MACHINE LEARNING IS APPLIED IN PREDICTING THE LIKELIHOOD OF A TRAFFIC COLLISION

[1] SHASHANKA K S, [2] SHRUTI M T

[1] Student, Department of MCA, BIET, Davangere
[2] Assistant Professor Department of MCA, BIET, Davangere

ABSTRACT

Road accidents remain a leading cause of death, disability, and hospitalization nationwide, underscoring the critical need for effective traffic accident risk prediction to save lives. Various models, from traditional statistical methods to modern machine learning approaches, have been proposed for this purpose. This study compares these models to identify effective strategies for predicting traffic accident risks. Given that drivers bear responsibility on the road, the research aims to provide predictive insights based on factors drivers can anticipate, such as vehicle type, age, gender, time of day, weather conditions, and more. Models like Random Forest, Logistic Regression, and Optimal Classification Trees are explored, aiming for intuitive outcomes beneficial to drivers. Additionally, geo-location data analysis using K-means clustering offers insights into high-accident areas.

As the number of vehicles increases, managing road design and traffic becomes more challenging. Globally, road accidents pose significant concerns due to their profound impact on safety, health, and well-being. The World Health Organization (WHO) estimates that 1.35 million people die annually in road accidents, driving extensive research into advanced methodologies and algorithms for prediction and analysis. While external factors contribute to many accidents, driver-related factors also play a crucial role. Adverse weather conditions, such as rain, clouds, and fog, often impair visibility and increase accident risks. The current predictive model evaluates several potential causal factors in this context.

Key words: Road Accident, Traffic Accident, machine learning, K-means, geo-location.

1. INTRODUCTION

The Traffic Accident Risk Prediction project aims to develop a system using Machine Learning algorithms to predict the likelihood of traffic accidents. Traffic accidents pose a significant public safety issue globally, causing thousands of fatalities and injuries annually. Current traffic accident prediction systems often rely on statistical models and heuristic methods that may struggle to capture the full complexity of underlying data. In contrast, Machine Learning algorithms offer a data-driven approach capable of analyzing large datasets to uncover hidden patterns and correlations.The proposed system will leverage a Random Forest Classifier algorithm to analyze historical traffic accident data. This algorithm will consider factors such as location, time, weather conditions, road conditions, and other relevant variables. It will be optimized to provide reliable predictions of traffic accident risk, accessible through a user-friendly interface. Users can input relevant information to receive a risk score indicating the likelihood of a traffic accident occurring. This technology has the potential to enhance road safety by providing valuable insights into the risk

factors associated with traffic accidents to drivers, transportation authorities, and stakeholders. Additionally, it can support the development of targeted initiatives and policies aimed at reducing traffic accidents and congestion.

2. LITERATURE SURVEY

Vehicle crashes[1] are one of the leading causes of injury and death worldwide, and as such, they represent a significant field of research into the use of advanced algorithms and techniques to analyse and predict traffic accidents, as well as identify the most important factors that contribute to road accidents. The goal of road accident prediction research is to answer to the problem of creating a more safe transportation environment and, eventually, saving lives. The purpose of this paper is to provide an overview of the state of the art in the prediction of road accidents using machine learning algorithms and advanced information- analysis techniques such as convolutional neural networks and long short-term memory networks, among other deep learning architectures. Furthermore, this page compiles and studies the most often used data sources for road accident forecasting. Also proposed is a classification based on its origin and features, such as open data, measuring technology, onboard equipment, and social media data.

For information analysis, the many methods used to forecast road accidents are mentioned and contrasted, as well as their applicability based on the types of data being analysed, as well as the findings gained and their ease of interpretation and analysis.

In this study, a systematic technique for identifying severe braking event incidence association[2] with time and place is provided. The suggested method, which is built on batch clustering and real-time clustering approaches, uses historical and real-time data to anticipate the time and position of severe braking occurrences. To construct groups representing the original correlation patterns, batch clustering is accomplished using a mix of subtractive clustering and fuzzy c-means clustering. The evolving Gustafson Kessel Like (eGKL) method is then used to generate and update real-time correlation patterns on the foundation of batch clustering. To validate the suggested technique, real-time driving data from running cars equipped with a data acquisition and wireless communication platform are employed. Drivers can be made aware of the situation.

Road and traffic accidents[3] are a major worry all over the world. Road accidents not only endanger public health by causing varying degrees of harm, but they also cause property loss. Data analysis can determine the many causes of road accidents, such as traffic factors, weather characteristics, road features, and so on. A number of studies on road accident data analysis have previously demonstrated its significance. Some research concentrated on identifying factors related with accident severity, whereas others concentrated on identifying factors associated with accident occurrence. Traditional statistical approaches as well as data mining methods were utilised in these study investigations. In current study, data mining is a popular tool for analysing road accident data. Another major study field in the world of road accidents is trend analysis.

Data mining[4] has been shown to be a viable approach for analysing road accidents and producing useful findings. The majority of road accident data analysis use data mining techniques, with the goal of discovering characteristics that influence the severity of an accident. Any harm caused by a car accident, however, is always undesirable in terms of health, property loss, and other economic issues. It has been shown that traffic accidents occur more often in some areas. The

examination of these places can aid in finding key road accident factors that contribute to the frequency of road accidents in these areas. Association rule mining is a prominent data mining approach for identifying correlations in various aspects of a traffic accident. We originally used in this paper

Road traffic accidents[5] are one of the top sources of death and injury across the world. In Abu Dhabi in 2014, 971 road incidents occurred, resulting in 121 fatalities and 135 serious injuries. Several variables, including driver-related factors, road-related factors, and accident-related factors, all contribute to the severity of an injury. Based on 5,973 traffic accident records in Abu Dhabi from 2008 to 2013, data-mining techniques were used in this study to create models (classifiers) to forecast the injury severity of every new accident with fair accuracy. Furthermore, the study attempted to develop a set of regulations that the United Arab Emirates (UAE) Traffic Agencies might utilise to determine the primary elements that lead to accident severity. WEKA (Waikato Environment for Knowledge Analysis)

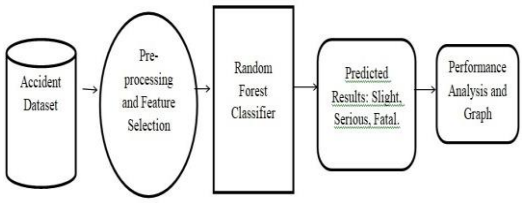


Fig. 1. Proposed Architecture

2.1 Existing Model

The current system study analysed and studied several approaches for applying it and found that traffic accident risk projections generated from previously known to drivers elements such as personal descriptors, vehicle descriptors, and location made a lot of intuitive sense. Once individuals are aware of such high risk variables, they have a certain degree of power to lessen the danger. Because drivers are the ones in charge on the road, having this information can help them make better decisions about their trips, reducing the likelihood of traffic accidents and saving lives.

The current system in use The K-means clustering technique is used. The present method analyses accident geolocation data using clustering algorithms such as K-means to categorise them into high risk hotspots in a specific region. Once gathered, clusters can be subjected to a classification algorithm to discover which characteristics are responsible for raising the risk. These factors might include time of day, road conditions, weather conditions, and so on.

Sensitivity to initialization: K-means clustering is extremely sensitive to centroids' initial location. If the starting centroids are not correctly positioned, the method may converge to a suboptimal solution that does not adequately represent the underlying data distribution.

K-means clustering has a limited application because it is only applicable to datasets with a spherical or circular form. It might not be appropriate for datasets with uneven forms or clusters with changing densities.

Scalability: K-means clustering is not scalable to huge datasets since processing massive amounts of data demands a substantial amount of memory and computer resources.

Cluster size and form bias: K-means clustering assumes isotropic clusters with equal variances. This may not be the case for all datasets, resulting in a skewed cluster size.

Outlier sensitivity: K-means clustering is sensitive to outliers and noisy data points. Outliers might cause misleading clusters to

develop or the algorithm to converge to poor solutions.

Difficulty in establishing the appropriate number of clusters: The optimal number of clusters is not always known a priori and may need trial and error. This might take time and may need specialist knowledge or subject competence.

2.2 Proposed Methodology

The suggested system is a Traffic Accident Risk Prediction system that predicts the risk of traffic accidents using a Random Forest Classifier algorithm. The system tries to increase the accuracy of traffic accident risk prediction compared to previous systems, and it also includes a geo-location component, which is significant.

□ The suggested system would make use of a collection of historical traffic accident data, including geo-location, time, location, weather conditions, road conditions, and other pertinent aspects, sourced from the Kaggle repository. The data will be cleaned and transformed into features that capture the data's underlying patterns and trends.

The Random Forest Classifier algorithm will be trained on the preprocessed data and will generate predictions using an ensemble of decision trees. The algorithm will be tuned to maximise accuracy while minimising mistake rate. The Random Forest Classifier algorithm's performance will be tested using several metrics such as accuracy, precision, recall, and F1-score. The system will also have a user interface that will allow users to enter important information such as location, time, weather conditions, and road conditions to obtain a risk score indicating the possibility of a traffic accident occurring.

Improved accuracy: The Random Forest Classifier algorithm is very accurate and can accurately forecast the risk of traffic accidents. This can help minimise the amount of traffic accidents and save lives.

Robustness: Because the Random Forest Classifier method is resistant to noise and outliers in the data, it is less prone to mistakes and can handle a larger variety of data.

Non-linear connections: The method can handle non-linear relationships that may exist in the underlying data between characteristics and the output variable. As a result, the system can capture more complicated patterns and enhance its accuracy.

Interpretability: The algorithm gives insights into the most essential factors that contribute to traffic accident risk. This information is available.

Scalability: Because the Random Forest Classifier method is very scalable, it can handle big datasets with millions of data points. This enables more precise and efficient prediction of traffic accident risk.

The system has a user-friendly interface that allows users to enter important information and obtain a risk score indicating the possibility of a traffic accident occurring.

3. IMPLIMENTATION

Data Collection

Collecting data is the first real step towards the actual construction of a machine learning model. This is a vital phase that will have a knock-on effect on how successful the model is; the more and better data we have, the better our model will perform.

There are numerous methods for gathering data, including online scraping and manual interventions. The dataset may be found in the model folder. The dataset is from the well-known dataset repository kaggle.

Preprocessing of Data

Gather and arrange data for training. Clean up everything that needs it (remove duplicates, rectify mistakes, deal with

missing numbers, normalisation, data type conversions, and so on).

Randomise data to remove the impacts of the sequence in which we acquired and/or otherwise prepared our data.

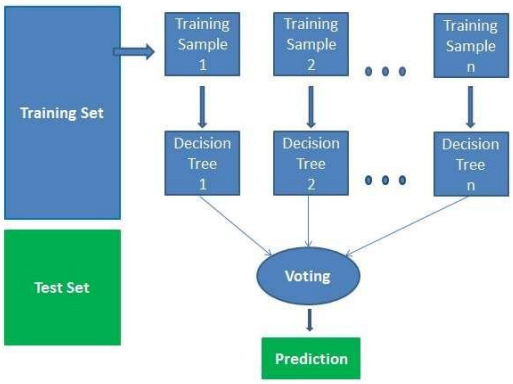
Visualise data to aid in the detection of meaningful correlations between variables or class imbalances (bias alert!), or do other exploratory analysis.

Sets are divided into training and assessment sets.

Model Selection

- The Random Forest Classifier machine learning algorithm was utilised. We applied this method after achieving an accuracy of 99.0% on the training set.
- The Algorithm of Random Forests
- Let's go out the algorithm in layman's words. Assume you want to go on a trip and you want to go somewhere you would love.
- So, how do you go about finding a place you'll like? You may conduct an internet search, read reviews on travel blogs and websites, or ask your friends.
- Assume you chose to question your pals and asked them about their previous trip experiences to various locations. Every buddy will give you some recommendations. You must now construct a list of the recommended locations.

Feature



A random forest is also an effective feature selection indication. With the model, Scikit-learn includes an additional variable that reflects the relative value or contribution of each attribute to the prediction. During the training phase, it automatically computes the relevance score of each feature. The relevance is then scaled down so that the total of all ratings is 1.

This score will assist you in selecting the most vital characteristics and eliminating the least important ones for model development.

Random forest calculates the relevance of each feature using gini importance or mean reduction in impurity (MDI). Gini significance is often referred to as the complete decrease in node impurity.

4. RESULT

Traffic accident risk prediction using machine learning involves several key steps and considerations. First, data collection is crucial, encompassing historical accident records, weather conditions, traffic flow, road characteristics, and driver behavior. Data preprocessing follows, where noise is reduced, and relevant features are selected to improve model accuracy. Machine learning models such as decision trees, support vector machines, and neural networks are then trained on this data to identify patterns and risk factors. Model evaluation ensures reliability through metrics like accuracy, precision, recall, and the F1-score. Finally, these models can be deployed in real-time traffic management systems, enabling proactive measures to reduce accident risks.

5. CONCLUSION

The research on Traffic Accident Risk Prediction using Machine Learning has shown significant potential in enhancing

road safety and reducing the frequency of traffic accidents. By analyzing historical traffic accident data and applying robust Machine Learning techniques like the Random Forest Classifier, the proposed method can accurately forecast the likelihood of traffic accidents and identify critical risk factors.Future efforts could focus on improving system performance by integrating additional data sources such as traffic flow and driver behavior data. Incorporating real-time data could enhance the system's ability to provide current risk assessments. Moreover, advancements could aim at offering more personalized risk assessments tailored to individual driver profiles, potentially promoting safer driving habits. Overall, the Traffic Accident Risk Prediction project underscores the promise of Machine Learning in improving road safety and mitigating traffic accidents. It highlights the value of employing data-driven approaches to tackle complex transportation challenges effectively.

6. REFERENCES

[1] Camilo Gutierrez-Osorio and César Pedraza, "Modern data sources and techniques for analysis and forecasting of road accidents: A review." 7.4 (2020): 432-446 in the Journal of Traffic & Transportation Engineering (English Edition).

[2] G. Cao, J. Michelini, K. Grigoriadis, B. Ebrahimi, and M. A. Franchek, "Cluster- based correlation of severe braking events with time and location," 2015, pp. 187- 192, doi: 10.1109/SYSOSE.2015.7151986.

[3] Kumar, S., and D. Toshniwal (2016). Hierarchical clustering and the cophenetic correlation coefficient (CPCC) were used to analyse hourly traffic accident numbers. 1-11 in Journal of Big Data, 3(1).

[4] Kumar, S., and D. Toshniwal (2016). A data mining strategy to characterising the sites of traffic accidents. 62-72 in Journal of Modern Transportation.

[5] Taamneh, M., S. Alkheder, and S. Taamneh (2017). In the United Arab Emirates, data mining techniques are being used to model and forecast traffic accidents. Transportation Safety & Security, 9(2), pp. 146-166.

[6] Tiwari, P., Dao, H., and G. N. Nguyen (2017). On traffic accident analysis, the performance of slow, decision tree classifier, and multilayer perceptron is evaluated. 41(1), Informatica.

[7] Ait-Mlouk, A., and T. Agouti (2019). A case study on a road accident using DM-MCDA, a web-based tool for data mining and multiple criteria decision analysis. SoftwareX, vol. 10, no. 100323.