A ROAD ACCIDENT PREDICTION MODEL USING DATA MINING TECHNIQUES

Mr.S. YAKHOOB ALI[1], **A.Sai Sree**[2], **S.Supriya**[3], **S.Thanuja**[4],**C.Naga Lakshmi Sneha**[5], **B.Tharunya**[6]

[1]Associate Professor, Dept of CSE, Gouthami Institute Of Technology and Management for Women, Andhra Pradesh, India

[2,3,4,5,6]U.G Students, Dept of CSE, Gouthami Institute Of Technology and Management for Women, Andhra Pradesh, India

Abstract

Road accidents are a major public health concern, leading to significant loss of life and property worldwide. The ability to predict road accidents can play a crucial role in enhancing traffic safety and implementing effective prevention strategies. This study proposes a comprehensive Road Accident Prediction Model utilizing advanced data mining techniques. By analyzing historical accident data, traffic patterns, weather conditions, and road infrastructure, the model aims to identify key factors contributing to accidents and predict potential accident hotspots.

Our approach integrates various data mining techniques, including classification, clustering, and regression analysis, to uncover hidden patterns and correlations within the data. We employ machine learning algorithms such as Decision Trees, Random Forest, and Support Vector Machines to develop predictive models. The performance of these models is evaluated using metrics such as accuracy, precision, recall, and F1-score.

The results demonstrate that the proposed model can effectively predict road accidents with high accuracy, enabling stakeholders to implement targeted interventions. The insights gained from this study can inform the development of intelligent transportation systems, enhance road safety measures, and ultimately reduce the incidence of road accidents. This research underscores the potential of data mining techniques in transforming raw traffic data into actionable knowledge, contributing to safer and more efficient road networks.

Keywords: Industries, Road traffic accidents, Government, Models that make accident predictions, Algorithm towards forecast, Patterns of data.

Introduction

Various research have looked into various elements of RTAs, with the majority of them focus on predicting or identifying the essential components that influence injury severity. Many data mining-related studies have been conducted to study RTA data locally and globally, with results vary widely based on the socioeconomic factors and technology of a particular region.

Various methods were employed to increase the accuracy of individual classifiers for two RTA intensity categories in order to investigate the association between RTA severity and operating surrounding parameters. Three alternative approaches have been used in neural and logistic regression individual classifiers: classifier fusion using the Participatory method, the Bayesian procedure, and the linear model; data ensemble fusion using sparking and dropping; and clustering using the k-means algorithm. However, it's among the world's biggest incidents, leading in death and physical damage. Identify key cause of road accidents will aid in the development of a suitable solution to reduce the negative impact of severity on people and damage to property. Severity on the road isn't accidental: it follows the pattern that can be foreseen and prevented. In a fraction of a second, human life and property were destroyed. It is one of the country's more frightening leading causes of mortality.

In the last couple of decades, one of the research areas in road safety has been the severity of RTAs. Just on road accident severity categorization based models, researchers used novel methodologies. The research looks at where to develop models using a standard statistical method. These methods aid in gaining insight into and identifying the underlying causes of automobile accidents and other issues that affect road safety. Machine learning now outperforms traditional statistical models in forecasting the model due to the large amount of available data.

There is, however, a scarcity of comparisons between state-of-the-art algorithms, Hybrid Machine Learning algorithms, and deep learning algorithms. Obtaining a suitable technique can make forecast accuracy more informative in some cases. As a result, selecting the best model aids in identifying important road accident elements. Furthermore, target-specific relevant aspects had not been discovered but was not a concern. To anticipate the seriousness of road accidents, the researchers used a combination of clustering and classification methods. Further, the suggested proposed method is compared to a deep learning network in order to compare it to other state-ofthe-art classification techniques. Depending on categorization and performance metrics, the suggested classifier outperforms other classifications in the testing.

The purpose of this research is to determine the most relevant characteristics that affect the degree of injuries suffered by individuals involved in traffic incidents on these roadways, whereas by reducing or managing these factors, overall safety can be improved. They employed the CART method (Classification and Regression Tree).

Literature Survey

For the past few decades, traffic deaths have been the leading cause both injuries and deaths globally. When a road collides with some other vehicle, a person, an animal, or a geographical or physical obstruction, it is called a traffic collision. It also has the potential to cause damage, damage to property, and death. The place where essential data about society is gathered and preserved is the traffic control system. We can identify risk factors for car accidents, injuries, and fatalities using this data, and take precautions that can save lives. The intensity of injuries has societal consequences. Conventional statistical model-based strategies were utilised to forecast accident mortality and severity in the field of road safety. Classical statistical-based research include the mixed logit model based, ordered Probit model, and logit model. According to certain studies, the traditional statistical approach is more effective at detecting direct and indirect accident variables.

