Predicting Severity of Traffic Accidents using Classification Algorithms

1st Palash Dange
Department of Computing
Dublin City University
Dublin Ireland
palash.dange2@mail.dcu.ie

2nd Anshika Sharma
Department of Computing
Dublin City University
Dublin Ireland
anshika.sharma27@mail.dcu.ie

3rd Aniruddha Kulkarni
Department of Computing
Dublin City University
Dublin Ireland
aniruddha.kulkarni4@mail.dcu.ie

***Abstract*—In many developed countries, traffic accident are the first cause of death. However, there is lot of public data available that can be used to determine the severity and cause of accidents by building a predictive model. This research paper compares three different prediction algorithms logistic regression, KNN and Decision Tree to predict the severity of accidents. In our approach feature extraction was performed where we applied PCA on dataset. There were 31 total features in dataset which were reduced to 15 to perform further analysis. All the three models are explained and compared using ROC curve that was calculated during the research.We found that the prediction of severity for Logistic Regression and KNN was approximately identical. Please follow the link to view google colab notebook. GitHub Link**

***Index Terms*—Road Accidents, Traffic, Congestion, Classification, Neural Network, KNN, Decision Tree, Regression**

I. Introduction

Traffic accidents in the current era have become severe threats for the human beings. Lot of research is being conducted in the automobile industry to make it more safe and secure, but these accidents are unavoidable. The cost of fatalities and injuries have a great impact on economic well being of the society. Although, many developed countries like USA has certain measures in place to avoid such accidents but the rate of accidents is still high. Many theories can be deployed to reduce the fatalities during road accidents, one of them being predicting severity of the accidents. Here, Data mining approach along with different technologies can be used to understand the severity of these accidents. Data Mining is a process to extract and understand the correlation, patterns between the data to predict the end results. It includes Collection of massive data, Different data mining algorithms. Given the significance of this technique, there have been number of researches to predict severity of accidents. The shortcoming of it being that the dataset was limited or small to conclude. The dataset that we are working on “US Accidents” includes 3 million traffic incidents that took place within USA between 2016 to 2019. It offers a wide range of attributes such as time data, description, weather conditions, location. The main objective of the research is to investigate the role of these variable factors for accident severity using predictive models. Dataset can be found on the given link: Dataset- US Accidents

II. Literature Review

Quaseem A, proposed an approach of predicting the severity of accidents in smart cities. Dataset used by the author has 49751 records over 31 features. Due to the skewness in data, this research uses imbalance techniques like oversampling and undersampling to deal with the majority and minority of classes. Even the hybrid approach is used to deal with the imbalanced class problem, which involves using both oversampling and undersampling. Once the attributes are finalized, then classification algorithms like Decision tree, SVM and ANN are used to design the predictive models. Evaluation measures like Accuracy, Recall, Precision, and F1 measures are used in the paper to give the best results. The experiment shows the Decision tree (Random Forest) has the highest accuracy of 80.65%, precision, and recall of .814% and 0.806% respectively. All these results were achieved by a hybrid sampling dataset. This paper also presented the PART algorithm which uses rules to present the knowledge. Several rules based on the dataset were designed to use the PART algorithm.

Hamzah Al Najada, predicted the severity of accidents on Big Data using the classification approach. Technology growth and human comfort have led to an increase in traffic on roads. Increasing traffic may lead to an increase in road accidents. This paper has used several tools for analysis like WEKA and H20 for Big Data. Two datasets are used by this paper one is of Accidents with 31 attributes and 146322 records and the other is causalities dataset with 15 attributes and 123988 records. Several preliminary tests were performed on the data, it was found that only 9 features of accidents data and 8 attributes of causalities data were of use for prediction. Naïve Bayes classifier was used to design and predict the accident severity. This research involves 10 fold cross-validation, means the data is tested and validated 10 times before publishing. The goal of this research was to build an efficient model instead of high accuracy, even though Decision tree was giving better accuracy value for 50 trees, the computation time of the model was very high as compared to Naïve Bayes and C4.5. Accuracy of C4.5 and Naïve Bayes was 84% and 80% respectively. For

AUC(Area under ROC curve) Naïve Bayes performed well as compared to C4.5 resulting in an accuracy of 70% and50% respectively. This paper concludes how human behavior and feature affects accidents and congestion on the road. E.g. attributes of the driver like sex, age and type can also affect accidents and traffic congestion greatly.

