Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 10(108)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10101879

Analysis of the severity of transport vehicle accidents by a comparative study of machine learning models

Mensouri Houssam[1*], Azmani Abdellah[2] and Azmani Monir[2]
Research Scholar, Department of Intelligent Automation Laboratory, FST of Tangier, Abdelmalek Essaadi University, Tetouan, Morocco[1]
Professor, Department of Intelligent Automation Laboratory, FST of Tangier, Abdelmalek Essaadi University, Tetouan, Morocco[2]

Received: 07-July-2023; Revised: 01-November-2023; Accepted: 03-November-2023

Abstract

Traffic accidents pose a significant global threat to public safety, with the World Health Organization (WHO) estimating that they claim the lives of approximately 1.25 million individuals each year. Without intervention, traffic accidents are projected to become the leading cause of death by 2030. Predicting accident severity and understanding their underlying causes represent crucial steps in developing effective strategies to prevent accidents and enhance overall traffic safety. This paper presented an in-depth analysis of accident severity prediction, considering a wide range of factors, including the vehicle, driver, environmental conditions, and more. The study aims to predict the extent of the severity of traffic accidents using a comprehensive dataset comprising over 4 million incidents that occurred across 49 states in the United States of America (USA) between February 2016 and December 2020. Various machine learning models, including logistic regression (LR), support vector machine (SVM), decision tree (DT), and random forest (RF), were implemented and rigorously evaluated against multiple performance metrics. The achieved results reveal that the RF model stands out with the highest accuracy of 91% in predicting accident severity. Additionally, this model demonstrates excellent performance across additional evaluation metrics, including a precision rate of 89%, a recall rate of 91%, a root mean square error (RMSE) of 18%, and an F1 score of 89%. These findings emphasize the exceptional predictive power and robustness of the RF model, making it a highly promising approach for real-world traffic accident scenarios. This research provides valuable insights into predicting accident severity, which is crucial for the development of effective accident prevention strategies and improvements in traffic safety.

Keywords

Accident, Severity prediction, Machine learning, RMSE.

1.Introduction

Due to the significant rise in both fatal injuries and property damage caused by road collisions, addressing this issue has become a pressing global concern [1]. These accidents have devastating consequences, leading to a significant loss of lives and extensive property damage worldwide. In 2020, according to the organisation for economic co-operation and development (OECD) (https://stats.oecd.org), road accidents claimed 3,347 lives in Morocco, while Algeria witnessed over 3,500 fatalities. France recorded 2,579 deaths and approximately 60,000 injuries due to road accidents in the same year. Canada experiences an annual average of 1,700 road accident fatalities, along with countless injuries. Similarly, Australia witnesses over 1,000 fatalities and numerous serious injuries annually. The United States (U.S.), with over 38,000 collision-related deaths in 2020 and millions of dollars in property damage, ranks among the countries with the highest rates of road accidents. Given these alarming statistics, several countries, including Canada [2], Australia [3], and the U.S. [4], have been actively striving to develop innovative systems that can effectively prevent accidents.

Effectively addressing the gravity of road crashes entails tackling several challenges. It is crucial to thoroughly consider factors such as the count of fatalities, injuries, and property losses to determine

[*] Author for correspondence

the severity of accidents. However, accurately forecasting accident severity remains a crucial yet complex task in the field of accident management. Driven by the imperative to mitigate the destructive consequences of road accidents, the aim of this research is to develop a robust model that can reliably forecast accident severity. Such a model plays a critical part in guiding the decision-making process and can facilitate the implementation of targeted prevention measures. To achieve this goal, the research has set forth the following objectives: (1) Undertake an extensive literature review to establish a solid foundation for the study, (2) Identify and analyze key factors contributing to the traffic accident severity, (3) Employ machine learning techniques to analyze and clean the collected data, (4) Implement and evaluate different models, including logistic regression (LR), support vector machine (SVM), decision tree (DT), and random forest (RF), to determine the most efficient approach for predicting accident severity.

To successfully apply a machine learning algorithm to a database of 4 million records, access to both robust hardware and efficient software capable of handling demanding computational tasks is essential. In our study, we recognized the computational challenges posed by the extensive dataset, and we utilized the resources and hardware provided by our institution's research computing facility. The experiment was conducted on a server equipped with a substantial computational capacity, boasting 48 cores derived from two AMD EPYC 7402 24-core processors. This high-performance infrastructure was essential for the efficient training of the machine learning models in our research. Given the sheer volume of data, comprising over 4 million traffic accident incidents, we were well aware of the resource-intensive nature of our study. To address this, we employed a cluster of high-performance graphics processing units (GPUs) known for their exceptional efficiency in handling deep learning tasks. Leveraging these GPUs enabled us to significantly reduce training times and effectively manage the extensive dataset. Furthermore, we implemented distributed training techniques across multiple GPUs and carefully optimized batch sizes to maximize the utilization of our computational resources. This approach allowed us to strike a balance between efficient resource utilization and training effectiveness.

As for the software requirements, we employed popular deep learning frameworks, namely TensorFlow and PyTorch, for implementing and training our machine learning models. These frameworks are widely recognized for their compatibility with GPU acceleration and their extensive libraries for building and optimizing deep neural networks. The entire research project was conducted in Python, which served as the primary programming language for data preprocessing, model development, and evaluation. We used various Python libraries for data preprocessing, including pandas for data manipulation, NumPy for numerical operations, sci-kit-learn for feature engineering and model evaluation, and Matplotlib and Seaborn for data visualization.

The article begins by offering a comprehensive overview of the relevant literature in Section 2, establishing a solid foundation for our study. It then conducts an in-depth analysis of key factors contributing to the severity of traffic accidents, shedding light on critical determinants. In Section 3, the article outlines the sophisticated machine-learning techniques employed for data analysis and cleansing, ensuring the robustness of our approach. Section 4 explores the evaluation criteria meticulously utilized to assess the performance of various models, providing insight into our rigorous analysis. The subsequent section showcases the results of our analysis, accompanied by a thorough discussion that delves into the nuances and implications of our findings in Section 5. The article concludes by summarizing key takeaways and offering insights into their practical implications for enhancing road safety measures.

2.Literature review

2.1Road safety

Road safety issues have become a global concern [5], as traffic accidents cause injuries, deaths, and major property loss [6]. In recent years, road accidents have garnered increasing attention across multiple disciplines [7]. Researchers in the field of accident prevention have made considerable efforts to forecast accident severity and understand its determinants, culminating in two distinct categories of findings. The first category underscores the significance of driver behavior and driving practices as substantial risk factors influencing both the likelihood and intensity of accidents [8, 9]. Studies within this category delve into the intricate interplay between driver conduct and accident outcomes. The second category of research encompasses a comprehensive examination of interactions among three essential elements: the driver, the vehicle, and the environment

[10, 11]. This section of literature explores the multifaceted relationships between these factors and accident severity, shedding light on their intricate dynamics. This study falls within the ambit of the second category of literature, which scrutinizes the determinants of road accident severity through a comprehensive consideration of multiple interacting elements.

Furthermore, much of the research in recent decades has primarily relied on small-scale datasets with limited coverage, often confined to a few road segments or a single city [12–14]. Notable examples include the work of Chang et al. [12], who utilized road geometry, annual average daily traffic, and weather data to predict accident frequency for a highway road using a neural network model. Kumar and Toshniwal [13] applied data mining techniques to extract association rules for causality analysis, utilizing a small-scale dataset. Similarly, Wenqi et al. [14] applied a convolutional neural network model for accident prediction on a road segment. While these insights and findings are intriguing, the limited scale of the datasets used raises questions about the applicability and generalizability of the results.

Certainly, there are numerous studies that have employed larger-scale datasets [15, 16]; however, in many cases, these datasets have either been private, not easily accessible, or outdated. To address these challenges, we utilize a large-scale accident dataset with countrywide coverage, featuring comprehensive data attributes, including location, time, weather, period of day, and points of interest (POI) annotations.

2.2Factors related to the severity of road accidents

The objective of the traffic accident severity prediction model is to correctly anticipate the gravity of the traffic accident, and identify the major contributing variables [17]. Determining these factors helps reduce accident rates, decrease accident severity, and mitigate injuries, fatalities and property loss [18]. The identification of relevant factors is a prerequisite before starting the predicting process since they serve as the input for the severity prediction models. Previous works [19] have demonstrated that variables about driver, vehicle, road, weather, and external factors are the primary ones related to accident severity.

2.2.1Driver related factors

Many driver characteristics such as gender, age, behaviours, and the psychological and psychic state of the driver are potential sources of accidents on the roads. Young age drivers are linked to elevated levels of accident severity. Previous research shows that younger drivers are more likely than experienced drivers to have accidents due to risky behaviours such as speeding, and alcohol and drug abuse [20]. Gender also influences accident severity; males are more prone to fatal injuries than females [21]. The ingestion of alcohol and drugs, including substances like cocaine, marijuana, or any other illicit drug, before or during driving, impairs driver judgment due to their euphoric effects. These impairments have an adverse effect on the severity of accidental injuries [22]. Furthermore, failure to follow safe driving rules contributes to an escalation in the gravity of an accident such as not wearing a seat belt while driving which are effective at avoiding deaths [20]. Lastly, the psychological state of the driver can quickly have a negative impact on the trajectory. It is the driver's visual abilities that allow him to see danger and transmit the information to his brain so that he can analyse the situation and make the appropriate decision. When a driver gets behind the wheel when he is tired or drowsy, his alertness and reaction time deteriorate.

2.2.2Vehicle-related factors

Other than driver factors, vehicle factors are also having an impact on crash severity. The age of the vehicle [23] is related to accident severity, as the age of a vehicle increases, the probability of it being implicated in a severe accident also rises. Additionally, mechanical malfunctions further amplify the severity of accidents [24]. Particularly, the cost of cars has an unfavorable correlation with accident severity. This objective result seems unexpected but can be explained by drivers not taking big risks if the car is expensive. Additionally, the size of the vehicle has an impact on how likely accidents are to occur; bigger vehicles are more likely to result in deadly accidents [21].

2.2.3Road related factors

The intensity of collisions is greatly influenced by the state of the roads. Several road characteristics could influence the intensity of an accident. The existence of traffic lights diminishes the occurrence of side collisions, thereby reducing the likelihood of severe accidents [25]. Moreover, roads equipped with cruise control are correlated with a reduced likelihood of causing significant injuries compared to roads without cruise control [26]. Moreover, accidents that occur at intersections are more prone to be severe compared to those happening on other roads. [27]. In addition, road width is also an important factor. Fewer serious and fatal accidents occur on wider roads [21].

2.2.4Weather related factors

In general, the major environmental factors that most affect weather conditions are temperature, atmospheric pressure, wind, humidity, precipitation, and nebulosity (clouds, fog). Together, these factors give the weather at a given location at a given time. Weather conditions also contribute to influencing the accident severity. Various weather factors affect the severity of accidents, serious accidents are more likely to happen on foggy days due to the visibility [28], and rain and snow, in turn, are strongly linked to it [29].

2.2.5Visibility related factors

Besides weather conditions, environmental factors also influence the severity of accidents. Reduced visibility during night time increases the fatality of accidents compared to those occurring during daylight hours [30].

3.Methodology

3.1Machine learning process

Figure 1 shows the proposed study process. To find and handle corrupted or absent records, data cleansing was first carried out. After that, most of the features underwent exploratory data analysis (EDA) and feature engineering. Four classifiers are implemented to develop predictive models. Finally, different evaluation criteria were used to evaluate those models.

3.1.1Data

The dataset utilized in this study was graciously made available by the National Highway Traffic Safety Administration (NHTSA), a division operating under the U.S. Department of Transportation with the objective of mitigating the consequences of vehicle crashes, including injuries, fatalities, and economic repercussions. Spanning all 49 states across the U.S., this extensive dataset comprises over four million records documenting traffic accidents that occurred between 2016 and 2020. It is openly accessible to the academic community at https://smoosavi.org/datasets/us_accidents and has been employed in several notable research articles [1, 31], underscoring its value as a valuable resource for academic investigation.

To ensure the dataset's comprehensiveness, data collection was conducted through a systematic approach, amalgamating information from various sources, including law enforcement reports, emergency medical service records, traffic cameras, and citizen reports. The dataset has been meticulously categorized into four principal categories: 'Environment' (encompassing weather conditions and visibility), 'Location' (providing geographical coordinates for precise accident locations), 'Infrastructure' (detailing road characteristics and traffic infrastructure), and 'Basic' (providing foundational accident information) as shown in *Figure 2*. For a comprehensive understanding of these categories and data attributes, *Tables 1 to 5* offer detailed explanations and exemplifications. This comprehensive dataset, coupled with its systematic categorization and extensive data collection, forms a robust basis for our research, ensuring the depth and reliability of our analysis.

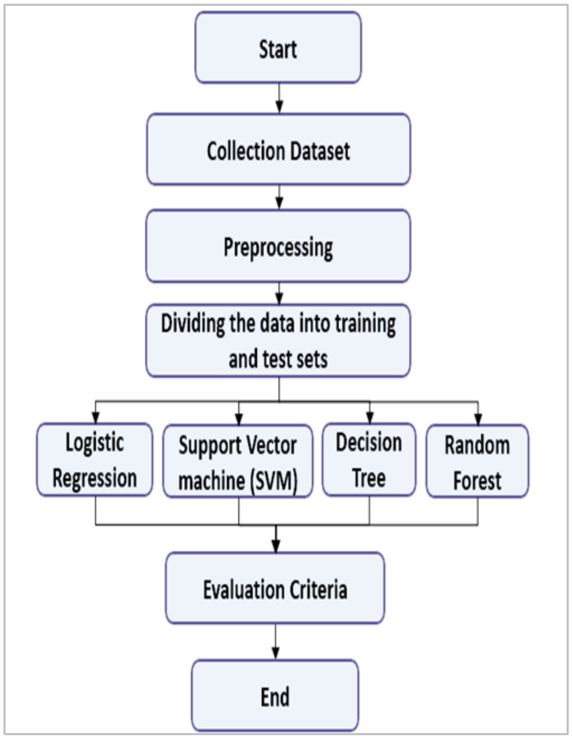


Figure 1 Process proposed for this study

3.1.1.1Dataset collection mechanism

Figure 3 provides an overview of the dataset creation process, while the subsequent sub-sections offer detailed descriptions of each individual step.

3.1.1.1.1 Traffic data collection

In their study, as documented in [31], the authors utilized real-time traffic data collection methods to gather streaming traffic data from two prominent real-time data providers, namely, 'MapQuest Traffic' [32] and 'Microsoft Bing Map Traffic' [33]. These data providers offer comprehensive APIs that disseminate real-time traffic events, encompassing accidents, congestion, and other relevant information sourced from diverse entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and road-network traffic sensors. In their research, the authors meticulously compiled a substantial dataset of 2.27 million traffic accident instances, with 1.73 million cases sourced from MapQuest and an additional 0.54 million cases from Bing.

Following real-time data collection, data integration involved the elimination of cases duplicated between the two sources and the construction of a cohesive dataset. The criteria for identifying duplicates involved considering two events as such if their haversine distance and recorded times of occurrence were both below predefined thresholds, specifically set at 250 meters and 10 minutes, respectively.

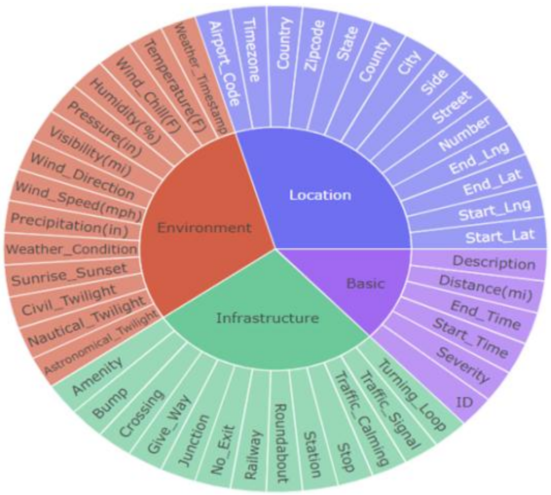


Figure 2 Data categorizing

3.1.1.1.2 Data augmentation

The collected data was enriched through a two-fold augmentation process:

- Reverse Geocoding: Initially, the raw traffic accident records contained solely GPS data. To provide a more comprehensive understanding of the accidents' locations the Nominatim tool is used. This tool enabled the conversion of GPS coordinates into detailed addresses, encompassing street number, street name, relative side (left/right), city, county, state, country, and zip code.
- Weather data and POI: Weather information plays a crucial role in providing context for traffic accidents. For this purpose, the weather underground API was employed to gather pertinent weather data for each accident. Beyond weather details, the dataset was further enriched by the inclusion of POI. These POI annotations were sourced from open street map (OSM) for the United States. Notably, only the POI annotations within a defined distance threshold were added to the traffic accident dataset.

3.1.2Pretreatment

Before running a machine learning algorithm on the data, the pre-processing phase must be completed. During this phase, the relevant features are selected, and the data is cleaned of any missing or abnormal values. The aim is to verify that the data is in an optimal state before applying the algorithm. The first step of the pre-processing phase is performed independently on each dataset. The objective is to develop a model that can aid in enhancing the overall system's efficiency during road accidents in the future. Consequently, only the pertinent variables that can be employed as input to assess the accident's severity are retained. This implies that the retained items consist solely of information available from the accident description provided during the emergency call or any deducible information derived from it.

