IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING BASED TRAFFIC ACCIDENT RISK PREDICTION USING RANDOM FOREST ALGORITHM

Dr. S. R. Menaka (Assistant Professor)
Department of Information Technology
K.S.R. College of Engineering,
Tiruchengode
Tamil Nadu, India

K. Boomika (Student)
Department of Information Technology
K.S.R. College of Engineering,
Tiruchengode
Tamil Nadu, India

M. Devadharshini (Student)
Department of Information Technology
K.S.R. College of Engineering,
Tiruchengode
Tamil Nadu, India

J. Ramsurya (Student)
Department of Information Technology
K.S.R. College of Engineering,
Tiruchengode
Tamil Nadu, India

Abstract--Traffic injuries own substantial demanding situations to public protection and transportation control structures worldwide. Predicting coincidence hazard correctly can assist in devising proactive measures to mitigate the prevalence of injuries and enhance typical street protection. In this look at, we advocate a device gaining knowledge of method utilising Random Forest set of rules for visitors' coincidence hazard prediction. The proposed version leverages ancient coincidence facts alongside diverse applicable functions along with climate conditions, street characteristics, visitors' density, and time of day to expect the chance of injuries at precise locations. Random Forest, recognized for its robustness and cap potential to address big datasets with complicated interactions, is hired to construct a green predictive version. Overall, our look at showcases the effectiveness of device gaining knowledge of techniques, mainly Random Forest, in visitors coincidence hazard prediction, presenting precious insights for policymakers, city planners, and transportation government to decorate street protection and decrease the prevalence of visitor's injuries.

Keywords: Machine learning, Road Accident, Traffic Accident, Geo - Location.

I. INTRODUCTION

A. Machine Learning

This creation units the level for exploring the function of gadget mastering in visitors' protection management, highlighting its cap potential to revolutionize twist of fate prediction, chance assessment, and proactive protection measures. By harnessing the strength of gadget mastering, we will paintings closer to growing more secure and greater sustainable transportation systems, in the long run saving lives and mitigating the societal and financial effects of visitor's accidents. The essence of gadget mastering lies in its cap potential to extract significant styles and insights from great and complicated datasets, hence uncovering precious understanding that won't be obvious via conventional analytical methods. By leveraging strategies along with supervised mastering, unsupervised mastering, and reinforcement mastering, gadget mastering algorithms can determine difficult relationships inside data, become aware of trends, and generalize from beyond reports to make knowledgeable predictions approximately destiny event.

B. Road Accident

Road injuries constitute a extensive worldwide public fitness and protection concern, inflicting great human suffering, monetary losses, and societal disruption. Every year, hundreds of thousands of lives are misplaced or completely altered because of avenue injuries, making it one of the main reasons of loss of life and harm worldwide. These injuries can arise because of a large number of factors, which include human error, inclusive of speeding, reckless using, and using below the have an impact on of alcohol or drugs; environmental factors, inclusive of damaging climate situations or poorly maintained roads; and vehicle-associated factors, inclusive of mechanical screw ups or defects

C. Traffic Accident

Traffic injuries, additionally referred to as street visitor's collisions or crashes, are incidents that arise on roadways regarding vehicles, pedestrians, cyclists, or different street users. These injuries can bring about numerous stages of harm to vehicles, accidents to individuals, or even fatalities. The effects of visitor's injuries may be intense, ensuing in accidents starting from minor cuts and bruises to intense trauma, everlasting disabilities, or fatalities. Moreover, visitor's injuries impose good sized monetary expenses on society, together with scientific expenses, assets harm, lack of productivity, and criminal fees.

D. Geo-Location

Geolocation, quick for geographic place, refers back to the identity or estimation of the real-global geographic place of an object, including a person, device, or place, the usage of diverse technology and techniques. Geolocation performs an important position in several programs and industries, which includes navigation, mapping, logistics, emergency offerings, marketing, and social networking. Overall, geolocation era has turn out to be an quintessential a part of contemporary life, allowing a extensive variety of programs and offerings that depend upon correct and well timed place information

II. LITERATURE REVIEW

Tarik Agouti[23], et al. has proposed in this paper, Today`s ultra-related global is producing a big quantity of statistics saved in databases and cloud surroundings mainly withinside the generation of transportation. These databases want to be processed and analyzed to extract beneficial data and gift it as a legitimate detail for transportation managers for similarly use, consisting of avenue protection, delivery delays, and delivery optimization. The capacity of statistics mining algorithms is essentially untapped, this paper indicates large-scale strategies consisting of institutions rule evaluation, more than one standards evaluation, and time collection to enhance avenue protection via way of means of figuring out hot-spots earlier and giving threat to drivers to keep away from the dangers. Indeed, we proposed a framework DM-MCDA primarily based totally on affiliation policies mining as a initial challenge to extract relationships among variables associated with a avenue accident, after which combine more than one standards evaluation to assist decision-makers to make their desire of the maximum applicable policies. The evolved gadget is bendy and permits intuitive advent and execution of various algorithms for an in depth variety of avenue visitors subjects. DM-MCDA may be multiplied with new subjects on demand, rendering information extraction extra strong and offer significant data that would assist in growing appropriate regulations for decision-makers.

Tiwari [22], et al. has proposed in this paper, Because it can show the relationship between the several exclusive types of characteristics that go into creating a street coincidence, analysis of street coincidences can be quite important. The characteristics of the street, the surroundings, the visitor, etc., can all have an impact on the street coincidence. Examining street coincidence can provide information about the role of those characteristics that can be used to overcome the coincidence rate. These days, a well-known method for examining the street coincidence dataset is data mining. We have implemented the street coincidence type on the concept of the street consumer category in this study. After organising the facts into homogeneous segments using the Self Organising Map (SOM) and K-modes clustering technique, we performed Support to classify the facts, decision trees, Naive Bayes (NB), and support vector machines (SVM) were used. Type has been applied to facts both with and without clustering. The final result shows that following fact segmentation using clustering, better type accuracy can be achieved.

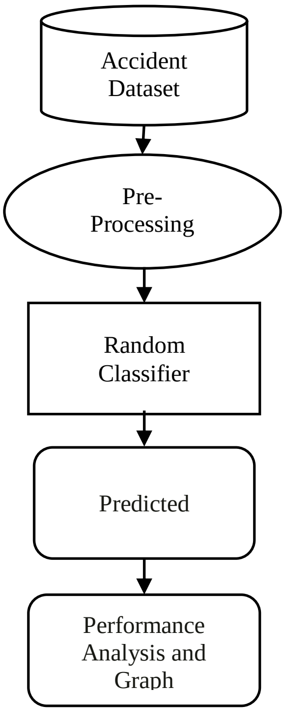
R.E. AlMamlook [19], et al. has proposed in this paper, One of the most important issues facing the industry is traffic injuries, which cause a great quantity of annual lives, injuries, and fatalities in addition to financial losses. For transit systems, predicting the visitor's coincidence severity accurately is an essential task. This research attempt sets trends for selecting a hard and fast set of influential factors and building a version for categorising injury severity. These styles are created using a variety of device learning methodologies. The coincidence records of visitors are subjected to supervised gadget mastery methods, which include AdaBoost, Logistic Regression (LR), Naive Bayes (NB), and Random Forests (RF). To handle record imbalance, the SMOTE set of rules is applied. The results of this analysis indicate that the RF version could be a useful tool for estimating how serious visitors' injuries will be. The RF set of rules has demonstrated superior overall performance, outperforming LR (74.5%), NB (73.1%), and AdaBoost (74.5%) in terms of accuracy.

J.Y. Wang [26], et al. has proposed in this paper, The duration of the visitor's twist of fate forecast is incredibly well-sized for quickly handling any injuries sustained by visitors, primarily for immediate injury rescue and removal of any risks to their safety. This research uses methods for predicting the twist of destiny length that are mostly based on artificial neural networks (ANN) and support vector machines (SVM). A case study using data on around 235 injuries that occurred on motorways between Dalian and Shenyang between 2012 and 2014 validates the suggested methodology. The two measures' performances are evaluated using the imply absolute error (MAE), the foundation imply rectangular error (RMSE), and the imply absolute percent error (MAPE). The following are the conclusions: Both ANN and SVM models could predict a visitor's twist of destiny length within exact bounds. For long-duration incident cases, the ANN variant achieves a better final result. For the purpose of predicting the length of site visitors' twists of destiny, the SVM version performs better overall than the ANN version.

Tiwari P [6], et al. has proposed in this paper, In every nation, traffic and street coincidence are a serious challenge. Road coincidence effects on a variety of items, including property damage, unique damage stages, and a high death toll. Technological data knowledge can be used to investigate unique causes behind traffic and street coincidences, in addition to weather, street, time, and other variables. We presented novel class and grouping approaches for research data in this work. We used unique classifiers such Decision Trees, Lazy Classifiers, and Multilayer Perceptron Classifiers to categorise the dataset according to casualty classes. Additionally, we used k-way and hierarchical clustering algorithms to group the datasets. Initially, we used those classifiers to analyse the dataset, and we performed accuracy tests at several stages, followed by clustering strategies and class strategies on the resulting clustered data. When clustering algorithms were used to a dataset as opposed to one that was categorised without clustering, our accuracy stage increased at a few points

III. EXISTING WORK

Vehicle injuries are a great challenge worldwide, main to lack of life, injuries, and monetary costs. Addressing this difficulty calls for powerful prediction fashions that may perceive cap potential dangers and mitigate them proactively. In latest years, gadget studying algorithms have proven exquisite promise on this domain, imparting the cap potential to examine significant datasets and extract significant insights to enhance street safety. In a latest have a look at carried out via way of means of researchers, Python turned into hired because the programming language to put into effect gadget studying algorithms for predicting car twist of fate dangers. The have a look at cantered on comparing the overall performance of various algorithms in diverse city street environments, which include situations concerning unique occasions including injuries and transient street controls.



Finally, the researchers used the educated prediction fashions to expect the real-time coincidence hazard of motors primarily based totally at the range of motors and the driver`s age band as category conditions. The consequences have been visualized the usage of color-coded regions and points, with blue indicating low coincidence probability, white indicating slight hazard, and crimson indicating excessive hazard. The visualization confirmed that because the newest release value (n) increased, the output of the prediction version become constantly optimized, main to clearer differences among extraordinary hazard levels

IV. PROPOSED WORK

Traffic injuries pose big threats to public protection and infrastructure worldwide. Addressing this problem calls for proactive measures, consisting of the prediction of twist of fate danger to mitigate their impact. Machine gaining knowledge of gives promising solutions, with the Random Forest set of rules status out for its effectiveness in coping with complicated datasets and imparting correct predictions. This precis outlines the important thing additives and standards worried in growing a system gaining knowledge of-primarily based totally machine for predicting site visitors twist of fate danger the use of the Random Forest set of rules, with a focal point on Python implementation.

E. DATA COLLECTION AND PREPROCESSING

Collect applicable facts bearing on site visitor's accidents. This ought to encompass elements consisting of climate conditions, avenue type, time of day, ancient twist of fate facts, car sorts involved, etc. Pre-procedure the facts to address lacking values, outliers, and inconsistencies. This would possibly contain facts cleaning, normalization, and function engineering.

F. FEATURE SELECTION

Identify the maximum applicable capabilities that would have an effect on the chance of a visitor's coincidence. These capabilities may consist of street conditions, weather, visitors'density, time of day, etc. Select capabilities primarily based totally on their significance in predicting coincidence risk.

G. MODEL TRAINING WITH RANDOM FOREST

Make use of the Random Forest method, an ensemble learning strategy that relies on selecting bushes. In order to increase accuracy and reduce overfitting, Random Forest constructs a few selection bushes and combines their predictions. Train the Random Forest version using functions identified as useful coincidence risk predictors on the pre-processed dataset.

H. MODEL EVALUTION

Use appropriate overall performance indicators, such as accuracy, precision, recall, and F1-score, to assess the educated version. Make sure the version is generalizable by using techniques like cross-validation to validate its overall performance.

I. HYPERPARAMETER TUNING

To maximise the Random Forest version's overall performance, adjust its parameters. This could involve modifying the number of trees, the highest tree intensity, and the least number of samples needed to slice up a node. Use appropriate overall performance criteria, such as accuracy, precision, recall, and F1-score, to assess the skilled version. Make sure the version is generalizable by using techniques like cross-validation to validate its overall performance.

J. PREDICTION

Once the version is educated and evaluated, it is able to be used to expect the probability of visitor's injuries primarily based totally on new enter data. Given a hard and fast of capabilities describing modern conditions (e.g., weather, street type), the version can expect the threat of a visitors twist of fate occurring.

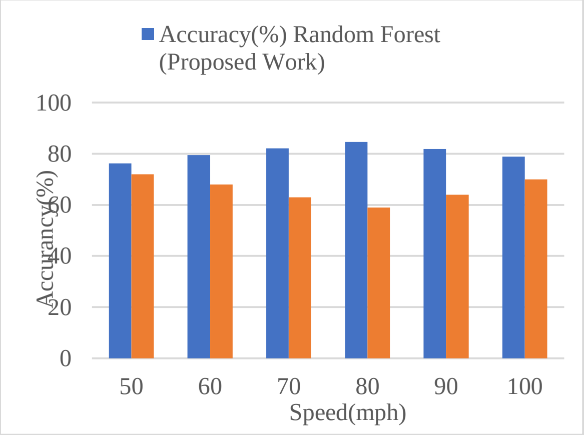
K. DEPLOYMENT AND MONITORING

Deploy the educated version in a manufacturing surroundings in which it could offer real-time predictions. Continuously reveal the version`s overall performance and replace it as important with new records to make certain its accuracy and relevance over time

V. RESULT ANALYSIS

In the proposed work utilizing Random Forest, the accuracy of predicting traffic accident severity increases with higher speeds. At 50 mph, the accuracy stands at 76.2%, climbing to 79.5% at 60 mph and reaching 82.1% at 70 mph. The accuracy peaks at 84.6% at 80 mph before slightly dropping to 81.8% at 90 mph and then to 78.9% at 100 mph. Conversely, in the existing work employing K-means clustering, the accuracy tends to decrease as speed rises. Starting at 72% at 50 mph, the accuracy diminishes to 68% at 60 mph and further to 63% at 70 mph. At higher speeds, the accuracy decreases even more significantly, with a drop to 59% at 80 mph, a slight recovery to 64% at 90 mph, and a subsequent increase to 70% at 100 mph.

Speed(mph)	Accuracy(%)	
	Random Forest (Proposed Work)	K means Cluster (Existing work)
50	76.2	72
60	79.5	68
70	82.1	63
80	84.6	59
90	81.8	64



VI. ALGORITHM DETAILS

One well-known machine learning technique that falls within the category of supervised learning is Random Forest. It can be applied to any ML problem involving classification and regression. It is mostly predicated on the concept of ensemble learning, which is a way to combine a few classifiers to improve the model's overall performance and tackle a complex problem.

According to the description, "Random Forest is a classifier that consists of some of selection bushes on diverse subsets of the given dataset and takes the common to enhance the predictive accuracy of that dataset." The random woodland area uses the prediction from each tree and is mostly dependent on it rather than counting on a single chosen tree entirely based on the collective votes of forecasts, and it forecasts the final product. A larger number of bushes inside the forested region improves accuracy and eliminates overfitting.

L. RANDOM FOREST ALGORITHM

The way Random Forest operates is in two parts: the first part involves creating the random woods by joining N selection trees, and the second part involves creating predictions for each tree that is formed in the first part.

Step 1: From the education set, choose K information factors at random.
Step 2: Construct the timber selection in accordance with the selected information variables (Subsets).

Step 3: Select the large variety N of wood that you need to build.
Step 4: Carry out Steps 1 and 2.
Step 5: Find each selection tree's predictions for new information factors, then allocate the newly discovered information factors to the class that receives the majority of votes.

VII. PSEUDO CODE

```
class DecisionTree:
 def __init__(self):
 def fit(self, X_train, y_train):
def predict(self, X_test):
class RandomForest:
 def __init__(self, n_trees):
  self.n_trees = n_trees
   self.trees = []
def fit(self, X_train, y_train):
  for _ in range(self.n_trees):
X_subset, y_subset = randomly_select_subset(X_train, y_train)
  tree = DecisionTree()
  tree.fit(X_subset, y_subset)
   self.trees.append(tree)
predict(self, X_test):
  predictions = []
  for tree in self.trees
predictions.append(tree.predict(X_test))
 final_predictions = []
        for i in range(len(X_test)):
            votes = [predictions[j][i] for j in range(self.n_trees)]
            final_predictions.append(majority_vote(votes))
        return final_predictions
def randomly_select_subset(X_train, y_train):
        def majority_vote(votes):
```

VIII. FUTURE WORK

Experiment with extraordinary hyperparameters of the Random Forest set of rules together with the variety of trees, tree depth, and minimal samples consistent with leaf. Additionally, take into account the usage of ensemble strategies like Gradient Boosting or XGBoost, which frequently outperform Random Forest in predictive accuracy. By specializing in those areas, destiny paintings can improve the todays in device learning-primarily based totally site visitors' coincidence threat prediction, main to extra correct fashions and in the long run enhancing avenue safety

IX. CONCLUSION

In conclusion, the utilization of machine learning techniques, specifically Random Forest, for predicting traffic accident risks presents a promising approach with significant potential for enhancing road safety measures. Through the analysis of various features such as weather conditions, road characteristics, and historical accident data, Random Forest models can effectively discern patterns and relationships that contribute to accident occurrence. The benefits of using Random Forest for site visitors twist of fate danger prediction consist of its cap potential to address massive datasets with several enter variables, its resilience to overfitting, and its capability to offer insights into the relative significance of various features. Moreover, the interpretability of Random Forest fashions lets in stakeholders to apprehend the elements influencing twist of fate risks, thereby facilitating centred interventions a coverage decision.

X. REFERANCE

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.
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 beingused 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), Informatics.
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. Software, vol. 10, no. 100323
8. Y. Zou, B. Lin, X. Yang, L. Wu, M. M. Abid, and J. Tang (2021). Application of the Bayesian model averaging in analyzing freeway traffic incident clearance time for emergency management. J. Adv. Transp., Pages 1–9.
9. J. Tang, L. Zheng, C. Han, W. Yin, Y. Zhang, Y. Zou, and H. Huang (2020). Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. Anal. Methods Accident Res., 27.
10. M. Umer, I. Ashraf, A. Mehmood, S. Ullah, and G.S. Choi (2021). Predicting numeric ratings for Google apps using text features and ensemble learning. ETRI J., 43(1): 95–108.
11. M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G.S. Choi, and B.W. (2020). Fake news stance detection using deep learning architecture (CNNLSTM). IEEE Access, 8: 156695–156706.
12. S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G.S. Choi, and B.W. (2021). Aggression detection through deep neural model on Twitter. Future Gener. Comput. Syst., 114: 120–129.
13. Z. Imtiaz, M. Umer, M. Ahmad, S. Ullah, G.S. Choi, and A. Mehmood (2020). Duplicate questions pair detection using Siamese MaLSTM. IEEE Access, 8: 21932–21942.
14. M.I. Sameen and B. Pradhan (2017). Severity prediction of traffic accidents with recurrent neural networks. Appl. Sci., 7(6): 476.
15. S. Seid and Pooja (2019). Road accident data analysis: Data preprocessing for better model building. J. Comput. Theor. Nanosci., 16(9): 4019–4027. [9] S.K. Singh (2017). Road traffic accidents in India: Issues and challenges. Transp. Res. Proc., 25(5): 4708–4719.
16. D. Delen, R. Sharda, and M. Bessonov (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Anal. Prevention, 38(3): 434–444.
17. D.W. Kononen, C.A.C. Flannagan, and S.C. Wang (2011). Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accident Anal. Prevention, 43(1): 112– 122.
18. P. Duan, Z. He, Y. He, F. Liu, A. Zhang, and D. Zhou (2020). Root cause analysis approach based on reverse cascading decomposition in QFD and fuzzy weight ARM for quality accidents. Comput. Ind. Eng., 147.
19. H.M. Alnami, I. Mahgoub, and H. Al-Najada (2021). Highway accident severity prediction for optimal resource allocation of emergency vehicles and personnel. In Proc. IEEE 11th Annu. Comput. Commun. Workshop Conf. (CCWC), Pages 1231–1238.
20. M. Umer, I. Ashraf, A. Mehmood, S. Kumari, S. Ullah, and G. S. Choi (2021). Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model. Comput. Intell., 37(1): 409– 434.
21. S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi (2021). Discrepancy detection between actual user reviews and numeric ratings of Google app store using deep learning. Expert Syst. Appl., 181.
22. P. Tiwari, S. Kumar, and D. Kalitin (2017). Road-user specific analysis of traffic accident using data mining techniques. In Proc. Int. Conf. Comput. Intell., Commun., Bus. Anal. New York, NY, USA, Pages 398–410.
23. R.E. AlMamlook, K.M. Kwayu, M.R. Alkasisbeh, and A.A. Frefer (2019). Comparison of machine learning algorithms for predicting traffic accident severity. In Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol., Pages 272–276.
24. T. Beshah and S. Hill (2010). Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia. In Proc. AAAI Spring Symp., Art if. In tell. Develop., Volume 24, Princeton, NJ, USA: Citeseer, Pages 1173–1181.
25. X. Ma, C. Ding, S. Luan, Y. Wang, and Y. Wang (2017). Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. IEEE Trans. Intell. Transp. Syst., 18(9): 2303–2310.
26. B. Yu, Y.T. Wang, J.B. Yao, and J.Y. Wang (2016). A comparison of the performance of ANN and SVM for the prediction of traffic accident duration. Neural Netw. World, 26(3): 271.