GUI Based Road Accident Prediction Using Ensemble Learning

[1]S Anirudh, [2]Varun S, [3]Rogit T S, [4]R Logeshwari,
[1, 2, 3] B.Tech, [4]Assistant Professor,
[1,2,3,4] Department of computer science and Engineering,
[1,2,3,4] SRM Institute of Science and Technology, Chennai, India.

***Abstract*— Each and every day population in the world increases which results in increase in vehicles. As vehicle increases, so do the accidents are also increases. The reason for the accidents can be the reckless driving and infractions served by the driver. However, not all the accident cases can be justified against the driver. Because there can be considerations like route, road conditions, weather and other factors such as age and traffic can also be the reason for accidents. The second reason can be reduced by analyzing the past incidents as data set and using them to predict the best route, speed and time to travel by the help of data manipulation methods, classification and regression machine learning algorithms.**

Keywords—Increase in vehicles, Route conditions, Data manipulation, classification algorithms.

I. Introduction

Countries like India have way big population strength which in turn increases the count of vehicles. As the number of vehicles grows up, the rate of accidents also grows. Each and every day metropolitan cities are experiencing so many accidents. A recent survey published; in a metropolitan city as an average of 7250 accident cases are filed in a year. The count is increasing day by day as the population grows on. Accidents occurs in daily basis can result in drastic loss in both social and economic loss and it occur are due to either Human error by serving infractions and poor route condition such as damaged roads, low lighting on the road or by nature for weather conditions. The accidents cannot be stopped; however, the accidents can be reduced by picking up the safest route and timing for travel or having good knowledge about the risky route. The already occurred incidents usually provide enough data for the research to identify the factors which leads to accidents. The data which extracted from the past incidents is used to predict whether an accident can occur or not. The proposed system provides the abovementioned information as a prediction with the help of past data as a data set and manipulating then classifying the with help of Machine learning algorithms.

II. Related Works

In this paper, the related works section provides a summary about the proposed system with which attributes are taken from the data set for the prediction. The data set is based on Road Accidents in 2017 from Gov.in website. It is found that the majority road crashes are happened due to human error.

The prediction is a process of using the past or previous occurrence information to explore or calculate the future possibility. To make the predication some of the data attributes must be fetched to the machine algorithms. Those attributes can be independent or as a group may contribute to an outcome. Also, the attributes have high chance on affecting the possibility of the accuracy. We choose attributes such as Gender, Road lighting conditions, Weather conditions, Number of vehicles, Type of vehicle and location for the prediction and observations. As these attributes can individually or collaboratively relate to the outcome [1].

A. *Road Lighthing Condition*

Lighting conditions in the environment contributing more possibility to the accidents [4]. There is a study in Road accidents report stating, that the greater number of accidents are occurred at darker environment. Due to poor visibility towards the path makes the rider to involve in accidents. One study in New Zealand investigated accident cases by improving lighting conditions and adding new lights reduced the accident cases over by 50 % [6].

B. *Weather Condtiions*

Another major factor which increases the accident is weather condition [3]. Accident occurring in rainy or in foggy weather is much more in rate compared to the accident occur in the clear or sunny weather. As these climatic conditions can alter the possibility of the accident occurrence [4]. Climatic conditions which have variation winds also a factor to be considered [1]. People who are taking Car or van like vehicle which have medium to protect the rider from outside world are much safer from accidents than two wheelers as they more open to the real world.

C. *Type of Vehicle and Gender*

Considering vehicle and gender as factor for the observation for the accident helps toward the prediction. Because vehicles like car, van are considered to safe than two wheelers [1]. As seconding the above statement, four wheelers are protecting the person from weather such as rain, and equipped with much more controlling features which are not available in two wheelers. However, two wheelers are having much risk as the person in driving is more exposed. Recent study made a statement; men are causing more accidents than women when they are behind the wheels. the same study shows that men make accidents due to overconfidence, excessive speeding and serving infractions like reckless driving behavior [2]. However, the accidents caused by women are lacking managing the traffic and they find trouble in navigation. These reports are supported by the fact Women's physical height is less compared to men [8], which affect their view when they drive and women have less confidence and experience in driving when compared to men.

III. Domain Overview

A. *Machine Learning*

Machine learning, a type of A.I. and a concept in computer science used to predict the future using past data. The proposed systems use machine learning algorithms by classify, train the available data and test them to predict the accidents.

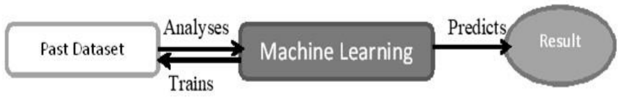

Fig 1: Overview of the proposed system

Each and every datum present in the dataset can affect the accuracy of the system in both positive and negative ways. Any incomplete or duplicate entries in the data set can alter the effectiveness of the proposed system, so data is validated and refined before training of data process by removing redundant data and incomplete data.

B. *Ensemble Learning*

Ensemble method is uses multiple machine learning classification algorithms for a better predictive performance. Using voting classifiers for every algorithm in training and testing the data, the accuracy is calculated and the most accurate result is taken by comparing the result with data set.

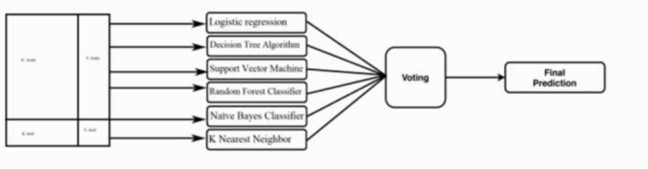

Fig 2: Ensemble learning method

IV. System Architecture Diagram

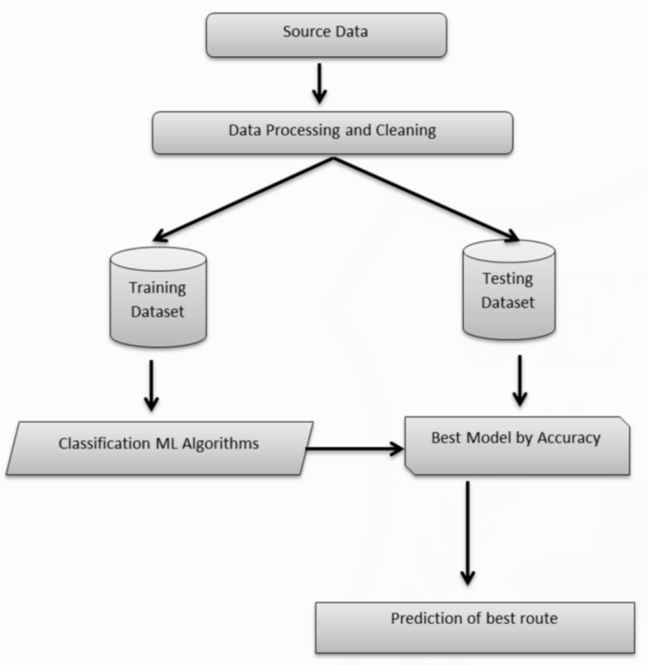

Fig 3: System Architecture of the proposed system

The Road traffic information is generated as a dataset in tabular format as comma separated values file. Then, the data set is processed to meaningful structured information for the proposed system which helps for prediction. Once the data set is converted to a meaningful structure, it is visualized for choosing the best attributes for the training, testing as well as for the entry of new observation. For the prediction of results, the data set is sliced for dataset training and dataset testing with multiple machine learning algorithms. Those machine learning algorithms analyze and classify the entries in the data set and calculate the probability of positive and probability of negative results. Each algorithm generates different level of prediction with different accuracy. Once the new observations are fetched into the system, the ensemble method will work for finding the best accurate algorithm for the observation to predict the result. Ensemble learning is used to cross validate the algorithms and best among them is picked by voting classifiers.

V. System Requirements

A. *Functional Requirements*

The proposed system undergoes data processing, data manipulation, mathematical, numerical processes and frontend GUI framework using python library packages such as Pandas, Numpy, Sk-learn, Matplotlib, Seaborn and Tkinter.

B. *Environmental Requirement*

Operating System	: Windows
Tool	: Anaconda, Jupyter Notebook
Processor	: Pentium IV/III
Hard disk	: minimum 80 GB
RAM	: minimum 2 GB

VI. Methodology

A. *Data Validation and Processing*

Data validation techniques are used in the system to refine the dataset which have numerous amounts of data. In real world conditions, the consistency of data may have redundancy and flaw which proportionally affect the algorithms performance. The process of eliminating the redundancy and flaws also fine tuning by ensuring a proper fit in the dataset.

B. *Data visualization and training the model* As the data is validated and processed, it is explored in graphical representation for finding the correlation between the parameters and finding the crucial attributes which results proportion in end result predictions. Then, dataset is divided in two halves with 70:30 ratios for training and testing the system by analyzing the 70% of the data from the data set including the result and testing the remaining 30 % data by having the attributes and predicting the result.

C. *Logistic regression and Decision Tree Algorithm* Logistic regression predicts something which lies between two given possible extents. Considering one or more attributes as genotype, a curve is generated with probability for the new data to be placed on one extent which is a prediction process.

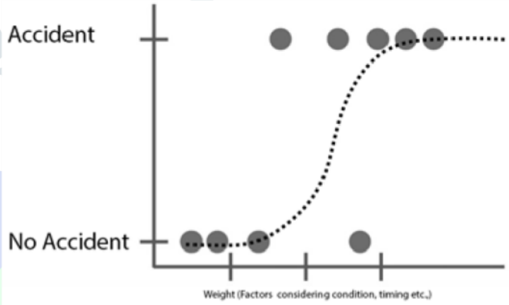

Fig 4: Logistic Regression

Decision Tree Algorithm, from a tabular data to root-tree structure having binary results to classify the data and generate the end result for prediction. Considering any of the attributes available in the data by calculating and comparing the minimal gini value is taken as root value to classify the data and to create a new leaf node. Similarly, other remaining attributes are considered with minimal gini value.

D. *Random Forest Classifier and Support Vector Machine* Decision trees are more sensitivity towards prediction can be affected by inaccuracy, it doesn't have great effect on classifying new samples. Random forest classifier combines simplicity and flexibility which results in accuracy in prediction. A bootstrapped dataset is created from the original, by choosing random entries. Then decision trees are created by choosing random attributes from the bootstrapped data. When a new sample is entered to the system, it is set to run with all the trees. Then result of the trees is selected by comparing votes.

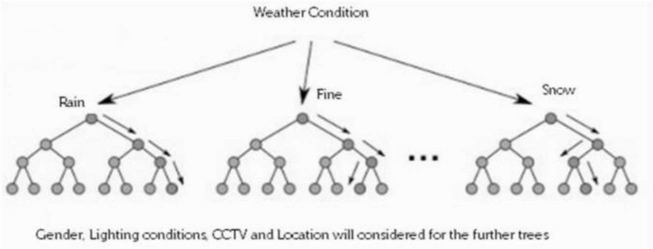
Weather Condition
Rain
Fine
Snow
...
Gender, Lighting conditions, CCTV and Location will considered for the further trees

Fig 5: Random Forest Classifier

Support Vector machine, a classifying machine learning algorithm, categorizes given data in a plane to yield high prediction rate by creating a threshold margin between the categories. When a new observation enters with comparing the range between the threshold margins the observation is classified.

E. Naïve Bayes Classifier and K Nearest Neighbor

Typically, any dataset will have a greater number of entries with results. Those results present in the data can be either in a positive or in negative way. In that case, the proposed system's dataset also contains various end results. Each or more than one attribute can generate the event. Each attribute is compared with other attribute occurrence and event and probability is generated.

The data given in the proposed system already have a lot of data that define the results and by using the same we can decide when a new observation reaches the algorithm to the result by looking at the nearest annotated event. With that we can classify the observation for prediction.

F. Ensemble Learning method

As all the algorithms trained and tested the dataset for the prediction. Each algorithm has its own prediction accuracy and they differ from each other. Those prediction accuracy rates are compared and cross validated with each other, the most effective is voted.

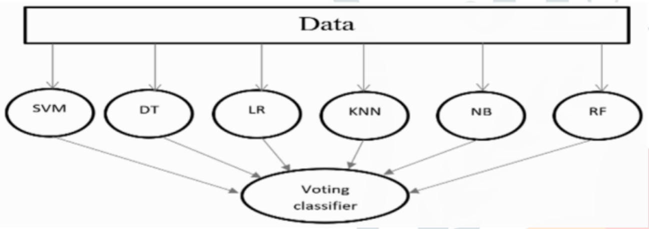

Fig 6: Ensemble Learning

G. Tkinter

To interact with the system, GUI is created with Tkinter, A python GUI Framework. The new observations are entered into the proposed system prediction through dropdown list buttons such as Gender, Lighting conditions, Weather, vehicles. To increase the potency of the system, the dataset is recreated with replacing the attributes which are used predict the results with integer values. Once the observations are entered, the system uses trained and tested data to check the accuracy of each algorithm for the new observation. With the assistance of ensemble learning the most effective rated algorithm's result is considered as prediction with accuracy.

VII. PERFOMANCE OF THE CLASSIFICATION ALGORITHMS.

The ability of the classification algorithm towards the prediction is generated with Confusion matrix. It is an abstract of the outcome results of the training and testing methods. The sensitivity of the confusion matrix of each classifying algorithm is used for comparing their effectiveness with the help of Precision, Recall and Accuracy performance rate.

Confusion Matrix

	Predicted Yes	Predicted No
Actual Yes	True positive	False Negative
Actual No	False Positive	True Negative

True Positive: Both actual and predicted to be Yes.

True Negative: Both actual and predicted to be No.

False Positive: The actual value in the matrix is negative. But the prediction is positive.

False Negative: The actual value in the matrix positive. But the prediction is Negative

Performance of the algorithms

	Precison	Recall	F Measure	Accuracy
LogisticRegression	50%	58%	50%	58.1%
Decision	100%	100%	100%	100%
SVM	32%	57%	41%	56%
Random Forest	100%	100%	100%	100%
Naïve	100%	100%	100%	100%

$$Precison = \frac{TP}{TP + FP}$$

Precision is the ratio of true positive case to sum of true positive and false positive.

$$Recall = \frac{TP}{TP + FN}$$

Recall is the ratio between True positive to aggregate of true positive and false negative.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy performance rate is the ratio between the sum of true positive and true negative to all the predictions in the matrix.

VIII. EXPERIMENTAL ANALYSIS OF RESULT

As discussed in the Related works section (Section II), It was difficult to find the factors which plays wide role in road accidents because of more amount of data which can have unknown entries in the data set. So, small of amount data is generated and it has been fine tuned by removing the non-available and redundant attributes. Then crucial attributes are selected such as Gender and Environment factors.

A. Gender

Gender is considered as a supporting factor to the accident occurrence. From the detail provided in the related works section, men are more likely to involve in road accidents than women. As the proposed system is based on the dataset which also follows the same details, the results generated are similar to the study. Out of 20 trails for each gender with setting observations on worst possibility, the system predicts around 13 Men can involve in accident. Similarly, with female as gender and all other conditions as same, the predicted result was 9 women can involve in crash.

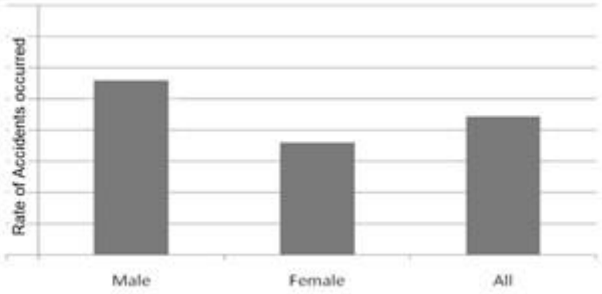

Fig 7: Gender Relationship Model to the road accidents

B. Road Lighting condition and Weather Conditons

As discussed in the related works section (Section II), Road lighting conditions are considered as one of the major factors for the vehicle crashes. We consider the following lighting aspects, 1) No street Light, 2) Street light Unknown 3) Street Light Present. Availability of light in vehicles is only enough for viewing the road but when comes to the obstacle in some distant, street lights are required. As the driver finds the obstacle in distance allows him/her to have time for react.

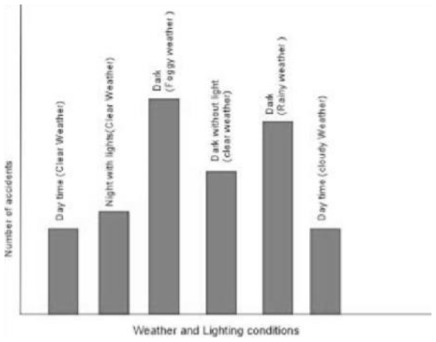

Fig 8: Gender Relationship Model to the road accidents

IX. Result and Conclusion

The entire process of predicting the result for the fetched information starts from data validation to ensemble learning by carrying the data visualization and classifying process with data training and testing in between. The predictions are performed to generate the result of a future occurrence; However, these results remain at general level and they are not at expert level. Because there are multiple factors which can be categorized as Human and environment factors, which individually or as a group contributing towards the accident. This methodology enables us to understand that these computational processes can be implemented in such traffic control departments for analyzing the past or already occurred data. There some cases such as sudden Age, Cardiac arrest, sleepiness and suicide attempts, where accidents can happen without providing all the factors are safe observations. In those events, the health factors aren't included as they cannot be determined and controlled for each and every individual. So, the proposed system doesn't use any of the abovementioned attributes. On the contrary, human factor like Age and environment factors such as lighting conditions are considered. By these processes and methodology one can predict the result and that can be used to prevent road accidents by following road discipline, road signs and regulations of the tracks.

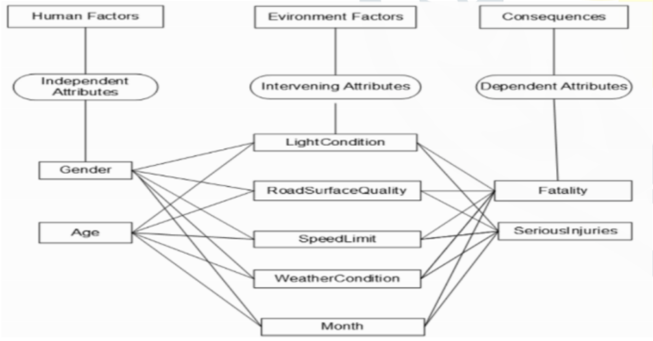

Fig 9: Accident Factors Relationship Model

With availability of much more data more accurate results can be determined. In other words, the entire process can be more versatile with greater accuracy when more attributes and information are available. Such as timing, road conditions, speed limit and etc.

In conclusion, most of accidents are occurred due to poor lighting and bad weather, by supporting those two factors gender and type of vehicles also contributing. Male riders behind the wheel are involved in accidents due to excessive speeding and other infractions which they served. Similarly, women lack the confidence and experience make them involved in such occurrences. In this case, the accidents can be reduced, if the person behind the wheel follows all the rules and regulations with unserving the infractions and wearing safety equipment. Also, road traffic department can lay and maintain good road and lighting conditions in order to decrease the rate of road accidents.

X. Future Work

As discussed in the conclusion, even with availability of greater number of data, the accuracy predication can be increased. It is important that the data should contain all the needed or at least most needed information to design a perfect prediction system. Creating a web portal will enable more people to enter update the present data, however some man power is required to check the entries are true enough with real world entity for accuracy. Updating the urbanizing process on a route by the government makes easy for the end using people to know about the route.

References

[1] A Data Mining Approach for Analysing Road Traffic accidents; Tariq Abdullah, Symon Nyalugwe, University of Derby, IEEE 2019.

[2] Generating Road Accident Prediction Set with Road Accident Data Analysis Using Enhanced Expectation-Maximization Clustering Algorithm and Improved Association Rule Mining; Sakham Nagendra Babu, Jebamalar Tamilselvi, IIETA, 2019.

[3] Data Mining Methods for Traffic Accident Severity Prediction; Qasem A. Al-Radaideh and Esraa J. Daoud, ISSN, 2018.

[4] Stack Denoising Convolutional Autoencoder Model for Accident Risk Prediction via Traffic Big Data; Chao Chen, Xiaoliang Fan, Chuanpan Zheng, Lujing Xiao, Ming Cheng, Cheng Wang ; IEEE 2018.

[5] A Review On Road Accident Data Analysis Using Data Mining Techniques; Prajakta S. Kasbe, Apeksha V. Sakhare, ICIIECS, IEEE2017

[6] Road lighting research for drivers and pedestrians: The basis of luminance and illuminance recommendations; S Fotios, R Gibbons, Lighting Res. Technol. 2018;

[7] Data Mining Methods for Traffic Accident Severity Prediction; Qasem A. Al-Radaideh and Esraa J. Daoud, ISSN 2018.

[8]https://www.syracuse.com/news/2011/07/women_worse_drivers_more_crashes_than_men_less_driving.html, Updated 2019.