Data mining is a new and powerful tool that can help firms focus on most critical information in their database systems by extracting hidden prediction information from large databases. It's an useful tool for dealing with the requirement to move useful data from a database, such as hidden patterns.

Analysis of traffic injury severity: An example of how nonparametric classification tree approaches can be used: The goal of this study is to create the CART model that will be used to find connections among injuries and motorist characteristics, highway /environmental variables, and crash variables. They employed Logistic Regression and Back propagation Systems. **A Data Mining Approach to Identify Key Factory of Traffic Injury Severity:** The purpose of this research is to determine the most relevant factors that influence overall degree of injuries sustained by drivers related to traffic incidents on these roads, so that by deleting or regulating these elements, overall safety can be improved. They employed using CART method (Classification and Regression Tree).

Existing System

Interface requirements- System permission is necessary for users at the start of the service. For all users, the login method is the same. They will provide a login and password that is legitimate or authorised. The user interface summary is detailed in general in the parts below.

Sign in- The user will be provided with a login screen whenever the Accident Analysis and Prediction System web address is opened. If an user has successfully registered in the system, he or she can log in using the username and password; if the user has not yet registered, the user should do so.

Upload data set- To evaluate or predict accidents based on parameters, the user should upload the data into the database server.

Data visualization- The system generates recommendations based on the uploaded data set, processes data, and predicts the outcomes. The proposal will just be displayed in the interface, as well as in the form of graphical visualisation.

Signout- When a people click the sign out button, the system's activities were stopped, and the user is routed to the login page.

Clustering Techniques:

A process of collecting data elements could be treated as a single entity. When undertaking clustering algorithm, we divide the data set into groups based on data similarity, and assign labels to the groups. Clustering has the advantage of being adaptive to changes & assisting in the identification of useful qualities that separate groups. Traffic accident data is now being collected in huge volumes. Multi - processor systems with a bunch of processing capacity. Various research have examined into various aspects of RTAs, with the majority of them focusing on anticipating or identifying the essential components that influence injury severity. Several database miningrelated research have been conducted to evaluate RTA data locally and internationally, with results differing widely based on the economic status and technology of a given place.

Proposed System

1. Data Collection

Accident-related datasets are collected from reliable sources such as government transportation departments, traffic surveillance systems, weather databases, and open data repositories. The data includes various attributes such as:

- Date and time of the accident
- Geographical location (latitude and longitude)
- Road and traffic conditions
- Weather conditions
- Vehicle type and speed
- Number of casualties and fatalities
-

2. Data Preprocessing

The raw data is often incomplete, noisy, and inconsistent. Therefore, preprocessing is a crucial step and includes:

- Handling missing or inconsistent data
- Removing duplicate or irrelevant entries
- Converting categorical variables into numerical form (label encoding or one-hot encoding)
- Normalizing numerical data for uniformity
- Temporal or spatial aggregation if necessary

3. Feature Selection and Transformation

Feature selection techniques such as correlation analysis, Chi-square tests, and feature importance ranking (e.g., using Random Forest or Information Gain) are applied to identify the most influential variables that contribute to road accidents. Dimensionality reduction techniques like PCA may be used to improve model efficiency.

4. Application of Data Mining Techniques

Various classification and prediction algorithms are employed to build the predictive model. The choice of algorithms includes:

- **Decision Trees**: For rule-based classification
- **Random Forest**: For ensemble-based prediction
- **Naïve Bayes**: For probabilistic classification
- **K-Nearest Neighbors (KNN)**: For instance-based learning
- **Support Vector Machines (SVM)**: For high-dimensional classification
- **Artificial Neural Networks (ANN)** (optional): For capturing non-linear relationships

Each model is trained on historical data to predict the likelihood of an accident occurring under specific conditions.

5. Model Evaluation

The performance of the models is evaluated using standard metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

Cross-validation techniques such as k-fold cross-validation are used to validate model performance and ensure generalizability.

6. Visualization and Interpretation

Visual tools like heatmaps, bar charts, and GIS-based mapping are used to interpret the prediction results and identify high-risk accident zones. This information can assist traffic authorities in decision-making and planning preventive measures.

UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors their goals.

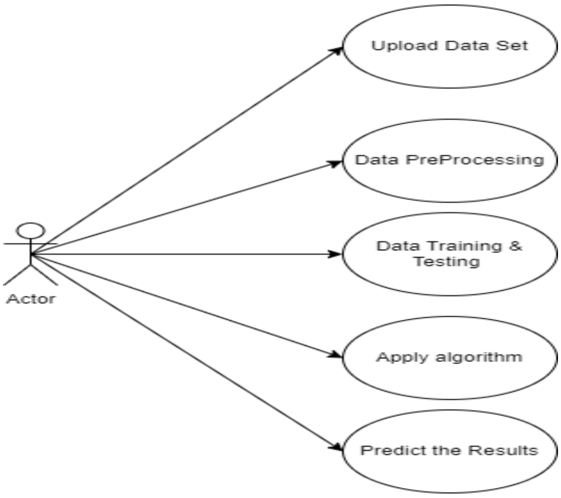


Fig-1: Use Case Diagram

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

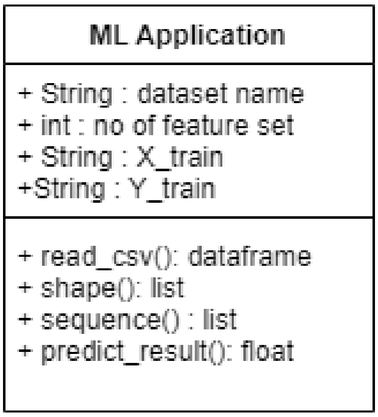


Fig-2:Class Diagram

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

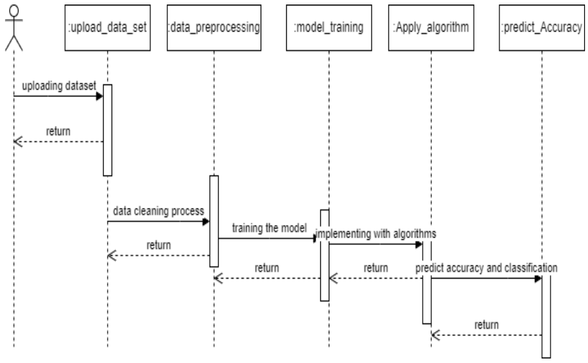


Fig-3: Sequence diagram

<u>Module Description:</u>

- **Gathering Data**
- **Data preparation**
- **Data Wrangling**
- **Analyse Data**
- **Train the model**
- **Test the model**

Gathering data:

Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.

In this step, we need to identify the different data sources, as data can be collected from various sources such as **files**, **database**, **internet**. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

Data preparation:

After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.

In this step, first, we put all data together, and then randomize the ordering of data.

This step can be further divided into two processes:

- **Data exploration:**

  It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data.
  A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.

- **Data pre-processing:**
  Now the next step is pre-processing of data for its analysis.

Data Wrangling:

Data wrangling is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues.

It is not necessary that data we have collected is always of our use as some of the data may not be useful. In real-world applications, collected data may have various issues, including:

- **Missing Values**
- **Duplicate data**
- **Invalid data**
- **Noise**

Data Analysis:

Now the cleaned and prepared data is passed on to the analysis step. This step involves:

- **Selection of analytical techniques**
- **Building models**
- **Review the result**

The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome. It starts with the determination of the type of the problems, where we select the machine learning techniques such as **Classification**, **Regression**, **Cluster analysis**, **Association**, etc. then build the model using prepared data, and evaluate the model.

Train Model:

Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.

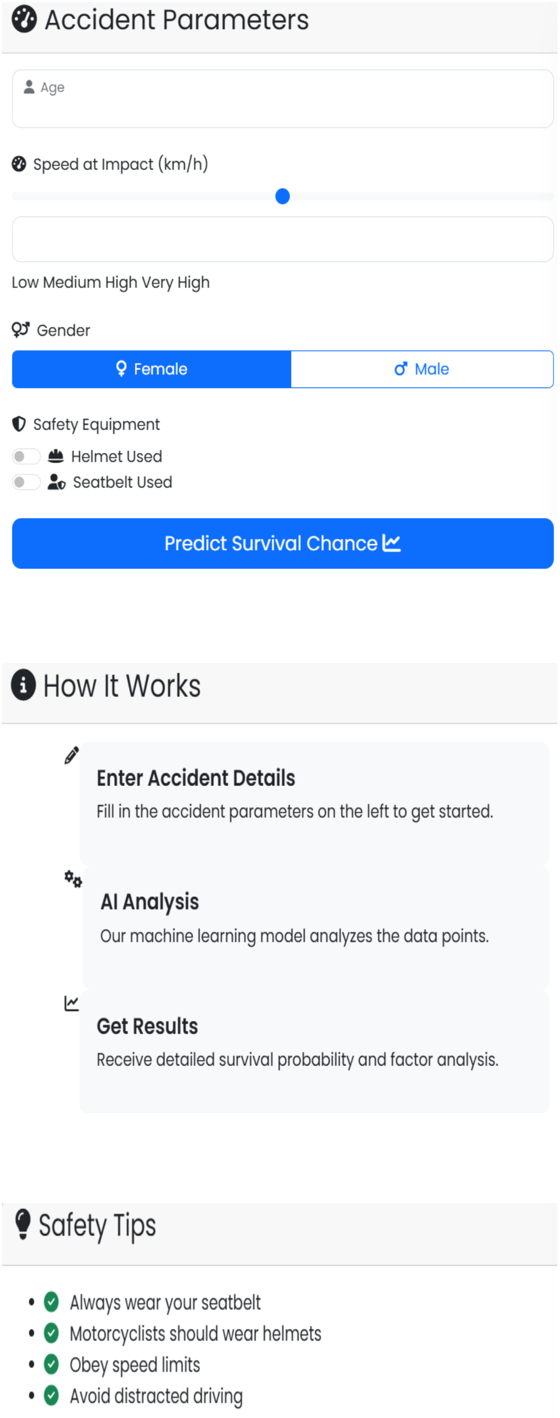
We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features

Test Model:

Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it.

Testing the model determines the percentage accuracy of the model as per the requirement of project or problem

OUTPUT



Accident Parameters

Age

Speed at Impact (km/h)

Low Medium High Very High

Gender

Female | Male

Safety Equipment

○ Helmet Used
○ Seatbelt Used

Predict Survival Chance

How It Works

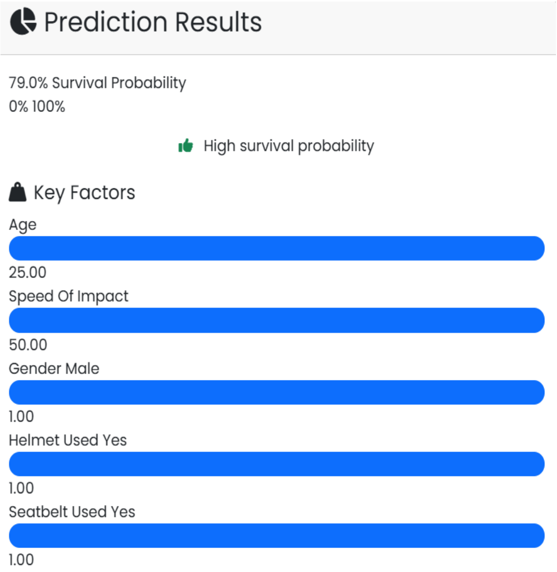
Enter Accident Details
Fill in the accident parameters on the left to get started.

AI Analysis
Our machine learning model analyzes the data points.

Get Results
Receive detailed survival probability and factor analysis.

Safety Tips

- ✓ Always wear your seatbelt
- ✓ Motorcyclists should wear helmets
- ✓ Obey speed limits
- ✓ Avoid distracted driving



Prediction Results

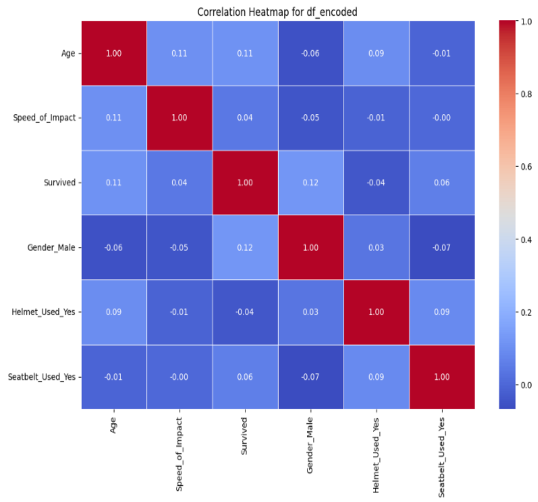
79.0% Survival Probability
0% 100%

High survival probability

Key Factors

Age
25.00
Speed Of Impact
50.00
Gender Male
1.00
Helmet Used Yes
1.00
Seatbelt Used Yes
1.00

Impact Analysis



Applications

1. Traffic Management and Planning

- Authorities can identify accident-prone zones (black spots) and take preventive measures such as installing warning signs, speed breakers, or traffic signals.
- Helps in optimizing traffic flow by understanding accident patterns during specific times or conditions.

