iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44300

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Road Accident Analysis Using Machine Learning

Hemanth Kanchi[1], Deepthi Kandagatla[2], Akshay Baheti[3], Dr. M. Shailaja[4]
[1, 2, 3]*Department of Electronics and Computer Engineering (ECM),* [4]*Associate Professor, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India*

Abstract: ***Road accidents are one among the key concerns within the country. The economic impact of traffic accidents cost many billions of dollars for US citizens per annum. Reducing the traffic accidents has always been a challenge. The most important objective of the project is to investigate what are the foundations for mishaps like effect of precipitation or any natural elements and to predict the probability of road accidents supported the present accidents records. By performing data visualization, we are able to identify the car accidents zones, factors effecting accident severity, and locations etc. Identifying the key factors of how these accidents occur can help in implementing well informed actions. Road safety analysts following up on street accidents information have seen outcome in street auto collisions examination through the applying information logical methods, however, little headway was the expectation of street injury. This report says progressed information investigation strategies to anticipate injury seriousness levels and assesses their presentation. The review utilizes prescient demonstrating methods to recognize chance and key factors that adds to mishap seriousness. The review utilizes openly accessible information from US branch of transport that covers the sum from 2005 to 2019. The report presents a methodology which is adequately general so are frequently applied to various informational collections from different nations. The outcomes recognized that tree based methods like Extreme Gradient Boosting beat relapse covers, as Artificial Neural Network (ANN) also to the paper, recognizes fascinating connections were recognized concepts related with nature of information.***
Keywords: Artificial Neural Network; Extreme Gradient Boosting ; Data Visualization

I. INTRODUCTION

Road Accident is an unexpected event that occurs causing injuries to the people involved in the accident, these accidents happen on the roads due to different types of reasons likes drink and driving, The twenty first century has been seeing a climb of road motorization because of quick development of populace, huge urbanization, and expanded versatility of the chic culture, risks of road site guests casualty can likewise furthermore also end up better and Road Accidents can be expected to be as a "current pandemic". This report bears the cost of an insightful system to are expecting incident seriousness for road site guests wounds. Past examinations on road site guests wounds assessment had specifically relied upon measurable methodologies like direct and Poisson relapse. This report manages the cost of a logical system to are expecting occurrence seriousness for road site guests wounds. In particular, the paper tends to inconveniences connected with realities pre-handling and schooling like realities conglomeration, change, work designing and imbalanced realities. Also, the report focuses to apply device concentrating on styles to permit extra right expectations. Thus, the analyzes the general exhibition of various device concentrating on calculations in foreseeing the happenstance hurt seriousness.

This project analyse the data and predicts the accidents that might occur in future at any place or according to weather conditions and also visualize the data to understand the data in less time. It uses machine-learning algorithms like KNN(K-nearest-neighbours), Artificial Neural Network, Decision Tree, Random Forest to show the accuracy of the predictions and these four algorithms have been compared to give out the highest accuracy as a result. Firstly, we clean the data by removing all the null values or filling the null values with the mean of the preceding and succeeding data of the null value. In data visualization we will be visualizing which the data that shows , in what conditions(i.e weather) the accidents has recorded more, What the the top accident prone zones in the country, which cities and states has the highest record of the accidents , In which year ,month, week, hour the accidents has recorded are done in the data visualization. Then, Feature selection is applied to the columns and changing the low severity to zero and the hih severity to 1. Next, Implementing the four algorithms mentions above in the paper and evaluating the algorithms using accuracy score. After comparing the models .Random forest gives out the accuracy of 84%.

