International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-2 (April 2020)
www.ijemr.net https://doi.org/10.31033/ijemr.10.2.3

Analysis of Machine Learning Algorithm with Road Accidents Data Sets

P Sumanth[1], P Sai Anudeep[2] and S Divya[3]
[1]U.G Student, Department of Computer Science, Sathyabama University, Chennai, INDIA
[2]U.G Student, Department of Computer Science, Sathyabama University, Chennai, INDIA
[3]Assistant Professor, Department of Computer Science, Sathyabama University, Chennai, INDIA

[1]Corresponding Author: pokalasumanth21@gmail.com

ABSTRACT

Beginning at now, street transport framework neglect to alter up to the exponential expansion in vehicular masses and to ascertaining the quickest driving courses and catastrophes inside observing differing traffic conditions is a critical issue right presently structures. To upset this issue is to explore the vehicle division dataset with bundle learning technique for finding the best street choice without calamity gauging by want aftereffects of best accuracy count by looking at oversaw AI figuring. In bits of information and AI, bundle strategies utilize diverse learning calculations to give indications of progress prudent execution. The assessment of dataset by facilitated AI technique (SMLT) to get two or three data takes after, factor perceiving proof, univariate evaluation, bivariate and multi-variate appraisal, missing worth medications and separate the information support, information cleaning/organizing and information perception will be done with everything taken into account given dataset. In addition, to look at and talk about the presentation of different AI figuring estimations from the given vehicle division dataset with assessment of GUI based street fiasco want by given attributes.

***Keywords*--** Dataset, Ensemble Method, GUI Results

I. INTRODUCTION

The proportion of setback data set aside by traffic the administrators workplaces has been creating at an ever-growing rate over the latest couple of years because of different road vehicle crashes. Government substances and some private parts were found social affair setback data at step by step bases. Data from accidents is every now and again among the most significant assets for neighborhood authorities and central governments, as it helps in arranging and utilization of approaches. It can in like manner help policymakers with settling on fitting decisions relating to foundation progress, arranging and social distinctions movement. All things considered, as the extent of this information is making, there is demand and a need of discovering strategies, procedure and contraptions to assessments such colossal volumes of information and discover an answer t o t he reason behind debacles. Street vehicle crashes are a vital as a rule danger that keeps causing challenges, wounds and fatalities on roadways reliably, accomplishing tremendous episodes both at the financial and social levels. This general issue needs more idea as for reduce the truth and the rehash of mishap event. The past information about past mishaps speaks to an impressive open door for analysts to distinguish the most persuasive factors in such mishaps, which thusly assume a key job in finding suitable answers for alleviate this issue later on. Lately, a few information mining systems have been successfully used to separate valuable information from enormous informational indexes containing data about car crashes. Street and car crashes are one of the significant reasons for casualty and weakening over the world. Street possibility can be considered as an event in which a conveyance collides with other conveyance, person or other objects. A road risk not only does road hazard cause property damage, it can also lead to partial or complete disability, and can sometimes be fatal to humans. Addition number of street mishaps is definitely not a decent sign for the transport security. The main arrangement requires the investigation of traffic possibility information to recognize various reasons for street mishaps and taking preventive measures. AI is one of the utilizations of Artificial Intelligence procedure where the machine learns the information certainly as opposed to express programming. These days, AI assumes an essential job in our everyday life. It is utilized nearly in each field like vehicle, clinical, banking and so forth in which transport and clinical field has more significance than others as they are identified with lives. In the field of transportation, AI can be applied in numerous segments like traffic stream forecast, mishap expectation, and traveler place proposal and so on. AI is utilized for computerization as well as for wellbeing.