2. Road Safety Improvement

- Enables proactive safety planning by predicting potential accident occurrences.
- Supports the design of safer road infrastructure based on data-driven insights.

3. Emergency Response Optimization

- Prediction of high-risk times and locations allows emergency services to position ambulances and rescue teams more efficiently.
- Reduces response time during accidents, potentially saving lives.

4. Driver Assistance Systems

- Can be integrated into navigation or vehicle systems to alert drivers about high-risk zones ahead.
- Enhances real-time decision-making and driving behavior, especially under adverse weather or road conditions.

5. Urban Development and Smart Cities

- Contributes to building intelligent transportation systems (ITS) in smart cities.
- Assists city planners in making data-informed decisions for safer road networks.

6. Insurance Risk Assessment

- Helps insurance companies assess the accident risk associated with certain routes, locations, or driving conditions.
- Can be used to customize insurance premiums based on predictive analytics.

7. Public Awareness Campaigns

- Insights from the model can be used to educate the public about dangerous driving behaviors or high-risk times (e.g., weekends, late nights).

- Promotes safer driving habits through targeted awareness initiatives.

Advantages

- Accurate Prediction: Data mining techniques can effectively identify complex patterns and predict accident-prone scenarios.
- Early Risk Detection: Helps in identifying high-risk zones, weather conditions, and time periods before accidents happen.
- Cost-Effective: Reduces the economic impact of road accidents by enabling preventive measures.
- Data-Driven Decisions: Supports government agencies and traffic departments with actionable insights.
- Automation: Can be integrated into real-time systems for continuous monitoring and alerts.

Disadvantages

- Data Quality Issues: Inaccurate, incomplete, or outdated data can negatively affect prediction accuracy.
- Limited Generalization: A model trained on data from one region may not perform well in another due to different traffic and road conditions.
- High Computational Cost: Some advanced models (like neural networks) require significant computational resources and time.
- Privacy Concerns: Collecting detailed data (e.g., from GPS or surveillance) may raise privacy issues.
- Real-Time Limitations: Implementing real-time prediction and alert systems can be technically complex.

Conclusion

The increasing number of road accidents poses a serious threat to public safety, economic stability, and infrastructure development. This project demonstrates how data mining techniques can be effectively utilized to analyze complex accident data and develop predictive models that forecast accident likelihood based on key factors such as location, weather, time, and traffic conditions.

By employing algorithms like Decision Trees, Random Forest, Naïve Bayes, and SVM, the system can identify hidden patterns and correlations within large datasets that are otherwise difficult to detect through traditional analysis methods. These insights can support proactive decision-making by traffic authorities, emergency responders, city planners, and insurance companies.

Moreover, the application of such models enables the development of real-time alert systems and the identification of high-risk zones (black spots), which can lead to timely interventions and the prevention of accidents. Integration with smart city infrastructure can also enhance urban mobility and ensure safer transportation networks.

However, the success of such models largely depends on the availability of high-quality and up-to-date data. Regional variations, privacy concerns, and computational limitations also present challenges that need to be addressed in future research.

In conclusion, road accident prediction using data mining techniques is a promising approach toward reducing accidents and saving lives. With ongoing advancements in machine learning, data availability, and IoT integration, such predictive systems will become increasingly accurate and accessible, making our roads safer for everyone.

References

Here are some sample references you can use (replace or expand as needed based on your actual sources):

1. Ministry of Road Transport and Highways (MoRTH). (2023). *Road Accidents in India*. https://morth.nic.in
2. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
3. Kumar, P., & Toshniwal, D. (2015). A data mining framework to analyze road accident data. *Journal of Big Data*, 2(1), 1-26.
4. Yadav, D., & Mishra, S. (2019). Road Accident Prediction using Machine Learning Algorithms. *International Journal of Computer Applications*, 178(42), 15-20.
5. U.S. National Highway Traffic Safety Administration (NHTSA). *Traffic Safety Facts*. https://www.nhtsa.gov