**ROAD ACCIDENT ANALYSIS: PREDICTING AND EXPLORING INJURY SEVERITY USING MACHINE LEARNING PARADIGMS**

A blue and white logo

Description automatically generated

**BY**

**MUHAMMAD SHUJAAT ABID**

**MS210400061**

**DEPARTMENT OF COMPUTER SCIENCE**

**FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY  
VIRTUAL UNIVERSITY OF PAKISTAN  
LAHORE, PAKISTAN**

**2024**

## TITLE

**ROAD ACCIDENT ANALYSIS: PREDICTING AND EXPLORING INJURY SEVERITY USING MACHINE LEARNING PARADIGMS**

**BY**

**MUHAMMAD SHUJAAT ABID**

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF REQUIREMENTS FOR THE DEGREE**

**OF**

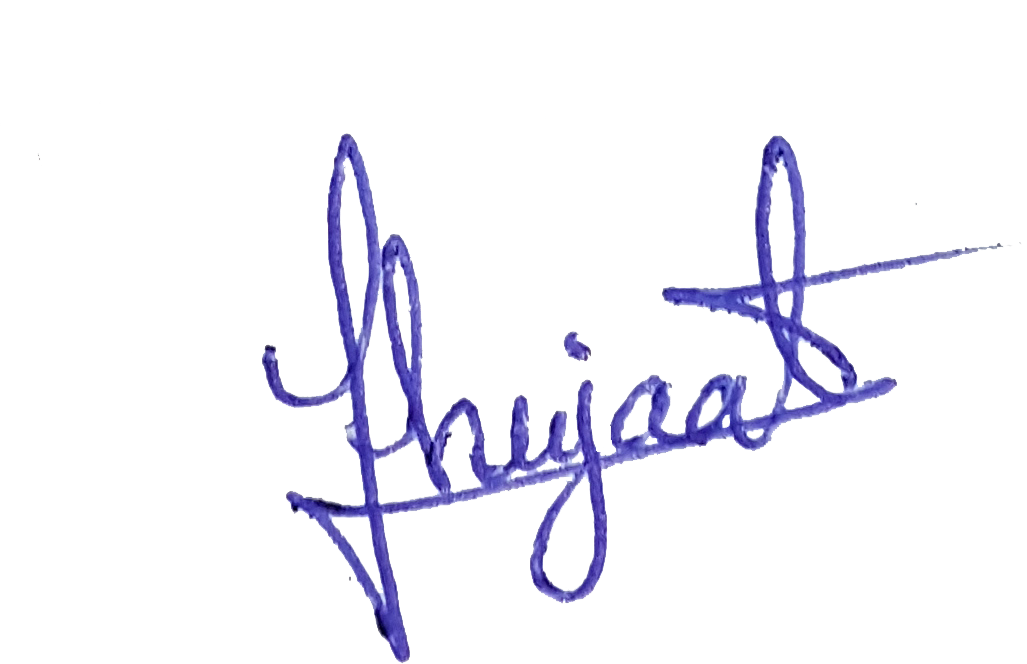
**MASTER OF SCIENCE   
IN  
COMPUTER SCIENCE**

**FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY  
VIRTUAL UNIVERSITY OF PAKISTAN  
LAHORE, PAKISTAN**

**2024**

## DECLARATION

I hereby declare that the contents of the thesis **“**ROAD ACCIDENT ANALYSIS: PREDICTING AND EXPLORING INJURY SEVERITY USING MACHINE LEARNING PARADIGMS**”** are the creation of my own research and no part has been copied from any published source (except the references, standard mathematical or geometrical models/equations/formulae/protocols etc.). I further declare that this work has not been submitted for the award of any other diploma/degree. The University may take action if the information provided is found inaccurate at any stage. (In case of default, the scholar will be proceeded against as per HEC plagiarism policy).



**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Muhammad Shujaat Abid (MS210400061)**

**To**

**The Controller of Examinations,**

**Virtual University of Pakistan,**

**Lahore.**

We, the supervisory committee, certify that the contents and forms of the thesis submitted by **MUHAMMAD SHUJAAT ABID**, VUID **MS210400061** have been found satisfactory and recommend that it be processed for evaluation by the External Examiner(s) for the award of degree.

**SUPERVISORY COMMITTEE:**

1. Dr. Mushtaq Hussain (Supervisor) Signature
2. Dr. Said Nabi (Member) Signature
3. Saad Ahmed (Member) Signature

## 

## DEDICATION

Dedicated to my beloved parents, whose constant encouragement and limitless affection have been the motivating factor behind every achievement in my life. Their persistent support and guidance have brightened my way, making the journey worthwhile. Their prayers have been the quiet strength driving me ahead, influencing my triumphs and promoting my growth.

And to my respected mentors, instructors and supervisor, your knowledge and guidance have been the foundation of my educational career. Your words of guidance and encouragement have been helpful, directing me through problems while pushing me to reach new heights.

## ACKNOWLEDGEMENTS

My profound gratitude first and foremost goes to Almighty Allah for granting me the strength, wisdom, and opportunity to embark on and successfully conclude this thesis work. This achievement would not have been possible without His blessings.

I also draw inspiration from the teachings of Prophet Hazrat Muhammad (P.B.U.H) and from the profound saying that emphasizes the importance of lifelong learning. This commitment to continuous education has been a guiding principle throughout my academic journey.

I extend my heartfelt appreciation to my esteemed supervisor, Dr. Mustaq Hussain. His invaluable guidance, unwavering patience, and insightful mentorship have been instrumental in navigating the complexities of this thesis work. His dedication and encouragement have propelled me toward academic success. My experience in pursuing MSCS at VU has indeed been an enlightening journey.

Lastly, my deepest thanks go to my parents for their unparalleled support. They hold a central place in my life, and this thesis is dedicated to them. Their unwavering assistance and guidance in solving challenges with patience and tolerance have been invaluable. In particular, I dedicate this thesis to my beloved mother.

M Shujaat Abid

## TABLE OF CONTENTS

Table of Contents

[TITLE 2](#_Toc166752060)

[DECLARATION 3](#_Toc166752061)

[DEDICATION 5](#_Toc166752062)

[ACKNOWLEDGEMENTS 6](#_Toc166752063)

[TABLE OF CONTENTS 7](#_Toc166752064)

[ABSTRACT 14](#_Toc166752065)

[CHAPTER 1 1](#_Toc166752066)

[INTRODUCTION 1](#_Toc166752067)

[1.1 Introduction 1](#_Toc166752068)

[1.2 Background 1](#_Toc166752069)

[1.3 Road Accidents 4](#_Toc166752070)

[1.4 Significance of Predicting Injury Severity 5](#_Toc166752071)

[1.5 ML Techniques Used in the Study 6](#_Toc166752072)

[1.5.1 Predictive approach: 7](#_Toc166752073)

[1.5.2 Descriptive Approach: 8](#_Toc166752074)

[1.6 Innovation of the Current Study 8](#_Toc166752075)

[1.7 Problem Statement 10](#_Toc166752076)

[1.8 Research Questions 10](#_Toc166752077)

[1.9 Research Objectives 11](#_Toc166752078)

[1.10 Strengths and Limitations of the Study 11](#_Toc166752079)

[1.11 Structure of Thesis 12](#_Toc166752080)

[CHAPTER 2 13](#_Toc166752081)

[REVIEW OF LITERATURE 13](#_Toc166752082)

[CHAPTER 3 23](#_Toc166752083)

[MATERIALS AND METHODS 23](#_Toc166752084)

[3.1 Dataset Description: 23](#_Toc166752085)

[3.1.1 Killed or Seriously Injured (KSI) Dataset: 23](#_Toc166752086)

[3.1.2 Road Traffic Accident (RTA) Dataset: 26](#_Toc166752087)

[3.2 Proposed Framework: 28](#_Toc166752088)

[3.2.1 Data Collection: 30](#_Toc166752089)

[3.2.2 Data pre-processing: 30](#_Toc166752090)

[3.2.2.1 Data cleansing: 30](#_Toc166752091)

[3.2.2.2 Feature Engineering and Transformation: 31](#_Toc166752092)

[3.2.2.3 Feature Selection: 31](#_Toc166752093)

[3.2.2.4 Class Imbalance Problem: 32](#_Toc166752094)

[3.2.2.5 Normalization: 32](#_Toc166752095)

[3.2.3 Data Splitting: 32](#_Toc166752096)

[3.2.4 Model training: 33](#_Toc166752097)

[3.2.5 Model Evaluation: 33](#_Toc166752098)

[3.2.5.1 Accuracy: 34](#_Toc166752099)

[3.2.5.2 Recall: 34](#_Toc166752100)

[3.2.5.3 Precision: 35](#_Toc166752101)

[3.2.5.4 F1-Score: 35](#_Toc166752102)

[3.2.5.5 Area Under the Curve (AUC): 36](#_Toc166752103)

[3.2.6 ML Techniques 36](#_Toc166752104)

[3.2.6.1 Classification 37](#_Toc166752105)

[3.2.6.1.1 Extra Trees Classifier 37](#_Toc166752106)

[3.2.6.1.2 Random Forest Classifier 37](#_Toc166752107)

[3.2.6.1.3 Gradient Boosting Classifier 38](#_Toc166752108)

[3.2.6.1.4 Extreme Gradient Boosting 38](#_Toc166752109)

[3.2.6.1.5 Ridge Classifier 38](#_Toc166752110)

[3.2.6.1.6 Linear Discriminant Analysis 39](#_Toc166752111)

[3.2.6.1.7 Decision Tree Classifier 39](#_Toc166752112)

[3.2.6.1.8 Light Gradient Boosting Machine 39](#_Toc166752113)

[3.2.6.1.9 AdaBoost Classifier 40](#_Toc166752114)

[3.2.6.2 Association Rule Mining 40](#_Toc166752115)

[3.2.6.2.1 Apriori Algorithm: 41](#_Toc166752116)

[CHAPTER 4 43](#_Toc166752117)

[RESULTS AND DISCUSSION 43](#_Toc166752118)

[4.1 Visualization 43](#_Toc166752119)

[4.2 Statistical Analysis: 48](#_Toc166752120)

[4.3 Experimental Setup: 50](#_Toc166752121)

[4.4 Results 54](#_Toc166752122)

[4.5 Discussion 68](#_Toc166752123)

[4.6 Deployment 87](#_Toc166752124)

[CHAPTER 5 88](#_Toc166752125)

[SUMMARY 88](#_Toc166752126)

[5.1 Summary 88](#_Toc166752127)

[5.2 Thesis Contributions 88](#_Toc166752128)

[5.3 Future Work 89](#_Toc166752129)

[LITERATURE CITED 90](#_Toc166752130)

**LIST OF FIGURES**

**FIGURE DESCRIPTION           PAGE**

Figure 3.1 Proposed Framework of the Study 29

Figure 4.1 Number of Collisions Happened over 17 Years (2006-2002) 44

Figure 4.2 Frequency of Collisions and the Injury Type against Weekday 45

Figure 4.3 Frequency of collisions and the injury severity against Traffic

Control Signal 45

Figure 4.4 Frequency of collisions in RTA over years (2020-July 2023) 46

Figure 4.5 Frequency of collisions in RTA against Gender 47

Figure 4.6 Frequency of collisions in RTA against Age Range and Education Title 48

Figure 4.7 Distribution of Target Attribute in KSI Dataset Before and After SMOTE 52

Figure 4.8 Distribution of Target Attribute in RTA Dataset Before and After SMOTE 52

Figure 4.9 Main Groups of Attributes to Obtain Association Rules 54

Figure 4.10 Visualization of Accuracy, Precision and Recall of ML models trained

over KSI 69

Figure 4.11 Visualization of F1-Score and AUC of ML models trained over KSI 70

Figure 4.12 ROC Curve of top performing ML models trained over KSI Dataset 71

Figure 4.13 Feature Importance for KSI Dataset 72

Figure 4.14 Visualization of Accuracy, Precision and Recall of ML Models

Trained over RTA 73

Figure 4.15 Visualization of F1-Score and AUC of ML Models Trained over RTA 74

Figure 4.16 ROC Curve of top performing ML models trained over RTA Dataset 75

Figure 4.17 Feature Importance for RTA Dataset 76

Figure 4.18 Top Association Rules of KSI and RTA Datasets 85

**LIST OF TABLES**

**TABLE DESCRIPTION PAGE**

Table 3.1 Input features of the KSI dataset 24

Table 3.2 Input features of the RTA dataset 27

Table 4.1 Statistical Analysis on KSI Dataset 49

Table 4.2 Statistical Analysis on RTA Dataset 50

Table 4.3 Confusion Matrix of LGBM Model Developed Over KSI Dataset 55

Table 4.4 Confusion Matrix of Extreme GB Model Developed Over KSI Dataset 55

Table 4.5 Confusion Matrix of RF Model Developed Over KSI Dataset 56

Table 4.6 Confusion Matrix of ET Classifier Model Developed Over KSI Dataset 56

Table 4.7 Confusion Matrix of GBC Model Developed Over KSI Dataset 57

Table 4.8 Confusion Matrix of DT Model Developed Over KSI Dataset 57

Table 4.9 Confusion Matrix of Ridge Classifier Model Developed Over KSI Dataset 58

Table 4.10 Confusion Matrix of LDA Model Developed Over KSI Dataset 58

Table 4.11 Evaluation Measures of Models Developed Over KSI Dataset 59

Table 4.12 Confusion Matrix of LGBM Model Developed Over RTA Dataset 59

Table 4.13 Confusion Matrix of GB Model Developed Over RTA Dataset 60

Table 4.14 Confusion Matrix of RF Model Developed Over RTA Dataset 60

Table 4.15 Confusion Matrix of ET Model Developed Over RTA Dataset 61

Table 4.16 Confusion Matrix of AdaBoost Model Developed Over RTA Dataset 61

Table 4.17 Confusion Matrix of DT Model Developed Over RTA Dataset 62

Table 4.18 Confusion Matrix of LDA Model Developed Over RTA Dataset 62

Table 4.19 Evaluation Measures of Models Developed Over RTA Dataset 63

Table 4.20 Top associations for ‘Nature of Accident’ over KSI 64

Table 4.21 Top associations for ‘Nature of Accident’ & ‘Environmental Factors’

over KSI 64

Table 4.22 Top associations for ‘Nature of Accident’ & ‘Driver Factors’ over KSI 65

Table 4.23 Top associations for ‘Nature of Accident’ & ‘Road Factors’ over KSI 65

Table 4.24 Top associations for ‘Nature of Accident’ & ‘Vehicle Factors’ over KSI 66

Table 4.25 Top associations for ‘Nature of Accident’ & ‘Involvement Factors’

over KSI 66

Table 4.26 Top associations for ‘Nature of Accident’ over RTA 67

Table 4.27 Top associations for ‘Nature of Accident’ & ‘Environmental Factors’

over RTA 67

Table 4.28 Top associations for ‘Nature of Accident’ & ‘Vehicle Factors’ over RTA 68

Table 4.29 Top associations for ‘Nature of Accident’ & ‘Involvement Factors’

over RTA 68

**LIST OF ABBREVIATIONS**

WHO World Health Organization

ML Machine Learning

KSI Killed or Seriously Injured

RTA Road Traffic Accident

SMOTE Synthetic Minority Over Sampling Technique

ET Extra Tree Classifier

RF Random Forest

GBC Gradient Boosting Classifier

XGBoost Extreme Gradient Boosting Classifier

LDA Linear Discriminant Analysis

DT Decision Tree

AdaBoost Adaptive Boosting

TP True Positive

TN Ture Negative

FP False Positive

FN False Negative

TPR True Positive Rate

FPR False Positive Rate

ROC Receiver Operating Curve

AUC Area Under the Curve

# ABSTRACT

Road safety is a serious concern in urban environments, involving the well-being of all road users, with a special emphasis on those considered vulnerable such as pedestrians, bicyclists and motorcyclists etc. The purpose of this study is to use the Traffic Accident datasets obtained from various online and local sources to uncover significant factors contributing to fatal or major injuries among road users and design effective solutions.

The research aims to analyze the injury severity of road users and the leading factors that lead to fatal or major injuries. This study implemented several Machine Learning (ML)-based algorithms for building predictor models. These ML models were evaluated on traffic dataset to check their performances based on several defined metrics such as accuracy, recall, precision, f1-score and Area Under the Curve (AUC).

The findings revealed that Light Gradient Boosting Machine, Extreme Gradient Boosting, Gradient Boosting Classifier and Random Forest demonstrated better performance as compared to the other models. The Light GBM exhibited overall excellent performance, achieving an overall accuracy of around 75% on both datasets.

The current study has utilized the Association Rule Mining technique to determine the underlying hidden factors that lead to collisions and fatal or major injuries. The extracted rules reveal that certain factors, such as over speeding, aggressive driving, pedestrian collisions, disobeying traffic rules, lost control, driver’s inattentiveness and the absence of traffic controls at major roads and intersections are associated with a higher risk of fatal or major injuries. Lastly, the study provides a comprehensive comparison of factors contributing to severe collisions in Pakistan and Canada. It is found that there is a need for targeted enforcement in both countries including stricter licensing and education initiatives for young drivers about responsible driving behaviors. In Pakistan, there is a need of proper measures to enforce the use of helmet for bike riders along with improving the lighting conditions. Whereas Canada should adopt the proper measures to ensure the pedestrian and cyclist safety.

The proposed framework can be deployed in a real environment for improving road safety in urban environments. The identification of specific risk factors can inform the development of targeted interventions to reduce the incidence of collisions and in turn fatal injuries in real time.

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

In today's fast-paced world, roads, transportation systems and vehicles play a crucial role in connecting people and transferring goods and services across cities, countries, and even continents. With the increase in the global population and urbanization, there is also a tremendous increase in the number of road vehicles over the years. The development and maintenance of road infrastructure has significant economic, social, and environmental impacts, affecting industries, governments, and societies at large.

However, along with the benefits of road infrastructure come risks and challenges. Despite efforts to improve infrastructure, implement education and awareness campaigns, and enforce traffic laws, road users remain at risk of collisions with vehicles. Road accidents, particularly those involving vulnerable road users, are a major public health concern worldwide.

As we come up with the complicated scenario of road safety, it becomes obvious that traditional techniques, however useful, may fall short in dealing with the dynamic nature of road accidents. The need for a more proactive and responsive solution is crucial. Hence, the necessity for an automated traffic safety system, powered by cutting-edge Machine Learning paradigms, is becoming increasingly important. Such a technology has the ability to not only anticipate injury severity but also inquire about the deep network of causes contributing to collisions. By utilizing the power of Machine Learning, we can pave the path for a better, more secure future on our roads, where the harmony of transportation is not overshadowed by the shadows of possible risk.

## 1.2 Background

Traffic accidents are a complicated subject that is impacted by numerous elements, such as road conditions, type of vehicle, atmospheric conditions, driver behavior, etc. (Yannis et al., 2016). These accidents, ranging from inadequate collisions to major wrecks, have profound consequences, affecting people, their loved ones, and their communities. Identifying the variables involved in determining the degree of injuries arising from these incidents is vital for creating appropriate preventative strategies and boosting overall road safety. Implementing before-and-after studies allows us to assess the impact of road safety interventions such as tolls or speed limits, and the implementation of legislation, like the safety belt act. A notable example is Percoco's (2016) investigation, he examined the consequences of implementing a road pricing scheme for motorists in Milan on accident rates. Although the total number of accidents decreased in the treated area compared to non-treated areas, there was no discernible effect on the number of fatalities resulting from the measure.

Traffic accidents prediction may be divided into macro and micro prediction. The macro prediction refers to the prediction of accidents on long terms such as long timeframes or large geographical area. On the other hand, micro predictions refer to the prediction of accidents in a short term such as short period of time or at a specific location, road or intersection (Shunshun et al., 2023).

Prior studies on road traffic accidents have largely applied standard statistical analysis to discover relevant factors and examine the influence of multiple variables on injury severity. These methods stand as one of the most prevalent approaches in predicting traffic accidents. Statistical models play a crucial role in harnessing limited historical data to make predictions (Li et al., 2020), assuming that past accident patterns bear some resemblance to what the future holds. By scrutinizing this historical data, one can make educated predictions about impending traffic accidents.

Kelly et al., (2008) states that the traffic accidents data that is analyzed through statistical methods uses the quantitative theory as its base. Shunshun et al in 2023 explained that the predictions based on historical trends and past data using several mathematical, statistical and analytical techniques is called Quantitative forecasting. Statistical analysis and machine learning methods come under the umbrella of quantitative forecasting techniques. Quantitative theory finds extensive utility across practical domains, extending beyond traffic safety to encompass fields like finance, environment, and healthcare (Keqiang et al., 2005).

Tongyuan et al., (2007) introduced the traffic accident prediction approach utilizing a grey prediction model. The methodology involves an initial analysis of historical traffic accident data, followed by the construction of a grey prediction model to forecast future trends in accident occurrences. This research also takes into account external factors to enhance prediction accuracy. This method demonstrates high feasibility and accuracy in practical application, offering valuable assistance to traffic management departments in implementing timely measures to mitigate the occurrence of traffic accidents.

Although these research efforts have produced significant insights, the intricate and dynamic nature of vehicular crashes necessitates a more comprehensive and sophisticated approach.

The introduction of Machine Learning (ML) has led to a paradigm change in the study of road traffic incidents. ML is the branch of Artificial Intelligence (AI) that enables the systems to learn from the existing data and make predictions or judgments without being programmed in any way. This capacity provides new options for comprehending the deep patterns and correlations within collision data, exceeding the constraints of standard statistical methodologies. The essence of machine learning-driven traffic accident prediction techniques lies in the analysis of historical traffic data. By employing machine learning algorithms, these techniques construct prediction models aimed at foreseeing the likelihood and impact of future traffic accidents (Pourroostaei et al., 2023).

In addition to, ML adds to the greater objective of guaranteeing road safety by automating the study of massive datasets in real time. The automated collision detection systems, anomaly detection algorithms, and risk assessment models powered by ML offer a flexible and adaptable framework for avoiding collisions and limiting their consequences. Pourroostaei Ardakani et al., (2023) conducted a study to predict road accidents. They used four machine learning algorithms namely decision trees, random forest, multinomial logistic regression, and naïve Bayes. The outcomes reveal satisfactory levels of accuracy ranges from 60% to 80% for predicting car accidents, with the exception of naïve Bayes.

The applicability of ML in road traffic accident investigation spans beyond prediction. ML approaches dive into the deep interaction of variables, discovering complicated correlations among these variables. ML approaches also help out in finding the most contributing aspect and factors that leads to collisions, this ML approach is known as association rule mining. The notion of Frequent Itemset Mining (FIM) was originally presented by Agrawal et al., (1993). The Apriori method, a major component of FIM and association rule learning, was initially described in the work of Agrawal and Srikant (1994). It allows the detection of hidden elements and related rules that could defy standard statistical approaches. So, the ML approaches give a completer and more dynamic picture by using sophisticated algorithms.

As we traverse this intersection between technological advancement and road safety, this research intends to contribute to the developing panorama of ML applications in the field of road traffic accidents. By leveraging modern ML algorithms for injury severity prediction and adopting not only the predictive but also the descriptive ML approaches to uncover unseen factors, this study aims to expand our awareness of road safety mechanisms and pave the path for better preventative tactics.

## 1.3 Road Accidents

Accident-related injuries are the greatest reason for fatalities, especially for the young ones. The World Health Organization (WHO) estimates that approximately 1.3 million lives are lost annually due to road accidents, with over half of these fatalities impacting vulnerable road users that includes pedestrians, bicyclists, and bikers. In addition to the lives cut short by these incidents, an estimated 20 to 50 million people suffer non-fatal injuries each year, often resulting in permanent impairments. Notably, while holding nearly 60% of the world's automobiles, low- and middle-income countries bore the brunt, accounting for 93% of the world's road traffic deaths (Wordl Health Organization, 2022).

Keeping in view the above fact, this study intended to use two datasets one of them is the KSI (Killed or Seriously Injured, 2023) dataset provided by the Toronto Police Services. This dataset contains detailed information on the characteristics of collisions that result in fatal injuries. It includes demographic data, collision location, time, environmental variables, road conditions and other factors that may contribute to the risk of injury. The other one is the local dataset collected from The Punjab Emergency Service, Rescue 1122 Rawalpindi (RTA, 2023). This dataset is enriched with various accidental features including time and date of collisions, environmental variables, location, number of involvements, demographic details of involved ones, cause of accident, injury type, patient status and the vehicle involved.

Both of these datasets serve as essential tools for analyzing vehicular collisions and fatalities, delivering extensive insights into the precise facts underlying these incidents. The KSI gives a complete summary of accident factors leading to fatal injuries in the developed countries whereas the local dataset supplements the study with a broad variety of incidental factors in the developing countries like Pakistan. The comparative examination of these datasets not only enriches our understanding of factors contributing to occurrences in developed places but also provides a nuanced perspective on those occurring in developing countries like Pakistan.

## 1.4 Significance of Predicting Injury Severity

Enhancing road safety is an important research topic because it is one of the leading reason of fatalities in the world. The analysis of the factors that lead to severe or fatal injuries is of utmost importance so that the proper measures can be taken to reduce the fatalities rate due to collisions.

Predictive models are extensively being used in every field of life, within the field of road accidents, the prediction of injury severity appears as a keystone in developing effective preventative tactics. From the literature, it can be seen that many researchers are working in this field and one notable example is the work of Percoco. In 2016, they investigated several leading factors of collisions and then implemented the proper road safety measures to reduce the number of collisions. After this, they conducted a post study to assess the implications of these interventions and found that the total number of collisions decreased in the targeted area compared to non-treated areas (Percoco et al., 2016). Such a type of developed model can easily be integrated into the real-time traffic monitoring system to identify the factors that may lead to a collision.

The ability to predict the probability of an accident gives us a proactive position, enabling early interventions and limiting the aftermath's impact. A well-informed decision can be made by analyzing the real-time data to avoid collisions. Furthermore, by identifying the variables leading to fatal or major injury severity, we acquire the capacity to modify road safety measures, saving lives and protecting the well-being of individuals using the road infrastructure. The critical variables identified can be effectively used by the officials to formulate evidence-based policies aimed at improving road safety.

## 1.5 ML Techniques Used in the Study

The cognitive capability of computers in certain disciplines is still below to that of humans, however owing to machine learning algorithms, computer skills are developing swiftly in domains such as e-learning, recommendation, pattern recognition, image processing, medical diagnosis, and many more. ML is an area of artificial intelligence. Machine learning algorithms are developed using sample data as inputs and then evaluated with new data (Hussain et al, 2018). ML algorithms can automatically uncover complicated patterns from features extracted from the dataset. These identified patterns enable us to make smart and informed decisions about the current scenarios (Holland et al., 1992).

ML is widely being used to build prediction models from large datasets; ML approaches have the ability to cover both numerical and categorical predictor variables. Decision trees (DTs) is one of the most extensively used machine learning algorithms that is used to create trees and find prediction rules based on existing data (Edelstein et al., 1998). Soheil Rezashoar, Ehsan Kashi and Soheila Saeidi in 2023 has also utilized the ML technique for the prediction of traffic accident severity. They have provided a comparison of several algorithms in their work and concluded that Naive Bayes, Support Vector Machine and Neural Network has predicted the accident severity with reasonable accuracy. Furthermore, they have also stated that it is not necessary to define the relationship between variables while implementing ML models and these models are more suitable where accuracies matters than the associations (Rezashoar et al., 2023).

Furthermore, conventional statistical models rely on predetermined relationships and underlying assumptions, rendering them subject to misleading outcomes when those assumptions are violated. In contrast, machine learning (ML) is a cutting-edge, non-parametric approach capable of adeptly capturing non-linear effects in both continuous and discrete data. Without relying on past assumptions, ML achieves greater prediction accuracy, providing a more flexible and robust modeling alternative.

The road accident data is multidimensional and quite heterogeneous, the traditional statistical analysis may not be able to identify hidden patterns and relationships in the data that are important for predicting the likelihood of fatal or major injuries. However, machine learning techniques can provide a solution to this problem.

In the domain of road safety research, the incorporation of advanced machine learning techniques offers a viable route for understanding and tackling the complexities connected with road traffic accidents. The study has applied multiple machine learning techniques as analytical learning approaches intended to predict the injury severity level and to find out the hidden associations among the critical factors.

The study aims to use predictive and descriptive machine learning approaches to get insights into traffic accident datasets.

### 1.5.1 Predictive approach:

It is an approach in machine learning that aims at developing a predictor model based on some algorithm. The primary goal of these models is to predict the targeted variable accurately based on some input data. The models are trained over the existing or historical datasets, which can then be used to predict the new or unseen data. It is an effective technique to get insights of the data and to make rational decisions in a variety of disciplines.

By exploiting the predictive powers of machine learning, this study seeks to identify profound connections among numerous accidental dataset features including demographic data, environmental variables, road conditions, and collision circumstances and to develop a robust model that can accurately predicts the injury severity sustained by the involvement in a collision.

### 1.5.2 Descriptive Approach:

It is another approach in machine learning that focuses on data analysis, summarization and understanding the patterns within the dataset to gain insights of the data rather than just predicting the future outcomes as in the case of predictive approach. This approach describes the correlations between variables, as well as their associations and underlying patterns that might otherwise go unnoticed. So, by using this method, one may make rational decisions based on factual facts and established patterns.

This study also presents the application of descriptive machine learning approach. The investigation of dataset through descriptive machine learning techniques adds a layer of depth to this research. This technique enables the detection of patterns within the dataset, giving light to intricate interactions between variables. It allows us to identify relationships and dependencies among variables, which can help to identify risk factors. This study aims to identify the factors that lead to fatal or major injuries of the involved.

This merging of predictive and descriptive machine learning not only aims to estimate injury severity but also to find hidden insights.

## 1.6 Innovation of the Current Study

The mortalities that happen due to road accidents is one of the main causes of death. This worldwide challenge also bears considerable economic ramifications, constituting around 3% of the total GDP in most countries. The UN General Assembly has established an extremely lofty objective to cut by half the worldwide mortality rate from road traffic accidents by 2030 (World Health Organization, 2022). The urgent need for comprehensive efforts to mitigate and relieve the far-reaching impacts of road traffic events is obvious.

So, the current study emerges as an inspiration for the merging of road safety with innovative ML paradigms. This research strives not only to the prediction of injury severity but also to examine the more subtle factors associated with accidents. In doing so, it establishes the framework for an in-depth understanding of road safety dynamics. The integration of predictive models and descriptive approach sets the stage for a transformational approach—one that strives not simply to respond to incidents but to proactively design a safer, more secure future on our roadways.

This study has used several basic and some ensemble machine learning algorithms to build classifiers that predict the injury severity level of the victims who got injured in a collision. These classification methods have been chosen because the road traffic accident dataset contains a combination of numerical and categorical features and these algorithms can handle both of these features. These algorithms also work well with noisy data as they aren't thrown off by nonlinear correlations between factors. These models are transparent, acting like white boxes, making it easy for users to understand and interpret their rules. Additionally, they enable us to identify key variables in the dataset effortlessly. The study has implemented various popular ML models including Light GBM, extreme GB, RF, ET, DT, LDA, GB, and AdaBoost on the road traffic accident datasets. It is concluded that the Light Gradient Boosting Machine is the best performing model in this problem having an accuracy of 75%. Next, the application of Association Rule Mining reveals certain factors, such as over speeding, aggressive driving, pedestrian collisions, disobeying traffic rules, lost control, driver’s inattentiveness and the absence of traffic controls at major roads and intersections are associated with a higher risk of fatal or major injuries.

Then the study has also provided a comprehensive comparison of risk factors between Pakistan and Canada that strengthens the study's innovative approach to cross-regional analysis with valuable policy implications. It is concluded that the Pakistan need to ensure the proper measures to enforce the use of helmet for bike riders along with improving the lighting conditions. On the other hand, Canada should adopt the proper measures to ensure the pedestrian and cyclist safety. Both of these countries need to focus on strict licensing and education of young drivers.

This study has important implications for urban planning and transportation policy, as well as for the development of targeted interventions. The obtained results can be used to reduce the incidence of fatal or severe collisions by using data-driven approaches and machine learning techniques and to address public safety issues. This study also demonstrates the value of applying innovative methods to complex problems.

## 1.7 Problem Statement

The increasing number of severe injuries and fatalities from road accidents underscores the need for reliable and effective models to predict injury severity. Currently, the absence of an affordable and dependable system for predicting injury severity presents a significant challenge for road safety authorities and urban planners. Existing models aim primarily at achieving high levels of accuracy, utility, and reliability. This research seeks to address this gap by developing models that predict injury severity with high accuracy, recall, precision, F1-score, and Area Under the Curve (AUC) using advanced machine learning algorithms. Additionally, this study incorporates Association Rule Mining to identify hidden factors that contribute to severe injuries, offering deeper insights into the underlying causes of fatal and major accidents. Moreover, the research includes a detailed comparison of risk factors between a developing country (Pakistan) and a developed country (Canada), taking into account variations in road safety regulations and enforcement practices. This comparison aims to inform targeted interventions and enhance overall road safety strategies in both contexts.

## 1.8 Research Questions

The study focuses on the following research questions:

**Question 1:** Can ML algorithms effectively model the injury severity level resulting from collisions in road accidents, and if so, which ML classifier offers optimal performance in predicting the injury severity level across different geographic regions?

**Question 2:** What are the underlying hidden factors that contribute to fatal or major injuries?

**Question 3:** Given the differences in road safety regulations and enforcement between Pakistan and Canada, how do the identified risk factors and their impacts on injury severity differ between these two countries? What policy recommendations can be derived from these differences to enhance road safety initiatives in both regions?

## 1.9 Research Objectives

1. To provide an injury severity predictor model with higher performance in terms of accuracy, sensitivity, specificity, precision, and ROC score.

2. To make the decision-making process easier for road traffic managers and policy makers.

3. To identify the underlying hidden factors that contribute to fatal or major injuries resulting from road accidents.

4. To conduct a comparative analysis of the identified risk factors associated with injury severity in road accidents between Pakistan and Canada, taking into account the differences in road safety regulations and enforcement practices in these two countries.

## 1.10 Strengths and Limitations of the Study

The study has the following strengths and weaknesses.

**Strengths:**

* The study provides an implementation of advanced machine learning algorithms including ensembles models to predict the injury severity with high accuracies.
* The use of both predictive and descriptive analysis provides a novel approach to uncover the hidden factors that contribute to severe injuries. It offers deeper insights into the root causes of fatal and major road accidents.
* The comparative analysis between a developing country and a developed country reveals how different approaches impact safety, informing targeted interventions and policy recommendations.

**Limitations:**

* This study leverages existing datasets, which may have inherent limitations in completeness, accuracy, and representativeness.
* The findings of the study may not be generalized as the performance of machine learning models may vary because they are inherently dependent on the quality and quantity of the input data.
* The comparison between Pakistan and Canada although is informative but not generalized as the cultural, economic, and infrastructural factors can significantly influence the outcomes.

## 1.11 Structure of Thesis

The first chapter focuses on the problem statement, road traffic accidents, the significance of predictor models and the use of machine learning algorithms in this modern age. Related Work is discussed in Chapter 2, and details about Material and Methods are presented in Chapter 3. Chapter 4 provides the Result and Discussion on the findings and lastly, Chapter 5 presents the summary and future directions.

# CHAPTER 2

# REVIEW OF LITERATURE

Considerable research has been conducted in the past to investigate road traffic accidents in different directions. The past researches have used different techniques on different datasets with varying number of input features to investigate the relationships between these features. Some of the related work is mentioned below:

Ahammad et al., (2023)conducted research with the aim of building a prediction model that is capable of automatically classifying the accident severity level. This research has applied one of the supervised learning techniques of machine learning known as classification for this purpose. They have developed models that are based on Random Forest (RF), Naïve Bayes, Gradient Boosting Classifier (GBC), Logistic Regression (LR) and K-Nearest Neighbor (KNN). The research has collected road traffic accident data from the UK. This dataset contains information about road class, weather conditions, road surface, vehicle speed, junction information, lighting conditions and many others. Overall, the dataset consists of 34 attributes and around 2 million records. The accident severity column is categorized into three values i.e., minor, serious and fatal. The research has selected 80% of the data for training purposes and the remaining 20% for evaluation purposes. The metrics of accuracy, precision, TPR, TNR and F1 score have been used for evaluation. After developing models with the above-mentioned algorithm, it has been concluded that the Random Forest classifier has achieved the best results among all of these with an accuracy of 63%. Following it, the Gradient Boosting algorithm is the second best performing predictor with an accuracy of 62%. Gradient Boosting has taken a longer time to build the model as compared to Random Forest, but it has achieved the highest F1 score and AUC among all others. Lastly, the research has concluded that the severity of an accident is related to many factors, not any single factor contributes to it. It is highly dependent on temporal data including weather, light conditions, road surface, traffic signal, speeding and many more.

Kuyumcu et al., (2023) conducted a study in order to find out the relationship between factors that increase the chance of accidents using machine learning methods. This research has applied the association rule mining techniques on traffic accident data. Their dataset contains 6 years of collision data of the city Sakarya, Turkey from 2015 to 2020 having 23,551 entries and 26 attributes. The necessary preprocessing techniques are applied on this data to clean it and then divide it into five sections as follows: nature of accidents, nature of accidents and external factors, nature of accidents and driver factors, nature of accidents and road factors, lastly, nature of accidents and vehicle factors. The research has taken one section at a time and tried to find out the hidden patterns between them. This research has found a lot of association rules, but they had taken the lift as a measure to consider the most important rules and discussed the top 9 associations from nature of accidents, top 10 association rules regarding accidents and external factors, top 4 rules related to accidents and driver features, top 11 associations with respect to accident and road features and lastly, top 4 features related to vehicle and accident. It has been concluded that speeding is highly correlated with the injury severity, rollover and run off road leads to major vehicle damage. Close following leads to rear end collisions, hitting pedestrians leads to non-injured drivers. There exist seasonal patterns in the dataset such that fatal injuries are most common in spring and summer seasons due to high speed of vehicles. Drivers with low educational level tend to be involved more in collisions due to rule violations and side collisions. It has also been found out that the young drivers lead to severe accidents due to over speeding whereas the middle-aged drivers are mostly involved in the side collisions. However, the limitation of the study is that it does not take into account the injury severity level of the persons involved in the accidents. It has been suggested by the research as future work to make predictor models by using machine learning methods that can predict the nature of accidents.

Neil and Datu (2023) have uncovered the insights of the traffic accidents happening in the Angeles city of Philippines. They applied the association rule mining technique on 1438 records of traffic collisions from January 2015 to June 2018. The data was collected from Angeles City Traffic Management and Enforcement Unit (ACTMEU). The research conducted to find out the most common characteristics of the suspected motorists and their attributes that contribute mostly to road accidents. They have incorporated the CRIPS methodology for data mining and performed the feature extraction technique to include some of the most important attributes in the analysis. Confidence, support and lift are the metrics that have been used to evaluate the association rules in the study. This research has found the four associations whose consequence is non-fatal accident. They have not found any satisfactory rule that leads to Fatal accident. They have revealed from association rules and descriptive analysis that majority of the road accidents involve male drivers with age range between 15 to 29. Balibago city is the most prone to collisions from 8 PM to 12 AM and on Sunday. However, the association rules point out that road accidents are more likely to happen from 8 Pm to 12 AM and on Saturday. A further analysis shows that Balibago having the greatest number of accidents does not have the higher probability of having collision during the 8Pm to 12Am on Saturday.

Emu M et al., (2022) developed a model that predicts the consequence of a collision as fatal or non-fatal. The focus of their study was the prediction of collision severity with the help of machine learning algorithms. For developing the required model, this research has used four machine learning methodologies that include Support Vector Machine, Random Forests, K-Nearest Neighbor, and Convolutional Neural Network. This paper has employed methods that are related to deep learning on a large dataset that is collected over 20 years and contains 5.8 million entries. The overall dataset contains 104 attributes that are reduced only to 67 without losing any performance of prediction models by using information content analysis technique. An in-depth analysis reveals that road traits, vehicle type, weather conditions, usage of safety devices, time of collisions, involvement class, and the status of traffic control are the aspects that participate most in the prediction models. The dataset has the problem of class imbalance and to tackle this issue in the study the method of under sampling has been used. The study has used accuracy, recall, precision and F1-score as performance measures for each method. The accuracy of all of these methods ranges from 69.38% to 75.56%. The KNN has the lowest accuracy of 69.38% and the CNN performed significantly higher than all of these methods with an accuracy of 75.56%. CNN also has better precision (75.74%) and recall (75.45%) than the other methods. So, these results show that Convolutional Neural Network (CNN) achieves the best performance in the prediction of accident seriousness by using both soft and majority voting. It has been concluded from the study that CNN is the most suitable option that can be deployed for the purpose of developing an ITS (Intelligent Transportation System) to perform the task of real-time fatality hazard predictor on various navigation apps.

Yadav et al., (2021) identified the factors that lead to road accidents. They tried to investigate the factors that impact on road accidents by using machine learning techniques and then based on these factors they have tried to give some valuable suggestions regarding road accidents. For this purpose, they have used the Killed or Seriously injured (KSI) dataset provided by the City of Toronto Police in their Open Data Portal. This dataset contains all the information of traffic accidents that have been reported from 2007 to 2017. The data has been cleaned before use in the analysis to avoid any misleading information. The input features are selected using feature selection technique to reduce the dimensionality and column ‘FATAL’ was chosen as target output. The research has applied several supervised machine learning techniques including Decision Trees, K-Nearest Neighbor (KNN), Naïve Bayes and AdaBoost. It has also applied several unsupervised machine learning techniques including K-mean clustering algorithm, CART and ROC value. This research has achieved 81.5% performance with the CART algorithm and AdaBoost also achieves the best performance among these. By analyzing the data, it has been revealed that the number of accidents has been slightly reduced over the years, and the greatest number of accidents happens from June to October. It has been concluded that most of the accidents that lead to fatal injuries happen in the east Toronto region and the victims are mostly pedestrians. Aggressive driving (62.9%) is the major cause of accidents, then over speeding (21.4%), red light signal violation (10.4%) and then alcohol (5.3%). It has been concluded that the results of this study are not very general, but they have classified the secure and risky zones very well.

Shanshal D et al., (2020) conducted research with the aim of building a prediction model that identifies the patterns in severe and fatal collisions happening in Toronto. This research tried to predict the fatalities of individuals including only drivers, pedestrians and cyclists using data analytics and machine learning techniques. The KSI dataset provided by the Toronto Police Services is used in this study. It has been analyzed that pedestrians, cyclists and motorcyclists are the most affected ones that get major or fatal injuries in the collisions. This research has used the data mining approach of association rules, classification algorithms including Lasso Regression and Random Forest. The research has found that Random Forest is the best predictor of injury severity with an accuracy of 80% for both drivers and pedestrians and 89% accuracy for cyclists. The research has used the apriori algorithm for finding association rules and to mine the dataset and uncover the patterns and rules between the variables. Many association rules for each category have been created by the research. It is concluded that the factors of aggressive driving, failing to yield right of way, improper turns and driver’s inattentiveness lead to severe collisions. Several environmental conditions such as clear and dry weather in summer and spring seasons lead to more accidents. The cyclist in the months of June and July gets fatal injuries. In the case of rainy weather, wet surfaces, and dark light causes pedestrian's risk of fatal or major injury. The cyclist gets fatal injuries when the driver sideswipes cyclists while driving in the same direction, a motorist turning left across the cyclist’s path and the cyclist struck by the opened vehicle door.

Gan J et al., (2020) got inspired by the Deep Forest algorithm based on the decision tree ensemble and conducted a study with the aim of developing a model that will predict the severity of traffic accidents based on this DF algorithm. According to the research, this is the first study that uses the Deep Forest algorithm to predict the severity of traffic accidents. This research used the UK road safety dataset in 2016 obtained from the Kaggle website. The dataset includes 18 items in total including latitude and longitude, time characteristics, vehicle type, gender, age, vehicle age, speed, light conditions, weather, road surface and other data characteristics. First, the research did the reliability verification of the dataset and then the preprocessing methods were applied. Then the research has used a combination of Randomized Search and Grid Search methods for the optimization of parameters that result in eight features, and these eight features are used as the main data features. For the validation of their proposed model and to compare the performance of the Deep Forest algorithm, this research has incorporated other famous machine learning algorithms also that including DNN, Random Forest, LightGBM, XGboost, K-Nearest Neighbor (KNN) and decision trees. Then after the comparison, it is concluded that Deep Forest has achieved the best performance among all of these. The Deep Forests achieved an accuracy of 90.69%, the recall is 0.92 and the highest F1 score and ROC of 0.91 and 0.93 respectively. Then the research explained several advantages of the Deep Forests algorithm as compared to traditional machine learning methods. The high prediction accuracy of Deep Forests algorithms reveals that it can be used more effectively for the prediction of accident severity, and it will be more conducive to the transplantation of models as it requires fewer hyper-parameters as compared to other models. This model can be easily adapted to solve lots of different traffic problems as well.

Hébert et al., (2019) conducted research aiming to build a model that can predict the likelihood of collision at various locations in the city of Montreal by using machine learning approach. Firstly, this research has used the Random Forest (RF) algorithm and then Balanced Random Forest (BRF) algorithm and lastly, they applied the XGBoost algorithm. The research has used Apache Spark for the implementation of the Random Forest algorithm. Interestingly, in the study, they have not confined themselves to use one dataset only, they have used three public datasets provided by the government of Canada and by the city of Montreal. The first one is the ‘Montreal Vehicle Collisions’ dataset that contains all the road collisions with important information such as location, date and time, details of the vehicles involved in the collisions and the severity of the accident. The next dataset is the ‘National Road Network’, it contains the information about the city’s road network including variables such as speed limits, road types and the presence of traffic signals and pedestrian crossings. Lastly, the ‘Historical Climate Dataset’ contains hourly weather information across Canada. The research has used some important variables from this dataset including temperature, precipitation, visibility and certain environmental variables. The research has applied the pre-processing on the dataset and saved the intermediate results on the disk for later use. They computed the feature importance by using the Random Forest algorithm and then removed those features that are not very important that have significantly improved the performance of their model. Then the research implemented the algorithms, developed the models and found out that the RF, BRF, XGB and base line model has 0.918, 0.916, 0.909 and 0.874 area under the ROC curve. This shows that RF and BRF have almost the same performance and perform better than the XGBoost and baseline model. The results show that RF has the best performance among all of these models, and it has successfully detected 85% of road collisions with a false positive rate of 13%. This model could be used to identify the most dangerous road segments every hour, and the necessary actions could be taken in order to avoid the risk of accidents.

Labib et al., (2019) conducted study to analyze and predict the severity of road accidents in Bangladesh using machine learning techniques. This research has used four supervised machine learning algorithms i.e. Decision Trees, K-Nearest Neighbors (KNN), naive Bayes and Adaptive Boosting (AdaBoost) because of their proven accuracy. The intensity of accidents has been classified into four categories in this paper that are fatal, grievous, simple injury and motor collision. They have used a traffic accident dataset of Bangladesh from 2001 to 2015 that contains 43000 records. 70% of this data has been used for training and the remaining 30% is used for testing. After the initial data collection and applying data pre-processing techniques, the research applied three different feature selection algorithms that includes Univariate Feature Selection, Recursive Feature Elimination and Feature Importance. Each of these techniques has given the top 15 features and this research has chosen the 11 most common. They have performed two experiments on the dataset and calculated the accuracy as 71%, 67%, 80%, 80%, precision as 68%, 68%, 63%, 68% and F1 score as 71%, 67%, 71%, 73% for DT, KNN, Naive Bayes and AdaBoost respectively. In the second experiment, the dataset divided into two classes called Fatal and Grievous, and found that the accuracy of DT and KNN increased. The performance of AdaBoost increases a lot as the F1 score and precision increase from 73% to 74% and 68% to 75% respectively. It is concluded from the above experiments that AdaBoost achieves the best performance among all of these machine learning techniques. In the end, they have done statistical analysis and found that as some features increase such as location type, movement, road geometry, light, weather and traffic control the number of traffic accidents also increases. Furthermore, it is found that buses and trucks are more involved in accidents, and the accident rate is high where no junction exists. When the surface condition is good, dry and sealed most of the accidents happen at that time. Lastly, the research has shown its intent to build a recommender system that can give the prediction of traffic accidents to make it more feasible.

Parathasarathy et al., (2019) develop several predictor models to predict the accident type. They have used machine learning approaches for their study that include K-Nearest Neighbor and Support Vector machines algorithms. The model proposed by them uses a combination of both KNN and SVM algorithms that have better accuracy as compared to the separate algorithms. Initially, the KNN algorithm was used to classify the dataset and predict the type of accident which was then along with other factors used to train the SVM algorithm. They obtained a dataset from the UCI Repository that contains 1000 records with 12 variables. The accident type is the target variable that is to be predicted by the model and it is divided into four classes: crash, drunk & drive, fire and skid. It is found that the KNN and SVM both give an accuracy of 60% whereas the hybrid approach used in this study gives an accuracy of 92%. So, it is concluded that the use of this hybrid approach proved more accurate in finding the accident class as compared to the individual use of the algorithm.

Kumeda et al., (2019)conducted research for the classification of road accidents using machine learning algorithms. The major goal of their research and this classification is to find out the major and key factors that cause traffic accidents, so that these factors can be used to form policies to reduce the traffic accident rate. They have talked about some crucial numbers that show the accident severity is unbalanced compared to the number of vehicles. They have used a dataset from UK road traffic accidents of the year 2016 that contains 12 attributes and 555 records. This research has divided the dataset and used 90% of the data for training and the remaining 10% for testing purposes. Road accidents are classified into three classes as fatal, serious and slight accident classes. They have applied 6 machine learning algorithms that include Fuzzy-FARCHD, Random Forest, Hierarchal LVQ, BRF Network, Multilayer Perceptron and Naïve Bayes on this dataset and obtained an accuracy of 85.94%, 83.42%, 82.16%, 84.14%, 79.27% and 80.90% respectively. It can be seen that the fuzzy-FARCHD classifier has obtained the top accuracy and further analysis shows that most of the causality’s ages lie between 20 to 40 years, most of them were male and the involved vehicle type is a car and mostly happens in daylight. It has been concluded that the Fuzzy-FARCHD algorithm can be applied to classify and predict the road accidents severity and its factors can be identified with the help of this model. Based on these factors and conclusions new policies can be made for accident prevention.

AlMamlook et al., (2019)established different models for classifying the injury severity during road accidents and to identify the factors behind traffic accident severity. This research has obtained the dataset from the Office of Highway Safety Planning (OHSP) that contains 271,563 road crashes that occurred in Michigan during the period 2010 to 2016. They have built several models using different machine learning algorithms and compared their performances to select the most suitable model. They have applied AdaBoost, Logistic Regression, Naïve Bayes (NB) and Random Forests (RF) algorithms on the data set. It has been concluded that the RF algorithm has achieved the best performance with an accuracy of 75.5% than LR with 74.5%, NB with 73.1% and AdaBoost with 74.5% accuracy. So, Random Forest has a good ability to resist noise as compared to other machine learning algorithms. The Area Under Curve (AUC) has also been calculated for all the algorithms and if it is greater than 0.7 it means this algorithm has an excellent prediction capability in pattern recognition. This is also statistical proof of the excellence of the RF algorithm used in the study. So, it is concluded by the research that RF seemed to be the best choice, and it could be implemented for prediction, monitoring and finding the factors of fatal injuries.

Gaurav and Zunaid Alam (2018) conducted study with the aim of reducing the road accidents in India. They state that according to a WHO report, it has been revealed that one serious accident occurs at every minute in India. So, this research has used the data mining approach of association rule on the dataset provided by National Highways Authority of India (NHAI) and analyzed it for finding the important patterns of accidents. The dataset contains the data from Panipat to Jalandhar. After applying different data preprocessing techniques and association rules algorithm it has been concluded that there is a high percentage of accidents in Punjab. Most of the accidents are injury based involving petrol vehicles as compared to other vehicles. The accident rate is high on four lane highways as compared to two lane highways. Late night and morning sessions are more susceptible to collisions. Over speeding, rear end collisions and collisions due to the fault of other drivers while overturning has strong confidence. Most of these accidents are only injury based and a very few serious accidents occur. So, it is suggested to avoid over speeding to further decrease the number of accidents.

Watkins E et al., (2017) have analyzed the traffic scenarios that could lead to fatal or major injuries or even deaths in their study. The main aim of their study was to identify the occurring patterns of fatalities and injuries to recommend some preventive measures that could help in reducing these occurrences. This research has used The National Collision Database (NCDB) provided by Transport Canada in this study for analysis. The dataset contains the data of collisions from 1999 to 2012. The research has cleaned and reduced the dataset in order to avoid skewness of the results. This research has used the data mining techniques of decision trees and neural networks in order to fulfill the research purposes by using Microsoft tools. The research has done various tests after data preprocessing and finds out that: If the driver's age is above 60 then there are six times more chances of incidents in which a person conceived death as compared to young drivers. A curved tunnel or underpass has 57.38% chances of single vehicle incidents. There are 81.60% chance of at least one fatality in collisions that happen at railway crossings where signs are only being used to control the traffic. It has also been found out that the new drivers feel nervousness when driving on a circular path or on express freeways that results in more collisions. The vehicle manufacturing year has an effect on the fatality rate, there is a general decrease of fatalities with the latest models. This shows that vehicle safety is advancing from time to time and their safety technology has a significant effect on the driver's safety too that leads to less chances of fatal injuries. Most of the collisions and fatalities happen at the areas where there is no traffic control and on highways due to increased speed limits. The recommendations given by the research are to make the tunnels and overpasses straight, addition of gates and proper light signals at railway crossings, and training of young drivers with regard to the rush hours. Spreading awareness among the people to use vehicles that have better safety measures.

After conducting a thorough literature review, it can be concluded that a significant amount of work has been done in the field of accident data mining, particularly in the domain of prediction models. However, most of the studies focus only on one domain i.e. building the prediction model or finding the factors of traffic accidents, only a few studies combining both of these aspects. The majority of these studies utilize machine learning methods on several different datasets containing attributes related to accidents such as year, month, time, road surface, weather, vehicle type, demographic information of the involvements with the injury severity level and many others. Most of the predictor models developed in the past focused only on the accident severity and their performances are very good. On the other hand, only a few research have focused on developing predictors that predict the injury severity level of the persons involved in the collision, but their performances are of average level.

This study aims to find out the best model that predicts the injury severity level of involvement and try to find out the associations between several factors that contribute to fatal or major injuries, which is a distinguishing factor of this study. The study will also focus on developing such a model that will perform better than the previous ones. Furthermore, the proposed model will be trained and tested against two large datasets of different countries. These datasets will be analyzed statistically, predictively and descriptively using several different machine learning algorithms to find the hidden patterns and extract the associations and factors that lead to severe collisions. Lastly, the study also intended to provide a comparison of these associations to understand the driving behavior and nature of accidents between these two countries. The study will also provide suggestions and recommendations to ensure the safety of vulnerable road users.

# CHAPTER 3

# MATERIALS AND METHODS

This chapter explains the full analytical procedure and methodology utilized to provide an injury severity predictor model. In the previous chapter, a review of literature contributing to various approaches to the research topic relevant to the design and development of an automatic injury severity predictor is described in depth. This chapter provides a thorough description of the specified system.

## 3.1 Dataset Description:

This section of the chapter describes the datasets used in the study, the dimensions of the datasets and the important features that are considered in the study with the targeted attribute. The study has used the two datasets described as follows:

### 3.1.1 Killed or Seriously Injured (KSI) Dataset:

The Toronto Police Services dataset on Killed or Seriously Injured (KSI) incidents (Toronto Police Services, 2023) provides valuable insights into public safety trends. This dataset contains all the traffic collision events that happened in Toronto from the year 2006 to 2022 that includes at least one person that is either seriously injured or killed. That dataset includes 57 variables and 18,194 rows. After performing preprocessing steps 16,578 rows and 45 features left. These variables can be classified as individual attributes and collision-related attributes.

The Individual attributes describe the characteristics and behavior of each individual involved in the collision such as involvement type, age, driver act and maneuver etc. The collision attributes describe the temporal, spatial and environmental conditions such as visibility, light and road surface conditions etc. Each row of the dataset represents an involvement type in the accident, these types include drivers, pedestrians, passengers, cyclists and motorcycle drivers etc.

Our focus in the study will be on the injury severity level of these involvements. The injury severity level has been classified as **Fatal, Major, Minor, Minimal and None**. The definitions of these levels are defined below:

**Fatal:** This class includes those body injuries that result in death. This class contains those cases in which death occurs as a result of the collision.

**Major:** These are non-fatal injuries but are severe enough that the person getting major injuries is required to be admitted to the hospital for treatment. The major injuries include fractures, burns and severe cuts etc.

**Minor:** These are injuries that do not require the involved person to be admitted to the hospital but can be treated in the emergency room of the hospital at the time of the collision.

**Minimal:** These include all of the non-fatal injuries that do not require an involved person to go to the hospital and can be treated on their own or none at all.

**None:** It includes all of those involved persons who do not get any of the above-mentioned injuries.

The final dataset consists of 43 attributes, Table 3.1 gives the description of these attributes.

**Table 3.1: Input features of the KSI dataset**

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Variable Type** |
| YEAR | Year in which collision occurs | Numerical |
| MONTH | Month in which collision occurs | Numerical |
| DATE | Date of collision occurrence | Numerical |
| DAY | Day of collision occurrence | Nominal |
| TIME | Time of collision occurrence | Numeric |
| ROAD CLASS | Road Classification | Categorical |
| DISTRICT | City District | Categorical |
| LOCCOORD | Location Coordinates | Categorical |
| ACCLOC | Collision Location | Categorical |
| TRAFFCTL | Traffic control type | Categorical |
| VISIBILITY | Environment condition | Nominal |
| LIGHT | Light condition | Nominal |
| RDSFCOND | Road surface condition | Nominal |
| ACCLASS | Class of accident | Ordinal |
| IMPACTYPE | Initial impact type | Categorical |
| INVTYPE | Role of the person involved in the collision | Categorical |
| INVAGE | The age range of the involved person | Nominal |
| INJURY TYPE | Level of injury the involvement sustains | Ordinal (Target Variable) |
| VEHTYPE | Type of Vehicle | Categorical |
| MANOEUVER | Vehicle Maneuver | Categorical |
| DRIVACT | Apparent Driver Action | Categorical |
| DRIVCOND | Driver Condition | Categorical |
| PEDTYPE | Pedestrian Crash Type - detail | Categorical |
| PEDACT | Pedestrian Action | Categorical |
| PEDCOND | Condition of Pedestrian | Categorical |
| CYCLISTYPE | Cyclist Crash Type - detail | Categorical |
| CYCACT | Cyclist Action | Categorical |
| CYCCOND | Cyclist Condition | Categorical |
| PEDESTRIAN | Pedestrian involvement in collision | Binary (Yes/No) |
| CYCLIST | Cyclist involvement in collision | Binary (Yes/No) |
| AUTOMOBILE | Automobile involvement in collision | Binary (Yes/No) |
| MOTORCYCLE | Motorcycle involvement in collision | Binary (Yes/No) |
| TRUCK | Truck involvement in collision | Binary (Yes/No) |
| TRSN\_CITY\_VEH | Transit or City Vehicle Involved in Collision | Binary (Yes/No) |
| EMERG\_VEH | Emergency Vehicle Involved in Collision | Binary (Yes/No) |
| PASSENGER | Passenger Involved in Collision | Binary (Yes/No) |
| SPEEDING | Speeding Related Collision | Binary (Yes/No) |
| AG\_DRIV | Aggressive and Distracted Driving Collision | Binary (Yes/No) |
| REDLIGHT | Red Light Related Collision | Binary (Yes/No) |
| ALCOHOL | Alcohol Related Collision | Binary (Yes/No) |
| DISABILITY | Medical or Physical Disability Related Collision | Binary (Yes/No) |

### 

### 3.1.2 Road Traffic Accident (RTA) Dataset:

The study has also incorporated the local dataset, the city of Rawalpindi has been selected as the study area. The researcher has collected the road traffic accident dataset from the District Office of Punjab Emergency Services, Rescue 1122 (RTA, 2023). They track the data of all the emergencies including the road accidents within the District Rawalpindi. Their dataset contains all the traffic collisions from January 2020 to July 2023. The dataset includes 25 variables and 46,189 records, where each record is an entry for one person involved in the accident. One deficiency of this dataset is that it lacks one important attribute of weather conditions. From previous studies, it has been concluded that it is an important feature and plays its role in the collision event. So, the weather data has been collected separately from the NASA website and integrated with this dataset. After performing preprocessing and feature extraction steps 46,187 rows and 29 features left.

The attributes contain the data about the demographic information of the victim such as gender, age and education and the collision-related data including the date and time of the collision, location of the collision, reason, injury type, patient status and the type of vehicle involved in the collision. The targeted feature is the type of injury that has been classified as Minor, Single Fracture, Multiple Fractures, Spinal Injury and Head Injury. The final dataset consists of 27 attributes, table 3.2 gives the description of these attributes.

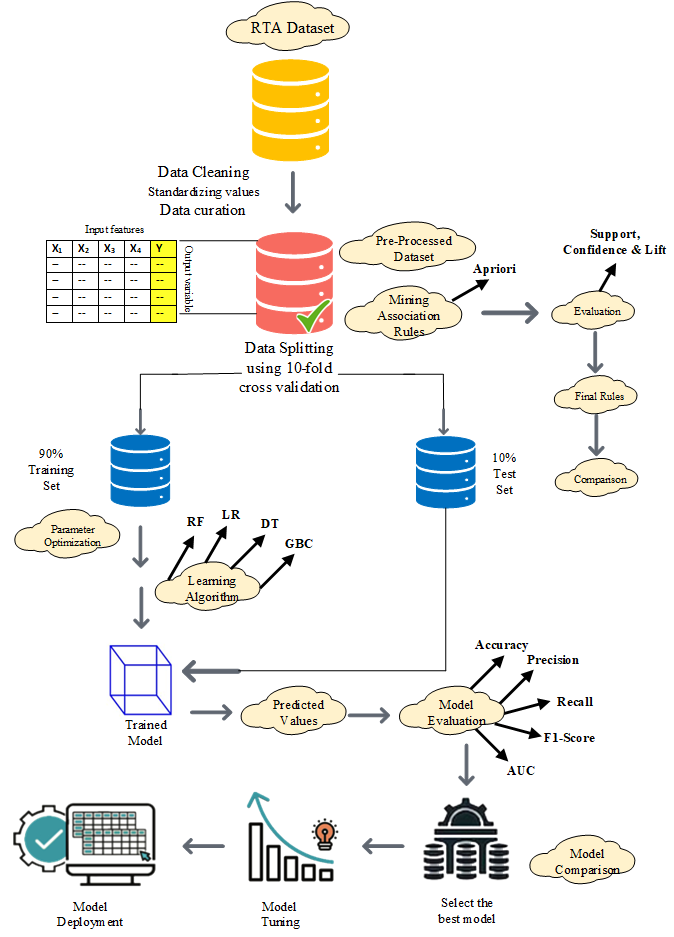
**Table 3.2: Input features of the RTA dataset**

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Variable Type** |
| Year | Year in which collision occurs | Numerical |
| Month | Year in which collision occurs | Numerical |
| Date | Year in which collision occurs | Numerical |
| Day | Day of collision | Nominal |
| Time | Time of collision | Nominal |
| Nature of Weekday | Specifies working day or weekend in which collision occurs | Nominal |
| Season | Season in which collision occurs | Nominal |
| Light | Lighting condition during a collision | Nominal |
| Weather | Weather conditions | Nominal |
| Total Patients | No of patients involved in a collision | Numerical |
| Gender | Victim’s gender | Nominal |
| Age Range | Victim’s age | Nominal |
| Education | Victims’ education level | Nominal |
| Acclass | Class of accident | Ordinal |
| Injury Type | Level of injury severity | Ordinal (Target Variable) |
| Cause | Reason of accident | Nominal |
| Patient Status | Condition of victim | Ordinal |
| Bicycle Involved | Bicycle involvement in collision | Binary (Yes/No) |
| Bikes Involved | Bikes involvement in collision | Binary (Yes/No) |
| Buses Involved | Buses involvement in collision | Binary (Yes/No) |
| Cars Involved | Cars involvement in collision | Binary (Yes/No) |
| Cart Involved | Cart involvement in collision | Binary (Yes/No) |
| Rickshaws Involved | Rickshaws involvement in collision | Binary (Yes/No) |
| Tractor Involved | Tractor involvement in collision | Binary (Yes/No) |
| Trains Involved | Trains involvement in collision | Binary (Yes/No) |
| Trucks Involved | Trucks involvement in collision | Binary (Yes/No) |
| Vans Involved | Vans involvement in collision | Binary (Yes/No) |

## 3.2 Proposed Framework:

This work proposes a novel framework for analyzing road traffic accident dataset. It consists of two phases: the first one is of making prediction models to predict the injury severity of involved ones in the collision and the second phase consists of finding the factors that lead to severe collisions using ML algorithms. By integrating these modules, the framework aims to achieve improved accuracy of the models and more interpretability of the dataset compared to existing methods. The proposed methodology for this study has been divided into five steps as follows:

The first step is data collection, in this step we will try to obtain dataset from different online sources and also from local departments. The study aims to utilize the ML techniques on two different datasets and provides a comparison of them. The second step is of data preprocessing, in this, the study will use several techniques including data cleansing, feature engineering, feature selection and normalization to remove the noisiness in the datasets. Furthermore, the class imbalance issue will also be addressed using SMOTE technique in this step. The third step is finding the association rules, the study intended to use the apriori algorithm in order to find out the hidden pattens and association among several factors that leads to severe collisions along with the comprehensive comparison. Next, the fourth step is of data splitting and model training, in this step, the study will develop several ML based predictor models that will predict the injury severity level. Lastly, the fifth step consists of model evaluation, in this step the developed models will be compared against some performance measures including accuracy, recall, precision, f1 score and AUC. The flow chart of the research design is shown below:



**Fig 3.1 Proposed Framework of the Study**

#### 3.2.1 Data Collection:

The collection of accurate and reliable data is a very crucial task in any study. If the data is not accurate then the results of the study would be faulty and misleading. There is always a need for special attention in understanding the contents and structure of the dataset.

Two datasets have been used in this study, the first one is the Killed or Seriously Injured (KSI) dataset obtained from the Public Safety Data Portal provided by Toronto Police Services (Toronto Police Services, 2023). This dataset contains all of the traffic collisions that happened from 2006 to 2022. The second dataset used in the study is collected from the District Emergency Office (Rescue 1122) Rawalpindi (RTA, 2023). This dataset contains all the traffic accident events that happened in the District Rawalpindi from January 2020 to July 2023.

Both of these datasets are enriched with important features such as demographic details of involvements, features related to accidents and several environmental variables.

#### 3.2.2 Data pre-processing:

Data preprocessing is the foremost step in the data mining approach that is done after collecting a dataset. Preprocessing the dataset takes 70% of the time to make the dataset useful, as per the requirement of algorithms. The goal of data preprocessing is to investigate the dataset to remove the noisy and erroneous data that could make the algorithms application useless to mine the data.

The following operations have been performed to make the data more accurate so that it can be used efficiently for our study purposes.

##### 3.2.2.1 Data cleansing:

This process helps us to clean the data by removing the incomplete, unnecessary and redundant data that is not useful for our study. The most common issue that comes during the process of knowledge discovery through data mining includes missing values. A dataset with 1–5% missing values may not impact as much, and from 5–15% range requires sensitive algorithms to apply (Saleem et al., 2014).

In this step, the missing values were replaced with the average or possible values. Furthermore, eliminated columns having a significant number of missing values or negligible contributions to the analysis.

In both of the datasets the targeted feature i.e. ‘INJURY TYPE’ contains some null entries, these entries have been removed from the analysis.

##### 3.2.2.2 Feature Engineering and Transformation:

The study has performed feature engineering to extract some additional information from the dataset that might be useful in finding out the hidden patterns. The Date, Month, Year, Time, Day and Weekday information has been extracted from the timestamp variable in both datasets. The season attribute has been derived from the month attribute in the RTA dataset. The study has also performed the type conversion where it is necessary such as the conversion of string data into nominal form so that the necessary algorithms can be applied on the dataset.

##### 3.2.2.3 Feature Selection:

Sometimes, our dataset contains a lot of extra information or features that are not relevant to our study. These extra features unnecessarily increase the size of our dataset and that results in slow processing of the algorithms. So, it is required to select only the required features so that the relevant and appropriate outcomes can be generated.

In this study, the feature selection technique has also been implemented to reduce the dimensionality of the dataset and some of the extra attributes including ACCNUM, INDEX, FATAL\_NO, DIVISION, HOOD\_140, HOOD\_158, NEIGHBOURHOOD\_158, NEIGHBOURHOOD\_140, and OBJECT\_ID have been removed from the KSI dataset as these are not related to our analysis. Similarly, the features of EcNumber, CallTime, EmergencyArea, Responsetime and Reason have been dropped from the RTA dataset.

##### 3.2.2.4 Class Imbalance Problem:

Sampling is a commonly employed technique for addressing a limited number of instances of targeted variable or class imbalance issues of datasets. After looking deep into the derived datasets, it is revealed that the data distribution of the targeted variable i.e. injury severity is highly imbalanced in both of our datasets.

In the KSI dataset, the number of fatal, minor and minimal classes of injury severity are far less (minority class) than the number of none and major classes (majority class). Similarly in the RTA dataset the injury severity of ‘minor’ is the major class and the remaining values of single fracture, multiple fractures, head injury and spinal injury are the minority classes.

To address such instances, the study choses the Synthetic Minority Oversampling Technique (SMOTE) as a sampling methodology, given its widespread presence in recent literature about similar research. The dataset used in the is a multiclass dataset that contains categorical data. So, SMOTE is the most suitable method to apply in our specific use case.SMOTE creates artificial samples of the minority class to make the amount of that class more even with the majority class. Then the predictor models will be developed based on this balanced class.

##### 3.2.2.5 Normalization:

The data normalization ensures that our features are on a similar scale so that our machine-learning models can make quick and accurate predictions. In this step, some of the attributes of the datasets are normalized that have transformed the domain of these attributes into a given range on the training set. The age feature has been normalized in the RTA dataset.

#### 3.2.3 Data Splitting:

Next, the study has divided the dataset into two parts, i.e. one for training purposes and the remaining one for testing our models. In this study, the k-fold cross-validation method has been employed. This technique divides the dataset into ‘k’ equal parts and iteratively uses the k-1 instances for training the model and the leftover kth chunk will be used for testing this model. This process is done k-times and at the end, the average result values are to be returned.

For example, if the value of k is chosen as 10, the 10-fold cross-validation technique divides the datasets into two parts such that 90% of the data will be used for model training and the remaining 10% will be used for model evaluation iteratively.

#### 3.2.4 Model training:

The phase of model training is the primary step in which the algorithm acquires knowledge of patterns and correlations in the data in order to generate predictions or to do the correct classifications. In this stage, the selected algorithm is provided with labelled data including the targeted attribute, so that the iterative adjustment of its internal parameters can be made through a process called optimization. The algorithm strives to get optimal parameter values that precisely represent the basic patterns in the data.

The training process usually consists of a number of epochs, with each epoch representing a full iteration over the complete training dataset. While iterating over the data, the algorithm improves its internal representations in order to more effectively apply to cases that have not been seen before. The training process comes to an end when the training process meets a predetermined terminating condition, Next, it is ready for evaluation or predicting the new, unseen data.

#### 3.2.5 Model Evaluation:

To measure the performance of classification model different metrics are used. This study has utilized the metrics of accuracy, recall, precision, F-1 score, Area Under the ROC Curve (AUC – ROC), MCC and Kappa measure for the assessment of predictive models.

To assess the performance, the study utilizes a confusion matrix, which compares actual outcomes with predicted ones from our models. This matrix reveals the frequency of correct and incorrect predictions made by each algorithm. There are four viable outcomes of the prediction as follows:

**True Positive (TP):** The Model predicts positive and it’s a true prediction.

**True Negative (TN):** The model predicts negative and it’s a true prediction.

**False Positive (FP):** The model predicts positive and it’s a false prediction.

**False Negative (FN):** The model predicts negative and it’s a false prediction.

##### 3.2.5.1 Accuracy:

Accuracy serves as a crucial metric for evaluating the effectiveness of a machine learning model. It measures the proportion of correct predictions (TP+TN) out of the total instances (TP+TN+FP+FN). This is particularly valuable in classification tasks, where each instance is assigned a distinct label. (Shanshal et al., 2020 and Kumeda et al., 2019).

Accuracy can be calculated in terms of the confusion matrix by using the following formula:

The formula above yields an accuracy value ranging from 0 to 1. To express accuracy as a percentage, it can be multiplied by 100. While accuracy provides a straightforward assessment of a model's performance, it may not suffice if the dataset distribution is uneven. Therefore, it's crucial to complement accuracy with other metrics such as precision, recall, and F1 score.

##### 3.2.5.2 Recall:

Recall is also an important measure that is used for evaluation purposes, it portrays a true positive rate and is also named as sensitivity. It measures the true positive (*TP*) predictions from all the actual positive instances (*TP+FN*) (Shanshal et al., 2020 and Kumeda et al., 2019).

Recall can be calculated in terms of a confusion matrix by using the following formula:

The formula provided above yields a recall value ranging from 0 to 1. To express recall as a percentage, simply multiply it by 100. Recall is particularly useful when aiming to reduce or eliminate false negatives. It aids in assessing the model's ability to identify positive items. A high recall vale indicates the model's proficiency in detecting numerous positive items, consequently minimizing false negatives.

##### 3.2.5.3 Precision:

Precision is a crucial metric that has been extensively used to evaluate the performance of machine learning models that work as classifiers. It measures the ratio between actual positive predictions (*TP*) and the overall positive predictions made by the predictor (*TP+FP*) (Kumeda et al., 2019).

Precision can be calculated in terms of confusion matrix by using the following formula:

The above formula gives the value of precision that is between 0 and 1, to get the percentage value, can be obtained by multiplying it with 100. Precision is very useful in situations where the aim is to minimize the false positives. It gives us an idea of how successful the model is at identifying positive cases without mistakenly labelling negative ones as positive. A high precision rate means the model has a low rate of false positives, which makes the predictions more consistent.

##### 3.2.5.4 F1-Score:

The F1 score, widely used in assessing machine learning models, particularly in classifications, provides a balanced and comprehensive evaluation of both precision and recall. It offers valuable insight into the overall effectiveness of the model (Kumeda et al., 2019).

F1-Score can be calculated by using the following formula:

The F1-Score ranges from 0 to 1 based on the formula provided, which can be converted to a percentage by multiplying by 100. A higher F1-Score indicates better model performance. In scenarios of highly imbalanced classes and uneven dataset distributions, relying solely on accuracy or recall may not yield meaningful results. In such cases, the F1-Score serves as a valuable metric, offering a balanced assessment by considering both precision (false positives) and recall (false negatives) simultaneously.

##### 3.2.5.5 Area Under the Curve (AUC):

The AUC is a popular tool used for assessing the performance of machine learning models specifically in binary classification. It measures how well the model can differentiate between the two different classes. This measure usually ranks the instances by calculating the probabilities for each given case (AlMamlook et al., 2019).

The Receiver Operating Characteristic (ROC) curve is used to calculate the Area Under Curve (AUC). This plot shows the relation between sensitivity (TPR) and specificity (FPR) for different thresholds of classification. Each point on the curve represents a unique setting, and the area underneath is the AUC. The ideal value of AUC for any classifier is 1, the higher the AUC value the better it performs. If the AUC value is 0, it means the model is predicting completely opposite. The acceptable value of AUC is between 0.5 and 1. A model having an AUC of less than 0.5 is not very accurate and not considered credible. The AUC metric is useful because it is not affected by the threshold chosen, and it can be applied easily to imbalanced datasets without any specific concern because it takes into account the overall grading of objects (AlMamlook et al., 2019).

#### 3.2.6 ML Techniques

This research explores the complex realm of road traffic accidents, employing predictive and descriptive data mining approaches to achieve a two-fold objective: gaining a thorough grasp of predictive modeling and uncovering patterns. Collectively, these machine learning methods play a crucial role in understanding the complexities of road safety, providing a comprehensive viewpoint.

##### 3.2.6.1 Classification

Classification is a core method in supervised learning that aims to give predetermined labels or groups to occurrences based on their attributes. These classification algorithms analyzed historical data to identify trends that might assist in predicting injury severity (Emu et al., 2022 and AlMamlook et al., 2019). By conducting detailed analysis of factors such as vehicle types, environmental conditions, and human factors, these algorithms help create predictive models that enable stakeholders to anticipate and reduce potential risks of serious injuries on roads. Some of the most commonly used and popular classification algorithms are ET, RF, GB, XGBoost, Ridge, LDA, DT, Light GB and AdaBoost that are explained below.

##### 3.2.6.1.1 Extra Trees Classifier

The Extra Trees Classifier is an ensemble learning algorithm, typically used to manage high-dimensional data and reduce computational overhead while producing accurate predictions. The algorithm works by randomly selecting parts of the training data called bootstrapping, then constructing decision trees for each subset by dividing the data based on numerous features, using arbitrary thresholds instead of probing for the perfect split point. Final predictions are made by combining the forecasts of all individual trees, usually through majority voting in classification tasks (Umer et al., 2020).

##### 3.2.6.1.2 Random Forest Classifier

Random Forest is a popular machine-learning algorithm proposed by Leo Breiman and Adele Cutler. It can be employed for both classification as well as regression. Random Forest Algorithm is a form of ensemble learning that operates by randomly selecting a subgroup of the training data (bootstrapping), to generate various subdivisions. These subsets are then used to construct individual decision trees by recursively splitting the data according to different features. The final prediction is made by aggregating the predictions of all the individual trees, either by selecting the most popular class in the case of classification or by taking the average/median of the predicted values in the case of regression. This aggregation produces more accurate and reliable results (AlMamlook et al., 2019). The key features of Random Forest include the handling of missing values by propagating them down to the tree, outlier detection and the highlighting of the important features that show which of the features are most contributing to the predictions (Ahammad et al., 2023).

##### 3.2.6.1.3 Gradient Boosting Classifier

The Gradient Boosting Classifier was founded by Leo Breiman and has become very popular due to its efficiency and ability to solve a variety of machine-learning problems.

It works initially by creating a weak model that is usually created using DT or LR algorithms, and then its error is calculated. This process of creating a weaker model continues until the desired performance is achieved. It is known in the world of data science due to its power, performance and versatility. It is capable of handling large datasets, multiple data types and imbalance classes efficiently (Ahammad et al., 2023).

##### 3.2.6.1.4 Extreme Gradient Boosting

XGBoost is an enhanced version of the Gradient Boosting Classifier created by Tianqi Chen. XGBoost utilizes a more efficient and optimized algorithm than GBC, it is able to manage larger datasets more efficiently. It also offers several regularization techniques, such as L1 and L2 regularization, to reduce overfitting. XGBoost can handle missing values and employs an improved level-wise tree production method. Additionally, XGBoost provides advanced regularization techniques such as dropout and stochastic gradient boosting. Furthermore, XGBoost is speedier and more scalable than GBC and includes a pre-built cross-validation framework. These features make XGBoost a popular and powerful tool for machine learning tasks (Chen et al., 2016).

##### 3.2.6.1.5 Ridge Classifier

The Ridge Classifier is an effective machine learning algorithm that belongs to the linear classifiers. It is an extended version of linear regression that includes a Ridge regularization also known as L2 regularization. The Ridge Classifier is usually used for binary classifications based on several features. The Ridge Classifier starts its work from data preparation that requires labelled samples with numerical or encoded features, and then feature scaling is done. During the Ridge regularization, the complexity of the model is controlled, and optimization is done by iteratively fitting the coefficients to decrease the cost function. After optimization, tuning of the regularization parameter is done using cross-validation or other methods. Then a decision boundary is defined, and lastly, the predictions are made that are based on the placement of samples across the defined boundary.

##### 3.2.6.1.6 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a powerful learning and pattern recognition technique that reduces the dimensionality of data and is used for classification tasks. Its goal is to create a set of linear features that maximally separate different classes in the dataset. It is assumed that the data follows a normal distribution and that the classes have equal variance. The process consists of preprocessing the data, reducing the dimensionality, calculating the mean vectors and scatter matrices, performing eigenvalue decomposition, projecting the data onto the discriminant axes, and finally classifying the data using a linear or support vector machine (Das et al., 2020).

##### 3.2.6.1.7 Decision Tree Classifier

The Decision Tree Classifier is an extensively used machine learning algorithm that is commonly employed for both classification and regression tasks. It employs a non-parametric method and works by partitioning the data into sections based on previously defined rules. It creates a structure that resembles a tree, where each internal node is a decision based on a feature, and each leaf node is a predicted class label or value. The algorithm works by first selecting a feature to split the data, then splitting the data into subgroups based on the feature values. This procedure continues until a specific condition is met. After this, leaf nodes are assigned class labels or values, and when new instances are input, the algorithm follows previously defined rules to traverse the tree and output a calculated value (Meißner, Katherina, 2022).

##### 3.2.6.1.8 Light Gradient Boosting Machine

LightGBM is a strong machine learning algorithm commonly employed for both classification and regression problems. It is a framework belonging to the framework for gradient boosting. It is optimized especially for distributed but efficient, and extremely fast computing. LightGBM repeatedly builds an ensemble of decision trees, with each tree aimed at rectifying the faults made by its predecessor. The learning process is optimized using a gradient-based technique, with an emphasis on specific erroneous instances. The method is renowned for its efficiency and scalability, rendering it well-suited for larger datasets. LightGBM excels in dealing with categorical attributes and is known for its capability to manage datasets that are imbalanced (Umer et al., 2020).

##### 3.2.6.1.9 AdaBoost Classifier

AdaBoost, also known as Adaptive Boosting, is a method of ensemble learning that integrates the predictions made by weak learners in order to produce a strong and precise model. Insufficient learners, which are usually decision trees, are trained in a sequential manner on the data. Each succeeding tree assigns greater importance to the cases that were incorrectly identified by the prior trees. By employing this adaptive method, AdaBoost is able to prioritize the cases that are challenging to categorize, hence enhancing its overall efficiency. The ultimate predictions are a computed aggregation of the results obtained from those who are weak learners. AdaBoost is praised for its adaptability and efficacy in increasing the performance of different base learners, thus making it an appealing option for many different kinds of problems related to classification (Umer et al., 2020).

##### 3.2.6.2 Association Rule Mining

This study has also applied the descriptive data mining approach of machine learning known as association rule mining to find out the hidden patterns in a high volume of data. Association rule mining is a type of rule-based machine learning algorithm that aims to discover interesting relations between items in large databases (Piatetsky-Shapiro et al., 1991). The idea is to produce rules that can predict the occurrence of an item based on occurrences of other items (Antonie and Luiza, 2008). Agrawal et al., (1993) provide a more formal definition of an association rule given as follows:

Let *I = {i1, i2, …, in}* be a set of n attributes and *T = {t1, t2, …, tn}* be a set of m transactions forming a database. Each transaction ti includes a subset of the items available in *I*. A rule can be defined as X => Y where X and Y ⊆ I. X is referred to as antecedent and Y is referred to as consequent of the rule respectively. Simply, a rule is a predictable transaction within the database.

Association rules can be useful in traffic collision data analysis as they can detect the correlation between different features within the dataset. In particular, they can be used to correlate the features that contribute most to traffic accidents. Furthermore, the rules can also elaborate the conditions that lead to fatal accidents or those that cause major injuries in the involvements. Association rules are often considered for those applications that have nominal datasets. Several association rules algorithms have been developed and the study has used the Apriori algorithm because it can be easily implemented on large datasets properly.

###### 3.2.6.2.1 Apriori Algorithm:

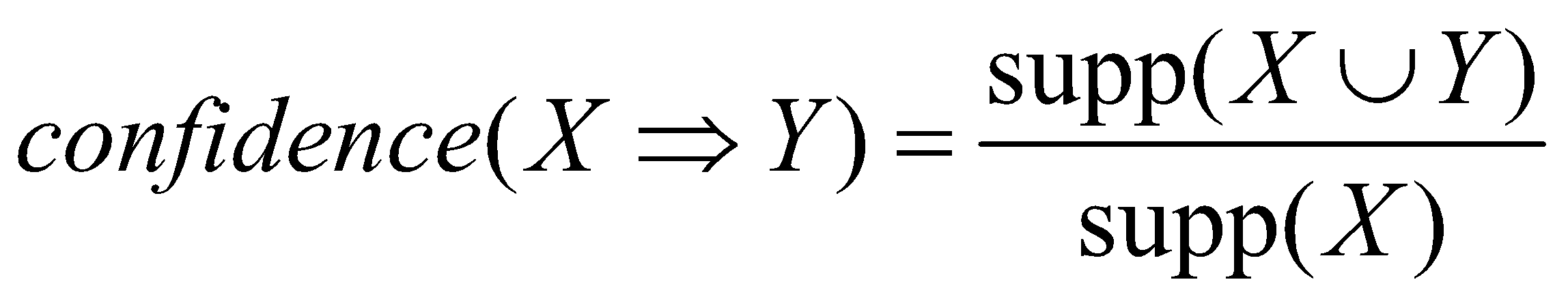
The Apriori algorithm is a BFS-based algorithm that depends upon the frequency of the item-sets to identify a set of association rules (Agrawal et al., 1994). This algorithm requires a minimum threshold *N* that is given by the user, and it tries to find out those item-sets that appear in at least *N* transactions within the database. It adopts an *N* time ‘bottom-up’ approach, in which a frequent subset is extended by one item in each iteration. This means that the algorithm starts with item-sets of length 1 (i.e. only one item within the itemset) and determines the itemset that has a frequency higher than the considered threshold *N*. This is repeated until no newer frequent item-sets are found. The length is then incremented by 1 and the same process is adopted again. This continues until there are no more possible extensions of the item-sets.

Moubayed et al., (2048) states that the apriori algorithm is popular in data mining due to the reason of its simplicity and easy applicability as well as its usefulness on large datasets. However, this algorithm may work slowly because it requires several dataset scans in order to produce rules. Its speed depends upon the size of the dataset, the number of data items in it and the choice of minimum measures (Antonie et al., 2008).

For an association rule to be of interest, several measures have been proposed, some of these have been discussed below:

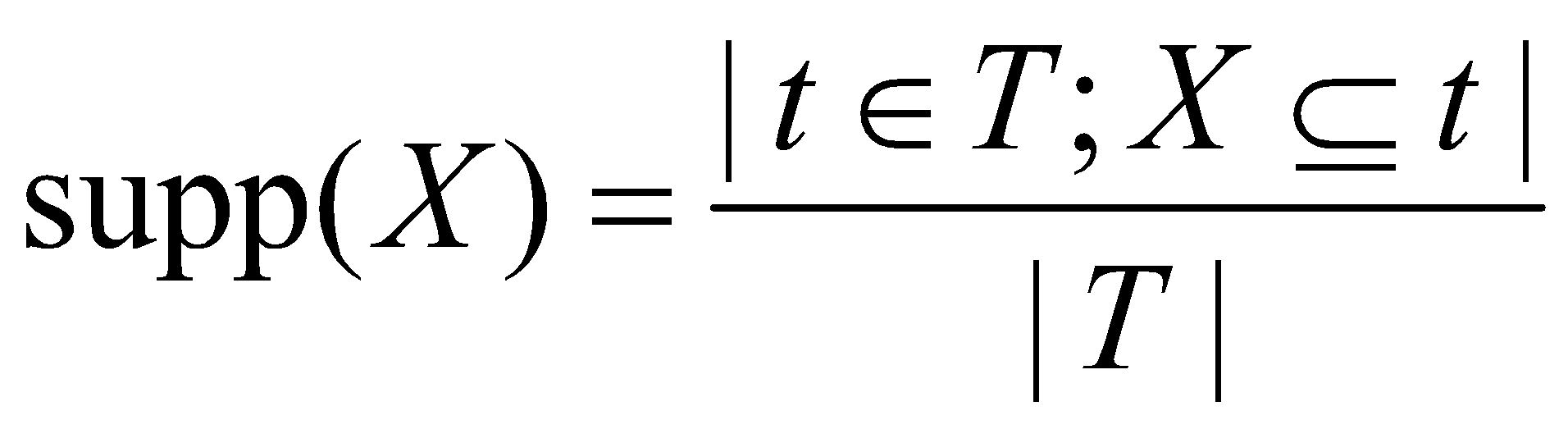
**(i) Confidence:**

It measures the frequency of the rule occurrence within the database (Alam et al., 2018 and KUYUMCU et al., 2023). It is defined as follows:

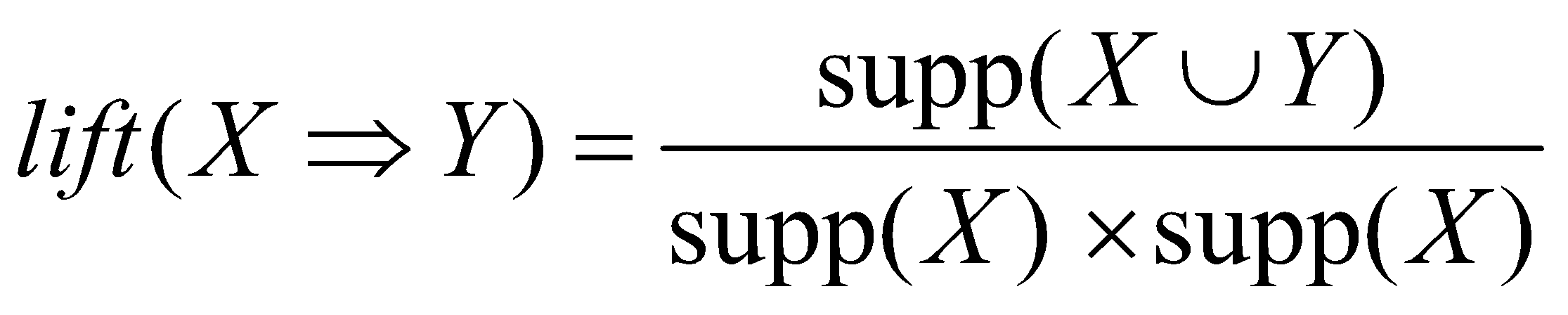


Hipp et al., (2000) states that the confidence of a rule can be thought of as a conditional probability of the rule.

**(ii) Support:** It indicates the frequency of occurrence of an item in the transaction (Alam et al., 2018 and KUYUMCU et al., 2023). Formally, the support of a data item X with respect to a set of transactions T is calculated as follows:



**(iii) Lift:** It determines the probability of the rule occurring relative to the probability of the antecedent and consequent being independent. It is calculated as follows (KUYUMCU et al., 2023 and Brin et al., 1997):



It is a measure of the interestingness of rules. The lift value >1 indicates that the two itemset are dependent on each other and form a true rule that can be useful in the future prediction of consequences given the defined antecedents. The lift measure is more significant as it takes into account both the confidence of the rule as well as the overall transaction database (Hahsler et al., 2005).

# CHAPTER 4

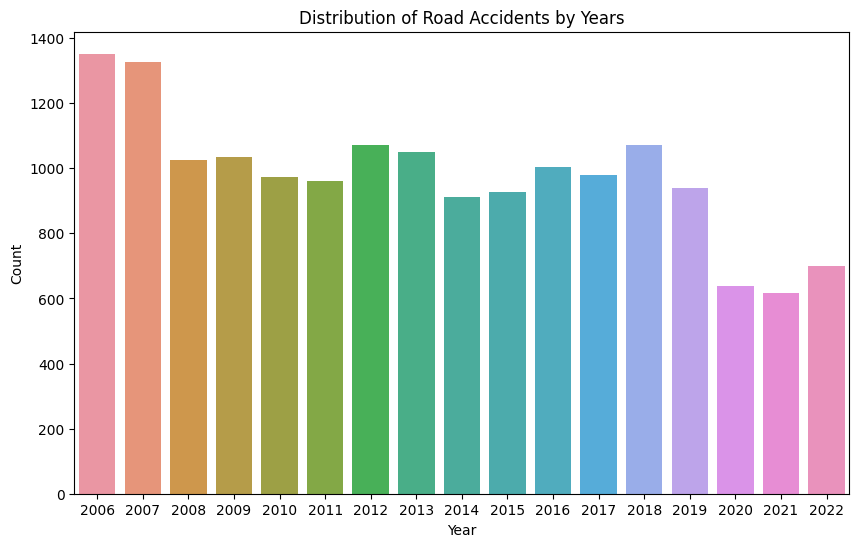
# RESULTS AND DISCUSSION

In this chapter of the study, the prediction of injury severity level of involvement was carried out using various features related to accidents. Both predictive and descriptive data mining tasks were performed to address the research questions of this study, utilizing datasets from KSI and RTA. The analysis involved the implementation of several ML classifiers and association rule mining techniques, as described below.

## 4.1 Visualization

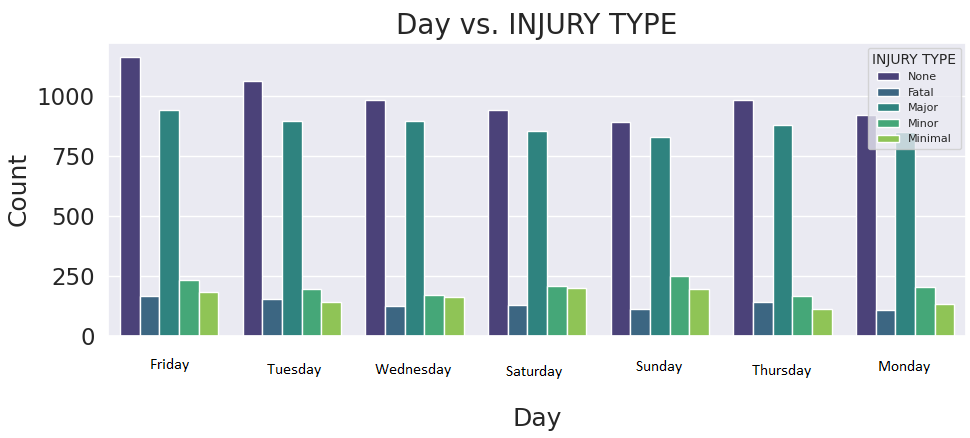
The study has utilized visualization approaches to elucidate the significance of crucial input variables in predicting degrees of injury severity and to examine our datasets in a more concise manner. These illustrations act as concise narratives, revealing the relationships between different data items or trends over time. Through this visual exploration, the study gives a simple yet comprehensive understanding of the dynamic interplay between input factors and accident severity.

Firstly, the study has analyzed the KSI dataset, the target variable has been analyzed against different input features. Fig 4.1 shows that the overall trend of accidents has been reduced over the years.



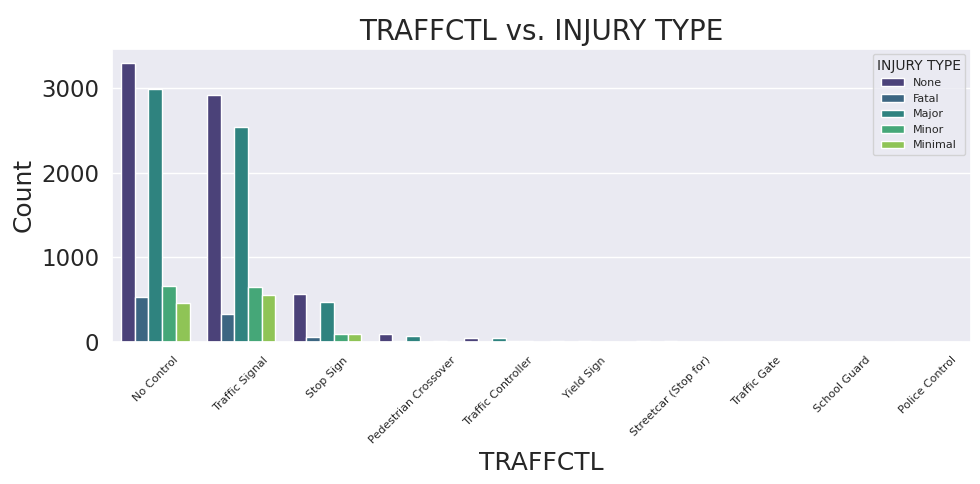
**Fig 4.1 Number of Collisions Happened over 17 Years (2006-2002)**

Upon delving deeper into the dataset, a distinct pattern emerges, highlighting Fridays as the day with the highest frequency of collisions, particularly those resulting in Fatal or Major injuries. In contrast, Sundays exhibit the lowest ratio in this regard. Fig 4.2 represents this observation and underscores the significance of temporal factors in influencing the severity of traffic accidents, with the end of the workweek seemingly contributing to a heightened risk scenario.



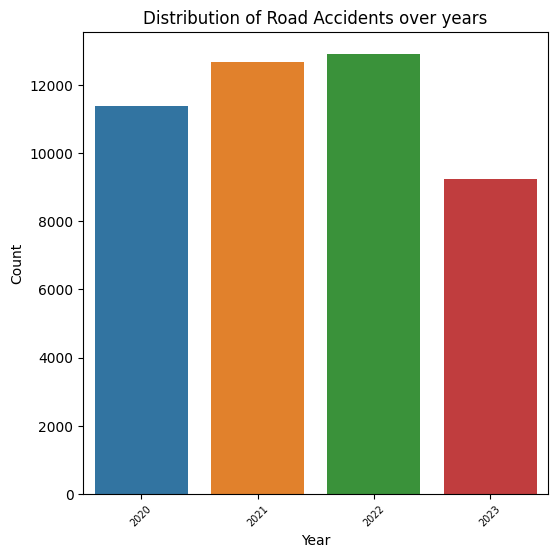
**Fig 4.2 Frequency of collisions and the injury type against weekday**

A noteworthy revelation within the dataset given in Fig 4.3 depicts that a significant portion of collisions resulting in Fatal or Major injuries occur regardless of the presence or absence of traffic signals. This prompts a critical examination of the contributing factors beyond signal regulation, suggesting that diverse elements may play a pivotal role in the severity of accidents.



**Fig 4.3 Frequency of collisions and the injury severity against Traffic Control Signal**

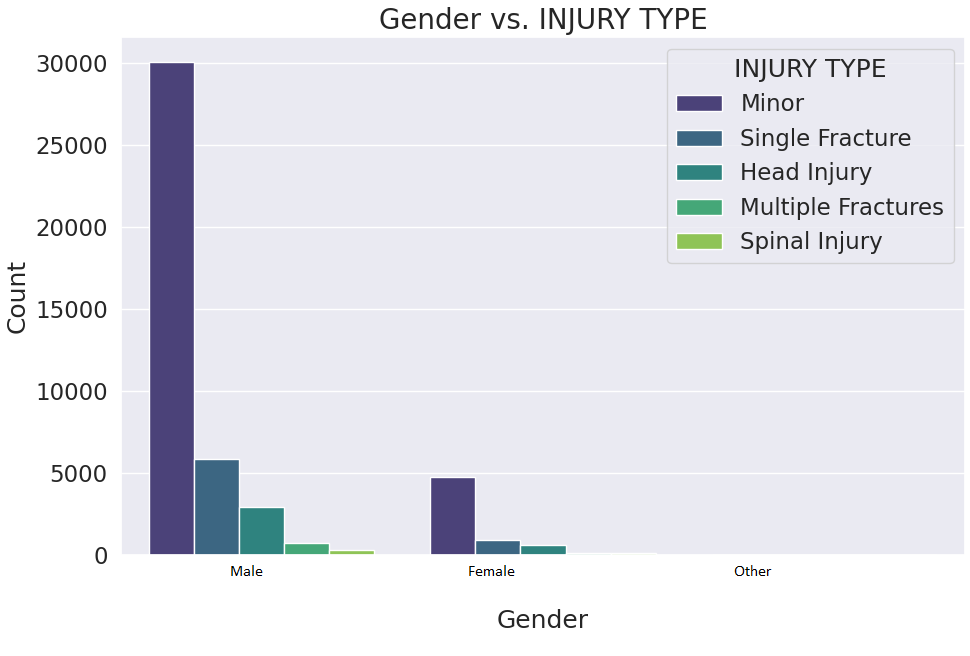
In analyzing the Road Traffic Accident (RTA) dataset spanning from 2020 to July 2023, fig 4.4 shows a clear pattern that becomes apparent when looking at the yearly frequency of accidents. Between 2020 and 2022, there is a clear and worrisome increase in the frequency of accidents, suggesting a significant rise in road occurrences throughout this time frame.



**Fig 4.4 Frequency of collisions in RTA over years (2020-July 2023)**

The data indicates that each subsequent year had a greater number of collisions than the previous year, implying a possible worsening of road safety issues throughout this period. Since the information only covers up until July 2023, the yearly results for that year are still incomplete and may not provide an accurate representation of the general pattern.

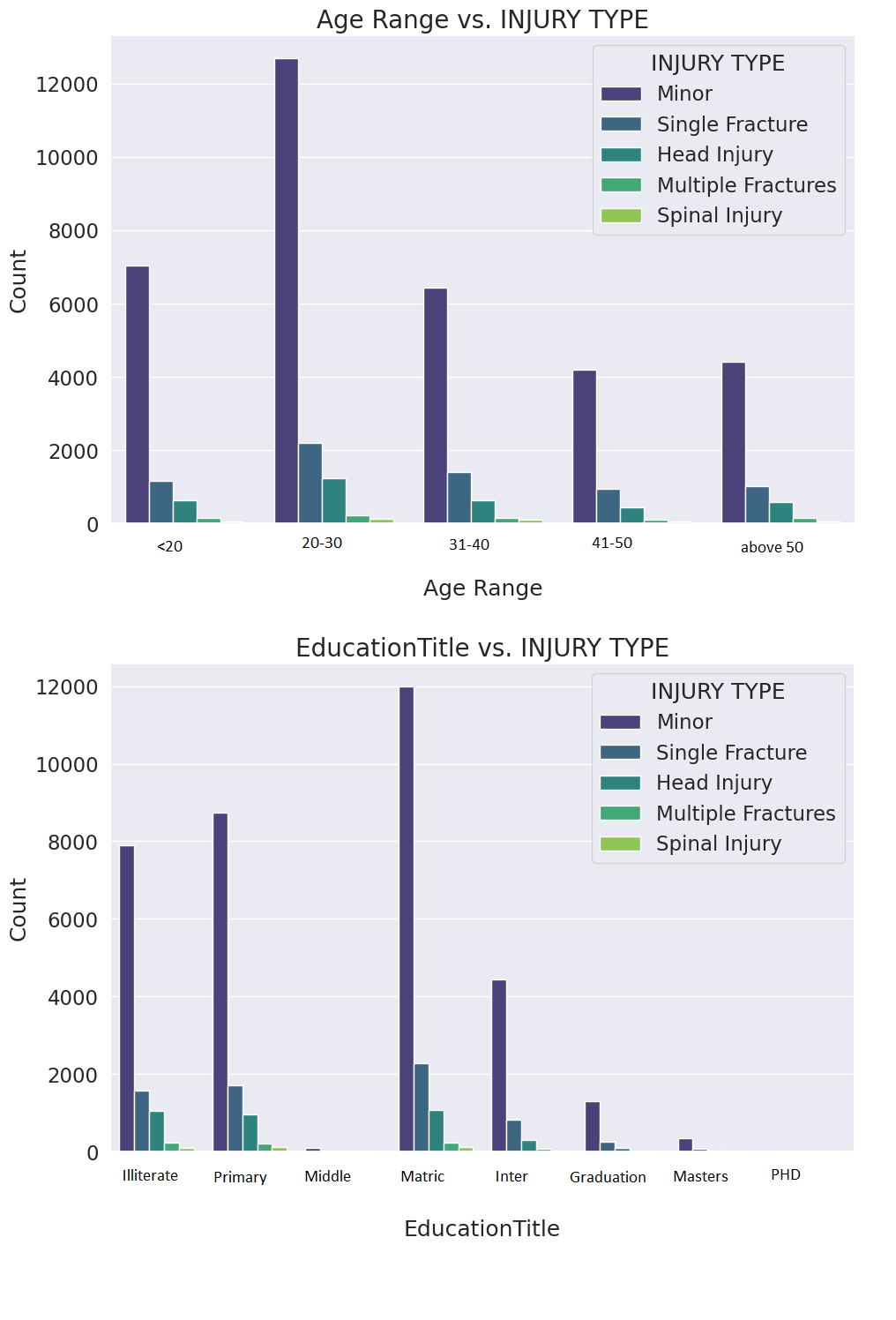
When examining the gender-specific aspects of injury severity in road traffic accidents, an obvious pattern can be seen from the graph shown in fig 4.5. The significant predominance of males in both the occurrence of accidents and the severity of injuries incurred is a convincing depiction of gender-related trends in road safety occurrences.



**Fig 4.5 Frequency of collisions in RTA against Gender**

The significant majority of males who suffer from small injuries and single fractures highlight the necessity for specific procedures and focused awareness campaigns. This finding stimulates a thorough investigation into the possible elements that are influencing these trends.

Fig 4.6 represents an analysis of the correlation between age and road traffic accidents, it indicates a significant concentration of collisions and injuries among those having an age range between 20-30. This particular age group is found to be the most vulnerable to road accidents, indicating the importance of implementing specific safety measures and educational programs for this population.



**Fig 4.6 Frequency of collisions in RTA against Age Range and Education Title**

Shifting emphasis to the association between education levels and accident engagement, a remarkable tendency arises. Contrary to expectations, fig 4.6 shows that persons with educational levels equal to Matric have the largest representation in traffic crashes, followed by those with a primary school education and then, shockingly, illiterate individuals. This surprising discovery contradicts conventional assumptions and motivates a closer examination of the factors influencing accident rates among various educational levels.

## 4.2 Statistical Analysis:

The study has also incorporated statistical analysis before applying the ML techniques on both of these datasets to identify how much the input features are correlated with our targeted attribute. The study has conducted the Chi-Square Contingency Test along with the calculation of Cramer’s V. The Chi-Square and Cramer’s phi are the most suitable statistical techniques according to the nature of these datasets (Ben-Shachar et al., 1982). After conducting the analysis each independent variable got Chi-Square statistic, p-value, degree of freedom and Cramer’s phi. These values reflect the strength and significance of the correlation between the tested pair of attributes. The topmost important correlations of the KSI dataset are shown in Table 4.1.

**Table 4.1 Statistical Analysis on KSI Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Chi-Square Statistic** | **P-value** | **Degree of Freedom** | **Cramer’s V** |
| INVTYPE | 8754.868189 | 0.000 | 64 | 0.363452 |
| ACCLASS | 6567.699552 | 0.000 | 8 | 0.445067 |
| VEHTYPE | 4435.858836 | 0.000 | 120 | 0.287356 |
| PEDESTRIAN | 954.76729 | 0.000 | 4 | 0.239984 |
| TRAFFCTL | 108.55384 | 3.34E**-09** | 36 | 0.040502 |
| DRIVCOND | 1096.557247 | 0.000 | 36 | 0.172358 |
| RDSFCOND | 90.977844 | 1.46E-07 | 32 | 0.037067 |
| ROAD\_CLASS | 101.925096 | 3.23E-08 | 36 | 0.03959 |
| LIGHT | 75.68876 | 2.12E-05 | 32 | 0.033785 |
| SPEEDING | 261.805108 | 1.86E-55 | 4 | 0.125668 |

The statistical analysis on KSI shows that INVTYPE, ACCLASS and PEDESTRIAN are highly significant and have strong association with injury severity type. The VEHTYPE, DRIVCOND and SPEEDING are also highly significant but have a moderate association, the TRAFFCTL, RDSFCOND, ROAD\_CLASS and LIGHT have a high significance and weak association with the targeted attribute.

Similarly, the same statistical analysis conducted on the RTA dataset, table 4.2 shows the topmost significant correlations.

**Table 4.2 Statistical Analysis on RTA Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Chi-Square Statistic** | **P-value** | **Degree of Freedom** | **Cramer’s V** |
| Patient Status | 14305.29379 | 0.000 | 8 | 0.393526 |
| ACCLASS | 5258.443386 | 0.000 | 4 | 0.337418 |
| Total Patients in Emergency | 509.534073 | 0.000 | 44 | 0.052517 |
| Trains involved | 557.56488 | 0.000 | 4 | 0.109872 |
| Trucks involved | 701.024602 | 0.000 | 12 | 0.071129 |
| Age Range | 183.039109 | 0.000 | 16 | 0.031476 |
| Cause | 283.857996 | 0.000 | 24 | 0.039198 |
| Education title | 201.455444 | 0.000 | 28 | 0.033022 |
| Gender | 111.796691 | 0.000 | 8 | 0.034789 |
| Light | 105.60159 | 0.000 | 4 | 0.047816 |

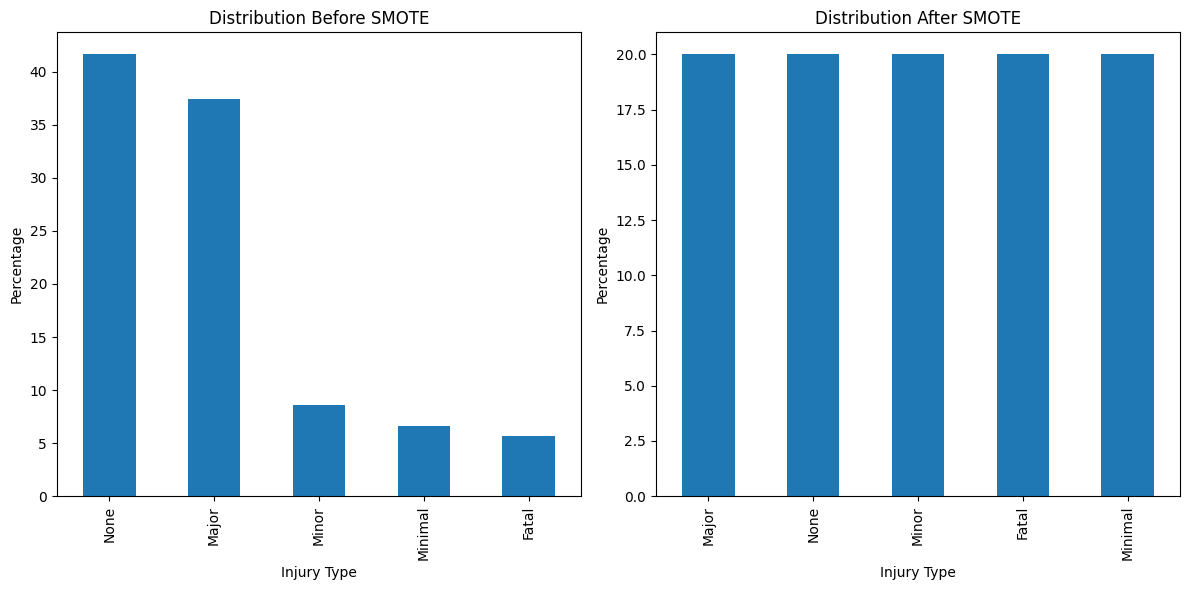
The statistics on RTA dataset shows that Patient Status and ACCLASS variables exhibit a very strong association and will provide a high contribution in prediction. The variables of total patients in emergency, trains involved and trucks involved show a moderate association and are statistically significant factors in predicting the injury severity. Lastly, age range, cause, education title, gender and light variables provide weaker association and are having statistically small impact that suggests that alone they may not be a strong predictor of injury severity.

## 4.3 Experimental Setup:

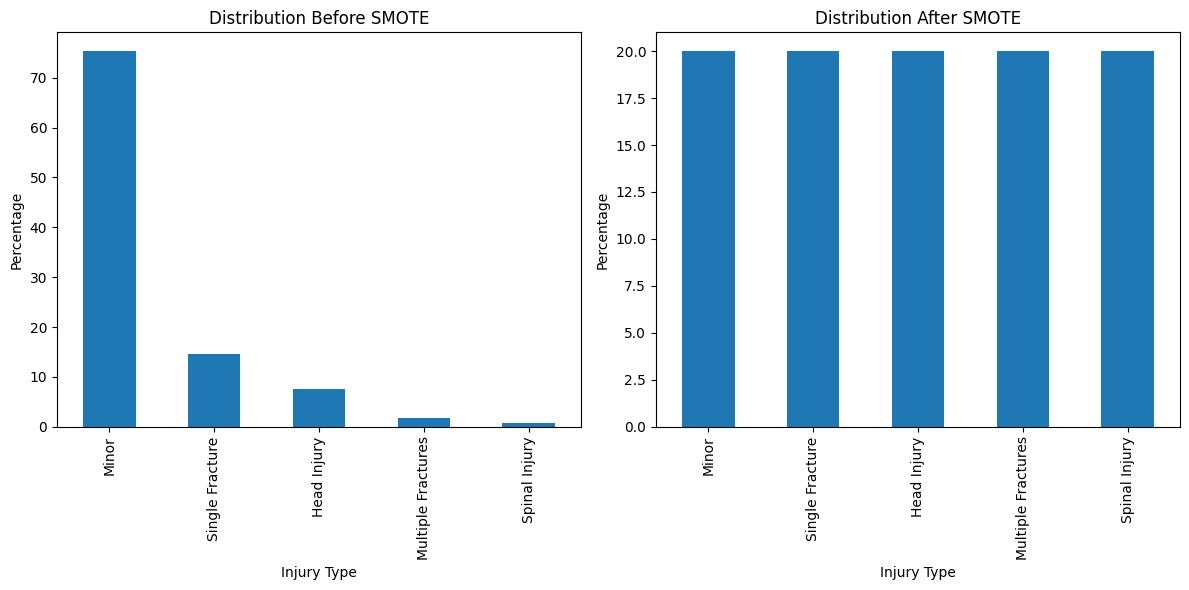
To implement our proposed framework, Python has been selected as the preferred tool. Python is notable for its extensive collection of machine learning libraries and user-friendly features, providing a variety of built-in packages for various purposes. Python is widely recognized for its simplicity and extensive use in the scientific world, making it convenient for conducting experiments. The datasets for the study were obtained in the xlsx format, Python's diverse utility packages facilitate the seamless reading and processing of such files.

The study utilized a collection of libraries, such as pandas, numpy, matplotlib, math, tensorflow\_decision\_forests, apyori, sklearn, and pycaret, to ensure smooth implementation. The Google Colab Notebooks, running on Python version 3.10.12, were employed on a Microsoft Windows 10 platform hosted on an HP machine equipped with a Core i7 processor and 8GB of RAM. This simple technological setup ensures a robust and efficient environment for the execution of our framework.

The study has employed two datasets to answer the research questions. The first one is the KSI dataset, sourced from the Public Safety Data Portal provided by Toronto Police Services (Toronto Police Services, 2023). The final KSI dataset consists of 45 attributes and 16,578 rows after applying data preprocessing techniques. The second one is the RTA dataset, obtained from Punjab Emergency Services, Rescue 1122 office Rawalpindi (RTA, 2023). The final RTA dataset consists of 29 attributes and 46,187 rows after applying data preprocessing techniques. After necessary data preprocessing, the analysis of the data revealed the class imbalance issue, so the SMOTE technique was employed to tackle this issue. Fig 4.7 and Fig 4.8 show the distribution of data of both datasets before and after applying the SMOTE technique.



**Fig 4.7 Distribution of Target Attribute in KSI Dataset Before and After SMOTE**



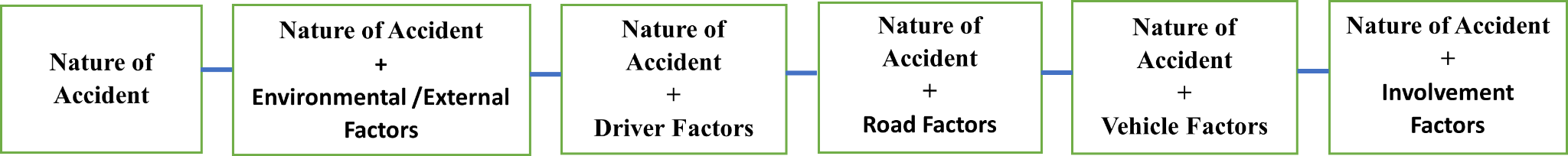
**Fig 4.8 Distribution of Target Attribute in RTA Dataset Before and After SMOTE**

The study utilizes Auto Machine Learning, a library of Python known as PyCaret for developing and comparing different predictor models. PyCaret is a robust and flexible tool in the field of machine learning, significantly speeding up the process of developing models (Ali et al., 2020). This Python library exhibits a dedication to simplicity and effectiveness in the field of data science, providing a complete set of functions that expedite various processes, from data pre-processing to model deployment. It offers a user-friendly framework that is easy to use for both novices and experienced practitioners. After applying SMOTE, the datasets have been distributed into two parts using the 10-fold cross-validation technique. The 90% of the data will be used for training purposes and the remaining 10% will be used for model evaluation. Then the models have been trained using the *compare\_models()* function. This results in the development of models including Extra Tree Classifier, Random Forest, Gradient Boosting Classifier, Light Gradient Boosting Machine, Extreme Gradient Boosting, Ridge Classifier, Linear Discriminant Analysis, Decision Tree Classifier, Logistic Regression and others. However, the study has shortlisted and considered only the top-performing models. After developing predictor models the study has used the metrics of Accuracy, Precision, Recall, F1-Score and AUC for the purpose of evaluation of these models.

Next to find out the hidden factors that lead to fatal or major injuries, the study has performed the next experiment on the same datasets using the apriori algorithm of the association rule mining technique. While performing the necessary pre-processing techniques, the study has confirmed that all the attributes are categorical as the apriori algorithm is only applicable to categorical attributes, so here in the current study, the numeric attribute of age has been converted into categorical form. To generate the effective associations some attributes such as INDEX, STREET1, STREET2, Year, Month, Date and Time in the KSI dataset and EcNumber, CallTime, Year, Month, Date and Time and ResponseTime in the RTA dataset have been removed as these attributes have individual values and do not create a group of values. Next, the study has divided the datasets into subsections to find out how the nature of accidents correlates with the factors under external and environmental, driver, road, vehicle and involvement attributes.

The nature of accidents includes timestamp-related data, accident class, injury severity, cause of accident and impact type. Environmental and external factors are related to lightning conditions, season and visibility. Driver factors include driver action, driver condition and maneuver. Road factors include road class, district, accident location, traffic control availability and road surface condition traits. The vehicle related info includes the type of vehicle and lastly involvement related factors including involvement type, age and education level.

This division of datasets has been merged into 6 groups as shown in Fig 4.9, to acquire association rules.



**Fig 4.9 Main Groups of Attributes to Obtain Association Rules**

The study has employed the Apriori algorithm in the same setup to perform the association rule mining technique. The study has utilized the `apyori` library, a Python library specially designed for association rule mining, to uncover significant patterns and associations within our dataset. The `apyori` library provides an easy and flexible interface for mining frequent item-sets and creating association rules. The current study has used the metrics of support, confidence and lift to shortlist the most important rules. Keeping in view the volume of datasets, the threshold values of confidence and lift are taken as 0.5 and 1.0 respectively.

## 4.4 Results

This section of the study depicts the outcomes of the proposed framework. This section contains the performance measures of models created on both datasets i.e. KSI and RTA. The models have been evaluated on accuracy, precision, recall, f1 score and AUC. Furthermore, the section also contains some of the interesting associations that are mined using the apriori algorithm on both of these datasets. The study answers the following questions:

**Question 1:** Can ML algorithms effectively model the injury severity level resulting from collisions in road accidents, and if so, which ML classifier offers optimal performance in predicting the injury severity level?

To answer our first research question, this study has performed the first experiment of developing predictor models to predict injury severity using different accident-related features. The study has used two datasets for this experiment i.e. Killed or Seriously Injured (Toronto Police Services, 2023)and Road Traffic Accident (RTA, 2023) datasets. This study has used Python for developing different predictor models using the PyCaret library and evaluated these models using 10-fold cross-validation. Table from 4.3 to 4.10 represent the confusion matrices of the top-performing models.

**Table 4.3 Confusion Matrix of LGBM Model Developed Over KSI Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| None | Fatal | Major | Minor | Minimal |
| Actual Value | None | 248 | 1 | 1 | 23 | 6 |
| Fatal | 17 | 1525 | 66 | 75 | 162 |
| Major | 8 | 91 | 65 | 26 | 147 |
| Minor | 26 | 153 | 31 | 96 | 121 |
| Minimal | 10 | 154 | 34 | 47 | 1841 |

The top performing model is the Light Gradient Boosting Machine, table 4.3 shows its results. The LGBM has an accuracy of 75.78% and a TPR rate of 0.7578.

**Table 4.4 Confusion Matrix of Extreme GB Model Developed Over KSI Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| None | Fatal | Major | Minor | Minimal |
| Actual Value | None | 254 | 2 | 1 | 16 | 6 |
| Fatal | 20 | 1462 | 109 | 104 | 150 |
| Major | 8 | 69 | 88 | 53 | 119 |
| Minor | 27 | 121 | 63 | 124 | 92 |
| Minimal | 18 | 115 | 87 | 88 | 1778 |

The 2nd model generated is the Extreme Gradient Boosting, this model has an accuracy of 75.54% and TPR of 0.7554. Table 4.4 shows the performance of this model.

**Table 4.5 Confusion Matrix of RF Model Developed Over KSI Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| None | Fatal | Major | Minor | Minimal |
| Actual Value | None | 243 | 5 | 1 | 12 | 15 |
| Fatal | 18 | 1491 | 56 | 56 | 224 |
| Major | 11 | 85 | 61 | 21 | 159 |
| Minor | 27 | 163 | 15 | 92 | 130 |
| Minimal | 9 | 142 | 25 | 37 | 1873 |

Next thirdly, the Pycaret has generated the Random Forest Model, the table 4.5 shows its performance. This model has an accuracy of 74.92% and TPR of 0.7492.

**Table 4.6 Confusion Matrix of ET Classifier Model Developed Over KSI Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| None | Fatal | Major | Minor | Minimal |
| Actual Value | None | 246 | 5 | 3 | 13 | 12 |
| Fatal | 19 | 1483 | 63 | 75 | 205 |
| Major | 12 | 83 | 70 | 33 | 139 |
| Minor | 27 | 162 | 26 | 97 | 115 |
| Minimal | 17 | 158 | 33 | 48 | 1830 |

Fourthly, the Extra Tree Classifier model has been generated with an accuracy of 74.68% and TPR of 0.7468. Table 4.6 shows its performance.

**Table 4.7 Confusion Matrix of GBC Model Developed Over KSI Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| None | Fatal | Major | Minor | Minimal |
| Actual Value | None | 254 | 2 | 1 | 16 | 6 |
| Fatal | 20 | 1462 | 109 | 104 | 150 |
| Major | 8 | 69 | 88 | 53 | 119 |
| Minor | 27 | 121 | 63 | 124 | 92 |
| Minimal | 18 | 115 | 87 | 88 | 1778 |

Next at fifth number, Gradient Boosting Classifier model has been generated. Table 4.7 represents its confusion matrix. The GBC model has a TPR rate of 0.7397 and has an accuracy of 73.97%.

**Table 4.8 Confusion Matrix of DT Model Developed Over KSI Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| None | Fatal | Major | Minor | Minimal |
| Actual Value | None | 217 | 21 | 5 | 17 | 19 |
| Fatal | 21 | 1395 | 104 | 143 | 182 |
| Major | 11 | 90 | 87 | 55 | 94 |
| Minor | 22 | 136 | 55 | 116 | 98 |
| Minimal | 16 | 158 | 115 | 132 | 1665 |

The table 4.8 represents the results of the Decision Tree model that has an accuracy of 68.55%. The DT predictor model has a TPR of approximately 0.6855.

**Table 4.9 Confusion Matrix of Ridge Classifier Model Developed Over KSI Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| None | Fatal | Major | Minor | Minimal |
| Actual Value | None | 274 | 2 | 0 | 2 | 1 |
| Fatal | 34 | 1302 | 271 | 189 | 49 |
| Major | 27 | 31 | 149 | 76 | 54 |
| Minor | 64 | 54 | 133 | 130 | 46 |
| Minimal | 54 | 57 | 248 | 134 | 1593 |

The Ridge Classifier model was generated next by Pycaret, table 4.9 represents its confusion matrix. This model has a TPR of 0.6817 and has an accuracy of 68.17%.

**Table 4.10 Confusion Matrix of LDA Model Developed Over KSI Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| None | Fatal | Major | Minor | Minimal |
| Actual Value | None | 277 | 0 | 0 | 1 | 1 |
| Fatal | 34 | 1247 | 310 | 219 | 35 |
| Major | 22 | 21 | 172 | 75 | 47 |
| Minor | 56 | 40 | 153 | 140 | 38 |
| Minimal | 54 | 29 | 324 | 141 | 1538 |

The last model considered for the comparison is Linear Discriminant Analysis. Table 4.10 represents its results, the model has an accuracy of 66.75% and a TPR of 0.6675.

A comparison with several evaluation metrics of these eight models has been presented in

Table 4.11.

**Table 4.11 Evaluation Measures of Models Developed Over KSI Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **AUC** | **Recall** | **Precision** | **F1** |
| Light Gradient Boosting Machine | 0.7578 | 0.9244 | 0.7578 | 0.7332 | 0.7419 |
| Extreme Gradient Boosting | 0.7554 | 0.9229 | 0.7554 | 0.7316 | 0.7400 |
| Random Forest Classifier | 0.7492 | 0.9116 | 0.7492 | 0.7211 | 0.7282 |
| Extra Trees Classifier | 0.7468 | 0.9015 | 0.7468 | 0.7231 | 0.7313 |
| Gradient Boosting Classifier | 0.7397 | 0.9190 | 0.7397 | 0.7395 | 0.7387 |
| Decision Tree | 0.6855 | 0.7819 | 0.6855 | 0.6945 | 0.6895 |
| Ridge Classifier | 0.6817 | 0.0000 | 0.6817 | 0.7797 | 0.7166 |
| Linear Discriminant Analysis | 0.6675 | 0.9110 | 0.6675 | 0.7986 | 0.7122 |

Similarly, the study has performed the same experiment again on RTA dataset. The tables from 4.12 to 4.18 represent the confusion matrices of these models developed over the RTA dataset.

**Table 4.12 Confusion Matrix of LGBM Model Developed Over RTA Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| Minor | Single Fracture | Head Injury | Multiple Fractures | Spinal Injury |
| Actual Value | Minor | 130 | 873 | 16 | 31 | 1 |
| Single Fracture | 29 | 10271 | 2 | 134 | 1 |
| Head Injury | 25 | 192 | 4 | 13 | 0 |
| Multiple Fractures | 5 | 1905 | 0 | 110 | 1 |
| Spinal Injury | 3 | 110 | 0 | 1 | 0 |

The Light Gradient Boosting Machine model is the top performing model on RTA dataset. Table 4.12 shows its results. This model has a TPR of 0.7585 and an accuracy of 75.85%.

**Table 4.13 Confusion Matrix of GB Model Developed Over RTA Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| Minor | Single Fracture | Head Injury | Multiple Fractures | Spinal Injury |
| Actual Value | Minor | 118 | 841 | 23 | 68 | 1 |
| Single Fracture | 16 | 10086 | 3 | 327 | 5 |
| Head Injury | 23 | 186 | 6 | 19 | 0 |
| Multiple Fractures | 3 | 1825 | 1 | 191 | 1 |
| Spinal Injury | 3 | 107 | 0 | 3 | 1 |

The second-best performer model is based on Gradient Boosting algorithm. This model has an accuracy of 75% and a TPR of 0.7539. Table 4.13 represents its results.

**Table 4.14 Confusion Matrix of RF Model Developed Over RTA Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| Minor | Single Fracture | Head Injury | Multiple Fractures | Spinal Injury |
| Actual Value | Minor | 162 | 800 | 5 | 83 | 1 |
| Single Fracture | 50 | 9993 | 7 | 384 | 3 |
| Head Injury | 28 | 184 | 3 | 18 | 1 |
| Multiple Fractures | 17 | 1776 | 3 | 225 | 0 |
| Spinal Injury | 6 | 97 | 1 | 10 | 0 |

The Random Forest Model is the third best predictor model. Table 4.14 shows its confusion matrix. This model has an accuracy of 75% and TPR of approximately equal to 0.7502.

**Table 4.15 Confusion Matrix of ET Model Developed Over RTA Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| Minor | Single Fracture | Head Injury | Multiple Fractures | Spinal Injury |
| Actual Value | Minor | 184 | 740 | 9 | 117 | 1 |
| Single Fracture | 145 | 9700 | 16 | 566 | 10 |
| Head Injury | 31 | 171 | 4 | 27 | 1 |
| Multiple Fractures | 48 | 1630 | 8 | 335 | 0 |
| Spinal Injury | 9 | 86 | 1 | 18 | 0 |

Fourthly, Pycaret has generated the Extra Tree Classifier model. It has an accuracy of 73.6% and a TPR of approximately equal to 0.7361. Table 4.15 shows its results.

**Table 4.16 Confusion Matrix of AdaBoost Model Developed Over RTA Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| Minor | Single Fracture | Head Injury | Multiple Fractures | Spinal Injury |
| Actual Value | Minor | 137 | 536 | 4 | 372 | 2 |
| Single Fracture | 34 | 8691 | 0 | 1706 | 6 |
| Head Injury | 28 | 120 | 1 | 84 | 1 |
| Multiple Fractures | 9 | 1249 | 1 | 761 | 1 |
| Spinal Injury | 5 | 65 | 0 | 44 | 0 |

Next, the AdaBoost model was generated that has achieved an accuracy of 68.83% and TPR of 0.69. Table 4.16 represents its confusion matrix.

**Table 4.17 Confusion Matrix of DT Model Developed Over RTA Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| Minor | Single Fracture | Head Injury | Multiple Fractures | Spinal Injury |
| Actual Value | Minor | 236 | 511 | 49 | 239 | 16 |
| Single Fracture | 617 | 8413 | 98 | 1227 | 82 |
| Head Injury | 44 | 113 | 21 | 52 | 4 |
| Multiple Fractures | 240 | 1145 | 65 | 532 | 39 |
| Spinal Injury | 22 | 58 | 2 | 31 | 1 |

The Decision Tree has occupied the sixth place in this list, with an accuracy of 66% and a TPR of approximately equals to 0.6605. Table 4.17 represents its results.

**Table 4.18 Confusion Matrix of LDA Model Developed Over RTA Dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Confusion Matrix* | | Predicted Value | | | | |
| Minor | Single Fracture | Head Injury | Multiple Fractures | Spinal Injury |
| Actual Value | Minor | 293 | 28 | 182 | 306 | 242 |
| Single Fracture | 959 | 6315 | 453 | 1572 | 1138 |
| Head Injury | 45 | 8 | 59 | 63 | 59 |
| Multiple Fractures | 308 | 201 | 274 | 783 | 455 |
| Spinal Injury | 27 | 8 | 8 | 31 | 40 |

Lastly, the Linear Discriminant Analysis is the last model discussed on this dataset. It has achieved the lowest accuracy of just 54% and TPR equals to 54.34. Table 4.18 represents its results. The reason to include the LDA model for comparison is the highest value of precision among all which is equal to 78%.

Table 4.19 represents a comparison of the results of these predictor models using the measures of accuracy, precision, recall, F1 score and AUC.

**Table 4.19 Evaluation Measures of Models Developed Over RTA Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **AUC** | **Recall** | **Precision** | **F1** |
| Light Gradient Boosting Machine | 0.7585 | 0.8004 | 0.7585 | 0.6867 | 0.6830 |
| Gradient Boosting Classifier | 0.7539 | 0.7902 | 0.7539 | 0.6869 | 0.6834 |
| Random Forest Classifier | 0.7502 | 0.7867 | 0.7502 | 0.6773 | 0.6892 |
| Extra Trees Classifier | 0.7361 | 0.7735 | 0.7361 | 0.6713 | 0.6915 |
| Ada Boost Classifier | 0.6883 | 0.7272 | 0.6883 | 0.7027 | 0.6816 |
| Decision Tree Classifier | 0.6605 | 0.6175 | 0.6605 | 0.6724 | 0.6662 |
| Linear Discriminant Analysis | 0.5434 | 0.7860 | 0.5434 | 0.7839 | 0.6278 |

**Question 2:** What are the underlying hidden factors that contribute to fatal or major injuries?

Next, to answer the second and third research question and to find out the hidden factors that lead to collisions and fatal or major injuries, the study has performed the second experiment to find out the association rules.

The study has used the same datasets for this experiment i.e. KSI and RTA datasets. This study has used Python for mining associations with apriori algorithm.

The algorithm was first applied to the subgroup ‘Nature of Accident' of the KSI dataset and it has generated numerous rules. The top 6 association rules with the highest values of lift are shown in table 4.20 as follows:

**Table 4.20 Top associations for ‘Nature of Accident’ over KSI**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | Pedestrian Collisions | No Red-light Violation, Weekend | 0.05 | 0.97 | 3.85 |
| 2 | Speeding | Aggressive Driving | 0.12 | 0.85 | 2.42 |
| 3 | Rear End, No Red-light Violation | Non-Fatal, Aggressive Driving | 0.06 | 0.70 | 2.26 |
| 4 | Red-light Violation | No Over speeding, Non-Fatal Injury, Aggressive Driving | 0.062 | 0.73 | 2.13 |
| 5 | No Over speeding, Pedestrian Collisions | None | 0.19 | 0.95 | 1.12 |
| 6 | SMV | No Red-light Violation, Non-Fatal | 0.06 | 0.82 | 1.04 |

Next, the apriori algorithm was applied on the combination of subgroups of ‘Nature of Accident’ and ‘Environmental/External Factors’ of KSI dataset. Table 4.21 represents the top 6 meaningful association rules with the highest values of lift.

**Table 4.21 Top associations for ‘Nature of Accident’ & ‘Environmental Factors’ over KSI**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | Clear, Daylight | Aggressive Driving,  Over speeding, | 0.05 | 0.89 | 1.96 |
| 2 | SMV, Artificial Dark | Major | 0.01 | 0.50 | 1.48 |
| 3 | Daylight, Cyclist Collisions, Clear | Non-Fatal | 0.06 | 0.61 | 1.33 |
| 4 | Rain | No Over speeding, No Aggressive Driving | 0.053 | 1 | 1.16 |
| 5 | Clear, Workday, Pedestrian Collisions | Major | 0.09 | 0.74 | 1.11 |
| 6 | Daylight, Pedestrian Collisions, Clear | Major | 0.07 | 0.60 | 1.05 |

Table 4.22 shows the top 8 associations rules extracted from the combination of subgroups of ‘Nature of Accident’ and ‘Driver Factors’ of KSI dataset.

**Table 4.22 Top associations for ‘Nature of Accident’ & ‘Driver Factors’ over KSI**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | Following too close | Read End | 0.01 | 0.860 | 8.96 |
| 2 | Pedestrian Collisions, Failed to Yield Right of Way, Inattentiveness | Turning Left | 0.01 | 0.70 | 7.09 |
| 3 | Stopped | Rear End | 0.01 | 0.56 | 5.88 |
| 4 | Exceeding Speed Limit | Going Ahead | 0.013 | 0.93 | 2.69 |
| 5 | Disobeyed Traffic Control | Going Ahead | 0.02 | 0.89 | 2.60 |
| 6 | Lost control | Major | 0.03 | 0.57 | 1.68 |
| 7 | Inattentive, Cyclist Collisions | Non-Fatal | 0.01 | 0.94 | 1.09 |
| 8 | Improper Turn, Inattentive | Non-Fatal | 0.01 | 0.92 | 1.07 |

Next, the apriori algorithm was applied to the combination of subgroups of ‘Nature of Accident’ and ‘Road Factors’ of KSI dataset. Table 4.23 represents the top 6 meaningful association rules with a maximum value of lift.

**Table 4.23 Top associations for ‘Nature of Accident’ & ‘Road Factors’ over KSI**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | Mid-Block, Etobicoke York | No Control | 0.08 | 0.96 | 1.99 |
| 2 | At Intersection | Traffic Signal | 0.33 | 0.70 | 1.67 |
| 3 | Stop Sign | At Intersection | 0.06 | 0.73 | 1.54 |
| 4 | No Control, Pedestrian Collisions | Fatal | 0.08 | 0.54 | 1.12 |
| 5 | Pedestrian Collisions, At Intersection | Major | 0.07 | 0.51 | 1.07 |
| 6 | Pedestrian Collisions | Intersection | 0.27 | 0.69 | 1.04 |

Then, the subgroup of ‘Nature of Accident’ and ‘Vehicle Factors’ are considered for mining the association rules, the table 4.24 represents the top 3 associations within this combination.

**Table 4.24 Top associations for ‘Nature of Accident’ & ‘Vehicle Factors’ over KSI**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | Rear End | Automobile, Station Wagon | 0.05 | 0.53 | 1.28 |
| 2 | Automobile, Pedestrian Collisions | Fatal | 0.05 | 0.75 | 1.03 |
| 3 | Automobile, Station Wagon | Fatal | 0.05 | 0.96 | 1.02 |

Lastly, the subgroups of ‘Nature of Accident’ and ‘Involvement Factors’ are grouped together and handed over to the apriori algorithm. Table 4.25 shows the top 5 association rules having the highest value of lift.

**Table 4.25 Top associations for ‘Nature of Accident’ & ‘Involvement Factors’ over KSI**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | Pedestrian, Crossing with right of way | Aggressive Driving | 0.05 | 0.87 | 11.30 |
| 2 | Pedestrian | Crossing with right of way | 0.05 | 0.99 | 5.80 |
| 3 | Aggressive Driving, Pedestrian Collisions | Major | 0.06 | 0.78 | 5.315 |
| 4 | Aggressive Driving, Driver | Major, Non-Fatal | 0.06 | 0.97 | 1.12 |
| 5 | Aggressive Driving, 20 to 24 | Non-Fatal | 0.05 | 0.86 | 1.003 |

The same experiment has been repeated with the RTA dataset. This dataset has been divided into 4 subgroups i.e. nature of accident, environmental factors, vehicle factors and involvement attributes. Table 4.26 shows the top 3 association rules for the subgroup ‘Nature of Accident’ of the RTA dataset that have the maximum value of lift.

**Table 4.26 Top associations for ‘Nature of Accident’ over RTA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | 12 to 18, Over Speed | Minor, Non-Fatal | 0.14 | 0.57 | 1.04 |
| 2 | Carelessness | Single Fracture, Non-Fatal | 0.05 | 0.99 | 1.01 |
| 3 | 12 to 18, Weekday, Carelessness | Non-Fatal | 0.08 | 0.72 | 1.001 |

Following this, a combination of subgroups from the categories 'Nature of Accident' and 'Environmental Factors' is chosen, and the algorithm is applied to it. Table 4.27 displays the top 3 association rules.

**Table 4.27 Top associations for ‘Nature of Accident’ & ‘Environmental Factors’ over RTA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | 18 to 24, Night | Single Fracture | 0.04 | 1 | 2.88 |
| 2 | 12 to 18, Rainy, Day | Minor, Non-Fatal | 0.05 | 0.78 | 1.55 |
| 3 | Over Speed, Rainy, Day | Minor | 0.05 | 0.52 | 1.05 |

Subsequently, a blend of subcategories of 'Nature of Accident' and 'Vehicle Factors' is chosen, and the algorithm is implemented on it. The resulting top 6 association rules are presented in Table 4.28.

**Table 4.28 Top associations for ‘Nature of Accident’ & ‘Vehicle Factors’ over RTA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | 1 Bike, Over speed, weekend | Minor | 0.05 | 0.50 | 3.67 |
| 2 | Over speed, 1 Car | Non-Fatal | 0.14 | 0.64 | 1.11 |
| 3 | 1 Rickshaw, Over speed | Minor | 0.06 | 0.71 | 1.10 |
| 4 | Carelessness, 1 Car | Non-Fatal | 0.05 | 0.95 | 1.06 |
| 5 | Over speed, 1 Van | Minor | 0.05 | 0.66 | 1.03 |
| 6 | Head Injury | 1 Bike | 0.04 | 0.60 | 1.00 |

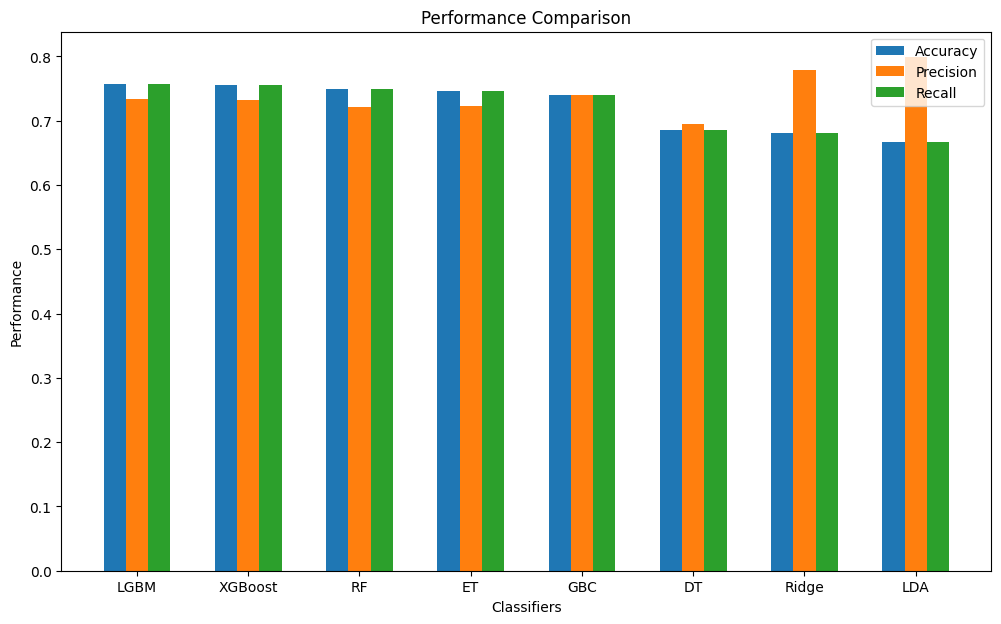
Lastly, the combination of ‘Nature of Accident’ and ‘Involvement Factors’ are combined and analyzed for association rule mining. Table 4.29 shows the top 4 most important associations.

**Table 4.29 Top associations for ‘Nature of Accident’ & ‘Involvement Factors’ over RTA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rule** | **Antecedent** | **Consequent** | **Support** | **Confidence** | **Lift** |
| 1 | Over speed, weekend, 1 Bike | Alive & Unstable | 0.044 | 0.60 | 3.33 |
| 2 | Alive & stable, 20-30, 12 to 18, Male, 1 Patient | Over Speed, Minor, Non-Fatal | 0.04 | 0.58 | 1.63 |
| 3 | Matric, Over speed, 1 Bike, 1 Patient | Alive & Unstable, Minor | 0.11 | 0.63 | 1.61 |
| 4 | <20, Male, 1 Patient, 12-18 | Over Speed, Minor, Alive & Stable | 0.05 | 0.53 | 1.49 |

## 4.5 Discussion

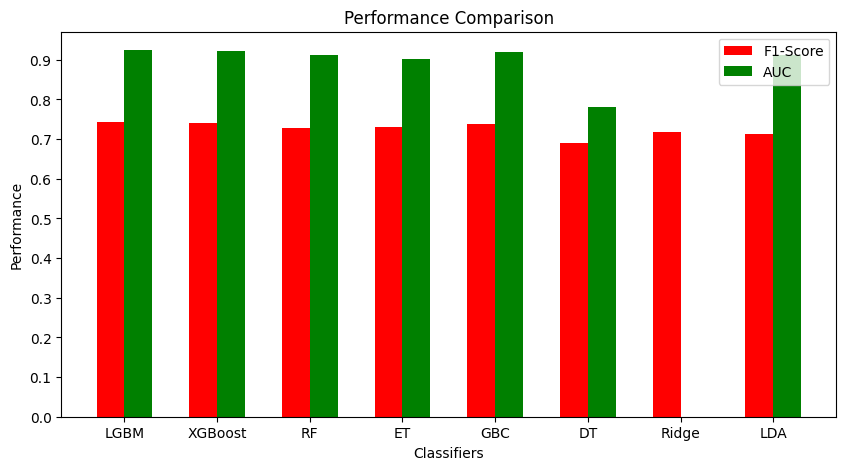
The results of the experiment performed on KSI dataset shows that Light Gradient Boosting Machine, Extreme Gradient Boosting, Random Forest Classifier and Extra Tree Classifier has predicted the injury severity with high accuracies of 0.7578, 0.7554, 0.7492 and 0.7468 respectively. The Gradient Boosting Classifier has also nearly equal performance but the Decision Tree, Ridge Classifier and Linear Discriminant Analysis are less accurate. A comparison of accuracy, precision and recall has been shown in the following fig, fig 4.10.



**Fig 4.10 Visualization of Accuracy, Precision and Recall of ML Models Trained over KSI**

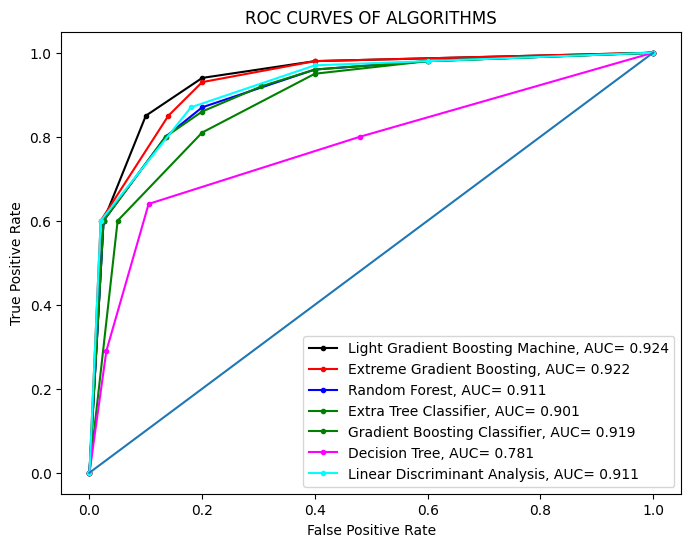
When the data is imbalanced, as in these datasets, then the accuracy alone is not considered a good measure of evaluation, therefore it is pertinent to consider the other metrics such as recall, precision, F1-score and ROC curve. Fig 4.10 shows that the LGBM, XGBoost, RF and ET have achieved higher recall values as compared to GBC, DT, Ridge and LDA. However, it can be seen that the LDA has achieved the highest precision value of 0.7986, after this Ridge Classifier with a precision of 0.7797. Then comes the GBC and LGBM with precision values equal to 0.7395 and 0.7332 respectively.

Next, the study has also compared the F1-Score and AUC of the models. Fig 4.11 represents a comparison of these measures, The F1-Score is the harmonic mean of precision and recall and it is considered most useful in the imbalanced class. It can be seen from Table 4.11 and Fig 4.11 that LGBM, XGBoost, GBC and ET classifier models have achieved the highest F1-Score measure equal to 0.7419, 0.7400, 0.7387 and 0.7313 respectively. The RF, Ridge Classifier and LDA have mediocre F1-Score and DT has a relatively lower F1 value. Fig 4.11 also shows a comparison of AUC. The AUC is the area under the ROC curve. It summarizes the performance of the predictor models. The AUC value ranges from 0 to 1, and the closer the value is to 1, the better it is considered.



**Fig 4.11 Visualization of F1-Score and AUC of ML models trained over KSI**

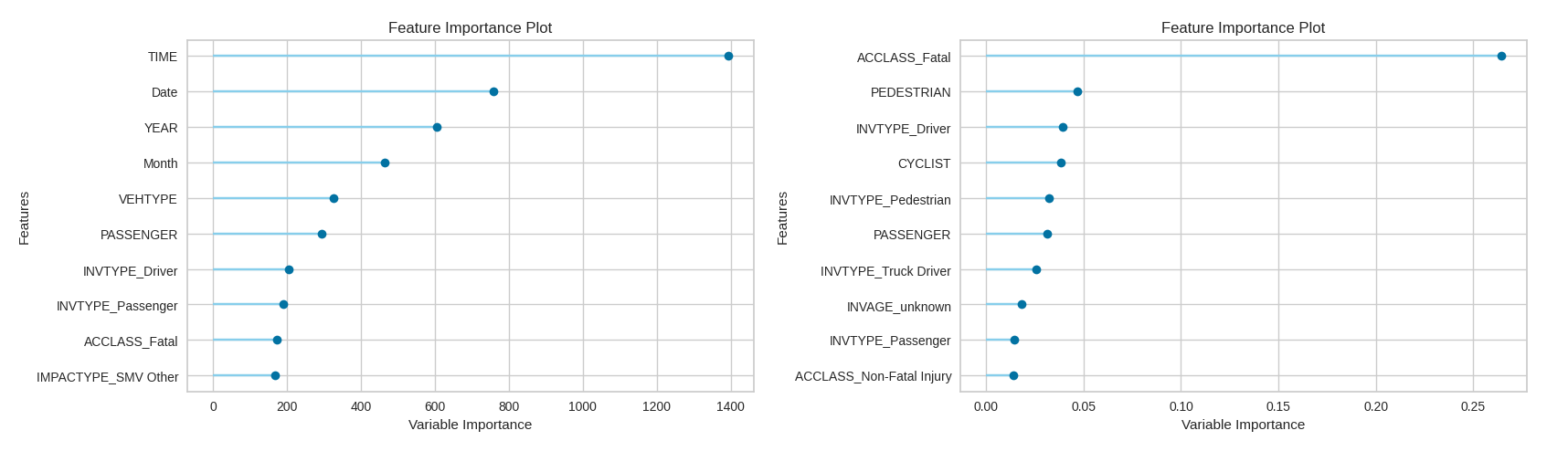
These developed models have also been evaluated using the AUC metric that denotes the area under the ROC curve. Generally, the AUC value below or equal to 0.50 is considered inadequate, and a value between 0.70 to 0.80 is considered sufficient. The fig shows that most of the models have a higher AUC value which is greater than 0.90, the DT has a relatively low AUC value of 0.78 which is also not bad but has an inferior performance than the other predictor models. Lastly, the Ridge Classifier has an AUC equal to zero which shows it is not performing well and its predictions are not very useful.



**Fig 4.12 ROC Curve of top performing ML models trained over KSI Dataset**

Fig 4.12 shows the detailed receiver operating characteristic (ROC) curve and area under the curve (AUC) of LGBM, Extreme GB, GBC and RF and as these models have achieved the highest AUC than other algorithms and thus performed better.

Next, to understand which factors are the most influencing in predicting the injury severity in the KSI dataset, the top 10 most important features in the top-performing models are enlisted in fig 4.13.

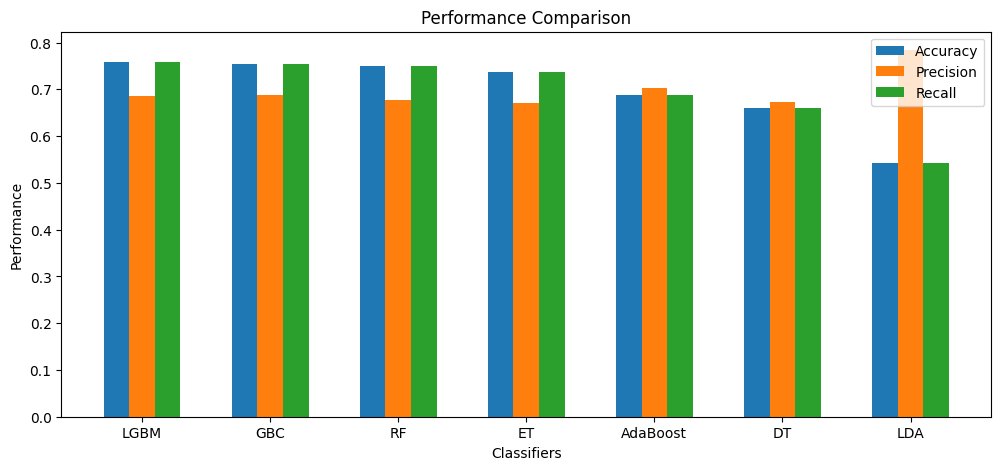


**Fig 4.13 Feature Importance for KSI Dataset**

The figure shows the feature importance of the Light Gradient Boosting Machine and Extreme Gradient Boosting Machine. The features of ACCLASS\_Fatal, PASSENGER, INVTYPE\_Driver, and INVTYPE\_Passenger are common in both of these models. It can also be seen that the LGBM model has focused mainly on Timeframe and GBM has focused mainly on accidental features and involvement. When analyzing the feature importance of GBM, it can be seen that most of the features are associated with fatal or major injuries of the involvement. For example, the two features of the presence of pedestrians and cyclists, and the category of ACCLASS as Fatal leads to fatal or major injuries among involvements especially the pedestrians and cyclists.

So, from these experimental results and their discussion, it may be concluded that the Light Gradient Boosting Machine, Extreme Gradient Boosting and Random Forest Classifier are the top three best injury severity predictors for the KSI dataset in terms of accuracy, recall and AUC. However, if the metric of precision is considered then the LDA model has achieved the highest precision of 79.8%.

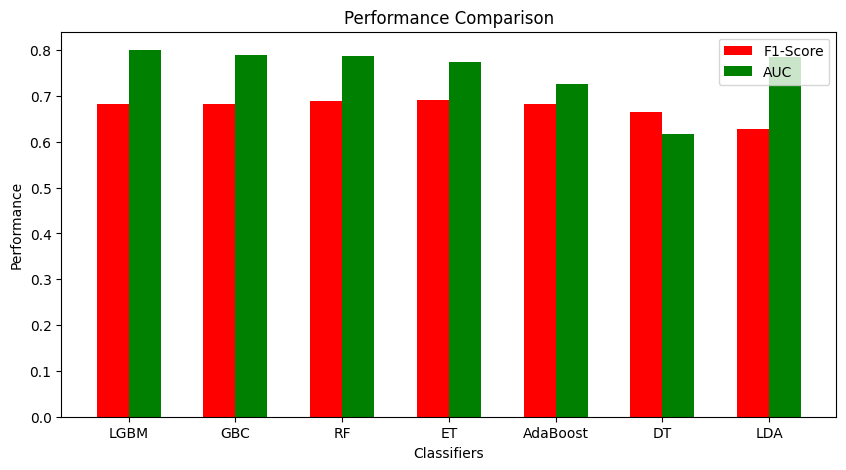
Next, the same experiment has been performed on the RTA dataset, and it has been seen from Table 4.19 that Light Gradient Boosting Machine, Gradient Boosting Classifier and Random Forest Classifier have achieved the high actuaries of 75.85%, 75.39% and 75.02% respectively. The Extra Tree Classifier has also nearly the same performance but Ada Boost, Decision Tree and LDA have achieved relatively low accuracy. A comparison of accuracy, precision and recall of these models has been shown in the following fig, fig 4.14.



**Fig 4.14 Visualization of Accuracy, Precision and Recall of ML Models Trained over RTA**

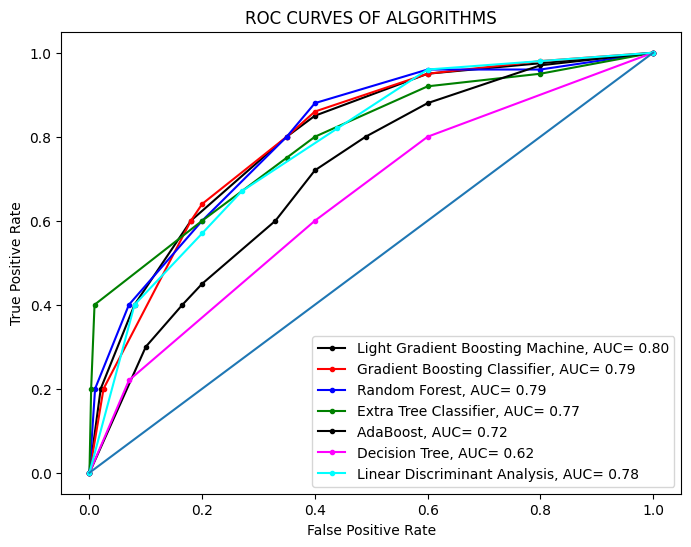
Considering the other performance measures of recall and precision, it can be seen from Fig 4.13 that LGBM, GBC and RF have achieved a higher number of recall values as compared to ET, AdaBoost, DT and LDA. Similarly, in terms of precision, it can be seen that the LDA has achieved a maximum precision of 0.7839, after this AdaBoost Classifier has achieved a precision of 0.7027. The rest of the models have approximately the same measure of precision.

Next, fig 4.15 represents a comparison of the f1-score and AUC of these models. These metrics are again considered as the RTA dataset has also an imbalance issue. It can be seen from Table 4.19 and Fig 4.14 that the Extra Tree Classifier has the highest f1-score of 0.6915, following the LGBM, GBC, RF and AdaBoost that are having approximately equal values of f1. The DT and LDA have the lower f1 score as compared to the rest of these models.



**Fig 4.15 Visualization of F1-Score and AUC of ML Models Trained over RTA**

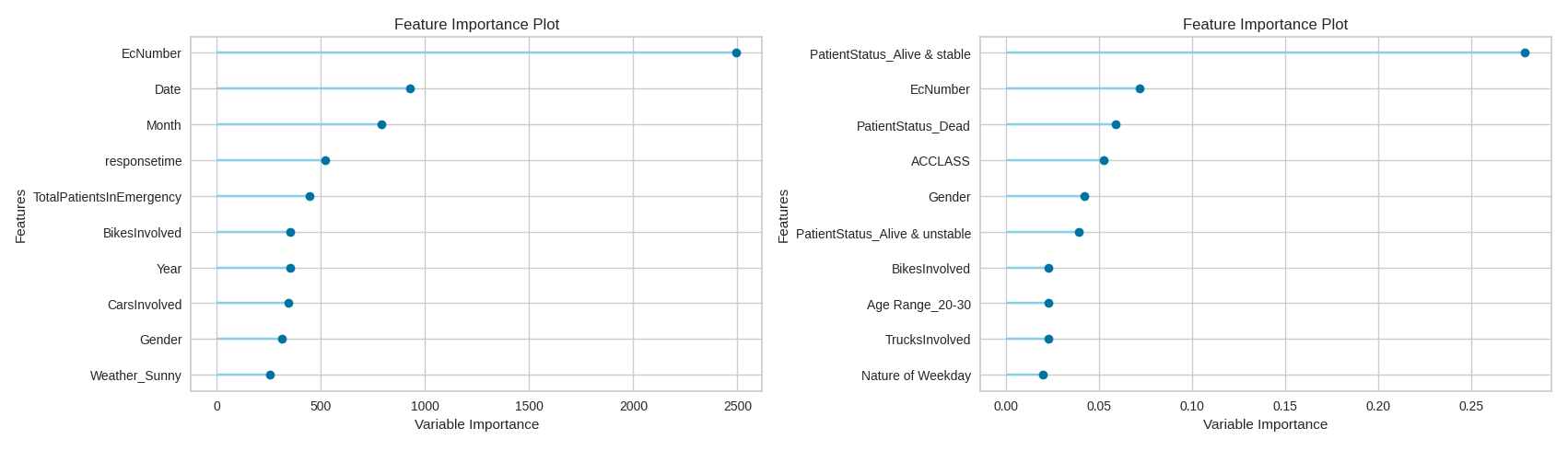
Fig. 4.15 shows that almost all models have achieved good AUC scores, except for the Decision Tree (DT) model, which has an AUC of 0.6175. The LightGBM (LGBM) model achieved the highest AUC of 0.8004, followed by Gradient Boosting Classifier (GBC) with 0.7902, Random Forest (RF) with 0.7867, Linear Discriminant Analysis (LDA) with 0.7860, Extra Trees (ET) with 0.7735, and AdaBoost with 0.7272.



**Fig 4.16 ROC Curve of top performing ML models trained over RTA Dataset**

The fig 4.16 represents the detailed receiver operating characteristic (ROC) curves and area under the curve (AUC) of LGBM, GBC, RF and LDA, these models have achieved the highest AUC than all other models and can be considered as good predictor models for predicting injury severity over RTA dataset.

Next, the importance of features has been visualized in Fig 4.17 for the top-performing models on the RTA dataset. The figure shows the top 10 most important features that have been used by Light Gradient Boosting Machine and Gradient Boosting Machine.



**Fig 4.17 Feature Importance for RTA Dataset**

It can be seen that the features of bikes involved and gender are the common ones in both of these algorithms. The LGBM has focused more on numeric attributes such as date, month, year, response time and total patients in emergency. Whereas the GBC has more focused on accidental features such as patient status, accident class, gender, age range and nature of the weekday. The Patient Status feature is the most important one in developing this model. The accidental data also shows that most of the fatal and major injuries are sustained by bikers, so the involvement of bikes is also of importance. The statistical analysis (fig 4.6) also shows that the age range of most of the persons involved in the accidents ranges from 20 to 30.

So, from this discussion, it may be concluded that the Light Gradient Boosting Machine, Gradient Boosting Classifier and Random Forest Classifier are the top three best injury severity predictors for the RTA dataset in terms of accuracy, recall and AUC. However, if the metric of precision is considered, the LDA has achieved highest precision equals to 78.3%.

Next, the study performed the second experiment of mining association rules using the apriori algorithm. The results of this experiment performed on the subgroup ‘Nature of Accident' of the KSI dataset are shown in Table 4.20, the analysis of these rules can be summarized as follows:

Rule 1 indicates a strong association between accidents involving pedestrians and the absence of red-light violations on weekends. The higher lift value shows a stronger association between antecedent and consequence, and the confidence value shows that there are 97% chances of pedestrian collisions on weekends with no red-light violation. The statistical analysis of the KSI dataset (Table 4.1) has also shown that the involvement of pedestrians has a strong association with injury severity.

Rule 2 suggests a strong association between speeding and aggressive driving. The support indicates that this pattern is present in 12% of the dataset. The confidence value shows that when there is speeding, there is an 85% likelihood of aggressive driving. The higher lift value suggests that the occurrence of aggressive driving is 2.42 times more likely when speeding is observed, compared to when these events are independent.

Rule 3 indicates that aggressive driving is the reason for 70% of rear-end collisions regardless of the red-light violations, and these collisions lead to non-fatal accidents. The higher lift value shows a stronger association between these attributes.

The next rule 4 indicates red light violations are associated with the absence of over speeding, coupled with aggressive driving that leads to non-fatal injuries. The support shows that this pattern occurs in 6.2% of the dataset. The confidence value suggests that when there is a red-light violation, there is a 73.3% likelihood of observing no over speeding and aggressive driving.

The slightly higher lift value of rule no 5 shows a mild association between accidents involving pedestrians with the injury severity as none. The rule also states that about 95% of such collisions exist in the dataset with no over speeding from the vehicle involved.

The last rule of this category is the rule no 6, which shows that the Slow-Moving Vehicle (SMV) collisions lead to non-fatal injuries. The confidence value shows that about 82% of collisions are of this nature, furthermore, these collisions do not break the red-light violation.

So, it can be concluded from these rules that there is a need to increase the safety measures for pedestrians, especially at the weekends. Furthermore, special attention should be paid to controlling the aggressive driving and over speeding behaviors of drivers.

Table 4.21 shows the top 7 association rules for the group ‘Nature of Accident’ and ‘Environmental Factors’. The analysis of these rules is explained as follows:

A higher lift and confidence value of rule 1 and rule 4 shows a strong association between the antecedent and consequent. Rule 1 shows that aggressive driving and over speeding are strongly correlated with clear weather and daylight conditions. Rule 4 elaborates its reverse, i.e. during rainy weather the drivers are more cautious and avoid over speeding and aggressive driving behaviors.

Next, rule 2 indicates that if collisions involving SMV vehicles happen in artificial dark conditions then it leads to major injury severity in the involvements.

Rules 3, 5 and 6 are of almost similar nature, which shows the strong association between the collisions involving cyclists or pedestrians in the daylight and clear weather and the accident being classified as non-fatal and the injury severity as major.

So, these rules suggest that there is a need to address aggressive driving behaviors through education and stricter enforcement, particularly in favorable weather conditions, and ensuring adequate lighting and infrastructure for vulnerable road users can help mitigate these concerning trends.

When the nature of accident factors are combined with the driver factors it generates some interesting and important associations as shown in Table 4.22, the analysis of these rules is given below:

The higher lift value of 8.96 in rule 1 is related to the following too close as an antecedent and read end as a consequence reflecting the relatively stronger correlation between these traits. The rule suggests that there are about 86% chances of rear-end collisions if the driver is following too close to a vehicle.

Next, rule 2 shows an association between the accidents in which the driver was inattentive while turning left, failed to yield the right of way and hit some pedestrians. The higher lift value shows a stronger association between these traits and the value of confidence shows a 70% likelihood of occurring collisions with this event.

Rule 3 explains another reason for rear-end collisions in which one vehicle is suddenly stopped and the following vehicle doesn’t have enough time to respond. The higher lift value suggests a stronger association between the antecedent and consequence and the value of confidence shows that if the vehicle is stopped there is a 56% chance of rear collision.

Next, rules 4 and 5 show the reasons for exceeding speed limits and disobeying traffic rules as going ahead. These rules show a stronger association between these traits. When the driver is going ahead, it leads to a 93% chance of exceeding speed limits or 89% likelihood of disobeying some other traffic rule.

Rule 6 indicates an association between accidents where the driver loses control and the consequence being classified as major. The higher value of lift shows a stronger association and the value of confidence shows a 57% chance of such incidences.

Rules 7 and 8 indicate a significant reason for non-fatal collisions. Rule 7 shows a 94% chance of non-fatal injury with inattentive drivers in cyclist collisions, and then rule 8 shows a 92% chance of non-fatal injury involving improper turns by inattentive drivers.

These findings emphasize the pressing necessity of prioritizing the resolution of driver inattention and the behavior of following too closely in order to reduce rear-end collisions and pedestrian accidents. Encourage the practices of safe turning methods and the cultivation of driver awareness in order to minimize collisions with bicycles. Foster responsible driving behaviors and improve infrastructure to minimize collisions due to sudden stops and loss of control.

After this, the subgroup of nature of the accident is combined with the road factors subgroup to find out the useful associations. Table 4.23 shows the association rule, the explanation of these rules is as follows:

The rule 1 shows that 96% of Mid-Block roads in the District Etobicoke York do not have any type of traffic control installed. So, the installation of traffic controls can greatly reduce the frequency of collisions there.

The rule 2 and 3 show a strong association between traffic signal/stop signs and location coordinates as intersections. These rules show that around 70% of intersections have traffic signals installed, and around 73% have stop signs. The rest of the intersections should also contain some type of traffic control mechanism installed to avoid collisions at intersections. The coming rules support this suggestion.

The rule 4,5 and 6 are the important ones, rule 4 shows that are a 54% chance of pedestrian collisions where there are no controls installed, and these collisions result in fatal injuries. The statistical analysis given in Table 4.1 also shows a stronger association of pedestrians and traffic controls with injury severity. Rule 5 shows that there is 51% chance of pedestrian collisions at intersections that lead to major injuries. Rule 6 explains that around 69% of pedestrian collisions happen at intersections.

These rules highlight the need for proper traffic control mechanisms at intersections to minimize pedestrian collision events to mitigate major or fatal injury incidents.

Table 4.24 shows the top association rules for the group nature of accident and vehicle factors. The description of these rules is as follows:

Rule 1 indicates that accidents involving rear-end collisions usually happened with automobiles, and station wagons rather than any other vehicle type. The confidence value shows that around 53% chances exist of involving these vehicles in read-end collisions.

Rule 2 suggests that accidents involving automobiles and pedestrian collisions are more likely to be fatal. The confidence is 0.75, which means that 75% of accidents involving automobiles and pedestrian collisions are fatal.

The last rule of this combination is rule 3, which also shows an important association of vehicles involving fatalities. The rule suggests that accidents involving automobiles and station wagons are more likely to be fatal than accidents involving other types of vehicles. The higher confidence value shows that 96% of accidents involving automobiles and station wagons are fatal.

So, these rules suggest that automobile and station wagon vehicles should be treated specially and there should be a targeted effort to enhance safety for occupants of these vehicles.

Lastly, the attributes from the nature of accident category are combined with the involvement factors and some interesting associations are mined as shown in table 4.25, the explanation is given below:

This rule 1 states that if a pedestrian is crossing with the right of way and it leads to a collision, then there is a high probability (87%) that aggressive driving was a factor in the accident. The higher lift value shows a very strong association between the antecedent and its consequence.

Rule 2 is also highly correlated and it states that if there is a pedestrian, then there is a high probability (99%) that they were crossing with the right of way. Then it means that such accidents occur due to the mistakes of the drivers not of the pedestrians.

The next rules 3,4 and 5 are related to the previous two rules. Rule 3 having a higher association state that if pedestrian collisions occur due to aggressive driving behavior, then the injury severity is more likely to be the major ones. Rule 4 is similar to this one, but it states that if aggressive driving behavior and driver involvement are factors in an accident, then the accident is more likely to be major and non-fatal. Lastly, rule 5 shows an association between aggressive driving behavior and if the age range of involved ones is 20 to 24 then it leads to non-fatal accidents.

Overall, these rules suggest that aggressive driving is a major factor in accidents involving pedestrians and young drivers. They also suggest that collisions involving aggressive driving are more likely to be serious ones.

The same technique of association rule mining has been implemented on the RTA dataset. Table 4.26 represents the useful associations related to the ‘nature of accident’ factors.

Rule 1 indicates that over speeding between 12-18 hours (12:00 PM to 6:00 PM) is linked to minor, non-fatal accidents. The confidence value implies that when accidents happen during the specified time range and involve over speeding, there is a 57% likelihood of the consequence being minor and non-fatal.

Rule 2 indicates careless driving behavior is strongly associated with single fractures and non-fatal injuries. The very high confidence of 0.99 implies that when the accident involves carelessness, there is a 99% chance of the consequence being a single fracture and non-fatal.

The rule 3 is similar to the rule 1 and states that carelessness during 12-18 hours on weekdays often results in non-fatal accidents. The confidence value shows that when accidents happen during the specified time range on weekdays and involve carelessness, there are 72% chances of non-fatal injuries.

In order to mitigate the higher rate of non-fatal accidents occurring between the hours of 12 PM and 6 PM, especially on weekdays, it is recommended to give priority to specific enforcement measures and awareness efforts that specifically target speeding and carelessness during these peak periods. It is necessary to encourage attentive and safe driving behaviors among all drivers.

Next, the subgroup of nature of accident is combined with the subgroup of environmental factors and table 4.27 represents the topmost interesting association rules, the explanation is given as follows:

Rule 1 shows a strong association between accidents that occur during the night hours 18 to 24 hours (6:00 PM to 12:00 AM) and the consequence being a single fracture. The value of confidence shows that in these conditions there is a 100% chance of a single fracture.

Rule 2 indicates that the injury severity of minor, non-fatal accidents are more common during rainy days between 12 PM and 6 PM. The value of confidence implies that when accidents happen during the specified time range on rainy days, there is a 78% chance of minor and non-fatal injuries.

The last rule no 3 shows that over speeding during rainy days is associated with minor injuries. The value of confidence states that when these conditions are met there is a chance of 52% minor injuries.

In light of the above rules, in order to reduce the increased risks of accidents during nighttime driving and rainy daytime circumstances, it is recommended to give priority to specific education and enforcement actions. There is a need to promote careful driving behavior in rainy weather, particularly from midday to 6 PM, and rigorously enforce speed restrictions under such circumstances.

Following this, the combination of the nature of accident and vehicle factors are undergone with this algorithm and table 4.28 represents the topmost associations whose analysis is given as follows:

The higher lift value of rule 1 states that over speeding on motorcycles during weekends is strongly associated with injury severity as minor. The rule 2 and 3 state that over speeding in cars and rickshaws leads to non-fatal and minor injuries. The high confidence value of rule 2 and rule 3 shows that there are 64% and 71% chances of non-fatal and minor injuries respectively.

Rule 4 is of much importance that shows careless driving behavior in cars is highly associated with non-fatal accidents. The high confidence value of these rules states that there are 95% chances of non-fatal collisions when careless driving behavior of cars is involved.

Rule 5 shows that over speeding in vans is linked to minor injuries with a confidence of 66%. Lastly, rule 6 is also important that shows head injuries are common in accidents involving motorcycles. The confidence value of this rule shows that around 60% of head injuries are related to motorcycle collisions.

From these rules, it is suggested to do targeted educational campaigns to address the risks of over speeding and careless driving especially to the car drivers. Furthermore, it is highly recommended to pay particular attention to motorcycle speeding during weekends and encourage the use of helmets among motorcyclists to mitigate the head injury risks.

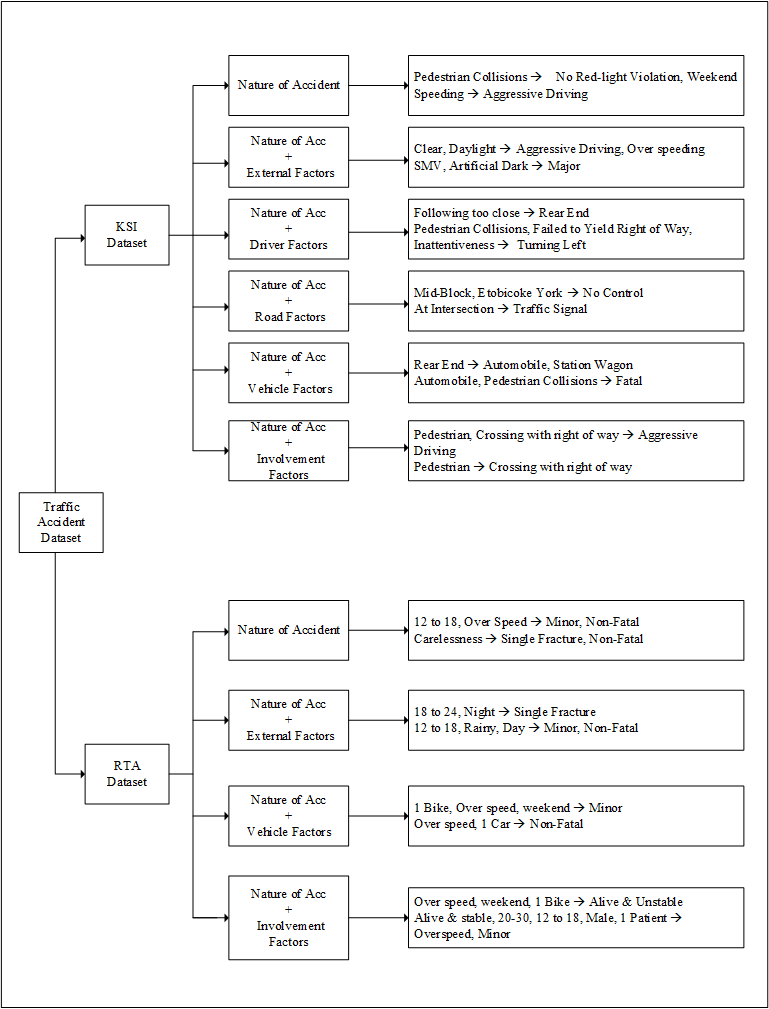
Lastly, the same algorithm has been applied to the combination of subgroups of nature of accident and involvement factors, the table 4.29 shows the top association rules and the explanation is given below:

Rule 1 indicates over speeding on motorcycles during weekends is strongly associated with "alive and unstable" condition of involvement. The value of confidence suggests that under these conditions there is a 60% likelihood of the patient status as ‘alive and unstable’. Rule 2 states that stable and alive male patients of age between 20-30 years involved in accidents between 12 PM and 6 PM are likely to have been over speeding, resulting in minor, non-fatal injuries. The confidence value of this rule shows that there are 58% chances of the consequence being over speeding, injury severity as minor, and accident class as non-fatal under the conditions specified.

Rule 3 shows that the persons with an education level of matric are involved in over speeding motorcycles often experience minor injuries with "alive and unstable" status. The confidence of 0.63 implies that when accidents involve individuals with the specified characteristics, there is a 63% likelihood of the consequence being alive and unstable with injury severity as minor. Lastly, rule 4 indicates that male drivers/riders of age under 20 years old involved in accidents during the peak hours of 12 PM and 6 PM are likely to have been over speeding, resulting in minor injuries and are found in stable conditions after the collision. The confidence value shows that there is 53% chance of the consequence under the specified characteristics of the collision.

These rules suggested that there is a need to curb motorcycle speeding, especially during weekends, to mitigate the risk of severe injuries. Target interventions toward young male drivers, who are more likely to engage in over speeding.

The following figure shows the top most important associations of both of these datasets:



**Fig 4.18 Top Association Rules of KSI and RTA Datasets**

The above associations provide a lucid depiction of the factors contributing to severe collisions in a developing country like Pakistan, as opposed to a developed nation like Canada. It can be seen that there are some common factors like over speeding, carelessness and driver’s inattentive that leads to fatal or major injuries. However, there are some distinct associations that also exist that are explained below.

In Pakistan, severe accidents predominantly involve young drivers below the age of 20. Conversely, in Canada, the minimum age bracket for drivers involved in severe collisions tends to be between 20 and 24 years old.

Another notable disparity lies in the nature of major accidents between Pakistan and Canada. In Pakistan, a prevalent cause of fatal and major collisions stems from head injuries resulting from motorcycle accidents. Conversely, in Canada, major accidents often entail collisions between pedestrians or cyclists and automobiles.

Next contrasting factor is the timing of severe collisions. In Pakistan, nighttime significantly contributes to higher probability of severe collisions compared to daytime incidents. Conversely, in Canada, daytime emerges as a more critical period for major collisions.

The final contrasting factor pertains to the impact of rainy weather, in Pakistan, rainy weather is associated with a higher incidence of major accidents, whereas in Canada, fewer accidents occur during rainy conditions. This is due to the fact that drivers in Canada exhibit greater caution and restraint during rainfall, leading to a reduction in aggressive driving behaviors compared to their counterparts in Pakistan.

Both countries can enhance road safety through targeted enforcement measures, including stricter licensing and education for young drivers. In Pakistan, enforcing helmet usage for bike riders and raising awareness about adverse weather conditions, alongside improving lighting conditions, are crucial. Meanwhile, Canada should prioritize educating young drivers and enhancing pedestrian and cyclist safety. Addressing behaviors such as inattentiveness, aggressive driving, and carelessness is imperative in both countries.

## 4.6 Deployment

The use of this system could be very beneficial for the world and can help greatly in saving the precious lives of people. Following are some of the suggestions for the deployment of this system:

Firstly, the main idea is to use this system as an in-vehicle and driver-focused learning system. The main vision of this system is an AI-powered system embedded in your car’s dashboard, ready to revolutionize your driving experience.

It actively learns your every move. Through GPS and sensors, it maps your usual routes, recognizing traffic patterns, speed limits, and even hidden bumps or tricky corners. The system goes beyond the roads. By analyzing your acceleration, braking, and turning style, it builds a unique profile of your driving personality, understanding your risk tolerance and comfort level. It becomes an extension of your mind, even factoring in real-time weather reports and sensor data to adjust its advice based on rain, fog, or slippery conditions.

It considers your driving style, the current weather, and even your past preferences to suggest the safest and most comfortable path. The system can actively scan the environment using external cameras and sensors and offers visual and audio alerts to keep you focused and aware.

The proposed system can prove itself as a revolution in road safety. It empowers the drivers to be confident and informed in real-time while contributing to a future where accidents are minimized, and lives are saved. So, the proposed system can become the proactive co-pilot that guides drivers toward a safer, smarter, and truly personalized driving experience.

Secondly, the proposed model can serve the purpose of traffic management systems. The system can dynamically allocate routing to avoid high-risk areas, targeted enforcement to prevent accidents, and real-time hazard warnings to alert drivers of potential dangers.

It is capable of adjusting the traffic lights automatically to prioritize safe flow, police focusing on areas predicted to have severe collisions, and drivers receiving dynamic warnings through billboards, mobile app or their own vehicles.

# CHAPTER 5

# SUMMARY

## 5.1 Summary

In the domain of accident analysis several studies have been conducted either based on binary classification using the base algorithms or using unsupervised machine learning techniques.

The aim of this research is to provide a reliable tool for injury severity prediction in traffic accident collisions and to find out the hidden factors that lead to collisions and fatal or major injuries in involvements. The research is conducted using two datasets, the first one is the Killed or Seriously Injured (KSI) dataset obtained from the Toronto Police Safety Public Safety Data Portal, and the second one is the local Road Traffic Accident (RTA) dataset obtained from Punjab Emergency Services, Rescue 1122 Rawalpindi. Each of these datasets consists of several collision-related features, demographics of involvement, vehicle factors, several environmental factors and road conditions that helped a lot in understanding the hidden factors.

In the first chapter of the study, the detailed introduction of the problem, its significance and innovation has been discussed. The second chapter consists of literature review, the third chapter consists of the proposed research methodology. The fourth chapter shows the actual implementation of the proposed methodology along with the statistical analysis. The study has implemented several ensemble machine-learning models to develop predictor models. It has been found that the Light Gradient Boosting Machine is the most suitable ensemble technique for this problem. The Light GBM has achieved an accuracy of 75% for both of these datasets. Next, the study has found several interesting associations that lead to fatal or major collisions and provide a comprehensive comparison of these factors found from both countries and lastly, based on these factors the study has also suggested some recommendations. The chapter also presents a discussion about these findings and the proposed deployment. Lastly, chapter five presents the summary.

## 5.2 Thesis Contributions

The thesis is comprised of data collection, the KSI dataset is obtained from online resources and the RTA dataset is obtained from the office of Punjab Emergency Services, Rescue 1122 Rawalpindi. Then, data preprocessing includes data cleansing, feature engineering, feature selection, normalization and lastly applying the sampling technique of SMOTE to solve the class imbalance issue. Then the study implemented several basic and ensemble ML models to train different ML predictor models on both datasets. These models are then evaluated using a 10-fold cross validation technique on several measures including accuracy, precision, recall, f1-score and AUC. The Light Gradient Boosting Machine algorithm has achieved the best performance among all. Lastly, the study has found out the factors that lead to collisions and fatal or major injuries and provides a comparison of these factors mined from both of these datasets. Then the study has also suggested some recommendations based on these factors to minimize the collisions and injury severities.

## 5.3 Future Work

In this proposed work, the study has considered several basic and ensemble machine learning classification algorithms, in future new and some advanced machine learning or deep learning methods can be tried on similar datasets. There are different sampling techniques available, current study has implemented SMOTE, and in future we can implement some other sampling technique or some other version of SMOTE. Furthermore, the study has used one international and one local dataset, the same can be implemented on more different datasets to generalize the findings.

# LITERATURE CITED

Yadav, J., Batra, K., & Goel, A. K. (2021, August). A Framework for Analyzing Road Accidents Using Machine Learning Paradigms. In Journal of Physics: Conference Series (Vol. 1950, No. 1, p. 012072). IOP Publishing.

Shanshal, D., Babaoglu, C., & Başar, A. (2020). Prediction of Fatal and Major Injury of Drivers, Cyclists, and Pedestrians in Collisions. Promet-Traffic&Transportation, 32(1), 39-53.

Watkins, E., Kloc, M., Weerasuriya, S., & El-Hajj, M. (2017, January). Collision analysis of driving scenarios. In 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 1-7). IEEE.

Alam, Z. (2018). Improving Road Safety in India Using Data Mining Techniques. In Data Science and Analytics: 4th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2017, Gurgaon, India, October 13-14, 2017, Revised Selected Papers 4 (pp. 187-194). Springer Singapore.

Hébert, A., Guédon, T., Glatard, T., & Jaumard, B. (2019, December). High-resolution road vehicle collision prediction for the city of montreal. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 1804-1813). IEEE.

Emu, M., Kamal, F. B., Choudhury, S., & Rahman, Q. A. (2022). Fatality prediction for motor vehicle collisions: mining Big Data using deep learning and ensemble methods. IEEE Open Journal of Intelligent Transportation Systems, 3, 199-209.

Gan, J., Li, L., Zhang, D., Yi, Z., & Xiang, Q. (2020). An alternative method for traffic accident severity prediction: using deep forests algorithm. Journal of advanced transportation, 2020, 1-13.

Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K., & Nawrine, F. (2019, June). Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. In *2019 7th international conference on smart computing & communications (ICSCC)* (pp. 1-5). IEEE.

Parathasarathy, G., Soumya, T. R., Das, Y. J., Saravanakumar, J., & Merjora, A. A. (2019, February). Using hybrid data mining algorithm for analysing road accidents data set. In 2019 3rd International Conference on Computing and Communications Technologies (ICCCT) (pp. 7-13). IEEE.

Kumeda, B., Zhang, F., Zhou, F., Hussain, S., Almasri, A., & Assefa, M. (2019, June). Classification of road traffic accident data using machine learning algorithms. In 2019 IEEE 11th international conference on communication software and networks (ICCSN) (pp. 682-687). IEEE.

AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019, April). Comparison of machine learning algorithms for predicting traffic accident severity. In 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT) (pp. 272-276). IEEE.

Saleem, A., Asif, K. H., Ali, A., Awan, S. M., & Alghamdi, M. A. (2014, December). Pre-processing methods of data mining. In *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing* (pp. 451-456). IEEE.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge discovery in database*, 229-248.

Antonie, L. (2008). INTRODUCTION TO DATA MINING ASSOCIATION RULES. *University of Guelph*.

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216).

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, *2*(1), 58-64.

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997, June). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data* (pp. 255-264).

Hahsler, M., Grün, B., & Hornik, K. (2005). arules-A computational environment for mining association rules and frequent item sets. *Journal of statistical software*, *14*, 1-25.

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2018, March). Relationship between student engagement and performance in e-learning environment using association rules. In 2018 *IEEE world engineering education conference (EDUNINE)* (pp. 1-6). IEEE.

Antonie, L. (2008). INTRODUCTION TO DATA MINING ASSOCIATION RULES. *University of Guelph*.

World Health Organization (WHO) 2022. [https://www.who.int/news-room/fact-](https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries)

[sheets/detail/road-traffic-injuries](https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries)

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational intelligence and neuroscience*, *2018*.

Edelstein, H. A. (1998). *Introduction to data mining and knowledge discovery*. Two Crows.

Percoco, M. (2016). The impact of road pricing on accidents: a note on Milan. *Letters in Spatial and Resource Sciences*, *9*(3), 343-352.

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216).

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

Yannis, G., Dragomanovits, A., Laiou, A., Richter, T., Ruhl, S., La Torre, F., ... & Li, H. (2016). Use of accident prediction models in road safety management–an international inquiry. *Transportation research procedia*, *14*, 4257-4266.

Shunshun, W., Changshun, Y., & Yong, S. (2023). A Review of Road Traffic Accident Prediction Methods. *American Journal of Management Science and Engineering*, *8*(3), 73-77.

Kelly, F. (2008). The mathematics of traffic in networks. *The Princeton companion to mathematics*, *1*(1), 862-870.

Keqiang, H., Jibao, Y., & Sijing, W. (2005). Analysis of dynamic factors of debris landslide by means of the model of quantitative theory—using the Xintan landslide, China, as an example. *Environmental geology*, *48*, 676-681.

Li, W., Zhao, X., & Liu, S. (2020). Traffic accident prediction based on multivariable grey model. *Information*, *11*(4), 184.

Tongyuan, H., & Yue, W. (2007, December). Forecasting Model of Urban Traffic Accidents Based on Gray Model GM (1, 1). In *Second Workshop on Digital Media and its Application in Museum & Heritages (DMAMH 2007)* (pp. 438-441). IEEE.

Pourroostaei Ardakani, S., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. *Sustainability*, *15*(7), 5939.

Toronto Police Services. (2023). KSI. Public Safety Data Portal. Retrieved from <https://data.torontopolice.on.ca/pages/total-ksi>

Rezashoar, S., Kashi, E., & Saeidi, S. (2023). Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity (Case Study: United Kingdom from 2010 to 2014).

Ahammad, S. H., Sukesh, M., Narender, M., Ettyem, S. A., Al-Majdi, K., & Saikumar, K. (2023). A Novel Approach to Avoid Road Traffic Accidents and Develop Safety Rules for Traffic Using Crash Prediction Model Technique. In *Micro-Electronics and Telecommunication Engineering: Proceedings of 6th ICMETE 2022* (pp. 367-377). Singapore: Springer Nature Singapore.

KUYUMCU, Z. Ç., Aslan, H., & YURTAY, N. (2023). Identifying Interrelated Factors of Fatal and Injury Traffic Accidents Using Association Rules. *Turkish Journal of Civil Engineering*, *34*(5), 55-80.

Datu, N. H. (2023, March). Road traffic accidents analysis using association rule mining and descriptive analytics. In *AIP Conference Proceedings* (Vol. 2508, No. 1). AIP Publishing.

Road Traffic Accident Dataset. (2023). RTA. The Punjab Emergency Service, Rescue 1122 Rawalpindi. Available at: <https://doi.org/10.7910/DVN/4VGTDR>

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Saher, N., & Madni, H. A. (2020). Comparison analysis of tree based and ensembled regression algorithms for traffic accident severity prediction. *arXiv preprint arXiv:2010.14921*.

Meißner, K. (2022). *Exploratory Road Accident Analysis-Identification of Interesting Relationships over Time to Support Road Safety Planning* (Doctoral dissertation, Stiftung Universität Hildesheim).

Das, S., Le, M., & Dai, B. (2020). Application of machine learning tools in classifying pedestrian crash types: A case study. *Transportation safety and environment*, *2*(2), 106-119.

Ben-Shachar, M. S., Patil, I., Thériault, R., Wiernik, B. M., & Lüdecke, D. (2023). Phi, Fei, Fo, Fum: Effect Sizes for Categorical Data That Use the Chi-Squared Statistic. *Mathematics*, *11*(9), 1982.

Ali, M. (2020). PyCaret: An open source, low-code machine learning library in Python. *PyCaret version*, *2*.