II. LITERATURE REVIEW

Examination models are divided into two classifications: prescient or logical models that endeavour to comprehend and evaluate crash chance and improvement strategies that attention on limiting accident risk through course/way determination and rest-break planning. Their work introduced a freely accessible information sources and clear scientific methods (information outline, representation, and aspect decrease) that can be utilized to accomplish more secure steering and give code to work with information assortment/investigation by experts/analysts. The paper likewise evaluated the factual and AI models utilized for crash risk displaying. [3] classified the enhancement and prescriptive logical models that attention on limiting accident risk. Ziakopoulos. [4] basically investigated the current writing on various spatial methodologies that remember aspect of room for its different angles in their examinations for street security. Moosavi . [5] recognized shortcomings with street car crashes research which include: limited scope datasets, reliance on broad arrangement of information, and being not pertinent for ongoing purposes. The work proposed an information assortment procedure with a profound brain network algorithm called Deep Accident Prediction(DAP); The outcomes showes critical upgrades to foresee interesting mishap occasions. Zagorodnikh .[6] fostered a data framework that shows the mishaps focus on electronic landscape map naturally mode for Russian Road Accidents to help improving on the RTA examination. Kononen . [7] broke down the seriousness of mishaps happened in United States utilizing strategic relapse model. They detailed execution 40% and 98%, for awareness and particularity separately. Additionally, they distinguished the main indicators for accidents level are: change in speed, safety belt use, and safety precautions.

The Artificial neural networks (ANN) is an algorithm that are used as one of the information mining instruments and non-parametric methods where specialists have dissected the seriousness of mishaps and wounds among those engaged with such crashes. Delen . [8] applied an ANNs to show the connections between injury seriousness levels and crash related factors. They utilized US crash information with 16 ascribes.

III. ARCHITECTURE

The Model includes different steps to follow to reach our aim. This type of analysis and visualization are done before but , our team aim is to give the best accuracy score so we will analyse the data and find the accuracy score of the predictions of four different machine-learning algorithms and then give out the highest accuracy as a result. Before building of model and evaluation there are few steps to be followed to give out the better result. First, the collection of data from different sites. Then, cleaning the dataset by removing the null values if the columns or rows have most of the null values and that column or row is important in analysis then the values are filled with the mean values of the preceding and succeeding of the null values and dropping the columns which are not very useful in building models. Next comes visualization of the data which helps in understanding the data easily and reduces the time period of analysing the data. The data visualization involves in ploting pie charts, bar graphs, line graphs , correlation matrix and in the process of data visualization we have used the package called seaborn which is built on mathplot library which adds more colours to the palette and makes the graphs more colourful and attractive. Then comes the feature selection which selects only few columns form the dataset and change the low severity to zero and the high severity to one in the data before building the model. The four models are build with different packages and the accuracy scores has been noted. Lastly ,these models are evaluated and the highest accuracy score is give out ad result.

Fig 1 : Architecture of the proposed system [10]

IV. MODEL EVALUATION

Evaluation is a basic step in the data assessment which overviews the perceptive limit of the model and recognize the model which performs effectively. A couple of methods ordinarily used to evaluate gathering models, for instance, the chaos system, recipient manager twist (ROC) and the locale under the curve (AUC). A chaos structure shows the right groupings real up-sides (TP) and certified negatives (TN) despite wrong request fake up-sides (FP), and deluding negatives (FN).

A confusion matrix is a table that shows how well a grouping model (or "classifier") performs on a bunch of test information for which the genuine qualities are known. The disarray grid itself is direct, yet the related terminology may confound.

A sum of 165 expectations were made by the classifier (e.g., 165 patients were being tried for the presence of that infection).

The classifier accurately anticipated "yes" multiple times and "no" multiple times out of 165 cases.

1) *True Positives* (TP): These are cases in which we anticipated yes and they do.
2) *True negatives* (TN): The cases we predicted no.
3) *False Positives* (FP): Although we predicted indeed, they don't have the illness. (This is likewise alluded to as a "Type I blunder.")
4) *False negatives* (FN): We anticipated that they wouldn't have the illness, yet they do. (This is likewise alluded to as a "Type II blunder.")
5) *TPR*: TPR approaches the quantity of genuine up-sides separated by the all out number of up-sides. In this way, the quantity of genuine positive focuses is TP, and the all out number of positive focuses is the amount of the sections containing TP, which is P. TPR = TP/P.
6) *FPR*: A False Positive Rate is a precision metric that can be estimated on a subset of AI models. To get a perusing on obvious precision of a model, it should have some idea of "ground truth", for example the genuine situation.

The bogus positive rate is communicated as FP/FP+TN, where FP addresses the quantity of misleading up-sides and TN addresses the quantity of genuine negatives.

A. *Precision*

Precision is one proportion of an AI model's exhibition since it estimates the precision of a positive expectation made by the model. Accuracy is determined by separating the quantity of genuine up-sides by the complete number of positive forecasts.

Precision= TruePositives/(TruePositives + FalsePositives)

B. *Accuracy*

Model accuracy is characterized as the quantity of right groupings anticipated by a model separated by the all out number of forecasts made. It is one strategy for assessing a model's exhibition, however it is nowhere near the one to focus on.

One measurement for assessing order models is exactness. Casually, precision is the level of right forecasts made by our model. Officially, exactness is characterized as follows: Number of right predictions = Accuracy Total number of figures.

C. *Recall*

The recall assessment technique is characterized as the negligible portion of accurately recognized positive cases. b) characterized as the extent of genuine positive cases versus all cases wherein the expectation is right. c) characterized as the extent of right forecasts mentioned out of every single observable reality

Ascertain the review. The review is determined as the proportion of tp/(tp + fn), where tp explains the quantity of distinctive up-sides and fn addresses the quantity of misleading negatives. The review is effectively the classifier's capacity to track down every example. The best worth is 1 and the absolute bad worth is 0.

D. *F-Measure*

The F-measure is determined as the symphonious mean of accuracy and review, with equivalent weighting for each. It empowers a model to be assessed involving a solitary score that records for both accuracy and review, which is helpful while portraying model execution and contrasting models.

Precision can be used when TP's(True Positives) and TN's(True Negatives) are significant, while F1-score is utilized when False Negatives and False Positives are significant. At the point when the class conveyance is comparable, exactness can be utilized.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Metric	Formula
True positive rate, recall	$\frac{TP}{TP+FN}$
False positive rate	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F-measure	$\frac{2 \cdot precision \cdot recall}{precision + recall}$

Fig 2 : Evaluation Metrics [9]

V. METHODOLOGY

The initial step for information examination process is information assortment which is viewed as the essential structure block for effective information investigation project. There are numerous information sources like neurons, visual information with cameras, and Internet of Things and cell phones that catches information in various organizations and should be put away realtime or disconnected. What's more, information gathered from various specialists connected with traffic volume, mishap subtleties and segment data. The capacity can be on the neighbourhood clouds or on servers. The main point of information the executives pyramid is information pre-processing. It requires pre-processing prior to playing out any investigation. The procured information might incorporate missing data that expected to amend as the need might have arisen to be taken out because of duplication. The pre-processing might include the information change as it helps in information standardization, property choice, discretization, and order age. Information decrease perhaps expected on the enormous size of information as the investigation of a tremendous measure of information is more diligently. Information decrease expands the productivity of capacity and the examination cost. Information Analysis utilize numerous AI calculations to get the knowledge of information. The information investigation is exceptionally pivotal for any association as it gives the definite data about the information and is useful in specific direction and forecasts about the business. Information can be introduced in different structures relying upon the kind of information being utilized. The information can be displayed into coordinated tables, outlines, or charts. Information show is vital for costumers as it gives the outcomes from the examination of information in a visual arrangement. This paper takes a gander at building prescient algorithm for mishap seriousness level and examines the method involved with developing a characterization model to anticipate mishap seriousness level, specifically the review:

• Presents the information the board system. This is trailed by conversation on how the information was ready preceding displaying.

VI. RESULTS

The main objective of this model is to visualize the data to understand detect which states and countries are has most accident prone area, in which weather the accidents are occurring and at what hour/day/week/month/year accident records are more and analyse the data with help of machine-learning algorithms and predict the accuracy of accidents that might occur in future.

The reason that we used US data instead of India accidents is that the available Indian data does not have much information to work on and with that data we would not have explored every perspective of the model, US accidents data consists all the data that we want to place in our model i.e time of accidents, place, states, cities of the accidents.

Our data of US accidents consists of columns that we are not going to use in the machine-Learning models so we apply feature selection to the data and select only few columns that gives out the effective outcome. During feature selection, we change the low severity data to zero and the high severity to one so that predictions are made right. In this model evaluation Random forest has the highest accuracy score of 90%.

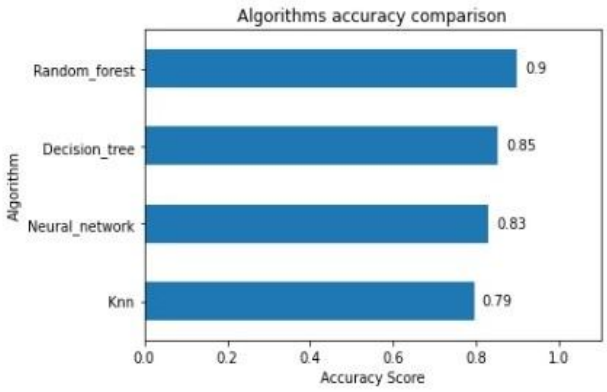


Fig 3 : Accuracy comparison of Machine Learning algorithms

VII. CONCLUSION

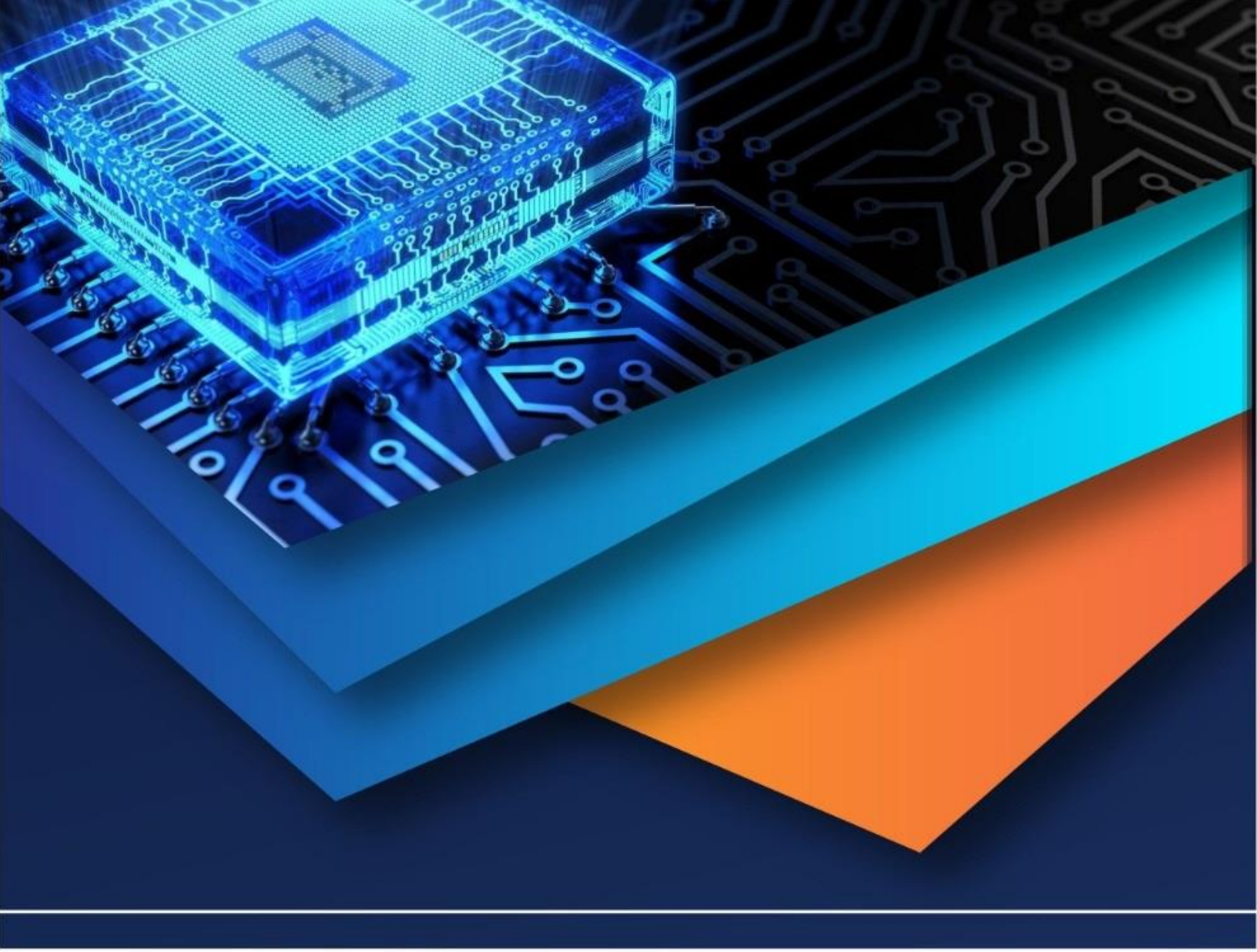
We loaded the 'US accidents' dataset from the Kaggle site and imported the necessary libraries. In the Data preprocessing step, null values were replaced with mean, mode, and 0 values. Turning loop column was removed because it only has one value, False, and isn't useful for our analysis. The Data Visualization step discovered some interesting things, such as which states have the most accident cases, whether accidents occur more during the day or at night, which zones have the most accidents, what the atmospheric conditions are during accidents, and so on. All of this information will help you travel more efficiently and listen while driving. The correlation of variables was plotted using a heatmap and selected features that are highly correlated. Using the label encoder, boolean values were converted to numeric and civil time zone data was chosen predicting the seriousness of accidents Following that, modeling was done using the -Knn, neural network, decision tree, and random forest algorithms. The models were evaluated using accuracy scores, a classification report, and a confusion matrix. Finally, a graph was created that compares the accuracy lots of all the algorithms. According to this graph, the Random forest model has the highest accuracy score of 90 percent, while Knn has the lowest accuracy score of 79 percent.