3.1.2.1Remove unnecessary features

The ID attribute lacks meaningful information regarding the event and can be ignored. Similarly, variables such as 'TMC', 'Distance', 'End-Time', 'Duration', 'End-Lat', and 'End-Lng' are not suitable for predicting serious accidents as they can only be observed or measured after the accident has already taken place. Additionally, 'Country' and 'Turning-Loop' are also removed from the analysis since they only consist of a single class and do not contribute to the prediction task. These variables are excluded from the dataset as they do not provide any valuable insights or predictive power for identifying severe accidents.

3.1.2.2Minimize categorical characteristics

Upon careful examination of the categorical features, it becomes evident that there is some disorderliness in "Wind_Direction" and "Weather_Condition". Consequently, it is necessary to perform data cleaning and organize these variables appropriately.

a) Wind direction: Wind direction characteristics before cleaning were [‘Calm’ ‘SW’ ‘SSW’ ‘WSW’ ‘WNW’ ‘NW’ ‘West’ ‘NNW’ ‘NNE’ ‘South’ ‘North’ ‘Variable’ ‘SE’ ‘SSE’ ‘ESE’ ‘East’ ‘NE’ ‘ENE’ ‘E’ ‘W’ nan ‘S’ ‘VAR’ ‘CALM’ ‘N’]. It is evident that there are repetitions. After cleaning treatment Wind direction was simplified to: [‘CALM’ ‘SW’ ‘S’ ‘W’ ‘NW’ ‘VAR’ ‘SE’ ‘E’ ‘NE’ nan].

b) Weather conditions: Each year, vehicle accidents associated with weather conditions claim more lives than major weather catastrophes. The road weather management program reports that the majority of weather-related accidents occur on wet roads and during periods of precipitation. Additionally, winter weather conditions and fog are identified as two significant factors contributing to weather-related accidents. The initial step in identifying these three weather conditions is to analyze the data contained within the "Weather_Condition" feature.

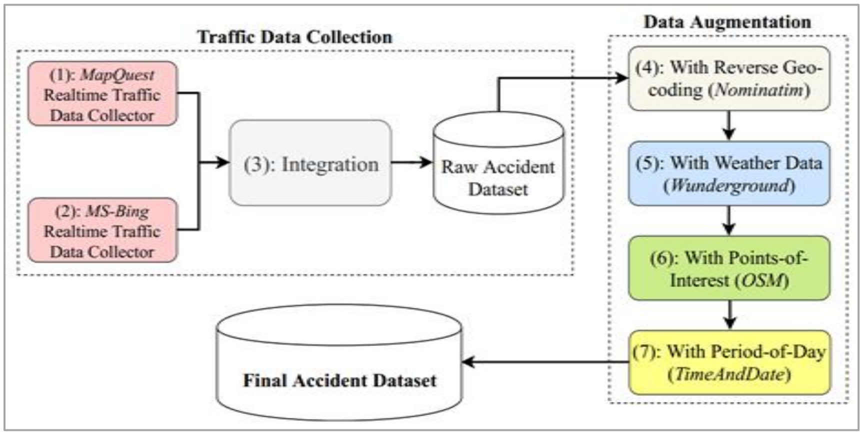

Traffic Data Collection
(1): MapQuest Realtime Traffic Data Collector
(2): MS-Bing Realtime Traffic Data Collector
(3): Integration
Raw Accident Dataset
Data Augmentation
(4): With Reverse Geo-coding (Nominatim)
(5): With Weather Data (Wunderground)
(6): With Points-of-Interest (OSM)
(7): With Period-of-Day (TimeAndDate)
Final Accident Dataset

Figure 3 Process of creating traffic accident dataset [31]

Weather Conditions feature are ['Clear', 'Cloudy', 'Drizzle', 'Dust', 'Dust Whirlwinds', 'Fair', 'Fog', 'Funnel Cloud', 'Hail', 'Haze', 'Heavy ', 'Heavy Drizzle', 'Heavy Ice Pellets', 'Heavy Rain', 'Heavy Rain Showers', 'Heavy Sleet', 'Heavy Snow', 'Heavy T-Storm', 'Ice Pellets', 'Light ', 'Light Drizzle', 'Light Fog', 'Light Hail', 'Light Haze', 'Light Rain', 'Light Snow Shower', 'Light Snow Showers', 'Light Thunderstorm', 'Low Drifting Snow', 'Rain', 'Rain Shower', 'Rain Showers', 'Sand', 'Scattered Clouds', 'Shallow Fog', 'Smoke', 'Snow', 'Snow Grains', 'Snow Showers', 'Squalls', 'T-Storm', 'Thunder', 'Thunderstorm', 'Thunderstorms', 'Tornado', 'Windy', 'Wintry Mix']

Considering the variety of meteorological conditions described earlier, various features were categorized for common weather conditions. Weather Conditions after grouping were reduced to [‘Clear’, ‘Cloudy’, ‘Fog’, ‘Rain’, 'Heavy_Rain', ‘Smoke’, ‘Snow’, 'Heavy_Snow', ‘Windy’].

3.1.2.3Fixing the date and time

The average time difference between 'Start-Time' and 'Weather_Timestamp' is calculated to be 0 days. Given that 'Weather_Timestamp' is almost identical to 'Start-Time', it is feasible to retain only the 'Start-Time' variable. Consequently, the 'Start-Time' can be associated with the 'Year', 'Month', 'Weekday', 'Day', 'Hour', and 'Minute' attributes, as demonstrated in *Table 6*. This simplification allows for a more streamlined and efficient representation of the temporal information associated with the accidents.

3.1.2.4Processing missing data

a) Features to remove: A significant proportion, specifically over 60%, of the data for the 'Number', 'Wind_Chill(F)', and 'Precipitation(in)' variables are missing. Based on previous research indicating a weak relationship between 'Number' and 'Wind_Chill(F)' variables with accident severity, they will be excluded. However, 'Precipitation(in)' is deemed a potentially valuable predictor and will be handled separately as a distinct function.

b) Distinct feature: A new feature is introduced to handle missing values in the "Precipitation" variable. The missing values are substituted with the median value, as indicated in *Table 7*.

c) Imputation of Value: A significant portion of the remaining columns shows just a few numbers of missing data, which could be addressed through filling or imputation methods.

Continuous weather data features: It includes various features such as temperature, humidity, pressure, visibility, and wind speed. However, some of these features may have missing values. To address this, the weather data is organized by location and time, since time is inherently linked to weather conditions. The location feature chosen for this purpose is 'Airport_Code' since the weather data is sourced from airport-based weather stations.

Next, the data is categorized based on the 'Start-Month' instead of the 'Start-Hour'. This decision is made to minimize computational costs and to deal with fewer missing values. By grouping the data in this way, we can establish distinct subsets based on the month in which the weather data was recorded.

Once the data is grouped accordingly, the missing values within each group are imputed using the median value. This means that for each specific location, period (start month), and weather feature, the missing values are exchanged with the median of that particular group. This approach allows us fill in the gaps in the weather data while minimizing the impact of outliers or extreme values. Categorical weather features: When it comes to categorical weather characteristics, a different approach is taken to handle missing values. Instead of using the median, the majority value is used to replace these missing values.

3.1.3Treatment

This article aims to create a web application that employs machine learning algorithms to accurately predict the severity of accidents. To do this, different models are applied and evaluated and the best model showing the best measure is selected. The measures used to assess each model are then loaded after the data has been divided into training and testing portions.

3.1.3.1Logistic regression(LR)

It is the most often used regression model [34], this model is used in machine learning to help create accurate predictions. It is a statistical model that predicts the probability of an event occurring (1) or not (0) based on regression coefficients. It shows the relationship between traits and calculates the probability of a particular outcome which always varies between 0 and 1. An event is more likely to happen when the expected variable is greater than a threshold but not when this value is less than the same threshold [35]. LR can be defined as shown in Equation 1 and Equation 2 [34].

$$P(x) = \frac{\exp(w.x+b)}{1+\exp(w.x+b)} \qquad (1)$$

$$P(x) = \frac{1}{1+\exp(w.x+b)} \qquad (2)$$

Where x $\in$ Rn is the input entity, Y $\in$ {0,1} is the label vector w denotes the weight, b represents the weight value, and w.x signifies the scalar product of the matrices (Equation 3). When two probability values are compared, LR allocates x to the range that possesses the highest probability value [36].

$$P(x) = \frac{\exp(w.x+b)}{1+\exp(w.x+b)} \qquad (3)$$

Figure 4 shows the LR graphically.

Table 1 Basic category (Traffic attributes)

Attributes	**Description**
ID	Each accident record has a unique identifier, which is the ID
Severity	Severity is a number between 1 and 4 indicating the gravity of the accident. When the accident has a significant impact on traffic the severity equal 4, and 1 indicates the least impact on traffic.
Start-Time	Represents the start time and end time.
End-Time	
Start-Lat	Represent the latitude and Longitude of the start and the end point of the accident in GPS coordinates.
Start-Lng	
End-Lat	
End-Lng	
Distance(mi)	The distance of the road where the accident occurred

Table 2 Localization category (Address attributes)

Attributes	**Description**
Number	Represent data related to the address
Street	
City	
State	
Zipcode	
Country	
Side	Represent the relative side of the road
Timezone	Shows the timezone of the accident location

Table 3 Environment category (Weather Attributes)

Attributes	**Description**
Airport_Code	Indicates the nearest weather station to the mishap site, which is located at an airfield.

Attributes	**Description**
Weather_Timestamp	Displays the weather data record's time tag.
Temperature	Represent the temperature
Wind_Chill	Represent the wind chill
Humidity	Represent the humidity
Pressure	Displays the atmospheric pressure at the scene of the mishap.
Visibility	Displays visibility in the accident road
Wind_Direction	Demonstrates the mishap site's wind direction
Wind_Speed	Displays wind speed in the accident location
Precipitation	Displays the quantity of rain in inches
Weather_Condition	Shows the weather conditions in the accident location

Table 4 Environment category

Attributes	Description
Sunrise-Sunset	**Displays the time of day**
Civil-Twilight	
Nautical-Twilight	
Astronomical-Twilight	

Table 5 Infrastructure category (Point-Of-Interest attributes)

Attributes	**Description**
Amenity	Points of Interest (POI) are various attributes or annotations that indicate the presence of specific features in nearby locations. These features include speed bumps or humps, intersections, railroads, terminals (bus, trains, etc.), stop signs, traffic calming measures, lights, and turning loops.
Bump	
Crossing	
Give-Way	
Junction	
No-Exit	
Railway	
Roundabout	
Station	
Stop	
Traffic-Calming	
Traffic-Signal	
Turning-Loop	

Table 6 'Start-Time' and the 'Year', 'Month', 'Weekday', 'Day', 'Hour', and 'Minute' attributes

	Start time	**Year**	**Month**	**Weekday**	**Day**	**Hour**	**Minute**
0	2016-02-08 05:46:00	2016	2	0	39	5	346
1	2016-02-08 06:07:59	2016	2	0	39	6	367
2	2016-02-08 06:49:27	2016	2	0	39	6	409
3	2016-02-08 07:23:34	2016	2	0	39	7	443
4	2016-02-08 07:39:07	2016	2	0	39	7	459

Table 7 Handling missing values in the "Precipitation"

	Precipitation	**Precipitation-NA**
0	00.02	0
1	00.00	0
2	00.00	1
3	00.00	1
4	00.00	1
5	00.03	0

LR hyperparameter tuning:

Hyperparameter tuning plays a pivotal role in the machine learning process, as it is instrumental in aligning a model with the desired performance metric. In the context of LR, several key hyperparameters warrant careful consideration. These include the choice of solver, the specification of penalty, and the determination of regularization strength.

The solver serves as the optimization algorithm that guides the LR model. Penalty, on the other hand, is a regularization technique utilized to counteract overfitting by introducing a penalty term into the loss function. When employing L2 penalty, the penalty term equates to the sum of the squares of the model's weights, thereby influencing the cost function. The strength of this penalty is governed by the hyperparameter 'C,' where smaller 'C' values correspond to more robust regularization. Scikit-learn, commonly referred to as Sklearn, is a prominent machine learning library extensively utilized in the realm of data science within the Python programming language. In the context of implementing the LR model, the model's performance is greatly influenced by the values of its hyperparameters. To precisely identify the optimal hyperparameter values for the LR model, the GridSearchCV function is employed, playing a pivotal role in the hyperparameter tuning process. To systematically explore the hyperparameter space, a comprehensive grid search was conducted, and the detailed results are provided in *Table 8*. The graphical representation of the grid search's outcomes can be observed in *Figure 5*.

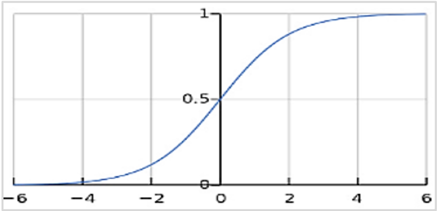


Figure 4 LR graph

Table 8 Hyperparameter grids used for LR

Model	grid size	Hyperparameter	Values range
LR	23	L2 penalty	{0.001, 0.03, 0.005, 0.007, 0.01, 0.03, 0.05, . . . ,1.3, 1.7, 5, 10, 50, 100, 500, 1000, 5000, 10000}

3.1.3.2Support vector machine

The SVM, a supervised machine learning technique, is grounded in the statistical theory of risk reduction. [37]. It was first developed by Vapnik et al [38] for classification problems, then it was updated to its latest version by Vapnik and Cortes in 1995 [39], and in 1998 it was developed and adapted to regression problems by Vapnik [40]. Today SVM is used in various areas that go from classification, and regression to character recognition and time series [41, 42]. Relative to alternative machine learning methods, SVM has some benefits. Focusing on reducing structural risk, makes it demand less data, have higher prediction accuracy, fewer adjusted parameters, and faster speed [43].

Support vector machine hyperparameter tuning:

Optimizing a machine learning model through hyperparameter tuning is a critical phase in the model development process. Specifically, in the case of SVM, the pivotal hyperparameters under consideration are 'C' and 'γ' (gamma).

- C: Acting as the penalty parameter for the error term, 'C' plays a fundamental role in striking a balance between achieving a smooth decision boundary and accurately classifying the training data points.
- γ (gamma): 'γ' determines the extent to which a single training example's influence extends, with low values indicating a broader influence and high values signifying a more localized impact.

In our pursuit of hyperparameter optimization, exhaustive hyperparameter grid search was executed, and the comprehensive results are presented in *Table 9*. GridSearchCV is the function used from Scikit-learn python library, to perform hyperparameter tuning in order to determine the optimal values for the model. The outcome of implementing the SVM model is thoughtfully illustrated in *Figure 6*.

Table 9 Hyperparameter grids used for SVM

Model	Total grid size	Hyperparameter	Values range
SVM	96	C	{0.01, 0.1, 1, 10, 100, 1000}
		Kernel	{linear, RBF}
		Loss	Squared hinge
		γ	{0.1, 0.3, 0.5, 1.0, 1.5, 2.0, scale, auto}

3.1.3.3Decision tree

DT algorithm is a supervised machine learning technique employed for solving regression or classification problems. It was initially proposed by Hunt et al. [44] in the 1960s. Research has proven that DTs are easy to understand and quick to learn and classify [45]. For this, it remains a fundamental and privileged tool to obtain classification rules [46]. This method has been employed extensively in fields like illness and mishap prediction [47].

DT has a hierarchical tree structure. Each tree consists of these four elements [48], as shown in *Figure 7*:

- Root: it is at the top of a Tree, and it represents the first node where the first split takes place.
- Internal Node: it is located between the leaf and root nodes; it signifies a test conducted on an attribute.
- Branch: it is each branch that connects the leaf node to the root one. it shows the test output.
- Edge: It indicates the direction to the next node.
- Leaf Node: Terminal nodes that predict the result of the DT.

The process of creating a DT model can be compared to a recursive division that divides a source set into subgroups according to a split criterion. Each derived subgroup is subjected to this procedure once more [49]. The recursion ends when splits are no longer useful or when all derived subsets can perform a single classification [50].‘‘If A, then the B’’ rule can be made by tracing the route from the parent node to the leaf node. *Figure 8* summarizes the DT model setting up process.

DT hyperparameter tuning:

Hyperparameter tuning for the DT model is a process that entails fine-tuning several parameters to enhance the model's performance. Within the realm of DTs, there are several critical hyperparameters that merit attention, including Max Depth, Min Samples Split, Min Samples Leaf, and Max Features.

To systematically explore the hyperparameter space, a comprehensive grid search is conducted, the results of which are presented in *Table 10*. The DT model featured in this article is crafted using the Sklearn library and leverages the GridSearchCV function for hyperparameter optimization. The outcome of implementing the DT model is thoughtfully illustrated in *Figure 9.*

3.1.3.4Random forest

As described in reference [51] the RF algorithm is a notable ensemble technique that draws inspiration from random fractional selection [52] and random subspace techniques [53]. It is a set of tree-based classifiers combined for classification [54]. The RF classifier is primarily an ensemble method that involves using a set of DTs, where each tree is trained separately on a different subset of the training dataset. This is accomplished by first aggregating different groups of the available predictive factors [51]. The distinctiveness of each DT in the RF stems from the random selection of predictive variable subsets. This randomization process contributes to lowering the overall variance of the classifier. In the end, the RF classifier combines the choices made by the individual trees for decision-making using a voting process analogous to majority voting. In other words, each DT selects a group for each observation, and the RF selects the category with the most votes.

Most classification methods use the RF machine learning algorithm. RF is an efficient and robust method that is known for its speed and ability to handle noisy data effectively. One of the key advantages of RF is its ability to detect non-linear patterns within the data. Another strength of RF is its capability to handle both numerical and categorical data with ease. It can efficiently process digital and categorical features, eliminating the need for extensive preprocessing or transformation of the data. Furthermore, RF has a built-in mechanism that prevents overfitting, even when additional trees are added to the ensemble [55].

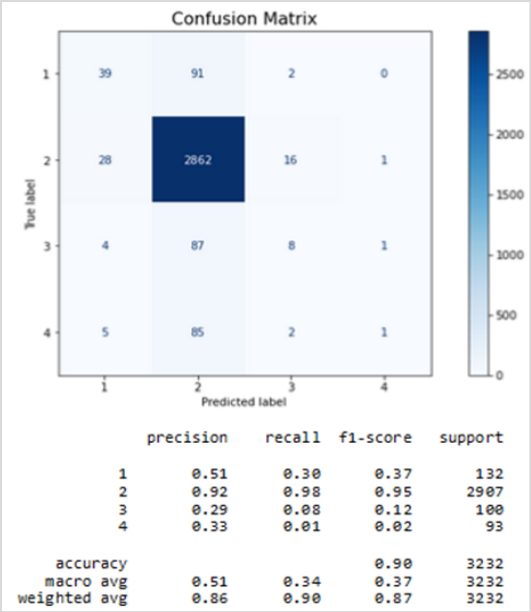
Random forest hyperparameter tuning:

The primary hyperparameters crucial for the RF model encompass the number of estimators, maximum features, minimum samples for splitting, minimum samples for leaf nodes, and the cost-complexity parameter denoted as α. These hyperparameters are detailed in *Table 11*.

For the purpose of hyperparameter tuning in the RF model, the same function, GridSearchCV, is employed as utilized previously. The outcome of implementing the RF model is visually represented in *Figure 10.*

3.2Evaluation criteria of the analysis

The most common criterion for evaluating a model is accuracy [56]. Accuracy, confusion matrix, precision, recall, root mean square error (RMSE), and f1 score are typical performance metrics for assessing each classifier's performance in this study, which works with a classification-based issue [57]. The objective of this evaluation is to determine the most performed machine learning algorithm that is precise, dependable and has the greatest accuracy.

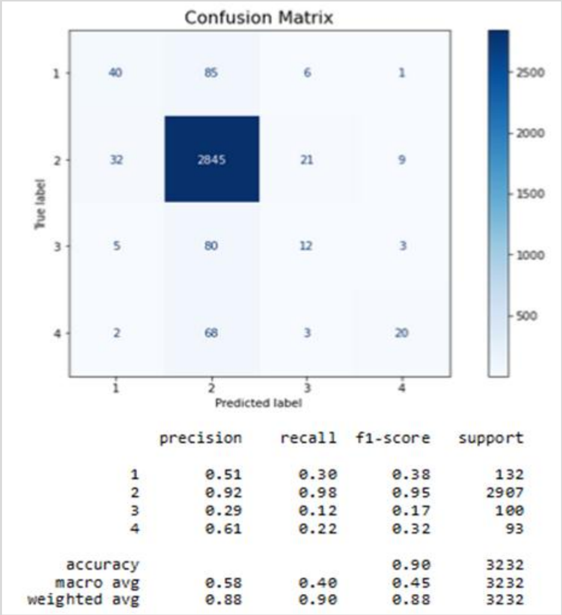


```
              precision    recall  f1-score   support

           1       0.51      0.30      0.37       132
           2       0.92      0.98      0.95      2907
           3       0.29      0.08      0.12       100
           4       0.33      0.01      0.02        93

    accuracy                           0.90      3232
   macro avg       0.51      0.34      0.37      3232
weighted avg       0.86      0.90      0.87      3232
```

Figure 5 Confusion Matrix of the " LR" model



```
              precision    recall  f1-score   support

           1       0.51      0.30      0.38       132
           2       0.92      0.98      0.95      2907
           3       0.29      0.12      0.17       100
           4       0.61      0.22      0.32        93

    accuracy                           0.90      3232
   macro avg       0.58      0.40      0.45      3232
weighted avg       0.88      0.90      0.88      3232
```

Figure 6 Confusion Matrix of the " SVM" model

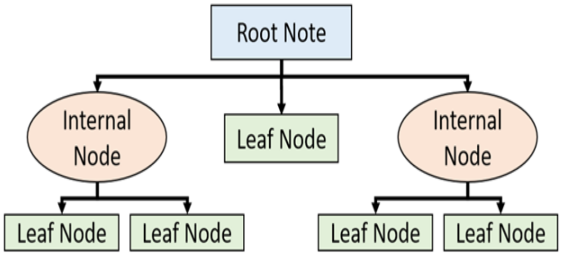


Figure 7 DT Structure

Table 10 Hyperparameter grids used for DT

Model	Total Grid Size	Hyperparameter	Values range
DT	135	Max depth	{2, 4, 6, 8, 10}
		Min samples split	{2, 5, 10}
		Min samples leaf	{1, 2, 3}
		Max features	{auto, sqrt, sqrt}

Table 11 Hyperparameter grids used for RF

Model	Total Grid Size	Hyperparameter	Values Range
RF	640	Num. Estimators	{32, 64, 128, 256, 512}
		Max Features	{sqrt, log2}
		Min. Samples Split	{2, 4, 8, 16}
		Min. Samples Leaf	{1, 2, 4, 8}
		Cost-Complexity α	{0., 0.001, 0.01, 0.1}

3.2.1Confusion matrix

The confusion matrix, presented in *Table 12*, provides a valuable tool for a more detailed analysis of the model's performance. The matrix is composed of four quadrants, each representing a distinct count of results or errors. The predicted values are at the top, while the actual values are depicted on the side.

- True positives (TP): The correct predictions that have been classified in the positive class '1'.
- False positives (FP): The incorrect predictions that have been falsely sorted in the class '1'.
- False negatives (FN): The incorrect predictions that have been falsely classified in the negative class '0'.
- True negatives (TN): The correct predictions that have been classified in the class '0'.

3.2.2Accuracy

Accuracy stands as the primary performance measure for classification models, calculated by summing the number of correct predictions (TN + TP) and dividing it by the total amount of predictions made (TP + TN + FP + FN)[57]. It is mathematically formulated as in Equation 4. As shown in *Figures 5*, *6*, *9*, and *10*, the accuracy value of LR, SVM, DT, and RF methods is respectively 90%, 90%, 90% and 91%.

$$\text{Accuracy} = \frac{\text{True Negatives+True Positives}}{\text{Number of Predictions}} \qquad (4)$$

3.2.3Precision

It represents the accuracy of positive detections about the ground truth [57] i.e., how the model predicted the positive values. This ratio, as represented in Equation 5, compares the number of true positives (TP) to the total number of predicted positive examples (TP + FP). Precision responds to the following question: " What percentage of positive identifications accurately corresponded to the actual instances? ». As shown in *Figures 5, 6, 9,* and *10*, the precision value of LR, SVM, DT, and RF methods is respectively 86%, 87%, 90%, and 89%.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}} \qquad (5)$$

3.2.4Sensitivity

Sensitivity or recall evaluates the ability of a model to accurately identify the positive predictions in comparison to the true reference values [57]. This is a performance metric that calculates the relationship between the number of TP and FN, as illustrated in Equation 6. The recall provides a response to the following question: "What proportion of actual positive results have been correctly identified? ». As shown in *Figures* 5, 6, 9, and *10*, the recall value of LR, SVM, DT, and RF methods is respectively 90%, 90%, 89% and 91%.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}} \qquad (6)$$

3.2.5Root mean square error

The RMSE serves to gauge the precision of a predictive model for target values [58]. It quantifies the typical disparity between the real value and the predicted value, computed as outlined in Equation 7. Here, 'n' signifies the count of estimates, 'yj' represents the present estimate, and 'ŷj' denotes the mean estimate value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(yj - \hat{y}j)^2} \qquad (7)$$

3.2.6F1 score

It is a method used to assess the performance of a model by utilizing the harmonic sum of sensitivity and precision [57]. It varies from 0 to 1. Termed as the F Score or F Measure, it serves as a metric to capture the equilibrium between accuracy and recall. Since accuracy is sometimes more important than recall or vice versa, it is still not possible to abandon one of them.

