WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Road Accident Severity Prediction Using Machine Learning

Dr. N. Kalaivani[1], Mr. R. Arun Prasath[2], Mr. M. Chandru[3]

[1]Assistant professor, Department of Information Technology, Sri Krishna Adithya College of Arts & Science, Coimbatore.
[2,3] III B.Sc. IT, Department of Information Technology, Sri Krishna Adithya College of Arts & Science, Coimbatore

ABSTRACT

Road traffic accidents are a major public safety concern worldwide, with significant social, economic, and environmental impacts. Predicting the severity of road accidents can help in the formulation of effective traffic management strategies, the design of safety measures, and the allocation of resources for emergency response. This study aims to develop a machine learning-based model to predict the severity of road accidents. Various factors such as weather conditions, road types, time of day, vehicle type, driver characteristics, and accident-specific details are considered as input features for the prediction model. A dataset containing historical accident data is used to train and evaluate multiple machine learning algorithms, including decision trees, random forests, support vector machines, and neural networks. The performance of these models is assessed based on metrics such as accuracy, precision, recall, and F1-score. The results indicate that machine learning models can effectively predict the severity of accidents, providing valuable insights for improving road safety measures and enhancing public awareness. Future work will focus on refining the model by incorporating real-time data and exploring additional algorithms to further increase prediction accuracy. By leveraging a dataset of historical accident records, several machine learning models, including logistic regression, random forests, and support vector machines, are trained to classify accident severity into categories such as minor, moderate, and severe. Model performance is evaluated using metrics like accuracy, AUC (Area Under Curve), and confusion matrix, demonstrating that machine learning can effectively identify high-risk accident scenarios. The study highlights the potential for these predictive models to inform traffic safety interventions and reduce the impact of road accidents. Future research will focus on integrating real-time data and expanding the model to account for more complex interactions between variables. XGBoost, a powerful gradient boosting algorithm, was applied to predict the severity of accidents by combining the predictions of multiple weak models to create a strong, accurate model. Its ability to handle large datasets, its robustness to overfitting, and its ability to deal with missing values and non-linear relationships made it an ideal choice. SVM, on the other hand, is a robust classifier that works well in high-dimensional spaces, making it effective for handling the complexity of accident severity prediction.

Keywords: XGBoost, SVM

I. INTRODUCTION

Road traffic accidents are a leading cause of death and injury globally, with devastating consequences for public health, infrastructure, and economic development. The World Health Organization (WHO) estimates that over 1.3 million people lose their lives every year in road traffic accidents, and tens of millions more suffer injuries, many of which are life-altering. The severity of these accidents can range from minor fender benders to catastrophic events resulting in fatalities or significant injuries. Accurately predicting the severity of road accidents is essential for designing proactive road safety measures, improving traffic management, optimizing emergency responses, and allocating resources more effectively.

Traditional methods for predicting accident severity generally rely on statistical models, which often struggle to handle large and complex datasets. These methods typically focus on limited sets of variables such as weather conditions, road type, and traffic volume, which might not capture the full range of factors that influence accident outcomes. Moreover, expert-based judgment often lacks the capability to identify subtle patterns and interactions between these factors, leading to less reliable predictions.

In recent years, machine learning (ML) techniques have demonstrated great potential in addressing these challenges. Machine learning, particularly supervised learning methods, can analyze large datasets to identify non-linear relationships and complex patterns in the data. By utilizing historical accident data, machine learning models can be trained to predict the severity of an accident based on various features, including but not limited to weather conditions, road attributes (e.g., road type, surface condition), time of day, traffic density, driver demographics, and accident-specific variables such as collision type and vehicle involvement.

Several machine learning algorithms, such as decision trees, random forests, support vector machines (SVM), and neural networks, are increasingly being explored for accident severity prediction. These models can not only enhance the accuracy of severity predictions but also provide more scalable and adaptive solutions compared to traditional methods. With the integration of these advanced predictive models, transportation authorities can take timely and informed actions to mitigate the impact of high-risk accidents.

The primary goal of this study is to develop and evaluate machine learning models to predict road accident severity, classifying accidents into categories such as minor, moderate, and severe. We aim to explore how different algorithms perform in terms of prediction accuracy, precision, recall, and F1-score, using a dataset that includes a variety of factors influencing accident outcomes. Furthermore, we aim to provide insights into the key determinants that contribute to the severity of accidents and identify patterns that can aid in the design of better safety policies.