VIII. LIMITATIONS AND FUTURE STUDIES

The limitations of the following design can be adjusted with minor changes in the foreseeable future. The First limitation being the lack of availability of data-sets for Indian Road Accidents. There are no data-sets for Indian Road Accidents whereas we can find numerous data-sets for the countries like USA, UK, etc. If the data-sets are found then one can easily train the data and can help in reducing fatalities in road accidents. The future scope for this model would be the inclusion of Indian road accident's data-sets which can increase the scope for understanding and contributing to the safety of road traffic in India.

REFERENCES

[1] H. Meng, X. Wang, and X. Wang, "Expressway crash prediction based on traffic big data," ACM Int. Conf. Proceeding Ser., pp. 11–16, 2018.
[2] A. Mehdizadeh, M. Cai, Q. Hu, M. A. A. Yazdi, N. Mohabbati-Kalejahi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, "A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling," Sensors (Switzerland), vol. 20, no. 4, pp. 1–24, 2020.
[3] Q. Hu, M. Cai, N. Mohabbati-Kalejahi, A. Mehdizadeh, M. A. A. Yazdi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, "A review of data analytic applications in road traffic safety. Part 2: Prescriptive modeling," Sensors (Switzerland), vol. 20, no. 4, pp. 1–19, 2020.
[4] A. Ziakopoulos and G. Yannis, "A review of spatial approaches in road safety," Accid. Anal. Prev., vol. 135, no. July, p. 105323, 2020. [Online]. Available: https://doi.org/10.1016/j.aap.2019.105323
[5] S. Moosavi, M. H. Samavatian, A. Nandi, S. Parthasarathy, and R. Ramnath, "Short and long-term pattern discovery over large-scale geospatiotemporal data," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 2905–2913, 2019.
[6] N. Zagorodnikh, A. Novikov, and A. Yastrebkov, "Algorithm and software for identifying accident-prone road sections," Transp. Res. Procedia, vol. 36, pp. 817–825, 2018. [Online]. Available: https://doi.org/10.1016/j.trpro.2018.12.074
[7] D. W. Kononen, C. A. Flannagan, and S. C. Wang, "Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes," Accid. Anal. Prev., vol. 43, no. 1, pp. 112–122, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.aap.2010.07.018
[8] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," Accid. Anal. Prev., vol. 38, no. 3, pp. 434–444, 2006.
[9] https://deepai.org/machine-learning-glossary-and-terms/evaluation-metrics
[10] https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY