Ritchie's Predictive Model of Accidents using Machine Learning Algorithm

Settyrajeshwar Rao[1]

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, TamilNadu, India

Banalata Bhunia[2]

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, TamilNadu, India

Divyansh Sahu[3]

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, TamilNadu, India

Rajkumar.R[4]

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, TamilNadu, India

Damandeep Kaur[5]

DaDa Analtix Pvt Ltd,

Jolly Masih[6]

Prestige Institute of Engineering Management and Research,

Abstract:

In previous years the road accident became a global problem which is the ninth important cause of death in the world. Due to the many numbers of road accidents in every year, it has become a major problem in India. It is fully unacceptable and sorrowful to allow its citizens to be killed by road accidents. So, to handle this huge problem, an accurate analysis is required. This paper has been done to analyze the accident on the road by using machine learning approaches in India, which are categorized by the following four type accidents -1:Unscathed 2: Killed 3: Hospitalized wounded 4: Light injury. We find out those important factors that have a clear effect on road accidents and provide some beneficent suggestions regarding this issue. Analysis has been done using data cleaning and classification and random forest. Big Data is a field that treats approaches to break down, methodically remove data from, or generally manage informational indexes that are excessively huge or complex to be managed by conventional information handling application programming. Information with numerous cases (lines) offer more noteworthy measurable power, while information with higher intricacy (more qualities or sections) may prompt a higher false disclosure rate. We wanted this predictive model to be used in a specific context, so we thought about a scenario for the cases which is needed. Our idea was that this model could be used by emergency services when they receive an emergency call to report an accident. Here the information given by the caller, which is usually the place of the accident and maybe more details, as well as using information that can be inferred from this, they would be able to predict the severity of the accident. The goal is to predict the Severity of Accidents.

Keywords: Data cleaning, Classification, Random forest.

1. Introduction

The Road Accident is the most unpredictable and unexpected thing to occur to a road user, though the users happen quite often. Unfortunately, the monotony of users rise of road accidents in all over India. The accidents have a massive impact on society as well as in the economy of our country as there is a huge cost of fatalities and injuries. According to an inspection, which was held recently , on an average 12,000 lives have been taken by road accidents and it leads to almost 35,000 injuries . This record indicates that every day, approximately 32 people were killed in road accidents in India and it is gradually increasing. Besides this, according to WHO (World Health Organization) the economic cost of road accidents to a developing country like US, is 2-3% of GDP. This is a significant loss for a country like ours. Moreover, to reduce this loss has become a great matter of concern for our country now. We were trying to solve this problem using *Big Data, Data mining with ML algorithm.* "Big Data" is a field that treats approaches to break down, methodically remove data from, or generally manage informational indexes that are excessively huge or complex to be managed by conventional information handling application programming. Information with numerous cases (lines) offer more noteworthy measurable power, while information with higher intricacy (more qualities or sections) may prompt a higher false disclosure rate. We will be mainly focusing on Data Mining which is a very important field of Big Data. Data mining is a procedure which discovers helpful examples of enormous measures of information. This paper is about the data mining procedures and calculations and a portion of the associations which have adjusted information mining innovation to improve their organizations and discovered brilliant outcomes. Data mining is a process which finds useful patterns from large amounts of data. This paper about the data mining approaches and the algorithms and some organizations which have accepted data mining technology for improving their businesses and it found excellent results. In this paper, we are going to study and analyze and review various advancements in the field of data mining and we will focus on challenges and issues which are currently unresolved.

The advancement of Information Technology has produced an enormous measure of databases and immense information in different zones. The exploration in databases and data innovation has offered a way to deal with store what's more, control this valuable information for further basic leadership. Information mining is a procedure of extraction of helpful data and examples

from enormous information. It is likewise called the information disclosure process, learning mining of information, information extraction or information/design examination. Data mining is an intelligent procedure that is utilized to look through huge measures of information so as to discover valuable information. The objective of this method is to discover designs that were beforehand obscure. Once these examples are discovered they can further be utilized to settle on specific choices for improvement of their organizations.

Three steps involved are:

a) *Exploration*: In the initial step of exploration, information is cleaned and changed into another structure, and significant factors and after that nature of information depending on the issue are resolved.

b) *Pattern Identification***:** When information is investigated, refined and characterized for the particular factors the subsequent advance is to shape the design distinguishing proof. Recognize and pick the examples which make the best forecast.

c) *Deployment***:** Examples are conveyed for wanted result.

2. Literature Review

In this paper [1] the author Debela Deme Jime and his team told that the Road traffic accident is one of the most common reasons of deaths and injuries nowadays in the world. Ethiopia is the leading country in road traffic accidents. This study was conducted to identify the major causes of road traffic accidents. To solve this problem descriptive and inferential statistical analyses were used to identify causes of road accidents. The multinomial probit regression model and the accident severity value were used to assess the causes of traffic accidents and identify the black spot region . The study revealed that on average there are 417 crashes reported since 2014Sept to 2016 Feb. As a result of the crashes over 672 accidents were registered . And caused 285 human injuries and 387 property damage of the total human injuries.There are 37 were fatal, 65 and 183 were serious and slight injuries.

In paper [3], the authors proposed a method to detect the possibility of accidents on the road by using vision-based techniques in the context of accidents. They used roadside video data as their learning materials and it achieved 85% accuracy in particular situations. The researchers in this paper

[4], here analyzed 892 traffic accidents. By using many decision tree induction algorithms and have tried to find out the traffic accident patterns. They also take out the rules for the

trees to decrease road accidents on roads. There exist many advanced and beneficial research works about this field.

In this paper [5], the authors proposed the status of the road accident happens by using machine learning approaches. They applied the CART algorithm which determines the risk of the accident and achieved above 81.5% of accuracy. In this paper [6] Two authors applied machine learning algorithms like k- mean clustering and association rules mining, these two techniques of data mining to figure out the major factors affiliated with traffic accidents. In the paper [7], author Richard Antoni Gosno highlighted that a hybrid supervised machine learning model was proposed for construction site accident classification. A natural language processing was adapted to pre – process the text data prior to classification process, then a searching algorithm was integrated to optimize the parameters of the machine learning model and concluded that the classification results can be used to aid safety assessment of construction projects. In the paper [8], the author Arun Venkat along with his team predicted that in 2030, traffic accidents will be the fifth leading cause of death globally. The actual cause of road accident is difficult to determine due to combination of different characteristics like mental condition of a driver, weather conditions, violation traffic rules and traffic with this the placement of machine learning classifiers have replaced traditional data mining techniques which were used for generating high results and accuracy. According to the author [9], road accidents have become a major problem in Bangladesh, the author used a precise analysis for predicting road accidents and the analysis were done by using Decision Tree, KNN (K- Nearest Neighbors) and the best performance was achieved by Ada Boost. Sangil K [10]won the random forest was applied to road accident data and it yielded the most accurate data, then three data sets containing geometry data, precipitation data and traffic accident data were used and for analyzing these data random forest algorithm was used as the machine learning tool and the results showed that traffic accident severity correlated with Dg and this work improves the ability of decision making, policy maker and transportation safety designers to take action for traffic safety control.

3. Methodology

Proposed Model: Proposed To implement this model we used three algorithms. These are: Classification, Data Cleaning and Random forest. In classification algorithms we use trained data to get better boundary conditions that are used to determine each target class. Once the boundary conditions are determined, then the next task is to predict the next target .

There are many classifications algorithms these are Classifier, Classification model, Feature, Binary Classification, Multi Label classification. Classification is a type of supervised learning.in which element belongs to the specific class and is best used when

the output has finite and discrete values. Classification predicts a class for an input variable as well.

The second algorithm is *Data Cleaning Algorithm*. In this algorithm a process is used to determine inaccurate or incomplete data and then improving the quality through correction and detecting the errors. It is one of the important parts of machine learning. It plays a significant part in building a model. Data cleaning is the process in which we are detecting and correcting corrupt or inaccurate records from a data set, tables, or databases and it refers to identifying inaccurate incomplete, incorrect, or irrelevant parts of the data and then we replace, modify, or delete the dirty or coarse data. It has many data qualities like validity, accuracy, completeness and consistency and integrity also.

The steps involved in Data Cleansing is –

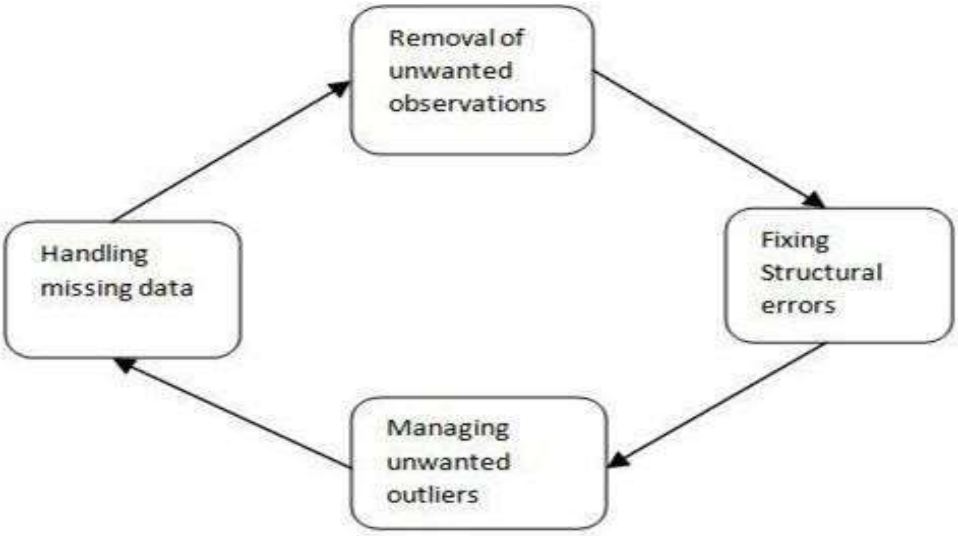

Fig. 1 Data Cleaning

The third algorithm is *random forest.* Random forest is the machine learning analysis which creates multiple decision trees and then merges all together to get an accurate and stable prediction. Random forest algorithms can be used for both classifications and regression tasks. It provides higher accuracy. Random forest classifiers will handle the missing values and maintain the accuracy of a large proportion of data. If there are more trees, it won't allow over fitting trees in the model. This algorithm for random forests applies the general technique of bootstrap aggregating or bagging, to tree learners.

Given a training set $X = x_1, \ldots, x_n$ with

responses $Y = y_1, \ldots, y_n$, bagging repeatedly (B times) selects a random

Sample with replacement of the training set and fits trees to these samples:

For b = 1, ..., B:

Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.

1. Train a classification or regression tree f_b on X_b, Y_b.

It has the power to handle a large data set with higher dimensionality predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

N estimate the uncertainty of the prediction can be made as the standard deviation of the production from all the individual regression tree on x.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

The number of sample B is a free parameter., a few hundred to several thousand trees are used here to depend on the size and nature of the training set. An optimal number of trees B can be found using cross validation . The mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample. The training and test error end to level off after some number of trees have been fit.

The first step in measuring the variable is

1. Predict whether someone will buy computer accessories. (Target class : yes or No)
2. Gender Classification from hair
3. Dn = {(Xi , Yi)} ˆ i = 1 forest to the data. is to fit a random
4. length(Target class: Male or Female)

Preparation of Data sets: Accurate and large accident data records are very important and primary need to get better performance by applying the algorithm. But, it is very difficult to get a 100% accurate dataset.it is very challenging. So process data according to the need, we are following the instructions.

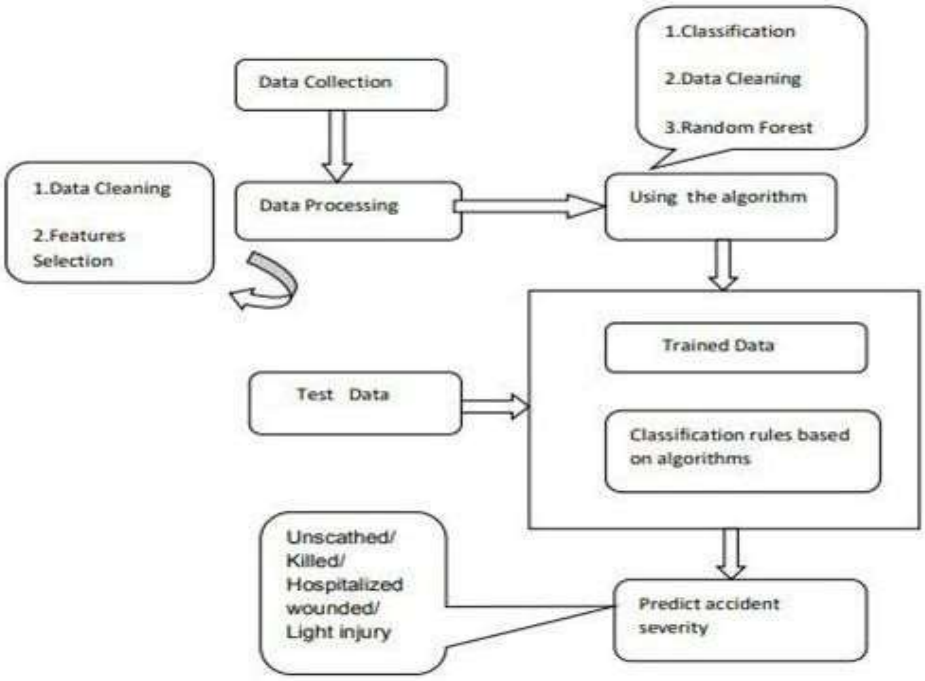

Fig.2 The working mechanism of proposed approaches

Data Colloection: For the accurate prediction of the severity of accidents, a considerable number of accident records with full information is needed to train by using the proposed algorithm. In this prediction we collected data from the Accident in France 2005 to 2016 and in this data set there are many entities present like accident_id, Day of the accidents, month and year of the accidents and time of the accident in hours and minutes.

Data Pre – Processing: In this collected dataset, all the accident records were written with formal words. We organize this total dataset properly according to feature. In total, firstly, we methodize all accident records by using features. After that, for many accident records, we have found missing values in the total dataset and the selective features have missing values also. As these missing values can affect the performance, on account of this, we have applied a method by using the mean value of that feature column to provide an amount where it is required. We use this method as it presents no extreme value which can affect the mean.

FEATURE SELECTION:

Working with a large number of features may affect the performance due to training time increases exponentially with the number of features .with the increasing number

of features it has also the risk of overfitting. So to get an accurate prediction and feature selection is a difficult factor here. To obtain the most necessary features, we operate an experiment by applying three different algorithms of feature selection. here there are four features of data

5. Results and Discussions:

In this paper, to evaluate the performance of the proposed approaches, we apply three different algorithms these are Classification algorithm, Data cleaning and random forest. Under the random forest we merge multiple decision trees. So we get this output after applying the merging.

After all the preprocessing done we able to merge all the decision trees into a single dataset, with the feature "Num_acc", the Accident ID, as the primary key.

	Num_Acc	mois	jour	hrmn
count	8.399850e+05	839985.000000	839985.000000	839985.000000
mean	2.010011e+11	6.679437	15.594687	13.559365
std	3.458009e+08	3.389489	8.750201	5.411096
min	2.005000e+11	1.000000	1.000000	0.000000
25%	2.007000e+11	4.000000	8.000000	10.000000
50%	2.010000e+11	7.000000	15.000000	14.000000
75%	2.013000e+11	10.000000	23.000000	18.000000
max	2.016001e+11	12.000000	31.000000	23.000000

Fig 3 : Merging Decision Trees

After importing the random forest classifiers it takes the time 61.2252068519.In feature selection after deleting the lowest importance features: "surf", "lum", "atm", "int", "nbv", "circ".

Then we get the accuracy: 0.6639161297, time taken: 52.39916706.

Testing for the accuracy means that the model provides true predictions of 71% of the times.Having almost the same results than with the validation set is reassuring about the fact that our model is not overfitting. It is an interesting measure to evaluate our model but in the context, other measures might be useful as well. The accuracy is 0.7112805585814033 and the time taken : 6.3936295509338.

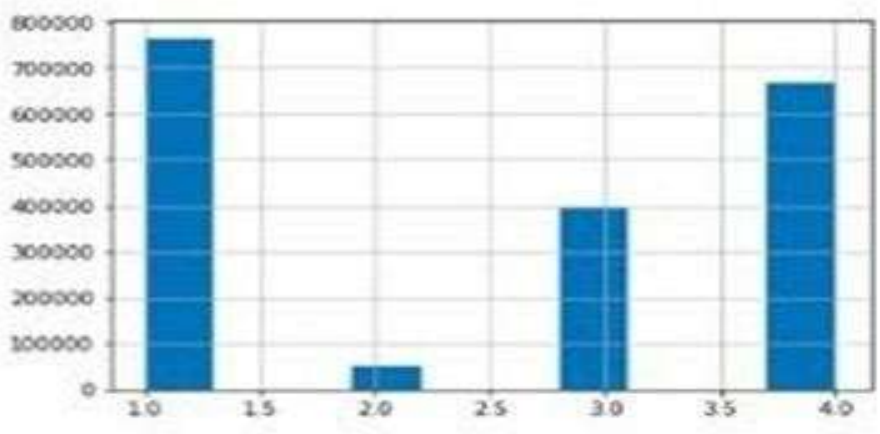

Fig 4: a. light injury b. Death c. Hospitalized wounded d.unscathed

After cleaning the data using clustering algorithm we get this final output table. This is the table for this huge data set after applying the algorithm.

	mois	jour	hrmn	lum	agg	int	atm	dep	catr	circ	nbv	surf	gra
0	2	1	14	1	2	1	8	59	3	2	2	1	1
1	3	16	18	1	2	6	1	59	3	1	2	1	1
2	7	13	19	1	1	1	1	59	3	2	2	2	1
3	8	15	19	2	2	1	7	59	4	2	2	1	1
4	12	23	11	1	2	3	1	59	4	2	2	1	1
5	12	23	11	1	2	1	7	59	3	2	2	1	1
6	5	1	11	1	2	1	7	59	3	2	2	1	0
7	5	14	19	2	1	1	1	59	3	2	2	1	1
8	9	23	19	1	2	1	1	59	4	2	2	1	1
9	12	30	10	1	1	1	9	59	4	2	2	7	1
10	1	25	8	2	2	1	8	59	3	2	2	2	1
11	1	28	18	3	1	1	1	59	3	2	2	1	1
12	2	5	16	1	2	1	1	59	3	2	2	1	1
13	4	17	12	1	2	1	1	59	3	2	2	1	1
14	8	17	19	1	1	6	7	59	3	1	1	1	1

Fig 5: The final output table.

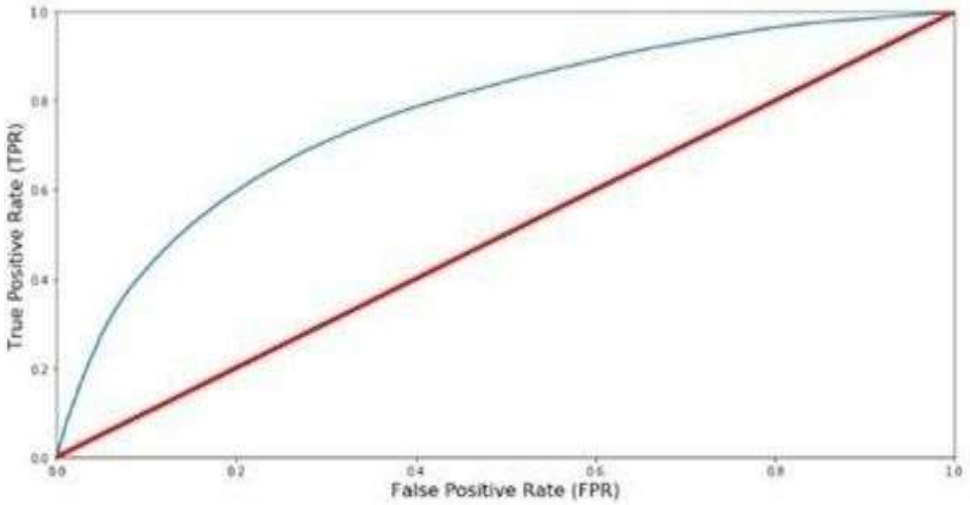

Fig 6: The graph of False positive rate and true positive rate.

We conclude in this project that, the accuracy score to take into consideration that injuries will be only light. Similarly, after knowing this information we can find some resources that can be more helpful to save other people that need it.

6. Conclusion

The losses in road accidents are increasing day by day to the society as well as a developing country. So, it is an essential requirement to control and arrange traffic with the advanced system. To decrease the road accident the traffic rules should be obeyed. By taking the simple precautions which are based on prediction or warnings of a sophisticated system. Hopefully it may prevent traffic accidents. Moreover, it is a primary need for all the country now, to tackle this situation, where every day so many people are killed in traffic accidents and day by day this rate is increasing. Implementation of machine learning is a functional and a great approach which gives an accurate decision with the experience to manage the current situation and helps to find the analysis part. It can be suggested to traffic authorities for reducing the number of accidents and it will tell the injuries type like badly injured or lightly injured so that the patient can go to a nearby hospital accordingly. Here We can use proposed approaches to implement the machine learning here because of their proven and higher accuracy to predict traffic accident severity. However to make it more feasible, we will try to make a recommender system by using these approaches in the near future. This approach can give a prediction to the traffic accident and it can warn the road user.

REFERENCES:

1. T. RahanRahman, T. (2012). Road accidents in Bangladesh: an alarming issue. *The World Bank*.
2. Elahi, M. M. L., Yasir, R., Syrus, M. A., Nine, M. S. Z., Hossain, I., & Ahmed, N. (2014, May). Computer vision based road traffic accident and anomaly detection in the context of Bangladesh. In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)* (pp. 1-6). IEEE.
3. Satu, M. S., Ahamed, S., Hossain, F., Akter, T., & Farid, D. M. (2017, December). Mining traffic accident data of N5 national highway in Bangladesh employing decision trees. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 722-725). IEEE.
4. Bülbül, H. İ., Kaya, T., & Tulgar, Y. (2016, December). Analysis for status of the road accident occurrence and determination of the risk of accident by machine learning in Istanbul. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 426-430). IEEE.
5. Nandurge, P. A., & Dharwadkar, N. V. (2017, February). Analyzing road accident data using machine learning paradigms. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 604-610). IEEE.
6. Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. *Journal of Modern Transportation*, *24*(1), 62-72.
7. Cheng, M. Y., Kusoemo, D., & Gosno, R. A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction*, *118*, 103265.
8. Venkat, A., KP, G. V., & Thomas, I. S. (2020). Machine Learning Based Analysis for Road Accident Prediction. *Machine Learning Based Analysis for Road Accident Prediction (March 7, 2020). IJETIE*, *6*(2).
9. Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K., & Nawrine, F. (2019, June). Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1-5). IEEE.
10. [10] Lee, J., Yoon, T., Kwon, S., & Lee, J. (2020). Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. *Applied Sciences*, *10*(1), 129.