# Predicting Severity of Road Accidents and using various Machine Learning Algorithms

*by* Umang Jain

**Submission date:** 05-May-2020 04:44PM (UTC+0530)

**Submission ID:** 1316507472

**File name:** idents_and_Comparative_Study_of_Machine_Learning_Algorithms.docx (175.91K)

**Word count:** 2633

**Character count:** 13788

# Predicting Severity of Road Accidents and using various Machine Learning Algorithms.

Umang Jain
19MCA0164
SITE, VIT Vellore
Umangj107@gmail.com

Rahul Sharma
19MCA0174
SITE, VIT Vellore
Rahul.Sharma2019@vitstudent.ac.in

Akshit
19MCA0020
SITE, VIT Vellore
akshitpanwar1234@gmail.com

Dr. Shynu P G
Associate Professor
SITE, VIT Vellore
pgshynu@vit.ac.in

*Abstract:* **During the past few years, there has been a tremendous increase in the cases of road accidents worldwide. The causes of such accidents could be anything like the condition of the roads, the climate and visibility conditions at that time as well as the age and health of the vehicle. Also various aspects relating to the driver also affect the chances of happening of road accidents like the mental state of the driver, the physical health and the age of the driver. The increase in the number of accidents has created an enormous amount of data, which could be used to perform analysis and visualizations in order to study the important facts and trends on these data. In this paper we have tried to visualize the dataset as well as build machine learning models for predicting the severity of these road accidents. This task is done involving a number of machine learning algorithms. Furthermore, the efficiency of these algorithms on this dataset is studied comparatively.**

***Keywords: Road accidents, Severity, Logistic Regression, Naïve Bayes, Random Forest, Decision Tree, Prediction, Comparative Study.***

## I. INTRODUCTION

There has been a tremendous increase in the number of road accidents all around the globe generating a huge amount of data in terms of volume as well as variety. Data today is among the most important asset we possess. This huge amount of data has been very beneficial in generating insights about the reasons or causes of these road accidents. The data also helps us know about how severe the accident was and what casualties have been in the mishap.

We have made this as the basis of our project. This paper deals with the problem of predicting the severity of the road accidents happening around the world. This has always been a bias process when it was done manually. Humans commit mistakes and this was a problem in deciding the severity of the accidents. In this paper, we have tried to solve this problem by building a machine learning model to do this task. The vast dataset of Road accidents is taken as input for the model and an efficient model is tried to build which could accurately predict the severity of road accidents without any human intervention as well as human bias.

For the purpose of building an efficient model, we have taken into account a number of machine learning models which seem to fit our purpose. Using each algorithm, a model is built and its efficiency is measured in comparison with the other models, which ultimately helps us find the best fitting model for our purpose of generating the severity of road accidents.

Utmost care is taken to avoid any manual mistakes and data set has been pre-processed according to our need. Attempts have been made to make the accuracy as high as possible and to avoid overfitting or underfitting of the model.

## II. LITERATURE SURVEY

There had been a very few works done on this dataset in the past. A few important ones among them are listed here. The paper [1] focussed on the analysing the factors that affect the road accidents. Furthermore the authors of this paper went ahead to estimate the factors which highly affect the happening probability of these accidents as well as the severity of these accidents, and concluded that the factors including the drugs and medicine contributed majorly to the accidents. The authors of paper [2] used data mining techniques to identify the areas where there are frequent accident cases and studied the road conditions in those areas using statistical tools. They studied different types of roads and concluded which road types contributed majorly to the accident count over the globe. The authors also came up with interesting statistical plots which helped generate important inferences about the road accidents. In paper [3], the authors studied the relationship between the casualty rate and the factors contributing to the happening of the accidents like weather conditions, light conditions, drunk driver at the time of happening of accidents.

Talking about the algorithms used in paper, In paper [4], they used logistic regression model and big data techniques to predict the accident between the ships. As we know logistic regression considered for more accuracy for small dataset and less number of variables it proved helpful. Accident data are classified variables of multi-level and multi-factor. Logistic Regression could be divided in to two or multilevel classifications here it is to classify the independent variables. Through this we can practically calculate the influenced factor responsible for the cause of such. In another work [5] , social data is analysed to find out the create a social data analysis and telling whether a person is more likely to live in south USA or not. The authors achieved the following research using the Logistic Regression and Decision Tree Algorithm. Results shows that with the rightful classification the logistic regression is obtaining much high accuracy then decision tree and neural networks. As a fact described by the models, people have differences between the salaries living in south and north region respectively. People whose father is having more salary tends to live in the south of the USA. In another paper [6] author managed to predict the accident hotspots with help of Logistic regression. They provided input of several variable such as driver state, vehicle condition, road condition, traffic in the area, etc. Logistic regression here showed the accuracy of 86.6% provided some suitable variable data used here is the Beijing traffic accident data. But the traffic accident is affected by the many other factors other than these such as weather condition, road alignment and traffic flows so we can say with decision tree we may have approached a different stats and accuracy but that would be more reliable. They also built a microscopic Bayesian networks with weather, time, traffic flow, vehicle speed etc.

## III. VISUALIZATION

Taking in view of the dataset in hand, we have a lot of possible columns that can be used for analysis purpose. Yet it is not feasible to take into consideration all the available columns or attributes and apply algorithms on them. So, in order to determine how each column is related to the required output column and which all columns to take into consideration for the purpose of analysis, we will now conduct some visualization of the available dataset columns.

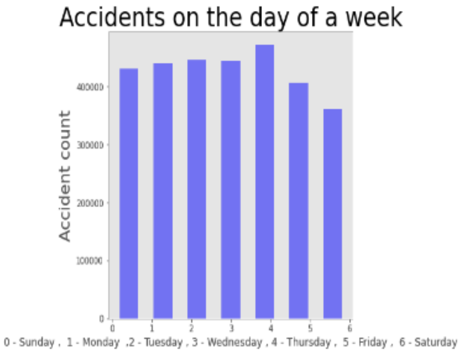

**Fig. 3.1**

1.) **Day of the week:** First column taken into consideration is the "Days of week" column. When we plot a histogram with Days of week v/s the accident count on each day (Fig. 3.1), we can see that the days of the week column does not cause any effect on the accident count. Yet we can see that that the accident count on Thursday is more as compared to other days of the week.

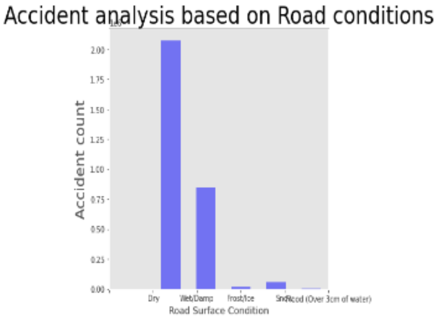

**Fig. 3.2**

2.) **Road Conditions:** Road conditions can differ according to seasons as well as is seen to play a major role in the determination of accident count as depicted in the Fig 3.2. In the plot we can see that the accident count on Dry and wet roads are much more as compared to that on frost and other kind of roads. Also, accident count is more on dry roads as compared to wet roads.

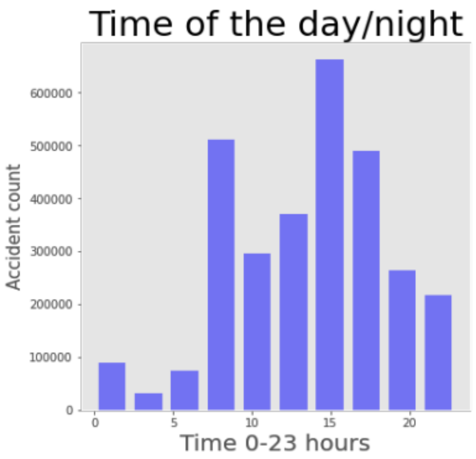

**Fig 3.3**

3.) **Time of the day / night:** From the graph (Fig 3.3) we can clearly see that the accident count increases in the rush hours in the morning as well as in the evening. The accident count is less in the night because of less traffic on the roads.

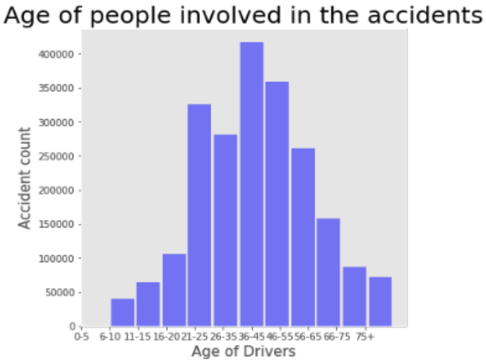

**Fig 3.4**

4.) **Age of people involved in the accidents**(Fig 3.4) play a major role in determining the accident count. As we can see, underage drivers get into accidents but their count is less as compared to other age groups. As seen in the graph, the people of age group 35-45 predominantly get into road accidents. On the whole, people of the middle age groups get into accidents more as compared to underage and aged drivers.

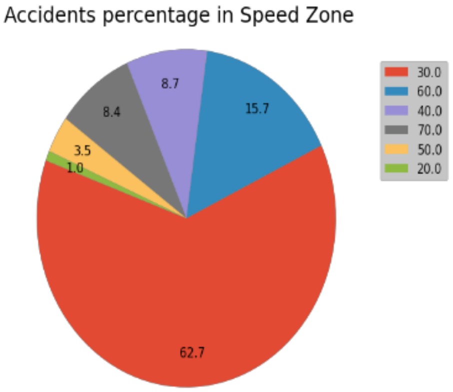

**Fig 3.5**

5.) **Speed Limit** has always been an important factor in determining the possibility of happening of an accident. As we can see from the pie chart (Fig 3.5), More than 50% of the accidents occur at a higher speed. So, when speed limits are crossed, accidents happen.

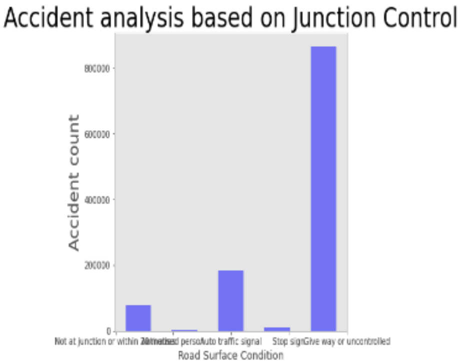

**Fig 3.6**

6.) **Junction Control:** The accident count increases with the change in the junction control technique. As we can see (Fig 3.6), accident count at junctions controlled by a traffic police is less as compared to other types of junctions. Accident count in auto traffic is more than in junction controlled by an authorized person. Also we can see that the accident count is the highest when we have an uncontrolled junction.

## IV. DATASET

After conducting pre-processing over the final dataset, we get a final pre-processed dataframe which we conduct further steps on. We merge the available dataset, drop unnecessary columns and inconsistent data rows, and finally get the pre-processed dataset for analyzation process. The final dataset contains the following columns:

- Age_of_Driver
- Vehicle_Type
- Age_of_Vehicle
- Engine_Capacity_(CC)
- Day_of_Week
- Weather_Conditions
- Road_Surface_Conditions
- Light_Conditions
- Sex_of_Driver
- Speed_limit

Further we conduct correlation analysis over the afore mentioned columns to see how they are related to each other, we get the following results:

| **Column names** | **Correlation_coefficient** |
|---|---|
| Accident_Severity | 1.000000 |
| Sex_of_Driver | 0.070637 |
| Weather_Conditions | 0.022946 |
| Road_Surface_Conditions | 0.011632 |
| Vehicle_Type | 0.008394 |
| Day_of_Week | 0.004186 |
| Age_of_Vehicle | -0.012951 |
| Engine_Capacity_(CC) | -0.018105 |
| Age_of_Driver | -0.024606 |
| Light_Conditions | -0.064616 |
| Speed_limit | -0.095826 |

By the result of the above correlation coefficients of the columns of the final dataframe with the accident severity, we can see that the columns Sex_of_Driver,Weather_Conditions,Road_Surface_Conditions, Vehicle_Type and Day_of_Week have a positive correlation with the Accident_Severity column. On the other hand, the columns Age_of_Vehicle, Engine_Capacity, Age_of_Driver, Light_Conditions and speed_Limit have a negative correlation with the Accident_Severity Column.

Using these columns, we further conduct our analysis on the final dataset.

## V. RESULTS

Now after we have the final ready dataset in hand, we go further to apply our algorithms to the dataset and perform our analysis on our data. Here we are planning to apply four different algorithms on our dataset and then compare their efficiency and accuracy in respect to the problem in hand.

The first algorithm apply is Random Forest Algorithm. After we apply Random forest algorithm on the dataset, we get the results of accuracy as in Table 5.1.1

| Accuracy 84.41 | | | | |
|---|---|---|---|---|
| | Precision | recall | f1-score | support |
| 1 | 0.056657 | 0.007271 | 0.012889 | 5501 |
| 2 | 0.219332 | 0.056924 | 0.090389 | 48591 |
| 3 | 0.866321 | 0.970538 | 0.915473 | 338947 |

**Table 5.1.1(Results)**

| Predicted | 1 | 2 | 3 | All |
|---|---|---|---|---|
| Actual | | | | |
| 1 | 40 | 387 | 5074 | 5501 |
| 2 | 138 | 2766 | 45687 | 48591 |
| 3 | 528 | 9458 | 328961 | 338947 |
| All | 706 | 12611 | 379722 | 393039 |

**Table 5.1.2(Confusion Matrix)**

Next we apply the Logistic Regression algorithm to perform the Severity prediction on our dataset. When we apply Logistic Regression to perform severity prediction, we the get results as in Table 5.1.1 and Table 5.1.2

| Accuracy 86.23 | | | | |
|---|---|---|---|---|
| | Precision | recall | f1-score | support |
| 1 | 0.000000 | 0.000000 | 0.000000 | 5501 |
| 2 | 0.200000 | 0.000021 | 0.000041 | 48591 |
| 3 | 0.862374 | 0.999959 | 0.926084 | 338947 |

**Table 5.2.1(Results)**

| Predicted | 1 | 2 | 3 | All |
|---|---|---|---|---|
| Actual | | | | |
| 1 | 0 | 0 | 5501 | 5501 |
| 2 | 1 | 1 | 48589 | 48591 |
| 3 | 10 | 4 | 338933 | 338947 |
| All | 11 | 5 | 393023 | 393039 |

**Table 5.2.2(Confusion Matrix)**

Next, we apply our third algorithm which is Decision Tree algorithm. When we apply Decision tree algorithm on our dataset, we get the results of accuracy as in fig 5.3

| Accuracy 74.8 | | | | |
|---|---|---|---|---|
| | Precision | recall | f1-score | support |
| 1 | 0.037211 | 0.048900 | 0.042262 | 5501 |
| 2 | 0.149981 | 0.179560 | 0.163443 | 48591 |
| 3 | 0.869822 | 0.840795 | 0.855062 | 338947 |

**Table 5.3.1(Results)**

| Predicted | 1 | 2 | 3 | All |
|---|---|---|---|---|
| Actual | | | | |
| 1 | 269 | 1166 | 4066 | 5501 |
| 2 | 1281 | 8725 | 38585 | 48591 |
| 3 | 5679 | 48283 | 284985 | 338947 |
| All | 7229 | 58174 | 379722 | 393039 |

**Table 5.3.2(Confusion Matrix)**

Now finally we try to perform the prediction and analysis using the last selected algorithm which is the Naïve Bayes Algorithm. This algorithm when applied to the dataset gives the results as in Table 5.4.1 and Table 5.4.2

| Accuracy 74.79 | | | | |
|---|---|---|---|---|
| | Precision | recall | f1-score | support |
| 1 | 0.035615 | 0.046719 | 0.040418 | 5501 |
| 2 | 0.150021 | 0.179663 | 0.163509 | 48591 |
| 3 | 0.869820 | 0.840780 | 0.855054 | 338947 |

**Table 5.1.1(Results)**

| Predicted | 1 | 2 | 3 | All |
|---|---|---|---|---|
| Actual | | | | |
| 1 | 414 | 36 | 5051 | 5501 |
| 2 | 1476 | 273 | 46842 | 48591 |
| 3 | 7911 | 716 | 330320 | 338947 |
| All | 9801 | 1025 | 382213 | 393039 |

**Table 5.1.2(Confusion Matrix)**

With this we are done with the process of result generation using different algorithms. Now we will try analyse the acquired results and generate important inferences.

## VI. DISCUSSIONS

Now that we have run different algorithms on our data and received the results of each algorithm, it is time that we analyse the results to arrive at a decision or final result. Looking at the results, we can see that the accuracy of Naïve Bayes algorithm is the least and its confusion matrix is also no so promising. So we eliminate the Naïve Bayes Algorithm.

Further, looking at the results from Logistic Regression algorithm, we can see that the accuracy of this algorithm is quite good, but when we look at its confusion matrix, it is not quite efficient. So we cannot use Logistic Regression as well.

Now comparing Random Forest algorithm and Decision tree algorithm, we can see that Random Forest algorithm has higher accuracy than Decision Tree algorithm. Now looking at the Confusion Matrices of both algorithms, we can see that Decision tree algorithm performs better than Random Forest algorithm. Also, when talking about the complexity and time taken, Random Forest algorithm is more complex and takes more time for predictions as compared to decision tree algorithm.

## VII. CONCLUSION

From the discussions as well as the points laid down, we can conclude that from among the four algorithms that we took into consideration in this paper including Random Forest algorithm, Decision Tree algorithm, Logistic Regression and Naïve Bayes algorithm; Decision tree algorithm functions best. The decision tree algorithm, although having lesser accuracy than a few other algorithms, has a better confusion matrix as well as is less complex and takes less time to run the predictions. Thus Decision Tree algorithm is best suited for our requirement.

## VIII. REFERENCES

[1] Suwarna Gothane,M.V. Sarode's "Analysing Factors, Construction of Dataset, Estimating Importance of Factor, and Generation of Association Rules for Indian Road Accident," IEEE, Feb 1, 2016

[2] Gagandeep Kaur, Er. Harpreet Kaur's "Prediction of the cause of accident and accident prone location on roads using data mining techniques," IEEE, Jul 1, 2017

[3] Liling Li, Sharad Shrestha, Gongzhu Hu's "Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques," IEEE, June 7-9, 2017.

[4] Y. Wang, Y. Ou, X. Deng, L. Zhao and C. Zhang, "The Ship Collision Accidents Based on Logistic Regression and Big Data," 2019 Chinese Control And Decision Conference (CCDC), Nanchang, China, 2019, pp. 4438-4440.

[5] R. Serban, A. Kupraszewicz and G. Hu, "Predicting the characteristics of people living in the South USA using logistic regression and decision tree," 2011 9th IEEE International Conference on Industrial Informatics, Caparica, Lisbon, 2011, pp. 688-693.

[6] T. Lu, Z. Dunyao, Y. Lixin and Z. Pan, "The traffic accident hotspot prediction: Based on the logistic regression method," 2015 International Conference on Transportation Information and Safety (ICTIS), Wuhan, 2015, pp. 107-110.

[7] Rishi Sai Reddy Sudireddy's "Prediction of Road Accidents using Correlation based on Map Reducing," IEEE

[8] Youhee Choi, Jeong-Ho Park, Byungtae Jang's "Developing safety checklists for predicting accidents," IEEE, Oct. 17-19, 2018
.

[9] Jihua Ye, Yanhui Zhou, Ming Li, Chunlan Wang's "Research and implement of traffic accident analysis system based on accident black spot," IEEE, Aug. 24-27, 2010.

[10] Yang Yanbin, Zhou Lijuan, Leng Mengjun, Sun Ling's "Early Warning of Traffic Accident in Shanghai Based on Large Data Set Mining," IEEE.

# Predicting Severity of Road Accidents and using various Machine Learning Algorithms

ORIGINALITY REPORT

| 2% | 0% | 2% | % |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Ramona Serban, Andrzej Kupraszewicz, Gongzhu Hu. "Predicting the characteristics of people living in the South USA using logistic regression and decision tree", 2011 9th IEEE International Conference on Industrial Informatics, 2011<br>Publication | 1% |
| 2 | Yi-han Wang, Yang Ou, Xu-dong Deng, Lu-ran Zhao, Chao-yu Zhang. "The Ship Collision Accidents Based on Logistic Regression and Big Data", 2019 Chinese Control And Decision Conference (CCDC), 2019<br>Publication | 1% |
| 3 | Zhao Yongyi, Pan Qiang. "Intelligent data mining for economie prediction and analysis", 2010 2nd IEEE International Conference on Information Management and Engineering, 2010<br>Publication | <1% |
| 4 | medium.com<br>Internet Source | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | < 10 words |
| Exclude bibliography | On | | |