



# Road Accident Analysis And Severity Prediction in Kerala using Machine Learning Techniques

Tintu Varghese, Aby Rose Varghese, Shy Mary Abraham

\*<sup>1</sup> Assistant Professor, Kristu Jyoti College of Management and Technology  
Kottayam

tinta.varughese@gmail.com<sup>1</sup>, 9188086651

<sup>2</sup> Assistant Professor, Kristu Jyoti College of Management and Technology  
Kottayam

abyrosev@gmail.com<sup>2</sup>, 9495849267

<sup>3</sup> Assistant Professor, Kristu Jyoti College of Management and Technology  
Kottayam

shyma85@gmail.com<sup>3</sup>, 9747517909

## ABSTRACT

In recent years, road traffic accidents have become a global problem and are listed as the ninth leading cause of death in the world. Due to the large number of road traffic accidents every year, this has become a major problem in Kerala. Allowing its citizens to die in road accidents is totally unacceptable and sad. Therefore, to deal with this overwhelmed situation, precise analysis is required. This research paper aims to analyze the traffic accidents in Kerala in more depth by using machine learning methods to determine the intensity of the accident. We have also identified those important factors that have a significant impact on road accidents, and provided some useful suggestions on this issue. Through the use of decision tree, K-nearest neighbor (KNN) and Naive Bayes, three supervised learning techniques are used for analysis, and the severity of the accident is divided into three categories: fatal, grievous and minor. Finally, the best performance is achieved by *Naïve Bayes*.

**Keywords:** Road Accident, Accident Severity, Decision Tree, Machine Learning

## I. INTRODUCTION

Road accidents have been proven to be one of the main causes of serious injuries and have been increasing these years. Compared to a few years ago, the number of vehicles on the road has almost doubled. The accident rate hit a record high; thus causing huge losses to health, finance and property. Although various laws and safety measures have taken effect, and there is always the possibility of accidents occurs for various reasons. Driver negligence, driver recklessness, road conditions, weather Conditions, driving skills and many other factors affect the safety of vehicles and vehicles surroundings.

The main goal of this research paper is to analyze the road Accident and determine the severity of the accident using advanced machine learning technology. In this research paper, the authors carry out traffic accident analysis by applying three advanced methods and the most popular machine supervised learning technology study because they have been proven to be accurate in the field. Those ones the methods are decision tree, K-nearest neighbor (KNN) and Naive Bayes.

In this paper, the authors grouped the intensity of an accident into three classifications Fatal, Grievous, and Minor. Furthermore to arrange the seriousness of any

traffic accident in these three classes, eleven fundamental factors that influence the greatest number of accident in kerala have been chosen as the component. Among these three procedures best execution is accomplished by *Naïve Bayes* also, its exactness was 80%. The essential objectives of this paper are to examine the contributory variables for road accidents and to group the seriousness level of accidents in Kerala.

## II. RELATED WORK

The connection between the accident cases and seriousness of accident has been analyzed. The result of the examination shows the reason for the accident, the environmental issues, citizen responsibility, vehicle type, and accident time. Log-linear model, driver attributes, pedestrian characteristics, street traffic, vehicle typologies have been considered in the task

so this gives a reasonable thought of what is influencing the accident in the school regions.

It is practically not possible to improve the road safety facilities at all junction points, therefore black spot identification method is analyzed and considered to identify the accident-prone zone. It is an efficient method to decrease road accidents by examines the cause of black spots. The Bayesian network was the best model to effectively identify road accident black spots. Accident occurring spot is known as black spot.[1] Implemented safety measures for highway transportation has more priority in transportation safety production. Proposed method for the estimation of crash rate and relative risk value is based on driver and passenger traits.[2]

A statistical analysis of serious traffic accidents based on the direct cause, accident type, accident responsibility, etc is used to provide support to strengthen the transportation industry management system. Many road safety acts were amended by the government authorities to reduce the severity in lane accidents. The Accident Prediction based on data mining helps us to analyze the different types of circumstantial information with slight complications [3]. Suggestions made for safe driving are based on rules and regulations, classification model, and clusters obtained.[4] Use of an Android Application called Crash prediction system(CPS) which analyzes the user's behavior and predicts whether the user is fit to ride or not.[5] Common problem in the crash analysis data is heterogeneous in nature. During the analysis of the collected data, heterogeneity has to be reviewed as some of the correlated data may be unsolved. Cluster Analysis is a key data mining technique that helps us to achieve the initial task like segregating the different types of road accidents. Regression Analysis is the most acceptable technique in the accident analysis, as the link between the crashes and factors affect them.[6]

The analysis results in the preventive and risk factors of motorcycle accidents. There are different injury patterns during the crashes like the general pattern.(head injuries, Lower-extremity injuries). Road crashes, Street transport frequently speaks to the most serious danger that we are presented to during an ordinary working day.[7] As indicated by the information from dataset, the death cases normally occur during weekend days between 6:00 a.m. to 6:00 p.m. that includes (HGVs)

Heavy Goods Vehicles. Bus accidents happen frequently on weekend days. On weekend days, during early morning and late evening usage of buses are more during the peak hours.

