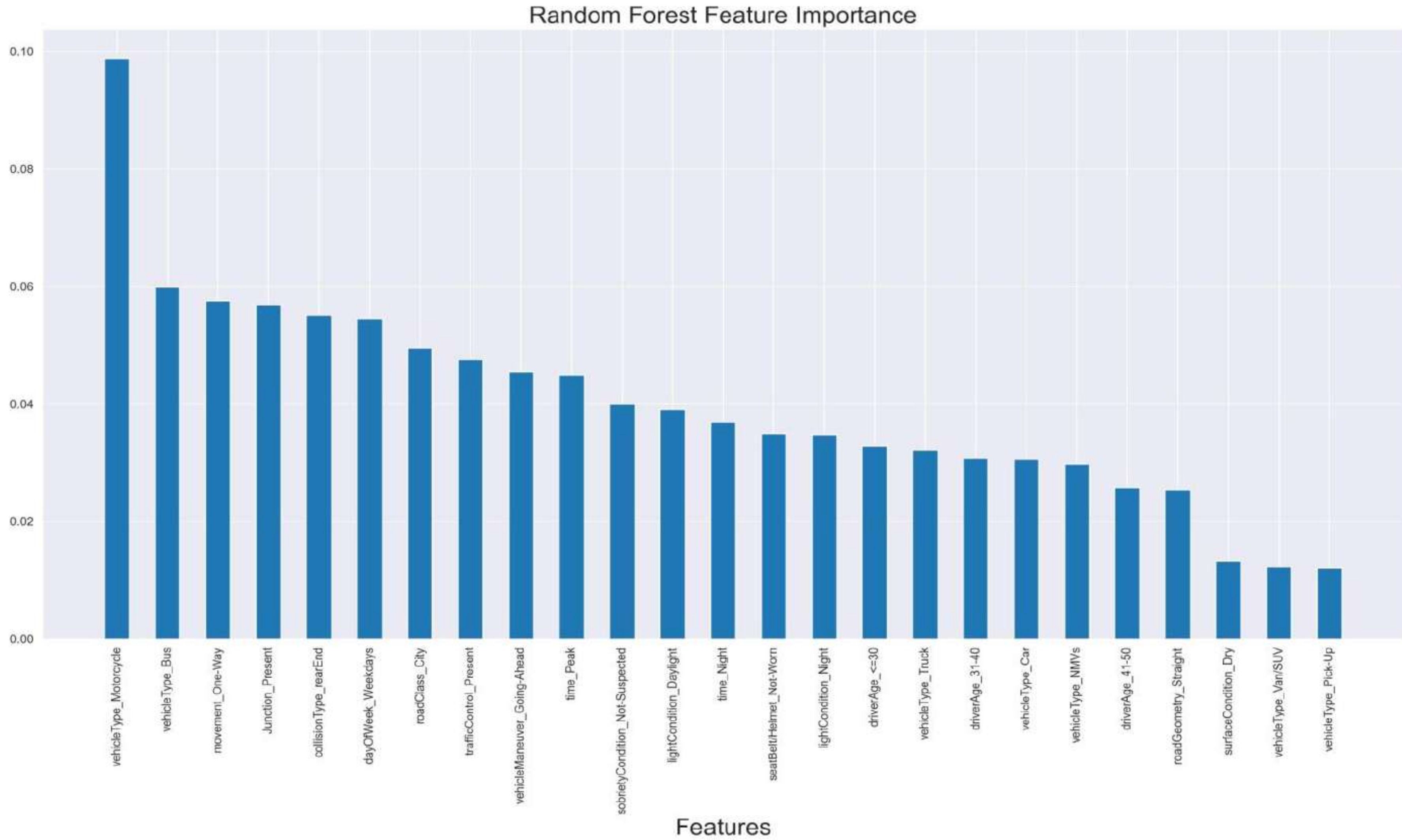


Figure 4.1: Useful Feature Selection Using RF

Vehicle type such as bus, car, motorcycle, NMV, pick-up, truck, van/SUV and any other types have been counted in raw data set. All the vehicle types have been selected in twenty five extracted feature set where motorcycle, bus, truck, car, NMVs, van/SUV and pick-up have been considered most important features respectively among them. There are two types of vehicle maneuver considered in the dataset one is “going straight” and another one is “not going straight”; where “going straight” maneuver has been selected as more important feature in case of two vehicle crashes. Considering driver characteristics driver age, sobriety condition and seatbelt/helmet use have been incorporated in data as important variables. Driver age have been categorized as driver age <= 30, driver age >50, driver age 31-40 and driver age 41-50; where driver age <= 30, driver age 31-40 and driver age 41-50 have been considered as important features for severity prediction of two vehicle crashes. There were two sobriety conditions depending on drunken suspecting whether “driver suspected drunk” or “driver not suspected drunk”. It has been found those drivers who were not suspected drunk had important contribution in predicting crash severities of two vehicle accidents. According to the feature seatbelt/helmet use, it has been identified that seatbelt/helmet not worn has been selected as an influential variable in severity prediction. In case of roadway characteristics road geometry (straight/not straight), surface condition (dry/not dry), junction (present/ not present), road class (within the city/outside of the city), traffic control (present/not present) and movement (one-way/two-way) have been used as independent variables. Among all these variables movement (one-way), junction (present), road class (city), traffic control (present), road geometry (straight), surface condition (dry) have been extracted as significant features for predicting two vehicle crash severities. In environmental characteristics, three light conditions (dawn-dusk/daylight/dark) that have been used as predicting variables; where daylight and dark/night condition have been identified as important features respectively. Crash characteristics (collision type), and temporal features (time, day of week) have been extracted from raw feature set. After feature selection it has been confirmed that rear end collision had significant contribution in two vehicle crashes. Besides them, temporal features peak-time and night-time along with weekdays have been contributed significant role in crash severity prediction.

4.3 Model Evaluation

Taking into account the importance levels of the twenty-five features shown in Figure 4.1, all possible combinations of the features were developed, of which five significant combinations such as beginning with the first five features, afterwards correspondingly first ten features, first fifteen features, first twenty features and at last ending with all twenty-five features have been demonstrated better prediction performance. To assess the performance of different predictors, the number of selected attributes was increased from the first five to the first twenty, and finally all twenty-five features were employed combined. Five distinct individual models (CART, AdaBoost, LR, RF, and SVM) have been developed, with these five feature combinations introduced for each model to assess the efficacy of the classifiers.

4.3.1 Accuracy and AUC Score of Different Models

This study compared the performance of LR and different ML classifiers using the evaluation criteria: Accuracy (ACC), Area under the Curve (AUC), Receiver Operating Characteristics (ROC) curve values. Receiver Operating Characteristics (ROC) curves have been plotted for LR, SVM, CART, RF, Adaboost and voting classifier in Figure 4.2, 4.3, 4.4, 4.5, 4.6 and 4.7 respectively for all five feature subset. In order to compare the all the classifiers, in this study, Receiver Operating Characteristics (ROC) curves of different classifies for all feature subset have been plotted as shown in Figure 4.8. It showed the AUC-ROC curves for six different methods, illustrating the trade-off between sensitivity and specificity for different classifiers. The accuracy and AUC scores of different models (LR, CART, Adaboost, RF, SVM and voting classifier) for each feature subset have been identified from the ROC curves shown in Table 4.1.

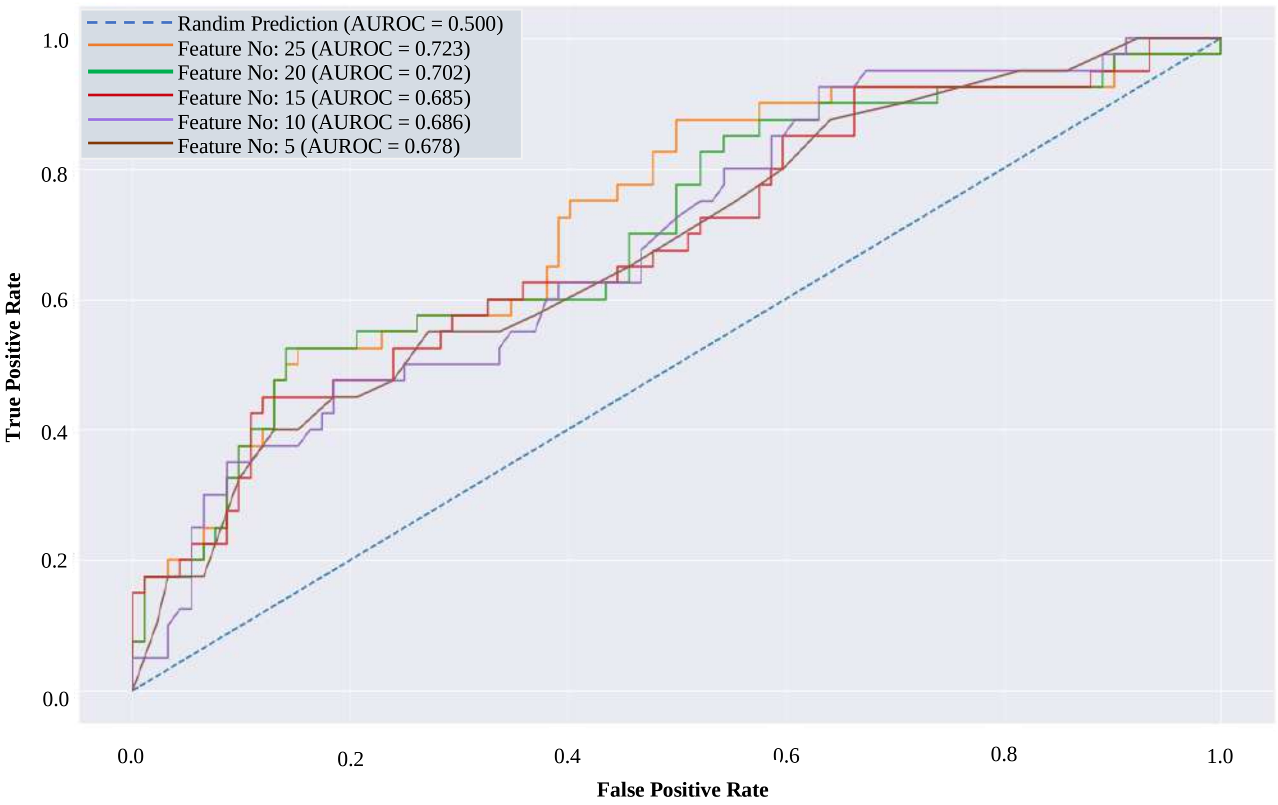


Figure 4.2: ROC Curve for Logistic Regression (LR)

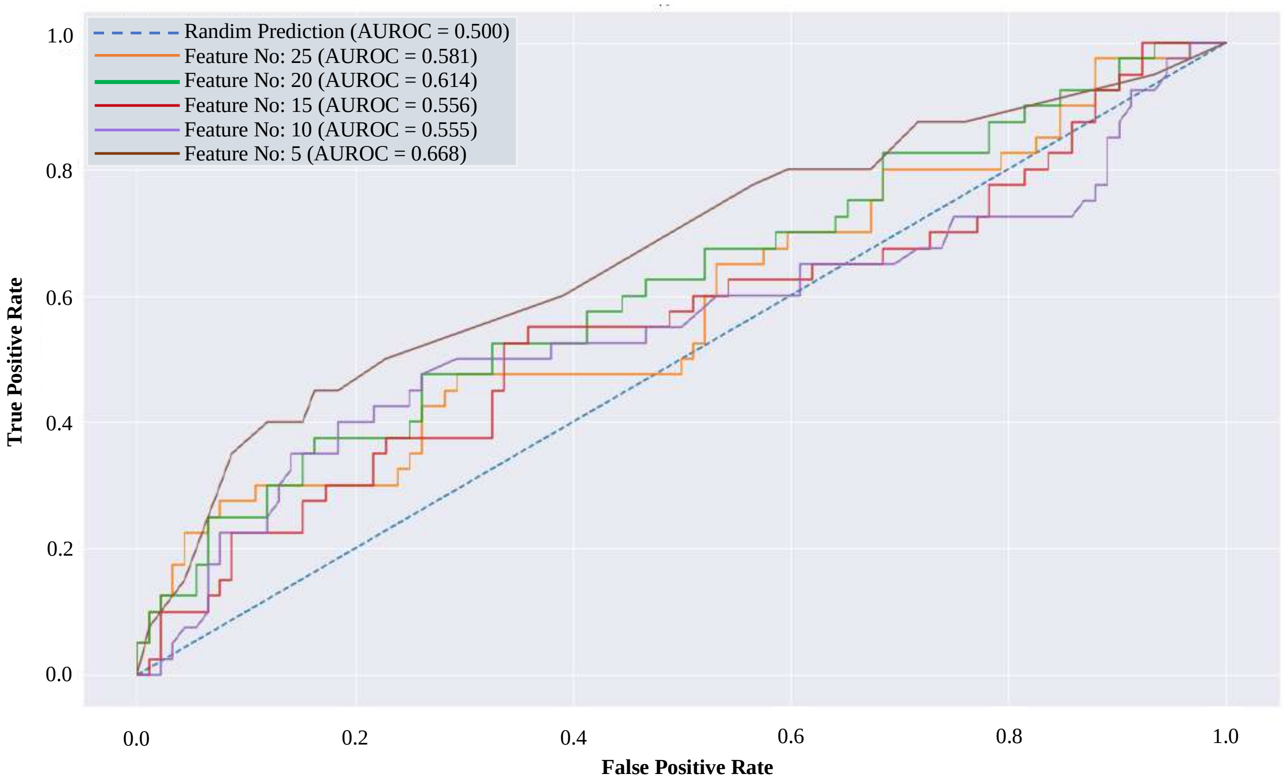


Figure 4. 3: ROC Curve for Support Vector Machine (SVM)

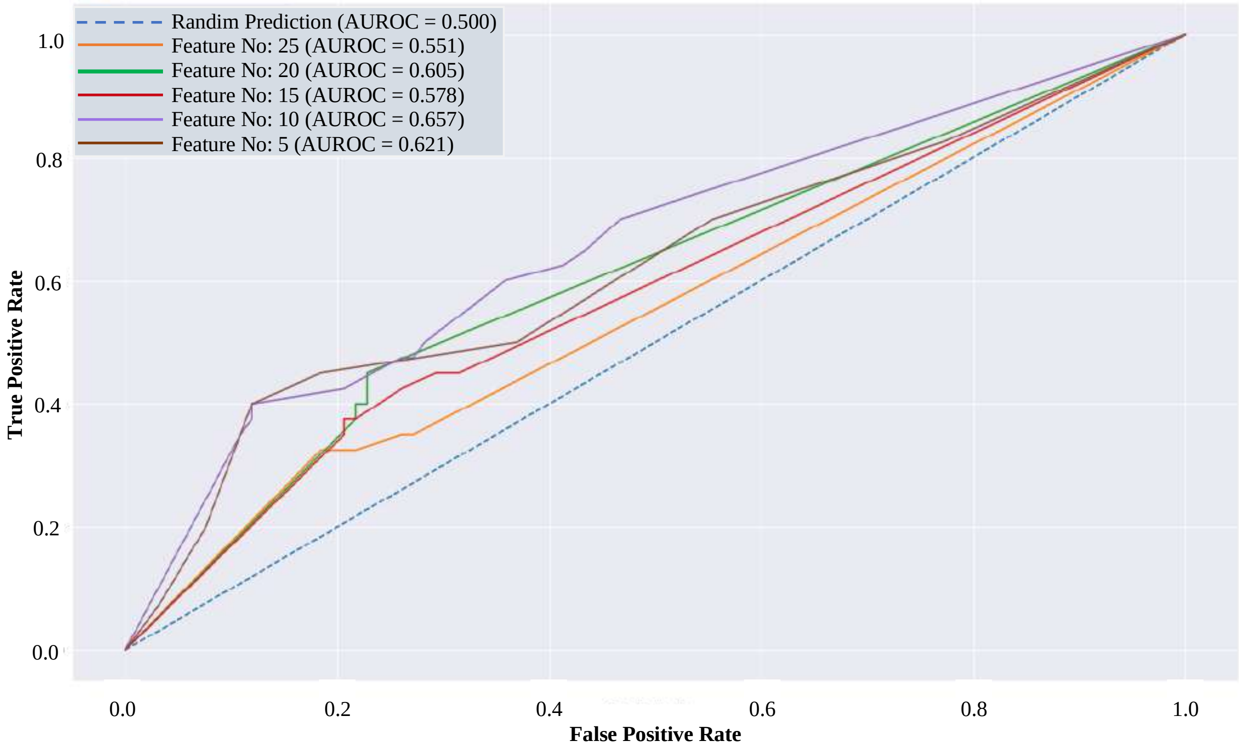


Figure 4.4: ROC Curve for Classification and Regression Tree (CART)

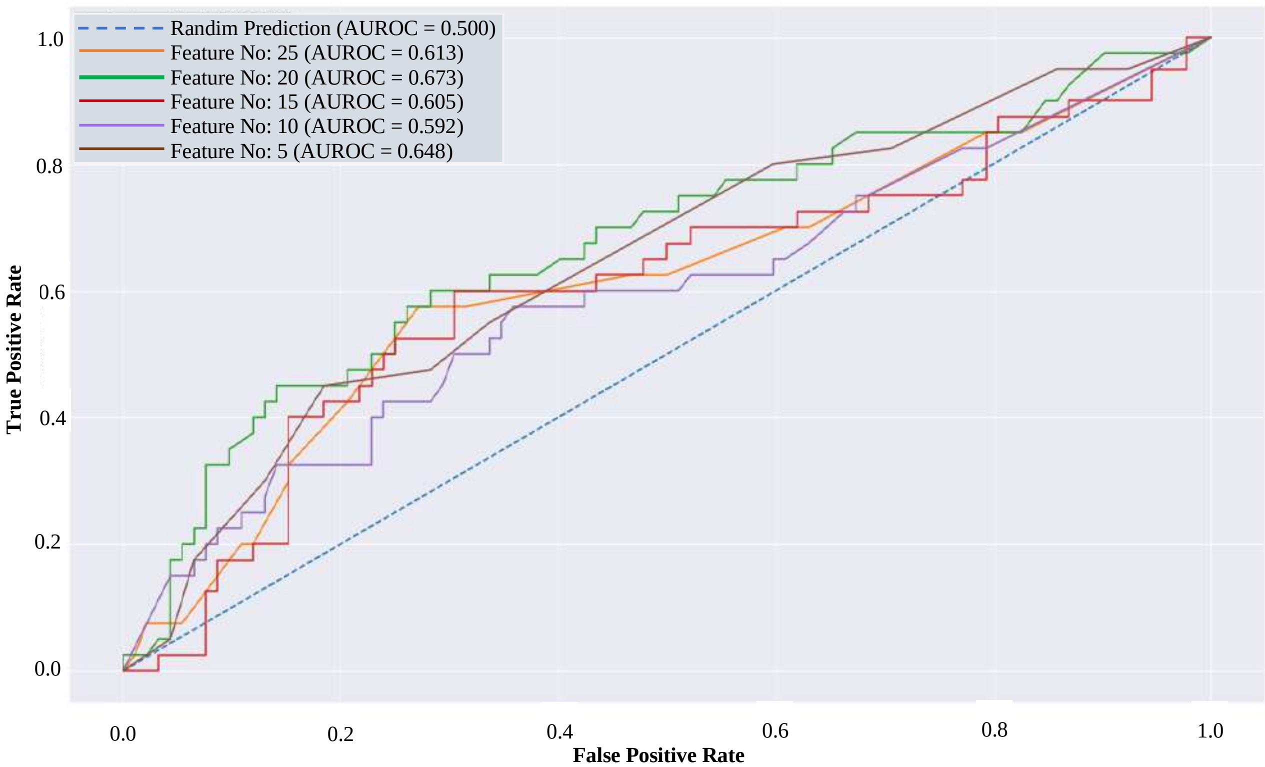


Figure 4.5: ROC Curve for Random Forest (RF)