This research contributes to the growing field of road safety and transportation studies by demonstrating how machine learning can be applied to accident severity prediction, which is crucial for improving traffic safety, reducing fatalities, and enhancing the efficiency of emergency response systems. The findings from this study could be a valuable tool for road safety authorities, city planners, and emergency services in reducing the risk of severe road accidents.

II.MATERIAL AND METHOD

Dataset and Datamining Process:

The dataset used in this study consists of historical road accident records, which contain detailed information about accidents, including factors that are believed to influence accident severity. The dataset is sourced from publicly available traffic accident data, provided by government agencies, law enforcement, or transportation departments. In this study, we have used a dataset that includes the following key features:

- Accident ID: A unique identifier for each accident.
- Severity: The target variable, which represents the severity of the accident (e.g., minor, moderate, severe, or fatal).
- Weather Conditions: Information regarding the weather during the accident (e.g., clear, rain, fog, snow).
- Road Conditions: Type of road (e.g., urban, rural), surface condition (e.g., dry, wet, icy).
- Time of Day: The time the accident occurred, categorized into intervals (e.g., morning, afternoon, evening, night).
- Day of Week: The day on which the accident occurred (e.g., Monday, Tuesday, etc.).
- Traffic Volume: The traffic conditions at the time of the accident (e.g., light, moderate, heavy).
- Vehicle Types: The types of vehicles involved in the accident (e.g., car, truck, motorcycle).
- Driver Demographics: Age and gender of the drivers involved in the accident.
- Location: Geographical information related to the accident's location (e.g., intersection, highway, residential area).
- Collision Type: Type of collision (e.g., rear-end, head-on, side-impact).
- Accident Time: Specific time of the day the accident occurred, providing insight into the temporal distribution of accidents.

The dataset contains thousands of records of accidents over multiple years, allowing for a robust analysis of the various factors contributing to road accident severity.

Data Preprocessing and Cleaning

Before applying machine learning algorithms, the dataset undergoes several preprocessing steps to ensure the data is clean, consistent, and ready for analysis:

1. Missing Data Handling: Incomplete records, such as missing values in key columns (e.g., missing weather conditions or accident time), are addressed. Missing data can be handled by imputing with mean/median values, using interpolation, or, if a significant portion of the data is missing for a particular variable, dropping the variable entirely.
2. Data Transformation: Categorical variables, such as weather conditions, road types, and vehicle types, are encoded using techniques like one-hot encoding or label encoding to convert them into numerical format, making them suitable for machine learning models.
3. Outlier Detection: Outliers that deviate significantly from the rest of the data are identified and addressed, as they can skew model performance. For instance, extremely high traffic volumes or unexpected road conditions may be treated as outliers and either removed or adjusted.
4. Normalization/Standardization: Continuous numerical features (e.g., vehicle speeds, traffic volumes) are normalized or standardized to a common scale, ensuring that no particular feature disproportionately affects the model's learning process.
5. Feature Selection: Irrelevant or highly correlated features are removed to improve the model's efficiency and reduce the risk of overfitting. Techniques such as correlation matrices, Recursive Feature Elimination (RFE), and Random Forest feature importance are used to select the most significant variables.

All machine learning algorithms are implemented using Python and libraries such as **scikit-learn**, **TensorFlow**, and **XGBoost**. Data preprocessing and feature engineering are carried out using libraries like **pandas** and **NumPy**, while **matplotlib** and **seaborn** are employed for data visualization.

Classification Algorithms:

Logistic regression (LR):

Logistic Regression (LR) is a widely used statistical method for binary and multi-class classification problems. Despite its name, it is a linear model for classification, not regression, that predicts the probability of an instance belonging to a particular class based on one or more input features.

Overview of Logistic Regression

In the context of road accident severity prediction, the goal of Logistic Regression is to model the relationship between a set of independent features (such as weather conditions, road type, vehicle type, etc.) and the target variable (accident severity). Logistic Regression outputs a probability value that indicates the likelihood of an accident falling into one of the severity categories, such as "minor," "moderate," or "severe."

Logistic Regression is particularly suited for predicting categorical outcomes that are mutually exclusive and typically binary (e.g., severe vs. non-severe accidents). However, it can be extended to multi-class classification problems using techniques such as **One-vs-Rest (OvR)** or **Softmax Regression**.

Random forests (RF):

Random Forest (RF) is a powerful ensemble learning method that is widely used for classification tasks, including road accident severity prediction. It builds upon decision trees by creating a collection (or "forest") of decision trees and combining their predictions to make more accurate and robust predictions. This method is especially useful for handling complex, high-dimensional datasets with multiple features and intricate relationships between them.

XGBoost:

XGBoost (Extreme Gradient Boosting) is one of the most popular and powerful machine learning algorithms, known for its efficiency, performance, and scalability. It is a type of gradient boosting algorithm, which is an ensemble learning technique that combines the predictions of multiple base models (typically decision trees) to improve predictive accuracy.In road accident severity prediction, XGBoost can be used to predict the severity of accidents (e.g., minor, moderate, severe) based on various factors like weather conditions, road type, traffic volume, time of day, and more. XGBoost is particularly effective in handling large and complex datasets and can model intricate relationships between features, making it well-suited for this problem

4. K-Nearest neighbors (KNN):

Support vector machine (SVM):

K-Nearest Neighbors (KNN) is a simple yet powerful machine learning algorithm used for both classification and regression tasks. In the context of road accident severity prediction, KNN can be used to classify the severity of an accident (e.g., minor, moderate, severe) based on various features like weather conditions, time of day, road type, vehicle type, traffic volume, etc.The KNN algorithm works by comparing a given data point (i.e., an accident record) to its nearest neighbors in the feature space and assigning the most frequent class (severity level) among the neighbors as the prediction. K-Nearest Neighbors (KNN) is a simple yet effective algorithm for predicting road accident severity, particularly when the relationships between features and outcomes are not overly complex. While KNN offers simplicity and interpretability, it does have limitations such as high computational cost and sensitivity to feature scaling. Therefore, KNN may be best suited for smaller datasets or cases where its simplicity and ease of understanding are priorities.

Deep learning:

Deep Learning (DL) is a subset of machine learning that involves the use of neural networks with many layers (hence the term "deep"). Deep learning has gained significant popularity in recent years due to its ability to automatically learn complex patterns and representations from large datasets. In road accident severity prediction, deep learning can be used to predict the severity of accidents (e.g., minor, moderate, severe) based on a variety of input features such as weather conditions, road type, traffic volume, time of day, and more.Deep learning models, especially Artificial Neural Networks (ANNs), are capable of capturing complex, non-linear relationships between features and outcomes, making them a powerful tool for predicting road accident severity, especially when dealing with large datasets with many features.

III.RESULTS

The results of a machine learning model for road accident severity prediction are crucial for evaluating the effectiveness of the chosen algorithm and understanding its performance. The results typically include model evaluation metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curves (for classification tasks). Additionally, the feature importance and the model's ability to generalize on unseen data are essential considerations.

In this section, we'll discuss the key results you might obtain when applying machine learning models (such as Logistic Regression, Random Forest, XGBoost, KNN, or Deep Learning) to predict road accident severity and explain how to interpret them.

IV.DISCUSSION AND CONCLUSION

The goal of predicting road accident severity using machine learning is to develop a system that can automatically assess the severity of accidents based on various factors such as traffic conditions, weather, time of day, and road type. By leveraging machine learning algorithms, we aim to improve road safety, optimize emergency response systems, and contribute to effective traffic management strategies.

In this study, various machine learning models were explored and evaluated for their ability to predict accident severity. These models included traditional algorithms like Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and XGBoost, as well as deep learning approaches such as Multilayer Perceptrons (MLP). The results from these models were compared in terms of classification accuracy, precision, recall, F1-score, and other relevant metrics.

The prediction of road accident severity using machine learning holds significant potential for improving road safety, optimizing emergency responses, and supporting traffic management policies. Among the various models evaluated in this study, XGBoos**t** and Random Forest demonstrated the best performance, offering a good balance between accuracy and interpretability. Deep learning models, particularly Multilayer Perceptrons (MLPs**)**, provided superior accuracy but required larger datasets and higher computational resources, making them less practical for every use case.

Key insights drawn from this study include the high importance of factors such as weather conditions, traffic volume, road type, and time of day in determining the severity of accidents. These findings can aid policymakers and traffic safety experts in prioritizing safety measures in high-risk conditions, such as during bad weather or at night.