REVA UNIVERSITY
Bengaluru, India

A Project Report on

Accident Severity Detection and Prediction Using Machine Learning Techniques

Submitted in Partial Fulfilment for Award of Degree of

Master of Business Administration
In Business Analytics

Submitted By

Sharon Joseph

R19DM053

Under the Guidance of

Mithun D J

Senior Manager -Data Science, RACE

REVA Academy for Corporate Excellence - RACE

REVA University

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064

race.reva.edu.in

August, 2022

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, **Sharon Joseph** hereby declare that I have completed the project work towards a Master of Business Administration in Business Analytics at, REVA University on the topic entitled **Accident Severity Detection and Prediction Using Machine Learning Techniques** under the supervision of **Mithun D J, Senior Manager -Data Science, RACE.** This report embodies the original work done by me in partial fulfilment of the requirements for the award of the degree for the academic year **2022.**

Place: Bengaluru
Date: 20th August 2022

Name of the Student: Sharon Joseph
Signature of Student

REVA UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the Project work entitled **Accident Severity Detection and Prediction Using Machine Learning Techniques** carried out by **Sharon Joseph** with **R19DM053,** is a bonafide student at REVA University, is submitting the project report in fulfilment for the award of Master of Business Administration (MBA) in Business Analytics during the academic year 2021. The Project report has been tested for plagiarism and passed the plagiarism test with a less than 15% similarity score. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said Degree.

Signature of the Guide	Signature of the Director
Name of the Guide **Mr Mithun D J**	Name of the Director **Dr Shinu Abhi**

External Viva

Names of the Examiners

1. Harsh Vardhan, Chief Digital Technology Architect, Capgemini
2. Pradeepta Mishra, Director – AI, L&T InfoTech
3. Rajib Bhattacharya, Director – Data & Analytics, Cargill

Place: Bengaluru

Date: 20th August 2022

REVA UNIVERSITY
Bengaluru, India

Acknowledgement

I would like to acknowledge the support provided by the founder and Hon’ble Chancellor, **Dr P Shayma Raju**, Vice-Chancellor, **Dr M. Dhanamjaya**, and Registrar, **Dr N Ramesh** for supporting the RACE program, and for providing the necessary facilities and infrastructure required for the best learning experience. I am immensely proud of being part of this program and REVA University.

I would like to thank **Dr Shinu Abhi**, Director, RACE (REVA Academics for Corporate Excellence) for her guidance and constant supervision in providing necessary information regarding the course, subjects, and a plethora of other topics, thus ensuring that the learning experience was immensely powerful.

My sincere gratitude to all my mentors who have taught and guided me in this journey of achieving my degree. Their vast knowledge and practical experience paved way for a great learning experience.

A special thanks to my project guide **Mr Mithun D J**, Senior Manager -Data Science RACE for his valuable support, encouragement, and guidance provided in understanding the concept for executing the project.

I am also thankful to my classmates for their constructive criticism and friendly advice during the course period.

Last but not the least, I would like to thank the Almighty, my family, and my colleagues who have extended their invaluable cooperation and encouragement throughout this project lifecycle.

Place: Bengaluru
Date: 20th August 2022

Name of the Student: Sharon Joseph
Signature of Student

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled "**Accident Severity Detection and Prediction Using Machine Learning Techniques**" was scanned for similarity detection. The process and outcome are given below.

Software Used: Turnitin

Date of Report Generation: 15th August 2022

Similarity Index in %: 16%

Total word count: 6288

Name of the Guide: Mr Mithun D J

Place: Bengaluru

Date: 20th August 2022

Name of the Student: Sharon Joseph

Signature of Student

Verified by: M N Dincy Dechamma

Signature

Dr Shinu Abhi,

Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	WHO	World Health Organization
2	RTA	Road Traffic Accidents
3	CNN	Convolutional Neural Networks
4	CRISP-DM	Cross-Industry Standard Process for Data Mining
5	EDA	Exploratory Data Analysis
6	UT	Union Territory
7	SVM	Support Vector Machine
8	KNN	K- Nearest Neighbour
9	NH	National Highway
10	WEKA	Waikato Environment for Knowledge Analysis

List of Tables

No.	Name	Page No.
Table 7.1	Variable Explanation	23
Table 7.2	Majority of Accident Occurrence analysis	27
Table 10.1.a	Confusion Matrix-Gaussian Naïve Bayes	31
Table 10.1.b	Confusion Matrix-Random Forest Tree	31
Table 10.2	Model Performance Metrics	32
Table 10.3	Classification Report Metrics	32
Table 10.4.a	Classification Report Metrics- Gaussian Naïve Bayes	33
Table 10.4.b	Classification Report Metrics – Random Forest Tree	33
Table 10.5	Accuracy Score	34

List of Figures

No.	Name	Page No.
Figure 5.1	CRISP-DM Methodology	19
Figure 5.2	Project Pipeline	20
Figure 7.1	Monthly accidents per State/UT	24
Figure 7.2	Timely accidents per State/UT	25
Figure 7.3	Season accidents	25
Figure 7.4	States with top accident occurrence	25
Figure 7.5	Accidents during different Seasons	26
Figure 7.6	Accidents occurrences at different Times of the Day	26
Figure 7.7	Snapshot of Accident Data	27
Figure 10.1	Confusion Matrix	31
Figure 11.1	Web application to predict the severity of accidents	35
Figure 11.2	Accident Severity Prediction Result	36

Abstract

Every year, traffic collisions in India are a major cause of death, injury, and property damage and approximately 450,000 accidents occur in India, with 150,000 people killed. India has the highest number of road fatalities, with 53 accidents occurring every hour and one death occurring every four minutes. Individuals, their families, and the nation suffer significant economic losses as a result of road traffic injuries.

Regardless of what has been done to enhance traffic safety in India, some such areas are always vulnerable. Because of the variety of these vulnerability-inducing causes, effective analysis is required to significantly reduce the alarming figures.

Data mining methods can assist in identifying the causative factors and assisting transportation authorities in enhancing safety standards. This study aims to propose a framework that analyses road accidents in India by mining frequent patterns and significant considerations causing the various types of accidents using the most suitable machine learning and classification techniques.

The Gaussian Naïve Bayes and Random Forest algorithms have been used to build the model. It analyses various factors like date and time, weather conditions, road conditions, vehicle speed, and road classifications to forecast the severity of traffic accidents.

Keywords: Road Traffic Accidents, Accident Severity, Data Mining, Machine Learning, Random Forest, Gaussian Classifier, Logistic Regression.

Contents

Candidate’s Declaration 2
Certificate 3
Acknowledgement 4
Similarity Index Report 5
List of Abbreviations 6
List of Tables 6
List of Figures 7
Abstract 8
Chapter 1: Introduction 10
Chapter 2: Literature Review 12
Chapter 3: Problem Statement 17
Chapter 4: Objectives of the Study 18
Chapter 5: Project Methodology 19
Chapter 6: Business Understanding 21
Chapter 7: Data Understanding 23
Chapter 8: Data Preparation 28
Chapter 9: Modeling 29
Chapter 10: Model Evaluation 31
Chapter 11: Deployment 35
Chapter 12: Analysis and Results 37
Chapter 13: Conclusions and Recommendations for future work 38
Bibliography 39
Appendix 42
Plagiarism Report 42
GitHub Link 47

Chapter 1: Introduction

Due to the ever-rising population and globalization, more people travel every day for various reasons. Worldwide, the most widely used modes of passenger transport are the Automobile followed by Bus and Rail. Roadways are most widely used and preferred across the globe thus increasing traffic movements and air pollution. RTAs are very common and each day, many accidents take place across the country.

Accidents on the road occur for a broad range of factors. Drivers are frequently distracted behind the wheel, diverting their focus from the road. In plenty of other cases, drivers might become tired after driving for several hours, resulting in avoidable errors. Accidents can happen because of a wide range of causes like poor visibility, hazardous road design or the lack of care of other drivers. Accidents can have a variety of causes, but the implications are often the same, likely to result in everything from vehicle and damage to property to serious injuries.

The ability to predict the intensity of a road accident is a crucial component in accident management. It provides critical information for emergency personnel to determine the extent of accidents and formulate effective accident systems and procedures. Nonetheless, two issues remain in the assessment of the severity of traffic accidents. The first is that traditional transportation methods cannot verify the magnitude of road accidents, and the second is that current road accident severity forecasting methods perform relatively poorly, both of which will result in untimely or unsatisfactory rescue facilities, resulting in casualties. As a result, accurate prediction of the severity of traffic accidents is critical in order to safeguard the lives of victims. Forecasting the intensity of road accidents is a significant problem in the field of road traffic safety (Zheng et al., 2019a).

Knowing the importance of road safety, authorities are trying to recognize the causes of road accidents to decrease the number of accidents. The exponential rise in the information about accidents is making it challenging to evaluate the barriers causing them. Data mining is the process of analysing data from different sources to obtain relevant information that can be utilized to make well-informed business decisions. It is not industry specific and is used in almost all fields to investigate the possibility of hidden information.

Data mining techniques such as classification, clustering, time series analysis, association rule mining, and regression are all important. Classification is an important data mining method that analyses data and categorises it into predetermined classes or groups. Linear regression, logistic regression, decision trees, SVM, Nave Bayes, KNN, Random Forest, and gradient boosting are just a few of the machine learning algorithms available.

Rising accident rates necessarily require a significant increase in road safety. Drivers are not always held responsible. Other factors that could have made a significant contribution to such an accident include automotive malfunctions, climate variability, and transportation conditions. According to a WHO report, dangerous commuters are motorcyclists, cyclists, and pedestrians who cause 50% of all accidents globally, with two-wheeler users responsible for approximately 31% of all fatalities. This continues by stating that adults are responsible for 59% of all fatalities nationally and internationally.

The proposed work includes a study of India's states and union territories for the contributing causes to make impactful decisions to facilitate traffic safety.

Chapter 2: Literature Review

For this study, research papers have been extensively reviewed on various topics related to Road traffic accidents and the different methodologies used to analyse accident data. Some of these research papers have been discussed herein.

The authors used data mining techniques in this study to link documented road characteristic features to traffic accidents in “Ethiopia”, and they developed a set of guidelines that the "Ethiopian Traffic Agency" could use to enhance safety on the road. They gathered and cleaned traffic accident data, tried to develop novel attributes, and tested several predictive models in this study. Road traffic accidents (RTAs) are simpler to carry on the study to discover subjects of interest that should be given traffic safety resources. The study's overarching goal is to examine the role of road-related aspects in accident severity using RTA data from “Ethiopia” and to develop predictive models (Beshah & Hill, 2010).

The purpose of this research is to investigate injury severity classification patterns in traffic accidents. The authors conduct experiments with road accident training data acquired from the Critical Analysis Reporting Environment System’s Fatality Analysis Reporting System. In this paper, they examined traffic accident training datasets using classification algorithms, feature ranking algorithms, and classifiers based on the Arc-X4 Meta classifier to identify road accident patterns based on injury severity. Among the methodologies, the Random Tree classifier with Arc-X4 Meta has an accurate result of 99.73% and a misclassification rate of 0.27%. Recall and precision values are used to assess accuracy (S. Shanthi & R. Geetha Ramani, 2012).

Artificial Neural Networks and Decision Trees are the data analysis techniques that were used in this work to discover new knowledge from past data about accidents on one of "Nigeria's busiest roads" to reduce death and destruction on highways. The performance of various Decision Tree algorithms and Artificial Neural Networks on a traffic accident dataset was contrasted in this paper (V.A & A.A, 2014).

This paper aims to infer some patterns from real-time accident data obtained from the "National Highway between Kharagpur and Kolkata (NH6) in India" using the spatial decision tree. It contains information such as the time of the accident, weather patterns, the side of the highway, transportation systems, the character of the accident, location, and type of accident. Accident data have a strong spatial context, which spatial decision trees use to bring in new findings. The objective is to recognize patterns over time and, as a result, help identify accident zones. The use of results obtained from these decision trees can improve accident management systems. Accident-causing constraints and patterns could be identified in certain places. This information can be utilized to help avoid accidents and put appropriate remedial measures in place. Spatial decision trees can also identify disaster centres and prioritise them to decrease the frequency and severity of potential accidents (Manasa et al., 2015).

In this paper, using the WEKA tool, the authors analysed that the “Info Gain Attribute Evaluator” function was used to assess the variables influencing accidents and determine which factor is more accident-prone. They also used the Apriori algorithm to perform association classification and determine the most appropriate rule for the accident dataset (Gothane & Sarode, 2016).

The purpose of this paper is to employ data mining to assist in the development of a prototype that not only provides a broad classification of information by combining related components with each other in order to identify the accident-prone regions of the country based on different accident-causing variables but also clearly signifies the relationship between these variables and casualties (Jain et al., 2016).

This paper discusses some data mining techniques that have been beneficial in resolving road accident severity problems and concludes which one may be the efficient strategy in a traffic accident scenario. The data mining technique is acknowledged as a dependable technique for assessing traffic accident severity issues and determining the factors that cause them. Clustering is a technique for grouping objects that are similar in some way. The k-means algorithm has high efficiency for clustering large data sets, but it is limited in its ability to form groups for real-world data while continuing to work only on numerical data because it helps to reduce the cost function by changing the meaning of the clusters (Gupta et al., 2017).

The existing methods for predicting the severity of traffic accidents primarily employ shallow severity forecasting and statistical models. An innovative traffic accident severity prediction-convolutional neural network model is proposed to improve prediction accuracy by considering combination relationships among traffic accident features. Based on the weights of traffic accident features, the "feature matrix to grey image" (FM2GI) algorithm is proposed to convert a single feature relationship of traffic accident data into grey images containing combination relationships in parallel as the model's input variables. Furthermore, experiments showed that the proposed model for predicting the severity of traffic accidents has a good performance (Zheng et al., 2019b).

This research effort establishes models for selecting the main determinants and developing a system for categorising the adversity of the injuries. “Several machine learning techniques are used to create these models. On traffic accident data, supervised machine learning algorithms such as AdaBoost, Logistic Regression (LR), Naive Bayes (NB), and Random Forests (RF) are used. To deal with data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is used. The results of this study suggest that the Random Forest model could be a useful tool for predicting the severity of traffic injuries" (Emhamed AlMamlook et al., 2019).

This study was based in Bangladesh to examine traffic accidents in depth to determine the seriousness of accidents using machine learning approaches. It also identifies the important variables that have an impact on accidents and makes beneficial recommendations on the matter. To classify the magnitude of accidents, Decision Tree, K-Nearest Neighbours (KNN), Nave Bayes, and AdaBoost were used (Md. Farhan Labib et al., 2019).

Thaduri proposes that a "Convolutional Neural Network(CNN) model for deep learning-based traffic accident prediction. It uses traffic accidents to create a state matrix that describes the traffic state and CNN model by influencing circumstances such as light, weather, and traffic flow. This paper examines the accuracy of the proposed system using examples. When compared to the traditional Backpropagation (BP) algorithm, the evidence-based findings indicate that the proposed model is much more effective than the prevailing neural network design in predicting traffic accidents" (Thaduri et al., 2021).

This study presents "RFCNN, a group of machine learning and deep learning models incorporating Random Forest and Convolutional Neural Networks for forecasting the severity of road accidents. The proposed approach's effectiveness is compared to that of several foundation learner classifiers. The analysis relied on accident records from February 2016 to June 2020 in the United States. The obtained results show that the RFCNN improved decision-making and outperformed other models" (Manzoor et al., 2021).

As per this study, the data points were supplied by the UK government and were utilized as chronological data to train the models from January 2015 to December 2018. The remainder examples from 2019 were used to evaluate the model's performance. Weather patterns, sun location, speed limit, and duration of the day were among the parameters observed. Given the circumstances and road conditions, a multiclass classification model was employed to predict the amplitude of the accident. Different datasets were combined to cover various traffic situations and circumstances and to predict the severity of accidents. Prior to training, the dataset data were normalised, and the training data and validation dataset were evaluated (Iveta et al., 2021).

The existing system included analysis of every single pattern of the data which was time-consuming. According to the literature Survey, some of the works include neural network methodology and data mining for the imbalanced data without handling it which is not the right way to complete the analysis. Hence, the motivation for this work includes handling the data and making it suitable for analysis and using the concept of machine learning on the same.

Chapter 3: Problem Statement

Accident-related death and injuries are recognised as a worldwide concern, and road safety has been a great problem since the beginning of the automobile era. The roads of India have continued to contribute to road accident fatalities and the rate of growing accidents in India has been rising since then. Therefore, it is necessary to take preventive measures to avoid Road accidents, thereby enabling rapid rescue operations to save the lives of people in our community.

The conventional method is manual, in which the government sector uses ledger data and manually analyse the information; based on the analysis, they take preventative measures to decrease the number of accidents.

Accident Severity Detection and Prediction using appropriate Machine Learning Algorithms and to explore the key drivers leading to accidents, thereby enabling stakeholders to reduce the occurrences of fatal accidents and take precautionary measures in accident-prone areas.

Chapter 4: Objectives of the Study

The main goal of this work is to utilize data mining techniques to help in creating a system that analyses the various factors that contribute to a Road traffic accident to detect a probable scenario where an accident could occur and anticipate the severity of a traffic collision depending on the intensity of factors like road type, road condition, date and time, weather condition, speed, etc.

Two major objectives of this study are,

1. Identify the key features which contribute to traffic accidents.

2. Predict the class of accident severity into Fatal, Serious and Minor.

Chapter 5: Project Methodology

This project uses the CRISP-DM methodology to implement and carry out the different phases of the data mining and machine learning process involved in this project.

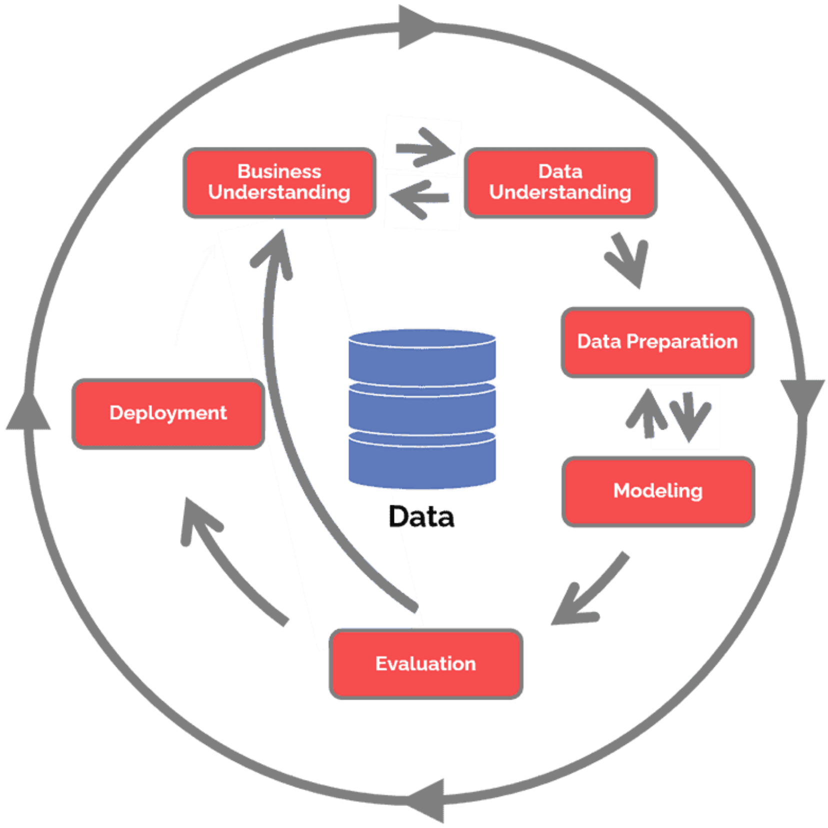

Figure 5. 1: CRISP-DM Methodology (Nick Hotz, 2022)

The "Cross-Industry Standard Process for Data Mining" (CRISP-DM) as seen in Figure 5.1, is a process model with 6 phases that naturally explains the life cycle of data science and performs like rules to assist you in implementing, planning and organizing your machine learning project.

The unique characteristics of CRISP-DM have directed to its status as a "de-facto" benchmark of data mining research techniques and a standard against which other methods are measured. One of the primary advantages of using this method in Data Science is that it is a cross-industry benchmark that can be used in any Data Science project, regardless of domain.

In our project, the six phases of the CRISP-DM methodology have been carried out in the following manner and as depicted in Figure 5.2 below.

- Business Understanding: Understand the problems faced by common people, while travelling on Indian Roads and the causes that lead to injuries and fatalities caused by RTAs.

- Data Collection: Various Data needs to be collected from various sources. Weather Data, Accidents Data, Road and Traffic data.

- Data Processing: Establish data connection, analyse, and clean the obtained data and perform exploratory data analysis. (EDA)

- Data Modelling: Apply data mining techniques to detect the accident contributing factors and apply Machine Learning algorithms to predict the accident severity based on various factors.

- Evaluation: Evaluate results and compare which algorithms give accurate results.

- Deployment: Web Application to show accident prediction and the severity of accidents.

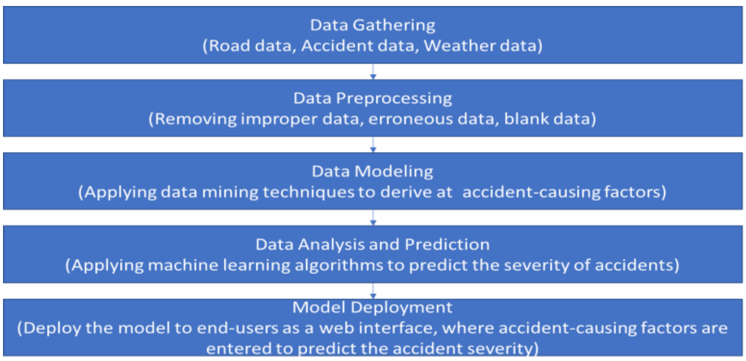

Figure 5. 2: Project Pipeline

Chapter 6: Business Understanding

Road traffic accidents are the eighth worldwide leading cause of death. Road accidents are caused by a variety of factors, which can be widely categorized as road and environmental aspects, human elements, and vehicular factors.

Some of the major concerns that ordinary people have when travelling on Indian roads today are as follows:

- <u>Accidents based on Road and Environment Factors.</u>

Road conditions, road junctions and type of traffic control, ongoing construction works, speed breakers, weather conditions, poor lighting, lack of adequate road signs, sidewalks and neighbourhood environment.

- <u>Accidents based on Human Factors.</u>

Disobeying of traffic rules, invalid driving license, not using safety devices like helmets and seat belts, triple riding, distracted driving, negligent parking, not crossing roads at the pedestrian crossings, road rage, overcrowding of passenger vehicles and improper use of headlights.

- <u>Accidents based on Vehicular Factors.</u>

Accidents caused by over-age vehicles, speed, vehicular faults, overloading.

- <u>Untimely rescue operations</u>

Delay in Ambulance services reaching the accident spot, hesitation of the public to get involved and help the victims, police and government formalities and public interference.

- <u>Law and Order</u>

Poor enforcement of traffic rules and regulations.

This project is mainly focused on saving people’s lives, who travel on Indian roads. It is a community problem than a business problem. The identification of different factors that play a critical role in road accidents, as well as the prediction of road accident severity, are essential components for safer travel on Indian roads.

For the right predictions of the intensity of the accidents, questions like what the time and date of the accident, how was the weather condition, were there any roadway hazards, what was the road Classification and road type and what was the vehicle's speed at the time of accident occurrence needs to be answered and studied.

Road transport authorities and the Police department will become the key beneficiaries of this study as it reduces their manual efforts and helps in better management of Road infrastructures and their maintenance.

Implementing such a project can save lives, time, and money. As a result, it is a useful study for the society we live in.

Chapter 7: Data Understanding

In CRISP-DM, the Data Understanding phase entails understanding the "gross" or "surface" properties of the received data and analysing the findings. The information for this work is gathered from various sources, including the “Ministry of Road Transport and Highways”.

Variables	Description
Carriageway/Roadway hazards	This variable talks about any obstacles on the roadway. It could be animals or pedestrians, potholes, etc. Here we are generally classifying them as hazards present and none.
Light	It indicates the lighting during the accident as Light present and Darkness.
Day of the Week	Days through Sunday to Saturday.
Special Conditions at the accident site	This includes malfunctioning traffic lights, improper signboards, road maintenance issues, road works and fuel-related issues.
Road Class	A-Principal Road, B-Secondary Road, C-Minor Road, Motorway and Unclassified Roads.
Junction Control	This field specifies how the junction is controlled.
Junction Details	Whether the accident occurred at a junction, crossroads, roundabout, private driveway, and so on.
Road Surface State	It is the state of the road during the accident like dry, wet, frost, snow and flood.
Road Type	This variable talks about the kind of roadway in which the accident occurred. Like: Single Road, Double Road, One-way, Slip Road, and U-turn.
Area Type	Whether Urban or Rural areas.
Weather	This indicates the climatic condition during the accident. Variables are Rain, Wind, Snow, Fog, Mist, Sunny, and Fine weather.
Speed	This variable talks about the speed of the vehicle at the time of the accident. In this dataset, it varies from 10-70 km/hr.
Time	This variable captures the time of the day (24 hours) and whether day or night.
Severity of Accident	The severity of the accident is the target variable, and represents three classes, called fatal, serious and minor.

Table 7.1: Variable Explanation

Table 7.1 explains the various features available in the dataset and what each of them means. It is important to understand the meaning and categories under each feature for further analysis.

Raw traffic accident datasets associated with road type and weather conditions had about 1048575 rows of data. The most challenging part of this research is collecting sample datasets from various organizations or departments. Since data gathering was time-consuming, crucial, and important, data samples were collected through different websites. Therefore, data was forced to be recorded in an excel spreadsheet to easily analyse and predict accident severity.

7.1 Exploratory Data Analysis (EDA)

The accident occurrences per state or union territory of India have been studied using various elements like weather, date and time and seasons. Figures 7.1, 7.2 and 7.3 indicates the count of accidents that have occurred during each month, time and season respectively.

	STATE/UT	YEAR	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER	TOTAL
0	A & N Islands	2001	8	23	15	15	14	19	14	19	7	12	13	22	181
1	A & N Islands	2002	12	10	14	16	10	7	16	11	23	21	11	17	168
2	A & N Islands	2003	19	13	15	13	13	12	8	16	17	25	14	15	180
3	A & N Islands	2004	21	14	22	17	13	18	16	19	16	20	15	24	215
4	A & N Islands	2005	19	21	22	17	13	19	21	14	15	19	10	16	206
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
485	West Bengal	2010	1245	1150	1349	1246	1172	1284	1231	1190	1128	1227	1251	1252	14725
486	West Bengal	2011	1350	1179	1314	1148	1220	1241	1185	1074	1112	1214	1161	1270	14468
487	West Bengal	2012	1346	1383	1357	1270	1352	1434	1349	1204	1112	1251	1179	1371	15608
488	West Bengal	2013	1564	1382	1474	1392	1629	1391	1315	1208	1228	1299	1335	1332	16549
489	West Bengal	2014	1516	1398	1473	1385	1527	1439	1416	1356	1391	1395	1415	1394	17105

Figure 7. 1: Monthly accidents per state/UT

	STATE/UT	YEAR	0-3	3-6	6-9	9-12	12-15	15-18	18-21	21-24	Total
0	A & N Islands	2001	2	6	29	40	39	40	18	7	181
1	A & N Islands	2002	2	6	22	41	33	33	23	8	168
2	A & N Islands	2003	2	8	31	35	28	36	25	15	180
3	A & N Islands	2004	2	5	29	42	43	43	37	14	215
4	A & N Islands	2005	0	8	27	28	38	42	50	13	206
...	...	...	...	...	...	...	...	...	...	...	...
485	West Bengal	2010	1241	1397	1721	2508	2272	2296	1831	1459	14725
486	West Bengal	2011	1200	1493	1687	2553	2182	2196	1812	1345	14468
487	West Bengal	2012	1346	1511	1837	2831	2328	2268	1966	1521	15608
488	West Bengal	2013	1442	1911	2136	2900	2246	2366	2137	1411	16549
489	West Bengal	2014	1455	1634	2022	2998	2570	2458	2132	1836	17105

Figure 7.2: Timely accidents per state/UT

	STATE/UT	YEAR	TOTAL	SUMMER	AUTUMN	WINTER	SPRING
0	A & N Islands	2001	181	52	32	53	44
1	A & N Islands	2002	168	34	55	39	40
2	A & N Islands	2003	180	36	56	47	41
3	A & N Islands	2004	215	53	51	59	52
4	A & N Islands	2005	206	54	44	56	52
..	...	...	...	...	...	...	...
485	West Bengal	2010	14725	3705	3606	3647	3767
486	West Bengal	2011	14468	3500	3487	3799	3682
487	West Bengal	2012	15608	3987	3542	4100	3979
488	West Bengal	2013	16549	3914	3862	4278	4495
489	West Bengal	2014	17105	4211	4201	4308	4385

Figure 7.3: Seasonal Accidents

The bar chart in Figure 7.4, below represents the accidents that have occurred in all the states of India, and it looks like Tamil Nadu has been reporting the largest number of RTAs. This can be easily found on google; on websites like Statista, Dowtoearth and so on.

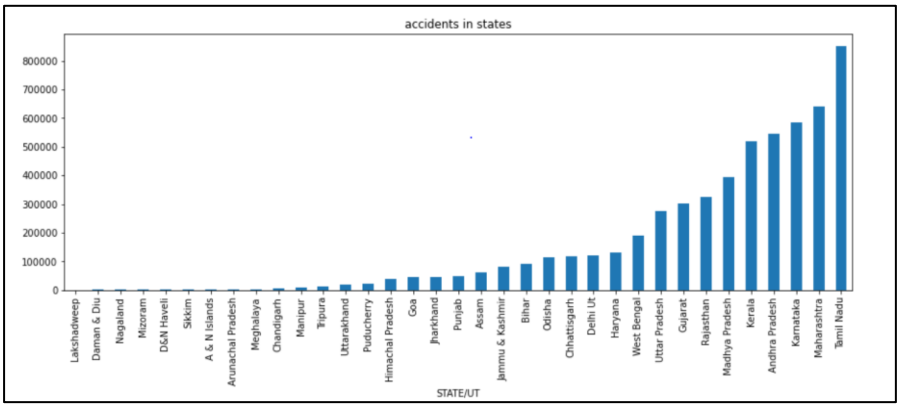

Figure 7.4: States with top accident occurrence

Data is analysed to better acknowledge the impact of weather on India's accident rate and figure 7.5 shows that majority of accidents from our dataset have occurred during the Spring season.

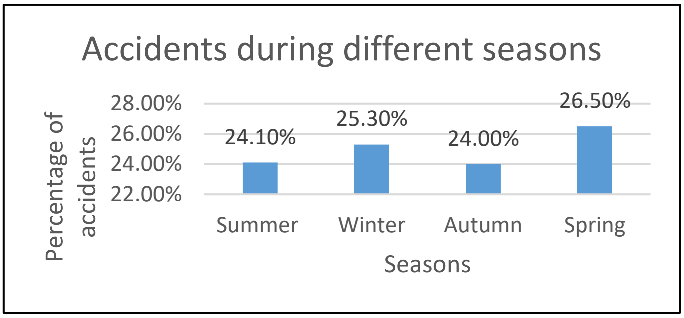

Figure 7.5: Accidents during different seasons

Similarly, data has been studied to comprehend the impact of time of occurrence on the accident rate in India. Figure 7.6 shows that most accidents in our dataset occurred during the afternoon.

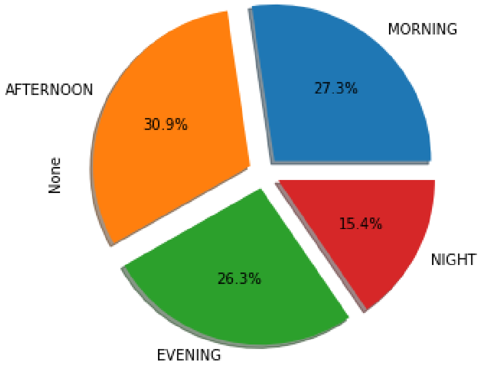

Figure 7.6: Accident occurrence at different times of the day

The state-by-state data has been analysed and investigated for various factors, the data is rightly handled to perform normalisation and utilize the data in the modelling phase. Accident data was gathered from different sources and clubbed together as getting the accident data for Indian roads was not possible.

There is no open-source data available for Indian Road accidents. Therefore, no data is readily available for analysis. Data had to be merged from different sources and the final data was prepared for analysis.

Accident_Index	1st_Road_	Accident_Severity	Carriageway_Hazards	Day_of_Week	Junction_Control	Junction_Detail	Light_Conditions
200501BS00001	A	Serious	None	Tuesday	Data missing or out of range	Not at junction or within 20 metres	Daylight
200501BS00002	B	Slight	None	Wednesday	Auto traffic signal	Crossroads	Darkness - lights lit
200501BS00003	C	Slight	None	Thursday	Data missing or out of range	Not at junction or within 20 metres	Darkness - lights lit
200501BS00004	A	Slight	None	Friday	Data missing or out of range	Not at junction or within 20 metres	Daylight
200501BS00005	Unclassifie	Slight	None	Monday	Data missing or out of range	Not at junction or within 20 metres	Darkness - lighting unknown
200501BS00006	Unclassifie	Slight	None	Tuesday	Data missing or out of range	Not at junction or within 20 metres	Daylight
200501BS00007	C	Slight	None	Thursday	Give way or uncontrolled	T or staggered junction	Darkness - lights lit
200501BS00009	A	Slight	None	Friday	Data missing or out of range	Not at junction or within 20 metres	Daylight
200501BS00010	A	Slight	None	Saturday	Auto traffic signal	Crossroads	Darkness - lights lit
200501BS00011	B	Slight	None	Saturday	Give way or uncontrolled	T or staggered junction	Daylight
200501BS00012	A	Slight	None	Sunday	Auto traffic signal	Crossroads	Darkness - lights lit
200501BS00014	A	Slight	None	Tuesday	Auto traffic signal	Crossroads	Darkness - lights lit

Figure 7.7: Snapshot of Accident data

Variables	**Type**	**Maximum %**
Road Class	A	45%
Carriageway Hazards	None	98%
Day	Friday	16%
Junction Control	Uncontrolled	48%
Junction Detail	Not at Junction	41%
Light Conditions	Daylight	73%
Road Surface	Dry	68%
Road Type	Single carriageway	74%
Special Conditions at the site	None	97%
Speed Limit	30	63%
Area	Urban	64%
Weather Conditions	No High Winds	79%
Accident severity	Minor	85%

Table 7.2: Majority Accident Occurrence analysis

From Table 7.2, the accident data which we have analysed shows the occurrences of accidents majorly when the above variables and types coincide. Most accidents have occurred in Road type A, which is on Main Roads, in Urban area where there is no junctions or junction control, there are no special conditions or hazards on the roads and the weather looks good, speed is controlled, there is enough daylight and accidents occurred on Friday. Therefore, the Accident severity of 85% is Minor from the data that we have. Using this data let us now prepare the data and split the data into train and test and predict the severity of accidents that may occur under different conditions.

Chapter 8: Data Preparation

Data preparation is a step that requires the most time and effort as raw datasets were not in an understandable format for computing machines and gave incomplete information. Using certain data reduced the model's efficiency in predicting accident severity. As a result, incorrect data must be removed in order to get quality data. An intensive data preprocessing technique is used in this study before building a model to derive meaningful and determinant risk factors such as data cleaning, missing value handling, outlier treatment, dealing with absolute value, encoding, and normalisation.

We have gathered a variety of data from different sources. This data has been cleaned and removed for missing values and the feature values have been encoded. This data is now stored in a clean excel file which we would use as an input data file.

Data Cleaning:

- Delete or drop data missing and data not found rows.
- Delete or drop data missing or out-of-range values

Encoding the data:

- Feature variables are encoded to create a simpler dataset for implementing the model.

Feature Engineering:

- Created a new column from the existing value for the variable Junction Control
- Create seasonal value from the date and time the accident occurred.

Split such that train and test data should have prediction classes in equal ratio

Data is split into a 75:25 ratio, where 75% of the data is used to train the model and 25% of the dataset is used to measure the performance of the prediction model. This prepared data set will be the basis of the analysis.

Chapter 9: Modeling

The raw datasets are divided into training and testing datasets. The training data aids in learning the proposed method and the testing data is used to assess the effectiveness of the newly proposed system. The raw dataset is split in the study using a 75:25 ratio where 75% of the dataset is used to train the prediction model, while the remaining 25% is used to evaluate the prediction accuracy.

9.1 Algorithms used:

In our study, we used the following algorithms to predict the severity of the accident.

9.1.1 Naïve Bayes Algorithm

"This algorithm is based on the Bayes hypothesis and uses the probability theorem of the Bayes to determine the class of an unknown data set. The existence of a particular feature in a class is assumed to be independent of the existence of any other features by a Naive Bayes classification algorithm. The Naive Bayes model is simpler to construct and useful while working with large amounts of data." The accident dataset is first trained, and then a model for prediction is created. Predictions can be made for required conditions, as well as the significance of the accident-causal factors (Sunil Ray, 2017).

For a class variable Y, and feature variables (X1, X2, …, Xn),

Bayes' hypothesis gives the following relation:

$$P\,(Y|X1\ldots, Xn) = P(Y)\,P\,(X1..., Xn|Y)\,/\,P\,(X1\ldots, Xn)$$

Using the Naïve Bayes assumption,

$$P\,(Xi|Y, X1..., Xi\text{-}1\ldots, Xn) = P\,(Xi|Y)$$

For all values of i, the relation can be given by,

$$P\,(Y|X1..., Xn) = P(Y)\,P\,(X1..., Xn|Y)\,/\,P\,(X1..., Xn) \propto P(Y)\,\Pi P(Xi|Y)$$

$$Y= \arg\max=P(Y)\,\Pi\,P(Xi|Y).$$

9.1.2 Gaussian Classifier

"The Gaussian Processes Classifier is a machine learning classification algorithm. Gaussian Processes are a generalized form of the Gaussian probability distribution that can be used to underpin advanced non-parametric algorithms for machine learning algorithms like regression and classification.

They are a type of kernel model, like SVMs, and unlike SVMs, the choice and structure of the kernel employed at the heart of the method can be difficult as the method produces highly calibrated class membership probabilities" (Jason Brownee, 2020).

9.1.3 Random Forest Classifier

"Random forest is a type of Supervised Machine Learning Algorithm that is commonly used in classification and regression problems. It constructs decision trees from various samples and uses their clear majority for classification and average for regression. One of the most important characteristics of the Random Forest Algorithm is its ability to handle data sets with both continuous and categorical variables, as in regression and classification. It outperforms other algorithms in classification problems" (Sruthi ER, 2021).

Chapter 10: Model Evaluation

In this section, an overview of the results obtained from different machine learning procedures is provided and explained.

10.1: Confusion Matrix

"A confusion matrix is a table that is used to calculate the performance of a prediction model. It contains 4 values namely, True Positive (TP) which is the correctly predicted event values, False Positive (FP) which is the incorrectly predicted event values, True Negative (TN) which is the correctly predicted no-event values, and False Negative (FN) which is the incorrectly predicted no event values."

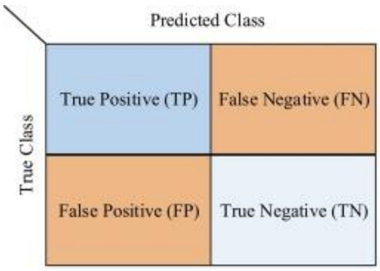

Figure 10.1: Confusion Matrix

Gaussian Naïve Bayes	
Confusion Matrix	
221520	2334
37554	736

Table 10.1.a: Confusion Matrix- Gaussian Naïve Bayes

Random Forest Tree	
Confusion Matrix	
151134	72720
19672	18618

Table 10.1.b: Confusion Matrix- Random Forest Tree

Tables 10.1 a and b denotes the confusion matrix with the TP, TN, FP and FN values. With these values, the model performance can be calculated with the following formulae derived from the confusion matrix.

Formulae to calculate the model performance:

"Accuracy = (TP+TN) / (TP+TN+FP+FN)"

"Precision = TP/ (TP+FP)"

"Recall = TP / (TP+FN)"

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

Model Performance	Gaussian Naïve Bayes	Random Forest Tree
Accuracy	0.85	0.65
Precision	0.86	0.88
Recall	0.99	0.68
F-1 Score	0.92	0.77

Table 10.2: Model Performance Metrics

Table 10.2 shows the model performance metrics calculated, the Naïve Bayes model has a higher accuracy score of 85% and an F1-score of 0.92 which is more closer to 1, proving better performance.

10.2: Classification Report

"A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of the trained classification model" (Aman Kharwal, 2021).

Metrics	Definition
Precision	"Precision is defined as the ratio of true positives to the sum of true and false positives"
Recall	"The recall is defined as the ratio of true positives to the sum of true positives and false negatives."
F1 Score	"The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is"
Support	"Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models; it just diagnoses the performance evaluation process."

Table 10.3: Classification Report Metrics (Aman Kharwal, 2021)

Classification Report Metrics - Gaussian Naïve Bayes				
	precision	recall	f1-score	support
0	0.86	0.99	0.92	223854
1	0.24	0.02	0.04	38290
accuracy			0.85	262144
macro average	0.55	0.50	0.48	262144
weighted average	0.77	0.85	0.79	262144

Table 10.4.a: Classification Report Metrics- Gaussian Naïve Bayes

Classification Report Metrics - Random Forest Tree				
	precision	recall	f1-score	support
0	0.88	0.68	0.77	223854
1	0.20	0.49	0.29	38290
accuracy			0.65	262144
macro average	0.54	0.58	0.53	262144
weighted average	0.79	0.65	0.70	262144

Table 10.4.b: Classification Report Metrics- Random Forest

The classification report from both the algorithms is as in Table 10.4.a and 10.4.b. It gives us various details like Support, Recall, Precision, F1 Score, Support, Macro average and Weighted average.

10.3: Accuracy Score

One metric for evaluating classification models is accuracy. In an informal way, "accuracy is the percentage of correct predictions made by our model". Formally, accuracy is defined as follows:

> "Accuracy = Number of correct predictions/ Total number of predictions"

Accuracy Score	
Gaussian Naïve Bayes	**Random Forest Tree**
84.78%	**64.75%**

Table 10.5: Accuracy Score

Table 10.5 shows the accuracy rate for our models, with the Gaussian Nave Bayes algorithm providing the highest overall accuracy of 84.78%. As a result, we can depend heavily on the model created using the Gaussian Nave Bayes algorithm to classify the severity of the accident as Minor, Serious, or Fatal.

Chapter 11: Deployment

The Python Flask framework is used to deploy the model on a web application. The prediction model is designed as a form-page in a web application, allowing for a more straightforward and simple method of foretelling the road accident intensity. The system can be installed in the end user domain as in the case of traffic police department.

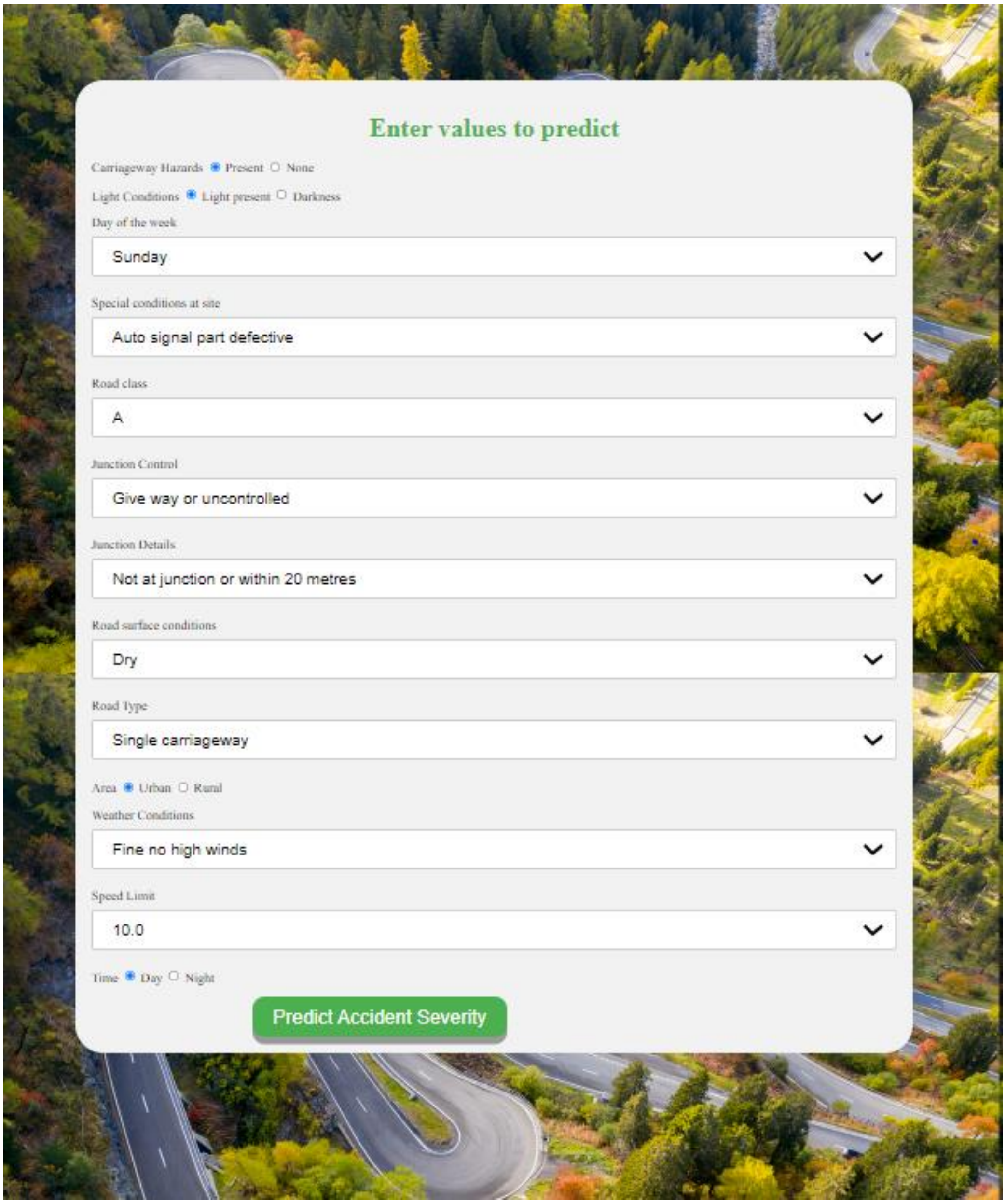

Figure 11.1: Web application to predict the severity of accident.

The end-user enters data for variables such as road and infrastructure conditions, climate and weather, time, date and so on. After the form is filled, when the button called Predict Accident Severity is clicked, there is a window which displays the severity class predicted as minor, serious or fatal.

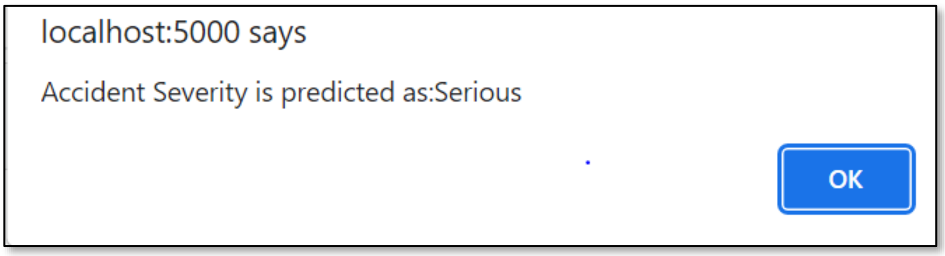

Figure 11.2: Accident Severity Prediction Result

Chapter 12: Analysis and Results

Each model's accuracy can be calculated separately utilizing the values from the confusion matrix, which is obtained separately for both algorithms that we have used in this study. The accuracy score of the Gaussian Naive Bayes algorithm is found to be higher than that of the Decision Tree algorithm indicating that the Gaussian Naive Bayes algorithm is the optimal system for the accident criticality forecasting system, which is used to predict the critical importance of an accident with the given attributes.

Gaussian Naïve Bayes : 85% Accuracy

The gaussian Naïve Bayes classifier gave the best results of 84.78% accuracy for our data and it is better and more optimal when lesser number attributes are used for analysis.

Chapter 13: Conclusions and Recommendations for future work

The goal of this study lies in the discovery of new insights into road accidents. These insights are extremely beneficial when devising strategies for increasing safety on roads, particularly when it comes to choosing the right methods and budgeting resources. With the size and variety of the dataset collected, using data-mining methods to model accident datasets can help in discovering how the time, date, road conditions and wind patterns are related to the various accidental severities caused. The government is thus able to develop better road safety initiatives, indicate roads with signboards, notifying pedestrians and drivers of potential hazards, and design better road infrastructure. Better understanding of the suitability of data mining methods for various applications in the safety research field is yet another expected outcome of this work.

In this project, data analysis can help us identify India's accident-prone states and union territories. The main goal of this study is to evaluate and investigate accident-causal factors and predict the severity of the accidents that have occurred and that may occur in the future. We have achieved this through machine learning algorithms like “Logistic Regression, Gaussian Classifier, and Random Forest Classifier” and tested for performance using the “Confusion matrix” and accuracy using “Classification report”. The system can classify accident severity into Serious, Fatal and Minor accidents based on the several undermining factors identified during the data mining activity.

Recommendations for future study:

In the future, we could perhaps work on developing an accident forecasting model that includes additional features such as driver and vehicle information. We could also investigate the precise location of the accident and recommend that traffic control and highway authorities take precautionary measures in these areas. Proper placement of signboards and warning messages or alarm systems in most accident-prone areas will significantly reduce road traffic accidents.

Bibliography

Aman Kharwal. (2021). *Classification Report*. https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/#:~:text=A%20classification%20report%20is%20a,this%20article%20is%20for%20you.

Beshah, T., & Hill, S. (2010). *Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia*.

Emhamed AlMamlook, R., Morgan Kwayu, K., Reda Alkasisbeh, M., & Ali Frefer, A. (2019). *Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity; Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity*.

Gothane, S., & Sarode, M. v. (2016). Analyzing Factors, Construction of Dataset, Estimating Importance of Factor, and Generation of Association Rules for Indian Road Accident. *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, 15–18. https://doi.org/10.1109/IACC.2016.13

Gupta, M., Solanki, V. K., & Singh, V. K. (2017). Analysis of Datamining Technique for Traffic Accident Severity Problem: A Review. *Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering*, *10*, 197–199. https://doi.org/10.15439/2017r121

Iveta, M., Radovan, A., & Mihaljevic, B. (2021). Prediction of Traffic Accidents Severity Based on Machine Learning and Multiclass Classification Model. *2021 44th International Convention on Information, Communication and Electronic Technology, MIPRO 2021 - Proceedings*, 1701–1705. https://doi.org/10.23919/MIPRO52101.2021.9596664

Jain, A., Ahuja, G., Anuranjana, & Mehrotra, D. (2016). Data mining approach to analyse the road accidents in India. *2016 5th International Conference on Reliability, Infocom Technologies and Optimization,* *ICRITO 2016: Trends and Future Directions*, 175–179. https://doi.org/10.1109/ICRITO.2016.7784948

Jason Brownee. (2020). *gaussian classifier*. https://machinelearningmastery.com/gaussian-processes-for-classification-with-python/

Manasa, J. M., Bhattacharjee, S., Ghosh, S. K., & Mitra, S. (2015, February 9). Spatial decision tree for accident data analysis. *9th International Conference on Industrial and Information Systems, ICIIS 2014*. https://doi.org/10.1109/ICIINFS.2014.7036565

Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Madni, H. A., & Bisogni, C. (2021). RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model. *IEEE Access*, *9*, 128359–128371. https://doi.org/10.1109/ACCESS.2021.3112546

Md. Farhan Labib, Ahmed Sady Rifat, Md. Mosabbir Hossain, Amit Kumar Das, & Faria Nawrine. (2019). *Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh*.

Nick Hotz. (2022). *Data Science Process*. https://www.datascience-pm.com/crisp-dm-2/

S. Shanthi, & R. Geetha Ramani. (2012). *Feature Relevance Analysis and Classification of Road Traffic Accident Data through Data Mining Techniques*.

Sruthi ER. (2021). *Random Forest*. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Sunil Ray. (2017). *Naive bayes*. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Thaduri, A., Polepally, V., & Vodithala, S. (2021). Traffic accident prediction based on CNN model. *Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021*, 1590–1594. https://doi.org/10.1109/ICICCS51141.2021.9432224

V.A, O., & A.A, E. (2014). Traffic Accident Analysis Using Decision Trees and Neural Networks. *International Journal of Information Technology and Computer Science*, *6*(2), 22–28. https://doi.org/10.5815/ijitcs.2014.02.03

Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Cui, Z., & Wang, Z. (2019a). Traffic accident's severity prediction: A deep-learning approach-based CNN network. *IEEE Access*, *7*, 39897–39910. https://doi.org/10.1109/ACCESS.2019.2903319

Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Cui, Z., & Wang, Z. (2019b). Traffic accident's severity prediction: A deep-learning approach-based CNN network. *IEEE Access*, *7*, 39897–39910. https://doi.org/10.1109/ACCESS.2019.2903319

Appendix

Plagiarism Report[1]

Accident Severity Detection and Prediction Using Machine Learning Techniques

by Sharon Joseph

Submission date: 15-Aug-2022 05:20PM (UTC+0530)
Submission ID: 1882760914
File name: n_and_Prediction_Using_Machine_Learning_Techniques_-_Sharon.docx (1.01M)
Word count: 6288
Character count: 35737

[1] Turnitn report to be attached from the University.

Accident Severity Detection and Prediction Using Machine Learning Techniques

ORIGINALITY REPORT

16%	9%	7%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Teesside Student Paper	3%
2	www.simonsfoundation.org Internet Source	2%
3	E. Suganya, S. Vijayarani. "Analysis of road accidents in India using data mining classification algorithms", 2017 International Conference on Inventive Computing and Informatics (ICICI), 2017 Publication	1%
4	Manasa, J. M., Shrutilipi Bhattacharjee, Soumya K. Ghosh, and Sudeshna Mitra. "Spatial decision tree for accident data analysis", 2014 9th International Conference on Industrial and Information Systems (ICIIS), 2014. Publication	1%
5	www.researchgate.net Internet Source	1%

6	Submitted to Sheffield Hallam University Student Paper	1 %
7	ijarcce.com Internet Source	1 %
8	Submitted to Zambia Centre for Accountancy Studies Student Paper	1 %
9	link.springer.com Internet Source	1 %
10	repository.tudelft.nl Internet Source	<1 %
11	Submitted to Middlesex University Student Paper	<1 %
12	Olutayo V.A, Eludire A.A. "Traffic Accident Analysis Using Decision Trees and Neural Networks", International Journal of Information Technology and Computer Science, 2014 Publication	<1 %
13	businessdocbox.com Internet Source	<1 %
14	Ming Zheng, Tong Li, Rui Zhu, Jing Chen, Zifei Ma, Mingjing Tang, Zhongqiang Cui, Zhan Wang. "Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network", IEEE Access, 2019	<1 %

Publication

15	www.bib.irb.hr Internet Source	<1 %
16	www.irjmets.com Internet Source	<1 %
17	Rabia Emhamed AlMamlook, Keneth Morgan Kwayu, Maha Reda Alkasisbeh, Abdulbaset Ali Frefer. "Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity", 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019 Publication	<1 %
18	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1 %
19	www.journaltocs.ac.uk Internet Source	<1 %
20	d.researchbib.com Internet Source	<1 %
21	digitalcommons.unf.edu Internet Source	<1 %
22	ijsrcseit.com Internet Source	<1 %
23	Submitted to Lovely Professional University Student Paper	<1 %

24	Submitted to Colorado State University, Global Campus Student Paper	<1 %
25	Mateja Iveta, Aleksander Radovan, Branko Mihaljevic. "Prediction of Traffic Accidents Severity Based on Machine Learning and Multiclass Classification Model", 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), 2021 Publication	<1 %
26	Salahadin Seid Yassin, Pooja. "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach", SN Applied Sciences, 2020 Publication	<1 %
27	repository.nwu.ac.za Internet Source	<1 %
28	www.coursehero.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

GitHub Link

https://github.com/sharonjoseph/Capstone-1-Accident-Severity-Detection