A few examinations show that organization vehicles regularly are associated with accidents at high speeds.[8-9] From the analyzed traffic accident dataset, data mining techniques helps to decrease the death rate. The database used for the road safety purpose helps to drop down the fatality rate by performing road safety events and creating awareness programs throughout the district. The information collected from the database gives the brief description about the accident like road conditions, gender, person involved, health issues and other information, that will be convenient for the evaluation and for fetching additional data for proper conclusion of the case.[10] The International Road Traffic and Accident Database (IRTAD), GLOBESAFE, site for ARC networks are the best assets to gather accident information. For the Investigation purpose, Information are taken from the internet to design and create the map. It helps to characterize the data as well as gives cautioning sound or video. Information from the dataset (Year:1988 to 1993) results describes that the risk of accident for drivers was around 9.5 fatalities per 100 million man hours and it was nearly 3 for other professions.[11] Some of the classification techniques help to foretell the severity level of lane accidents. Naive Bayes classifier,

Decision Tree classifier, AdaBoostM1 Meta classifier, PART Rule classifier, and Random Forest are looked at for characterizing the severity level road accidents. The end-product reveals that Random Forest technique beats the other four models.[12] Study of severity levels in accident helps to scrutinize the link between the severity level, bunch of parameters which include driver details, vehicle details, road pattern and cause of accidents. The impact of different factors on seriousness of the injury like day, time, speed limit, traffic details, climatic conditions, and driver details also used in the investigation.[13-15].

The study of data helped to evaluate the attributes of road accidents seriousness information and approached most common data used for the analysis. The logistic regression model helps to evaluate the autonomous commitment of accident and vehicle characteristics encouraged by the drivers. The variables, for example, driver's age, gender (female), liquor concentration are more prominent than 0.30, drivers without seatbelts, drivers health issues, poor quality of driving, speed limit more than 65 to 70 mph, and more established vehicles were related to greater chances of accident. When the wheel size is increased by 25 cm helps to decrease the death rate.[16]

Few models help to recognize major cause of accident that improves the chances of severity in accidents. The three levels of severity are minor, grievous, or dead injuries. They have been divided into three classes by their severity levels – for example deaths, severe and minor injuries, and they used negative binomial (NB) models on every class of road setbacks independently, bringing about three independent univariate models.[17] The authors in the paper , utilized logistic regression dependent on the vehicle condition after an accident to sort out the effect of road surrenders on accident severity.

### III. PROPOSED MODEL

There are three types of machine learning algorithms-supervised and semi-supervised learning, unsupervised learning, and reinforcement learning [18]. Among these three broad categories of machine learning classification approaches the supervised learning approach has been used in this research paper because of its competency in modeling and regulating dynamic systems. Here, the authors have used the three most popular machine learning techniques for road accident analysis [19]. Those are *Decision Tree*, *KNN* and *Naïve Bayes* . Figure 1 presents a simple view of the overall working process.

#### A. Preparation of dataset

Precise and broad accident information records are the most significant and prime need to improve execution by applying machine learning approaches. However, getting an ideal and 100% exact dataset is very difficult. Consequently, to deal with information dependent on the need the authors have adhered to the accompanying guidelines.

##### 1) Data Collection:

For the exact prediction of the seriousness of accident, an extensive number of traffic accident records with full data is needed to prepare by utilizing the proposed approaches. In this research work, the authors have gathered a data set from official website of Kerala

Police. We split our whole data set into two sections Training Data set and Test Data set. 70% of the entire dataset has been picked arbitrarily by utilizing a python library as a preparation informational index and the excess 30% has been utilized as our test dataset. We have utilized the 70-30 proportion for splitting dataset due to its demonstrated precision.

#### 2)Data Pre-Processing:

In this data set, all the accident records were composed with formal words. We appropriately arrange this all out data set dependent on the component. Altogether, we have discovered 34 components that influence past accidents somehow or another. First and foremost, we methodize all accident records by utilizing these 34 features. From that point onward, for some, accident records, we have discovered 8.7% missing qualities in the absolute dataset and the particular 11 features has 1.65% missing value. As these missing qualities can influence the presentation, by virtue of this, we have applied a strategy by utilizing the mean value of that include section to give a sum where it is required. We utilize this technique as there presents no extreme value which can influence the mean.

### IV. THE METHODOLOGY OF PROPOSED APPROACHES

- a) Decision Tree: For classification problems, the decision tree is extensively used the supervised algorithm. Decision Tree is a Supervised learning method that can be utilized for both classification and Regression problems, yet generally it is liked for solving Classification problems. It is a tree-structured classifier, where inner nodes address the features of a dataset, branches address the decision rules and each leaf node addresses the result. In a Decision tree, there are two hubs, which are the Decision Node and Leaf Node. Decision nodes are utilized to settle on any decision and have numerous branches, though Leaf nodes are the output of those decision and don't contain any further branches. The decision or the test are performed based on features of the given dataset.

b)KNN: KNN is a classification algorithm which is based feature similarity. It analyzes the data and measure the distance and similarities between data and cluster them based on K values. Distance is calculated in many ways, for this research, we used Euclidean distance measurement.

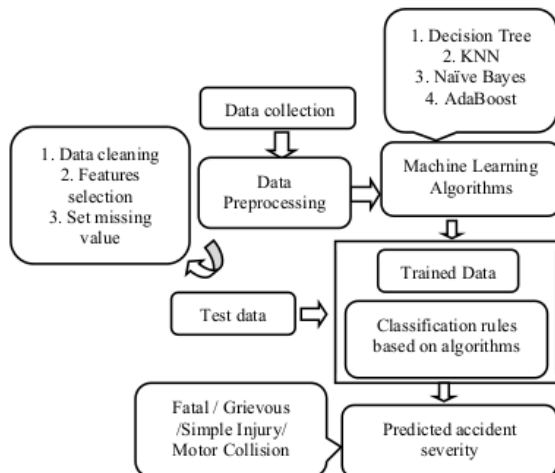


Figure 1. The working mechanism of proposed approaches

The class of new input data is classified by calculating the distance between the clusters and assigned it to the closest one. Euclidean distance is calculated by, [20]

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \dots + (x_n - x'_n)^2}$$

c) Naïve Byes: Naïve Byes is another classification technique based on Bayes theorem. It predicts the probability of different class based on several attributes and assigns the new class to the highest probability. The algorithm working rule:

$$\text{Posterior probability, } P(a|b) = \frac{P(b|a) * P(a)}{P(b)}$$

The posterior probability is mainly the probability of “a” being true given that “b” is true.[21]

## V.RESULT ANALYSIS AND DISCUSSION

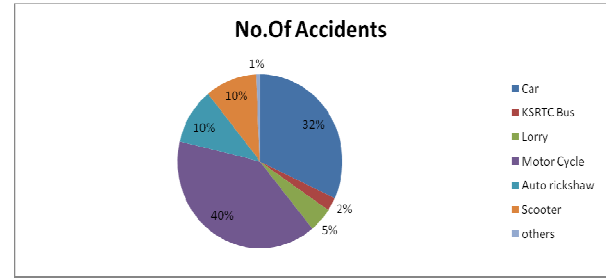
Number of accidents at different road patterns and various contributory factors that affects the accident has been identified from the historical data in Kerala. we have determined the performance of each algorithm, for three accident severity classes (Fatal / Grievous //minor). Naïve Bayes achieve the high accuracy among these four approaches, and their accuracy is 80% (Table II).

Algorithm	Accuracy(%)	Precision(%)	F1 score(%)
Decision Tree	73	70	71
KNN	68	67	67
Naïve Bayes	80	75	74

Table II: Severity prediction results of algorithms

we analyzed the effect of vehicle types on a traffic accident, and it is seen that the Motor Cycle is responsible for 40% traffic accident and cars for 32% (Figure 3).

Figure 3. Effect of vehicle type on accident



In Figure 4, it is found that most of the accidents have happened on the National road.

