ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse127862019.pdf
https://doi.org/10.30534/ijatcse/2019/127862019

Analysis of Road Accidents to Indentify Major Causes and Influencing Factors of Accidents – A Machine Learning Approach

Tirumuru Ketha [1], S Sagar Imambi [2]
[1]Student, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Guntur District, Andhra Pradesh-522502, India.: tketha.1@gmail.com
[2]Professor, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Guntur District, Andhra Pradesh-522502, India.: simambi@kluniversity.in

ABSTRACT

As indicated by WHO roughly 1.35 million individuals die every year because of road traffic crashes. Accidents are the fundamental issues going up against the world as they cause various wounds, fatalities as well as financial misfortunes consistently**.** This circumstance has found road accidents issue, affected general wellbeing and nation economy. Precise models to anticipate the reason for a accidents and to give safe driving proposals are a basic assignment for road transportation frameworks. This examination exertion sets up models dependent on the variables, such as Weather condition, Causes, Road Features, Road Condition, Type of Accident, that lead to accidents likewise choose a lot of influential components from best model and to develop a model for characterizing the reason for accidents. These are detailed by different Supervised Machine Learning methods like, Logistic Regression (LR), K- Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT) and Random Forests (RF) are implemented on accidents information like to discover how each component is impacting the accidents variables and this gives a safe driving proposals to limit the accidents. The discoveries of this investigation demonstrate that the Decision Tree can be a promising model for anticipating the reason for accidents. Decision Tree demonstrated better performance on all the components, namely Weather condition, Causes, Road Features, Road Condition, Type of Accident, with 99.4%.

Key words: Logistic Regression (LR), K- Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT) and Random Forests (RF), Traffic accident, Influencing factors, Machine Learning.

1. INTRODUCTION

In India, more than 150,000 individuals are died every year in accidents. That is around 400 fatalities per day. Every year more than 1 million vehicles are added to traffic averagely. There is a massive monetary misfortunes are a budgetary weight for creating nations.

Machine Learning includes the investigations which will find data which is systematic and purposeful data structures acquired from disordered information and summarizing it into helpful data and it very well may be utilized for smart business decision.

Classification is an important Machine Learning technique which extracts valuable information from huge datasets. In this work we applied different supervised ML classification methods like Logistic Regression (LR), K- Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT) and Random Forests (RF) on the factors such as Weather condition, Type of Accident, Causes, Road Features, Road Condition which lead to major causes of accidents also gives set of influential factors from the best model and gives a safe driving proposals to limit the mishaps additionally examines the presentation of the models. The exhibition measures are true positive, false positive, accuracy, precision, recall, rate of error and level of accurately classified models. Trial results uncover that, DT beat other ML technique with more accuracy and a less mistake rate.

1.1 Objectives

The main objective of our analysis are:

- To preprocess the accidents information.
- To produce the best model utilizing ML techniques like LR, KNN, NB, DT and RF.
- To identify the major impacting variables of major accidents from the best model that gives protected driving proposals to limit the accidents.

2. LITERATURE SURVEY

Rabia Emhamed Al et al. [1] used Machine Learning techniques to build a system for classifying the harshness of an mishap using Ada boost, LR,NB,RF. This work shows that Random Forest which gives better performance with 75% accuracy. The limitation of this project is that considering only few features which the other features may also have impact on the severity of accident.

Liling Li et al. [2], to give safe driving suggestions by using the algorithms such as, Apriori, NB, K- Means were applied on FARS data. There was an relationship between fatal rate and weather, light, surface, drunk and drive conditions were investigated. Using K Means they have suggested the regions which are more risky to drive. Here the main drawback is that they have considered only the fatal (people die in the accidents) accidents data but not the non fatal accidents data if it

could be considered then the suggestions could be made from large amount of data with more accuracy.

Md. Shahriare Satu et al. [3], applied Decision Tree algorithms on the Eight hundred and ninety two accidents happened on the N5 Highway where the data was collected from Accident research Institute Balbladesh. From this critical analysis they have found best classifier among the 12 decision trees classifier which helps to reduce the accidents in the N5 highway in the Bangladesh.

Ms. E. Suganya et al. [4] analyzed the road accidents and find the year where there is increase in the accidents in the regions of India using Linear and Logistic Regression, SVM, DT, NB, RF, K- Nearest Neighbour and gradient boost. From experimental results K-Nearest Neighbour is the best when compared with remaining algorithms with 93.7% accuracy using R tool. It is found that 2016 bearing the highest No.of accidents in Tamil Nadu.