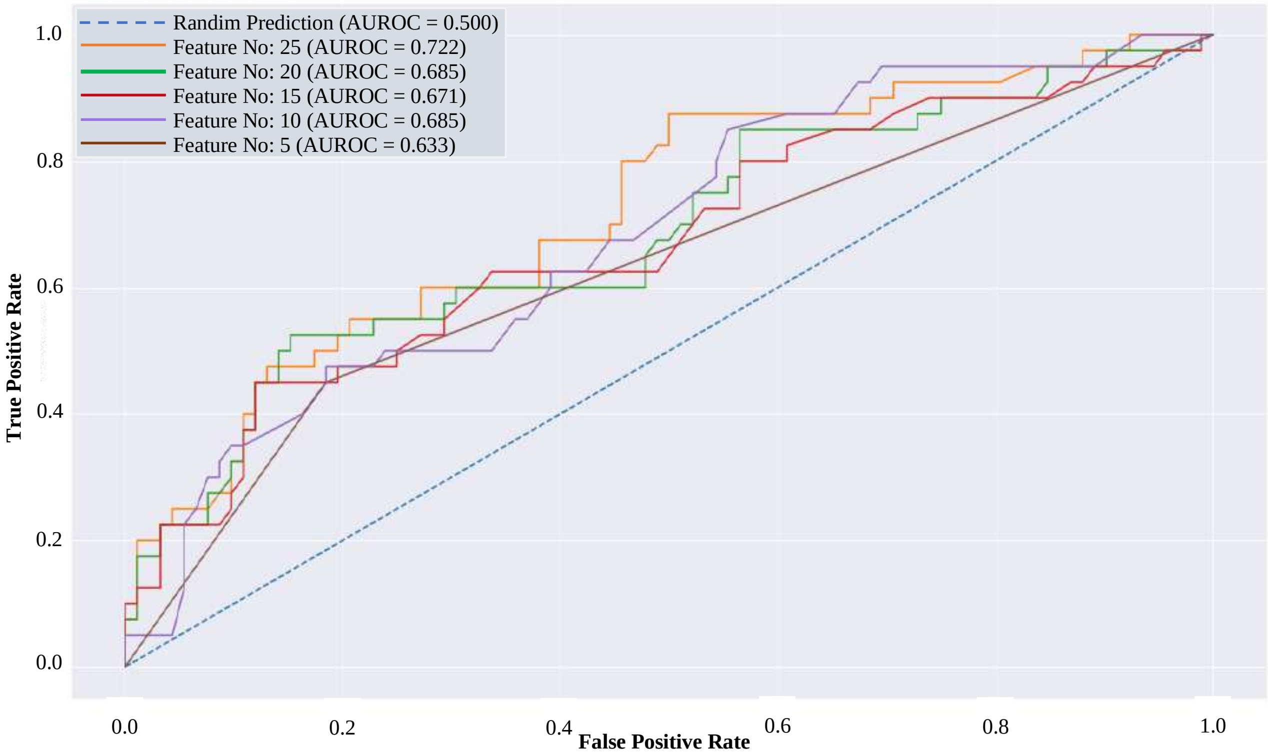


Figure 4.6: ROC Curve for Adaptive Boosting

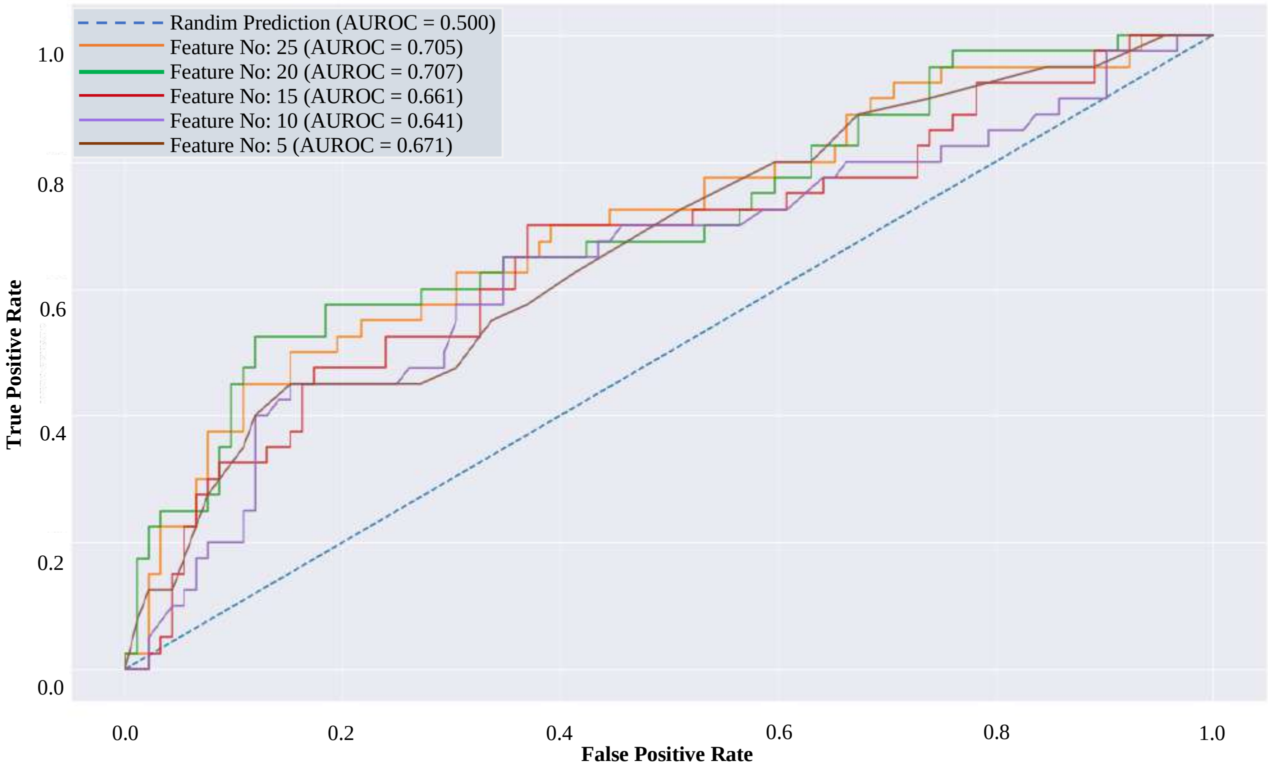


Figure 4.7: ROC Curve for Voting Classifier

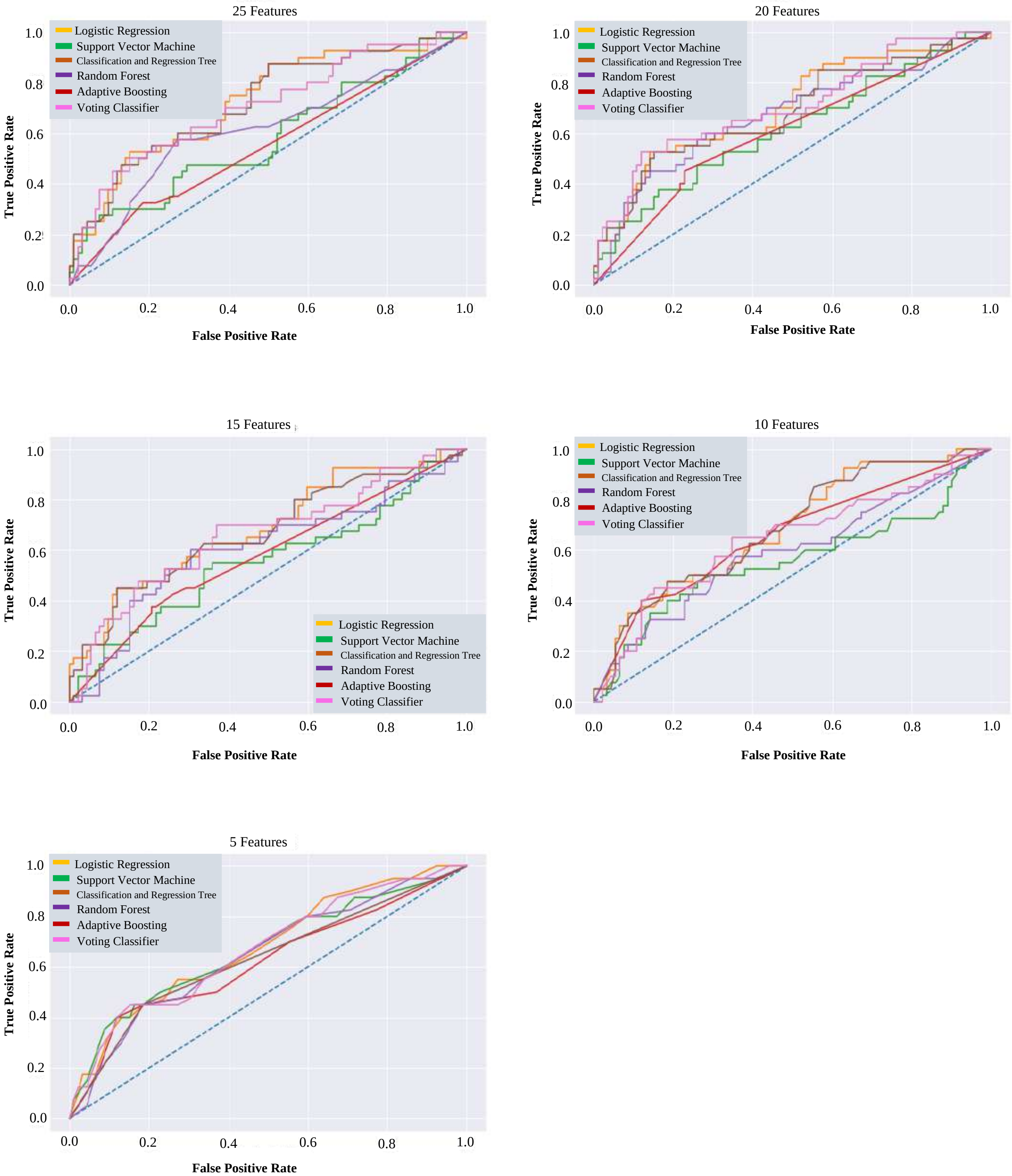


Figure 4.8: ROC Curve for Different Models Used in This Study

Table 4.1 Accuracy and AUC Score of Different Models

Model Name	No. of Features	Accuracy	AUC Score
CART	**5**	0.70	0.62
	10	0.65	0.66
	15	0.64	0.58
	20	0.66	0.60
	25	0.61	0.55
AdaBoost	**5**	0.70	0.63
	10	0.69	0.69
	15	0.70	0.67
	20	0.68	0.69
	25	0.67	0.70
LR	**5**	0.70	0.68
	10	0.69	0.69
	15	0.71	0.69
	20	0.68	0.68
	25	0.68	0.70
RF	**5**	0.70	0.65
	10	0.66	0.59
	15	0.69	0.61
	20	0.70	0.69
	25	0.69	0.61
SVM	**5**	0.70	0.67
	10	0.64	0.55
	15	0.61	0.56
	20	0.64	0.61
	25	0.63	0.58
Classifier-specific soft voting	**5**	0.70	0.67
	10	0.70	0.64
	15	0.70	0.66
	20	0.75	0.71
	25	0.71	0.71

According to Table 4.1, when the first five most significant characteristics have been incorporated into the models, the accuracy values have been found 0.70 for five distinct individual models (CART, AdaBoost, LR, RF, and SVM) and AUC scores of CART, AdaBoost, LR, RF, and SVM have been identified 0.62, 0.63, 0.68, 0.65 and 0.67 respectively. It has been observed that for first five most significant characteristics, LR and SVM outperformed other individual models with the same accuracy value of 0.70 and AUC score of 0.68 and 0.67, respectively. For the top ten most significant features, accuracy values of CART, AdaBoost, LR, RF, and SVM have been found 0.65, 0.69, 0.69, 0.66 and 0.64 respectively. AUC scores of CART, AdaBoost, LR, RF, and SVM have been identified 0.66, 0.69, 0.69, 0.59 and 0.55 respectively. AdaBoost and LR fared better for the top ten most significant features, with an accuracy value of 0.69 and an AUC score of 0.69. When the top fifteen most significant features have been put into the models, CART, AdaBoost, LR, RF, and SVM have been performed with accuracy value of 0.64, 0.70, 0.71, 0.69 and 0.61 respectively. AUC scores of CART, AdaBoost, LR, RF, and SVM have been identified 0.58, 0.67, 0.69, 0.61 and 0.56 respectively. When the top fifteen most significant features have been put into the models, LR performed better than other individual models, with an accuracy value of 0.71 and an AUC score of 0.69. When the top twenty most significant features have been put into the models, CART, AdaBoost, LR, RF, and SVM have been performed with accuracy value of 0.66, 0.68, 0.68, 0.70 and 0.64 respectively. AUC scores of CART, AdaBoost, LR, RF, and SVM have been identified 0.60, 0.69, 0.68, 0.69 and 0.61 respectively. For the first twenty most significant features, RF performed better, with an accuracy value of 0.70 and an AUC score of 0.69. There is further evidence of RF's predictive superiority in the research literature (Ji, A., and D. Levinson, 2020; Hagenauer, J., and M. Helbich, 2017 and Yassin, S. S., and Pooja, 2020). For all twenty-five selected features, accuracy values of CART, AdaBoost, LR, RF, and SVM have been found 0.61, 0.67, 0.68, 0.69 and 0.63 respectively; where AUC scores of CART, AdaBoost, LR, RF, and SVM have been identified 0.55, 0.70, 0.70, 0.61 and 0.58 respectively. It was discovered that, for all twenty-five selected features, AdaBoost and LR performed the best, with accuracy values of 0.67 and 0.68, respectively, and the identical AUC score of 0.70. Above all, this means that when the top fifteen and twenty most significant features have been picked as model predictors, accurate prediction of two vehicle crash severity has been discovered. The soft voting classifier, which combines three separate approaches AdaBoost, RF, and LR, outperforms individual models in relation to accuracy and AUC score. For each feature subset, the accuracy of soft voting classifier have found higher.

4.3.2 Graphical Representation of Accuracy and AUC Score of Different Models

The accuracy and AUC score for different combination of feature set have been expressed graphically in this study.

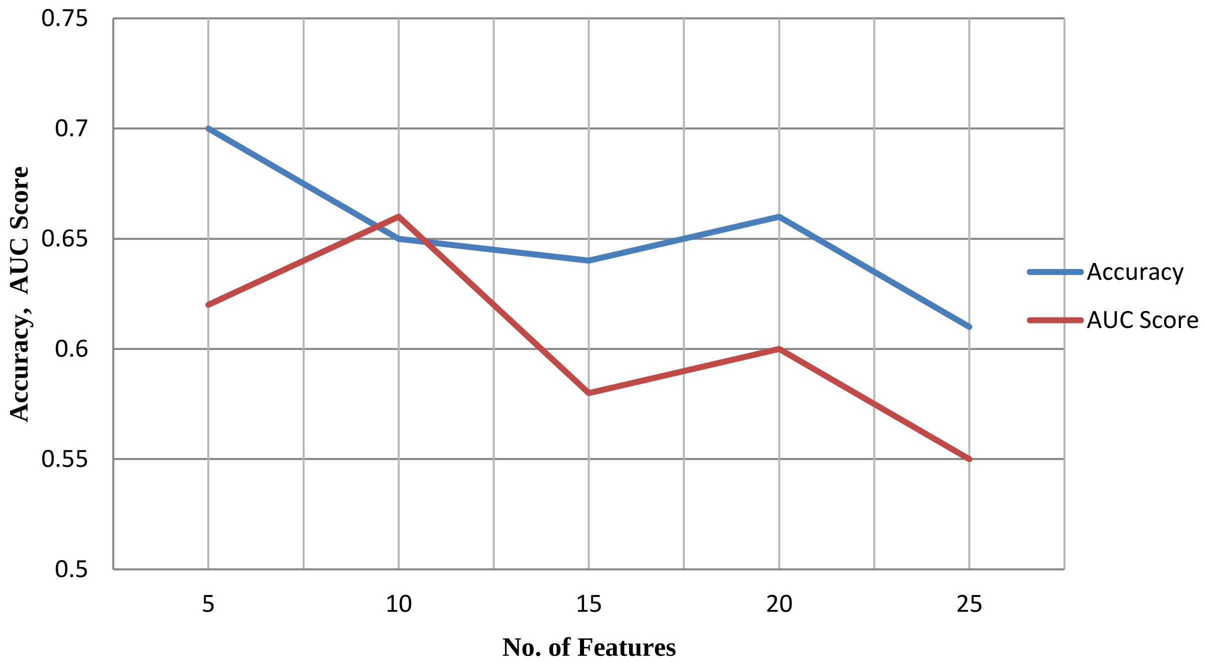


Figure 4. 9: Accuracy and AUC score for CART

When CART has been used as classifier, from Figure 4.9 it has been identified that this model has performed better when first five most significant characteristics have been incorporated into the model with an accuracy value of 0.70 and AUC score of 0.62.

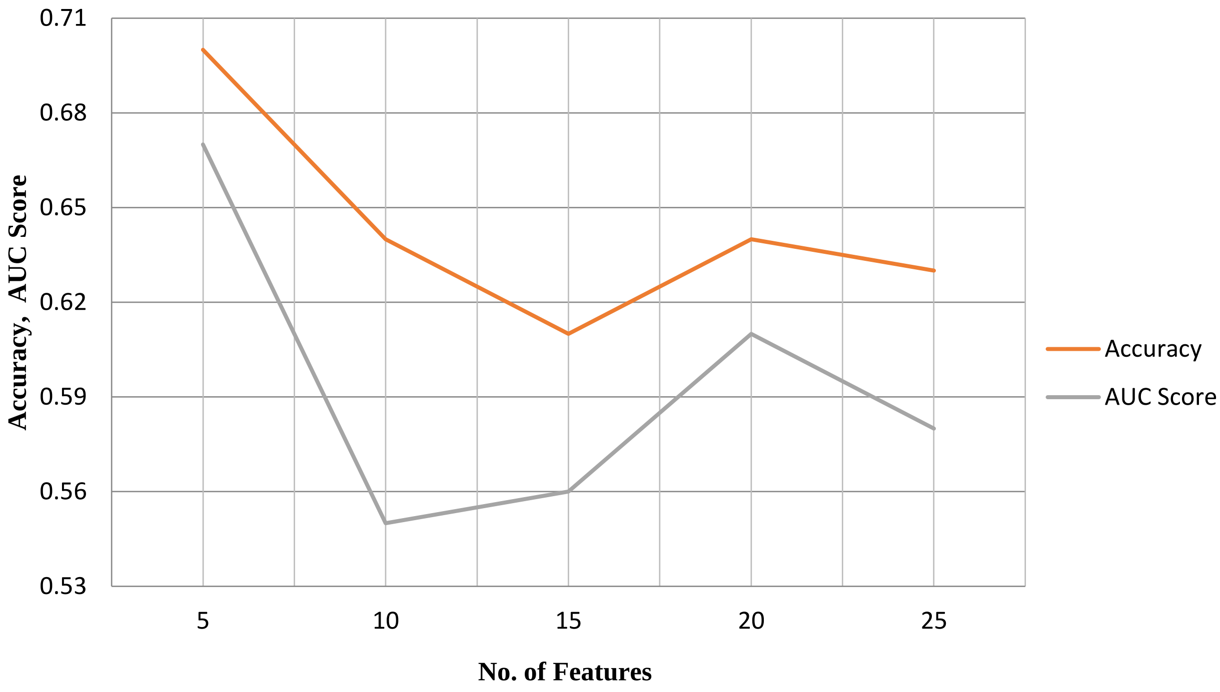


Figure 4.10: Accuracy and AUC score for SVM

When SVM have been used as classifiers, from Figure 4.10 it has been identified that this model has performed better when first five most significant characteristics have been incorporated into the model with an accuracy value of 0.70 and AUC score of 0.67.

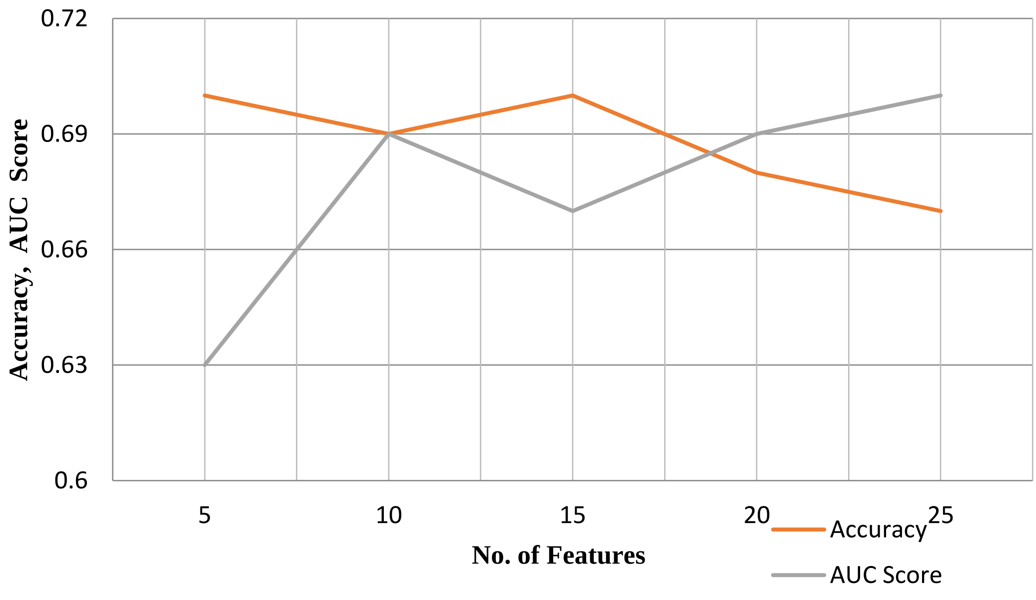


Figure 4.11: Accuracy and AUC score for AdaBoost

When AdaBoost has been used as individual classifier, from Figure 4.11 it has been identified that this model has performed better when first fifteen most significant characteristics have been incorporated into the model with accuracy value and AUC score of 0.70 and 0.67 respectively.

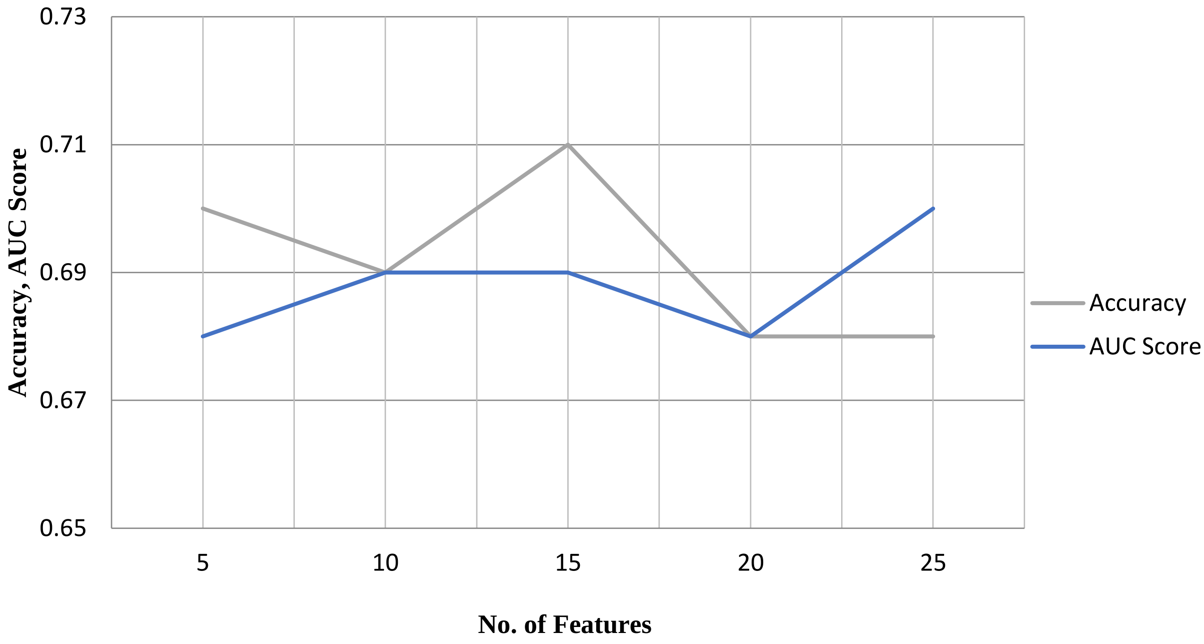


Figure 4.12: Accuracy and AUC score for LR

Similarly, when LR has been used as individual classifier, from Figure 4.12 it has been identified that this model has also performed better when first fifteen most significant characteristics have been incorporated into the model with accuracy value and AUC score of 0.71 and 0.69 respectively.

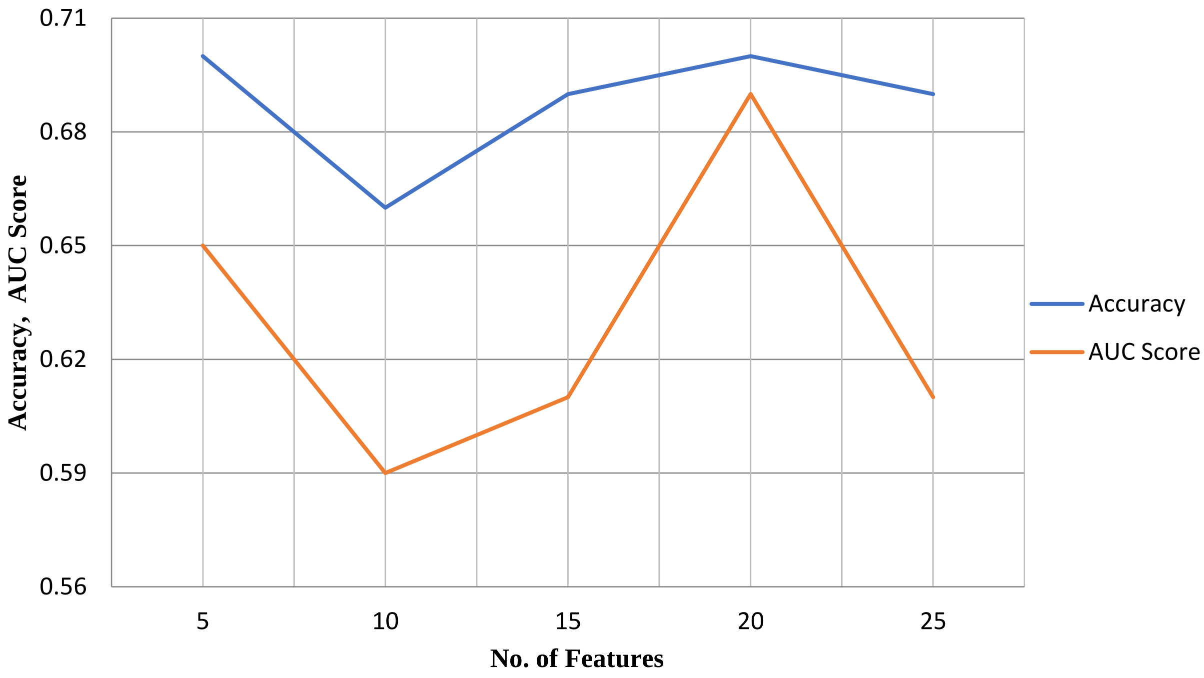


Figure 4.13: Accuracy and AUC score for RF

When the classifier RF has been used, from Figure 4.13 it has been identified that the model has performed better for first twenty feature sub-set with accuracy value and AUC score of 0.70 and 0.69 respectively.

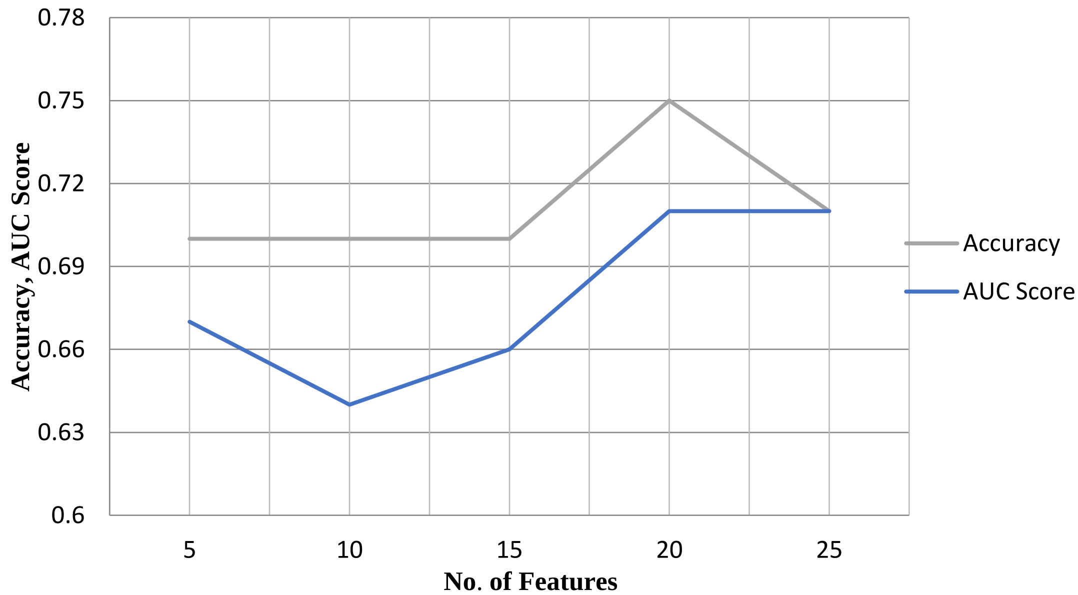


Figure 4.14: Accuracy and AUC score for hybrid model using soft voting classifier

Additionally, a hybrid model using soft voting classifier has been built with all conceivable combinations of separate models, among which a voting classifier with the combination of AdaBoost, LR, and RF demonstrated the best prediction accuracy. It was observed that a combination of different approaches based on classifier-specific soft voting performed satisfactorily (Liu et al., 2020). As shown in Figure 4.14, for the first five set of features the accuracy value and AUC score of voting classifier have been identified 0.70 and 0.67 respectively. The accuracy value and AUC score of voting classifier have been found 0.70 and 0.64 respectively for the first ten feature set. While for the first fifteen set of features, the accuracy value and AUC score of voting classifier have been identified 0.70 and 0.66 respectively. For the first twenty set of features, the accuracy value and AUC score have been identified 0.75 and 0.71 respectively. At the end, for all twenty-five feature set, both the accuracy value and AUC score have been identified 0.71 for voting classifier. It has been recognized that the voting classifier performed better in terms of accuracy (0.75) and AUC score (0.71) for the first twenty set of features.

For this reason, this study has examined the accuracy and AUC scores of each model for the first twenty characteristics and discovered that for that subset of features, RF had the best accuracy (0.70) and AUC score (0.69) of any of the models tested. As a result, this study has interpreted the global feature importance for twenty characteristics by RF selection and found the influential factors affecting two vehicle crash severities.

4.4 Model Interpretation using SHAP Methodology

This study uses the SHAP methodology to determine how well the features contribute to the severity prediction, thus finding influential factors.

4.4.1 Contributing Factors of Two vehicle Crash Severity According to SHAP Global Feature Importance

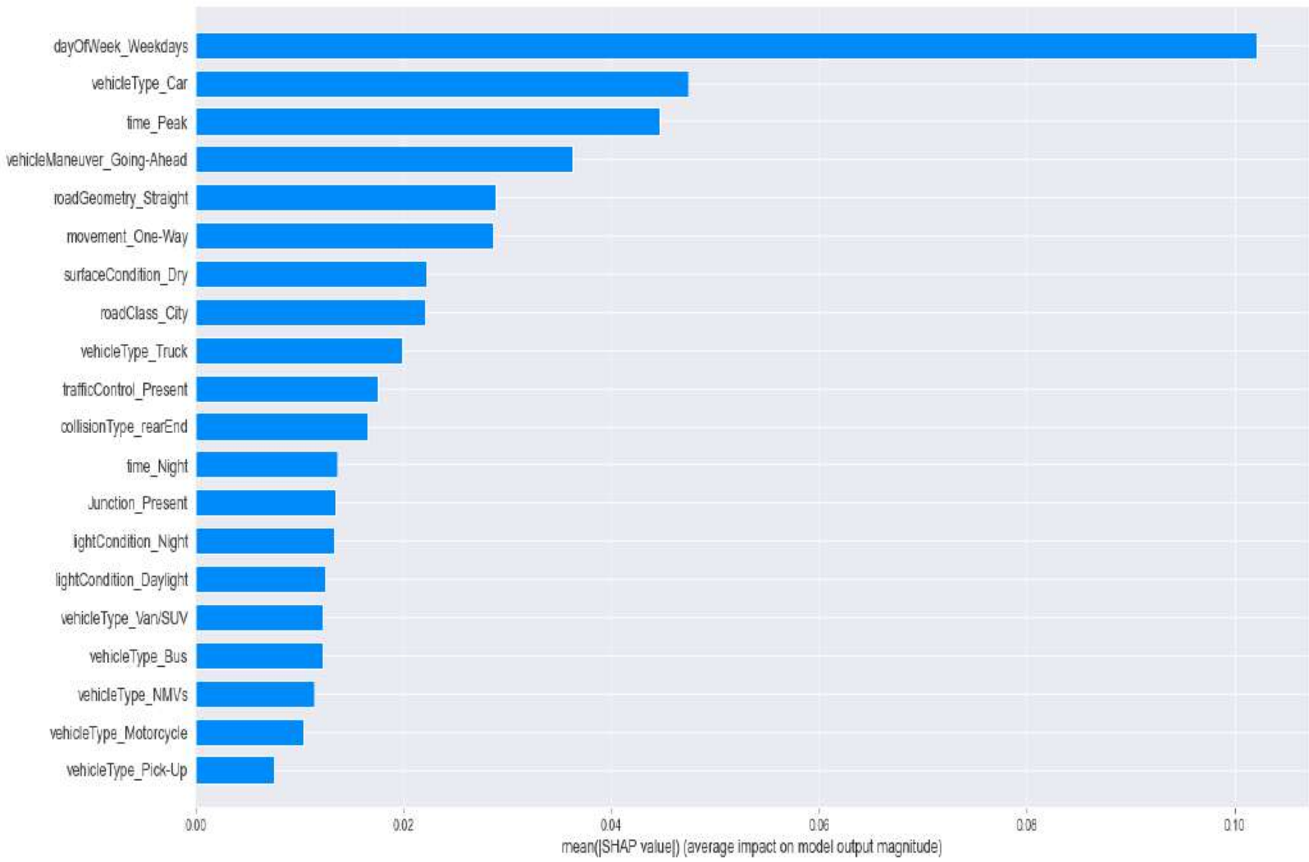


Figure 4.15: Contributing factors of two vehicle crash severity according to SHAP global feature importance

The global importance has been achieved by averaging the absolute Shapley values for each feature. This represents the marginal contribution of each feature in the prediction. For example, Shapley value for the first feature day of week (weekdays) can be determined by sampling a correlation that contains the first feature day of week (weekdays) and a correlation form by removing that feature. The difference between the respective values of these two correlations is known as marginal contribution of the first feature day of week (weekdays). This means how much the first feature day of week (weekdays) contributes to the correlation consisting the other nineteen features. From Figure 4.15, it has been identified that, according to the SHAP approach, the day of week (weekdays) has the most important contribution in predicting the severity of crashes between two vehicles. This factor alone contributed in predicting whether the two-vehicle crash severity would be fatal or not by an average of 10.2 percentage points. In addition, vehicle type (car) is another most critical variables in predicting two vehicle crash severities whether it would be fatal or not by an average of 4.8 percentage points. Other most critical variables in predicting two vehicle crash severities have been identified peak time period, straight vehicle maneuver, straight-road geometry, one-way movement, dry-surface condition, road class (city), vehicle type (truck), traffic control, collision type (rear end), night time period, junction, night-light condition, day-light condition, vehicle type (van/SUV), vehicle type (bus), vehicle type (NMVs), vehicle type (motorcycle) and vehicle type (pick-up) which have been contributed in predicting whether the two-vehicle crash severity would be fatal or not by an average of 4.3, 3.8, 2.9, 2.9, 2.2, 2.2, 2.0, 1.8, 1.7, 1.5, 1.5, 1.5, 1.4, 1.4, 1.4, 1.3, 1.2 and 0.9 respectively. However, it is until unclear the ways in which these features impact the prediction of crash severities.

It has been found that, the day of week, vehicle type, time of the day, vehicle maneuver, and road geometry are the most significant explanatory variables and have important contribution in predicting crash severities which is consistent with our previous studies (Chiou et al., 2020; Yuan et al., 2022; Lee, C., and X. Li, 2014 and Ji, A., and D. Levinson, 2020).

4.4.2 Contributing Factors of Two vehicle Crash Severity According to SHAP Local Explanation

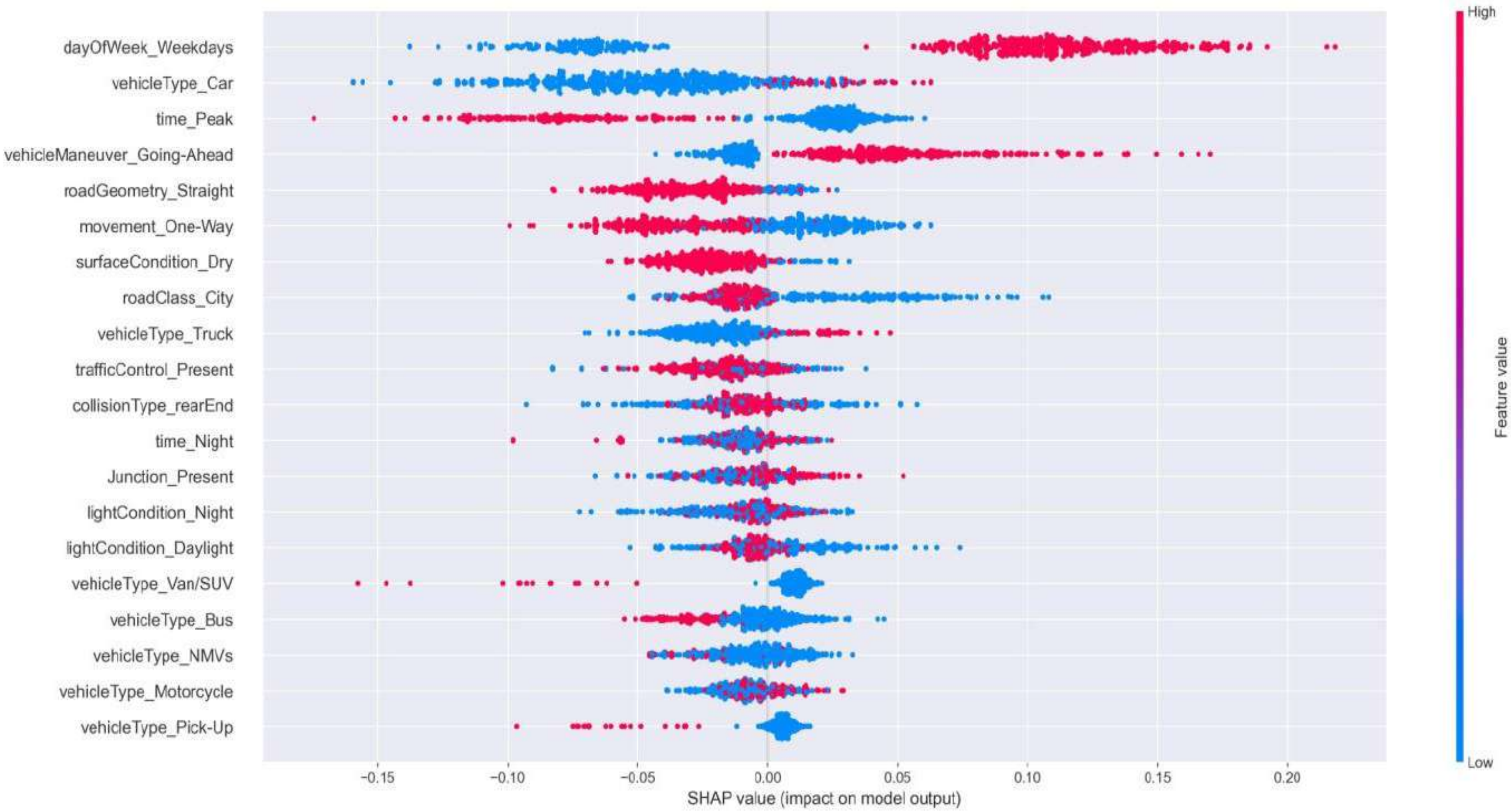


Figure 4.16: Contributing Factors of Two Vehicle Crash Severity According to SHAP Local Explanation

To identify how the contributing factors affect two vehicle crash severities, Figure 4.16 depicts the SHAP values of the elements determining the severity of a two-vehicle crash. It should be noted that SHAP values more than zero indicate positive effects on the risk of a fatal accident occurring, whilst SHAP values less than zero indicate negative consequences. The graphic depicts the local size, distribution, and direction of the contributing elements in determining whether a two vehicle crash will be fatal or non-fatal. Weekday, for example, has a red tail on the right and a short blue tail on the left, as illustrated in Figure 4.16. It means that if a two-vehicle collision occurs during the week, the accident is more likely to be fatal. Furthermore, the lengthy red tail implies that weekend accidents are not nearly as important as weekday accidents in affecting two vehicle crash severities. It can deduce from the vehicle type variable that if at least one of the vehicles involved in the collision is a private car, truck, or motorcycle, the probability of a fatal accident increases, because the red tail of a private car, truck, or motorcycle only appears on the right, whereas the red tail of an NMV, bus, van/SUV and pick-up appears on the left. Because larger vehicles like bus, pick-up and SUV provide better protection, resulting in fatal crashes for private cars and motorcycles, which is consistent with previous research (Yuan et al., 2022 and Yu, R., and M. Abdel-Aty, 2014). The vehicle maneuver depicts a long red tail on the right and a short blue tail on the left. This means that while straight-moving vehicles are more likely to be involved in a fatal accident, vehicles that are not travelling straight are less likely to do so. The fact behind this could be the unconsciousness of the drivers while moving straight without any turning on their way. The data also reveals that multiple vehicle crashes during off-peak hours seem to be more likely to result in mortality. The road geometry depicts a short blue tail on the right and a long red tail on the left. This means that, the accidents are less likely to be fatal if the roads are straight. The result is consistent with a previous study (Lee, C., and X. Li, 2014). The movement depicts a blue tail on the right and a red tail on the left. This means that, the accidents are more likely to be fatal for two-way movement and less likely to be fatal for one way movement. The surface condition has a short blue tail on the right and a long red tail on the left. It means that when the surface is wet, the accidents are more likely to be a fatal accident and less likely to be fatal when the surface is dry. The data also reveals that if the road flows to outside of the city the accidents are more likely to be a fatal accident as it depicts a long blue tail on the right for city road. The traffic control variable shows a short blue tail on the right and a red tail on the left. That means when there are no traffic control the accidents are more likely to be fatal. It has also identified that two vehicle crashes during night time seem to be more likely to result in mortality, because the red tail of night time variable has appeared on the right. At junction, the two vehicle crashes are more likely to be fatal due to a red tail on the right, whereas the blue tail on the left which means the accidents are less likely to be fatal when there is no junction. If a two vehicle crash occurs during the night light condition, the accidents are more likely to be fatal. Furthermore, the blue tail on the left implies that the accident occurs in day light are not likely to be fatal.

4.5 Overview

The analytical portion of this thesis can be divided into three stages: understanding the importance of analyzing two vehicle crashes severities of the country, applying different machine learning (ML) methods and traditional statistical regression method to identify the contributing factors responsible for such crashes and also compare the prediction accuracy of different models.

Chapter 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 General

This research analyzed crash severity prediction of two vehicles using LR as a statistical method and some popular ML methods such as CART, SVM, Adaboost, RF, and soft voting classifier. Different independent variables like vehicle characteristics (vehicle type, vehicle maneuver), driver characteristics (age, sobriety condition, seatbelt/helmet usage), roadway conditions/environment (road geometry, surface condition, traffic control, movement, light condition, junction, road class,), crash characteristics (collision type) and temporal features (time, day of week) have been used in predicting two vehicle crash severities. Using RF, this study employs a feature engineering strategy to extract valuable features for boosting ML classifier performance. This is the first study to compare classifier-specific Soft Voting with individual classifiers to see if individual or hybrid models can forecast the severity of a two vehicle crash with greater accuracy. The SHAP method, a brand new ML model interpretation technique, was used to identify possible factors influencing two vehicle collision severity, as well as their relative size, distribution, and direction in estimating two vehicle crash severities.

5.2 Conclusions

It is the first to use data-driven ML techniques to predict two vehicle crash severity in a low-income country context, Bangladesh, specifically Dhaka city. The major conclusions and findings of the research are summarized below:

(i) It has been found that when twenty features are utilized to evaluate the severity of a two-vehicle crash, RF achieves the highest accuracy and AUC score.

(ii) Between hybrid and individual classifiers, the soft voting classifier, which combines three separate approaches AdaBoost, RF, and LR, outperforms individual models in relation to accuracy (0.75) and AUC score (0.71).

(iii) According to the SHAP approach, it has been found that, the day of week, vehicle type, time of the day, vehicle maneuver, and road geometry are the most significant explanatory variables and have important contribution in predicting crash severities by an average of 10.2, 4.8, 4.6, 3.8 and 2.9 percentage points respectively.

(iv) If a two-vehicle crash occurs during the weekdays, the accident is more likely to be fatal probably due to heavy traffic flow.

(v) If at least one of the vehicles included in the crash is a private car, truck or motorcycle, the probability of a fatal accident increases. Other types of vehicles, such as buses and vans/SUV, are less likely to be seriously injured than passenger cars and motorcycles because larger vehicles provide better protection, resulting in fatal crashes for private cars and motorcycles.

(vi) Furthermore, in Dhaka city, after a certain time period, the movement of trucks prevails over other types which initiate the chances of truck-truck collisions may lead to fatal injuries.

(vii) Time is another critical variable in predicting two vehicle crash severities. Off-peak hours seem to be more likely to occur in fatal vehicle crashes. This may be due to the fact that, at off peak hours, multiple vehicles move with higher speed than peak hours and results in fatal injuries.

(viii) Another variable is vehicle maneuver that has much contribution in predicting two vehicle crash severities. In this study it has found that, straight-moving vehicles are more likely to be involved in a catastrophic accident than those are not travelling straight.

(ix) Another factor showed that, the vehicles moving on a straight road are less likely to be involved in a fatal accident than those are on a curved road.

5.3 Recommendations for Future Studies

It is being perceived that by overcoming the study limitations new research horizons would be yielded. These can be abridged as follows:

- This study has been concerned with how accident severities are attributed to vehicle related factors (vehicle type, vehicle maneuver), driver characteristics (age, sobriety condition, seatbelt/helmet usage), roadway and environment conditions (road geometry, surface condition, junction, road class, traffic control, movement, light condition), crash characteristics (collision type), and temporal features (time, day of week). But there was no information in the data to indicate which of the two-vehicle crashes were caused by pedestrian activity. The primary cause of the two-vehicle crash that may have been caused by a pedestrian action cannot be identified because of improper reporting system in Bangladesh. However it is important to include pertinent information on pedestrian activity before an accident.

- In-depth analyses of this research finding for developing countermeasures and policy level decisions would provide enormous scopes for future endeavors.

- For better prediction accuracy, deep machine learning can be applied to the data set.

- Advanced modelling techniques viz., artificial neural network, can also be applied on the same crash data used in this thesis.

- This study has done only for Dhaka city. It can be conducted for other cities of Bangladesh as well.

REFERENCES

Abdel-Aty, M. A., and Abdelwahab, H. T., (2004). Predicting injury severity levels in traffic crashes: A modeling comparison. *Journal of Transportation Engineering*, vol. 130, no. 2, pp. 204–210. https://doi.org/ 10.1061/(ASCE)0733-947X(2004)130:2(204).

Abdelwahab, H. T., and M. A. Abdel-Aty (2001). Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1746, No. 1, pp. 6–13. https://doi.org/10.3141/1746-02.

Abay, K. A., Paleti, R., & Bhat, C. R. (2013). The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity. *Transportation Research Part B: Methodological*, 50, 74–89. https://doi.org/10.1016/j.trb.2013.01.007.

Abellán, J., López, G., & de Oña, J. (2013). Analysis of traffic accident severity using decision rules via decision trees. *Expert Systems with Applications*, 40(15), 6047–6054. https://doi.org/10.1016/j. eswa.2013.05.027.

Aguero-Valverde, J., & Jovanis, P. P. (2009). Bayesian multivariate poisson lognormal models for crash severity modeling and site ranking. *Transportation Research Record: Journal of the Transportation Research Board*, 2136(1), 82–91. https://doi.org/10.3141/2136-10.

Al-mahmood, S.Z., 2007. Death on The Asphalt, Star Weekend Magazine. *The Daily Star*, Dhaka.

Amiri, A. M., Sadri, A., Nadimi, N., & Shams, M. (2020). A comparison between artificial neural network and hybrid intelligent genetic algorithm in predicting the severity of fixed object crashes among elderly drivers. *Accident Analysis and Prevention*, 138, 105468. https://doi.org/10.1016/j.aap.2020.105468.

Anderson, T. K. (2009). Kernel density estimation and K-Means clustering to profile road accident hotspots. *Accident Analysis and Prevention*, 41(3), 359–364. https://doi.org/10.1016/j.aap.2008.12.014.

Anvari, M. B., Kashani, A. T., and Rabieyan, R. (2017). Identifying the most important factors in the at-fault probability of motorcyclists by data mining, based on classification tree models. *International Journal of Civil Engineering*, vol. 15, no. 4, pp. 653–662.

Arhin, S. A., & Gatiba, A. (2020). Predicting crash injury severity at unsignalized intersections using support vector machines and naïve bayes classifiers. *Transportation Safety and Environment*, 2(2), 120– 132. https://doi.org/10.1093/tse/tdaa012.

Azimi, G., Rahimi, A., Asgari, H., & Jin, X. (2020). Severity analysis for large truck rollover crashes using a random parameter ordered logit model. *Accident Analysis and Prevention*, 135, 105355. https:// doi.org/10.1016/j.aap.2019.105355.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. *Springer*, ISBN 978-0- 387-31073-2.

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). A Training Algorithm for Optimal Margin Classifiers.

Breiman, L. Random Forests (2001). *Machine Learning*, Vol. 45, No. 1, pp. 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and Regression Trees. *Wadsworth Inc., Pacific Grove, CA.*

Budiawan, W., Saptadi, S., Tjioe, C., & Phommachak, T. (2019). Traffic accident severity prediction using naive Bayes algorithm-a case study of Semarang toll road. *IOP Conference Series: Materials Science and Engineering*, 598, 012089. IOP Publishing. https://doi.org/10.1 088/1757-899X/598/1/012089.

Cai, Q., M. Abdel-Aty, O. Zheng, and Y. Wu. (2022). Applying Machine Learning and Google Street View to Explore Effects of Drivers’ Visual Environment on Traffic Safety. *Transportation Research Part C: Emerging Technologies*, Vol. 135, https://doi.org/10.1016/j.trc.2021.103541.

Cao, J., S. Kwong, R. Wang, X. Li, K. Li, and X. Kong (2015). Class-Specific Soft Voting Based Multiple Extreme Learning Machines Ensemble. *Neurocomputing*, Vol. 149, No. Part A, pp. 275–284. https://doi.org/10.1016/J.NEUCOM.2014.02.072.

Champahom, T., Jomnonkwao, S., Watthanaklang, D., Karoonsoontawong, A., Chatpattananan, V., & Ratanavaraha, V. (2020). Applying hierarchical logistic models to compare urban and rural roadway modeling of severity of rear-end vehicular crashes. *Accident Analysis and Prevention*, 141, 105537 https://doi. org/10.1016/j.aap.2020.105537.

Chang, L. Y., & Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, 38(5), 1019–1027. https://doi.org/10. 1016/j.aap.2006.04.009.

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, 90, 128-139.

Chen, C., Zhang, G., Huang, H., Wang, J., & Tarefder, R. A. (2016). Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. *Accident Analysis and Prevention*, 96, 79–87. https://doi.org/10.1016/j. aap.2016.06.015.

Chen, F., and Chen, S. (2011). Injury Severities of Truck Drivers in Single- and Multi-Vehicle Accidents on Rural Highways. *Accident Analysis & Prevention*, Vol. 43, No. 5, pp. 1677–1688.

Chen, C., Zhang, G. H., Qian, Z., Tarefder, R. A., & Tian, Z. (2016a). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, 90, 128–139. https://doi.org/10.1016/j.aap.2016.02.011.

Chen, C., Zhang, G. H., Yang, J. F., Milton, C. J., & Alcántara, A. D. (2016b). An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. *Accident Analysis & Prevention*, 90, 95–107. https://doi.org/10.1016/j.aap.2016.02.00.

Chiou, Y. C., Fu, C. and Ke, C. Y. (2020). Modelling Two-Vehicle Crash Severity by Generalized Estimating Equations. *Accident Analysis and Prevention*, Vol. 148, https://doi.org/10.1016/j.aap.2020.105841.

Chong, M., Abraham, A., and Paprzycki, M. (2005). Traffic Accident Analysis Using Machine Learning Paradigms. *Accident Analysis and Prevention*, No. 29, pp. 89–98.

Christoforou, Z., Cohen, S., & Karlaftis, M. G. (2010). Vehicle occupant injury severity on highways: An empirical investigation. *Accident Analysis and Prevention*, 42(6), 1606–1620. https://doi. org/10.1016/j.aap.2010.03.019.

Cortes, C., Vapnik, V., and Saitta, L. (1995). *Support-Vector Networks Editor*. *Kluwer Academic Publishers*.

Das, A., Abdel-Aty, M., and Pande, A. (2009). Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. *Journal of Safety Research*, vol. 40, no. 4, pp. 317–327.

Ding, C. and Peng, H. (2008). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2): 185–205.

Dong, N., Huang, H., & Zheng, L. (2015). Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects. *Accident; Analysis and Prevention*, 82, 192–198, https://doi.org/10.1016/j.aap.2015.05.018.

Duncan, C. S., Khattak, A. J. and Council, F. M. (1998). Applying the Ordered Probit Model to Injury Severity in Truck-Passenger Car Rear-End Collisions. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1635, No. 1, pp. 63–71. https://doi.org/10.3141/1635-09.

Elamrani Abou Elassad, Z., Mousannif, H., & Al Moatassime, H. (2020). Class-imbalanced crash prediction based on real-time traffic and weather data: a driving simulator study. *Traffic Injury Prevention*, 21(3), 201–208. https://doi.org/10.1080/15389588.2020. 1723794.

Elamrani A. E., Mousannif, Z. H., and Moatassime, H. A. (2020). A Proactive Decision Support System for Predicting Traffic Crash Events: A Critical Analysis of Imbalanced Class Distribution. *Knowledge-Based Systems*, Vol. 205. https://doi.org/10.1016/j.knosys.2020.106314.

Eluru, N., & Bhat, C. R. (2007). A joint econometric analysis of seat belt use and crash-related injury severity. *Accident Analysis and Prevention*, 39(5), 1037–1049. https://doi.org/10.1016/j. aap.2007.02.001.

Fan, Z., Liu, C., Cai, D., and Yue, S. (2019). Research on Black Spot Identification of Safety in Urban Traffic Accidents Based on Machine Learning Method. *Safety Science*, Vol. 118, pp. 607–616. https://doi.org/10.1016/J.SSCI.2019.05.039.

Farhat, Z., Karouni, A., Daya, B., Chauvet, P., and Hmadeh, N. (2020). Traffic Accidents Severity Prediction Using Support Vector Machine Models. *International Journal of Innovative Technology and Exploring Engineering*, Vol. 9, No. 7, pp. 1345–1350. https://doi.org/10.35940/ijitee.F4393.059720.

Ferrari, E., & Bacciu, D. (2021). Addressing Fairness, Bias and Class Imbalance in Machine Learning: the FBI-loss. arXiv:2105.06345.

Fiorentini, N., & Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures*, 5(7), 61. https://doi.org/10.3390/infrastructures5070061.

Freund, Y., and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, Vol. 55, No. 1, pp. 119–139. https://doi.org/10.1006/jcss.1997.1504.

Geedipally, S., and Lord, D. (2010). Identifying Hot Spots by Modeling Single-Vehicle and Multivehicle Crashes Separately. *Transportation Research Record: Journal of the Transportation Research Board,*. 2147: 97–104.

Geedipally, S. R., and Lord, D. (2010). Investigating the Effect of Modeling Single-Vehicle and Multi-Vehicle Crashes Separately on Confidence Intervals of Poisson–Gamma Models. *Accident Analysis & Prevention*, Vol. 42, No. 4, pp. 1273–1282.

Griffith, M. S. (1999). Safety Evaluation of Rolled-In Continuous Shoulder Rumble Strips Installed on Freeways. *Transportation Research Record: Journal of the Transportation Research Board*, 1665(1): 28–34.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–82.

Hagenauer, J., and Helbich, M. (2017). A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice. *Expert Systems with Applications*, Vol. 78, pp. 273–282. https://doi.org/10.1016/J.ESWA.2017.01.057.

Hartigan, J.A. (2001). Statistical Clustering, International Encyclopedia of the Social & Behavioral Sciences.

Hill, T., and Lewicki, P. (2006) Statistics: Methods and Applications, A Comprehensive Reference for Science, Industry, and Data Mining. *StatSoft Inc., Tulsa, OK*.

Hossain, M. (2011). Development of a Real-Time Proactive Road Safety Management System for Urban Expressways. *Ph. D. Thesis, Department of Built Environment, Tokyo Institute of Technology*.

Hoque, M. M., Anowar, S., & Raihan, M. A. (2008). Towards sustainable road safety in Bangladesh. *Paper presented at the International Conference on Sustainable Transport for Developing Countries: Concerns, Issues and Options, Dhaka, Bangladesh*, 63–69.

Huang, H., Chin, H. C., & Haque, M. M. (2008). Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis. *Accident Analysis and Prevention*, 40(1), 45–54. https://doi.org/10.1016/j.aap.2007.04.002.

Huang, H., Siddiqui, C., & Abdel-Aty, M. (2011). Indexing crash worthiness and crash aggressivity by vehicle type. *Accident Analysis and Prevention*, 43(4), 1364–1370. https://doi.org/10.1016/j. aap.2011.02.010.

Huan L., and Hiroshi M. (2007). Computational Methods of Feature Selection. *CRC Press*.

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

Jamal, A., Zahid, M., Rahman, T., Hassan, M., Al-Ahmadi, H. M., Almoshaogeh, M., Farooq, D., and 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*, Vol. 28, No. 4, pp. 408–427. https://doi.org/10.1080/17457300.2021.1928233.

Ji, A., and Levinson, D. (2020). Injury Severity Prediction From Two-Vehicle Crash Mechanisms With Machine Learning and Ensemble Models. *IEEE Open Journal of Intelligent Transportation Systems*, Vol. 1, pp. 217–226. https://doi.org/10.1109/ojits.2020.3033523.

Jiang, L., Xie, Y., and Ren, T. (2019). Modelling highly unbalanced crash injury severity data by ensemble methods and global sensitivity analysis. *In Procceding Transportation Research Board 98th Annual Meeting, Washington, DC, USA*, pp. 13–17.

Khattak, A. J., Kantor, P., & Council, F. M. (1998). Role of adverse weather in key crash types on limited-access: Roadways implications for advanced weather systems. *Transportation Research Record: Journal of the Transportation Research Board*, 1621(1), 10–19. https:// doi.org/10.3141/1621-02.

Khorashadi, A., Niemeier, D., Shankar, V., and Mannering, F. (2005). Differences in rural and urban driver-injury severities in accidents involving largetrucks: An exploratory analysis. *Accident Analysis & Prevention*, vol. 37, no. 5, pp. 910_921.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868.

Kitali, A. E., and Sando, T. (2017). A Full Bayesian Approach to Appraise the Safety Effects of Pedestrian Countdown Signals to Drivers. *Accident Analysis & Prevention*, Vol. 106, pp. 327–335.

Kitali, A. E., Kidando, E., Raihan, M. A., Kutela, B., Alluri, P., and Sando, T. (2021). Exploring the Need to Model Two- and Multiple-Vehicle Crashes Separately. *Transportation Research Record*: *Journal of Transportation Research Board*, Vol. 2676, No. 1, pp. 622–636. https://doi.org/10.1177/03611981211037882.

Laureshyn, A., Svensson, Å., & Hydén, C. (2010). Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation. *Accident Analysis & Prevention*, 42(6), 1637–1646. https://doi.org/10.1016/j.aap.2010.03.021.

Lee, C., & Abdel-Aty, M. (2008). Presence of passengers: Does it increase or reduce driver's crash potential? *Accident Analysis and Prevention*, 40(5), 1703–1712. https://doi.org/10.1016/j. aap.2008.06.006.

Lee, C. and Abdel-Aty, M. (2005). Comprehensive analysis of vehicle_pedestrian crashes at intersections in Florida. *Accident Analysis & Prevention*, vol. 37, no. 4, pp. 775_786.

Lee, C., and Li, X. (2014). Analysis of Injury Severity of Drivers Involved in Single- and Two-Vehicle Crashes on Highways in Ontario. *Accident Analysis and Prevention*, Vol. 71, pp. 286–295. https://doi.org/10.1016/j.aap.2014.06.008.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1), 1–30. https://doi.org/10.1186/s40537-018-0151-6.

Liao, Y., Zhang, J., Wang, S., Li, S., and Han, J. (2018). Study on Crash Injury Severity Prediction of Autonomous Vehicles for Different Emergency Decisions Based on Support Vector Machine Model. *Electronics (Switzerland)*, Vol. 7, No. 12, https://doi.org/10.3390/electronics7120381.

Liu, L., Zhang, X., Liu, Y., Zhu, W., and Zhao, B.(2020). An Ensemble of Multiple Boosting Methods Based on Classifier-Specific Soft Voting for Intelligent Vehicle Crash Injury Severity Prediction. *IEEE Sixth International Conference on Big Data Computing Service and Applications,* pp. 17-24, DOI 10.1109/BigDataService49289.2020.00011.

Li, X., Lord, D., Zhang, Y., and Xie, Y. (2008). Predicting Motor Vehicle Crashes Using Support Vector Machine Models. *Accident Anaysis & Prevention*, vol. 40, no. 4, pp. 1611–1618.

Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention*, 45, 478–486. https://doi.org/10.1016/j.aap.2011.08.016.

Li, Z., Wang, H. X., Zhang, Y. W., and Zhao, X. H. (2020). Random Forest–Based Feature Selection and Detection Method for Drunk Driving Recognition. *International Journal of Distributed Sensor Networks*, Vol. 16, No. 2, https://doi.org/10.1177/1550147720905234.

Li, Z., Chen, C., Ci, Y., Zhang, G., Wu, Q., Liu, C., & Qian, Z. S. (2018). Examining driver injury severity in intersection-related crashes using cluster analysis and hierarchical bayesian models. *Accident Analysis and Prevention*, 120, 139–151. https://doi. org/10.1016/j.aap.2018.08.009.

Li, Z., Wang, W., Liu, P., Bigham, J. M., and Ragland, D. R. (2013). Using geographically weighted poisson regression for county-level crash modeling in California. *Safety Science,* vol. 58, no. 10, pp. 89_97.

Li, Z., Wu, Q., Ci, Y., Chen, C., Chen, X., & Zhang, G. (2019). Using latent class analysis and mixed logit model to explore risk factors on driver injury severity in single-vehicle crashes. *Accident Analysis and Prevention*, 129, 230–240. https://doi.org/10.1016/j. aap.2019.04.00.

Lombardi, D. A., Horrey, W. J., & Courtney, T. K. (2017). Age-related differences in fatal intersection crashes in the United States. *Accident Analysis & Prevention*, 99, 20‑29. https://doi.org/10.1016/j.aap.2016.10.030.

Louis Burger Group-BCL (2005). Strategic Transport Plan. *Dhaka Transport Coordination Board, Ministry of Communications, Government of Bangladesh.*

Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis & Prevention*, Vol. 37, No. 1, pp. 35–46.

Lu, P., Zheng, Z., Ren, Y., Zhou, X., Keramati, A., Tolliver, D., & Huang, Y. (2020). A gradient boosting crash prediction approach for highway-rail grade crossing crash analysis. *Journal of Advanced Transportation*, 2020, 1–10. https://doi.org/10.1155/2020/ 6751728.

Lundberg, S. M., and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions.

Ma, J., and Kockelman, K. M. (2006). Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity. *Transportation Research Record: Journal of the Transportation Research Board*, 1950(1): 24–34.

Mahmud, S. M. S., Hoque, M. S., Qazi, A. S. (2009). Road Safety Problems in Bangladesh: Some Major Initiatives, Constraints and Requirements. *Transport and Communications Bulletin (TCB) for Asia and the Pacific, Economic and Social Commission for Asia and the Pacific (ESCAP)*, No.79, pp. 47-67.

Ma, C., Hao, W., Xiang, W., & Yan, W. (2018). The impact of aggressive driving behavior on driver-injury severity at highway-rail grade crossings accidents. *Journal of Advanced*.

Ma, C., Peng, Y., Wu, L., Guo, X., Wang, X., & Kong, X. (2022,). Application of Machine Learning Techniques to Predict the Occurrence of Distraction-Affected Crashes with Phone-Use Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2676, No. 2, pp. 692–705. https://doi.org/10.1177/03611981211045371.

Ma, X., Chen, S., & Chen, F. (2016). Correlated Random-Effects Bivariate Poisson Lognormal Model to Study Single-Vehicle and Multivehicle Crashes. *Journal of Transportation Engineering*, p. 04016049.

Malyshkina, N. V., & Mannering, F. L. (2009). Markov switching multinomial logit model: An application to accident-injury severities. *Accident Analysis and Prevention*, 41(4), 829–838. https://doi. org/10.1016/j.aap.2009.04.006.

Martensen, H., and Dupont, E. (2013). Comparing Single Vehicle and Multivehicle Fatal Road Crashes: A Joint Analysis of Road Conditions, Time Variables and Driver Characteristics. *Accident Analysis & Prevention*, Vol. 60, pp. 466–471.

Mauro, R., De Luca, M., & Dell'Acqua, G. (2013). Using a K-means clustering algorithm to examine patterns of vehicle crashes in before-after analysis. *Modern Applied Science*, 7(10), 11. https://doi. org/10.5539/mas.v7n10p11.

Mesa-Arango, R., Valencia-Alaix, V. G., Pineda-Mendez, R. A., & Eissa, T. (2018). Influence of socioeconomic conditions on crash injury severity for an urban area in a developing country. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(31), 41–53. https://doi.org/10.1177/0361198118758684.

Milton, J. C., Shankar, V. N., & Mannering, F. L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis and Prevention*, 40(1), 260–266. https://doi.org/10.1016/j.aap.2007.06.006.

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2019). Improved support vector machine models for work zone crash injury severity prediction and analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2673(11), 680–692. https://doi.org/10.1177/0361198119845899.

Mondal, A. R., Bhuiyan, M. A. E., & Yang, F. (2020). Advancement of weather-related crash prediction model using nonparametric machine learning algorithms. *SN Applied Sciences*, 2(8), 1–11.

Moons, E. A. L. M. G., Wets, G. P. M., Aerts, M., Arentze, T. A., and Timmermans, H. J. P. (2016). The Impact of Simplification in a Sequential Rule-Based Model of Activity-Scheduling Behavior. *SAGE Journal*, Vol. 37, No. 3, pp. 551–568. http://dx.doi.org/10.1068/a36167.

Mussone, L., Ferrari, A., and Oneta, M. (1999). An analysis of urban collisions using an artificial intelligence model. *Accident Analysis & Prevention*, vol. 31, no. 6, pp. 705_718.

O'Donnell, C. J., & Connor, D. H. (1996). Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis & Prevention*, 28(6), 739–753. https://doi. org/10.1016/S0001-4575(96)00050-4.

Osman, M., Paleti, R., Mishra, S., & Golias, M. M. (2016). Analysis of injury severity of large truck crashes in work zones. *Accident Analysis and Prevention*, 97, 261–273. https://doi.org/10.1016/j.aap.2016.10.020.

Oña, J. d., López, G., & Abellán, J. (2013). Extracting decision rules from police accident reports through decision trees. *Accident Analysis and Prevention*, 50, 1151–1160. https://doi.org/10.1016/j. aap.2012.09.006.

Oña, J. de, Mujalli, R. O., and Calvo, F. J. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention*, vol. 43, no. 1, pp. 402–411.

Paleti, R., Eluru, N., and Bhat, C. R. (2010). Examining the in_uence of aggressive driving behavior on driver injury severity in traf_c crashes. *Accident Analysis & Prevention*, vol. 42, no. 6, pp. 1839_1854.

Pan XY and Shen HB (2009). Prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. *Protein & Pepttide Letters*, 16(12): 1447–1454.

Parsa, A. B., A. Movahedi, H. Taghipour, S. Derrible, and Mohammadian A. (2020),. Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis. *Accident Analysis & Prevention*, Vol. 136, p. 105405. https://doi.org/10.1016/J.AAP.2019.105405.

Pasupathy, Ivan, R. J. N., and Ossen, P. J. (2000). Single and Multi-Vehicle Crash Prediction Models for Two-Lane Roadways. *NEUTC UCNR9-8, Final Report, Durham, NH*.

Pradhan, B., and Sameen, M. I., (2020). Predicting Injury Severity of Road Traffic Accidents Using a Hybrid Extreme Gradient Boosting and Deep Neural Network Approach. *In Advances in Science, Technology and Innovation, Springer Nature*, pp. 119–127.

Prasetyo, E., (2012). Data Mining konsepdanAplikasimenggunakan MATLAB. Yogyakarta: Andi.

Qin, X., Ivan J. N., & Ravishanker, N. (2004). Selecting Exposure Measures in Crash Rate Prediction for Two-Lane Highway Segments. *Accident Analysis & Prevention*, Vol. 36, No. 2, pp. 183–191.

Raihan, M. A., Hossain, M., & Hasan, T. (2017). Data Mining in Road Crash Analysis: The Context of Developing Countries. *International Journal of Injury Control and Safety Promotion*, Vol. 25, No. 1, pp. 41–52. https://doi.org/10.1080/17457300.2017.1323929.

Rifaat, S. M., & Chin, H. C. (2007). Accident severity analysis using ordered probit model. *Journal of Advanced Transportation*, 41(1), 91–114. https://doi.org/10.1002/atr.5670410107.

Russo, B. J., Savolainen, P. T., Schneider, W. H., IV, & Anastasopoulos, P. C. (2014). Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. *Analytic Methods in Accident Research*, 2, 21–29. https://doi.org/10.1016/j.amar.2014.03.001.

Sameen, M. I., & Pradhan, B. (2017). Severity Prediction of Traffic Accidents with Recurrent Neural Networks. *Applied Sciences*, 7(6), 476. https://doi.org/10.3390/app7060476.

Sarkar, S., Pramanik, A., Maiti, J., & Reniers, G. (2020). Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. *Safety Science*, 125, 104616. https://doi.org/10.1016/j.ssci.2020.104616.

Savolainen, P. T., Mannering, F. L., Lord, D., and Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention,* vol. 43, no. 5, pp. 1666_1676.

Savolainen, P., and Mannering, F. (2007). Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 955_963.

SAl-Ghamdi, A. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, vol. 34, no. 6, pp. 729_741.

Shafizadeh-Moghadam, H., Tayyebi, A., Ahmadlou, M., Delavar, M. R., and Hasanlou, M. (2017). Integration of Genetic Algorithm and Multiple Kernel Support Vector Regression for Modeling Urban Growth. *Computers, Environment and Urban Systems*, Vol. 65, pp. 28–40. https://doi.org/10.1016/j.compenvurbsys.2017.04.011.

Shankar, V., & Mannering, F. (1996). An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. *Journal of Safety Research,* 27(3), 183–194. https://doi. org/10.1016/0022-4375(96)00010-2.

Shankar, V., Mannering, F., & Barfield, W. (1996). Statistical Analysis of accident severity on rural freeways. *Accident Analysis & Prevention* , 2 8 (3), 391–401. https://doi. org/10.1016/0001-4575(96)00009-7.

Shao, X., Ma, X., Chen, F., Song, M., Pan, X., & You, K. (2020). A Random Parameters Ordered Probit Analysis of Injury Severity in Truck Involved Rear-End Collisions. *International Journal of Environmental Research and Public Health*, 17(2), 395https://doi. org/10.3390/ijerph17020395.

Siddiqui, N., Chu, X., and Guttenplan, M. (2006). Crossing locations, light conditions, and pedestrian injury severity. *Transportation Research Record, Journal Transportation Research Board*, vol. 1982, no. 1, pp. 141_149,.

Sobhani, A., Young, W., Logan, D., & Bahrololoom, S. (2011). A kinetic energy model of two-vehicle crash injury severity. *Accident Analysis & Prevention, 43*(3), 741–754. https://doi.org/10.1016/j.aap.2010.10.021.

Taamneh, M., Alkheder, S., and Taamneh, S. (2016). Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *Journal of Transportation Safety Security*, pp. 1–21.

Tang, J., Liang, J., Han, C., Li, Z., & Huang, H. (2019). Crash injury severity analysis using a two-layer Stacking framework. *Accident Analysis & Prevention*, vol. 122, 226-238.

Trevor, H., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: Data mining, inference and prediction. Mathematical Intelligencer, 27, 2, 83–85.

Ullah, I., Liu, K., Yamamoto, T., Zahid, M., & Jamal, A. (2021). Electric vehicle energy consumption prediction using stacked generalization: An ensemble learning approach. *International Journal of Green Energy*, 1–14.

UNESCAP, (2007). THE STATUS PAPER ON ROAD SAFETY PROBLEMS IN BANGLADESH: The Bangladesh Country Paper. *United Nations Economic and Social Commission for Asia and the Pacific*.

Vilaça, M., Macedo, E., & Coelho, M. C. (2019). A rare event modelling approach to assess injury severity risk of vulnerable road users. *MDPI, Safety*, 5(2), 29. https://doi.org/10.3390/safety5020029.

Wahab, L., and Jiang, H. (2019). A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PLoS ONE*, vol. 14, no. 4, Art. no. e0214966.

Wang, C., Xu, C., & Dai, Y. (2019). A crash prediction method based on bivariate extreme value theory and video-based vehicle trajectory data. *Accident Analysis and Prevention*, 123, 365–373. https:// doi.org/10.1016/j.aap.2018.12.013.

Wang, J., Huang, H., Xu, P., Xie, S., & Wong, S. C. (2020). Random parameter probit models to analyze pedestrian red-light violations and injury severity in pedestrian–motor vehicle crashes at signalized crossings. *Journal of Transportation Safety & Security*, 12(6), 818– 837. https://doi.org/10.1080/19439962.2018.1551257.

Wang, Q., Sun, H., & Zhang, Q. (2017). A Bayesian Network Model on the Public Bicycle Choice Behavior of Residents: A Case Study of Xi'an. *Mathematical Problems in Engineering*, Vol. 2017, https://doi.org/10.1155/2017/3023956.

Wang, X., & Abdel-Aty, M. (2006). Temporal and spatial analyses of rear-end crashes at signalized intersections. *Accident Analysis & Prevention*, 38(6), 1137–1150. https://doi.org/10.1016/j.aap.2006.04.022.

Wang, Z., Cao, B., Deng, W., Zhang, Z., Lu, J. J., and Chen, H. (2011). Safety evaluation of truck-related crashes at freeway diverge areas. *In Proceeding of 90th Annual Meeting Transportation Research Board,Washington, DC, USA*, pp. 1_15.

Washington, S., Karlaftis, M., and Mannering, F. (2010). Statistical and Econometric Methods for Transportation Data Analysis. *Boca Raton, FL, USA: CRC Press*.

Wen, X., Xie, Y., Jiang, L., Pu, Z., & Ge, T. (2021). Applications of Machine Learning Methods in Traffic Crash Severity Modelling: Current Status and Future Directions. *Transport Reviews*, Vol. 41, No. 6, pp. 855–879. https://doi.org/10.1080/01441647.2021.1954108.

World Health Organization (WHO), (2022). Road Traffic Injuries. https://www.who.int/news-room/fact sheets/detail/road-traffic-injuries. Accessed Aug. 1.

World Health Organization (WHO), (2018). Global Status Report on Road Safety 2018.

World Health Organization (WHO), (2018). World report on road traffic injury prevention, Geneva.

World Health Organization (WHO), (2009). Global status report on road safety: Time for action. Geneva: Author.

Xie, Y., Zhang, Y., and Liang, F. (2009). Crash injury severity analysis using Bayesian ordered probit models. *Journal of Transportation Engineering*, vol. 135, no. 1, pp. 18_25.

Xie, Y., Lord, D., and Zhang, Y. (2007). Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 922–933.

Xiong, Y., Tobias, J. L., & Mannering, F. L. (2014). The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity. *Transportation Research Part B: Methodological*, 67, 109–128. https://doi.org/10.1016/j. trb.2014.04.007.

Xu, P., and Huang, H. (2015). Modeling crash spatial heterogeneity: random parameter versus geographically weighting. *Accident Analysis & Prevention*, vol. 75, pp. 16_25.

Yassin, S. S., and Pooja (2020). Road Accident Prediction and Model Interpretation Using a Hybrid K-Means and Random Forest Algorithm Approach. *SN Applied Sciences*, Vol. 2, No. 9, pp. 1–13. https://doi.org/10.1007/S42452-020-3125-1/TABLES/5.

Yang, J., Ren, P., and Ando, R. (2019). Examining Drivers' Injury Severity of Two-Vehicle Crashes between Passenger Cars and Trucks Considering Vehicle Types. *Asian Transport Studies,* Vol. 5, Issue. 4, pp. 720-733,

Yau, K. K. W., Lo, H. P., and Fung, S. H. H. (2006). Multiple-vehicle traf_c accidents in Hong Kong. *Accident Analysis & Prevention*, vol. 38, no. 6, pp. 1157-1161.

Yuan, Q., Lu, M., Theofilatos, A., & Li, Y. B. (2017). Investigation on occupant injury severity in rear-end crashes involving trucks as the front vehicle in Beijing area, China. *Chinese Journal of Traumatology = Zhonghua Chuang Shang Za Zhi*, 20(1), 20–26. https://doi. org/10.1016/j.cjtee.2016.10.002 28162916.

Yuan, R., Gan, J., Peng, Z., and Xiang, Q. (2022). Injury Severity Analysis of Two-Vehicle Crashes at Unsignalized Intersections Using Mixed Logit Models. *International Journal of Injury Control and Safety Promotion*. https://doi.org/10.1080/17457300.2022.2040540.

Yu, R., and Abdel-Aty, M., (2014a). Using Hierarchical Bayesian Binary Probit Models to Analyze Crash Injury Severity on High Speed Facilities with Real-Time Traffic Data. *Accident Analysis & Prevention*, Vol. 62, 2014, pp. 161–167. https://doi.org/10.1016/J.AAP.2013.08.009.

Yu, R., and Abdel-Aty, M. (2014). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Safety Science*, vol. 63, pp. 50–56.

Zahid, M., Chen, Y., Jamal, A., Al-Ahmadi, H. M., & Khalaf Al-Ofi, A. (2020). Adopting machine learning and spatial analysis techniques for driver risk assessment: Insights from a case study. *International Journal of*

Environmental Research and Public Health, 17(14), 5193– 5115. https://doi.org/10.3390/ijerph17145193.

Zahid, M., Chen, Y., Jamal, A., & Memon, Q. M. (2020). Short Term traffic state prediction via hyperparameter optimization based classifiers. *MDPI, Sensors*, 20(3), 685. https://doi.org/10.3390/s20030685.

Zeng, Q., Wen, H., and Huang, H. (2016). The Interactive Effect on Injury Severity of Driver-Vehicle Units in Two-Vehicle Crashes. *Journal of Safety Research*, Vol. 59, pp. 105–111. https://doi.org/10.1016/J.JSR.2016.10.005.

Zeng, Q., Wen, H., Huang, H., Pei, X., & Wong, S. C. (2017). A multivariate random-parameters tobit model for analyzing highway crash rates by injury severity. *Accident Analysis and Prevention*, 99(Pt A), 184–191. https://doi.org/10.1016/j.aap.2016.11.018.

Zeng, Q., & Huang, H. (2014). A stable and optimized neural network model for crash injury severity prediction. *Accident Analysis and Prevention*, 73, 351–358. https://doi.org/10.1016/j. aap.2014.09.006.

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. https:// doi.org/10.1109/ACCESS.2018.2874979.

Zhang, S., Khattak, A., Matara, C. M., Hussain, A., and Farooq, A. (2022). Hybrid Feature Selection-Based Machine Learning Classification System for the Prediction of Injury Severity in Single and Multiple-Vehicle Accidents. *PLoS ONE*, Vol. 17, No. 2, https://doi.org/10.1371/journal.pone.0262941.

Zhang, Y., and Xie, Y. (2007). Forecasting of short-term freeway volume with v-support vector machines. *Transportation Research Record, Journal of Transportation Research Board*, vol. 2024, no. 1, pp. 92–99.

Zhou, Z. H. (2012). *Ensemble Methods: Foundations and Algorithms*. *Chapman and Hall/CRC, Boca Raton, FL, USA.*

APPENDIX-A

ACCIDENT REPORT FORM (ARF)

Form No. 34
Bengal Form No. 403Q

গণপ্রজাতন্ত্রী বাংলাদেশ সরকার
বাংলাদেশ পুলিশ
সড়ক দূর্ঘটনার রিপোর্ট ফরম
Regulation 254 (b)

1. দূর্ঘটনার ক্রমিক নম্বর

2. প্রাথমিক তথ্য বিবরণী নম্বর

3. থানা

4. জিলা / মেট্রো পুলিশ

5. দূর্ঘটনা কবলিত গাড়ীর সংখ্যা

6. হতাহত ড্রাইভারের সংখ্যা

7. হতাহত যাত্রীর সংখ্যা

8. হতাহত পথচারীর সংখ্যা

9. দূর্ঘটনার মাত্রা
F. মৃত্যু ঘটিত দূর্ঘটনা
G. মারাত্মক ক্ষত জনিত দূর্ঘটনা
S. সাধারন ক্ষত জনিত দূর্ঘটনা
M. মটর সংঘর্ষ

10. দিন [সোম, মঙ্গল, বুধ, বৃহস্পতি, শুক্র, শনি, রবি]
1 2 3 4 5 6 7

দূর্ঘটনার তারিখ 11. তারিখ 12.মাস 13.বছর
--------- / ---------- / -----------

14. দূর্ঘটনার সময়

রিপোর্ট করার তারিখ

রিপোর্ট করার সময়

15. সংযোগ স্থলের ধরণ
1. সংযোগ স্থল নয়
2.
3.
4.
5.
6. রেলওয়ে ক্রসিং
7. অন্যান্য ---------

16. ট্রাফিক নিয়ন্ত্রণ ব্যবস্থা
1. অনিয়ন্ত্রিত
2. রোড ডিভাইডার দ্বারা নিয়ন্ত্রিত
3. পথচারী পারাপার
4. পুলিশ নিয়ন্ত্রিত
5. ট্রাফিক বাতি নিয়ন্ত্রিত
6. পুলিশ ও ট্রাফিক বাতি নিয়ন্ত্রিত
7. থামা / যেতে দিন সংকেত দ্বারা নিয়ন্ত্রিত
8. অন্যান্য -------

17. সংঘর্ষের ধরন
1. মুখোমুখি
2. পশ্চাদ ভাগ
3. সমকোন
4. পার্শ্ব ঘর্ষণ
5. উল্টে যাওয়া
6. রাস্তার উপরস্থ বস্তুকে আঘাত
7. রাস্তার ধারের বস্তুকে আঘাত
8. দাঁড়ানো গাড়ীকে আঘাত
9. পথচারীকে আঘাত
10. জন্তুকে আঘাত
11. অন্যান্য ----(যেমন গাড়ীর ছাদ থেকে পড়ে যাওয়া)

18. গাড়ী চলাচলের দিক
1. এক মুখী চলাচল
2. উভয়মুখী চলাচল

19. রোড ডিভাইডার
1. আছে
2. নাই

20. আবহাওয়া	21. আলো	22. রাস্তার জ্যামিতিক বিবরণ	23. রাস্তার উপরিভাগের অবস্থা	24. রাস্তার প্রকারভেদ	25. রাস্তার প্রকৃতি	26. রাস্তার শ্রেনী
1. ভাল	1. দিন	1. সোজা - সমতল	1. শুকনা	1. পাকা	1. ভাল	1. ন্যাশনাল
2. বৃষ্টি	2. ভোর / সন্ধ্যা	2. বাঁকানো	2. ভেজা	2. ইটের রাস্তা	2. এবড়ো থেবড়ো (রাফ)	2. রিজিওনাল
3. ঝড়	3. আলোকিত সড়ক (রাতে)	3. অসমতল	3. কর্দমাক্ত	3. কাঁচা	3. মেরামত কাজ চলছে	3. ফিডার রোড
4. কুয়াশা	৪. অনালোকিত সড়ক (রাতে)	4. বাঁকা- অসমতল	4. জলমগ্ন /প্লাবিত			4. রুরাল রোড
		5. চূড়া	5. অন্যান্য			5. সিটি রোড

27. রাস্তার বৈশিষ্ট্য
1. সাধরণ রাস্তা
2. সেতু
3. কালভার্ট
4. সংকীর্ণ / বাধা প্রাপ্ত
5. স্পীড ব্রেকার

28. এলাকার ধরন
1. শহর
2. গ্রাম এলাকা

অফিসে ব্যবহারের জন্য

29. XY MAP
30. X
31. Y
32. ROUTE
33. KM
34. 100m
35. NODE MAP
36. NODE 1
37. NODE 2

অবস্থান
নগর/শহর/গ্রাম এর নাম...থেকে দুরত্ব :-------------------------------(কি: মি/মি)

রাস্তার নাম : .... মধ্যে
রোড / স্থান (১)..................................দুরত্ব : ------------------------------(কি: মি/মি)
রোড / স্থান (২)..................................দুরত্ব : ------------------------------(কি: মি/মি)

দ্বিতীয় রাস্তার নাম (শুধু মাত্র সংযোগ স্থানের দূর্ঘটনার ক্ষেত্রে) : ....থেকে দুরত্ব : ------------------------------(কি: মি/মি)

<u>দূর্ঘটনা স্থানের রেখা চিত্র :</u> দূর্ঘটনার স্থান থেকে নিকটবর্তী কিমি পোষ্ট,সেতু বা রাস্তার সংযোগ স্থান বা অন্যান্য যে কোন স্থায়ী গুরুত্বপূর্ন স্থাপনা হইতে দুরত্ব দেখাইয়া চিত্র

<u>সংঘর্ষের রেখা চিত্র :</u> দূর্ঘটনা কবলিত গাড়ী / পথচারী সমুহের চলাচলের দিক এবং অবস্থান সহ সংঘর্ষের পূর্ণ চিত্র

দূর্ঘটনার সংক্ষিপ্ত বিবরণ
....
....
....
....
....
....
....
....
....
....
....
....
....
....

সাক্ষী
১. নাম ও ঠিকানা --------------------------------
----
২. নাম ও ঠিকানা --------------------------------

বিবরণ লিপিবদ্ধকারী অফিসার
নাম/ পদবি ....তারিখ

অনুসন্ধানকারী অফিসার
নাম/ পদবি ....তারিখ

তত্ত্বাবধানকারী অফিসার
নাম/ পদবি.....................................তারিখ

আইনের ধারা

কেসের অবস্থা
1. চার্জশীট
2. ফাইনাল রিপোর্ট
3. তদন্তাধীন

দুর্ঘটনায় ২ এর অধিক যানবাহন, ৬ এর অধিক যাত্রী অথবা ৩ এর অধিক পথচারী হতাহত হইলে অতিরিক্ত ফরমের দরকার হইবে। অতিরিক্ত ফরমে দূর্ঘটনার ক্রমিক নম্বর থানা ও জেলা/মেট্রোপুলিশ এবং দুর্ঘটনার বৎসর উল্লেখ করিয়া এক সাথে গাথিয়া দিতে হইবে।

যানবাহন ১	মালিকের নাম	চালক ১	নাম
মালিকের ঠিকানা		ঠিকানা	
যানবাহন প্রস্তুতকারী	রেজিষ্ট্রেশন নম্বর 38. জেলা 39. নম্বর	ড্রাইভিং লাইসেন্স 46. জেলা 47. নম্বর	
40. বৈধ ফিটনেস সার্টিফিকেট 1. আছে 2. নাই 3. প্রযোজ্য নয়	বীমা কৃত 1. ৩য় পার্টি 2. কমপ্রিহেনসিভ	লাইসেন্সের ধরণ এবং যানের শ্রেণী	মেয়াদ অতিক্রান্তের তারিখ

41. যানবাহনের ধরণ
1. বাইসাইকেল
2. রিকসা / ভ্যান
3. ঠেলাগাড়ী
4. মটর সাইকেল
5. বেবী টেক্সী/CNG
6. টেম্পো
7. মাইক্রোবাস
8. মিনি বাস
9. বাস
10. কার
11. জীপ
12. পিক আপ
13. ছোট ট্রাক
14. ভারী ট্রাক
15. আর্টিকুলেটেড ট্রাক
16. ট্যাংকার
17. ট্রাক্টর
18. পশু চালিত
19. অন্যান্য ---- (নসিমন/ করিমন ইত্যাদি)

42. যানবাহন চলাচলের ধরণ
1. বামে মোড়
2. ডানে মোড়
3. 'U' মোড়
4. সংযোগ স্থলে আড়াআড়ি অতিক্রম
5. ওভারটেকিং
6. সোজাসুজি চলা
7. পিছনের দিকে চলা
8. হঠাৎ যাত্রা করা
9. হঠাৎ থামা / গতি কমানো
10. পার্ক/দাঁড়ানো অবস্থায়
11. অন্যান্য ------------

48. চালকের লিঙ্গ
1. পুরুষ
2. স্ত্রী

49. চালকের বয়স

50. চালকের ক্ষত
F. মৃত্যু
G. মারাত্মক ক্ষত
S. সাধারণ ক্ষত
N. অক্ষত

43. যানবাহনের মালামাল বোঝাই
1. আইনানুগ
2. বেআইনী/ বিপজ্জনক বোঝাই

44. যানবাহনের ত্রুটি
1. ত্রুটি মুক্ত
2. লাইট
3. ব্রেক
4. ষ্টিয়ারিং
5. টায়ার
6. বহুবিধ
7. অন্যান্য ---------

45. যানবাহনের ক্ষতি (দুর্ঘটনা জনিত)
1. নাই
2. সামনে
3. পিছনে
4. ডানে
5. বামে
6. ছাদে
7. বহুবিধ
8. অন্যান্য----------

51. মদ্যপ কিনা
1. সন্দেহ আছে
2. সন্দেহ মুক্ত

52. সীট বেল্ট / হেলমেট
1. পরিহিত
2. পরিহিত নয়

যানবাহন ২	মালিকের নাম	চালক ২	নাম
মালিকের ঠিকানা		ঠিকানা	
যানবাহন প্রস্তুতকারী	রেজিষ্ট্রেশন নম্বর 38. জেলা 39. নম্বর	ড্রাইভিং লাইসেন্স 46. জেলা 47. নম্বর	
40. বৈধ ফিটনেস সার্টিফিকেট 1. আছে 2. নাই 3. প্রযোজ্য নয়	বীমা কৃত 1. ৩য় পার্টি 2. কমপ্রিহেনসিভ	লাইসেন্সের ধরণ এবং যানের শ্রেণী	মেয়াদ অতিক্রান্তের তারিখ

41. যানবাহনের ধরণ
1. বাইসাইকেল
2. রিকসা / ভ্যান
3. ঠেলাগাড়ী
4. মটর সাইকেল
5. বেবী টেক্সী/CNG
6. টেম্পো
7. মাইক্রোবাস
8. মিনি বাস
9. বাস
10. কার
11. জীপ
12. পিক আপ
13. ছোট ট্রাক
14. ভারী ট্রাক
15. আর্টিকুলেটেড ট্রাক
16. ট্যাংকার
17. ট্রাক্টর
18. পশু চালিত
19. অন্যান্য ---- (নসিমন/ করিমন ইত্যাদি)

42. যানবাহন চলাচলের ধরণ
1. বামে মোড়
2. ডানে মোড়
3. 'U' মোড়
4. সংযোগ স্থলে আড়াআড়ি অতিক্রম
5. ওভারটেকিং
6. সোজাসুজি চলা
7. পিছনের দিকে চলা
8. হঠাৎ যাত্রা করা
9. হঠাৎ থামা / গতি কমানো
10. পার্ক/দাঁড়ানো অবস্থায়
11. অন্যান্য ------------

48. চালকের লিঙ্গ
1. পুরুষ
2. স্ত্রী

49. চালকের বয়স

50. চালকের ক্ষত
F. মৃত্যু
G. মারাত্মক ক্ষত
S. সাধারণ ক্ষত
N. অক্ষত

43. যানবাহনের মালামাল বোঝাই
1. আইনানুগ
2. বেআইনী/ বিপজ্জনক বোঝাই

44. যানবাহনের ত্রুটি
1. ত্রুটি মুক্ত
2. লাইট
3. ব্রেক
4. ষ্টিয়ারিং
5. টায়ার
6. বহুবিধ
7. অন্যান্য ---------

45. যানবাহনের ক্ষতি (দুর্ঘটনা জনিত)
1. নাই
2. সামনে
3. পিছনে
4. ডানে
5. বামে
6. ছাদে
7. বহুবিধ
8. অন্যান্য----------

51. মদ্যপ কিনা
1. সন্দেহ আছে
2. সন্দেহ মুক্ত

52. সীট বেল্ট / হেলমেট
1. পরিহিত
2. পরিহিত নয়

হতাহত যাত্রীর বিবরণ — একজন যাত্রীর জন্য একটি লাইন পুরণ করুন — * = নীচের বক্স দেখুন

নাম ও ঠিকানা	53. যানবাহন নং	54. লিঙ্গ	55. বয়স	56. * ক্ষত	57. * অবস্থান	58. * কার্যক্রম
1.						
2.						
3.						
4						
5.						
6.						

হতাহত পথচারীর বিবরণ — একজন পথচারীর জন্য একটি লাইন পুরণ করুন — * = নীচের বক্স দেখুন

নাম ও ঠিকানা	59. যানবাহন নং	60. লিঙ্গ	61. বয়স	62. * ক্ষত	63. * অবস্থান	64. * কার্যক্রম
1.						
2.						
3.						

দুর্ঘটনার সহায়ক কারণ

১. মাত্রাতিরিক্ত গতি
২. বেপরোয়া চালান
৩. চালকের ক্লান্তি
৪. সামনের গাড়ির অতি সন্নিকটে চালান
৫. চালকের ভুল সংকেত
৬. ভুল ওভারটেকিং
৭. ভুল ভাবে মোড় নেয়া
৮. মদ্যপ চালক
৯. পথচারীর কার্যক্রম
10. যাত্রীর কার্যক্রম
11. খারাপ রাস্তার জন্য
12. রাস্তার জ্যামিতিক ত্রুটি
13. আবহাওয়া
14. গাড়ীর যান্ত্রিক ত্রুটি
15. বিপজ্জনক বোঝাই
16. টায়ার বার্ষ্ট
17. পশুর কার্যক্রম
18. অন্যান্য ---------- (যেমনঃ রাস্তার উপর দালাদার/ পিচ্ছিল জিনিস পড়ে থাকা, গতি রোধক, দুর্বল ব্রিজ / কালভার্ট ইত্যাদির কারণে)

65. ☐☐
66. ☐☐
67. ☐☐

*** 56--58 এবং 62--64 এর সহায়ক বক্স**

শুধুমাত্র নমুনার জন্য বৃত্ত দিবেন না	56. যাত্রীর ক্ষত 62. পথচারীর ক্ষত	57. যাত্রীর অবস্থান	58. যাত্রীর কার্যক্রম	63. পথচারীর অবস্থান	64. পথচারীর কার্যক্রম
	F. মৃত্যু G. মারাত্মক ক্ষত S. সাধারণ ক্ষত	1. গাড়ীর ভিতরে 2. গাড়ীর বাইরে 3. গাড়ীর ছাদে	1. নাই 2. যানে উঠিতেছিল 3. যান হইতে নামিতেছিল 4. যান হইতে পড়িয়া যাওয়া 5. অন্যান্য	1. পথচারি পারাপারে 2. পারাপারের ৫০ মিঃ মধ্যে 3.সড়ক দ্বীপ / ডিভাইডারে 4.রাস্তার উপরে 5.ফুটপাতে 6.রাস্তার পাশে/ সোল্ডারে 7.বাস ষ্টপে	1. নাই 2. রাস্তা পারাপার হওয়া 3. রাস্তার উপর দিয়ে চলা 4. রাস্তার পাশ/ সোল্ডার দিয়ে চলা 5. রাস্তার উপরে খেলা করা

APPENDIX-B

GUIDELINES FOR FILLING ROAD ACCIDENT REPORT FORM

SECOND EDITION, JANUARY 2010

সড়ক দুর্ঘটনার রিপোর্ট ফরম পুরণের নির্দেশিকা

দ্বিতীয় সংস্করণ, জানুয়ারী ২০১০

ARI

দুর্ঘটনা রিসার্চ ইন্সটিটিউট (ARI)

বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় (BUET)
ঢাকা-১০০০

ভূমিকা

বাংলাদেশ পুলিশের নতুন সড়ক দুর্ঘটনার রিপোর্ট ফরম যথাযথভাবে পুরণের সুবিধার্থে এই নির্দেশিকা প্রকাশ করা হলো। এই পুস্তিকার শেষে একটি পূরণকৃত দুর্ঘটনার রিপোর্ট ফরম দেয়া হলো।

সড়ক দুর্ঘটনার রিপোর্ট ফরমটিতে দুই পৃষ্ঠায় তথ্য লিখার জন্য সর্বমোট ৬৭টি ঘর আছে। এই ঘরসমূহ পুরণের সময় প্রায় ক্ষেত্রেই প্রযোজ্য উত্তরে শুধু গোলদাগ দিতে হবে। অনুসন্ধানকারী অফিসার (Investigating Officer) ফরমটির সম্পূর্ণ অংশ পড়ে, প্রতিটি ঘর ক্রমানুযায়ী যথাযথভাবে পূরণ করবে।

থানা থেকে পূরণকৃত ফরমের অনুলিপি রির্পোটকারী থানায় সংরক্ষণ করতে হবে। মূল ফরমটি পুলিশ সুপার অফিসে পাঠাতে হবে। পুলিশ সুপারগণ পূরণকৃত ফরমসমূহ সংশ্লিষ্ট রেঞ্জ এর এক্সিডেন্ট ডাটা ইউনিটে (ADU) পাঠাবেন। মেট্রোপলিটন এলাকায় থানা থেকে পূরণকৃত ফরম সরাসরি মেট্রোপলিটন পুলিশ কমিশনারের অফিসে অবস্থিত এক্সিডেন্ট ডাটা ইউনিটে (ADU) পাঠাবেন। প্রত্যেক ডিআইজি/মেট্রোপলিটন পুলিশ কমিশনারের অফিসে অবস্থিত এক্সিডেন্ট ডাটা ইউনিট (ADU) দুর্ঘটনার ফরমগুলো থেকে ডাটা MAAP5 Software-এর মাধ্যমে কম্পিউটারে এন্ট্রি করবে। ডিআইজি/মেট্রোপলিটন পুলিশ কমিশনারের দপ্তর থেকে এন্ট্রিকৃত Database CD/Pendrive/E-mail-এর মাধ্যমে ঢাকাস্থ পুলিশ সদর দপ্তরে পাঠাবে। পুলিশ সদর দপ্তর হতে এন্ট্রিকৃত Database CD/Pendrive-এর মাধ্যমে রোড সেফটি সেলে পাঠাতে হবে। রোড সেফটি সেল, জাতীয় সড়ক নিরাপত্তা কাউন্সিলের দায়িত্ব পালনের অংশ হিসাবে তথ্যগুলো সংগ্রহ, বিশ্লেষণ এবং বার্ষিক রিপোর্ট তৈরী করে থাকে এবং এরপর তা' বিভিন্ন সংস্থায় পাঠানো হয়। সড়ক দুর্ঘটনার তথ্য সরকারের নীতি নির্ধারণসহ বিভিন্ন সংস্থার প্রয়োজনে এবং সড়ক দুর্ঘটনা রোধ করার লক্ষ্যে বিভিন্ন গবেষণা প্রতিষ্ঠানের প্রয়োজনে সরবরাহ করা হয়।

অসম্পূর্ণ ও ভুলভাবে পূরণকৃত ফরম সম্পূর্ণ ও শুদ্ধভাবে পূরণ করার জন্য সংশ্লিষ্ট থানায়/অনুসন্ধানকারী কর্মকর্তার নিকট ফেরত পাঠাতে হবে। রিপোর্টকারী থানা তদন্ত নথির জন্য আরও বিস্তারিত মানচিত্র, মৃত্যুর পরবর্তী রিপোর্ট, গাড়ীর পরিদর্শণ রিপোর্ট ইত্যাদির প্রয়োজন হ'তে পারে, তবে এগুলি রিপোর্টকারী থানায় রেখে দিতে হবে।

ফরমটির যেসব ঘরের প্রথমে নম্বর যুক্ত আছে (১ হইতে ৬৭ পর্যন্ত) ঐগুলি কম্পিউটারে সংরক্ষিত হবে। তা' ছাড়াও দুর্ঘটনার লিখিত বিবরণ ও দুর্ঘটনার স্থান কম্পিউটারে সংরক্ষিত থাকবে।

এই রিপোর্ট ফরমটি অনুসন্ধানকারী অফিসার কর্তৃক দুর্ঘটনার স্থানেই অথবা যত তাড়াতাড়ি সম্ভব পূরণ করতে হবে।

(ডঃ মোঃ সামছুল হক)	**(কিউ.এ.এস.এম.জাকারিয়া ইসলাম)**
পরিচালক	ডাটা বেইজ স্পেশালিষ্ট
দুর্ঘটনা রিসার্চ ইন্সটিটিউট ও	দুর্ঘটনা রিসার্চ ইন্সটিটিউট
অধ্যাপক, পুরকৌশল বিভাগ	বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়
বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়	

তারিখ : জানুয়ারী ২০১০

বি.দ্র. যেসব মেট্রোপলিটন এলাকায় এক্সিডেন্ট ডাটা ইউনিট স্থাপিত হয়নি, সেসব এলাকায় রেঞ্জ অফিসে স্থাপিত ইউনিটই ডাটা এন্ট্রির কাজ করবে।

সড়ক দুর্ঘটনার রিপোর্ট ফরম পূরণ করার পদ্ধতি

(১) দুর্ঘটনার বিস্তারিত বিবরণ ঃ

1| দুর্ঘটনার রিপোর্ট নম্বর ঃ দুর্ঘটনার রিপোর্ট নম্বর রিপোর্টকারী থানা বা আঞ্চলিক হেড কোয়ার্টার কর্তৃক দেয় রিপোর্টের ক্রমিক নম্বর। প্রত্যেক থানা বা আঞ্চলিক অফিস প্রতি বৎসর ০০০১ হতে শুরু করে এই ক্রমিক নম্বর দিবে। প্রতিটি থানা একটি করে সড়ক দুর্ঘটনার হিসাব বই রাখবে যাতে দুর্ঘটনার সময়ানুক্রম পাওয়া যায় ও হেড কোয়ার্টারে ফেরত না দেয়া রিপোর্টের হদিস পাওয়া যায়। এই প্রশিক্ষণ ম্যানুয়েলের শেষে একটি হিসাব বই-এর নমুনা দেয়া হলো। এই দুর্ঘটনার রিপোর্ট নম্বরের সাথে এফ.আই.আর বা এম.সি.আর নম্বর গুলিয়ে ফেলা যাবে না।

2| প্রাথমিক তথ্য বিবরণী নম্বর ঃ থানা কর্তৃক কেস প্রতি দেয়া প্রাথমিক তথ্য বিবরণী (FIR) নম্বর।

3| থানা ঃ দুর্ঘটনার রিপোর্টকারী থানা/পুলিশ ষ্টেশন সমূহের নামের তালিকা প্রতিটি জেলা ও মেট্রোপলিটন পুলিশ বাহিনীতে রক্ষিত আছে।

4| জেলা/মেট্রোপলিটন ঃ পুলিশ জেলা বা মেট্রোপলিটন পুলিশ বাহিনীর নাম।

5| দুর্ঘটনা কবলিত গাড়ির সংখ্যা ঃ দুর্ঘটনা কবলিত সর্বমোট গাড়ির সংখ্যা। এর প্রতিটি গাড়ির জন্য অত্র ফরমের সম্পৃক্ত যানবাহন/চালক অংশ পূরণ করতে হবে।

6| হতাহত চালকের সংখ্যা ঃ দুর্ঘটনায় নিহত বা আহত চালকের মোট সংখ্যা।

7| হতাহত যাত্রীর সংখ্যা ঃ দুর্ঘটনায় নিহত বা আহত যাত্রীর মোট সংখ্যা। এর প্রতি যাত্রীর জন্য অত্র ফরমের সম্পৃক্ত যাত্রীর লাইন/অংশ পূরণ করতে হবে।

8| হতাহত পথচারীর সংখ্যা : দুর্ঘটনায় নিহত বা আহত পথচারীর মোট সংখ্যা। এর প্রতি পথচারীর জন্য অত্র ফরমের সম্পৃক্ত যাত্রীর লাইন/অংশ পূরণ করতে হবে।

9| দুর্ঘটনার মাত্রা : F = মৃত্যুঘটিত দুর্ঘটনা। যেখানে দুর্ঘটনার ৩০ দিনের মধ্যে কোন ব্যক্তি মৃত্যুবরণ করে। G = মারাত্মক ক্ষতজনিত দুর্ঘটনা। যেখানে দুর্ঘটনায় কোন ব্যক্তি মারাত্মকভাবে আহত হয়, তবে কেউ মৃত্যুবরণ করে না। S = সাধারণ ক্ষতজনিত দুর্ঘটনা। যেখানে কোন ব্যক্তি সাধারণভাবে আহত হয়। তবে কেউ মৃত বা মারাত্মকভাবে আহত হয় না। M = মোটর দুর্ঘটনা। যেখানে দুর্ঘটনায় কেউ হতাহত হয় না, কিন্তু গাড়ি বা সম্পদের ক্ষতি সাধিত হয়।

দুর্ঘটনার মাত্রা হতাহতের সংখ্যার উপর নির্ভর করে না বরং হতাহতদের মধ্যে সর্বোচ্চ আঘাতের মাত্রার উপর নির্ভরশীল। যেমন, কোন দুর্ঘটনায় যদি ২০ জন লোক সাধারণভাবে আহত (S) হয় ও ১ জন মারাত্মকভাবে আহত (G) হয় তবে দুর্ঘটনার মাত্রা মারাত্মক ক্ষতজনিত দুর্ঘটনা ধরতে হবে।

10| দিন : সপ্তাহের যে দিন/বারে (সোম,মঙ্গল,বুধ -----) দুর্ঘটনা সংঘটিত হয়।

দুর্ঘটনার তারিখ :

11| তারিখ : মাসের যে তারিখে দুর্ঘটনা সংঘটিত হয়।

12| মাস : যে মাসে দুর্ঘটনা সংঘটিত হয়।

13| বৎসর : যে বৎসর দুর্ঘটনা সংঘটিত হয়।

14| দুর্ঘটনার সময় : দুর্ঘটনা যে সময় সংঘটিত হয়। ২৪ ঘন্টার দিনকে ব্যবহার করতে হবে। উদাহরণ স্বরূপঃ সকাল ৯টা = ০৯.০০, রাত্র ৯টা = ২১.০০। তবে **এ পদ্ধতিতে যদি কোন দুর্ঘটনা রাত ঠিক ১২:০০টায় সংঘটিত হয় তবে**

দুর্ঘটনার সময় ০০:০০ বা ২৪:০০ না লিখে ০০:০১ লিখতে হবে।

রিপোর্ট করার তারিখ ঃ পুলিশের নিকট দুর্ঘটনার রিপোর্ট (FIR) করার দিন, মাস ও বৎসর।

রিপোর্ট করার সময় ঃ পুলিশের নিকট দুর্ঘটনার রিপোর্ট করার সময়।

15| সংযোগ স্থলের ধরণ ঃ দুর্ঘটনার স্থানের ধরণ বুঝে যথাযথ নম্বরে গোল দাগ দিতে হবে। যদি দুর্ঘটনাটি কোন রাস্তার সংযোগস্থলে সংঘটিত হয় তবে এই ফরমের দুর্ঘটনার অবস্থান অংশে দ্বিতীয় সড়কের নাম লিখতে হবে। এছাড়া সংঘর্ষের রেখা চিত্রের ঘরে রাস্তার সংযোগস্থলের যে রেখাচিত্র আঁকা হবে তা এই ঘরের রাস্তার সংযোগ স্থলের ধরনের সাথে অবশ্যই মিল থাকতে হবে।

উল্লেখ্য, দুর্ঘটনাটি সংযোগ স্থলের ২০ মিটারের মধ্যে সংঘটিত হয়ে থাকলে তা' সংযোগ স্থলে হয়েছে ধরে চিহ্নিত করতে হবে।

16| ট্রাফিক নিয়ন্ত্রণ ব্যবস্থা ঃ দুর্ঘটনার স্থানে অবস্থিত যানবাহন নিয়ন্ত্রণ ব্যবস্থার সাথে মিল রেখে যথাযথ ঘরে গোল দাগ দিতে হবে।

17| সংঘর্ষের ধরণ ঃ দুর্ঘটনার সংঘর্ষের ধরণ বুঝে যথাযথ চিহ্নে গোল দাগ দিতে হবে। সংঘর্ষের রেখাচিত্রে এই ঘরের সংঘর্ষের ধরনের দাগের সাথে মিল থাকতে হবে। এটা মনে রাখতে হবে যে, মুখোমুখি, পশ্চাদভাগে, সমকোন ও পার্শ্ব ঘর্ষণ জাতীয় সংঘর্ষের জন্য অন্ততঃ দুটি গাড়ি জড়িত থাকবে। একটি মাত্র গাড়ি কোন বস্তু বা পথচারীকে আঘাত করলে অথবা রাস্তার উপর উল্টে গেলে বা পাশে খাদে পড়ে গেলে এ চারটি ধরণ ব্যবহৃত হবে না।

মুখোমুখি ঃ যখন দু'টি গাড়ি মুখোমুখি সংঘর্ষে নিপতিত হয়।

পশ্চাদভাগঃ যখন একটি গাড়ি আরেকটি গাড়ির পশ্চাদভাগে আঘাত করে।

সমকোন ঃ যখন একটি গাড়ি অন্য গাড়ির পার্শ্বে প্রায় ৯০ ডিগ্রী কোণাকুনি আঘাত করে।

পার্শ্ব ঘর্ষণ ঃ যখন দুটি গাড়ি পরস্পরের পার্শ্ব ঘর্ষণে লিপ্ত হয়। গাড়ি দুটি একই দিকে বা বিপরীত দিকে গতিশীল থাকতে হবে।

18| গাড়ী চলাচলের দিক ঃ দুর্ঘটনাস্থলের রাস্তায় গাড়ি চলাচলের দিক নির্দেশের যথাযথ ঘরে গোল দাগ দিতে হবে।

একমুখি রাস্তা ঃ যখন রাস্তায় গাড়ি শুধু একদিকে চলাচল করে।

উভয়মুখি রাস্তা ঃ যখন রাস্তায় গাড়ি শুধু উভয়দিকেই চলাচল করে।

19| রোড ডিভাইডার ঃ দুর্ঘটনাস্থলের রাস্তার অবস্থা দেখে যথাযথ ঘরে গোল দাগ দিতে হবে।

আছেঃ রাস্তার মাঝ বরাবর কম উচ্চতার দেয়াল (সড়ক দ্বীপ) থাকলে এবং গাড়ি বিপরীত দিকে যেতে না পারলে।

নাই ঃ উপরের অবস্থার বিপরীত।

20| আবহাওয়া ঃ দুর্ঘটনার সময় আবহাওয়ার অবস্থা দেখে যথাযথ ঘরে গোল দাগ দিতে হবে।

21| আলো ঃ দুর্ঘটনার সময় আলোর অবস্থা দেখে যথাযথ ঘরে গোল দাগ দিতে হবে।

22| রাস্তার জ্যামিতিক বিবরণ ঃ দুর্ঘটনার সময় রাস্তার বাস্তব অবস্থা দেখে যথাযথ ঘরে গোল দাগ দিতে হবে।

চূড়া ঃ এটা পাহাড়ের সর্বোচ্চ অবস্থানকে বোঝায় যেখানে উভয় দিক থেকে আগত গাড়িগুলির দৃষ্টিসীমা কমে যায় অর্থাৎ ড্রাইভার সামনে বেশি দুর দেখতে পায়না।

23। রাস্তার উপরিভাগের অবস্থা ঃ দুর্ঘটনাস্থলের রাস্তার উপরিভাগের অবস্থা দেখে যথাযথ ঘরে গোলদাগ দিতে হবে।

24। রাস্তার প্রকারভেদ ঃ দুর্ঘটনা স্থলের রাস্তার উপরিভাগের প্রকারভেদ দেখে যথাযথ ঘরে গোলদাগ দিতে হবে।

25। রাস্তার প্রকৃতি ঃ দুর্ঘটনা স্থলের রাস্তার গুণাগুণ বিচার করে যথাযথ ঘরে গোলদাগ দিতে হবে।

26। রাস্তার শ্রেণী ঃ দুর্ঘটনা স্থলের রাস্তার শ্রেণী বিন্যাস নির্দেশক ঘরে গোলদাগ দিতে হবে। গুরুত্ব নির্বিশেষে প্রধান প্রধান শহরের সকল রাস্তাকে সিটি রোড হিসাবে দেখাতে হবে।

27। রাস্তার বৈশিষ্ট্য ঃ দুর্ঘটনাস্থলের রাস্তার বিশেষ বৈশিষ্ট্য নির্দেশক ঘরে গোলদাগ দিতে হবে।

সাধারণ রাস্তা ঃ যাতে বিশেষ কোন বৈশিষ্ট্য নেই।

সেতু ঃ দুর্ঘটনাটি যদি সেতুর উপর অথবা তার ২০ মিটারের মধ্যে সংঘটিত হয়ে থাকে তবে এই ঘরে গোলদাগ দিতে হবে। দাগের উপর সেতুর / নদীর নাম লিখতে হবে।

কালভার্ট ঃ দুর্ঘটনাটি যদি কোন কালভার্টের উপর অথবা কালভার্টের কারণে হয়ে থাকে তবে এই ঘরে গোলদাগ দিতে হবে।

সংকীর্ণ/বাধাপ্রাপ্ত ঃ দুর্ঘটনা স্থলে যদি কোন অস্থায়ী কারনের (যেমন হাট বাজার/গাড়ী থামানো/রাস্তা মেরামত কাজ ইত্যাদি) জন্য রাস্তা সংকীর্ণ হয়ে গাড়ী চলাচলে বাধাগ্রস্থ হয় তবে এই ঘরে গোলদাগ দিতে হবে।

28। এলাকার ধরণ ঃ দুর্ঘটনাস্থলের ধরণ বিবেচনা করে যথাযথ ঘরে গোল দাগ দিতে হবে।

শহর এলাকা ঃ যেখানে দুর্ঘটনাটি শহর বা নগরের মত বসতিপূর্ণ এলাকায় সংঘটিত হয়ে থাকে। যদি জায়গাটি শহরের সীমানার বাইরেও হয় তবুও বর্ণনাকারী অফিসার তা' শহর এলাকা বিবেচনা করতে পারেন যদি রাস্তার পার্শ্বে জনবসতি থাকে।

গ্রাম এলাকা ঃ যেখানে দুর্ঘটনাটি বসতিপূর্ণ এলাকার বাইরে সংঘটিত হয়ে থাকে। এর মধ্যে রাস্তাটি বন, আবাদী জমি বা ছোট গ্রামের মধ্য দিয়ে যেতে পারে।

(২) দুর্ঘটনার অবস্থানের তথ্য ঃ

দুর্ঘটনার উপযুক্ত অনুসন্ধান করতে হলে দুর্ঘটনা স্থলের অবস্থান-বৈশিষ্ট্য লিখতে হবে। এটা খুবই প্রয়োজনীয়, এই অংশে দুর্ঘটনাস্থলের বিস্তারিত তথ্যটি লিপিবদ্ধ করবেন, যাতে ভবিষ্যতে যে কেউ ঘটনাস্থল খুঁজে বের করতে পারেন। শুধুমাত্র অফিস ব্যবহারের জন্য ৯টি ঘর আছে। এগুলো কম্পিউটারে বিশ্লেষণের জন্য দুর্ঘটনার অবস্থানের বৈশিষ্ট্যসমূহ কোডভুক্ত করা হবে। এই ঘরগুলো পূরণ করা এই অংশের বিস্তারিত তথ্যাদির উপর নির্ভরশীল। অনেক জায়গায় কোন রাস্তা বা বস্তু বা বসতি থেকে দূরত্ব লিখতে হয়। এই দূরত্ব কিলোমিটার বা মিটারে লিখতে হবে। দূরত্ব লিখতে অপ্রয়োজনীয় কিঃ মিঃ অথবা মিঃ কেটে (অর্থাৎ প্রয়োজনীয় মিঃ অথবা কিঃ মিঃ রেখে) লিখতে হবে।

নগর/শহর/গ্রামের নাম ঃ এই ঘরে দুর্ঘটনা স্থলের নগর, শহর বা গ্রামের নাম লিখতে হবে। বসতি কেন্দ্র থেকে এর দূরত্ব লিখতে হবে। দুরত্ব শূণ্য হতে পারে, তখন ঐ ঘরে শূণ্য (০) লিখতে হবে। যদি দুর্ঘটনাস্থল বসতি থেকে অনেক দূর হয়, তা'হলে সবচেয়ে কাছের নগর/শহর/গ্রমের নাম লিখতে হবে। এই বসতি থেকে দূরত্ব ফরমের ঘরে লিখতে হবে।

দুর্ঘটনার অবস্থান ঃ

রাস্তার নাম ঃ এখানে দুর্ঘটনাস্থলের রাস্তার নাম লিখতে হবে। ন্যাশনাল রাস্তা হলে দুই প্রান্তের নগর/শহরের নামসহ একটি আদর্শ নাম

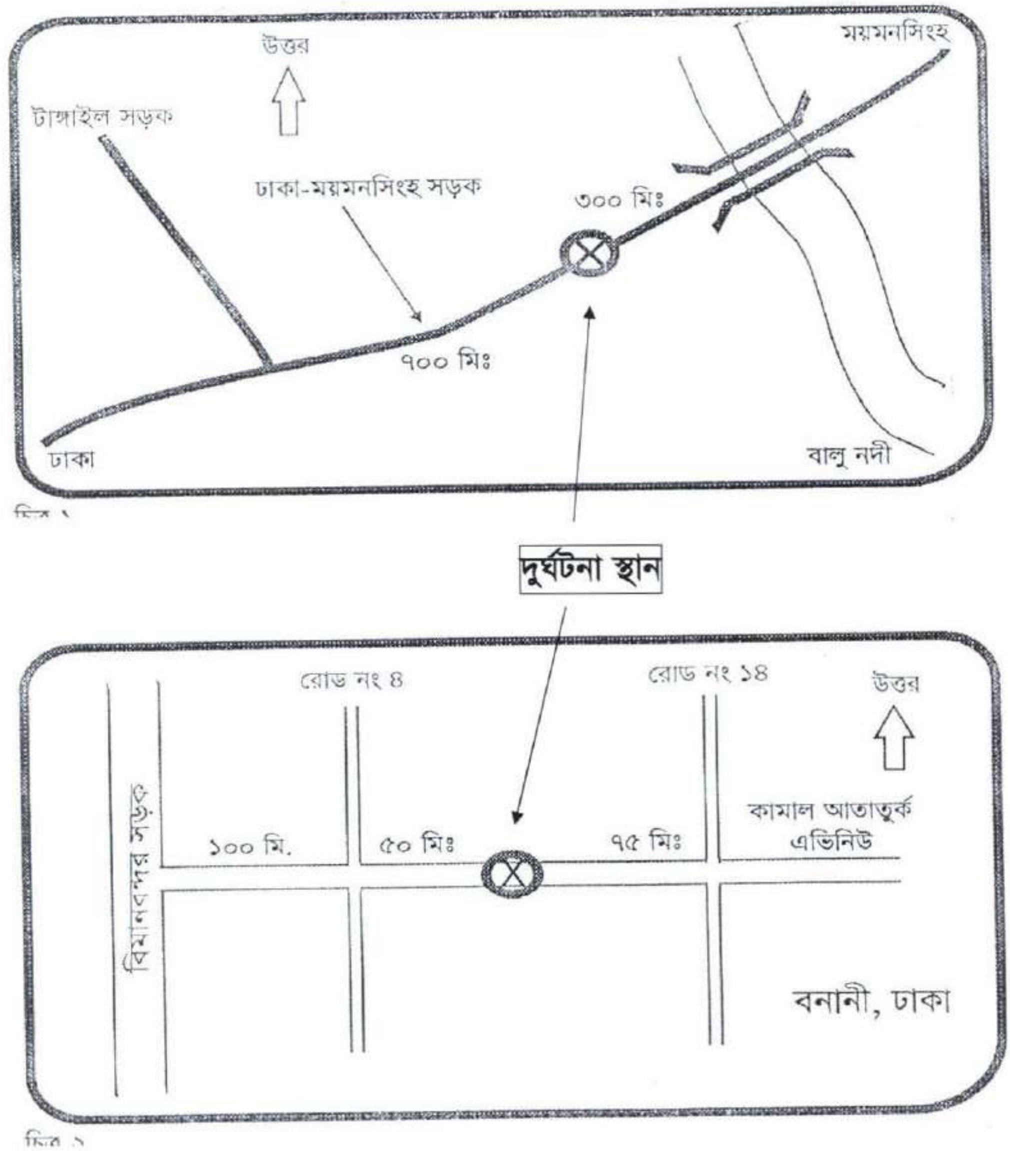
পদ্ধতি ব্যবহার করতে হবে অথবা সড়ক ও জনপথ দপ্তর কর্তৃক ব্যবহৃত সড়ক নম্বর ব্যবহার করতে হবে।

দৃষ্ট বস্তু-১ ঃ এখানে দুর্ঘটনাস্থলের রাস্তার উপর কোন লক্ষণীয় বস্তু / স্থায়ী স্থাপনা যেমন- কিলোমিটার পোষ্ট, সেতু, স্কুল, মাদ্রাসা, মসজিদ, রাস্তার সংযোগ স্থল ইত্যাদির নাম লিখতে হবে। এই লক্ষণীয় বস্তুর/স্থাপনার অবস্থানের দূরত্ব ফরমে জায়গামত লিখতে হবে।

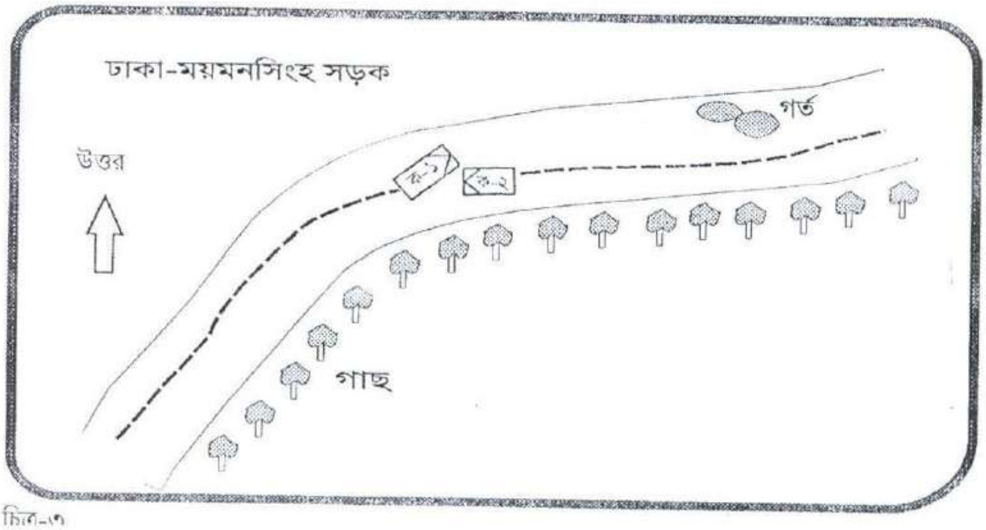
দৃষ্ট বস্তু-২ ঃ এখানে দৃষ্ট বস্তু-১ এর বিপরীত দিকের রাস্তায় অবস্থিত কোন লক্ষণীয় বস্তু / স্থায়ী স্থাপনা যেমন- কিলোমিটার পোষ্ট, সেতু, স্কুল, মাদ্রাসা, মসজিদ, রাস্তার সংযোগ স্থল ইত্যাদির নাম লিখতে হবে। দুর্ঘটনার স্থান থেকে ঐ লক্ষণীয় বস্তুর দূরত্ব ফরমে জায়গামত লিখতে হবে।

শুধুমাত্র সংযোগ স্থানের দুর্ঘটনা ঃ রাস্তার সংযোগ স্থলের দুর্ঘটনার ক্ষেত্রে দুইটি রাস্তারই নাম লিখতে হবে। দুর্ঘটনার স্থান থেকে এই সংযোগ স্থলের দূরত্ব ফরমে জায়গামত লিখতে হবে। দুর্ঘটনাটি যদি এই রাস্তা দুইটির ঠিক সংযোগ স্থলে হয়ে থাকে তবে দূরত্ব শূণ্য লিখতে হবে।

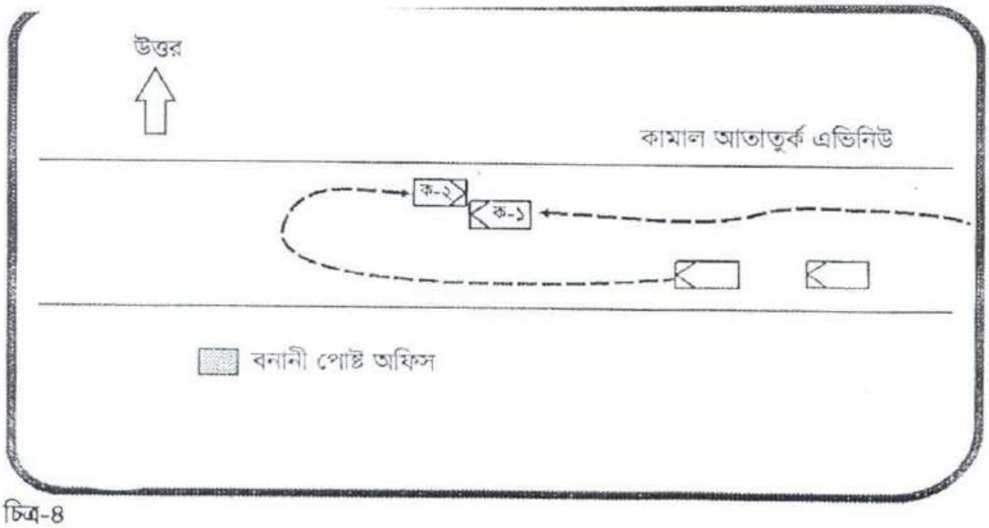
দুর্ঘটনাস্থলের রেখা চিত্র ঃ এই চিত্র অত্যন্ত দরকারী, যাতে ভবিষ্যতে যে কেউই চিত্র দেখে দুর্ঘটনার স্থানটি চিহ্নিত করতে পারে। এখানে শুধুমাত্র রাস্তাটির (বা রাস্তাগুলোর) রেখা চিত্র আঁকলেই চলবে এবং আশে-পাশের দৃষ্ট স্থাপনা সমূহ থেকে দুর্ঘটনার স্থানটির দূরত্ব দেখাতে হবে। মনে রাখতে হবে যে, এই রেখা চিত্রটি শুধুমাত্র দুর্ঘটনাস্থলের অবস্থান জানতে ব্যবহৃত হবে, কাজেই এতে দুর্ঘটনার ধরণের খুঁটিনাটি দেখানোর প্রয়োজন নেই। সংঘর্ষের ধরণের বিবরণ পরে বর্ণিত সংঘর্ষের রেখাচিত্রে দিতে হবে। নিম্নে দুইটি দুর্ঘটনাস্থলের রেখা চিত্রের নমুনা দেয়া হলো।



সংঘর্ষের রেখা চিত্র ঃ এই রেখাচিত্রটি দুর্ঘটনা তদন্তকারীদের জন্য খুবই গুরুত্বপূর্ণ এবং বহু বৎসর পরও তালিকাভুক্ত দুর্ঘটনা-প্রবণ স্থানসমূহের বিশ্লেষণের জন্য প্রয়োজন হয়। এটা একটি সংঘর্ষের রেখা চিত্র মাত্র, আগে বর্ণিত দুর্ঘটনা স্থলের চিত্র নয়। এখানে দুর্ঘটনায় জড়িত প্রত্যেকটি গাড়ির ও পথচারীর রাস্তার উপর অবস্থানস্থল ও চলাচলের পথ দেখাতে হবে। দুর্ঘটনার আগে প্রত্যেকটি গাড়ির গমনপথ ভাঙ্গা দাগ দিয়ে দেখাতে হবে। সংঘর্ষের সময় গাড়িগুলি যে যে দিকে যাচ্ছিল, তা' তীর চিহ্ন দিয়ে দেখাতে হবে। দুর্ঘটনাস্থলের রাস্তার অবস্থানের বিস্তারিত তথ্যাদি সংরক্ষণ করতে হবে। গাড়িগুলোকে ক-১, ক-২ ইত্যাদি প্রতীকে দেখাতে হবে। নিম্নে কয়েকটি সংঘর্ষের রেখাচিত্রের নমুনা দেয়া হল।



চিত্র-৩



চিত্র-৪

(৩) পুলিশের কার্যাদি ঃ

দুর্ঘটনার সংক্ষিপ্ত বিবরণ ঃ এখানে দুর্ঘটনার স্পষ্ট/সঠিক বিবরণ দিতে হবে। গাড়িগুলোকে ক-১, ক-২ ইত্যাদি বলে উল্লেখ করতে হবে। এখানে গাড়ি, পথযাত্রী বা অন্য কিছু, যা দুর্ঘটনার জন্য দায়ী সবই উল্লেখ করতে হবে।

সাক্ষী ঃ এখানে দু'জন সাক্ষীর নাম ও ঠিকানা লিখতে হবে।

বিবরণ লিপিবদ্ধকারী অফিসার ঃ এখানে দুর্ঘটনার বিবরণ লিপিবদ্ধকারী অফিসারের নাম ও পদবী লিখতে হবে।

অনুসন্ধানকারী অফিসার ঃ এখানে দুর্ঘটনার রিপোর্ট ফরম পূরণকারী ও অনুসন্ধানকারী অফিসারের নাম ও পদবী লিখতে হবে।

তত্ত্বাবধানকারী অফিসার ঃ এখানে দুর্ঘটনার রিপোর্ট ফরম পরীক্ষাকারী ও এর সম্পূর্ণতা ও নির্ভুলতা সম্পর্কে অনুমোদনকারী তত্ত্বাবধায়ক অফিসারের নাম ও পদবী লিখতে হবে।

আইনের ধারা ঃ এখানে সড়ক দুর্ঘটনার সংশ্লিষ্ট আইনের ধারা উল্লেখ করতে হবে।

কেসের অবস্থা ঃ তিনটি উল্লেখিত অবস্থার নির্দিষ্ট একটিতে গোল চিহ্ন দিতে হবে।

(৪) যানবাহন/চালক এর বিস্তারিত তথ্য ঃ

দুর্ঘটনা কবলিত প্রতিটি যানবাহনের জন্য এই যানবাহন/চালক অংশ পূরণ করতে হবে। দুর্ঘটনায় ২টির অধিক যানবাহন জড়িত থাকলে অতিরিক্ত ফরম পূরণ করতে হবে ও মূল ফরমের সাথে গেঁথে দিতে হবে। অতিরিক্ত ফরম ব্যবহৃত হলে তাতে দুর্ঘটনার ক্রমিক নং, থানা, জেলা/মেট্রোপুলিশ ও সন লিখতে হবে যাতে তা' সনাক্ত করা যায়। অতিরিক্ত ফরম ব্যবহৃত হলে সবগুলো একসাথে গেঁথে দিতে হবে।

৪.১ যানবাহন এর বিস্তারিত তথ্য ঃ

মালিকের নাম ঃ যানবাহনের মালিকের নাম লিখতে হবে।

মালিকের ঠিকানা ঃ যানবাহনের মালিকের যোগাযোগের ঠিকানা লিখতে হবে।

যানবাহনের প্রস্তুতকারী + তৈরি সন ঃ গাড়িটির বিস্তারিত বিবরণ যথা প্রস্তুতকারী, গঠন প্রকৃতি ও তৈরির সন লিখতে হবে।

38| জেলা ঃ যে জেলায় গাড়িটি রেজিস্ট্রেশন করা হয়েছে। অর্থাৎ ঢাকা, চট্টগ্রাম ইত্যাদি লিখতে হবে।

39| নম্বর ঃ এখানে গাড়িটির কেবলমাত্র রেজিস্ট্রেশন নম্বর লিখতে হবে। এতে গাড়িটির ধরণের সহিত মিল থাকতে হবে।

40| বৈধ ফিটনেস সার্টিফিকেট ঃ প্রযোজ্য ঘরে গোলদাগ দিতে হবে।

আছে ঃ গাড়িটির বৈধ ফিটনেস সার্টিফিকেট আছে।

নাই ঃ গাড়িটির বৈধ বা কোন রকম ফিটনেস সার্টিফিকেট নেই।

প্রযোজ্য নয় ঃ এই ধরণের গাড়ির জন্য ফিটনেস সার্টিফিকেটের প্রয়োজন নেই। (যেমন যন্ত্রবিহীন

গাড়ি এবং নসিমন/করিমন/ভটভটি এই ধরনের স্থানীয়ভাবে তৈরী গাড়ী)।

বীমাকৃত ঃ কৃত বীমার ধরণ বুঝে প্রযোজ্য ঘরে গোল দাগ দিতে হবে।

41| যানবাহনের ধরণ ঃ যানবাহনের ধরণের সাথে মিল রেখে গোল দাগ দিতে হবে। নসিমন/করিমন ধরনের যানবাহনকে অন্যান্য লেখা ঘরে পূরণ করতে হবে।

42| যানবাহন চলাচলের ধরণ ঃ দুর্ঘটনার সময় গাড়িটি যে কৌশলে চলছিল (বা চলার চেষ্টা করছিল) তার সাথে মিল রেখে যথাযথ ঘরে গোল দাগ দিতে হবে। এটা মনে রাখতে হবে যে, পার্ক অবস্থার অর্থ গাড়িটিকে দেখাশুনার কেউ নেই বা গাড়িটি সচল নয়। এতে রাস্তার ভীড়ের/জ্যামের মধ্যে দাঁড়ানো গাড়ি বা রাস্তায় সংযোগ স্থলে পারাপারের সারিবদ্ধ গাড়ি বুঝায় না।

আড়াআড়ি অতিক্রম ঃ এতে গাড়িটি অন্য একটি আড়াআড়ি বড় রাস্তা অতিক্রম করে সম্মুখে যাওয়া বুঝায়।

ওভার টেকিং ঃ যদি গাড়িটি অন্য গাড়িকে অতিক্রম করা অবস্থায় থাকে তবে তাকে অগ্রগমন না বলে ওভার টেকিং বলতে হবে।

43| যানবাহনে মালামাল বোঝাই ঃ গাড়িটিতে মালামাল বোঝাই করার ধরণ দেখে যথাযথ ঘরে গোল দাগ দিতে হবে। যদি অনুসন্ধানকারী অফিসারের মতে মালামাল বোঝাই নিরাপদ ও আইনানুগ হয় তবে প্রথম ঘর চিহ্নিত করতে হবে। কিন্তু যদি মালামাল বোঝাই বিপদজনক ও বে-আইনী হয় তবে দ্বিতীয় ঘর চিহ্নিত করতে হবে। বিপদজনক ও বে-আইনী বলতে অতিরিক্ত মালামাল বহন, ছাদে যাত্রী বহন ইত্যাদি বোঝায়।

44| যানবাহনের ত্রুটি ঃ বিআরটিএ কর্তৃক মটরযানের পরিদর্শন রিপোর্ট দাখিল করার পর এই ঘর পূরণ করতে হবে।

45| যানবাহনের ক্ষতি ঃ দুর্ঘটনার জন্য যানবাহনের যে ক্ষতি হয়েছে তার সাথে মিলিয়ে যথাস্থানে গোল দাগ দিতে হবে। দুর্ঘটনার আগে কোন ক্ষতি থাকলে তা' বিবেচনা করা যাবে না। যদি কোন ক্ষতি দেখা না যায় তা'হলে প্রথম ঘরে দাগ দিতে হবে।

4.2 চালকের বিস্তারিত তথ্য ঃ

নাম ঃ এখানে চালকের নাম লিখতে হবে।
ঠিকানা ঃ এখানে চালকের সঙ্গে যোগাযোগের ঠিকানা লিখতে হবে।

46| জেলা ঃ এখানে চালকের ড্রাইভিং লাইসেন্স যে জেলা হইতে ইস্যু করা হয়েছে তা' লিখতে হবে।

47| নম্বর ঃ এখানে চালকের ড্রাইভিং লাইসেন্স নম্বর লিখতে হবে।

লাইসেন্সের ধরণ ঃ লাইসেন্সের ধরণ ও যানবাহনের শ্রেণী লিখতে হবে।

48| চালকের লিঙ্গ ঃ চালক পুরুষ হলে "১" ও স্ত্রীলোক হলে "২" ঘরে গোল দাগ দিতে হবে।

49| চালকের বয়স ঃ এখানে চালকের বয়স বৎসরে লিখতে হবে।

50| চালকের ক্ষত ঃ নিম্নে বর্ণিত যথাযথ অক্ষরযুক্ত ঘরে গোলদাগ দিতে হবে।

F (মৃত্যু) ঃ দুর্ঘটনায় বা দুর্ঘটনার ৩০ দিনের মধ্যে যদি চালক মৃত্যুবরণ করে।

G (মারাত্মক) ঃ দুর্ঘটনায় যদি চালক মারাত্মক আঘাত প্রাপ্ত হয়।

S (সাধারণ) ঃ দুর্ঘটনায় যদি চালক সাধারণ আঘাত প্রাপ্ত হয়।

N (অক্ষত) ঃ দুর্ঘটনায় যদি চালক আঘাত প্রাপ্ত না হয়।

51| মদ্যপ অবস্থা ঃ এখানে চালক মদ্যপ বা সন্দেহ মুক্ত কিনা লিখতে হবে।

52| সীট বেল্ট/হেলমেট ঃ এখানে চালক সীট বেল্ট বাঁধা অবস্থায় ছিল কিনা এবং দ্বিচক্রযানের ক্ষেত্রে চালক হেলমেট পরিহিত ছিল কিনা লিখতে হবে।

(5) হতাহত যাত্রীর বিবরণ ঃ

দুর্ঘটনায় হতাহত প্রত্যেক যাত্রীর জন্য একটি করে লাইন পূরণ করতে হবে। অক্ষত যাত্রীকে অন্তর্ভুক্ত করা যাবে না।

যদি দুর্ঘটনায় ছয় জনের অধিক হতাহত যাত্রী থাকে তবে অতিরিক্ত ফরম পূরণ করতে হবে। যদি অতিরিক্ত ফরম ব্যবহৃত হয়, তবে তাতে দুর্ঘটনার ক্রমিক নম্বর, থানা, জেলা/মেট্রোপুলিশ ও সন উল্লেখ করতে হবে। অতিরিক্ত ফরম ব্যবহৃত হলে সবগুলো ফরম একসাথে গেঁথে দিতে হবে।

এই অংশ পূরণ করতে ফরমের পথচারীর বিবরণ অংশের পাদটিকার 'বি' নির্দেশ দেখা যেতে পারে। এতে গোল দাগ দিতে হবে না, কারণ যাত্রী সংখ্যা বেশি হতে পারে।

<u>দুর্ঘটনায় নিহত/আহত একজন যাত্রীর জন্য এই ফরমের একটি লাইন পূরণ করতে হবে।</u>

53| যানবাহন নম্বর ঃ যাত্রী যে যানবাহনে ভ্রমণরত ছিলেন সেই নম্বর লিখতে হবে (যেমন ১ নং যানবাহন/ ২ নং যানবাহন বা শুধুমাত্র ১,২ ইত্যাদি)। মনে রাখতে হবে যানবাহনের এই নং গাড়ির নম্বর প্লেট/রেজিস্ট্রেশন নং না।

54। যাত্রীর লিঙ্গ ঃ যাত্রী পুরুষ হলে "১" ও স্ত্রীলোক হলে "২" লিখতে হবে।

55। যাত্রীর বয়স ঃ এখানে যাত্রীর বয়স বৎসরে লিখতে হবে।

56। যাত্রীর ক্ষত ঃ এখানে যাত্রীর ক্ষতের সাথে মিলিয়ে নিচের যে কোন একটি অক্ষর লিখতে হবে।

F (মৃত্যু) ঃ দুর্ঘটনায় বা দুর্ঘটনার ৩০ দিনের মধ্যে যদি যাত্রী মৃত্যুবরণ করে।

G (মারাত্মক) ঃ দুর্ঘটনায় যদি যাত্রী মারাত্মক আঘাত প্রাপ্ত হয়।

S (সাধারণ) ঃ দুর্ঘটনায় যদি যাত্রী সাধারণ আঘাত প্রাপ্ত হয়।

57। যাত্রীর অবস্থান ঃ এই জায়গায় যাত্রীর অবস্থানের উপর বর্ণিত নিচের ছকে দেওয়া নম্বরের সাথে মিলিয়ে শুধু একটি নম্বর লিখতে হবে। কোন নম্বরে গোল দাগ দিতে হবে না। কারণ এটা শুধু নির্দেশিকার জন্য। এখানে যাত্রীর অবস্থান "গাড়ীর বাইরে" বলতে - বাসে উঠাকালীন/বাস বা ট্রাকের ছাদে/ রিক্সা ভ্যান ধরনের উন্মুক্ত যানের আরোহীকে বোঝায়।

58। যাত্রীর কার্যক্রম ঃ এই জায়গায় যাত্রীর কার্যক্রমের উপর বর্ণিত নিচের ছকে দেওয়া নম্বরের সাথে মিলিয়ে শুধু একটি নম্বর লিখতে হবে। কোন নম্বরে গোল দাগ দিতে হবে না। কারণ এটা শুধু নির্দেশিকার জন্য। এখানে যাত্রীর কার্যক্রম বলতে - যাত্রী দুর্ঘটনার সময় কি করছিল তা বোঝায়।

(6) হতাহত পথচারীর বিবরণঃ

দুর্ঘটনায় হতাহত প্রত্যেক পথচারীর জন্য একটি করে লাইন পূরণ করতে হবে। অক্ষত পথচারীকে অন্তর্ভুক্ত করা যাবে না।

যদি দুর্ঘটনায় তিন জনের অধিক হতাহত পথচারী থাকে, তা'হলে অতিরিক্ত ফরম পূরণ করতে হবে। যদি অতিরিক্ত ফরম ব্যবহৃত হয়, তবে তাতে দুর্ঘটনার ক্রমিক নম্বর, থানা, জেলা/মেট্রো-পুলিশ ও সন উল্লেখ করতে হবে যাতে তা' সহজেই সনাক্ত করা যায়। এই অতিরিক্ত ফরমে পথচারী সংখ্যা (পথচারী ৪, পথচারী ৫...) উল্লেখ করতে হবে। অতিরিক্ত ফরম ব্যবহৃত হলে সবগুলি ফরম একসাথে গেঁথে দিতে হবে।

এই অংশ পূরণ করতে নিচের পাদটিকার 'বি' নির্দেশ দেখা যেতে পারে। এতে গোলদাগ দিতে হবে না। কারণ পথচারীর সংখ্যা বেশি হতে পারে।

<u>দুর্ঘটনায় নিহত/আহত একজন পথচারীর জন্য এই ফরমের একটি লাইন পুরণ করতে হবে।</u>

59| যানবাহন নম্বর ঃ যে গাড়ি দ্বারা পথচারী আঘাত প্রাপ্ত হয় সেই গাড়ি নম্বর লিখতে হবে (যেমন গাড়ি নং ক-১, ক-২ অথবা শুধু ১,২)। মনে রাখতে হবে যানবাহনের এই নং গাড়ির নম্বর প্লেট/রেজিষ্ট্রেশন নং না।

60| পথচারীর লিঙ্গ ঃ পথচারী পুরুষ হলে "১" ও স্ত্রীলোক হলে "২" লিখতে হবে।

61| পথচারীর বয়স ঃ এখানে পথচারীর বয়স বৎসরে লিখতে হবে।

62| পথচারীর ক্ষত ঃ এখানে পথচারীর ক্ষতের সাথে মিলিয়ে নিচের যে কোন একটি অক্ষর লিখতে হবে।

F (মৃত্যু) ঃ দুর্ঘটনায় বা ৩০ দিনের মধ্যে যদি পথচারী মৃত্যুবরণ করে।

G (মারাত্মক) ঃ দুর্ঘটনায় যদি পথচারী মারাত্মক আঘাত প্রাপ্ত হয়।

S (সাধারণ) ঃ দুর্ঘটনায় যদি পথচারী সাধারণ আঘাত প্রাপ্ত হয়।

63| পথচারীর অবস্থান ঃ এই জায়গায় পথচারীর অবস্থানের উপর বর্ণিত নিচের ছকে দেয়া নম্বরের সাথে মিলিয়ে শুধু একটি নম্বর লিখতে হবে। কোন নম্বরে গোল দাগ দিতে হবে না। কারণ এটা শুধু নির্দেশিকার জন্য।

64| পথচারীর কার্যক্রম ঃ এই জায়গায় পথচারীর কার্যক্রমের উপর বর্ণিত নিচের ছকে দেয়া নম্বরের সাথে মিলিয়ে শুধু একটি নম্বর লিখতে হবে। কোন নম্বরে গোল দাগ দিতে হবে না। কারণ এটা শুধু নির্দেশিকার জন্য।

(7) সম্ভাব্য সহায়ক কারণ ঃ

নিচের দেয়া তিনটি ঘরে দুর্ঘটনার সম্ভাব্য সহায়ক কারণ নির্দেশ করা যেতে পারে। এই তিনটি ঘরের ছকে দেয়া সংখ্যা সমূহ থেকে সম্ভাব্য সহায়ক কারণ নির্দেশক সংখ্যা লিখতে হবে। যদি সহায়ক কারণ তিনটির কম হয়, তবে বাকি ঘরগুলি খালি রাখতে হবে।

65| সহায়ক কারণ ১ ঃ দুর্ঘটনার জন্য গুরুত্বপূর্ণ সহায়ক কারণ নির্দেশক সংখ্যা লিখতে হবে।

66| সহায়ক কারণ ২ ঃ দুর্ঘটনার জন্য দ্বিতীয় গুরুত্বপূর্ণ সহায়ক কারণ নির্দেশক সংখ্যা লিখতে হবে। যদি দ্বিতীয় কোন সহায়ক কারণ না থাকে, তা'হলে এই ঘর খালি রেখে দিতে হবে।

67| সহায়ক কারণ ৩ ঃ দুর্ঘটনার জন্য তৃতীয় গুরুত্বপূর্ণ সহায়ক কারণ নির্দেশক সংখ্যা লিখতে হবে। যদি তৃতীয় কোন সহায়ক কারণ না থাকে, তবে এই ঘর খালি রেখে দিতে হবে।

-- সমাপ্ত --