Ethiopia suffers from the highest number of road traffic accidents (RTAs) worldwide each year. Tibebe Beshah and Shawndra Hill analyzed the traffic system of Ethiopia. It was found that not only drivers but also other road-related elements contributed to RTAs. The authors extended the previous work and studied the severity of accidents based on road-related factors. They divided the research into three parts; explored the underlying variables (especially road-related ones) impacting car accident severity, used data mining techniques to predict accident severity, and compared standard classification models for this task. The relevant attributes were selected through feature selection. The classification was successfully achieved with an approximately similar accuracy of 80.818% by K-Nearest Neighbours, 80.221% by Decision trees and 79.996% by Naïve Bayes Classifier. The ROC curve was created as the results cannot just rely on the accuracy of the prediction, also the accuracies were all roughly similar. K-nearest neighbor displayed the closest AUCs value to 1.0. To make it easier for the end-users to understand, classification rules were generated using PART. The accuracy of this algorithm in making rules was 79.942. The outputs of the models were presented for analysis to domain experts for feedback. This would eventually help the stakeholders in the right decision making.

Liling Li, Sharad Shrestha, Gongzhu Hu in the analysis of road traffic accidents applied statistical analysis and implemented data mining techniques on the FARS dataset. In the data preparation, they replaced the numeric values with the nominal ones, removed around 99 records which had missing values whereas others were calculated using some independent variables. Three modeling techniques namely K-Means Clustering, Naive Bayes classifier, and Apriori algorithm were performed. Through association rule mining significant relations were identified in the dataset using the Apriori algorithm. 13 rules were generated with a minimum support= 0.4 and minimum confidence= 0.6 Records were classified into different categories based on Naïve Bayes Classifier with an accuracy of 67.95%. 48 US states were clustered into 3 clusters and then suggested driving measures to diminish the number of accidents. Different factors played an important role in contributing to the severity of the accidents. Head-on collisions resulted in a high number of fatal accidents as compared to others. The average speed limit of 55-70 km/hr, weather conditions like bright daylight, clear sky and dry road individually resulted in a greater number of accidents than others.

S.Ramya, SK. Reshma, V. Manogna, Y. Saroja, Dr. G.S. Gandhi proposed a classification technique called the Random Forest algorithm which is used to identify patterns and classify types of accident severity. This severity is identified using environmental features of RTAs which later can be used to build the model. This model was implemented on a realtime dataset with features such as Number of vehicles, casualties, Road type, Speed limit, light conditions. TO enhance the decision-making process, a decision system was built using the above-generated model. The authors here compared 3 different models J48, ANN and Random Forest Classifier. Random Forest had the highest accuracy with results 78.9%, 62.5% and 49.8% on hybrid, oversampling and Undersampling respectively. This research offers a prediction model by analyzing the relationship between accidents and parameters affecting them.

Laura Garcia Cuenca, Enrique Puertas, Nourdine Aliane, Javier Fernandez Andres proposed a case study of traffic accidents, their classification and severity prediction in Spain. Three machine learning algorithms such as Gradient Boosting Trees, Deep Learning, Naive Bayes have been compared here. The Dataset used for this case study is 6 years Raw data from Spanish traffic agencies ranging from 2011 to 2015. To evaluate the results from a comparison of three algorithms 3 metrics i.e. Precision, Accuracy and F-measure were used. The best result to obtain traffic accident classification is given by the Deep Learning algorithm. Here authors are looking towards building a prediction model to determine the severity of an accident.

This paper is based on the prediction of the severity of accidents in Hong Kong. This research used accident data from the TRADS a new computerized to replace the existing one which is TRACIS - Traffic accident statistical system. After implementing this system the transport department has seen a decreasing trend of accident i.e. from 15 deaths to 11 deaths per year in 2001. This indicates well-maintained traffic after implementing TRADS. In general, the TRADS database includes all the information on accidents including factors like District Factor, Human Factor, Site Factor, and Environmental Factor. This information is taken and linked to the accident report booklet and tracked further from there. The statistical analysis on TRADS began with a chi-square contingency test where the author used vehicle type as a reference to access association between injury severity and predictor variable. This paper uses multivariate logistic regression for determining the significant factors affecting the injury severity of Hong Kong. This paper plotted a table of Chi-square and Cramer's V and plotted rules and suggestions on it. This paper concluded with a discussion describing how important various factors are for predicting the severity of injury using the TRADS data

III. Methodologies

A. Data Preprocessing

Dataset used in the research is of accidents occurred between 2016 to 2019 in US. Visualizing the data we found that the frequency of number of accidents is higher in few stats of US compared to other. Frequency of congestion and accidents is seen in the data of 5 major cities. These cities include California(CA), Texas(TX), Florida(FL), Southern Carolina(SC), and New York(NY).From the Fig 1 we can see that two of our continous variables from the dataset are normally distributed, other variables are skewed variable (Please refer pairplot from GitHub for more details).

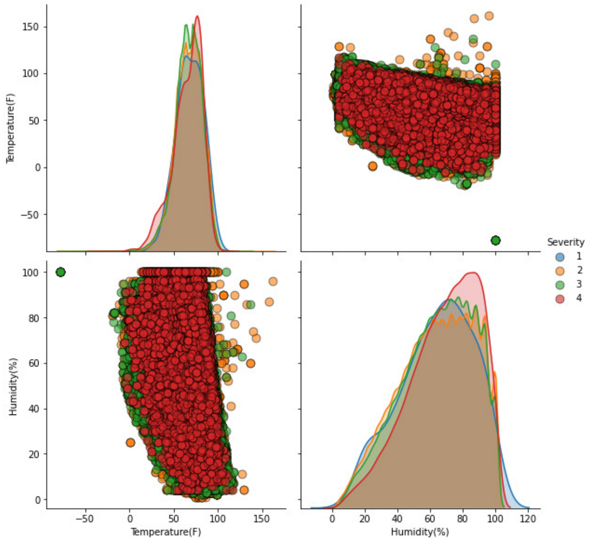

Fig. 1. Humidity and Temperature distribution w.r.t Severity

Data cleaning involves filling of missing values using the mean and mode imputer. The categorical and binary values are filled by the mode of column. A descriptive analysis of variables was done to check the Q1,Q2 and Q3 of the data. After analysis of descriptive data, continuous variable are filled by mean after visualizing the boxplot of those columns. One of the model KNN is implemented using the KNN imputer to fill the values. Total number of features are 31 in the dataset. Categorical data is encoded using LabelEncoder(). Pair plot was created to check the distribution of Severity of accidents using seaborn library of Python.

B. Feature Extraction

Total number of features in data were 31 to reduce the number of variables in dataset we applied PCA on the dataset. As the dataset was mix of both categorical and continuous variable we need to convert the Categorical to continuous features and then apply PCA on the dataset. After checking the explained variance ratio, 15 number of features cover around 91

C. Classification Algorithms

Different models were applied on the dataset once the dataset is clean and ready for training. The dataset was split into 70-30 using python. The models applied on the dataset include - Decision Tree using classifier and Regressor, KNN and Logistic Regression.

I Decision tree :
Decision Tree used for both categorical and continuous data. This algorithm is a set of if-else rules. It is used for both regression and classification models. This model deals with the categorical variable, in our dataset as well we have severity as our classification/categorical variable this algorithm can be used. As the name suggests, it is in the form of a tree structure, which has both the root node and leaf nodes. The topmost node is the root node which is the best predictor variable. The roots predict the In accidents dataset we have used the decision tree classifier to predict the severity.

II Multinomial/Multiclass Logistic Regression :
This algorithm is used to predict the classification variable present in the dataset, as severity ranges from 1-4 which is also a classification variable. Multinomial Logistic Regression is also called a multiclass model. This model is preferred over other models as this does not require any scaling or any other tuning to apply. Feature engineering plays an important role in this model. We have implemented this algorithm after applying PCA on the dataset and selecting features on it. we have also plotted the confusion matrix and ROC curve for the same which can be seen in the collab file shared. This algorithm uses the sigmoid function.

$$\sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Fig. 2. Sigmoid function

III KNN - (K Nearest neighbor)
KNN is a supervised learning algorithm that works on the idea of clustering. It is termed as a non-parametric algorithm, which does not make any assumption regarding the data distribution pattern. For the new data point, this algorithm identifies points closest to it and calculates the distance. For classification, it works on the count majority way, i.e. it counts the number of points in each category among the neighbors. The new point considered will belong to the class which has the highest number of neighbors. In our problem, the severity is plotted in cluster format and the KNN algorithm gives a good result with 6 neighbors (by hit and trial). The algorithm then checks the number of neighbors around it and plot the new data point which has the highest count of class around it. This algorithm uses feature similarity to predict what category the new class falls into.

D. Results

After performing a series of experiments, thorough data preprocessing resulted in a set of 1.4 million records with no missing values. The Severity had 4 class labels namely: 1, 2, 3, and 4. The models applied to predict the dataset are Decision Tree Classifier; K Nearest Neighbours and Logistic Regression. Decision Trees are easy to construct and understand, can handle continuous as well as categorical data, and can perform both classification and regression. Both the decision tree and KNN can be implemented as a regressor and classifier. But Logistic Regression is regression analysis for categorical variables. When analyzing the accuracy, the result generated were approximately similar for all the classifiers. As can be seen in the table the accuracy of the decision tree is 0.64 whereas it is similar for KNN and Logistic Regression, 0.68.

The analysis of precision, recall, f1-score, and support imply that Logistic Regression resulted in better outputs as compared to the other two. Since the results are still not clear we plotted the ROC curve, known as the Receiver Operating Characteristic curve. It compares two operating characteristics FPR and TPR as the criterion changes.

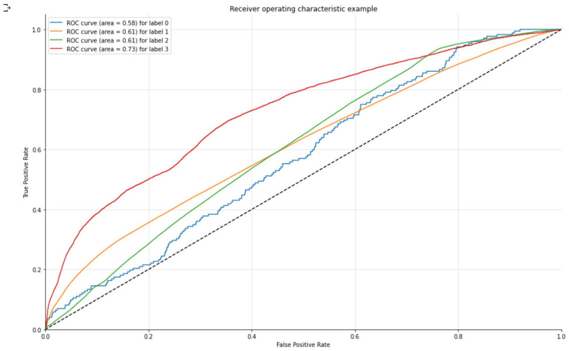

Fig. 3. ROC for Logistic Regression

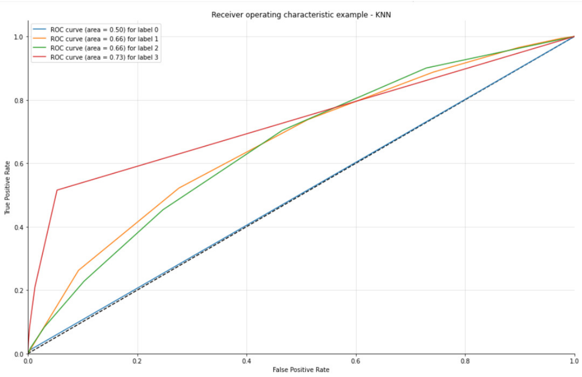

Fig. 4. ROC for KNN

It is represented by plotting the fraction of true positives (TPR = true positive rate) versus the fraction of false positives (FPR = false positive rate). It helps to select the optimal model and discard the suboptimal ones. The Area Under the ROC Curve (AUC) determines the strength of prediction. The individual class's AUC is present in the tables below for each of the three models.

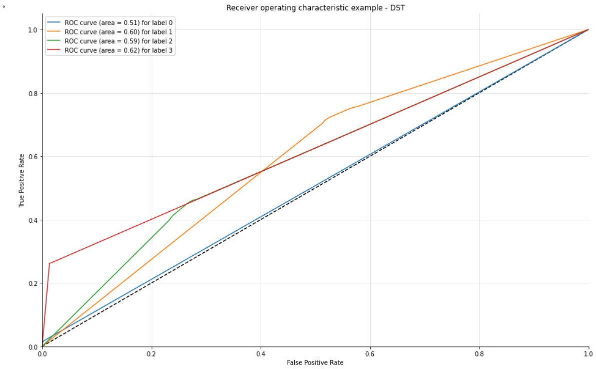

Fig. 5. ROC for Decision Tree

TABLE I
ACCURACY FOR EACH MODEL

Sr. No.	Accuracy
Logistic Regression	64%
K Nearest Neighbours	68%
Logistic Regression	68%

TABLE II
ROC AUC VALUES FOR EACH MODEL

Models		Classes		
	0	1	2	3
Logistic Regression	0.58	0.61	0.61	0.73
K Nearest Neighbours	0.5	0.66	0.66	0.73
Decision Tree Clasifier	0.51	0.6	0.59	0.62

The mean AUC value of the three models is 0.6325 for Logistic Regression, 0.6375 for K Nearest Neighbours, and 0.58 for Decision Tree Classifier. We noticed a similar pattern when we compared the models based on their accuracy. So, we can conclude that the prediction of classes of Severity for both the Logistic Regression classifier and KNN was approximately identical.

E. Conclusion

Road traffic accidents are one of the major concerns these days. The magnitude of its impact is huge, prediction of its impact has become a necessity. In this paper we analysed the traffic accident data to detect the severity of accidents considering various factors. The two grounds on which this research was carried out were: first, the implementation of 3 different models to identify the severity of accidents examining several factors; second, comparison of the performance of the respective models based on the precision of prediction. It was not possible to incorporate all the explanatory variables features in the models due to multi collinearity in the traffic accident variables So, through PCA 30 features were reduced to 15 for the final analysis. The results can further be used to enhance road safety and incorporate several measures. The next step can be to integrate human factors. The age, sex, experience in driving vehicles can also be considered as risk factors. Apart from environmental factors, analysing the data based on 4 seasons of the year might provide better and new results which can be utilized to improve the current situation. Through this task, we realised that no data is enough to make a strong decision.

REFERENCES

[1] Al Najada, H. and Mahgoub, I., 2016, September. Big vehicular traffic data mining: Towards accident and congestion prevention. In 2016 International Wireless Communications and Mobile Computing Conference (IWCMC) (pp. 256-261). IEEE.

[2] Sobhan Moosavi (2019, December). US-Accidents: A Countrywide Traffic Accident Dataset, Version 2. Retrieved February 16, 2020 from https://www.kaggle.com/sobhanmoosavi/us-accidents

[3] Al-Radaideh, Q.A. and Daoud, E.J., 2018. Data mining methods for traffic accident severity prediction. Int. J. Neural Netw. Adv. Appl, 5, pp.1-12.

[4] Beshah, Tibebe, and Shawndra Hill. ”Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia.” 2010 AAAI Spring Symposium Series. 2010.

[5] Li, Liling, Sharad Shrestha, and Gongzhu Hu. ”Analysis of road traffic fatal accidents using data mining techniques.” 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA). IEEE, 2017.

[6] Ramya, S., Reshma, S.K., Manogana, V.D. and Saroja, Y.S., 2019. Accident Severity Prediction using Data Mining Methods. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 5(2).

[7] Cuenca, L.G., Puertas, E., Aliane, N. and Andres, J.F., 2018, September. Traffic accidents classification and injury severity prediction. In 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE) (pp. 52-57). IEEE.

[8] Palash Dange, Aniruddha Kulkarni, Anshika Sharma (2020, April 12). US Accidents Dataset. https://github.com/dangepalash2/US_Accidents_Dataset

[9] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. sklearn.linear_model.LogisticRegression retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model. LogisticRegression.htmlsklearn-linear-model-logisticregression

[10] Renu Khandelwal, 2018, November. K-Nearest Neighbors(KNN) from https://medium.com/datadriveninvestor/k-nearest-neighbors-knn-7b4bd0128da7

BIBTEXb11 Afroz Chakure 2019, July. Decision Tree Classification. from https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac