International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-10 Issue-2, December 2020

Real Time Efficient Accident Predictor System using Machine Learning Techniques (kNN, RF, LR, DT)

P. Tamije Selvy, M. Ragul, G. Naveen Vignesh, M. Anitha

Abstract: ***Real time crash predictor system is determining frequency of crashes and also severity of crashes. Nowadays machine learning based methods are used to predict the total number of crashes. In this project, prediction accuracy of machine learning algorithms like Decision tree (DT), K-nearest neighbors (KNN), Random forest (RF), Logistic Regression (LR) are evaluated. Performance analysis of these classification methods are evaluated in terms of accuracy. Dataset included for this project is obtained from 49 states of US and 27 states of India which contains 2.25 million US accident crash records and 1.16 million crash records respectively. Results prove that classification accuracy obtained from Random Forest (RF) is96% compared to other classification methods.***

Keywords: ***Machine Learning, Accident Prediction, classification Techniques,***

I. INTRODUCTION

Road accident is one of the most important ongoing issues in the modern times traffic on roads. WHO has reported top ten disastrous reasons for taking human's life, and unfortunately road accidents come at ninth place. Crash predictor system is used to predict the accidents in the roads with the help of machine learning algorithms [1]. Accidents cause a huge impact on the society, where there is a great cost of casualties and injuries to the people. To avoid those accidents crash predictor systems can be used.

A. Machine Learning

Machine learning is used to discover new knowledge from the large amount of database[2].There are four different categories in machine learning and those categories are Supervised Learning, Unsupervised Learning, Semi-Supervised Learning and Reinforcement Learning. The steps followed in machine learning are as follows: 1.Data gathering 2.Data preparation 3.Choosing a model and training, 4.Evaluation, 5.Tuning and prediction. Artificial Intelligence is a program that can sense, reason, act and adopt. Machine learning algorithmsperformance increases as they are exposed to more data over time.

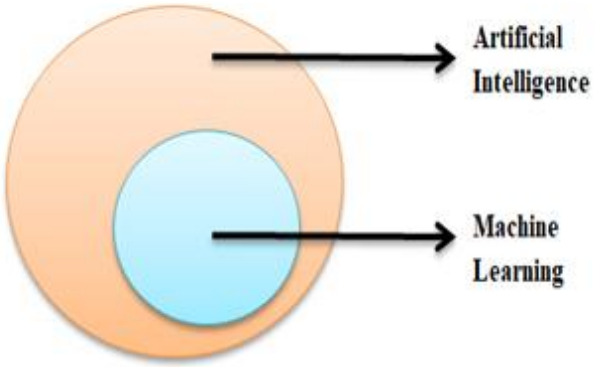

Fig.1 Concepts of Machine Learning

Machine learning which a subset of artificial intelligence isis depicted in fig 1.

II. LITERATURE SURVEY

TibebeBeshahTesemaet al [3] implemented the complex combination of characteristics like mental state of driver, road condition, weather conditions, traffic and other attributes. They used algorithms such as Naïve Bayes, k-Nearest Neighbors (kNN) and Decision Tree(DT). Then the final outcome provides the maximum performance of 86.25% for DT classifier. Ramani, R. G.,et al**.** [4], research work was in exploring the application of data mining techniques to aid in the prediction of road patterns related to pedestrian characteristics. In their study they implemented the DT algorithms viz, Random tree, C4.5, J48 and Decision Stump are applied to a database fatal accident occurred.

Mohamel et al [5], identified the cause of accident and the driver who committed for the accident and he implemented Data mining techniques to predict the cause of road accidents. Results obtained by him shows that his model can predict the road accident with the accuracy greater than 75% (SVM).

Navadaet al [6], used Neural network, DT, SVM and hybrid decision tree – neural network based approaches used to predict road accidents. The final outcome for their study was hybrid approach performance was better than neural network.

Revised Manuscript Received on December 15, 2020.
* Correspondence Author
Dr. P. Tamijeselvy*, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore-641042, India. Email: p.tamijeselvy@skct.edu.in
M. Ragul, Department of Computer Science and Engineering, Sri Krishna College of Technology,Coimbatore-641042, India. Email: 16tucs146@skct.edu.in
G. Naveen Vignesh, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore-641042, India. Email: 16tucs129@skct.edu.in
M. Anitha, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore-641042, India, Email: anitha88.kool@gmail.com

They mainly showed importance for fatal and non – fatal injuries. Then their experiment demonstrated that the model for fatal and non-fatal injury performed better than other classes.

V. Patel et al[7],she made a study on combination of the attributes (Fatal Vs weather, Time Vs Day etc.). Number of accidents reported to the ambulance was also recorded in her study. K. Thirunavukkarasuet al[8], used three different approaches: classifier fusion based on the Dempster-Shafer algorithm, the Bayesian procedure and logistic model and clustering based on k-means algorithm.

III. PROPOSED FRAMEWORK

In this study, four different machine learning models were used for classification logic. Those used models are supervised learning algorithm. The crashes are classified different levels according to the injury level.

The accident dataset has been collected, and the collected data has been analysed, integrated and grouped together based on different constraints using the best suited algorithm. This can be used to analyse and identify the flaws and reason for the accidents and can avoid those accidents in future.

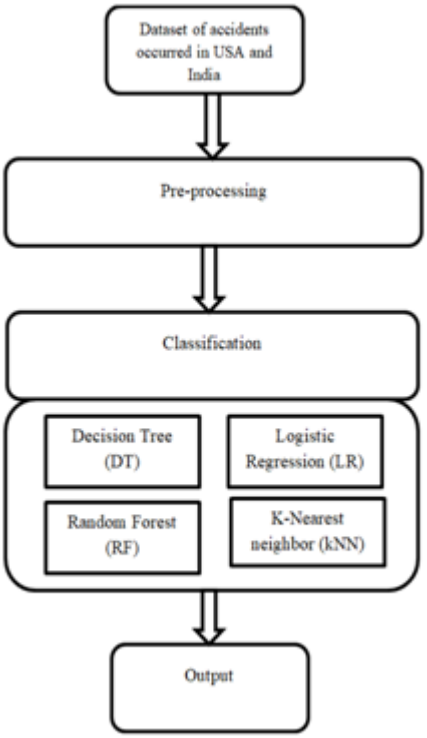

Fig. 2 Flowchart of proposed system

The above flow chart fig. 2 represents the process involved in efficient crash predictor system using machine learning techniques

A. Dataset

In our study, crash injury data has been obtained from 49 states of US and 27 states of India. Three year crash dataset was collected, a total of 1155332 & 28533 crashes were used to analyse the study.

Different attributes has been collected for the each dataset like weather, road condition, area type, No. Of vehicles involved pedestrians, sign boards etc.

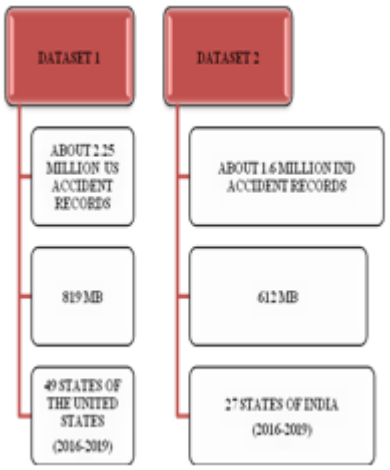

Fig 3 Accident dataset of US

Fig 4 Accident dataset of India

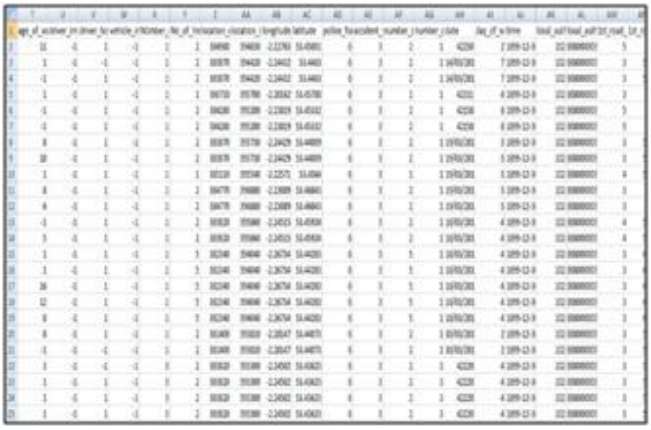

Fig 5 Collected accident record for US

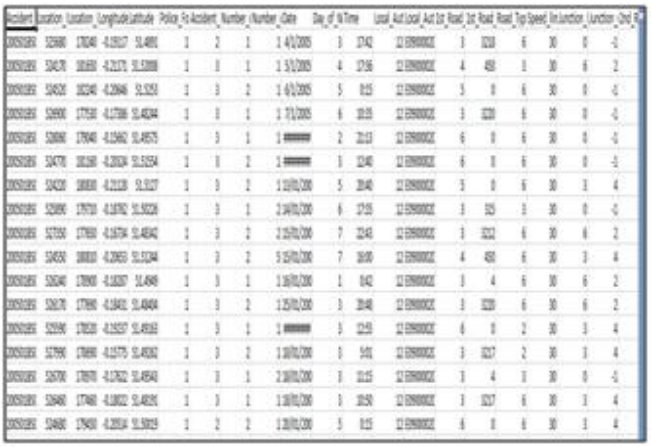

Fig 6 Collected accident record for India

Above fig 5 represent the collected dataset 1and fig 6 represent the collection of dataset 2.

B. Pre-processing

Data Pre-processing is the most important part of Machine Learning system. In this proposed work implementation for pre-processing is done with the help of python programming. The major advantages of preprocessing are increase in accuracy. Data visualization is increased while the data are processed using preprocessing methods. In our study two different datasets has been collected and preprocessed.

Python programming has been used to remove the invalid data from the dataset, which was helpful for gathering valid data.

IV. CLASSIFICATION OF ML METHODS

A. K-Nearest Neighbours

The K-Nearest Neighbors(KNN) algorithm is used to classify similar group of accidents. KNN is used to classify the dataset with their similarity level, similar classes will be grouped together [8]. This is done by a majority vote by the k closest points that have been observed. KNN algorithms use data and classifies new data points based on some similarity measures like distance function. The accuracy graph of KNN algorithm is shown in below fig 7.

Pseudo code:

1. Load and train the dataset and test data.
2. Choose the k value.
3. Find the Euclidean distance.
4. Store the Euclidean distance.
5. Choose the first k point.
6. Assign the class to test point.
7. End.

Fig 7Performance analysis of kNN

B. Decision Tree

Decision tree is one of the oldest algorithms in machine learning models which is commonly used of classification purpose. A flow chart like structure is used for DT classification for better understandings.

Decision Tree algorithm is also a supervised learning model. DT algorithm can be used for solving certain regression and classification problems [9]. The main purpose of using DT is to create a training model, which is used to predict class or value of target variables by learning from prior data available. Decision Tree includes two entities namely decision nodes and decision leaves. Execution of DT is comparatively easy. As DT is depicted using tree representationclassification problems can be solved using Decision tree.

Pseudo code:

1. Best attribute of dataset is root of the tree.
2. Training set is split into subsets.
3. Subset contains data with the same value of an attribute.

Repeat steps 1 and 2 until you find leaf nodes in all branches of tree.

C. Logistic Regression (LR)

Logistic Regression (LR) which is an extension of Linear Regression is used to find the probability of a particular event to occur. It belongs to supervised learning model.This method is named as LR because its basic functionalities are similar to Linear Regression. In generalLogistic regression includes three main types. They are binary, multimodal and ordinal logistic regression [10].Logistic regression is not only considered as efficient classification model but also provides probabilities. When this type of regression is applied for binary classification it is termed as Multinomial Regression.

D.Random Forest (RF)

Random Forest (RF)which is a supervised learning algorithm is used for classification and regression. This method is most effective for large databases. The RF algorithm isusually trained by the method called bagging. When compared to Decision Tree (DT), random forest is the best because, in decision tree the dataset will be split into a single tree but in random forest the data set is split into multiple trees for better classification. Hence in terms of performance analysis, RF provides better accuracy rate of 0.96 compared to other classification models.

V. PERFORMANCE ANALYSIS

This study made an analysis between the classification algorithms where the Random Forest (RF)provides the highest accuracy for crash prediction for datasets obtained from US and India. The prediction accuracy of each algorithm (kNN, DT, LR, RF) was compared and analysed. Performance analysis in terms of accuracy is as follows: Random forest provides highest accuracy of 96%, Logistic Regression provides accuracy of 94%, Decision Tree and K Nearest Neighbors with accuracy level of 93% as shown in fig 8. The resultsreveal that various machine learning algorithms can achieve a similar predicting accuracy.

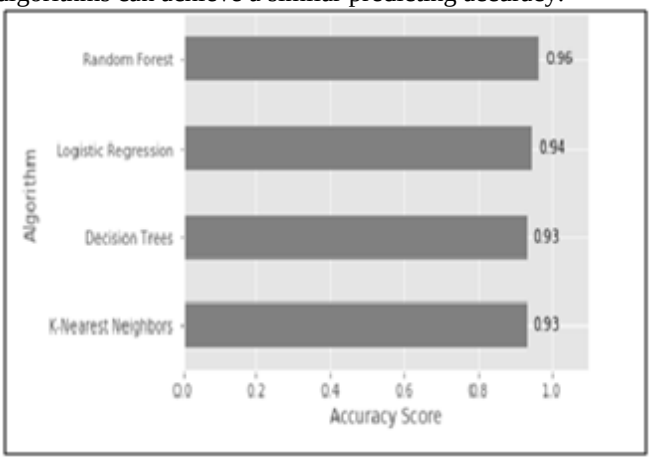

Fig 8.Performance analysis of RF, LR, DT, KNN.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

VI. CONCLUSION

The study is based on comparison of crash injury predictive performance between various machine learning algorithms. Based on crash data collected at freeway areas each and every accident was noted. The predicting accuracy of each training set and test set was calculated and compared. Machine learning models always produced better performance then statistical models. The cost of fatal injuries and driver injuries due to traffic accidents greatly affect the people of the society. This paper contributes various existing survey and works carried out by researchers. RF and Logistic Regression were found to be the best models that had the highest overall predicting accuracy which were 96% and 94% respectively. The future scope of this work is to consider sentiment analysis of road accidents using ensemble classifiers and deep neural algorithms.

REFERENCES

1. F. Galatioto, M. Catalano, N. Shaikh, E. McCormick and R. Johnston, "Advanced accident prediction models and impacts assessment," in IET Intelligent Transport Systems, vol. 12, no. 9, pp. 1131-1141, 11 2018.
2. P. Tamijeselvy, V. Planisamy, S. Elakkiya, “A novel approach for the prediction of epilepsy from 2D medical images using case based reasoning classification model, WSEAS TRANSACTIONS on COMPUTERS-2013.
3. TibebeBeshahTesema, Ajith Abraham, DejeneEjigu, "Learning the Classification of Traffic Accident Types", *copyright IEEE*, 2012
4. Ramani, R. G., & Shanthi, S. (2012). Classifier prediction calculation in modeling road traffic accident data. International Conference on Computational Intelligence and Research.
5. Mohamed, E. A. (2014). Predicting causes of traffic road accidents using multi-class support vector machine algorithm, 11(5), 441–447.
6. A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," 2011 IEEE Control and System Graduate Research Colloquium, Shah Alam, 2011, pp. 37-42.
7. RANDOM FOREST-S. V. Patel and V. N. Jokhakar, "A random forest based machine learning approach for mild steel defect diagnosis," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-8.
8. KNN-K. Thirunavukkarasu, A. S. Singh, P. Rai and S. Gupta, "Classification of IRIS Dataset using Classification Based KNN Algorithm in Supervised Learning," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4.
9. P. Tamijeselvy, V. Palanisamy, MS.Radhai, ”A proficient clustering technique to detect CSF level in MRI brain images using PSO algoritm”, WSEAS TRANSACTIONS on COMPUTERS -2013.
10. Iranitalab, A., &Khattak, A. (2017) comparison of machine learning methods for crash severity prediction. Accident analysis & prevention, 108, 27-36.

AUTHORS PROFILE

Dr. P. Tamijeselvy, is working as professor in the Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore Her areas of interests are image processing, data mining and artificial intelligence. Her research work includes medical imaging and social data mining.

M. Ragul, UG student in Sri Krishna College of Technology in the Department of Computer Science and Engineering.

G. Naveen Vignesh, UG student in Sri Krishna College of Technology in the Department of Computer Science and Engineering.

M. Anitha, Full-Time Ph.D Scholar, in Sri Krishna College of Technology in the Department of Computer Science and Engineering. Her areas of interest are image processing and data mining.