The combination of K- Means clustering where the information segmentation is done and association rule where situations appears for whole data which produces a large amount of information used to analyze the data in heterogeneity type of road. The main goal is to find main factors associated for the cause of accident. [5]

Shristi Sonal, Saumya Suman et al. [6] found the severity of accident using Machine Learning techniques which relay not only on internal factors like driver but also the external factors like weather conditions. The aim of their study is to help authorities to take actions on accident zone traffic and weather conditions.

Here several Machine Learning algorithms like DT and NB are used for determination of the harshness of the accident using WEKA tool. From the Result analysis it shows that J8 classifier gives the better accuracy compared to other algorithms to determine the severity of an accident.[7]

Table 1: Work related to Area of research

No	Techniques Used	Advantages	Drawbacks
[1]	Ada boost, LR, NB, RF where RF got better accuracy with 75%.	To find the severity of an accident.	The limitation of this work is that considering only few features where the other features may also have impact on the severity of accident.
[2]	apriori,Naïve Bayes,K Means	to give safe driving suggestions and risky regions to take precautions.	Here they have considered only the fatal accidents but not the non fatal accidents data.
[3]	Decision Tree algorithms	Generated best classifier which helps to reduce the accidents	Considered only Decision tree algorithms but not another which would give better chance to reduce accidents.
[4]	Linear and Logistic Regression, SVM, DT, NB, RF, K- Nearest Neighbour and gradient boost.	To analyze road accidents and find year where there is increase in the accidents in the regions of India	The main drawback is that here considered only the increase in the accidents year but not the other factors which cause the accidents, severity etc.
[5]	The combination of K- Means clustering.	Goal is to find main factors associated for the cause of accident.	Considered of K-Means clustering but not another which would give better chance to find more factors associated for cause of accident.
[6]	Apriori, FP-growth	Aim of this study is to help authorities to take actions on accident zone traffic and to find severity of accident.	The drawback is that considered only rule mining algorithms
[7]	DT and NB	To determine the severity of an accident.	The main drawback is that here they considered only severity of accident but not the other factors which can reduce the cause of accident.

3. METHADOLOGY

The approach that was utilized in this project is to build the forecasting classification for the best accurate model (LR, NB, KNN, DT and RF). General strategy is to explore this project, includes dataset selection, data preprocessing, attribute selection, data splitting, extracting the necessary knowledge to analyze the metrics of the LR, NB, KNN, DT and RF Classifier models which give better performance on prediction to utilize the most precise as a prescient model to get most influenced factors and give safe driving recommendations to limit the accidents. The performance matrices are compared.

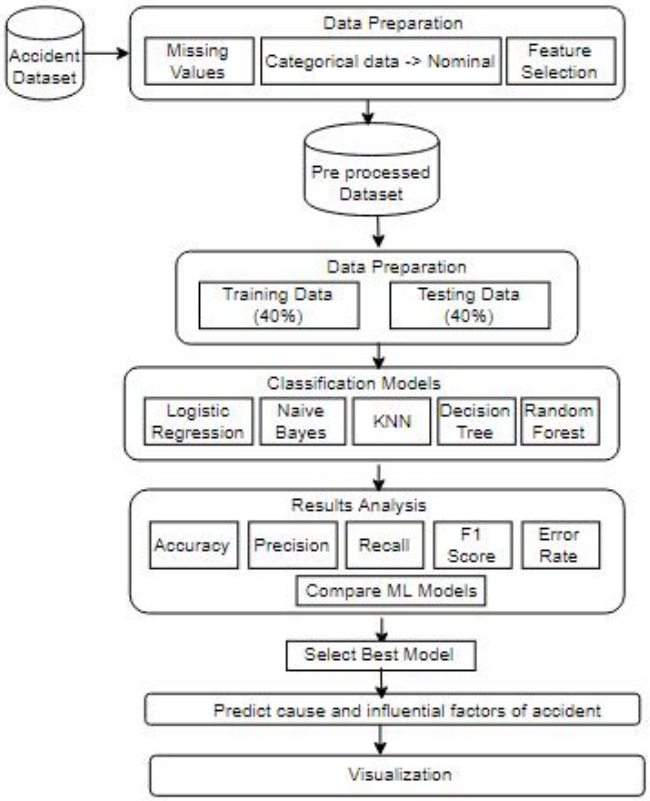

Figure 1: Block Diagram

3.1 Dataset Selection

Data is the most import part when you take a shot at prediction systems. It assumes an extremely essential job our entire project i.e., our system relies upon that information. So selection of data is the first and the basic advance step which should be performed appropriately. For our project we got the dataset from the govt. site. These datasets were accessible for all. There are different huge amounts of sites who give such datasets. The dataset we pick had chosen dependent on the different variables and requirements that we were going to take under the thought for our prediction system.

3.2 Dataset Description

This dataset holds the data about accidents happened during the period ranging from 2012 to 2019 up to June. This dataset contains 31661 road accidents records with 19 variables which are the cause for the accident.

3.3 Data Pre-processing

Data preprocessing is a significant stage for taking care of the information before using in machine learning algorithm or developing a model. All records with missing esteem usually replaced by mean estimation of column in the dataset All categorical values were converted to nominal values. This procedure includes different steps including cleaning, normalization, attribute selection, splitting and transformation.

Accidents are classified into 4 classes:

1. Fatal
2. Grievous Injury
3. Minor Injury
4. Non Injury

3.4 Attribute Selection

Attribute selection, otherwise called Feature choice or variable choice, is a methodology of picking a subset of significant properties for using in model advancement. We chose different edge for the No. of most powerful attributes to be used in the experiments. At that point the algorithms were applied to the dataset on these chose attributes, and the accuracies of them were contrasted and rehashed this procedure with multi edges to acquire the most noteworthy accuracies.

3.5 Dataset Splitting

Above we have explained about the preprocessing steps for our dataset, after completing the process of preprocessing, we need to part this dataset into a Training (60%) and Testing datasets (40%).

3.6 Analysis and Implementation of Machine Learning Classification Model:

For this project we utilized Anaconda which contains Jupyter notebook it is a free source conveyance of R and Python programming languages for enormous data processing, prediction, analysis etc where we utilized Python for coding. The accompanying five popular machine learning classification methods were utilized in this project:

3.6.1 Logistic Regression (LR)

Logistic Regression is a classifier. The probability of the calculation is to plot results of straight capacities to sigmoid capacities. The linear regression model is a basic scientific model and easy to execute. In straightforward, it predicts the likelihood of event of an occasion by isolating information to a logic function. Thus generally called a Logistic Regression.

3.6.2 Naïve Bayes (NB)

The Naive Bayesian classifier relies upon Bayes theory with independent assumptions between the predicators which is used for large datasets. Bayes classifiers are a gathering of clear "probabilistic classifiers" in perspective on applying Bayes hypothesis with strong opportunity acceptance between the attributes.

3.6.3 K-Nearest Neighbors (KNN)

KNN model frequently used to build classification and Regression models. This model is a primary set of rules that stores each accessible case and characterizes new instances by using a majority vote of its k neighbors. This instance is being allocated to the class is normally most ordinary amongst its k closest neighbors predicted using a distance function, as an instance Euclidean, Manhattan, Minkowsi and Hamming distance.

3.6.4 Decision Tree (DT

It is one of the supervised ML algorithm that's most often utilized for developing classification models. DT algorithm can be applied for datasets which deals with both categorical and continuous variables. This model relies upon maximum significant factors to make distinct organization viable.

3.6.5 Random Forest (RF)

The Random Forest version is an ensemble learning method which develops a chain of selection trees at training time and yields the magnificence that is the mode of the classes (characterization) or means prediction (relapse) of individual trees. Growing a group of trees and choosing the elegance kind through casting a vote have progressed accuracy. The minimal No. Of samples required to break up a node was set to two, and minimum samples according to leaf are set to at least one.

3.7 Performance Measurement

The progression, assessment of the performance of the type LR, NB, KNN, DT and RF algorithms are accomplished and in comparison. Accuracy, Precision, Recall and F1-measure had been utilized in the assessment procedure.

4. RESULT ANALYSIS

For the result analysis, information gathered from Open Government Platform. We utilized 3 execution factors such as Accuracy, Rate of Error and Influencing Factors which is cause for an accident. By contrasting 5 unique classifiers to be specific LR, KNN, NB, DT and RF calculations on 5 features like Weather condition, Type of Accident, Causes, Road Features, Road Condition. The Decision Tree algorithm has produced more accuracy in all the 5 cases with minimum error rate than remaining algorithms.

4.1 Classification of Accuracy

The accompanying table describes that the Accuracy, Precision, Recall, F1-score measures of classification algorithms. Table 2 and fig 2 shows the result of classification metrics on algorithms such as LR, KNN, NB, DT and RF which is applied on Weather Conditions.(This is also done for remaining 4 features that are described above).

Table 2: Classification Metrics

ALGORITHMS	WEATHER CONDITIONS			
	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Logistic Regression (LR)	94.9	90	95	93
Naive Bayes (NB)	89.2	90	89	90
K Nearest Neighbors (KNN)	94.9	92	95	93
Decision Tree (DT)	**99.4**	**99**	**99**	**99**
Random Forests (RF)	94.9	90	95	93

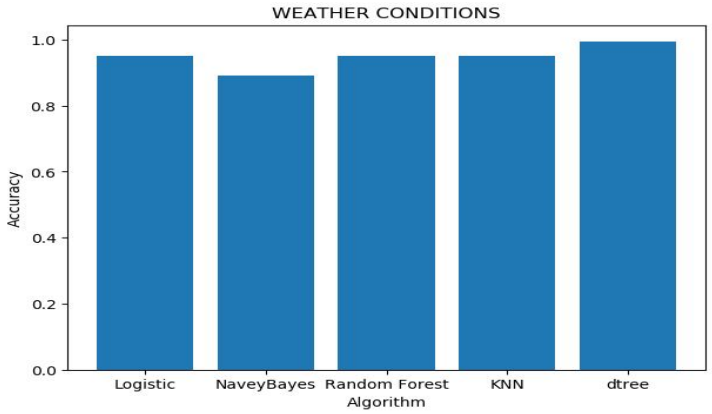

Figure 2: Graph for Classification of Accuracy

4.2 Error Rate

RMSE will be found in numeric worth as it were. RMSE esteem determined on the weather conditions for which DT with most minimal RMSE of 0.14 and other algorithms as pursues LR with 0.45, NB with 0.5, KNN with 0.45, RF with 0.15. Figure 3 speaks to (RMSE) root mean squared error. Here the DT characterization calculation has the most lowest root mean squared error.

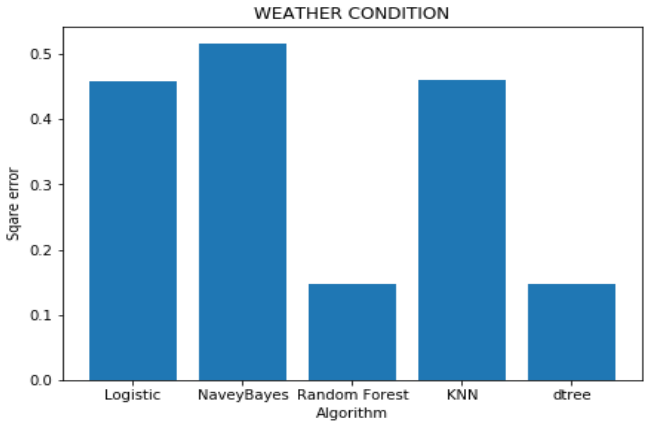

Figure 3: Graph for RMSE values

4.3 Influencing Factors

Here we are considering the influencing factors that are affecting on the Weather Conditions. From Figure 4 and Figure 5 we came to the conclusion that Time of accident is highly influencing on Time of Accident and very low influencing on Area (urban/Rural).

	importance
Time of Accident	0.193921
KM && M	0.183211
Vehicle Responsible	0.173336
Nature of Accident	0.143061
Location	0.049093
Vehicle Involved	0.048157
Help Provided by Ambulance/Patroling Vehicle	0.047390
Type of Accident	0.037240
Causes	0.036256
Nature of Accident Position	0.033174
Road Feature	0.019866
Position	0.015316
Road condition	0.014166
No of Affected Persons	0.002922
Classification of Accident	0.002156
Rural/Urban	0.000736

Figure 4: Influencing Factors for Weather Conditions

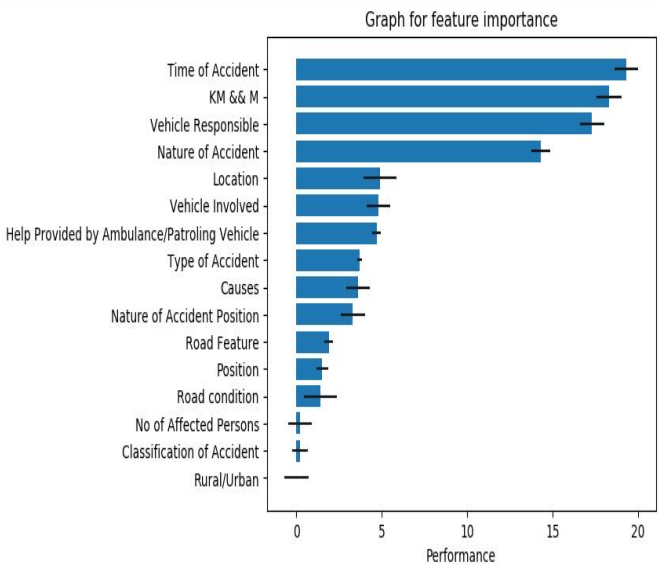

Figure 5: Graph for Feature Influencing Factors

5. CONCLUSION AND FUTURE SCOPE

Road Accidents are caused about by different elements. A progressively expansive examination of the road mishap can be made which can help improve the predictions, making them more accurate. Likewise this is very efficient path than the prior methodology which didn't cover wide scope of elements making it less compelling in current accident prediction. We made the examination much more extensive, remembering the consistently flooding of accidents. This proposed strategy utilizes the information from government information stage from 2012 to 2019 with 19 factors.

The key discovering rose up out of this investigation is the proficiency of five ML techniques is to build classifiers which are accurate and reliable. This incorporates LR, KNN, NB, DT and RF on the features such as Weather condition, Causes, Road Features, Road Condition, and Type of Accident. In view of confusion matrix, the outcomes shows that DT exhibited better execution for every one of the components, Weather condition, Causes, Road Features, Road Condition, Type of Accident, with 99.4%,97.8% , 99.5% , 99.1% , 73.2% exactness than LR with 94.9%, 57.4% , 93.4%, 91.9%, 44.7%; NB with 89.2%, 48.2%, 82.5%, 82.3%, 40.0%; KNN with 94.9% ,64.7%, 93.3%, 91.9%, 53.2%; and RF with 94.9%, 55.1%, 93.4%, 91.9%, 49.9% accuracy. This investigation can assist us with discovering how every component is impacting every accident variables and this gives safe driving recommendations to limit the accidents.

Future scope of this work would be assortment of information directly from the camera introduced about road side which alarms the driver about the distance and probability for event of an accident and if any accident occurred that straightforwardly gives ready sign to the closest police headquarters and the emergency clinic for the salvage activity.

REFERENCES

[1] Rabia Emhamed Al, Keneth Morgan Kwayu, Maha Reda Alkasisbeh, Abdulbaset Ali Frefer, **Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity**, 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019, IEEE.

[2] Liling Li, Sharad Shrestha, Gongzhu Hu, **Analysis of Road Traffic Fatal accidents Using Data Mining Techniques**, Jun 2017, pgno. 363-370, IEEE. https://doi.org/10.1109/SERA.2017.7965753

[3] Md. Shahriare Satu, Sharif Ahamed, Faruk Hossain, Tania Akter, Dewan Md. Farid, **Mining Traffic accident Data of N5 National Highway in Bangladesh Employing Decision Trees**, , Jan 2018, pgno. 722-725 IEEE.

[4] Ms. E. Suganya, Dr. S. Vijayarani, "**Analysis of Road Accidents in India Using Data Mining Classification Algorithms**", Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017), 2017, pgno. 1122-1126, IEEE. https://doi.org/10.1109/ICICI.2017.8365315

[5]Priyanka A. Nandurge, Nagaraj V. Dharwadkar, **Analyzing Road Accident Data using Machine Learning Paradigms**, International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), 2017, pgno.604-610, IEEE. https://doi.org/10.1109/I-SMAC.2017.8058251

[6] Shristi Sonal, Saumya Suman, **A Framework for Analysis of Road Accidents**, Proceedings of 2018 International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR), 2018, pgno.1-5, IEEE. https://doi.org/10.1109/ICETIETR.2018.8529088

[7]Tadesse Kebede Bahiru, Prof. Dheeraj Kumar Singh, Engdaw Ayalew Tessfaw, **Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity**, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018), 2018,pgno.1656-1660, IEEE.

[8] Vidyullatha, P., and D. Rajeswara Rao. **Machine learning techniques on multidimensional curve fitting data based on R-square and chi-square methods**, International Journal of Electrical and Computer Engineering 6.3 (2016): 974. https://doi.org/10.11591/ijece.v6i3.9155

[9] Greeshma, L., and G. Pradeepini. **Mining Maximal Efficient Closed Itemsets Without Any Redundancy,** Information Systems Design and Intelligent Applications. Springer, New Delhi, 2016. 339-347. https://doi.org/10.1007/978-81-322-2755-7_36

[10] R. Bhimanpallewar, **A Machine Learning Approach to Assess Crop Specific Suitability for Small Marginal Scale Croplands**, 2017,12, 23, pgno.13966-13973.

[11] Sakhare, N.N, Sagar Imambi, S, **Performance analysis of regression based machine learning techniques for prediction of stock market movement**, International Journal of Recent Technology and Engineering, 2019, vol.7.