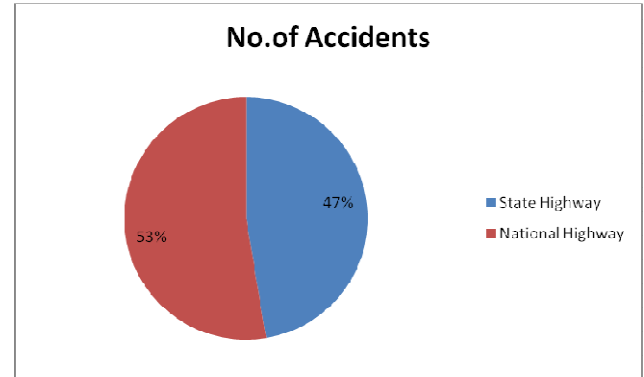


Figure 4. Effect of road class on accident

We analyzed the cause of accident and found that around 74% cases are reported on fault of driver .

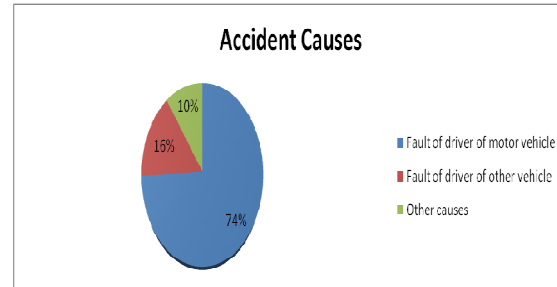


Figure 5:Accident Causes

Thereafter, we try to figure out, at which time traffic accidents occur more often. Based on the result, it is excogitated that in both urban area and rural area in early morning (6am-9am) accident rate is very high compared to other time (Figure 6).

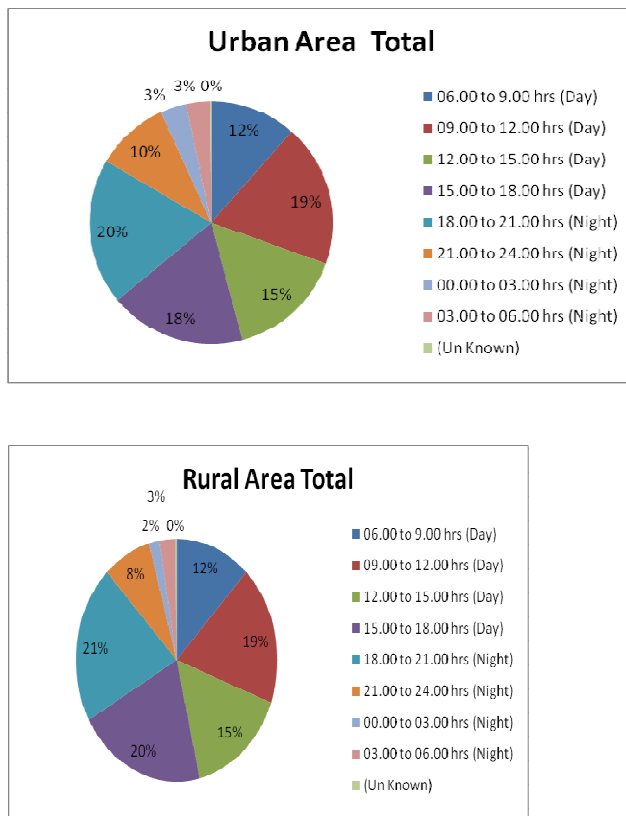


Fig 6:Effect of time on accident

## VI. CONCLUSION

Losses in road accidents are agonizing to the society as well as a developing country like us. Thus, it's gotten an fundamental necessity to regulate and arrange traffic with an advanced system to diminish the number of street accidents in our country. By taking precautions, based on

prediction or warnings of a classy system may prevent traffic accidents. Additionally, it's an important requirement for our country currently, to handle this circumstance where consistently so numerous individuals were killed in road accident and step by step this rate is getting increased. The execution of machine learning is a n effective way to take a particular decision with the experience to affect the present circumstance also, the findings of the analysis part (Figure 2-Figure 6) can be recommended to traffic experts for lessening the number of accidents. we will utilize proposed ways to affect perform machine learning here in sight of their demonstrated and better precision to foresee traffic accident severity.

Also, to make it more reliable, we will attempt to make a recommender framework by utilizing these methodologies that can give prediction to the traffic accident and can caution the road user. Later on, it will

be our attempt to make a versatile application by executing this technique to give a precise forecast to the client and make it extremely valuable and useful too.

## VII. REFERENCES

- [1] C. Zhang, Y. Shu, and L. Yan, "A Novel Identification Model for Road Traffic Accident Black Spots: A Case Study in Ningbo, China.," IEEE Access. Vol. 7, pp. 140197–140205, 2019.
- [2] T. Jian, L. Zhi-qiang, X. Jian-feng, and G. Hong-yu, "Cause Analysis and Countermeasures of fatal Traffic Accidents on Road Passenger Transportation Based on Typical Cases.," In: 2019 5th International Conference on Transportation Information and Safety (ICTIS). pp. 951–955. IEEE (2019).
- [3] Y.-K. Ki and D.-Y. Lee, "A Traffic Accident Recording and Reporting Model at Intersections.," IEEE Transactions on Intelligent Transportation Systems. Vol. 8, No. 2, pp. 188–194, 2007
- [4] L. Li, S. Shrestha, and G. Hu, "Analysis of road traffic fatal accidents using data mining techniques.," In: 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA). pp. 363–370. IEEE (2017).
- [5] E. Grant, P.M. Salmon, N.J. Stevens, N. Goode, and G.J. Read, "Back to the future: What do accident causation models tell us about accident prediction?," Safety Science. Vol. 104, pp. 99–109, 2018.
- [6] S. Kumar and D. Toshniwal, "A data mining framework to analyze road accident data.," Journal of Big Data. Vol. 2, No. 1, p. 26, 2015.
- [7] T.-O. Nævestad, R.O. Phillips, and B. Elvebakk, "Traffic accidents triggered by drivers at work – A survey and analysis of contributing factors.," Transportation Research Part F: Traffic Psychology and Behaviour. Vol. 34, pp. 94–107, 2015.
- [8] D. Deva Hema, S. Nandhini, S. VisnuDharsini, and J. ShivaNandhini, "Intelligent Speed Control in Motor Bikes for Accident Prevention Using Internet of Things.," Journal of Computational and Theoretical Nanoscience. Vol. 17, No. 4, pp. 1600–1605, 2020.
- [9] S. Newnam, B. Watson, and W. Murray, "Factors predicting intentions to speed in a work and personal vehicle.," Transportation Research Part F: Traffic Psychology and Behaviour. Vol. 7, No. 4–5, pp. 287–300, 2004.
- [10] F. Fan and Y.-B. Qian, "Analysis of Factors Affecting the Severity of Car Accidents at iNtersections Based on Cumulative Logistic Model.," In: 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC). pp. 310–313. IEEE (2017).
- [11] G. Yannis, A. Dragomanovits, A. Laiou, et al., "Use of Accident Prediction Models in Road Safety Management – An International Inquiry.," Transportation Research Procedia. Vol. 14, pp. 4257–4266, 2016.
- [12] S. Krishnaveni and M. Hemalatha, "A Perspective Analysis of Traffic Accident using Data Mining Techniques.," International Journal of Computer Applications. Vol. 23, No. 7, pp. 40–48, 2011.
- [13] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks.," Accident Analysis & Prevention. Vol. 38, No. 3, pp. 434–444, 2006.
- [14] H. Manner and L. Wünsch-Ziegler, "Analyzing the severity of accidents on the German Autobahn.," Accident Analysis & Prevention. Vol. 57, pp. 40–48, 2013.
- [15] P.T. Savolainen, F.L. Mannering, D. Lord, and M.A. Quddus, "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives.," Accident Analysis & Prevention. Vol. 43, No. 5, pp. 1666–1676, 2011.
- [16] M. Bédard, G.H. Guyatt, M.J. Stones, and J.P. Hirdes, "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities.," Accident Analysis & Prevention. Vol. 34, No. 6, pp. 717–727, 2002.
- [17] R.C. Gray, M.A. Quddus, and A. Evans, "Injury severity analysis of accidents involving young male drivers in Great Britain.," Journal of Safety Research. Vol. 39, No. 5, pp. 483–495, 2008.
- [18] M. A. A. Mamun, J. A. Puspo and A. K. Das, "An intelligent smartphone based approach using IoT for ensuring safe driving.," 2017 International Conference on Electrical Engineering and Computer Science (ICECOS), Palembang, 2017, pp. 217–223.
- [19] T. Adhikary, A. K. Das, M. A. Razaque, A. Almogren, M. Alrubaian, and M. M. Hassan, "Quality of Service Aware Reliable Task Scheduling in Vehicular Cloud Computing.," Mobile Networks and Applications, Volume 21, Issue 3, pp.482–493, June 2016.
- [20] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," Annals of Translational Medicine (2016), vol 4, issue no. 11, pp.218–218.
- [21] I. Rish, "An Empirical Study of the Naïve Bayes Classifier.," IJCAI 2001 Work Empir Methods Artif Intell (2001), vol 3.