INTERNATIONAL JOURNAL · ENGINEERING RESEARCH ·

International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

IJIEMR Transactions, online available on 20th July 2020. Link :http://www.ijiemr.org/downloads.php?vol=Volume-09&issue=ISSUE-07

Title: **ROAD ACCIDENTS ANALYSIS USING MACHINE LEARNING ALGORITHMS**

Volume 09, Issue 07, Pages: 159 - 165

Paper Authors

K.V.KIRAN, SARAGADAM MADHAVI, CHUKKA MOUNIKA

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

ROAD ACCIDENTS ANALYSIS USING MACHINE LEARNING ALGORITHMS

K.V.KIRAN[1*], SARAGADAM MADHAVI[2**], CHUKKA MOUNIKA[3***]
[1, 2, 3] Department of CSE, Welfare Institute of Science Technology And Management, Pinagadi, Pendurthi, Visakhapatnam, Andhra Pradesh, India
[*] kasi.kvk@gmail.com ** saragadamadhavi.cse@gmail.com , *** mounikach727@gmail.com

Abstract: Today, traffic safety is one among the most priorities of governments. Considering the importance of topic, identifying the factors of road accidents has become the main aim to reduce the damage caused by traffic accidents. Data mining is that the process of analysing data from different perspectives and summarizing it into useful information. Data mining allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically, data mining is that the process of finding correlations or patterns among dozens of fields in large relational databases. One of the key objectives in accident data analysis is to identify the main factors associated with a road and traffic accident. The analysis can be done by using supervised machine learning algorithms such as Gaussian Naïve Bayes, Logistic Regression and Random forest. We are developed an web application for prediction of accident severity

1. Introduction

Road accidents are a serious threat to the people. There is a huge impact on the society due to traffic accidents where there is a great costs of fatalities and injuries. In recent years, there is a increase in the researches attention to determine the significantly affect the severity of the drivers injuries which is caused due to the road accidents. Accurate and comprehensive accident records are the idea of accident analysis. the practical use of accident records depends on some factors, like the accuracy of the info, record retention, and data analysis. The residential and shopping sites are more hazardous than village areas.as may need been predicted , the frequencies of the casualties were higher near the zones of residence possibly because of the higher exposure.

There are many approaches are applied for the prediction of accident severity but it does not give the accurate values.

In this, To get the better values Advanced scientific methods are used that is Machine learning Techniques, For getting the accurate and better result we are using three supervised machine learning classification algorithms that are Gaussian Naïve Bayes, Logistic Regression, and Random Forest.

The main goal of our paper is to get the better accurate value by analyzing the road accident data set and determines the severity prediction using classification techniques .

Data mining Techniques such as Data preprocessing is used to remove the noisy data while collecting the data .To avoid these situations We are preprocessing the data for determining the accurate value.

Finally, The Logistic Regression classification can be achieved best performance with the accuracy of 84%.

By analyzing the data set and prediction of accident severity we can reduce the accidents by taking precautions based on road conditions, weather conditions, Road junction type etc., before occurrence of accidents.

2. Literature Survey

Road accidents analysis can be done using different approaches previously. Clustering and pattern techniques are used to analyze data but it does not give better result as compare to classification techniques.

We are surveyed the different types of research papers and distinct techniques applied for the accidents analysis, from that papers we are analyzed that peculiar conditions are taken for get the better result.

Some of the research papers we studied are:
Mussone et al. used neural networks to research vehicle accident that occurred at intersections in Milan, Italy . They choose feed-forward MLP using BP learning. This miniature had ten input nodules for eight variables (accident type, paved surface condition, and weather conditions). The output node was called an accident index and was calculated on the point of ratio between the number of accidents for a given intersection and therefore the number of accidents at the most dangerous intersection. Results showed that the very best accident index for running over of pedestrian occurs at non-signalized intersections at night.
Dia et al. used physical-world info for developing a multi-layered MLP neural network freeway incident detection model [5]. They compared the performance of the neural network model and therefore the incident detection model operational on Melbourne's freeways. Results showed that neural network model could provide faster and more reliable incident detection over the model that was operational. They also found that failure to supply speed data at a station could significantly deteriorate model performance within that section of the freeway.

Sohn et al. applied data fusion, ensemble and clustering to enhance the accuracy of individual classifiers for two categories of harshness (bodily injury and property damage) of road traffic accidents .The respective classifiers used were neuralnetwork and decision tree. They applied a clustering algorithm to the dataset to divide it into subsets, then used each subset of knowledge to convoy the classifiers. They found that classification supported clustering works better if the variation in observations is comparatively large as in Korean road traffic accident data.
Bedard et al. applied a multivariate logistic regression to work out the independent contribution of driver, crash,

and vehicle characteristics to drivers' fatality risk [3]. They found that increasing seatbelt use, reducing speed, and reducing the amount and severity of driver-side impacts might prevent fatalities. Evanco conducted a multivariate population-based statistical analysis to work out the connection between fatalities and accident notification times [6]. The scrutinity demonstrated that accident notice time is an important determinant of the amount of fatalities for accidents on rural roadways.

Yang et al. used semantic network approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs [17]. They performed the Cramer's V Coefficient test [18] to spot significant variables that cause injury to scale back the aspect of the data. Then, they applied data transformation method with a frequency-based scheme to rework categorical codes into numerical values. They used the Critical Appraisal Reporting Environment (CARE) system, which was developed at the University of Alabama, employing a Backpropagation (BP) neural network.

3. Existing System

Data Mining techniques are used to identify the locations where high frequency accidents are occurred and analyze them to identify the factors that have an effect on road accidents at that locations. The first task is to divide the accident location into k groups using the k-means clustering algorithm supported road accident frequency counts. Then, association rule mining algorithm applied so as to seek out the connection between distinct attributes which are in accident data set and consistent with that know the characteristics of locations.

4. Proposed System

Classification techniques will be using for identifying the accident prone area's. The accident data records which may help to know the characteristics of the many features like drivers behavior, roadway conditions, light condition, weather situation and so on. This can help the users to compute the security measures which is beneficial to avoid accidents. The data set can be analyzing by comparing Gaussian naïve Bayes, Logistic regression, Random Forest algorithms we are analyzing which algorithm will gives the accurate dataset . The models are performed to identify statistically significant factors which can be able to predict the probabilities of crashes and injury that can be used to perform a risk factor and reduce it.

5. System Architecture

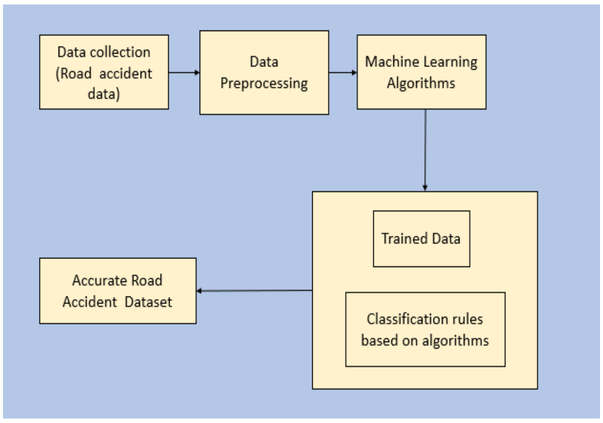

Data collection (Road accident data)
Data Preprocessing
Machine Learning Algorithms
Trained Data
Classification rules based on algorithms
Accurate Road Accident Dataset

fig: system architecture

This architecture describes the work flow of overall data analysis. Firstly the data can be collected then we can preprocessing the data to remove any occurences of noisy data in the data set. After preprocessing applying the suitable machine learning algorithms on trained data which gives the accurate data set.

Data collection:

To get the prediction of accurate accident severity. Number of road accidents data sets with complete information are needed for getting the accurate data value. we are collecting the data from the kaggale which consists of some constraints like Road type, light presentation, junction type etc.,

Data preprocessing:

It is the process of preparing the raw data and making it suitable for a machine learning model. We are preprocessing the dataset for clean data. while gathered the data from different sources there is chance for occurrence of noisy ,inconsistence data.

Trained Data:

It is the actual dataset which is used train the model for performing various actions.

6. Methodology

A) Gaussian Naïve Bayes Algorithm:

It is a classification technique supported Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature., A naive Bayes classifier would consider all of these properties to independently contribute to the probability Naive Bayesian model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is understood to outperform even highly sophisticated classification methods.

Bayes theorem provides how of calculating posterior probability

P(c|x) from P(c), P(x) and P(x|c)

Where,

P(c|x) is that the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P(x|c) is that the likelihood which is that the probability of predictor given class.

P(x) is the prior probability of predictor.

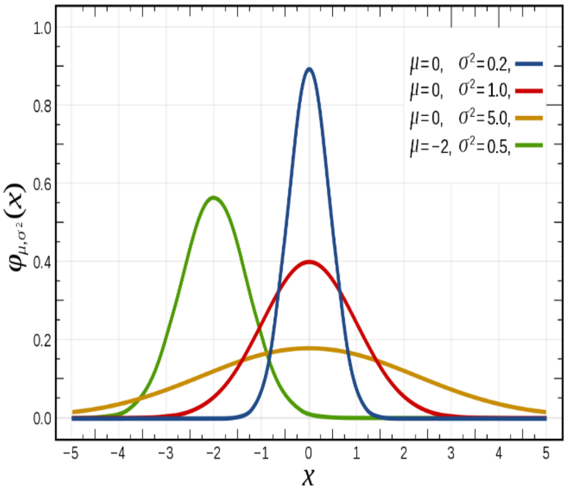


Fig: Gaussian Naïve Bayes

A) Logistic regression Algorithm:

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.It uses a more complex cost function, this cost function are often defined just as the '**Sigmoid function**' or also referred to as the 'logistic function' rather than a linear function.

The interpretation of logistic regression gravitate it to limit the cost function between 0 and 1. Thus linear functions fail to perform it as it can have a worth greater than 1 or but 0 which is not feasible as per the hypothesis of logistic regression.

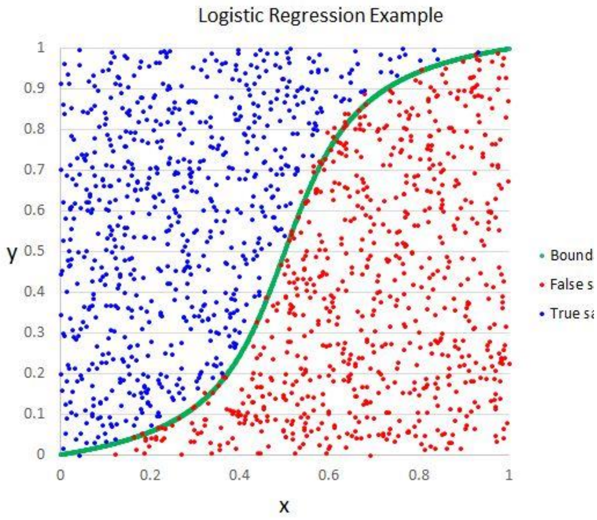


Fig: Logistic Regression.

A) Random Forest Algorithm:

Random Forest is a supervised learning algorithm which is used for both classification and as well as regression.This method creates Decision Trees on data fragments and then gets prognosis from each of them and finally selects the best solution by means of voting.It is a better algorithm because it reduces the over-fitting by averaging the result.It consists of a large number of individual decision trees. Each individual tree within the random forest spits out a category prediction and therefore the class with the foremost votes becomes our model's prediction .

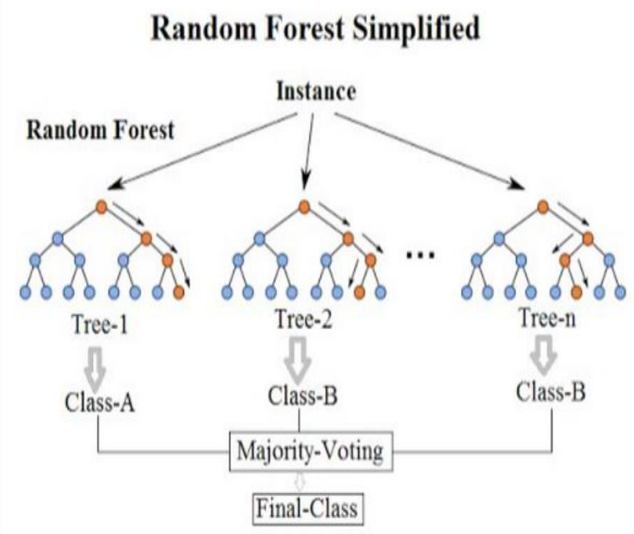
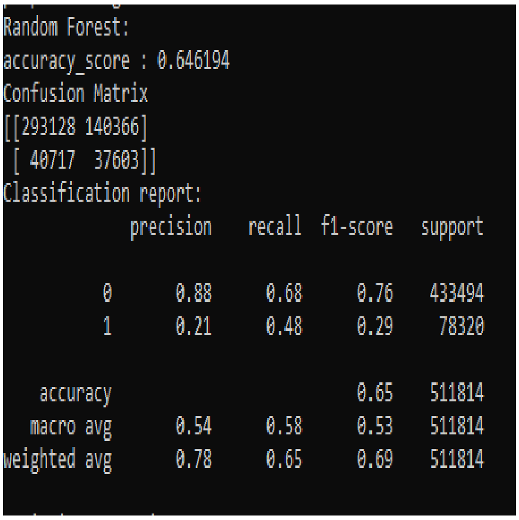


Fig: Random Forest

7. Result analysis

We are done with the analysis of road accident dataset and the output we are getting accuracy for each classifier and statistical representation of the accuracy for the dataset.



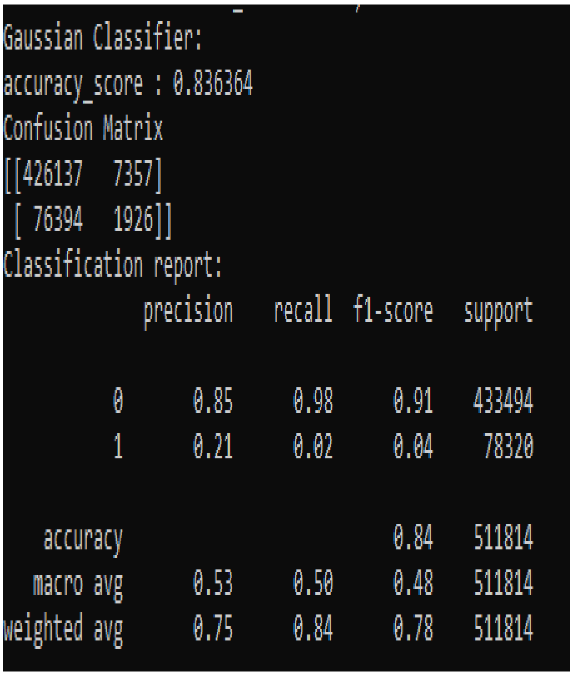
```
Random Forest:
accuracy_score : 0.646194
Confusion Matrix
[[293128 140366]
 [ 40717  37603]]
Classification report:
              precision    recall  f1-score   support

           0       0.88      0.68      0.76    433494
           1       0.21      0.48      0.29     78320

    accuracy                           0.65    511814
   macro avg       0.54      0.58      0.53    511814
weighted avg       0.78      0.65      0.69    511814
```

Fig: The accuracy we are getting for Random Forest classifier is 64%.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

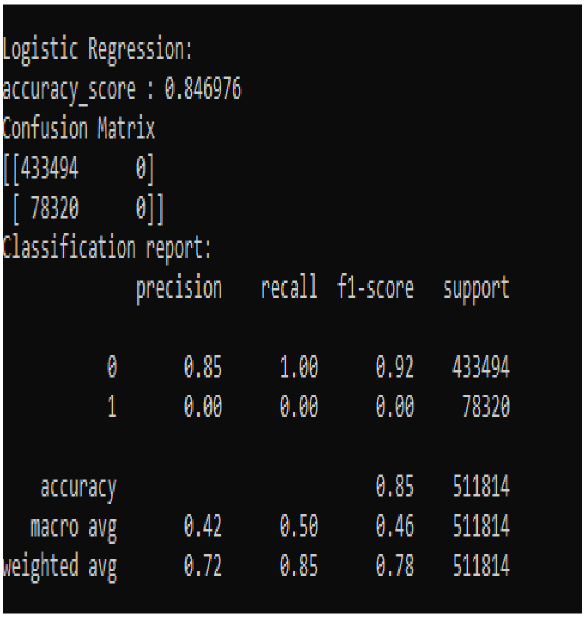


```
Gaussian Classifier:
accuracy_score : 0.836364
Confusion Matrix
[[426137   7357]
 [ 76394   1926]]
Classification report:
              precision    recall  f1-score   support

           0       0.85      0.98      0.91    433494
           1       0.21      0.02      0.04     78320

    accuracy                           0.84    511814
   macro avg       0.53      0.50      0.48    511814
weighted avg       0.75      0.84      0.78    511814
```

Fig: The accuracy we are getting for Gaussian classifier is 83%.



```
Logistic Regression:
accuracy_score : 0.846976
Confusion Matrix
[[433494      0]
 [ 78320      0]]
Classification report:
              precision    recall  f1-score   support

           0       0.85      1.00      0.92    433494
           1       0.00      0.00      0.00     78320

    accuracy                           0.85    511814
   macro avg       0.42      0.50      0.46    511814
weighted avg       0.72      0.85      0.78    511814
```

Fig: The accuracy we are getting for Logistic Regression is 84%.

Statistical Analysis

Fig: statistical representation of performance analysis.

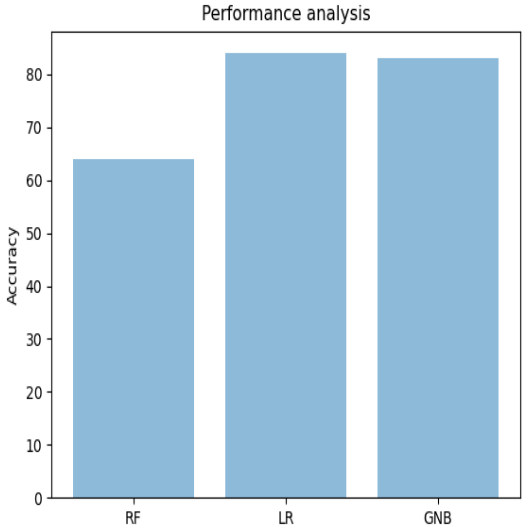
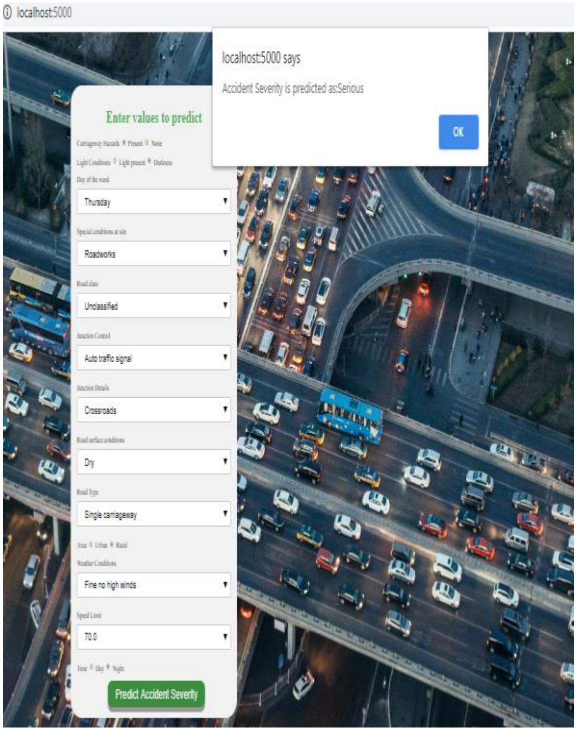


Fig: Accident Severity prediction as Serious.



localhost:5000

localhost:5000 says

Accident Severity is predicted as:Serious

Enter values to predict

Carriageway Hazards ● Present ○ None

Light Conditions ○ Light present ● Darkness

Day of the week: Thursday

Special conditions at site: Roadworks

Road class: Unclassified

Junction Control: Auto traffic signal

Junction Details: Crossroads

Road surface conditions: Dry

Road Type: Single carriageway

Area ○ Urban ● Rural

Weather Conditions: Fine no high winds

Speed Limit: 70.0

Time ○ Day ● Night

8. Conclusion

Road Accidents are caused by various factors. By going through all the research papers it can be concluded that Road Accident cases are hugely affected by the factors such as types of vehicles, age of the driver, age of the vehicle, weather situation, road structure and then on. Thus we've build an application which gives efficient prediction of road accidents supported the above mentioned factors.

Due to analyzing and severity prediction we can reduce the road accidents by taking some precautions before occurrence of accidents.

9. Future Enhancement

We can design an application to alert the users before the occurrence of accidents and give precautions to the user. We can also take the reasons of accidents from users to reduce the reason for that area to be accident prone and we can also inform the respective authorities so that necessary steps should be taken. We can provide the feature in which a user enters the starting point and destination point and application shows the different paths with the number of accidents occurred in each path.

10. References

[1] Abdel-Aty, M., and Abdelwahab, H., inquiry and Prophecy of Traffic Fatalities Resulting From Angle Collisions Including the Effect of automobiles' Contour and affinity. Accident Analysis and Prevention.

[2] Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds,J. P., The Independent supplement of Driver, Crash, and Vehicle distinctive to Driver Fatalities. Accident analysis.

[3] Buzeman, D. G., Viano, D. C., & Lovsund, P., Car Occupant Safety in Frontal Crashes: A framework Study of automobile Mass, Impact Speed, and essential Vehicle Protection. Accident inquiry and Prevention.

[4] Abdelwahab, H. T. and Abdel-Aty, M. A., Development of Artificial semantic Network Models to anticipate Driver Injury Severity in Traffic Accidents at Signalized Intersections. Transportation Research Record.

[5] Mussone, L., Ferrari, A., & Oneta, M., An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention.