A Comparative Study on Traffic Collisions Severity using Machine Learning Approaches

by

Rajvi Rathod

A thesis submitted in partial fulfillment of the requirements for the degree of

MSc Computational Sciences

The Faculty of Graduate Studies

Laurentian University

Sudbury, Ontario, Canada

© Rajvi Rathod, 2021
THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
Laurentian Université/Université Laurentienne
Faculty of Graduate Studies/Faculté des études supérieures

Title of Thesis
Titre de la thèse   A Comparative Study on Traffic Collisions Severity using Machine Learning Approaches

Name of Candidate
Nom du candidat   Rathod, Rajvi

Degree
Diplôme   Master of Science

Department/Program
Département/Programme   Computational Sciences

Date of Defence
Date de la soutenance   June 02, 2021

APPROVED/APPROUVÉ

Thesis Examiners/Examinateurs de thèse:

Dr. Kalpdrum Passi
(Supervisor/Directeur(trice) de thèse)

Dr. Ratvinder Grewal
(Committee member/Membre du comité)

Dr. Hafida Boudjellaba
(Committee member/Membre du comité)

Approved for the Faculty of Graduate Studies
Approuvé pour la Faculté des études supérieures
Tammy Eger, PhD
Vice-President Research (Office of Graduate Studies)

Dr. Rossitza Marinova
(Vice-rectrice à la recherche (Bureau de la recherche))

(External Examiner/Examinateur externe)

Laurentian University / Université Laurentienne

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

I, Rajvi Rathod, hereby grant to Laurentian University and/or its agents the non-exclusive license to archive and make accessible my thesis, dissertation, or project report in whole or in part in all forms of media, now or for the duration of my copyright ownership. I retain all other ownership rights to the copyright of the thesis, dissertation or project report. I also reserve the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that this copy is being made available in this form by the authority of the copyright owner solely for the purpose of private study and research and may not be copied or reproduced except as permitted by the copyright laws without written authority from the copyright owner.
Abstract

Road Traffic collisions and congestion are amongst one of the most crucial issues in the modern world. Every year, traffic collisions cause multiple deaths and injuries. It leads to economic losses as well. According to WHO, approximately 1.35 million people are losing their lives, with 20 to 50 million people face non-severe injuries every year because of road collisions. Hence, there is a need to create a prediction system that can help determine relations between various factors such as climate, types of automobile, driving pattern etc., to predict the severity of the collisions. It helps to improve public transportation, allowing safer routes and thus avoid the chances of high severity cases to make the roads safer. Smart cities concept can be helpful to handle modern problems. Accurate Models for predicting collision severity has become a significant challenge for transportation systems. This research establishes a procedure for identifying important parameters affecting collision severity and creates a relationship between human and environmental factors using several Machine Learning (ML) techniques. Among different types of ML techniques, classification algorithms have been applied for categorizing the level of severity. Supervised algorithms such as Random Forest (RF), Decision Trees (DT), Logistic Regression and Naïve Bayes have been used. A comparative study among performance and accuracies of various algorithms is also mentioned. These algorithms were tested on a dataset that contains historic data for collisions in the U.S and their severity levels. This study's findings show Random Forest with the best accuracy and identify the time of day, duration of an collision, and Point of Interest (POI) features as the influential parameters.

**Keywords:** Traffic Collision, Machine Learning, Collision Severity, Logistic Regression, Naïve Bayes, Random Forest, Python.
Acknowledgements

Throughout the thesis completion process, I would be glad to give credits to my thesis supervisor, Dr. Kalpdrum Passi as he has a big hand in assisting me throughout the task. His innovative ideas and his corrective solutions have given me guidance and a sense of direction through the process.

Additionally, when it comes to emotional support, my friends and family have a big hand in maneuvering the project. I am fortunate enough to have the right people to lead me and support me when I required direction. Therefore, I am sincerely grateful for the needful intervention and collaboration of my supervisor and my close ones.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thesis Defense Committee</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>ix</td>
</tr>
<tr>
<td>Chapter 1</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Causes of Road Collisions</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Objective</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Motivation</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Outline</td>
<td>5</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>6</td>
</tr>
<tr>
<td>Literature Review</td>
<td>6</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>15</td>
</tr>
<tr>
<td>Data and Preprocessing</td>
<td>15</td>
</tr>
<tr>
<td>3.1 Dataset</td>
<td>15</td>
</tr>
<tr>
<td>3.2 Data Extraction and Selection</td>
<td>17</td>
</tr>
<tr>
<td>3.3 Data Pre-processing</td>
<td>22</td>
</tr>
<tr>
<td>3.3.1 Missing Values</td>
<td>22</td>
</tr>
<tr>
<td>3.3.2 Time of the Day</td>
<td>24</td>
</tr>
<tr>
<td>3.3.3 Weather features</td>
<td>26</td>
</tr>
<tr>
<td>3.3.4 POI Features</td>
<td>30</td>
</tr>
<tr>
<td>3.3.5 Description of Collisions</td>
<td>32</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>34</td>
</tr>
<tr>
<td>Classification Methods</td>
<td>34</td>
</tr>
<tr>
<td>4.1 Overview of Methodology</td>
<td>34</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 3.1: Missing Values. 23
Table 3.2: Weather parameter conversion 28
Table 5.1: Confusion Matrix. 50
Table 5.2: Logistic Regression Performance Metrics 53
Table 5.3: Naïve Bayes Performance Metrics 55
Table 5.4: Random Forest Performance Metrics. 57
Table 5.5: K-fold cross validation 58
Table 5.6: Neural Networks Performance Metrics 59
Table 5.7: Summary Table 62

LIST OF FIGURES

Figure 3.1: Extract Transform and load Process 17
Figure 3.2: Sources of Data 18
Figure 3.3: Severity count by source 19
Figure 3.4: Count of Collisions by Year (MapQuest) 19
Figure 3.5: Count of Collisions by Year(BIng) 20
Figure 3.6: State-Wise Counts of Collisions 21
Figure 3.7: Count of collisions by weekday 25
Figure 3.8: Comparison of collisions during day and night 26
Figure 3.9: Count of collisions by weather features 29
Figure 3.10: Count of collisions in POI features 31
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>POI</td>
<td>Point of Interest</td>
</tr>
<tr>
<td>SEM</td>
<td>Structure equation model</td>
</tr>
<tr>
<td>HBDM</td>
<td>hazard-based duration model</td>
</tr>
<tr>
<td>DCM</td>
<td>discrete choice model</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Networks</td>
</tr>
<tr>
<td>CTM</td>
<td>classification trees model</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-Separated Values</td>
</tr>
<tr>
<td>EDA</td>
<td>exploratory data analysis</td>
</tr>
<tr>
<td>CART</td>
<td>classification and regression tree</td>
</tr>
<tr>
<td>MLR</td>
<td>Multinominal Logistic Regression</td>
</tr>
<tr>
<td>REST</td>
<td>Representational State Transfer</td>
</tr>
<tr>
<td>NHTSA</td>
<td>National Highway Traffic Safety Administration</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Traffic collisions are the main cause of death and injuries. Property damages caused by collisions further adds to economic losses as well. Due to population growth, there is rise in the number of vehicles on the road, which results in increased collisions. Traffic congestion is also one of the outcomes due to road collisions. It further leads to negative impacts on the environment, time, and money. Due to urbanization, there is a steady development of cities around the globe. According to a study [1], by the year 2050, cities will be home to 70% of the world's population. Thus, urbanization leads to improving the lives of people, healthcare, education, and transportation. However, at the same time, it gives rise to multiple issues and challenges. The concept of smart cities will help us to address these issues. It provides an infrastructure to collect data and improve the living environment.

U.S transportation is majorly facilitated by all four means: rail, road, air and water. The majority of travel occurs by air for long distances and roads for short distances. Understanding Traffic collisions' primary characteristics is an important step for developing effective countermeasures in traffic safety. Although there is a downtrend in the number of deaths due to road collisions in developed countries, the cost of traffic-related injuries and property loss remains a significant concern. Road collisions result in the death of 1.35 million people around the globe each year. In the last few years, many researchers are studying the impact of influential parameters affecting the collision severity and related information. Road collisions have negative impacts on people, nature, and society. Air pollution and time
delay for essential services like ambulance are considered the external parameters involved in a traffic collision. To minimize the financial and physical cost of traffic collisions, prediction models can be developed to determine the collision types or severity levels. These severity levels can create an infrastructure/facility as a countermeasure for the collision situation [1]. For example, if there are many collisions near a junction with no stop signal, the authorities can install a signal board.

In recent years, due to the development of traffic detection techniques, researchers can easily gather data and do a thorough review on traffic collisions with duration time to predict collision duration in a highly accurate manner [2]. Many statistical methods have been implemented for analyzing and predicting collision time. Structure equation model (SEM) [3], hazard-based duration model (HBDM) [4], discrete choice model (DCM) [5], linear regression [2], [6], [7], Bayesian classifier [8] and probabilistic distribution analyses [9], [10], are several of these approaches. Different ML (Machine Learning) methods have been implemented to estimate and predict the traffic incident duration time. Some of these methods are the following: Support Vector Machine (SVM/RVM) [11], Artificial Neural Networks (ANN) [12], [13], Genetic Algorithm (GA) [13], Decision Trees (DT) and Classification Trees Model (CTM) [1], [14]. A hybrid method [15] has been created to predict the traffic incident duration and, hence, apply the methods described in previous statements.

Various factors determine the traffic collision duration that includes few parameters, which cannot be observed. These factors turn out to be the traffic incident duration very heterogeneous by nature. One of the possible approaches for improving prediction accuracy is to use a dataset with many records. Thousands of incident reports are utilized in previous studies, while few datasets contain over 30,000 incidents.
1.1 Causes of Road Collisions

Human error is one of the major causes of a road collision. The driving pattern also determines the collision severity [11]. Some of the common driving behavior that results in collisions are listed below:

- **Over Speeding**
  
  - A vehicle moving at a higher speed will have more chances of impact during the crash, and hence it has higher chances of severity.
  
  - While driving at a higher speed, the drowsiness (or fatigue) makes an error of judgement to judge the forthcoming events. Due to sudden braking, the vehicle skids and results in a crash or collision.

- **Drunken Driving**
  
  - This is the second most common reason for a collision. Driving of vehicle under the impression of alcohol increases the chances of losing control over the vehicle, and an event of an collision is most likely to happen.

- **Distractions**
  
  - Sometimes drivers try to attend calls while driving. This hampers brain activity of handling vehicles and results in loss of control over the vehicle in sudden braking or other activity.

- **Jumping the signal**
  
  - To save some time, drivers usually skip the traffic signal. This directly results in an collision since other vehicles are coming from the intersection.
  
  - The main cause of traffic congestion is collisions at the intersection.
Weather Conditions
- Weather conditions perform a major part in the incident of a collision.
  - Rainfall has a positive correlation with high severe road collisions.
  - Snow and fog also come on the list.
- However, most of the collisions (High and Low severity) occur during the daytime and in clear weather.

Vehicle Factors
- A small number of injuries are caused by a vehicle's mechanical failure. It can be any type of component or part failures like brake, tire or steering. Worn tires are also one of the reasons for vehicle failure.

1.2 Objective

The key goal of this study is to develop a Machine Learning model that can predict the severity of collisions based on historical data of similar events. The outcome of this study will describe the influential parameters that affect the severity of collisions. The data is taken from the U.S collisions database, including parameters related to weather, amenities, description of the collision, geographical values and other details related to collisions. Statistical analysis is carried out to find relations between different parameters. The parameters that directly correlate with the collision severity are further taken for model building. For a given collision, if the data is delivered without any additional information, such as driver characteristics or vehicle specifications, this model is intended to predict whether the collision will be serious or minor. The models can then be used to take further actions on high severity collision-prone areas.
1.3 **Motivation**

Traffic collisions kill and injure people on a regular basis. It costs a considerable amount of financial loss to a nation. To ensure the safety of the public and reduce road collisions, there is a need to create a predictive model. It is essential to perform this research to study all the available parameters to reduce the financial loss caused due to damaged property. Road collisions result in the death of 1.35 million people around the globe each year. In the last few years, many researchers are studying the impact of influential parameters that are affecting the collision severity and related information. Severe injuries are responsible for a large portion of the damages. Reducing traffic collisions, especially serious ones, is often a challenging task. One of the two main approaches, the proactive approach is a method for resolving road safety issues. It works on reducing risky situations in the first place. Collision prediction and severity prediction is must for effective implementation of this approach. We may be able to do this work in real-time situations if the patterns could identify how these serious collisions happen and the main parameters.

1.4 **Outline**

The outline of this thesis work is as follows: Chapter 1 briefly introduces the current traffic collision scenario. Moreover, it highlights various economic and physical losses that millions of people suffer from. Chapter 1 also focuses on the different factors that are common among the drivers for a high severity collision. Chapter 2 describes some of the previous works carried out on a similar type of problem statement. Chapter 3 shows EDA (exploratory data analysis) and some important inferences that are used for model building. Chapter 4 demonstrates various supervised learning algorithms applied and their accuracies. It also shows statistical tests for showing the relationship between variables. Chapter 5 concludes
and discusses on future work that can be done to enhance the model.

Chapter 2

Literature Survey

The cost of collisions-related incidents and deaths has a negative impact on society. Due to the rapid growth of vehicles, severe crashes and collisions have increased in recent years. Factors affecting the severity of collisions has been studied in different contexts [16]. Various factors related to traffic collisions include users, vehicle type, road signs and weather parameters. These factors provide useful information to find the relationships that can be used to prevent high severity collisions [17]. Pakgohar et al. [4] used CART (classification and regression trees) and MLR (Multinominal Logistic Regression) for analyzing human factor roles in road collisions using SPSS based on three output variables (Fatal, Injury and Non-Injury). They concluded that the CART method resulted in higher accuracy when compared with MLR. Moreover, it was simpler and easier to interpret the results.

In recent times, social networks such as Facebook and Twitter data have been widely used as sources for traffic events. An example of this could be an event of traffic collision or congestion. With this data's help, we can get the areas where most of the collisions occur. SUM (Status Update Message) includes geographical details and current traffic state, which users can obtain while driving. This helps to avoid heavy traffic areas [1], [18].

Due to continuous increase in data generation from different applications, data mining has proved to be a very important research areas in recent times. This data can prove to be very useful for analyzing and obtaining meaningful inferences. Data mining is also known by the
term KDD (Knowledge Discovery in Databases), which analyzes and divides data from various dimensions and gets patterns from it [15], [19].

Liu et al. [20] suggested a new clustering approach based on the overcrowding of regions to define high traffic areas with the help of various information in the region. This approach is known as mobility-based clustering. In this method, high speed may show low congestion and low speed may show high congestion. There are no percentage results that show a relationship of the vehicle speed with other factors. Lécué et al. [21] showed a STAR-CITY (Semantic Traffic Analytics and Reasoning for City) framework. This is a real-time surveillance application throughout the city of Dublin’s public transportation. This work is useful in scenarios that uses a sensor data stream. The STAR-CITY framework calculates crowding by studying all the road traffic situations as weather information and collisions. An a priori algorithm is used to find association rules between different traffic scenarios by analyzing historical and real-time data.

Twitter, Facebook, and other social networks have been recently applied as sources of information related to traffic incidents. These events and users of the networks form a sensor. Anastasi et al. [1] proposed a SMARTY concept (social-sensing approach) to develop an Information and Communication Technology (ICT) Platform for smart cities that provides transport and mobility related tools and services. This work uses social networks to monitor and analyze activities of the network. It also utilizes real-time information obtained from social networks like Twitter and Facebook to find traffic events like collisions and congestion. It also suggests optimal tracks to the users. Heterogeneous data is pre-processed and analyzed to find knowledge patterns from the data. The authors developed a labelled dataset containing 500 tweets classified into two SUM (Status Update Messages) classes.
They used the C4.5 classifier along with 10-fold Cross-Validation and achieved a classification percentage of around 93.73%.

The traffic collisions predictive models are crucial in decision-making to reduce collisions on highways. Machine learning has become a very important tool in the predictive power of decision-making. Numerous studies [12], [13], [14] use ML algorithms to identify collision causes due to transport operations and management. Krishnaveni and Hemalatha [14] analyzed the traffic collisions in the city of Hong Kong using ML algorithms to predict the severity of injury and causes of collisions. Random Forest classifier outperformed all other algorithms. Beshah and Hill [12] employed Decision Tree (J48), Naive Bayes, and K-Nearest Neighbours to evaluate the prediction models for causes to traffic related collisions and the environmental reasons for severity of the injuries. Highest accuracy of 79.94% was achieved by the prediction models and the knowledge was represented in the form of rules using PART algorithm [12].

Dong et al. [22] used a supervised fine-tuning module and an unsupervised feature module to predict traffic collisions. The feature selection module reduces the dimensionality and preserves the information in the original data and finds relationships between the variables and the feature representations. The relationships between important parameters and the collision outcomes provide insight into the collision causes. Chen et al. [9] used two-year data on vehicle crashes in New Mexico city to predict the severity of injuries due to collision using SVM classifier. The outcomes indicated that polynomial kernel in the SVM gives better performance than the Gaussian RBF kernel.

Zhang et al. [23] created a hybrid method combining factor analysis method with an Ordered
Probit Model (OPM) for comprehensive analysis of work-zone collisions due to crashes. The analysis showed that there was a strong association between the crash type and the work zone severity. Osman et al. [24] applied OPM and logit models for detecting factors in injuries due to large truck collision in the work zones. The authors further noticed that more severe injuries were caused during daytime while speeding in rural areas. Ghasemzadeh and Ahmed [25] likewise created the OPM to investigate the relationship between the weather and work zone crashes in the severity of injuries. The results indicated that weather and lighting conditions caused severe injuries due to crashes in work zones.

Bharadwaj et al. [6] conducted a study that showed driving behavior as a major risk factor in severe injuries due to crashes in the work zone. Wei et al. [7] found in the study that injury rate increased by 72.7% during nighttime when the vehicles were speeding under poor lightning conditions in the work zone. Sze and Song [5] employed a multinomial logistic regression model to check at the common factors of work-zone related collisions to determine how much of a connection there is between the factors and the severity of the injuries. The authors concluded that the vulnerability of road users, time of the day and heavy vehicles were significantly related to the severity of injuries in work zone crashes.

Long et al. [8] studied the factors that influence work zone collisions. According to the results, rear-end type crashes was the most important factor, which intensifies the crash severity. Harb et al. [3] concluded in their study that high risk factors in work zone collisions include gender, age, the type of lane, the weather and lighting conditions, and the presence of drugs and alcohol. Yan et al. [10] investigated risk factors for rear-end crashes on main roads with traffic signals using logistic regression. Several environmental conditions were discovered to be directly related to the probability of a rear-end collision.
Effati et al. [26] proposed a geospatial technique based on machine learning. Artificial Neural Networks and fuzzy systems (CANDIS) were combined for a geospatial analysis to discover spatial and non-spatial factors to predict the severity of injuries in collisions from a dataset collected over a four-year period in Iran. The CANDIS was trained using Neurodimension's Neuro Solutions software and MATLAB. SVM outperforms CANFIS with an accuracy of 85.49% against 76.44%. Mohamed (2014) [27] used a multi-class SVM with a Gaussian Radial Basis Function kernel for predicting the causes of road traffic collisions in Dubai. The SVM model achieved an accuracy of more than 75%.

Perone [28] conducted an analysis using Support Vector Machines, Logistic Regression, Random Forests (RF), K-Nearest Neighbors and Naive Bayes on real-life collisions to develop collision severity prediction models. The author created a framework using Django. Also, the Pandas library was used for data analysis, and the scikit-learn framework was used for pre-processing and building predictors. Logistic Regression and Support Vector Machines gave the best AUC (Area Under the Curve) scores. The dataset used in this research lacked sufficient information on drivers, vehicles, and targets which may have a major impact on the experimental results. In addition, the author did not apply feature selection methods. Liang [29] suggested an ITS (intelligent transportation system) based on cloud computing and IoT platform. The complete analysis was performed using MATLAB. The Ant Colony Algorithm (ACA) was used to improve an SVM model for automated collision detection.

The study in [30] focuses on traffic collisions of Shanghai Expressway from April to June 2014. In this research, a different type of parameters responsible for road collisions was studied using association rule mining. The strict rules are hidden in item sets generated by
associative mining techniques that often reveal the connection between collision influencing factors, which can then be used to prevent collisions. These rules could have been used to investigate common road collision scenes. Weak rules will evolve from association mining. These methods were effective according to the results of the experiments. As a result, an automated modelling algorithm based on association rules was developed to facilitate the successful implementation of association rule mining in the intelligent transportation system.

The study in [31] establishes models for evaluating the severity of injuries in order to determine a list of important factors in traffic collisions. Various ML techniques formulate these models. Various Supervised ML algorithms such as Random Forests (RF), AdaBoost, Naive Bayes (NB) and Logistic Regression (LR) are applied on the dataset. To deal with data imbalance SMOTE technique was used. According to the results, the Random Forest gave the best performance with an accuracy of 75.5% in predicting the severity of injuries in traffic collisions. The paper [32] uses various data mining algorithms to present various models for predicting collision severity. The study focused on real data from previous research and predicted the injury severity level of traffic collision data.

Gakis et al. [33] introduced an AID (Automatic Incident Detection) system based on a Support Vector Machine. The data was collected from output retrieved from inductive-loop detectors. The research aimed to decide on the best effective feature and parameters of the Support Vector Machine that resulted in the most precise collision prediction. The proposed feature selection method was implemented using OpenCV. Moreover, the speed attributes are also used in incident detection. Garber & Zhao (2002) [34] researched important parameters of crashes at different locations within the work zones. The study indicates that angle and rear-end crashes, fixed objects, fatal crashes, and multiple vehicle crashes were more at night.
Li & Bai (2008) [35] showed a crash severity index model where they studied injury and fatal crashes as the dependent variable and categorized the explanatory variables into five categories: driver at fault, road conditions, environmental conditions, crash information, and time. Comparative analysis was conducted between injury severity and crash severity. They found that a large percent of truck involvement, head-on, disregarded traffic control, unfavorable light conditions, alcohol impairment, speeding, and complicated road geometries as the major causes for major fatal crashes while light-duty vehicles, rear-end and followed too close as the major factors causing injury crashes.

Traynor [36] studied the impact of drunk driving on the severity of collisions. Data for the analysis included time, road conditions, weather conditions, location, road type, parties involved, driver characteristics, crash severity, vehicle type at the time of the collision. The results indicated that drunk drivers were more likely to cause serious damage or death than sober drivers. It was estimated that with higher alcohol levels, there are chances of high impacts on injury crashes.

Karlaftis & Golias [37] have studied the collision rates on the rural roads with respect to the road geometry and traffic volume. They found that the most important factors impacting collision rates are pavement condition and geometric design. A comparative study of motor vehicle crashes between ages less than 16 and 25-49 reveals that novice drivers are involved more in fatal crashes than other groups [38].

Arditi et al. [39] have done a comparative study on fatal crashes between night-time and daytime on highway construction work zones for the years 1996-2001 in Illinois. For
determining the frequency of fatal crashes, lighting and weather conditions were selected as independent variables. Results showed that at night-time, construction was more hazardous. Li & Bai [35] studied the effectiveness of measures to control the traffic in highway construction zones. To find the significant risk factors, they build a logistic regression model. The research showed that the centerline, flasher, and flagger effectively reduced traffic crashes in work zones.

Harb et al. [27] studied characteristics for freeway work zone crashes. Multiple and conditional regression modelling was used to conduct this research. The study found that weather conditions, age, roadway geometry, alcohol, and drugs, lightning condition, gender and residence code were important factors affecting crash severity in Florida. According to one of the researches by Williams et al. [40], it showed that drivers without a learner's permit were involved in 57% of fatal collisions, and that the average age of the drivers was about 15 years old. The study also showed speeding, single-vehicle collisions, failure to drive in the proper lane, late-night driving, and driving without a license as crucial parameters for fatal crashes.

Lardelli-Claret et al. [41], Kim et al. [42], and Massie et al. [43] presented the impacts of age and gender on crashes. The drivers of age group 18-20, 60-64 and gender male were high-risk factors for fatal crashes compared to other age groups. Redondo-Calderón et al. (2000) [44] conducted a study comparing traffic collisions under different environmental conditions for single and multiple crashes. The study indicates that drivers of age 18-24 years and driving under alcohol are the significant risk factors.

Pillajo-Quijia et al. [66] have examined influential factors on injury severity for drivers of
light trucks and vans on Spain’s dataset. The goal of this work is to apply machine learning
approaches to explain the severity of driver injuries in run-off-roadway and rollover
incidents. The required categorical variables (driver, vehicle, infrastructure, and
environmental factors) are selected using a Random Forest (RF)-classification tree (CART)
methodology to generate models that classify, explain and predict the severity of such
incidents with high accuracy. Random Forest gave an accuracy of 77% and recall of 0.57.
Chapter 3

Experimental Design and Setup

3.1 Dataset

The U.S collision dataset is a piece of countrywide collision information that logs 49 States of the U.S. It has more than 3 million collisions for all categories (0, 1, 2, 3 ordered from low to high severity). The latest included information is until June 2020. There are three sources of data: MapQuest [45], Bing [46] and through other multiple channels.

Parameters Description:

Parameters in the dataset are divided into the following categories:

1) Weather
   - Weather conditions: snow, fog, rain, thunderstorm, etc.
   - Precipitation: It shows the condensation of atmospheric water vapor in Inches
   - Wind speed: speed of the wind in mph
   - Wind direction
   - Visibility: In miles
   - Pressure: Air pressure in Inches
   - Humidity: In percentage
   - Temperature: In Fahrenheit
   - Wind Chill: In Fahrenheit
   - Airport Code: Closest weather station based on the airport.
2) Address

- Street number, name, side of the road (right/left), city, country, state, zip code, time zone (eastern, central etc.)

3) Traffic

- ID: Unique record for each collision
- Source: MapQuest, Bing, and Others
- TMC: Traffic Message Channel
- Severity: Collision severity (0,1,2,3)
- Start and End time: shows the duration of the collision.
- Longitudes and Latitudes
- Distance(mm): The total impacted road distance due to collision
- Description: Description of an collision in words

4) POI Attributes

- Traffic Signal: shows whether a traffic signal is present or not.
- Turning: Shows presence of loop in that location
- Traffic Calming: Indicates the presence of traffic calming means.
- Stop: whether stop signal board is present or not
- Roundabout: the presence of roundabout
- Railway
- No Exit
- Junction
- Give way
- Crossing
- Amenities
- Bump

5) Time of Day

- Astronomical twilight
- Nautical twilight
- Civil twilight
- Sunrise/Sunset

3.2 Data Extraction and Selection

Data extraction is a process of obtaining data to meet end goals. ETL (Extract, Transform and Load) process is considered as a first step in data ingestion. An ETL tool can extract data from various sources and load it in data warehouse where it can be monitored and analysed for various insights.

![Extract Transform and Load Process](image)

Figure 3.1 Extract Transform and Load Process [47]
As mentioned earlier, there are 3 sources of data: MapQuest, Bing and Others. MapQuest is a free online web mapping service. It provides street level detail information for a variety of countries. MapQuest provides GPS navigations and it also provides features for comparing prices of nearby services like gas. Bing data services provides a REST (Representational State Transfer) interface store, geocode and query spatial data. Bing provides a web search engine which is owned by Microsoft. Among the data sources, MapQuest has highest number of cases logged. Bing is another source of data that has comparatively less data. Since both the data sources have their own methods of classifying the severity of collision, it would not be possible to consider both the sources. To classify the data between low and high severity levels, it is expected to have even distribution of records to avoid “bias” scenario. For selection of data source, we need to further analyse the record counts for each severity levels and for all data sources. Figure 3.2 describes the data records distribution for various sources. However, we need to analyse few other criteria in order to determine the best data source for achieving our end goal.

![Figure 3.2 Sources of Data](image)

Figure 3.2 Sources of Data
We can clearly observe that MapQuest has 68.7%, Bing has 29.5% and other sources has 1.8% of data. However, we are not sure about the severity level distribution of different data sources. As mentioned earlier to avoid bias scenario, we need to check the distribution of severity level before selection of data source.

![Figure 3.3 Severity count by source](image)

**Figure 3.3 Severity count by source**

From Figure 3.3, it can be noticed that “Bing” has all 4-severity levels while “MapQuest” has only 3. It would be interesting to see the year wise distribution for both the sources.

![Figure 3.4 Count of Collisions by Year (MapQuest)](image)

**Figure 3.4 Count of Collisions by Year (MapQuest)**
We can clearly observe from Figure 3.4 that severity 2 and 3 are present in equal manner for all years but high severity cases are only present for year 2019 and 2020. Since severe cases causes more damage financially and physically, it would be unfair to consider this data source.

For Bing data source, we can find both high and low severity cases for all consecutive years i.e., from 2016 to 2020. We proceed with the rest of the analysis activity using Bing data. As previously said, the primary goal of this study is to predict the severity of collisions. MapQuest data has very less data related to severity level 4. Bing has even distribution of severity levels for all consecutive years.

Since severity 0 is low severity and we have good distribution for severity1, we consider low severity for category 0 and 1 whereas we consider severity 3 and 4 as high severity. Therefore, we are going to merge data for respective categories. As mentioned in Section 3.1, the collision records consist of 49 states of U.S. It is important to check the distribution of state-wise number of collisions (Figure 3.6). California, Oregon and Florida have recorded
maximum number of collisions. However, it is unclear whether these are low severity or high severity ones. We analyze further about the low and high severity cases with respect to different states. Few states have negligible data with only 1 type of severity (i.e. either low or high cases for all records).

Such states should be handled with different strategy as the model predicts based on patterns and calculations within the data. In the dataset ‘Bing’, low severity cases are 768,405, and high severity cases are 266,394. In the dataset ‘MapQuest’, low severity cases are 1,594,325, and high severity cases are 819,976.

With 39.37 million residents, California was the state with the greatest resident population in the United States in 2020. With a population of around 580,000 people, Wyoming was the state with the least population. As a result, we can say that the density of population could be an affecting parameter for causing accidents and should be investigated in future.

![Accidents per state](image)

Figure 3.6 State-Wise Counts of Collisions
3.3 Data Pre-processing

Data pre-processing is the most important phase in the data analytics life cycle. Data processing helps to convert raw data into meaningful information. Data contains garbage values, missing values and parameters that do not have any relation with the end goal (For example, unique identifier of road collision: since each collision has unique ID, it is no longer helpful for prediction)

3.3.1 Missing Values

Missing data is a tough problem and there is no single perfect solution to handle it. Everything depends on the context and the end goal. If the business needs are very critical then we need to handle the data accordingly. There are various types of missing data: N/A, nan, garbage value (99999) or an empty string. Domain knowledge is the most important factor in handling missing value. Dropping missing rows seems to be safe at first, but it will reduce the data size. If the dataset is small, it could potentially damage the model performance. As a first step, we check the percentage split of missing values in our dataset. The distribution is shown in Table 3.1.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Missing Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMC</td>
<td>100</td>
</tr>
<tr>
<td>Number</td>
<td>74.14</td>
</tr>
<tr>
<td>Precipitation</td>
<td>46.51</td>
</tr>
<tr>
<td>Wind-chill</td>
<td>41.42</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>10.67</td>
</tr>
<tr>
<td>Weather Condition</td>
<td>2.62</td>
</tr>
<tr>
<td>Visibility</td>
<td>2.59</td>
</tr>
<tr>
<td>Humidity</td>
<td>2.51</td>
</tr>
<tr>
<td>Temperature</td>
<td>2.38</td>
</tr>
<tr>
<td>Wind Direction</td>
<td>2.26</td>
</tr>
<tr>
<td>Pressure</td>
<td>2</td>
</tr>
<tr>
<td>Weather Timestamp</td>
<td>1.5</td>
</tr>
<tr>
<td>Airport Code</td>
<td>0.25</td>
</tr>
<tr>
<td>Time zone</td>
<td>0.17</td>
</tr>
<tr>
<td>Zip code</td>
<td>0.073</td>
</tr>
<tr>
<td>City</td>
<td>0.005</td>
</tr>
<tr>
<td>Sunrise Sunset</td>
<td>0.005</td>
</tr>
<tr>
<td>Civil Twilight</td>
<td>0.005</td>
</tr>
<tr>
<td>Nautical Twilight</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Once we obtain the missing values percentage, we need to decide how to deal with missing values. The two most common strategies to handle missing values are 1) Removal of missing values 2) Imputation. The first method is the simplest way to get rid of missing values. Before deleting complete row that contains missing values, we need to make sure that we are not losing much data. For example, if the dataset consists of 100,000 rows and we have missing values in 100 rows, we can remove it without causing major impact on the dataset.

Imputation methods fills the missing place with some number. There are various methods for value imputation. The most common way of imputing a value is the mean. We can impute by mean, median or mode depending on context. In our dataset, we remove all the missing values since we have good amount of data and the missing percentage is more than 40% of the total data. Here we remove TMC, number, wind chill, precipitation and wind speed that has high percentage of missing values. After removing the missing values, we are left with approximately 980,000 rows of data which includes 730,000 rows as low severity and 250,000 records as high severity.

3.3.2 Time of the Day

One of the studies [39] in the literature survey indicated that times and days were included in the important parameters list. These parameters have a very common relation with collisions both fatal and non-fatal. There could be several factors which contributes to high number of collisions. Since most of the drivers are found drunk on weekend, there are chances of
collisions under the influence of alcohol as per the sources, 28% of the fatalities. Another reason could be more relaxed driving due to weekends. Saturday is the most dangerous day of the week for drivers according to the NHTSA (National Highway Traffic Safety Administration). Tuesday is considered to be the safest of all. In the graph shown in Figure 3.7, the day-wise counts of collisions are plotted using our dataset.

Figure 3.7 Count of collisions by weekday

Figure 3.7 clearly indicates that there are a greater number of severe cases during weekends when compared to weekdays. This justifies the reasons for collisions mentioned above. When compared to weekends without holidays, the rates of serious and non-serious collisions were slightly lower during holidays. There were fewer collisions per day on holidays than on weekends, but more collisions per day on weekdays. Holiday traffic was slightly higher than on weekends or weekdays.

High Severity collisions were found more during the hours of darkness which is shown in Figure 3.8. The percentage of collisions involving injuries increased, but fatal collisions have increased the most. As compared to daylight, the number of fatal injuries at night was almost
four times higher. The severest collisions occurred at night-time followed by those during
dawn and dusk and the least severe were during the day-time.

![Figure 3.8 Comparison of collisions during day and night](image)

**3.3.3 Weather features**

The first and foremost thing to consider before analyzing weather data is the condition of the
road on which the incident took place. Poor road conditions cause more collisions than bad
weather. When these two criteria are combined, the results can be deadly. If the crash
occurred on a road that was particularly windy or steep, this may be an important factor in
your bad weather collision case. In addition to that, if the road’s pavement quality was poor
or the signage was not clear due to fog the result could be severe. Potholes, missing
gaurdrails, or unmarked construction zones are significant dangers for Connecticut drivers,
and are only intensified by bad weather. According to FHWA (Federal Highway
Administration), approximately 22% out of 6 million collisions are due to bad weather
conditions. A weather-related collision can be caused due to any adverse condition such as
rain, fog, sleet, winds or slick pavement. Higher chances of collisions occur especially when the roads are wet following rainy weather. No question how cautiously you drive, another driver may still slip and slide into you and cause a serious injury.

There are few categorical parameters which have multiple values with less data and having similar meaning. For example: ‘Light Rain’, ‘Light Rain Shower’, ‘light Rain Showers’ and ‘Light Snow’, ‘Light Snow Shower’, 'Light Snow Showers’ can be converted to rain and snow, respectively. By doing this, we reduce the complexity of machine learning model. Such features are handles in this section. Out of many weather parameters, we now have only 7 main categories that are clear, cloud, rain, heavy rain, snow, heavy snow, fog. Table 3.2 shows the conversion of weather parameters.

One of the studies indicates that weather-related road collisions kill more people compared to weather disasters like Tsunami, tornado etc. Weather-related collisions occur on wet pavement and during rainstorms according to the Road Weather Management Program. Some factors that contribute to weather-related collisions include fog and the winter season.
Table 3.2. Weather parameter conversion

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Similar Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear</td>
<td>Clear</td>
</tr>
<tr>
<td>Cloudy</td>
<td>Cloudy, Funnel Cloud, Scattered Clouds</td>
</tr>
<tr>
<td>Rain</td>
<td>Light Rain, Light Rain Shower, Rain, Rain Showers</td>
</tr>
<tr>
<td>Heavy Rain</td>
<td>Heavy Rain, Rain Shower, Heavy T-Storm, Heavy Thunderstorms</td>
</tr>
<tr>
<td>Snow</td>
<td>Drifting Snow, Light Snow, Light Snow Showers, Snow, Snow Grains, Light Sleet</td>
</tr>
<tr>
<td>Heavy Snow</td>
<td>Heavy Snow, Heavy Sleet, Heavy Ice Pellets, Snow Showers, Squalls</td>
</tr>
<tr>
<td>Fog</td>
<td>Fog, Light Fog, Patches of Fog, Partial Fog, Shallow Fog</td>
</tr>
</tbody>
</table>

Figure 3.9 describes various weather features and the counts of collisions with low and high severity. We can see that there are more chances of high severity collisions when the weather is cloudy, rainy and snowy (a bit). Heavy fog can make roads nearly invisible. It makes it difficult for other drivers to see a few yards in front of their vehicle. Using fog lights becomes essential when driving in these conditions. An inexperienced driver may be driving without fog lights, which makes it more likely that an collision could happen. Having fog lights can help other drivers know about the vehicles in the front.
Types of Bad Weather:

a) Rain

It includes drizzles, mild rain and downpours. Rain, in all its three forms can be dangerous. Freezing rain may cause ice to form on roads.

b) Black Ice

When snow or rain melts on pavement and freezes, it creates a layer of ice on the roads known as "black ice". It is almost impossible to find and it can form if the road isn't correctly handled for water drainage. It's especially dangerous near bridges and overpasses, where temperatures drop rapidly on elevated surfaces.
c) High winds

Strong winds will also blow debris onto traffic lanes. Due to this it makes vehicles hard to control. It can also cause rollovers. The conditions can be disastrous if it couples with rains, snows, or sleets. Some weather conditions include sleet, wind chill and below-freezing temperatures.

3.3.4 POI Features

A POI (Point of Interest) is a specific location that is useful or interesting. It is generally used for hotels, fuel stations or several other categories used for general navigation. It is a quick, easy, and accurate way of mapping important places, maps or other interesting locations. Nowadays, some applications may handle other detailed data like description, photos of the place, contacts & even working hours too. POI data is used to enable digital navigation, improved routing items, and database validation. Furthermore, several websites specialize in the processing, maintenance, and authentication of POIs that users can view on their mobile devices.

Advantage of POI Mapping:

- Travelling to an unfamiliar route
- Navigation system
- Statistical data generation
- Data that is both spatial and non-spatial is combined.

POI information supports users in decision making and planning activities that are designed for specific requirements. One can track and assess the rapid expansion in navigation systems over the last decade. Companies have started to create their own navigation maps and
distribute them to their customers. In addition to that, Remote sensing and high-definition cameras are two big innovations in this area that are supporting both enterprises and map users. Any map-making company tries to improve its map data by using street maps, addresses, points of interest, and base maps, but there are two key considerations to consider i.e., as a source and destination, the address point and point of interest play an important role in any navigation service from one location to another. As a result, the real quest begins with POI.

Here in our dataset, we have several POI features which includes amenities, bump, crossing, give way, junction, no exit, railway, roundabout, station, stop signal, traffic calming and traffic signal. High and low severity collisions have been analyzed when the specific POI feature is available or not. Figure 3.10 shows the number of collisions in POI features.

![Count of Accidents in POI Features](image)

Figure 3.10. Count of collisions in POI features
It can be noticed in Figure 3.10 that there are some high severity cases in junction, traffic signal and crossings. However, there are few severe cases near station. Junction has maximum number of severe cases. Other POI features did not show much collision counts and we can call them imbalanced. Road collisions near crossing and traffic signal have less chances of major injuries. When they are at an intersection, they are more likely to have a major collision. As an instance, people typically slow down in front of traffic signals and crossings, but the severity of the intersection and the severity of the crossing are strongly related to speed.

3.3.5 Description of Collisions

Word clouds are the collection of words in a graphical way. It describes the information in cluster of words based on the repetition of a specific word. The more it repeats, the bigger and bolder size will be. It helps business users to compare different pieces of text and finding similarities between the two. In our dataset, we have a column that describes the collision. These descriptions are then turned into word clouds and checked for the most emphasized word. Clusters for both severe and non-severe collisions are plotted separately. Figure 3.11 is plotted for severe collisions. It can be observed that the word “closed” is used very frequently. This denotes that the road was temporarily closed for service since the collision was very severe. We can use this information to identify the areas where there are more severe collisions and the conditions at that instance.
Non-severe collisions show “blocked”, “lane” and “avenue” to be the most repeated words (Figure 3.12). This denotes that the road is only blocked and could be cleared based on the collision impact. It usually occurs on lane and avenue. These conditions can be checked, and precautionary measures can be taken to avoid those collisions.
Chapter 4

Classification Methods

4.1 Overview of Methodology

The data analytics methodology of this research is shown in Figure 4.1 as a flow diagram. There are six stages of methodology that are shown in Figure 4.1. The selection of the data source, data pre-processing and data visualization using graphs and charts have been included in Chapter 3. In this chapter we explain the techniques for data planning and building. In Chapter 5, the analysis based on the model results will be explained.

![Figure 4.1. Methodology](image)

a) Data Discovery

The first step of any analysis is to understand the problem, find the data sources, and collect them. Since the problem statement for this research is to predict the severity of collisions,
data repositories that log collision data were chosen. The dataset is vast and contains information on various scenarios that will be explained later. This step also includes the tools and language needed to obtain the end goal. Python is used for conducting the implementation.

b) Data Pre-Processing

After the selection of data sources, the second step is to prepare the data for further use. This step consists of everything that we need to do with the data. Data cleaning falls under this step. Various techniques are used to retrieve information from the data. This step is crucial for deciding the parameters to consider and continuing for the rest of the analysis process.

c) Data Visualization

After the data is cleaned (pre-processed), it is ready for some visual inference. The data visualization step is used to reduce the machine learning model's load since it gives some strong data patterns through graphs. Bar chart, histograms, pie chart, correlation etc., are some of the examples of visualization. By the end of the visualization phase, we can conclude some inference regarding our problem statement.

d) Data Balancing

When giving input data from a class distribution with a number of unbalanced classes, the most important challenge is that the model is likely to be biased towards most samples (Low Severity in Bing dataset). After performing data pre-processing the size of data is for low severity are 572023 and for high severity are 196535. To handle this problem, SMOTE is used to balance the dataset.
e) Model Planning and Model Building

After setting up business goals and pre-processing data, the next step is to decide which algorithms will accomplish the end goal. Here our goal is to classify the collision data based on its severity. Therefore, we are going to choose classification algorithms that fall under supervised learning. Model Building involves the actual building of a model. The algorithms/techniques that were decided in the previous step are going to be implemented here. The data is divided into trained and unseen data (test). Model building is performed on the training data, on the other hand, the model’s performance is evaluated in the test data for stability and efficiency.

f) Analyze the results of the Models

The last step is to move the model to a live environment and monitor whether the results match the end goal. If it matches, the model with its key findings is finalized. Else the analysis is moved backward to any phase of the lifecycle to perform the required changes.

4.2 Data Balancing

In our dataset (Bing), the target variable is a binary response which shows the severity of collisions. Here, the collisions are divided into two groups. Low-severity injuries fall into the first category, while high-severity collisions fall into the second. Machine Learning applied for this type of data, where the class labels are observable, is called supervised learning. Classification is used to create a predictive model which classifies the observations in two groups and it predicts the probability of a low or high severity injury. The most significant problem where the input data has an extremely unbalanced class distribution is that the model is likely to be biased towards the majority samples (Low Severity in Bing dataset). After
applying pre-processing, the dataset, the size of data for low severity cases are 572,023, and high severity cases are 196,535. To balance the data, we have multiple approaches that can be divided into Random Oversampling, Random Under-sampling and SMOTE (Synthetic minority over-sampling Technique). The majority class is reduced in random under-sampling and the majority and minority classes are balanced. Random oversampling works in similar manner for minority classes. It oversamples the minority class and brings the records equal or near to records of majority class. SMOTE has hybrid method of both under and over sampling techniques. It performs up-sampling on minority class and down-sampling on majority class. The synthetic model is created by choosing one amongst the k-nearest neighbors randomly and connecting it to the selected example to make data points within the feature space. The only difference between SMOTE and Random Oversampling is that random oversampling will pick some random records and increase the size of dataset by duplicating the values, while SMOTE will try to create new values by calculating the parameters belonging to same class.

In this study, we have used SMOTE for oversampling on the minority class to balance the category distribution. After applying SMOTE the size of data for low severity cases are 572,023 and high severity cases are 486,219. The oversampling ratio was given as 0.85 to increase the minority class count.

During the classification analysis, the existing dataset is divided into training-testing groups. Various algorithms are applied on the training dataset. In this study, five classification methods are used for learning and building the classification models, namely Random Forest (RF), Logistic Regression (LR), Naïve Bayes, and Neural Network, k-fold Cross Validation. These methods are explained briefly in the following sections.
4.3 Logistic Regression (LR)

Since we have binary target variable, linear regression will not be useful to meet the end goal. Logistic regression provides relaxation for few assumptions which are made by linear regression [48]. Logistic Regression itself is not a classifier, however, it provides probability of 0 and 1, so it is used in this analysis to predict the severity of an collision. A logistic model, for example, predicts a person's probability of attending a play as a function of weather or temperature. It might be predicted by the model that a change of 10 degrees may increase or decrease the chances of playing by two times. The expression "twice as likely" applies to the chances of doubling. Here the odds are doubling from 2:1 to 4:1, and so on. Such a logistic model is known as a log-odds model. As a result, LR is also known as the logistic model or logit model in statistics. By fitting data to a Logistic Function curve, this technique is used to predict the outcome of the occurrence of a particular event.

\[ P(y = 1) = \text{sigm}(a_0 + a_1 \cdot x) \]  

(1)

where \(a_0\) and \(a_1\) are the parameters of the LR model which are learned through training. Sigm is the sigmoid function that predicts values between 0 and 1 and is defined as

\[ \text{sigm}(z) = \frac{1}{1 + e^{-z}} \]  

(2)

When the probability of a record is greater than 0.5, the record will be classified as a high severity collision. Thus, with a threshold of 0.5, the output can be represented as follows.

\[ y = 1 \text{ if } P(y=1) \geq 0.5 \]
\[ y = 0 \text{ if } P(y=1) < 0.5 \]  

(3)

Figure 4.2 shows the sigmoid function graph for the above relation.
To test the validity of the LR, a generalized linear model is tested. So, the removal of less important parameters will not have any effect on the accuracy of the model. The equation of a regression line can be written as:

\[ y = a_0 + a_1 \times x \]  \hspace{1cm} (4)

where \( a_1 \) is the weight for an input variable \( x \) and \( a_0 \) is the bias expression. Logistic regression uses sigmoid function to predict the probability of each class's outcome.

### 4.4 Naïve Bayes

The Naïve Bayes classifier is a probabilistic classification method that determines the posterior probability of each class and groups them accordingly using Bayes theorem [50]. For this classifier, the main assumption is the “assumption of independence” which makes the probability of likelihood function of each feature separately and multiplication of the separated likelihood functions will be used for the calculation of the conditional probability. Naïve Bayes classifier also assumes the parameters used for training are equally important [51]. The Bayes Theorem is used for formulation of the Naïve Bayes classification method and is given below:
\[
P(y = High \mid features) = \frac{P(y=High) \cdot P(features \mid y=High)}{P(features)} \quad (5)
\]

The Bayes theorem calculates the posterior probability of the target variable. Considering the feature is equal to the prior probability, it is multiplied by the likelihood of the features in the collision severity group and then divided by the total probability of the features. This formula can be used for other feature groups,

\[
P(y = Low \mid features) = \frac{P(y=Low) \cdot P(features \mid y=Low)}{P(features)} \quad (6)
\]

In both formulas, the denominator is the same. The probability can be calculated based only on the numerator by omitting the denominator. The decision to assign features to a class will be based on the maximum probability derived. Hence, it can be written as:

\[
P(y = High \mid features) \propto P(y = High) \cdot P(features \mid y = High) \quad (7)
\]

\[
P(y = Low \mid features) \propto P(y = Low) \cdot P(features \mid y = Low) \quad (8)
\]

Whichever class has the higher posterior probability, the observation will be assigned accordingly to that severity level.

When considering the prior probability, the observation can be entered based on prior knowledge. If there exists no prior knowledge about the category, the prior probability can be calculated for the collision group by dividing the number of the collision observations in the training data by the total observations in the training sample, \(P(y = High)\). For other group, the severity can be calculated in the same way by finding the proportion of the observation to the total number of observations, \(P(y = Low)\).

As mentioned earlier, the likelihood function in Naïve Bayes assumes independence of features. According to probability rules, when two sets are independent, the intersection probability is equal to the multiplication of the probability of each set. This can be
expressed as follows:

\[ P(feature_1 \cap feature_2) = P(feature_1) \times P(feature_2) \]  \hspace{1cm} (9)

The likelihood function can also be written as:

\[ P(\text{features} \mid y = \text{High}) = P(feature_1 \mid y = \text{High}) \times \ldots \times P(feature_n \mid y = \text{High}) \]

Or,

\[ P(\text{features} \mid y = \text{High}) = \prod_{i=1}^{n} P(feature_i \mid y = \text{High}) \]  \hspace{1cm} (10)

The likelihood functions can be written separately for each feature. The assumption can be used to calculate the density for each feature of specific observations. If the features have normal distribution, the likelihood function can be written as:

\[ P(feature_{n_i} \mid y = \text{High}) = \frac{1}{\sqrt{2\pi \sigma_{High}^2}} \exp \left( - \frac{(feature_{n_i} - \mu_{High})^2}{2\sigma_{High}^2} \right) \]  \hspace{1cm} (11)

The \( n \)th feature and index \( i \) in the formula shown above refers to the \( i \)th observation of the sample. The mean and variance of the feature \( n_i \) are \( \mu_{High} \) and \( \sigma_{High}^2 \) respectively.

The posterior probability for each feature is calculated and the testing sample is assigned to the group with the maximum probability.

\[ y = \arg\max_y P(y) \prod_{i=1}^{n} P(feature_i \mid y) \]  \hspace{1cm} (12)

The above formula signifies that the probability for each class will be calculated for a specific observation and the observation will be assigned to a group according to its maximum posterior probability.

**4.5 Random Forest (RF)**

Random Forest is a supervised learning algorithm. This method was developed by Breiman L. (2001) [52], where the algorithm of the random decision forest (Ho T K, 1995) [53] is
extended. It uses ensemble learning strategy since it is done by bagging or aggregating technique. It fits the classification model using decision trees. In most cases, the accuracy given by decision tree model is lower than other classification algorithms because of high variance and low bias and hence are very flexible models. This will lead to the model being over-fitted. However, the prediction accuracy of the test data is low. To avoid this scenario, random forest algorithm is used. Random Forest is implemented by fitting multiple decision trees at a time and results are selected on the basis of selecting a subset from the training dataset where they are averaged.

This algorithm uses Bagging method for training the model. The Bagging method is done by selecting random samples from the training dataset. Each randomly selected subset includes \( n \) observations from the total \( N \) observations in the training samples. The model will then fit \( k \) decision tree models on each of these \( k \) samples. Predictions for the new observation is calculated by averaging the results or by finding the most frequent category in each of these \( k \) samples. The variance of this model is reduced when compared to individual prediction results from decision tree model and has a much better performance. The main difference between Bagging and random forest algorithm is that Bagging requires only bootstrap aggregation to be implemented while the random forest selects features randomly from each selected subset (feature Bagging). Hence in random forest both the observation and feature Bagging are used in each subsample. This is done in order to reduce the correlation between the subsets, due to the existence of a strong feature in the dataset (Ho TK, 2002) [54]. It leads to a decrease in both correlation between models and also variance of the model with no increase in the bias. Selecting \( k \) number of Bagging is done by fitting the out-of-bag error or performing cross validation on the dataset. Therefore, the number for \( k \) is optimal when the classification testing data error in the cross validation is minimal. The out-of-bag error
from the random forest model for observation \( x_i \), will be found only in the subsamples that do not include observation \( x_i \).

For each \( k \), a decision tree model will be fitted to the data. This is done by calculating the gini index, information gain, and chi square. The formula for calculating the information gain which is used in computing the decision tree is as follows:

\[
\text{Information Gain}(A, S) = H(S) - \sum_{t \in T} p(t)H(t)
\]  

(19)

In the above formula \( S \) is the target variable (severity indication), \( H \) is the entropy function, where \( t \) is the number of subsets created by splitting the feature, and \( p \) is the fraction of the number of elements of \( t \) to that of \( S \). The union of these elements in each bin will be sum of elements in \( S \). \( \sum t \in T p(t)H(t) \) the average entropy information for the feature \( A \). The entropy formula is as below:

\[
H(S) = \sum_{t \in T} -p(c)p(c)
\]

(20)

Lowercase \( c \) are the classes in the response variable and \( S \) is the response variable. \( p(c) \) is the fraction of the number of elements in class \( c \) to that in class \( S \). After calculating the information gain for each feature, the feature with the maximum information gain will be the parent node and the rest will be the sub nodes according to their information gain sorted in descending order. The Gini index can be used as the metric when deriving the decision tree. The formula for Gini index for each feature is as follows:

\[
GL(bin) = 1 - \sum_{c \in C} p(c)^2
\]

(21)

The Gini index for each feature is calculated after splitting the feature into several bins. In each bin, the proportion of negative to total cases in that bin will be the \( p(\text{negative}) \) and proportion of positive to total cases will be \( p(\text{positive}) \). \( C \) is the set of negative and positive
cases. Next, the Gini index for the feature will be multiplied by the bin’s Gini index and the sum of the weight of each bin is the total Gini index.

\[
Gini \text{ Index}(S, A) = \sum_{b \in B} W(b) GI(b)
\]  

(22)

For the feature A, the weight of the bin (b) is multiplied by the Gini index of bin b. The summation of all bins will be the Gini index for that feature where \( W(b) = \) number of elements in b / number of elements in S. A lower value for the Gini index is better because the decision tree will be split according to the sorted Gini index in ascending order. The feature with the lowest Gini index will be a parent node and all other will be as sub nodes according to their sorted gini index respectively. Pruning increases the performance of the decision trees. It can be done by removing the less important leaves contributing the prediction model and setting the majority class or the most frequent class as the target for that leaf. The size of pruning can be found by doing cross validation and should not reduce the classification accuracy in the testing data set.

4.6 Neural Network

Neural networks are theoretically based on the neuron and bodily activities. The mathematical and computational model for the neural network was developed by (McColloch A. & Pitts W. 1943) [55]. This technique uses a multi-layered network of neurons in order to make a classification model that connects the input features to the neurons in each layer. The final layer is the output layer, which includes the classification result or prediction of the category for that model.

To model the network, choose the number of hidden layers and the number of neurons for each hidden layer. Using one hidden layer is enough in most of the cases. Two hidden layers
can be used when the model must represent the result of an arbitrary decision boundary that gives an arbitrary accuracy. The suggested hidden layer sizes are to be considered between the number of input features and the output classes. The size of the hidden layer can be considered by including two-thirds the number of input features plus the number of output classes. It should also be less than twice of the number of input features. The size can be entered as a range or a hyper parameter when solving the neural network model.

Pruning the hidden layer size is done after finding the values for the weights of the connections. The neurons with a very low weight can be removed from the network and the accuracy of the updated model should be checked. The final model will be the one with the highest classification accuracy with the testing data.

The idea behind neural connections is based on the least square method. It uses a transforming function that maps the output to a value between the class labels similar to a logistic regression. The sigmoid activation function is one of the activation functions used to transform the linear model of each neuron. The equation of an individual connection to a neuron is as below:

\[
Sigmoid(B1 \ast F + B0) = predicted \ probability
\]  

(23)

In the above formula, F is a feature and B1 is the weight of the connection between a neuron and the input F. B0 is the bias of the neuron. The sigmoid function is:

\[
Sigmoid(Z) = \frac{1}{1 + e^{-Z}}
\]  

(24)

The complete equation for a neuron, including all the connection between the input and that neuron, is given below:

\[
Z1 = B_1 \ast F_1 + B_2 \ast F_2 + \cdots + B_p \ast F_p + Bias \ neuron1
\]  

(25)

\[
Activation \ of \ neuron1 = sigmoid(Z1)
\]
This is the formula for all connections for input features from 1 to p. Each neuron has its own bias term. The predicted probability for the neuron is found by using the sigmoid activation function.

After writing all the equations for each neuron across all hidden layers, the output layer, weight and bias for each neuron will be found by completing the minimization of the mean squared error. The cost function for the mean squared error is:

$$C(w, b) = \frac{1}{2n} \Sigma \mathbf{x} \in \mathbf{X} || y - a(x) ||^2$$ (26)

In the above equation, y represents the neuron's output, a represents the activation at x and is derived through w, b and x, while n represents the total number of inputs. The cost function is minimized by using the Gradient Descent algorithm.

The Gradient descent algorithm is based the idea that a multi variable function, which is differentiable at a point w, will decreases by moving from point w in the direction of the negative gradient. Therefore, the gradient descent approach will reach a local minimum after enough iterations. It uses any initial value for the weights and the bias. In the next iteration, it will find the updated weights and bias by using the gradient descent iteration on the cost function. Hence, the updated weights in the gradient descent will be found as:

$$w_{n+1} = w_n - \eta \nabla C(w_n)$$ (27)

The variable vector at iteration n+1 is equal with the variable vector at iteration n by subtracting a small value multiplied by the gradient of the cost function at iteration n.

C is the cost function and \(\nabla C\) is the gradient of the cost function. For k variables, from \(w_1\) to \(w_k\), the gradient C will be a vector of k parameters:

$$\nabla C = \left( \frac{\partial C}{\partial w_1}, \frac{\partial C}{\partial w_2}, ..., \frac{\partial C}{\partial w_k} \right)^T$$ (28)

\(\eta\) is a small value and can be found in each iteration. If the absolute difference between the weights at each iteration is a small value E, then the parameter \(\eta\) can be considered as \(E/||\nabla C||\).
Finding optimal solution of the network is done by using forward or back propagation. In the forward propagation, the output from the input layer is fed into the input of the hidden layer and the output of the hidden layer is input into the output layer and finally the output layer gives the output classes.

However, in back propagation, the solution starts from the output layer and returns step by step until it reaches to the input layer.

### 4.7 k-fold Cross Validation

The k-fold cross validation method is used to do the classification analysis. As a result, the selected features are split k-fold at random in each loop. The split for the 10-fold cross validation is done in the following process (Figure 4.3):

1. 60% of the data is used to train the classification algorithm, and 40% is used to validate the model. This would be accomplished by arbitrarily splitting the data k times to create a 10-fold of 60:40. In each fold, 60% of the data will be used for training and 40% for testing.

2. 70% of the data is used for training and 30% is used to validate the classification model, the 10-fold has a 70:30 proportion randomly split.

3. Training data is 80%, while testing data is 20%. The 10-fold cross validation will be implemented by an 80:20 proportion split randomly for each fold.

4. Finally, 90% of the data were used for training purposes and 10% for testing purposes. A 90:10 proportion split is being used in the 10-fold cross validation.
Figure 4.3. k-fold cross validation [56]
Chapter 5

Results and Discussion

5.1 Evaluation Metrics

In machine learning models, training data is used to train the algorithm, and test data is used to test the model's predictive performance. In general, we focus at how well the model performs when evaluated on unseen data. The question is, how do we evaluate and determine the model's performance? To do so, we use evaluation metrics to assess the model's results. This section analyses the different types of calculation metrics.

5.1.1 Confusion Matrix

This is the most often used evaluation metric since it is easy to manage and can be used to calculate a range of essential metrics such as accuracy, precision, recall and so on. The confusion matrix is an N x N matrix that defines a model's overall results, where N is the number of classification categories. When it applies to binary classification as shown in Figure 5.1, we have a 2x2 confusion matrix.

True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) statistical measures compose a confusion matrix. These numbers are derived from both actual and predicted data.

**True Positive (TP):** A situation in which the actual value and the predicted value are both positive.

**False Positive (FP):** In this case, the actual value was negative, but the predicted value is positive.

**True Negative (TN):** A case in which the actual value was negative, and the predicted value is negative as well.
False Negative (FN): A situation where the actual value was positive, but the predicted value is negative.

5.1.2 Recall (True Positive rate or sensitivity)

Sensitivity is another term for recall. It's the proportion of TP that correlates to actual positive cases. In simple terms, recall refers to how many true positives were discovered out of all the positive cases. It can be written as:

\[
\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]

For instance, think about a scenario where we have 7 black boxes and 3 white boxes. Suppose a machine learning model predicts 5 black boxes and out of 5 predicted boxes, just 3 were actually black (TP), with the others being white boxes (FP). Recall is 3/5 in this case.

Table 5.1: Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Black</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>White</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

5.1.3 Specificity (True Negative rate)

True negative rate or specificity is the number of cases labeled as negative that were actually negative. It is given by the following relation.

\[
\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}
\]

5.1.4 Precision

Precision is the ratio of TP (true positive) to TP (true positives) and FP (false positives). Simply put, precision refers to how many of the instances labelled as positive by the classifier is truly positive. It's possible to write it as
Precision = TP / (TP + FP)

The precision value for identifying black boxes in the example above is 3/5.

5.1.5 Accuracy

The most basic performance metric is accuracy, which is essentially the number of correctly expected observations to all observations. One might believe that if our model is accurate, it is the best.

Accuracy = TP+TN/TP+FP+FN+TN

5.1.6 F1 Score

The harmonic mean of precision and recall is the F1 Score, also known as F score or F-measure. The value varies from 0 to 1, with 0 being the worst and 1 being the best. It's possible to write it as:

F1 = 2 * (Precision * Recall) / (Precision + Recall)

5.1.7 AUC

When it comes to predictive analysis, the AUC, or Area Under Receiver Operating Characteristic (ROC) curve, is one of the most commonly used evaluation metrics. When used at various probability thresholds, this approach shows us how well a model does. For the classification problem, a probability threshold of 0.5 is set as the default value.

The ROC diagram is a curve of the TPR (True Positive Rate), also known as sensitivity, and the FPR (False Positive Rate). The region under the curve (AUC) can be calculated using the ROC curve, and it is believed that a formula would score a randomly chosen positive data point higher than a randomly chosen negative data point. A ROC curve is represented in Figure 5.2. The ROC of the sample is shown in blue, while the ROC of a random model with a region under the curve of 0.5 is shown in red.
5.2 Data Balancing

Before applying sampling, low severity cases are 572,023 and high severity cases are 196,535. After balancing the data with SMOTE low severity cases are 572,023 and high severity cases are 486,219. We applied our classifier on this balanced data obtained by SMOTE. Figure 5.3 shows the unbalanced and the balanced data after applying SMOTE.
5.3 Classification Results

Below are the results for few classification techniques applied on our dataset. The ROC-AUC curve is shown along with confusion matrix for a given algorithm. A total of 4 algorithms are evaluated on the dataset, Logistic Regression, Random Forest, Naïve Bayes, and Neural Networks. Moreover, these models are applied SMOTE data resampling technique as the dataset is highly imbalanced. The ROC-AUC curve is then used to measure the performance of all the models.

5.3.1 Logistic Regression

The Logistic Regression performed “not satisfactory” than it was expected. The performance of this model is not good while dealing with both high and low severity classes. Precision and recall are respectively 0.69 and 0.71. In addition to this, the ROC-AUC is not good.

<table>
<thead>
<tr>
<th>Training: testing</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1-Score</th>
<th>AUC Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>60-40</td>
<td>0.67</td>
<td>0.69</td>
<td>0.7</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>70-30</td>
<td>0.67</td>
<td>0.69</td>
<td>0.7</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>80-20</td>
<td>0.69</td>
<td>0.71</td>
<td>0.72</td>
<td>0.7</td>
<td>0.57</td>
</tr>
<tr>
<td>90-10</td>
<td>0.69</td>
<td>0.71</td>
<td>0.72</td>
<td>0.7</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Confusion matrix and classification report for training: testing ratios 90:10, 80:20, 70:30 and 60:40 is shown in Figure 5.4. From the figure we can say that the performance of Logistic Regression model remained constant throughout different sample sizes. It is very important to select the model with higher overall performance since higher severity cases are most critical.
The AUC for Logistic Regression is 0.57 which is considered “not good”. However, we can explore other classification algorithms and compare all the algorithms.
5.3.2 Naïve Bayes

The performance metrics of the Naive Bayes classifier on the dataset for predicting the severity of collisions as seen in Table 5.3. The results show that Naïve Bayes performs worse than Logistic Regression.

Table 5.3 Naïve Bayes Performance Metrics

<table>
<thead>
<tr>
<th>Training: Testing</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1-Score</th>
<th>AUC Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>60-40</td>
<td>0.77</td>
<td>0.38</td>
<td>0.66</td>
<td>0.51</td>
<td>0.78</td>
</tr>
<tr>
<td>70-30</td>
<td>0.77</td>
<td>0.39</td>
<td>0.66</td>
<td>0.51</td>
<td>0.77</td>
</tr>
<tr>
<td>80-20</td>
<td>0.76</td>
<td>0.4</td>
<td>0.67</td>
<td>0.52</td>
<td>0.77</td>
</tr>
<tr>
<td>90-10</td>
<td>0.76</td>
<td>0.4</td>
<td>0.67</td>
<td>0.52</td>
<td>0.77</td>
</tr>
</tbody>
</table>

The confusion matrix and performance chart for Naïve Bayes classifier is shown in Figure 5.5. From the figure it can be seen that the model performs “fair” for cases with severity “0” (low severity) but when it comes to cases with “1” (high severity), the model fails to predict as per the expectations.

The average performance of all the selected samples lies between 0.51 to 0.78. Since high severity cases are our priority, we are not going to consider this model as our best performing model. The F1 score, which shows balance between precision and recall is also very low.
The ROC curve for Naïve Bayes classifier shows the AUC value of 0.77 which is high when compared to logistic regression model. Logistic regression model’s accuracy is much better than Naïve Bayes but the performance for predicting high severity cases is low in both the classifiers.
5.3.3 Random Forest

The performance of Random Forest in predicting severity level is seen in Table 5.4. The results show that Random Forest gives the best results compared to Logistic Regression, and Naïve Bayes classifiers.

<table>
<thead>
<tr>
<th>Training: Testing</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1-Score</th>
<th>AUC Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>60-40</td>
<td>0.79</td>
<td>0.8</td>
<td>0.81</td>
<td>0.79</td>
<td>0.87</td>
</tr>
<tr>
<td>70-30</td>
<td>0.79</td>
<td>0.81</td>
<td>0.81</td>
<td>0.8</td>
<td>0.87</td>
</tr>
<tr>
<td>80-20</td>
<td>0.79</td>
<td>0.81</td>
<td>0.82</td>
<td>0.8</td>
<td>0.88</td>
</tr>
<tr>
<td>90-10</td>
<td>0.8</td>
<td>0.81</td>
<td>0.82</td>
<td>0.8</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Of all the algorithms, Random Forest performed the best. With a 94% training accuracy and an 81.8% test accuracy, Random Forest has a precision of 0.80 and a recall of 0.81. The confusion matrix of this classifier in all the sample sizes is shown in Figure 5.6.

This technique has very good precision and recall for all the sample sizes and strong F1 score which shows balance in the performance of the model. We can select this algorithm as the best performing classifier when compared to all the algorithms discussed in this thesis.
The AUC value for Random Forest is 0.89 which is considered as “very good” for a model performance. As training accuracy of 94% and test accuracy of 81.8%, the model is overfitting so 10-fold cross validation is applied to the model. The results are shown in the Table 5.5.

Table 5.5 10-fold cross validation Random forest

<table>
<thead>
<tr>
<th>K-fold</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>f1-score</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.82</td>
<td>0.80</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
</tr>
</tbody>
</table>
5.3.4 Neural Networks

Table 5.6 shows the performance of Neural Networks for predicting severity level of collisions. The results show that Neural Networks performs well as compared to most of the other classifiers except Random Forest.

<table>
<thead>
<tr>
<th>Training: Testing</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1-Score</th>
<th>AUC Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>60-40</td>
<td>0.7</td>
<td>0.86</td>
<td>0.77</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>70-30</td>
<td>0.74</td>
<td>0.79</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>80-20</td>
<td>0.75</td>
<td>0.75</td>
<td>0.77</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>90-10</td>
<td>0.72</td>
<td>0.81</td>
<td>0.77</td>
<td>0.77</td>
<td>0.77</td>
</tr>
</tbody>
</table>

The confusion matrix of Neural Networks is shown in Figure 5.7. This algorithm was tested on a sampled dataset as neural network requires high processing power and extra resources (like G.P.U) to accelerate the computation. There are 79 input dimensions for which the neural network has to process. The activation function is selected as ‘Relu’ and output activation function is selected as ‘Sigmoid’ since we have binary output for our model. Hidden layer with 8 nodes is selected in the model. The loss function is “binary entropy” and “Adam” (adaptive model) optimizer is selected to optimize our neural network model. The number of epochs selected are 50 and batch size is taken as 10 to compute the model effectively.
The AUC value of Neural Network is 0.77 which is less than Random Forest (0.88). Thus, we can say that Random Forest performed best among all the algorithms. A combined ROC Graph for all the algorithms is shown in Figure 5.8.
Random forest worked well in terms of precision and recall, as can be shown. The AUC curve of the algorithms is slightly higher than all other algorithms.

### 5.4 Results Comparison

Table 5.7 shows a comparison of all the models along with accuracy, precision, recall, f1-score and AUC values. It can be observed that Random Forest presents the best accuracy, precision, recall, F-score and AUC values. Figure 5.9 shows a graphical visualization of the performance of all the classifiers. It can be observed that Random Forest performs the best in all the metrics. Neural Networks has the second-best performance in all the performance metrics which is followed by Logistic Regression. Naïve Bayes has low recall, F1-score and accuracy. Figure 5.10 shows the comparison of performance metrics for each classifier. It can be observed that Logistic Regression has the lowest AUC value whereas Random Forest has the highest values in all the metrics.
Table 5.7 Summary Table

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>F1-Score</th>
<th>AUC Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.67</td>
<td>0.69</td>
<td>0.7</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.79</td>
<td>0.81</td>
<td>0.81</td>
<td>0.8</td>
<td>0.87</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.77</td>
<td>0.39</td>
<td>0.66</td>
<td>0.51</td>
<td>0.77</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.74</td>
<td>0.79</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Fig. 5.9 Classifier result comparison
Fig. 5.10 Performance Evaluation
Chapter 6
Conclusions and Future Work

6.1 Conclusions

We applied machine learning methods to determine whether the collision was of high or low severity in this study. After that we have referred some reference papers and figure out what we can do to make this model robust. We used a publicly available dataset for this, which contains the record of collisions from 2016 to 2020. It contains more than 25 million records and only 900,000 were selected for this research. The data was selected based on its quality and the fact that it came from a database with a large number of serious collisions that were reported in a consistent manner. We also perform some preprocessing steps to clean the dataset. The dataset was imbalanced since there are less high severity cases as compared to low severity records. The main challenge when using input data from an extremely unbalanced class distribution is that the model tends to support the largest samples. Consequently, it misrepresents a high severity collision as a low severity. To address this problem, we used a SMOTE technique to balance the dataset. As a bagging tool, we used the Random Forest model, which is considered robust in such scenarios. In comparison to these models, we evaluated the performance of Naive Bayes, Neural Networks, and logistic regression models in predicting severity levels. Then, we analyzed all the models and showed the results when tested on different sample sizes. The Random Forest outperformed the other models, according to the comparative results.

In this research, the four levels of severity were converted into two severity levels of high and low by labelling Severity 1 and Severity 2 cases as low severity and Severity 3 and Severity 4 as High Severity cases. Different regions and states are then categorized into high severity
and low severity cases. In one [65], Moosavi et al. worked on the “US Accidents” dataset which contains data till June 2020 and applied Logistic Regression classifier by states and achieved F1-score of 0.58 for Austin and 0.56 for Charlotte states, respectively. Deep Neural Network (DNN) gave F1-score for Austin as 0.64 and for Charlotte as 0.63. In this thesis, the same dataset of “US Accidents” was used to predict the traffic collisions severity. Our results gave the F1-score of 0.68 using Logistic Regression. In our study we have applied several other classifiers which gave higher accuracy, AUC, precision, recall and F1-score as compared to [65]. Random Forest achieved an accuracy of 81%, F1-score of 0.8 and AUC of 0.87. Naïve Bayes achieved an accuracy 66%, F1-score of 0.51 and AUC of 0.77. Neural Network achieved an accuracy of 78%, F1-score as 0.76 and AUC of 0.77. The results show an improvement of 26.5% from the best results of [65].

6.2 Future Work

It can be clearly seen that the classification algorithms could be well improved if the records in the dataset would be higher. More parameters may be added to the dataset for further analysis. When we increase the number of data points, we therefore increase the number of fatal or high-severity collisions. The major challenges with increasing the number of parameters and data points are that machine learning methods like Neural Networks are computationally expensive. Larger datasets require more computation time and resources. Exploring other sources of data and datasets of other countries, such as Spain could also help improving the classification process where high severity collision rate is much higher. In this thesis some assumptions have been made.
References


road incidents detection. In Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on (pp. 930-935). IEEE.


Redondo-Calderón, José Luis, Juan de Dios Luna-del-Castillo, José Juan Jiménez-Moleón, Miguel García-Martín, Pablo Lardelli-Claret, and Ramón Gálvez-Vargas.
"Application of the Induced Exposure Method to Compare Risks of Traffic Crashes among Different Types of Drivers under Different Environmental Conditions."


[45] https://www.mapquest.com/

[46] https://www.bing.com/

[47] https://www.stitchdata.com/


Workshop on Empirical Methods in AI


[57] https://towardsdatascience.com/


