Analysis and Prediction of Severity of United States Countrywide Car Accidents Based on Machine Learning Techniques

Lahiru S. Boyagoda
Deaprtment of Statistics and Computer Science
Faculty of Science
University of Peradeniya
Peradeniya, Sri Lanka
lahirus@sci.pdn.ac.lk

Lakshika S. Nawarathna
Deaprtment of Statistics and Computer Science
Faculty of Science
University of Peradeniya
Peradeniya, Sri Lanka
lakshikas@pdn.ac.lk

2022 7th International Conference on Information Technology Research (ICITR) | 979-8-3503-3203-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICITR57877.2022.9993371

***Abstract*— The number of vehicles and road transportation increases rapidly daily. Hence the frequency of road accidents and crashes also gradually increase with it. Analysing traffic accidents is one of the essential concerns in the world. Due to the considerable number of casualties and fatalities caused by those accidents, taking necessary actions to reduce road accidents is a vital public safety concern and challenge worldwide. Various statistical methods and techniques are used to address this issue. Hence, those statistical implementations are used for multiple applications, such as extracting cause and effect to predict real-time accidents. In this study, a United States (US) Countrywide car accidents data set consisting of about 1.5 million accident records with other relevant 45 measurements related to the US Countrywide Traffic Accidents were used. This work aims to develop classification models that predict the likelihood of an accident is severe. In addition, this study also consists of descriptive analysis to recognise the key features affecting the accident severity. Supervised machine learning methods such as Decision tree, K-nearest neighbour, and Random forest were used to create classification models. The predictive model results show that the Random Forest model performs with an accuracy of 83.95% for the train set and 80.69% for the test set, proving that the Random forest model performs better in accurately detecting the most relevant factors describing a road accident severity.**

Keywords—classification, decision tree, k-nearest neighbour, random forest

I. INTRODUCTION

Road accidents are widespread all over the world. Because of these road accidents, approximately 1.3 million people die annually [1]. Unfortunately, this ensures that men, women, or children playing in the streets or outdoors will never be safe, and there is no guarantee that anyone can reach their destination safely [2]. Due to the critical infrastructure damage and health injuries, accident prediction is a significant problem for governments and researchers.

The first road accident was recorded in 1771 due to the crashing of a steam-powered vehicle into a wall [3]. The first road casualty went back two centuries ago. In 1786 Anglican pastor died of fright at the sight of a loud and fast-moving model of "a steam engine on wheels" [4]. The world's first fatality caused by a road accident happened in 1869 in County Offaly in the UK [5]. Various independent factors are affecting the frequency of accidents. A few major reasons are traffic flow, lighting, weather, and infrastructure conditions [6]. Many researchers have developed numerous statistical models using road traffic accident data in past decades related to road accidents. This research was conducted using a Negative Binomial regression to analyse road accidents using four variables, namely traffic flow, speed limits, weather, and lighting conditions [7]. A study on the relationship between Accident Rates and Road Geometric Design was conducted using a variety of statistical modelling approaches such as linear regression models, multiple linear regression models, and multivariate models [8].

The United States safety performance seems less effective than many other European countries [9]. Many studies have been conducted using large-scale data sets throughout the period. But the dataset has been kept private rather than given access to everyone [10]. The founded data set corresponds to countrywide traffic accidents in the United States. It covers 49 states of the United States [11].

Regardless of the number of studies that have been carried out, the number of casualties due to road accidents are witnessed a rapid increase. Therefore, it is necessary to carry out further studies of road accidents to identify the factors that cause an accident. Hence preventive actions can overcome the accident rate and severity of accidents consequences. Successful development of strategies and countermeasures can potentially disrupt the chain of factors leading to an increase in the number of road accidents by generating a more secure mobility environment.

In this paper, the first phase is dedicated to performing an overall descriptive analysis of the data set utilised for the study to recognise the key features affecting the accident severity. Under that scenario, significant features such as which city/state has the highest number of accidents, which period has the least number of accidents and which is safe to travel to, and many other significant features were extracted and demonstrated the outcomes. Finally, classification models were developed and expected to find the most accurate model predicting accident severity. More specifically, the model is expected to predict the likelihood of the accident being a severe one.

The study is organised as follows. Section 1 provides the introduction of the study along with the historical background of similar studies. Section 2 focused on the data and the statistical methods used to perform the study. Further, the data set used to implement this study is described. A more comprehensive description of the nature of the sample is provided, and then the classification techniques are described. Detailed descriptions of the findings are interpreted in Section 3 concerning descriptive analysis and the findings of the classification algorithms. To sum up, in Section 4, the conclusion states the main findings of the analysis.

979-8-3503-3203-2/22/$31.00 ©2022 IEEE

II. DATA AND METHODS

A. Data

The data set corresponds to countrywide car accidents in the United States. It covers 49 states of the United States. The data is continuously collected between 2016 to 2020. Furthermore, traffic event data are provided by several data providers, and these events are captured by US government state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The data set consists of about 1.5 million accident records with other relevant 45 measurements related to the US countrywide traffic Accidents. Furthermore, the severity of the accidents is classified into four levels from 1 to 4, where 1 indicates the most negligible impact on traffic.

B. Severity Prediction Using Classification-Based Methods

In this study, classification-based methods, namely decision tree, K-nearest neighbour, and random forest, are used to predict a particular accident's likelihood of a severe accident.

C. Decision Tree

The decision tree is a popular classification algorithm used in many real-world applications constructed using the known class labels of a set of training data [12]. One of the vital features of decision tree classifiers is converting complex problems into simple decision-making problems [13]. Decision trees are composed of nodes and branches. Nodes represent the features in the data set, and each subset defines a value that can be taken by the node [14]. The root node is chosen at the beginning and then divides each input data, and the tree is formed by using values and features [15]. Decision trees are capable of solving problems, disregarding whether the input is discrete or continuous [16].

Gini Index and Entropy are used to calculate the information gained to split the nodes in the decision tree. The equations used to evaluate the Gini Index are given in Equation (1), and Entropy is given in Equation (2).

$$Gini\ Index = 1 - \sum_{i=1}^{n} p_i^2 \tag{1}$$

where p_i represents the probability of a given element being classified for a distinct class [17]. To measure the impurity of data, Entropy is used and shown in Equation (2).

$$Entropy\ = \sum_{i=1}^{n} p_i log p_i \tag{2}$$

where p_i represents the probability of a given element being classified for a distinct class [18].

D. K-Nearest Neighbour Algorithm

K-Nearest Neighbour is identified as one of the ten most common algorithms in classification tasks [19]. The K-Nearest-Neighbors (KNN) is an effortless, non-parametric classification algorithm categorised as a supervised learning algorithm [20]. Using a labelled training dataset comprising several data points categorised into different classes, unlabeled data can be predicted using this method [20]. As an extension of the Nearest Neighbour Classification, K-nearest neighbours are considered, and class labels are assigned based on majority voting decisions [21]. Both distance metrics applied and the value of K will primarily determine the performance of the KNN classifier [22]. To measure the dissimilarity, most KNN classifiers use the Euclidean metric in the absence of prior knowledge [23]. Euclidean distance is given in Equation (3).

$$d(x_i, x_j) = \sqrt{\sum_{i=1}^{n} w_r\left(a_r(x_i) - a_r(x_j)\right)^2} \tag{3}$$

where $x = (a_1, a_2, \dots, a_n)$is the n-dimensional vector, and w_r is the weight of the r^{th} attribute where r is from 1to n [24]. This algorithm assumes that all the instances in the data set correspond to points which is in n-dimensional space [25].

E. Random Forest

This algorithm consists of a collection of tree structure classifiers [26]. This classification algorithm consists of a collection of tree classifiers generated using a random vector. Based on the combination of trees, each tree casts a unit vote for the most popular class to classify the input data [27]. The divide and conquer principle grows a randomised tree predictor and aggregates predictors together [28]. This technique is applied to a range of applications not because of this algorithm's wide range of applications but because of the ability to deal with small sample sizes and high dimensional feature space with great accuracy [29].

F. Model Validation

Model validation can be defined as evaluating a model generated using a training data set against the testing data set. This helps to determine the ability of a trained model. Training and testing data sets are portions of the same data set. Each model's accuracy, precision, recall, and F1 score are considered to select the best model. A confusion matrix in Table I is used to display the results given below.

TABLE I. CONFUSION MATRIX

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

Accuracy is the ratio of the number of correct predictions to the total number of predictors.

$$Accuracy = \frac{TP + TN}{Total} \tag{4}$$

Recall can be defined as the proportion of positives that are correctly predicted. This evaluates the model's ability to detect positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Precision measures the accuracy of the model in classifying the sample as positive.

$$Precision = \frac{TP}{FP + TP} \tag{6}$$

The F1 score is a value between 0 and 1 used to measure a test's accuracy. When the value of the F1 score is closer to 1, the better the model's performance.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

III. RESULTS AND DISCUSSION

The data set is analysed city-wise to get an overall idea about the spread of the accidents across the US. The data set

consists of 10658 unique cities. As it is challenging to evaluate all 10658 cities, the top 10 cities that are accounted for the highest number of accidents in the United States is assessed to get a definitive idea of the distribution of casualties. After the analysis, Los Angeles accounted for the highest number of accidents in the US, with 39948 accidents.

The data are analysed under a time-related frame to provide further details on the accidents' characteristics (Figure 2). As the first step number of accidents related to a particular hour in the 24-hour day was evaluated using the whole data set. The evidence suggests that, in the mornings, accidents start to increase at 5 am and reach their peak at 7 am at about 5%. In the afternoon, accidents begin to grow at 1 pm and acquire a higher point at 5 pm at about 8%. 8 am is the riskiest hour in the morning, and in the evening, it is 5 pm.

The behaviour of the number of accidents during weekdays and weekends is also evaluated using the data set. Regarding the daily distribution of the accidents, significantly more accidents were observed during the weekdays. Regular working days of the week have almost two times higher accident proportions than the Weekend Days. Weekends illustrate a fewer number of accidents compared to a weekday. It is interesting to find out the hourly distribution of the weekends to check whether any significant pattern exists. The outcomes suggest a higher chance of an accident occurring from 2.00 pm to midnight.

Weather conditions such as wind direction, temperature, and many related factors might directly or indirectly affect the number of accidents. Temperature doesn't seem to have a direct impact on accidents. Yet, it is affected to decide the other crucial factors such as weather conditions, precipitation, wind directions, etc. Most accidents have occurred on days with temperatures between 50°F and 75°F. Since it is difficult to evaluate each weather condition, the top 10 most common weather conditions were figured out, affecting the accidents. Most accidents happen on days with Fairweather, followed by days with Mostly Cloudy or Clear weather.

Another significant finding is the severity analysis. It represents the severity of the accident, a number between 1 and 4, where 1 indicates the most negligible impact on traffic (i.e., short delay as a result of the accident) and 4 shows a significant impact on traffic (i.e., long delays). Most accidents have a severity level of 2 at about 80%, which means a moderate impact on traffic. Level 4, high impact on traffic, comes on third place with 7.5%.

TABLE II. ACCURACY MEASURES

Classifiers	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
Decision Tree	79.01%	74.34%	74.15%	74.24%	74.07%
Nearest Neighbors	65.44%	61.07%	60.53%	60.99%	60.42%
Random Forest	83.95%	80.69%	80.63%	80.61%	80.57%

Classification models were developed in the next phase of the data analysis to predict the likelihood of that particular accident being a severe one. Instead of using the same data set, several data preprocessing techniques were applied before

For each missing variable, values are calculated, and then the variables accounted for the highest proportions of the missing values were dropped from the data set. Further, there were a few missing values, and those rows also dropped before analysing the data set. A standard scale min-max normalisation is used to change the values of numeric columns in the dataset.

There are 45 different variables in the data set. This classification model is developed using 12 predictor variables: Start Latitude, Start Longitude, End Latitude, End Longitude, Distance, Wind Chill, Temperature, Humidity, Pressure, Visibility, Precipitation, and Windspeed. Only the numerical variables are considered. Since the objective is to build a classification model to predict a severe accident, balancing the data for all the different categories under the variable severity is essential. But according to the previous analysis, the data set is highly unbalanced because most accidents have a severity level of 2 at about 80%. To overcome this issue, all four categories are under-sampled to the number of records of the minority category, which is severity 1.

Model validation is performed to check whether the classification models can achieve the desired purpose. For model validation, the under-sample data set is divided into two data sets called training, and testing data set such that 70% of the observations are for the training data set. The rest is for the testing data set.

As the first classification model decision tree is used, the model is evaluated for different depths of the tree from 5 to 40. The highest accuracy of the model is recorded as 74.34% for the test set and 79.01% for the train set with a depth of the tree of 10. Precision, Recall, and F1 score are recorded as 74.15%, 74.24%, and 74.07%, respectively.

The K-nearest neighbour method is used for different K values. Model accuracy is studied in testing and training sets under different k values, and the best accuracy value was reported as 61.07% and 65.44% for the testing and training sets, respectively. As for the values for other accuracy measures, 60.53% is recorded for precision, 60.99% for recall, and the F1 score, 60.42%, is recorded.

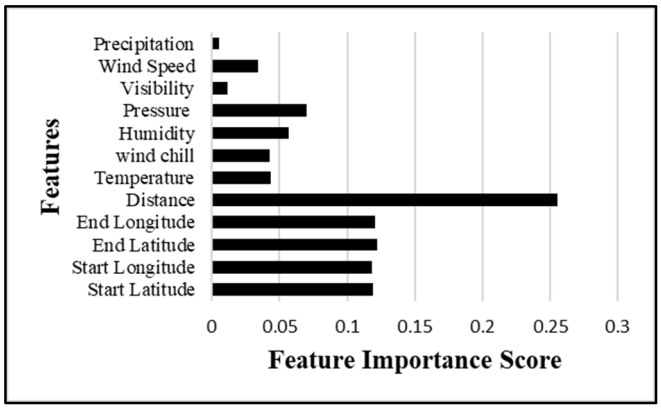

Fig. 1. Feature importance score

Finally, a random forest classification technique is used. The model ended up with the highest accuracy of 80.69% for the test set and 83.95% for the train set. Precision, Recall, and

F1 scores showed the highest values compared to the other two classification models.

The Random Forest model extracts the essential features that significantly affect the model used to predict the severity level of accidents. For each variable quality, the importance is calculated, and the usefulness of the input features are assessed. The feature importance scores are given in Figure 1.

The results suggest that the Distance variable, which means the length of the road extent affected by the accident, has a more significant effect on the model used to predict a severity level of an accident.

IV. Conclusion

Various research studies have been developed and undertaken over the years, focusing on identifying factors that can influence the level of severity of road accidents. In the studies related to road safety and accident severity research, continuous efforts are required to identify the hidden patterns and develop the appropriate methodologies to prevent future road accidents. Due to the development of new technologies and vehicles, it may not be possible to identify the unique patterns associated with accident severity levels using past studies. Hence continuous improvement must be required and aiming for those goals. This particular study was undertaken using the updated data set.

Predicting the likelihood of a severe accident based on available factors becomes an essential task in many real-world situations. This paper proposes a computational model using machine learning techniques known as Random Forest as the best model to predict the likelihood of the accident being a severe one using associated factors. The predictive model results show that it performs with an accuracy of 80.69% for the test set and 83.95% for the train set. Further, for other accuracy measures, 80.63% is recorded for precision, 80.61% is recorded for recall, and as the F1 score, 80.57% is recorded. These measures prove that this model can predict a given accident's severity.

The study helps to identify the significant factors affecting road accidents, such as time of the day, location, days, and weather factors. By deriving the critical elements, researchers

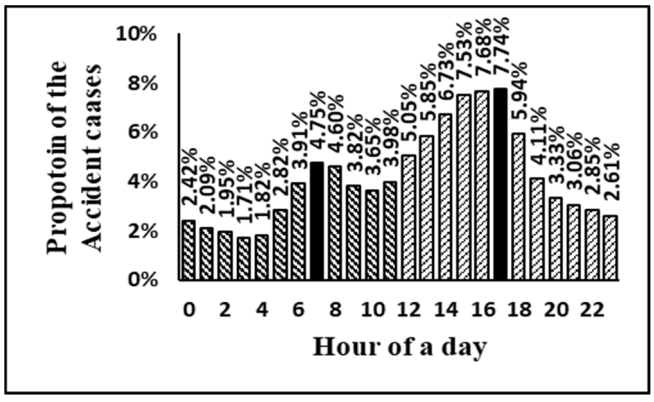

Fig. 2. Hourly Distribution of The Accident Cases

can utilise different strategies to implement transportation infrastructures and urban planning, which can benefit social and economic development. Conversely, the study suggested that random forest classification is the finest machine learning algorithm to find the generalisable predictive patterns related to this data set. Selecting the superior model in predicting accidental severity can be used to make insights into preventing road traffic accident injuries. Based on the features of machine learning models predicting the selected factors will effectively engage in implementing necessary road safety policies. In terms of future work, it is expected to combine machine learning with deep learning algorithms. Hence the effectiveness of the models will be increased to perform real-time road accident predictions.

References

[1] J. Zarocostas, "Road safety plan aims to save five million lives in next ten years," BMJ, vol. 342, May 2011.

[2] P. L. Jacobsen, F. Racioppi, and H. Rutter, "Who owns the roads? How motorised traffic discourages walking and bicycling," Injury Prevention, vol. 15, no. 6, pp. 369–373, Dec. 2009.

[3] K. Goniewicz, M. Goniewicz, W. Pawłowski, and P. Fiedor, "Road accidents in the early days of the automotive industry," Polish Journal of Public Health, vol. 125, no. 3, pp. 173–176, Sep. 2015.

[4] Y. Sun, Y. Cui, and Z. Tao, "Evaluating the Coordinated Development of Economic, Social and Environmental Benefits of Urban Public Transportation Infrastructure: The Case of Four Chinese Autonomous Municipalities," Mar. 2017.

[5] M. S. Thabet, "Anderson Heteropolymolybdates Cluster Loaded on Zeolite Materials: Preparation and Characterization," Journal of Encapsulation and Adsorption Sciences, vol. 09, no. 01, pp. 1–12, 2019.

[6] C. Caliendo, M. Guida, and A. Parisi, "A crash-prediction model for multilane roads," Accident Analysis & Prevention, vol. 39, no. 4, pp. 657–670, Jul. 2007.

[7] D. Eisenberg and K. E. Warner, "Effects of Snowfalls on Motor Vehicle Collisions, Injuries, and Fatalities," American Journal of Public Health, vol. 95, no. 1, pp. 120–124, Jan. 2005.

[8] M. H. Islam, L. Teik Hua, H. Hamid, and A. Azarkerdar, "Relationship of Accident Rates and Road Geometric Design," IOP Conference Series: Earth and Environmental Science, vol. 357, p. 012040, Nov. 2019.

[9] P. Holló, V. Eksler, and J. Zukowska, "Road safety performance indicators and their explanatory value: A critical view based on the experience of Central European countries," Safety Science, vol. 48, no. 9, pp. 1142–1150, Nov. 2010.

[10] D. Eisenberg, "The mixed effects of precipitation on traffic crashes," Accident Analysis & Prevention, vol. 36, no. 4, pp. 637–647, Jul. 2004.

[11] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident Risk Prediction based on Heterogeneous Sparse Data," Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL '19, 2019.

[12] C. Kingsford and S. L. Salzberg, "What are decision trees?," Nature Biotechnology 2008 26:9, vol. 26, no. 9, pp. 1011–1013, Sep. 2008.

[13] Priyanka and D. Kumar, "Decision tree classifier: A detailed survey," International Journal of Information and Decision Sciences, vol. 12, no. 3, pp. 246–269, 2020.

[14] N. Sandhu and S. Kumar, "Decision Tree Problem Solving Techniques A Review," International Journal of Computer Sciences and Engineering, vol. 6, no. 7, pp. 599–604, Jul. 2018.

[15] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision Trees: An Overview and Their Use in Medicine," Journal of Medical Systems 2002 26:5, vol. 26, no. 5, pp. 445–463, Oct. 2002.

[16] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," International Journal of Computer Sciences and Engineering, vol. 6, no. 10, pp. 74–78, Oct. 2018.

[17] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "The CART decision tree for mining data streams," Information Sciences, vol. 266, pp. 1–15, May 2014.

[18] B. Taha Jijo and A. Mohsin Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, Mar. 2021.

[19] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbour and modified k-nearest neighbour algorithm for data classification," Proceedings - 2017 2nd International

Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017, vol. 2018-January, pp. 294–298, Feb. 2018.

[20] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of the nearest neighbour algorithm for learning and classification," 2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019, pp. 1255–1260, May 2019.

[21] R. Raja Kumar, P. Viswanath, and C. Shobha Bindu, "Nearest Neighbor Classifiers: A Review," International Journal of Computational Intelligence Research, vol. 13, no. 2, pp. 303–311, 2017, Accessed: Dec. 29, 2021.

[22] B. W. Silverman and M. C. Jones, "E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)," International Statistical Review / Revue Internationale de Statistique, vol. 57, no. 3, p. 233, Dec. 1989.

[23] S. Sun and R. Huang, "An adaptive k-nearest neighbour algorithm," Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010.

[24] S. Sun and R. Huang, "An adaptive k-nearest neighbour algorithm," Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010, vol. 1, pp. 91–94, 2010.

[25] L. Wang, "Research and Implementation of Machine Learning Classifier Based on KNN," IOP Conference Series: Materials Science and Engineering, vol. 677, no. 5, Dec. 2019.

[26] Y. Liu, Y. Wang, and J. Zhang, "New Machine Learning Algorithm: Random Forest," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7473 LNCS, pp. 246–252, 2012.

[27] M. Pal, "Random forest classifier for remote sensing classification," http://dx.doi.org/10.1080/01431160412331269698, vol. 26, no. 1, pp. 217–222, Jan. 2007.

[28] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[29] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, vol. 51, no. 1, pp. 107–113, Jan. 2008.