The F1 score can be defined as shown in Equation 8 below. As shown in *Figures 5, 6, 9,* and *10*, the F1 value of LR, SVM, DT, and RF methods is respectively 87%, 88%, 90% and 89%.

$$\text{F1} = \frac{2\times\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \qquad (8)$$

4.Result

Table 13 illustrates the comparison of results obtained from the machine learning models discussed earlier when applied to the accident data set. The differences in their performance are minimal, but the RF model stands out with the highest accuracy of 91% in predicting accident severity. Additionally, this model demonstrates excellent performance across additional evaluation metrics, such as a precision rate of 89%, a recall rate of 91%, an RMSE of 18%, and an F1 score of 89%.

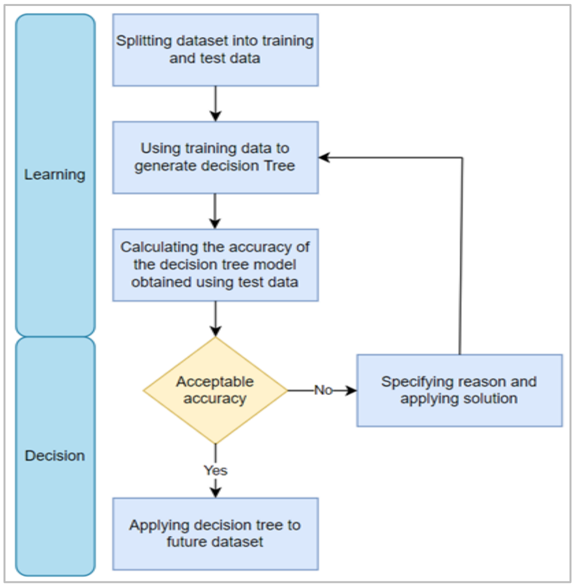

Learning
Decision
Splitting dataset into training and test data
Using training data to generate decision Tree
Calculating the accuracy of the decision tree model obtained using test data
Acceptable accuracy
No
Specifying reason and applying solution
Yes
Applying decision tree to future dataset

Figure 8 Procedure for setting up DT model

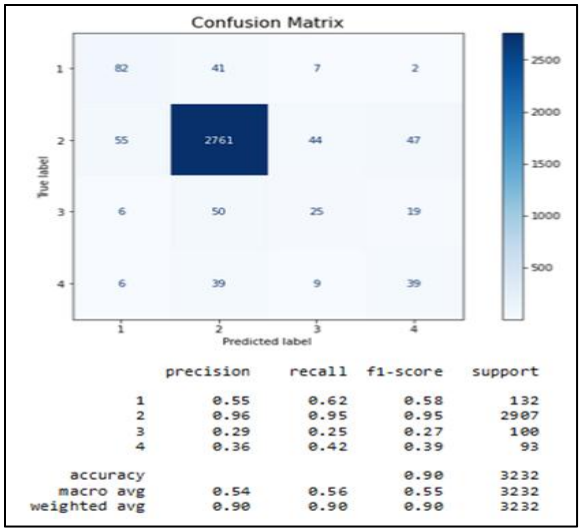
The scope of this work revolves around accidents and proposes a comparative study of four methods. While the process followed is specifically applied to the accident domain, its applicability extends across various sectors, encompassing workplace accidents, loss of goods, academic failure, and more. Apart from LR, SVM, DT, and RF, several other techniques can be employed.

In the field of accident severity prediction, several previous studies have delved into various methodologies, with neural networks being one of the prominent approaches [59, 60]. Despite the commendable efforts made by these earlier investigations, our present study stands out by demonstrating superior accuracy compared to its predecessors. Our work builds upon the foundation laid by these prior endeavours, showcasing notable improvements in both predictive performance and robustness. Our model excels in accurately assessing accident severity, outperforming existing methods, and setting a new benchmark in this critical domain. The comparative study of machine learning models presented in this research demonstrates promising results, particularly with the RF model exhibiting exceptional performance in predicting accident severity. These findings highlight the significance and effectiveness of the selected methods in addressing the challenges associated with accident severity prediction.

4.1Optimizing model performance

Optimizing the performance of our RF model was a critical aspect of our research. To achieve this goal, we conducted an extensive hyperparameter tuning process. We utilized a combination of grid search and random search techniques to systematically explore a wide range of hyperparameter configurations. Specifically, we fine-tuned several key hyperparameters for the RF model, including the number of trees in the forest, maximum tree depth, minimum samples required to split a node, and the maximum number of features considered for each split. These hyperparameters were varied within predefined ranges, and we employed cross-validation to rigorously assess their impact on model performance.

As a result of this iterative process, we were able to optimize our model's predictive accuracy, enhancing it from an initial 91% to a more robust 93%. This improvement underscores the model's ability to capture meaningful patterns in the data and make accurate predictions regarding traffic accident severity. The chosen evaluation metrics guided us in selecting the hyperparameter combinations that maximized the model's utility for our research.

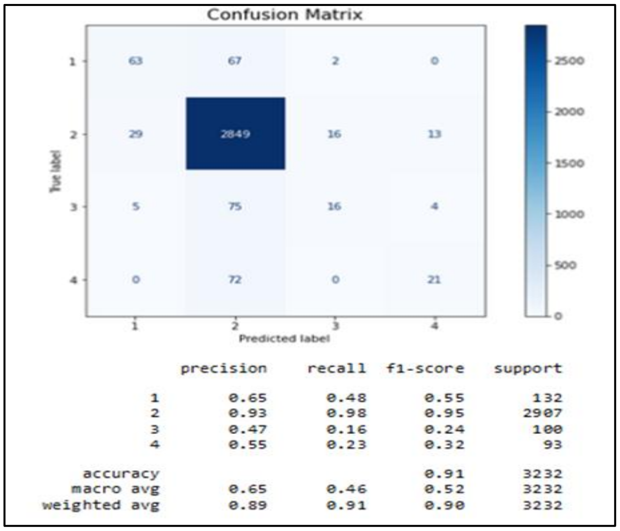


```
              precision    recall  f1-score   support

           1       0.55      0.62      0.58       132
           2       0.96      0.95      0.95      2907
           3       0.29      0.25      0.27       100
           4       0.36      0.42      0.39        93

    accuracy                           0.90      3232
   macro avg       0.54      0.56      0.55      3232
weighted avg       0.90      0.90      0.90      3232
```

Figure 9 Confusion matrix of the " DT " model



```
              precision    recall  f1-score   support

           1       0.65      0.48      0.55       132
           2       0.93      0.98      0.95      2907
           3       0.47      0.16      0.24       100
           4       0.55      0.23      0.32        93

    accuracy                           0.91      3232
   macro avg       0.65      0.46      0.52      3232
weighted avg       0.89      0.91      0.90      3232
```

Figure 10 Confusion matrix of the " RF " model

Table 12 Matrix of confusion and performance rates

	Predicted	
	Positive	**Negative**
Actual True	TP	FN
Actual False	FP	TN

5.Discussion

The contributions of this article are multifold. Firstly, it provides valuable insights into the factors influencing the severity of accidents, contributing to a better understanding of the problem. Secondly, it proposes a model that accurately forecasts accident severity, enabling informed decision-making and effective accident management procedures. Lastly, by evaluating different models and their performance, this research helps identify the most efficient approach for predicting accident severity, paving the way for further advancements in accident prevention and road safety.

In our pursuit of developing predictive models for accident severity assessment, it is essential to acknowledge both overarching limitations and specific challenges that emerged during our study. We recognize that despite the overall efficacy of our models, there are nuanced scenarios where they encounter difficulties. One notable instance arose in situations where accidents transpired under exceptionally rare or unprecedented conditions, such as extreme weather events or atypical traffic patterns. In these cases, our models, trained on historical data, occasionally grappled with accurate predictions due to the lack of precedent within the training dataset. Moreover, our models faced limitations when confronted with rapidly changing road conditions, such as unexpected road closures or construction activities. These dynamic variables, extending beyond the scope of our historical training data, presented challenges for real time predictions. Additionally, complexities emerged in cases involving multiple vehicles, each contributing distinct characteristics to the incident, making it challenging to pinpoint precise severity levels. Despite these challenges, our models continued to offer valuable insights in the majority of cases. These specific instances, however, underscore the importance of ongoing refinement and the integration of real-time data to address unique and evolving road safety scenarios. Furthermore, it is paramount to transparently acknowledge overarching limitations in our study, including our reliance on available data, which may contain inherent biases or limitations in terms of accuracy and coverage. The precision of our accident severity predictions is undeniably linked to the caliber and comprehensiveness of the data utilized for both model training and evaluation. Dynamic factors such as evolving road infrastructure, fluctuating weather conditions, and shifting driver behaviors inherently introduce variability into the prediction process. Our models' effectiveness is inevitably influenced by their ability to capture and adapt to these changing variables in real-time, which is an ongoing challenge. Additionally, while our research aims to contribute valuable insights, it cannot comprehensively account for all possible variables and scenarios that might influence accident severity. Factors beyond our analytical scope, such as specific local policies, traffic management practices, or unforeseen events, can also impact prediction accuracy. These limitations are inherent in the domain of accident severity prediction, and our study serves as a foundational step toward addressing these complex challenges. By transparently acknowledging these limitations, we strive to provide a balanced perspective on the capabilities and boundaries of our research, paving the way for future studies to build upon our findings and enhance the accuracy and applicability of accident severity prediction models.