II. RELATED WORK

Many different types of systems have come up earlier to predict traffic on lanes, control accidents on busy roads and to detect the rate of traffic too. All of them work in independent forms. They follow their own individual algorithms and methodologies to solve their complexities. One of them is to detect the traffic on the road by taking input from the cc cameras available. This system uses

machine learning algorithms. They use the detecting mechanism that has gained high scope lately. By using almost similar techniques, a detection of the number of people on roads is carried out. Rather than focusing on the vehicles, this methodology focuses on the people rate to determine the accident prone regions. Through machine learning prediction algorithm, huge amount of data about the regions, places, kind of accidents, number of deaths, time of accident etc. are given to the system and trained. Thus, this system attains the capability to determine the places that are more prone to accidents. This is one of the traditional ways that has been worked upon.

III. LITERATURE SURVEY

Title: A Data Mining Approach for Analyzing Road Traffic Accidents
Author: Zhaoqiang Chen; Qun Chen; Zhanhuai Li
Year: 2019
Description

It completed a data mining structure to recognize, analyze and choose credits adding to road accidents. The guideline purpose of this investigation adventure is to realize a data burrowing structure for dismembering the association between accident characteristics and make proposition for thwarting the high occasion of these setbacks. It evaluated with road incidents data from Khomas area, Namibia. The results show that the use of such a coherent gadget can help in making a data base. The results find that male drivers have enormously added to the higher risk of setbacks, especially, at unions and during light.

Title: Forecasting of Road Accident in Kerala: A Case Study
Author: Ihab F. Ilyas; Xu Chu
Year: 2018
Description

Vehicle crashes are the central explanation behind death and wounds far and wide, fatalities are still moving in many making countries including India. Information Analysis of street fiascos has a solid effect for taking preventive measure to beat the trouble. It kept an eye out for the guess issue of street difficulties utilizing time strategy evaluation over all zones of Kerala. Time course of action appraisal is valuable in finding the models in street mishaps which connects with the figure of future models. In the present MS, it utilized the time plan street mishaps information in Kerala, India for the period January 1999 – December 2016 to comprehend the models in the information and to make fitting model to foresee about future models which may empower specialists to make preventive advances.

Title Predicting the Number of People for Road Traffic Accident on Highways by Hour of Day
Author: Xu Chu; Ihab F. Ilyas; Paolo Papotti
Year: 2019
Description

Predicting the accidents through the number of vehicles on roads in an already researched approach. However, in this paper, the authors have determined the newer method. They predict the number of people on the roads and how they influence the rate of accidents. The focus has been shifted from vehicles to people. Another important parameter, the time of the day is also considered important. The integrated moving model, multiplicative decomposition model, and holtwines are used to detect the people passing. The screw up rate in the end helps perform the final prediction.

Title: Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity
Author: Joeri Rammelaere; Floris Geerts; Bart Goethals
Year: 2018
Description

As demonstrated by World Health Organization report the amount of passing by road fender bender is more than 1.25 million people and reliably with non-deadly setbacks impacting more than 20-50 million people. A couple of segments are contributed on the occasion of road car crash. At this moment, mining portrayal strategies applied to develop models (classifiers) to recognize accident factors and to anticipate auto collision earnestness using as of late recorded traffic data. The accuracy of J48 classifier is higher than various classifiers, anyway simple Bayes gathering execution is better than anything the others whether or not its precision is under J48 and CART classifiers as it is showed up in AUC and ROC results. The outcome of the examination showed that speed limit, atmosphere condition, number of way, lighting condition, and accident time are commonly huge and most remarkable road disaster factors. In the other hand, sex, age, incident zone, and vehicle type are factors that have less impact on road disaster reality.

Title: A Survey on Analyses of Factors Related to Road Accidents Using Data Mining Techniques
Author: Philip Bohannon; Wenfei Fan; Floris Geerts; Xibei Jia; Anastasios Kementsietsidis
Year: 2018
Description

These Analyses made concerning factors influencing street mishaps. Street catastrophes cause vital harms to people. It can cause lifetime wounds. There is a spike in the measure of misfortunes over the advancing years. So it is a basic worry for working environments that administer street security and for the occupants. Street

traffic information must be on an essential level inquired about to pick the fragments that are enduringly identified with street episodes. Fragments like impact type, street condition, light impact, air and alcoholic driver must be considered. Places close to neighborhoods, zebra crossing, and school locale are the basic zone of debacles. This paper makes near appraisal on the assessments done in that limit with regards to the present minute. It proposes sensible methods for analyzing the street debacle dataset. Existing work basically fixates on essentially isolating and discovering association between the parameters influencing street mishaps utilizing the dataset for existing roads. It doesn't consider the starting late coordinated streets and paths a work in progress. This paper targets finding gawky districts for streets that are made game arrangements for future and paths that are being taken a shot at.

IV. EXISTING SYSTEM

There are so many existing systems to control traffic and prevent accidents using previous data sets. Government still follows basic and old traffic control system. There are chances of happening accidents in such cases. To prevent these types of incidents we are going to propose our system. The existing systems have the timer enabled traffic control system, which may lead to accidents in a certain types of conditions like bad weather conditions that is the case that it cannot be controlled manually. It cannot alert the accident zones based on weather condition, type of vehicle. There are some systems that alert the accident zones manually by passing the sign boards through that route but this is an old approach although it can alert the persons who are travelling in that particular route. Automatic systems are not much available in all the places for the prediction of accident zones.

V. PROPOSED SYSTEM

Social occasion learning improves AI results by consolidating two or three models. This methodology permits the creation of better smart execution stood apart from a solitary model and it is the specialty of joining distinctive strategy of understudies (specific models) together to ad lib on the security and insightful power of the model. In the area of Statistics and Machine Learning, Ensemble learning methods endeavor to make the presentation of the farsighted models better by improving their exactness. Outfit learning is a system utilizing which different AI models, are intentionally attempted to manage a specific issue.

VI. MODULE DESCRIPTION

1. DATA VALIDATION AND PRE-PROCESSING TECHNIQUE (MODULE-01)
2. EXPLORATION DATA ANALYSIS OF VISUALIZATION AND TRAINING A MODEL BY GIVEN ATTRIBUTES (MODULE-02)
3. PERFORMANCE MEASUREMENTS OF LOGISTIC REGRESSION AND DECISION TREE ALGORITHMS (MODULE-03)
4. PERFORMANCE MEASUREMENTS OF SUPPORT VECTOR CLASSIFIER AND RANDOM FOREST (MODULE-04)
5. PERFORMANCE MEASUREMENTS OF KNN AND NAIVE BAYES (MODULE-05)
6. CALCULATING THE VOTES BY ENSEMBLE LEARNING METHOD (MODULE-06

DATA VALIDATION AND PRE-PROCESSING TECHNIQUE (MODULE-01)

Machine learning validation technology is applied to get the Machine Learning (ML) model error rate close to the real data set error rate. Data validation is not mostly required for all kinds of data. However, a system that needs to work on real time scenarios, need data validation. Real time scenarios consists of huge amount of data. This type of large volume data holding multiple data types definitely needs to be validated. After data validation, we require to gather information about our data. The process of data collection, and its analysis gives better insights of what exactly the data is. It basically standardizes the data. Finds missing data, handles outliers, classifies and creates labels. It further extracts only the data that is required for our process. The dataset is well balanced and this increases the accuracy rate of our research. Therefore, the entire process of data validation and pre-processing gives us an idea of which algorithm the system must work with.

EXPLORATION DATA ANALYSIS OF VISUALIZATION AND TRAINING A MODEL BY GIVEN ATTRIBUTES (MODULE-02)

Information perception is a significant aptitude in applied measurements and AI. Measurements do in reality center around quantitative depictions and estimations of information. Information representation gives a significant suite of apparatuses for increasing a subjective comprehension. This can be useful when investigating and finding a workable pace dataset and can help with recognizing plans, degenerate data, inconsistencies, and generously more. With a little space data, data portrayals can be used to discuss and show key associations in plots and charts that are more intuitive and accomplices than extents of alliance or centrality. Data portrayal and exploratory data assessment are whole fields themselves and it will recommend an increasingly significant bounce into some the books referenced at the end.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-2 (April 2020)
https://doi.org/10.31033/ijemr.10.2.3

PERFORMANCE MEASUREMENTS OF LOGISTIC REGRESSION AND DECISION TREE ALGORITHMS (MODULE-03)

It is a statistical method for analyzing a data set in which there are one or more distinct outcome-specific variables. The result is calculated using a dichotomo us equation (in which only two possible results exist). The relapse is calculated. Its main target is to make the best model that fits to depict the link between result variable and the illustrative. Calculated relapse is a Machine Learning grouping calculation that is utilized to anticipate the likelihood of an all-out ward variable. In different to the calculated relapse, the strategic relapse, the parallel variable is assigned with the value 1 or 0 accordingly. It is one of the most impressive and famous calculation. Choice tree calculation falls under the class of directed learning calculations. It works for both consistent just as clear cut yield factors.

```
Classification report of Logistic Regression Results:

              precision    recall  f1-score   support

           0       0.99      1.00      1.00       431
           1       1.00      0.98      0.99       207
           2       1.00      0.99      1.00       122

    accuracy                           0.99       760
   macro avg       1.00      0.99      0.99       760
weighted avg       0.99      0.99      0.99       760

Accuracy result of Logistic Regression is: 99.3421052631579

Confusion Matrix result of Logistic Regression is:
 [[431   0   0]
 [  4 203   0]
 [  0   1 121]]

Sensitivity :  1.0

Specificity :  0.9806763285024155
```

Figure 1: Result of Logistic Regression

```
Classification report of Decision Tree Classifier Result

              precision    recall  f1-score   support

           0       1.00      1.00      1.00       431
           1       1.00      1.00      1.00       207
           2       1.00      1.00      1.00       122

    accuracy                           1.00       760
   macro avg       1.00      1.00      1.00       760
weighted avg       1.00      1.00      1.00       760

Accuracy result of Decision Tree Classifier is 100.0

Confusion Matrix result of Decision Tree Classifier is:
 [[431   0   0]
 [  0 207   0]
 [  0   0 122]]

Sensitivity :  1.0

Specificity :  1.0
```

Figure 2: Result of Decision Tree Classifier

PERFORMANCE MEASUREMENTS OF SUPPORT VECTOR CLASSIFIER AND RANDOM FOREST (MODULE-04)

A classifier that sorts the informational index by setting an ideal hyper plane between information. I picked this classifier as it is extraordinarily flexible in the quantity of various kernelling capacities that can be applied and this model can yield a high consistency rate. Bolster Vector Machines are maybe one of the most mainstream and discussed AI calculations. They were incredibly main stream around the time they were created during the 1990s and keep on being the go-to strategy for a high-performing calculation with small tuning.

```
Classification report of Support Vector Machines Results:

              precision    recall  f1-score   support

           0       0.57      1.00      0.72       431
           1       0.00      0.00      0.00       207
           2       0.00      0.00      0.00       122

    accuracy                           0.57       760
   macro avg       0.19      0.33      0.24       760
weighted avg       0.32      0.57      0.41       760

Accuracy result of Support Vector Machines is: 56.71052631578948

Confusion Matrix result of Support Vector Machines is:
 [[431   0   0]
 [207   0   0]
 [122   0   0]]

Sensitivity :  1.0

Specificity :  0.0
```

Figure 3: Result of Support Vector Machine

```
Classification report of Random Forest Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00       431
           1       1.00      1.00      1.00       207
           2       1.00      0.99      1.00       122

    accuracy                           1.00       760
   macro avg       1.00      1.00      1.00       760
weighted avg       1.00      1.00      1.00       760

Accuracy result of Random Forest is: 99.86842105263159

Confusion Matrix result of Random Forest is:
 [[431   0   0]
 [  0 207   0]
 [  0   1 121]]

Sensitivity :  1.0

Specificity :  1.0
```

Figure 4: Result of Random Forest Algorithm

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-10, Issue-2 (April 2020)
https://doi.org/10.31033/ijemr.10.2.3

PERFORMANCE MEASUREMENTS OF KNN AND NAIVE BAYES (MODULE-05)

K-Nearest Neighbor is a managed AI calculation which stores all cases compare to preparing information focuses in n-dimensional space. At the point when obscure discrete information is gotten, it breaks down the nearest k number of occasions spared (closest neighbors) and returns the most widely recognized class as the expectation and for genuine esteemed information it restores the mean of k closest neighbors. Out yonder weighted closest neighbor calculation, it loads the commitment of every one of the k neighbors as indicated by their division using the going with question giving progressively essential burden to the closest neighbors The Naive Bayes count is an instinctual method that uses the probabilities of each acknowledge having a spot for each class to make a figure. It is the overseen learning approach you would consider in case you expected to exhibit a judicious showing issue probabilistically. The probability of a class regard given an estimation of a quality is known as the prohibitive probability. By expanding the prohibitive probabilities together for every quality for a given class regard, we have a probability of a data event having a spot with that class. To make a figure we can register probabilities of the case having a spot with each class and select the class a motivating force with the most raised probability.

```
Classification report of K-Nearest Neighbor Results:

              precision    recall  f1-score   support

           0       0.57      0.72      0.64       431
           1       0.28      0.24      0.26       207
           2       0.38      0.11      0.18       122

    accuracy                           0.49       760
   macro avg       0.41      0.36      0.36       760
weighted avg       0.46      0.49      0.46       760

Accuracy result of K-Nearest Neighbor is: 49.21052631578947

Confusion Matrix result of K-Nearest Neighbor is:
 [[310 103  18]
 [152  50   5]
 [ 82  26  14]]

Sensitivity :  0.7506053268765133

Specificity :  0.24752475247524752
```

Figure 5: Result of K-Nearest Neighbor

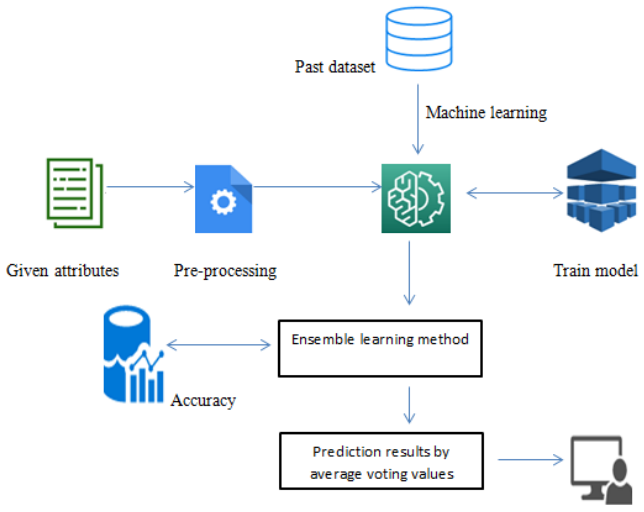
```
Classification report of Naive Bayes Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00       431
           1       1.00      1.00      1.00       207
           2       1.00      1.00      1.00       122

    accuracy                           1.00       760
   macro avg       1.00      1.00      1.00       760
weighted avg       1.00      1.00      1.00       760

Accuracy result of Naive Bayes is: 100.0

Confusion Matrix result of Naive Bayes is:
 [[431   0   0]
 [  0 207   0]
 [  0   0 122]]

Sensitivity :  1.0

Specificity :  1.0
```

Figure 6: Result of Naïve Bayes

CALCULATING THE VOTES BY ENSEMBLE LEARNING METHOD (MODULE-06)

Tkinter is a python library for creating GUI (Graphical User Interfaces). We utilize the tkinter library for making a use of UI (User Interface), to make windows and all other graphical UI and Tkinter will accompany Python as a standard bundle, it tends to be utilized for security motivation behind every client or bookkeepers. There will be two sorts of pages like enlistment client reason and login passage motivation behind clients.

VII. SYSTEM ARCHITECTURE



The proposed system initially gathers all the required parameters. All the gathered data is further pre-processed. Data cleaning, transformation and integration is

done one after the other. The pre-processed data is analyzed and then visualized accordingly. This data is used to train the model and then come up with a prediction algorithm. The performances of all these algorithms are calculated. For accuracy, the ensemble learning method is applied.

VIII. FUTURE ENHANCEMENT

Transport Department needs to automate the recognizing the best course by not disaster from capability process (consistent) in perspective on the record detail. To automate this strategy by show the desire achieves web application or work territory application.

IX. CONCLUSION

A more flexible traffic signal has been updated for a mixed type of traffic. Mixed traffic consists of both manual and automated vehicles. For AVs, an efficient speed changing procedure is advanced responsively. The proposed system maintains the present traffic to be smooth and content with crossing point of genuinely decided vehicles. A number of parameters are considered in the system. These parameters are the main source on determining whether that traffic is an accident prone region or non-accident prone region. The user interface has been well developed where the user has the facility to choose the algorithm he wants to work with. However, the system is capable of working on all algorithms, a few being K-nearest, Random forest, Logistic Regression, and chooses the one that produces the highest accuracy. This facilitates the user to receive a highly reliable response. Thus the system turns out to be more trustworthy. Based on the result the user can avoid going to the accident prone region. This system works well for even normal people in their day to day routine. It can be developed to a smaller scale for the user's easy accessibility.

REFERENCES

[1] K. G. Boldness & S. M. Parapar. (1975). Postponement and fuel utilization at traffic signals. *Inst. Traffic Eng., 45*(11), 23–27.
[2] T.- Q. Tang, Z.- Y. Yi, & Q.- F. Lin. (2017). Impacts of sign light on the fuel utilization and emanations under vehicle following model. *Physics A, Statistical Mechanics and its Applications, 469*, 200–205.
[3] N. H. Gartner. (1982). Advancement and testing of an interest responsive system for traffic signal control. *In Proceedings of the American Control Conference*, pp. 578–583.
[4] N. H. Gartner, F. J. Pooran, & C. M. Andrews. (2001). Usage of the OPAC versatile control methodology in a rush hour gridlock signal system. *In Proceedings of Intelligent Transpiration Systems*, pp. 195–200.
[5] A. G. Sims & K. W. Dobinson. (1980 May). The Sydney facilitated versatile traffic (SCAT) framework reasoning and advantages. *IEEE Transactions of Vehicular Technology, VT-29*(2), 130–137.
[6] D. I. Robertson & R. D. Bretherton. (1991 Feb). Upgrading systems of traffic flags progressively the SCOOT strategy. *IEEE Trans. Veh. Technol., 40*(1), 11–15.
[7] S. Sen & K. L. Head. (1997). Controlled streamlining of stages at a crossing point. *Transp. Sci., 31*(1), 5–17.
[8] B. Asadi & A. Vahidi. (2011 May). Prescient journey control: Utilizing up and coming traffic signal data for improving mileage and diminishing excursion time. *IEEE Trans. Control Syst. Technol., 19*(3), 707–714.
[9] M. A. S. Kamal, S. Taguchi, & T. Yoshimura. (2015 Sep). Convergence vehicle helpful eco-driving with regards to somewhat associated vehicle condition. *In Proc. Int. Conf. Intell. Transp. Syst. (ITSC)*, pp. 1261–1266.
[10] S. I. Guler, M. Menendez, & L. Meier. (2014 Sep). Utilizing associated vehicle innovation to improve the proficiency of crossing points. *Transp. Res. C, Emerg. Technol., 46*, pp. 121–131.
[11] J. Rios-Torres & A. A. Malikopoulos. (2017 May). A study on the coordination of associated and mechanized vehicles at crossing points and converging at thruway entrance ramps. *IEEE Trans. Intell. Transp. Syst., 18*(5), 1066–1077.