Despite these limitations, this research provides a valuable foundation for understanding and predicting accident severity, and it serves as a starting point for further advancements in the field. Future studies can build upon these findings and explore innovative approaches to address the identified limitations, ultimately striving for enhanced road safety and accident prevention. Our research has embraced the constructive feedback of this review process, leading to a deeper exploration of the practical implications arising from our findings. Our study on accident severity prediction carries substantial real-world relevance, with the RF model, in particular, demonstrating noteworthy potential for practical application in enhancing road safety measures. This model's ability to accurately predict accident severity can facilitate timely and targeted interventions, offering the potential to reduce the severity of accidents and mitigate their adverse consequences. Transportation authorities and emergency services can leverage this predictive tool to allocate resources more efficiently, prioritize accident response based on severity, and proactively address accident-prone areas. Furthermore, our study envisions the development of predictive tools that can serve as integral components of intelligent transportation systems (ITS), designed to enhance overall traffic management and road safety. Our research insights contribute to informed decision-making, offering the potential to save lives, reduce injuries, and curtail economic losses associated with road accidents. In essence, this work not only enriches our understanding of accident severity prediction but also opens avenues for practical implementations and policy recommendations, underscoring its tangible impact on road safety and the well-being of communities at large. A complete list of abbreviations is shown in *Appendix I.*

Table 13 Models comparison

	Accuracy (Balanced)	**Accuracy**	**Precision**	**Recall**	**RMSE**	**F1 score**
DT	0.560086	0.899443	0.901062	0.899443	0.293	0.900048
RF	0.460782	0.912438	0.893591	0.912438	0.334	0.897549
SVM	0.404189	0.902537	0.878315	0.902537	0.302	0.884959
LR	0.342682	0.900371	0.863136	0.900371	0.182	0.873299

6.Conclusion

Accurately anticipating accident severity and identifying the crucial factors influencing it are crucial endeavours in bolstering road safety and curbing the incidence of fatalities and injuries arising from traffic accidents. The ability to predict accident severity, as showcased in this paper, constitutes a significant stride toward establishing safer and more sustainable transportation systems. This underscores the pivotal role played by data and analytics in shaping the future of transportation.

In this paper, before investigating the different models of accident severity classification, the data is first prepared and cleaned. Data cleaning helps to ameliorate the models since irrelevant data affects the model's performance. The different outcomes of the four models are analyzed and evaluated based on different evaluation criteria. The results indicate that, when compared to other models, the RF model exhibits superior suitability for accident severity prediction, initially achieving an accuracy of 91%. Following our rigorous hyperparameter tuning process, the accuracy of our RF model was further improved to an impressive 93%.

Through the comparison of various models, a methodological contribution to enhance the precision of severity estimates can be derived from this study. The model devised in this study enables the prediction of accident severity, whether it is from existing models or for accidents with limited information that have recently transpired. Therefore, in future works, the proposed model can be further developed for real-time implementation and integration into existing traffic management systems. This would allow for the immediate prediction of accident severity and enable proactive response strategies, such as timely dispatching of emergency services and dynamic traffic control measures. The utilization of a real-time implementation model will enhance the advancement of accident management systems, leading to increased effectiveness and efficiency, ultimately aiming to reduce fatalities, injuries, and property damage caused by road accidents.

Acknowledgment

This research is supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA), and the National Center for Scientific and Technical Research (CNRST) of Morocco (Smart DLSP Project - AL KHAWARIZMI IA-PROGRAM).

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statement

Mensouri Houssam: Conceived and designed the study, conducted experiments, collected and analyzed data, and drafted the manuscript. **Azmani Abdellah:** Contributed to the study design, data analysis, and reviewed the manuscript for important intellectual content. **Azmani Monir:** Provided guidance on the research direction, reviewed the manuscript. All authors have read and approved the final manuscript.

References

[1] Moosavi S, Samavatian MH, Parthasarathy S, Teodorescu R, Ramnath R. Accident risk prediction based on heterogeneous sparse data: new dataset and insights. In proceedings of the 27th SIGSPATIAL international conference on advances in geographic information systems 2019 (pp. 33-42). ACM.

[2] Mills BN, Andrey J, Hambly D. Analysis of precipitation-related motor vehicle collision and injury risk using insurance and police record information for Winnipeg, Canada. Journal of Safety Research. 2011; 42(5):383-90.

[3] Thompson JP, Baldock MR, Dutschke JK. Trends in the crash involvement of older drivers in Australia. Accident Analysis & Prevention. 2018; 117:262-9.

[4] Hao W, Kamga C, Daniel J. The effect of age and gender on motor vehicle driver injury severity at highway-rail grade crossings in the United States. Journal of Safety Research. 2015; 55:105-13.

[5] Zhao L, Wang C, Yang H, Wu X, Zhu T, Wang J. Exploring injury severity of non-motor vehicle riders involving in traffic accidents using the generalized ordered logit model. Ain Shams Engineering Journal. 2023; 14(5):101962.

[6] Pereira V, Bamel U, Paul H, Varma A. Personality and safety behavior: an analysis of worldwide research on road and traffic safety leading to organizational and policy implications. Journal of Business Research. 2022; 151:185-96.

[7] Wu J, Lu Y, Shi S, Zhou R, Liu Y. Research on the prediction model of hazardous chemical road transportation accidents. Journal of Loss Prevention in the Process Industries. 2023:105103.

[8] Alkheder S, AlRukaibi F, Aiash A. Risk analysis of traffic accidents' severities: An application of three data mining models. ISA Transactions. 2020; 106:213-20.

[9] Ji A, Levinson D. Injury severity prediction from two-vehicle crash mechanisms with machine learning and ensemble models. IEEE Open Journal of Intelligent Transportation Systems. 2020; 1:217-26.

[10] Lane MN. Pricing risk transfer transactions1. ASTIN Bulletin: The Journal of the IAA. 2000; 30(2):259-93.

[11] Miyajima C, Nishiwaki Y, Ozawa K, Wakita T, Itou K, Takeda K, et al. Driver modeling based on driving behavior and its evaluation in driver identification. Proceedings of the IEEE. 2007; 95(2):427-37.

[12] Chang LY. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Safety Science. 2005; 43(8):541-57.

[13] Kumar S, Toshniwal D. A data mining framework to analyze road accident data. Journal of Big Data. 2015; 2(1):1-8.

[14] Wenqi L, Dongyu L, Menghua Y. A model of traffic accident prediction based on convolutional neural network. In 2nd international conference on intelligent transportation engineering 2017 (pp. 198-202). IEEE.

[15] Yu L, Du B, Hu X, Sun L, Han L, Lv W. Deep spatio-temporal graph convolutional network for traffic accident prediction. Neurocomputing. 2021; 423:135-47.

[16] Yuan Z, Zhou X, Yang T. Hetero-convlstm: a deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In proceedings of the 24th international conference on knowledge discovery & data mining 2018 (pp. 984-92). ACM.

[17] Wahab L, Jiang H. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS one. 2019; 14(4):1-17.

[18] Zhu L, Lu L, Zhang W, Zhao Y, Song M. Analysis of accident severity for curved roadways based on bayesian networks. Sustainability. 2019; 11(8):1-17.

[19] Kopelias P, Papadimitriou F, Papandreou K, Prevedouros P. Urban freeway crash analysis: geometric, operational, and weather effects on crash number and severity. Transportation Research Record. 2007; 2015(1):123-31.

[20] Soderstrom CA, Dischinger PC, Kufera JA, Ho SM, Shepard A. Crash culpability relative to age and sex for injured drivers using alcohol, marijuana or cocaine. In annual proceedings/association for the advancement of automotive medicine 2005 (pp. 327-31). Association for the Advancement of Automotive Medicine.

[21] Zajac SS, Ivan JN. Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural connecticut. Accident Analysis & Prevention. 2003; 35(3):369-79.

[22] Beirness DJ, Simpson HM, Williams AF. Role of cannabis and benzodiazepines in motor vehicle crashes. Drugs and Traffic. 2005:12-21.

[23] Zhang G, Yau KK, Chen G. Risk factors associated with traffic violations and accident severity in China. Accident Analysis & Prevention. 2013; 59:18-25.

[24] Islam M. Multi-vehicle crashes involving large trucks: a random parameter discrete outcome modeling approach. Journal of the Transportation Research Forum. 2015; 54(1):77-104.

[25] Rifaat SM, Tay R, De BA. Effect of street pattern on the severity of crashes involving vulnerable road users. Accident Analysis & Prevention. 2011; 43(1):276-83.

[26] Christie SM, Lyons RA, Dunstan FD, Jones SJ. Are mobile speed cameras effective? a controlled before and after study. Injury Prevention. 2003; 9(4):302-6.

[27] Moore DN, Schneider IVWH, Savolainen PT, Farzaneh M. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accident Analysis & Prevention. 2011; 43(3):621-30.

[28] Al-ghamdi AS. Experimental evaluation of fog warning system. Accident Analysis & Prevention. 2007; 39(6):1065-72.

[29] Edwards JB. The relationship between road accident severity and recorded weather. Journal of Safety Research. 1998; 29(4):249-62.

[30] Behnood A, Al-bdairi NS. Determinant of injury severities in large truck crashes: a weekly instability analysis. Safety Science. 2020; 131:104911.

[31] Moosavi S, Samavatian MH, Parthasarathy S, Ramnath R. A countrywide traffic accident dataset. ArXiv Preprint Server. 2019:1-6.

[32] http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0. Accessed 24 October 2023.

[33] https://data.brla.gov/Transportation-and-Infrastructure/Baton-Rouge-Traffic-Incidents/2tu5-7kif. Accessed 24 October 2023.

[34] Zong F, Xu H, Zhang H. Prediction for traffic accident severity: comparing the Bayesian network and regression models. Mathematical Problems in Engineering. 2013; 2013:1-10.

[35] Satri J, El MC, Hachimi H. Artificial intelligence and machine learning for a better decision making in the public sector. In 8th international conference on optimization and applications 2022 (pp. 1-5). IEEE.

[36] Hidayat TH, Ruldeviyani Y, Aditama AR, Madya GR, Nugraha AW, Adisaputra MW. Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. Procedia Computer Science. 2022; 197:660-7.

[37] Güven I, Şimşir F. Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods. Computers & Industrial Engineering. 2020; 147:106678.

[38] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In proceedings of the fifth annual workshop on computational learning theory. 1992 (pp. 144-52). ACM.

[39] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20:273-97.

[40] Vapnik VN. An overview of statistical learning theory. IEEE Transactions on Neural Networks. 1999; 10(5):988-99.

[41] Li K, Zhou G, Yang Y, Li F, Jiao Z. A novel prediction method for favorable reservoir of oil field based on grey wolf optimizer and twin support vector machine. Journal of Petroleum Science and Engineering. 2020; 189:1-11.

[42] Sahana M, Rehman S, Sajjad H, Hong H. Exploring effectiveness of frequency ratio and support vector machine models in storm surge flood susceptibility assessment: a study of Sundarban Biosphere Reserve, India. Catena. 2020; 189:104450.

[43] Yu Y, Shao M, Jiang L, Ke Y, Wei D, Zhang D, et al. Quantitative analysis of multiple components based on support vector machine (SVM). Optik. 2021; 237:166759.

[44] Hunt EB, Marin J, Stone PJ. Experiments in induction. Academic Press. 1996.

[45] Jin C, Li F, Ma S, Wang Y. Sampling scheme-based classification rule mining method using decision tree in big data environment. Knowledge-Based Systems. 2022; 244:108522.

[46] Chang MY, Chiang RD, Wu SJ, Chan CH. Mining unexpected patterns using decision trees and interestingness measures: a case study of endometriosis. Soft Computing. 2016; 20:3991-4003.

[47] Clarke DD, Forsyth R, Wright R. Machine learning in road accident research: decision trees describing road accidents during cross-flow turns. Ergonomics. 1998; 41(7):1060-79.

[48] https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-decision-tree-classification-using-python/. Accessed 24 October 2023.

[49] Shorabeh SN, Samany NN, Minaei F, Firozjaei HK, Homaee M, Boloorani AD. A decision model based on decision tree and particle swarm optimization algorithms to identify optimal locations for solar power plants construction in Iran. Renewable Energy. 2022; 187:56-67.

[50] Maimon OZ, Rokach L. Data mining with decision trees: theory and applications. World Scientific; 2014.

[51] Breiman L. Random forests. Machine Learning. 2001; 45:5-32.

[52] Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning. 2000; 40:139-57.

[53] Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20(8):832-44.

[54] Hosseinpour M, Ghaemi S, Khanmohammadi S, Daneshvar S. A hybrid high- order type- 2 FCM improved random forest classification method for breast cancer risk assessment. Applied Mathematics and Computation. 2022; 424:127038.

[55] Chaudhary A, Kolhe S, Kamal R. An improved random forest classifier for multi-class classification. Information Processing in Agriculture. 2016; 3(4):215-22.

[56] Kim Y, Kim Y. Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models. Sustainable Cities and Society. 2022; 79:103677.

[57] Egwim CN, Alaka H, Toriola-coker LO, Balogun H, Sunmola F. Applied artificial intelligence for predicting construction projects delay. Machine Learning with Applications. 2021; 6:1-15.

[58] Mensouri D, Azmani A. A new marketing recommendation system using a hybrid approach to generate smart offers. Applied Computer Systems. 2022; 27(2):149-58.

[59] Alkheder S, Taamneh M, Taamneh S. Severity prediction of traffic accident using an artificial neural network. Journal of Forecasting. 2017; 36(1):100-8.

[60] Shaik ME, Islam MM, Hossain QS. A review on neural network techniques for the prediction of road traffic accident severity. Asian Transport Studies. 2021; 7:1-11.

Mensouri Houssam received the engineering diploma in Electronics and Automatic Electronics at the University of Science and Technology, Abdelmalek Essaâdi University, Tangier, Morocco, in 2018. He is a PhD student at the Abdelmalek Essaâdi University, Tangier, Morocco, and a member of the Laboratory of Informatics, System and Telecommunication (LIST). He has contributed to many scientific research projects and articles. His research interests include Logistics, Artificial Intelligence, and Maintenance.
Email: houssam.mensouri@etu.uae.ac.ma

Azmani Abdellah earned his PhD degree in Industrial Computing with a focus on Dynamic System Modelling and Artificial Intelligence from the University of Science and Technology of Lille in 1991. He has held positions as a Professor at the Ecole Centrale de Lille, France, and at the Institute of Computer and Industrial Engineering in Lens, France. Currently, he serves as a Professor at the Faculty of Science and Technology of Tangier, Morocco. Abdellah is a distinguished member of the Laboratory of Informatics, System, and Telecommunication (LIST), and he established and coordinates the Intelligent Automation team. His significant contributions include supervising numerous theses and participating in various scientific research projects. Furthermore, he has developed and implemented several IT and decision-support solutions for public administration, business management, marketing, and logist
Email: a.azmani@uae.ac.ma

Azmani Monir is a highly accomplished expert in Automation, Signal Processing, and Computer Science. He holds a PhD from the University of Littoral Côte d'Opale and has supervised numerous doctoral theses. Currently, he is a professor at the Faculty of Science and Technology of Tangier, Morocco. Dr. Monir is the Head of the Electrical Engineering Department and a member of the Smart Automation and Bio Laboratory. His current focus is on applying artificial intelligence in diverse fields such as Logistics, Maintenance, and Embedded Systems. Dr. Monir's contributions to research and his expertise in AI have earned him great respect in both academia and industry.
Email: m.azmani@uae.ac.ma

Appendix I

S. No.	Abbreviation	Description
1	DT	Decision Tree
2	EDA	Exploratory Data Analysis
3	FN	False Negatives
4	FP	False Positives
5	GPUs	Graphics Processing Units
6	ITS	Intelligent Transportation Systems
7	LR	Logistic Regression
8	NHTSA	National Highway Traffic Safety Administration
9	OECD	Organisation for Economic Co-Operation and Development
10	OSM	Open Street Map
11	POI	Points of Interest
12	RF	Random Forest
13	RMSE	Root Mean Square Error
14	SVM	Support Vector Machine
15	TN	True Negatives
16	TP	True Positives
17	USA	United States of America
18	WHO	World Health Organization