Classifying the Severity Levels of Traffic Accidents Using Decision Trees

Zamira Hasanah Zamzuri(✉) and Khaw Zhi Qi

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia
zamira@ukm.edu.my

Abstract. Road accident is one of the main causes of deaths in Malaysia as well as heart disease and cerebrovascular disease. This study aims to identify the main factors that drive the occurrence of road accidents in Malaysia. Thus, preventive measures can be designed to reduce the incidence of road accidents. The relationship between the severity of road accidents and influencing factors such as vehicle movement, traffic system, marking and road geometry are also studied. The Classification and Regression Tree (CART) and Chi-square Automatic Interaction Detector (CHAID) techniques are used to identify the effects of factors in this study. The results from the decision tree show that the main factors that determine the severity of the accident are the type of vehicle, the type of violation, lighting, and severity of the driver's injuries. The performances of the two classification models are compared based on the prediction accuracy and models reliability. It is found that CHAID performs slightly better than CART and offers richer information in terms of influential factors and decision rules. The information in this study is important with the hope that road users can be vigilant and avoid being exposed to causes that allow them to be involved in accidents.

Keywords: Traffic accidents · Classification · Decision trees · Severity level

1 Introduction

One of the major causes of death in Malaysia is road accidents (Department of Statistics Malaysia 2014). Based on this report, road accidents recorded 5.6% or 4304 from 77365 deaths. Accidents not only causing injury and death to the road users, but at the same time increasing the government's burden to bear all surveillance and related costs. Hence, research on traffic accidents is essential in order to plan strategies for traffic safety purposes. Based on Ariyathilake and Rathnayaka (2019), road accidents killed 1.3 million people in the world and caused 20 to 50 million injured and long term or permanent physical disabilities. Therefore, identifying factors causing traffic accidents is vital for safety measures strategies planning. For example, the study conducted by Ramli (2011) identifies that weather, the road conditions, the vehicle's state, drivers and pedestrians' behaviour contributes to the occurrence of traffic accidents. Wang et al. (2019) assess the severity of traffic accidents from a macro perspective and suggest efficient safety

© The Author(s) 2023
N. Wahi et al. (Eds.): ICMSS 2022, ACSR 98, pp. 173–181, 2023.
https://doi.org/10.2991/978-94-6463-014-5_17

measures from the findings. This study uses epidemiology indices to indicate the level of severity and human loss per accident. Among identified contributed factors are types of accidents, time of the accident occurrence and speeding exceeding limit. Zhang and Fan (2013) investigated the contributing factors of accidents occurrence and found that probabilistic distributions provides great benefit to understand and avoid such events. Having said that, the data volume and complexity may make the process difficult, hence data mining techniques can be implemented to provide meaningful results and interpretation of the analysis.

Jain et al. (2016) fits binomial negative and Poisson regression models to identify accidents hotspots in India. In this study, the decision trees approach is used to identify the contributing factors and this information is further used in the development of Bayesian regression model to determine the frequency of the traffic accidents. Referring to Muhammad et al. (2017) in which the roots of accident occurrences are predicted using the decision trees, it is found that this approach provides better accuracy compared to the Neural Network (NN) technique. Li et al. (2019) also implement the same approach to model traffic accidents data in Wenli highway for 8 years using the C4.5 algorithms. These past studies shown the importance of considering machine learning technique in traffic accident analysis.

In this study, we use the data sourced from Malaysia Royal Police (PDRM) obtained through the Malaysia Institute of Road Safety (MIROS) to identify main causes of traffic accidents based on the levels of severity. The decision trees methods chosen are based on Classification and Regression Trees (CART) and Chi Square Automated Interaction Detector (CHAID). The accuracy between these two algorithms is also compared. Till date, the number of studies on applying decision trees to traffic accident data in Malaysia is still very limited. Rusli et al. (2018) focused on the accidents along rural mountainous highway, while Azhar et al. (2022) emphasize on the accidents involving heavy vehicle drivers. Hence, this paper aims to fill the gap through the construction of decision trees for the the set of accident data in Malaysia.

2 Materials and Method

2.1 Data

This study uses data from the police record of traffic accidents in Malaysia in the year of 2013 to 2015 sourced from the Malaysian Institute of Road Safety Research (MIROS) based on the reports to the Royal Malaysian Police (RMP). The process of data extraction and recording contribute to the delay on obtaining the updated data sets. Although the data are not the current ones, it is still significant to infer from this past data as the yearly patterns may help in future analytics. There are two sets of data based on the reports which are the accident profile and the driver's profile. These data sets are then merged based on the report ID before being used in the analysis. Referring to Table 1, for all three years considered, the number of accidents occurred are around 16000 for 2013 and 2014 and decreases to around 14000 on 2015. Meanwhile, Table 2 listed the variables in the data sets, accident, and driver. The accident data set consists of 13 variables mainly on the location and environment details on the accident and the driver data set have 12

Table 1. The sample size for the data sets (2013–2015)

Year	Sample size
2013	16147
2014	16645
2015	16458

variables mainly on demography and individual driving details. For this study, the data set is partitioned into 70:30 ratios for training and testing sets.

2.2 Decision Trees

As the name given, a decision tree is a structured tree that look similar with a flow chart. Each node in the tree represents the attribute and the last node is called as leaf that represent the class. The earliest node is known as root node and the branches from these nodes to the leaves are determined by the classification rules. There are many algorithms in decision trees, but the main idea is the induction process from top to bottom. Attributes chosen to be the nodes in the tree are vital since it will determine the accuracy of the tree.

Two important steps in the development of a decision tree, which are branching and pruning. Let *S* be the data set, *A* is the attribute set and *D* is the decision attribute set, the steps in the process are as follows:

1. Let *S* be the root node, if all data in *S* in the same class, *S* become the leaf node.
2. If not, choose one attribute in *A* and partition the node based on different values from the attribute chosen. *S* has *m* number of nodes on the lower level, while branches represent values different than the attribute chosen.
3. Repeat steps 1 and 2 for branch node *m.*
4. If the attribute in a node belongs to the same class or there are no more nodes to be partitioned, then the process ended.

Two main issues in the decision tree process are how to determine the best branch nodes and when should we stop the partitioning for a particular node. These issues will lead to the problem of overfitting that can decrease the accuracy rate of the decision tree. To overcome such issue, the pruning step is needed.

Classification and Regression Tree (CART)

The main essence of CART is the usage of Gini Index in order to compute the authenticity of the node. The Gini index value is between 0 and (1–1/n) in which n is the number of categories in the dependent variable. The classification and regression trees are very similar; the difference is classification trees used to predict discrete and categorical dependent variable whereas regression trees are for a continuous response variable. Formula for the Gini index is given in Eq. (1) that measure the totalvariance in all K

Table 2. The variables in the accident and driver data sets

Data set	Variables
Accident	Police station number Report number Month Hour Day of week Accident severity Traffic system Road geometry Lane marking Collision type Weather Light condition Location
Driver	Report number Vehicle type Vehicle damage Vehicle movement Sex Age Race Injury Belt wearing Part of body injured Errors Qualification

classes, in which $\hat{p}_{mk}$ refers to the proportion presenting the observations in mth region originated from kth class.

$$\mathrm{Gini} = \sum_{k=1}^{K} \hat{p}_{mk}\left(1 - \hat{p}_{mk}\right) \tag{1}$$

Chi Square Automated Interaction Detection (CHAID)
For CHAID, the algorithm builds the decision tree determined by the interaction of the explanatory variables in explaining the dependent variable. A chi-square test is used to determine the best partitioning of the node when the response variable is categorical. In the cases of continuous response variable, CHAID uses F-test for the same purpose of determining the best partitioning of the node. The chi square statistic formula is given in Eq. (2), in which o_{ij} is the observed frequency for the *i*th row and *j*th column and e_{ij} is the expected frequency for the *i*th row and *j*th column. The degrees of freedom is given by $(r - 1)(c - 1)$.

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{o_{ij} - e_{ij}}{e_{ij}}\right) \tag{2}$$

Table 3. The frequency (percentage) of the number of accidents based on the levels of severity and gender

Levels of Severity	Male	Female	Total
Fatal	8026 (24.35%)	955 (2.90%)	8981 (27.25%)
Serious	7418 (22.50%)	1299 (3.94%)	8717 (26.44%)
Minor	12600 (38.22%)	2665 (8.09%)	15265 (46.31%)
Total	28044 (83.07%)	4919 (14.93%)	32963 (100%)

Confusion Matrix

A confusion matrix is a commonly used measure used to assess the performance of a classification procedure. The four cells in the matrix are true positive, false positive, true negative and false negative. The accuracy rate is computed as the number of true cases (true negative and true positive) divided with the total number of observations.

3 Results and Discussion

3.1 The Accidents and Drivers' Profile

In this section, we discuss the profile of the drivers and accidents occurred based on the source of the data. Referring to Table 3, 28044 male drivers involved in accidents whereas only 4919 female drivers involved in accidents occurred around 2013–2015, which means the chance of a male driver to involve in an accident is 5.56 times more compared to the female drivers. This is due to the fact that more male drivers on the road, hence the exposure for them is greater, as pointed out by Liew et al. (2017). The same pattern was observed for the three levels of severity: fatal, serious, and minor.

Figure 1 represents the biplot of levels of severity with respect to the types of vehicles. It is observed that most of the minor accidents are the motorcycle riders with engine capacity below 251 cubic centimeters (cc), whereas for serious accidents, the riders are the ones with the engine capacity greater than 250 cc. Majority of fatal accidents are the ones involving taxis and vans. In relation to that, the findings by Talib and Gerhad (2000) and Talib et al. (2013) highlight those bad habits of the riders such as not wearing helmets may increase the risk of occurrence for more severe accidents.

3.2 The Classification and Regression Tree (CART)

Based on Fig. 2, the results of the CART analysis establish seven decision rules in classifying the levels of severity of the accidents. The rules are as follows:

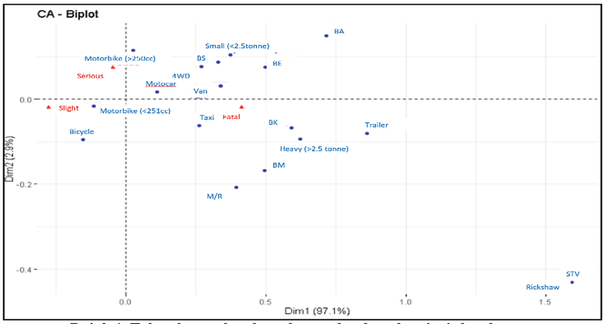

Fig. 1. The biplot of levels of severity and vehicle type.

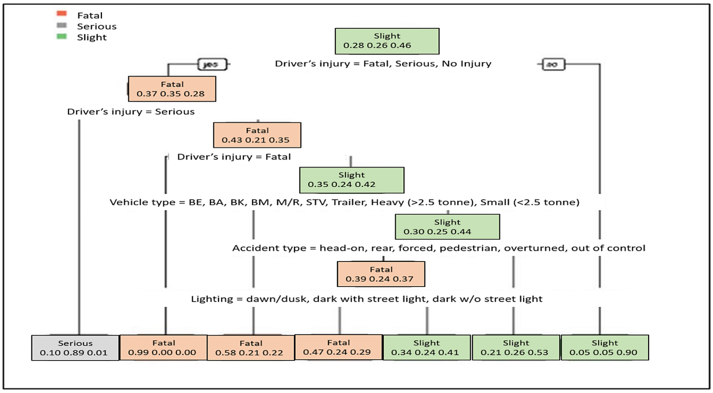

Fig. 2. The CART decision tree of the traffic accidents based on the levels of severity.

Rule 1: IF (driver's injury = no injury) AND (vehicle type = car / motorcycle > 250cc / motorcycle < 251cc / taxi / van / bicycle) AND (types of accident = head on / overhead / rear / forced / pedestrians / overturned / out of control) AND (lighting = dawn / dusk / dark with street lights / dark witho ut street lights), THEN fatal.

Rule 2: IF (driver's injury = no injury) AND (vehicle type = bus / trailer lorry / Lorry >2.5 tonne / small lorry < 2.5 tonne), THEN fatal.

Rule 3: IF (driver's injury = fatal), THEN fatal.

Rule 4: IF (driver's injury = serious), THEN serious.

Rule 5: IF (driver's injury = minor), THEN minor.

Rule 6: IF (driver's injury = no injury) AND (vehicle type = car / motorcycle > 250cc / motorcycle < 251cc / taxi / van / bicycle) AND (types of accident = head on / overhead / rear / forced / pedestrians / overturned / out of control) DAN (lighting = day), THEN minor.

Rule 7: IF (driver's injury = no injury) AND (vehicle type = car/ motorcycle > 250cc / motorcycle < 251cc / taxi / van / bicycle) AND (type of accident = right angle side / angular / side swipe / hitting animal), THEN minor.

Table 4. The confusion matrix and performance measures for CART model.

Actual/Predicted	Fatal	Serious	Minor
Fatal	1915	169	1217
Serious	431	1511	1224
Minor	516	22	5018
Model classification accuracy 70.23%			
95% confidence interval (0.6941, 0.7105)			
Kappa 0.5125			
p-value McNemar test < 0.000			

3.3 Chi Square Automated Interaction Detection (CHAID)

The decision rules produced by CHAID is more complicated and different than CART. The number of rules produced is 21, which is greater than seven decision rules established using CART. With more decision rules, we can say that the rules are more detailed and specific, but the downside is difficulty in terms of interpretation. Due to the space limitation, the whole decision trees based on CHAID is not included in this section. From the tree constructed using CHAID, four additional variables that were not identified in CART appear as important for CHAID's decision rules. These variables are the road marking, the status of helmet wearing and the gender and race of drivers.

3.4 The Comparison of Accuracy

Next, we compare the performance between these algorithms using a confusion matrix of the predictions on the test data set. Referring to Table 4, the accuracy rate for CART model is 70.23% with a confidence interval of (0.6941, 0.7105). The Kappa value of 0.5125 explains the reliability of the model at a moderate level. The p-value of McNemar test suggests that the relationships between the explanatory variables with the levels of severity are significant. Based on these values, it is concluded that the classification model performance is acceptable and satisfactory. Table 5 highlights the same information but for the CHAID algorithm. The values are almost similar in which we can conclude that the performance of these two algorithms is comparable, with CHAID performs slightly better than CART algorithm.

Table 5. The confusion matrix and performance measures for CHAID model.

Actual/Predicted	Fatal	Serious	Minor
Fatal	2280	170	851
Serious	654	1512	1000
Minor	740	24	4792
Model classification accuracy 71.40%			
95% confidence interval (0.7058, 0.7220)			
Kappa 0.5412			
p-value McNemar test < 0.000			

4 Conclusions

In this study, decision trees have been used to classify the severity of an accidents based on other factors. The findings based on the CART algorithm identify those important factors that determine the level of severity are driver's injury, vehicle type, type of accident and lighting. Meanwhile, based on the CHAID algorithm, all factors are important except the day of the week. The most important variable is the driver's injury followed by the vehicle type and the type of accident. The confusion matrix indicates that CHAID's performance is slightly better than CART. This is due to the fact that CHAID uses double partitioning at the nodes while CHART uses binary partitioning, Furthermore, CHAID can avoid the overfitting problem in which that only a node will be partitioned if the significant criteria are fulfilled.

Acknowledgments. We would like to thank Malaysia Institute of Road Safety Research (MIROS) for providing the data used in this study.

References

Ariyathilake, P.B.S.N., Rathnayaka, R.M.K.T.: Comparative analysis of machine learning algorithms for road accident forecasting. In: Proceedings of the 7th International Conference of Sabaragamuwa University of Sri Lanka (2019)

Jain, A., Ahuja, G., Anuranjana, Mehrotra, D.: Data mining approach to analyse the road accidents in India. In: 5th International Conference on Reliability Infocom Technology Optimization, pp. 175–179 (2016)

Li, J., He, J., Liu, Z., Zhang, H., Zhang, C.: Traffic accident analysis based on C4.5 algorithm in WEKA. In: MATEC Web of Conference, vol. 272, pp. 1–8 (2019)

Muhammad, L.J., et al.: Using decision tree data mining algorithm to predict causes of road traffic accidents, its prone locations and time along Kano-Wudil highway. Int. J. Database Theory Appl. **10**(1), 197–206 (2017)

Liew, S., Hamidun, R., Mohd Soid, N.F.: Differences of driving experience and gender on traffic offences among Malaysian motorists. In: MATEC Web of Conferences, vol. 103 (2017)

Talib, R.J., Gerhad, F.: Kes Sebenar Kemalangan Jalan Raya Antara Motosikal Dengan Kereta. Jurnal Teknologi. **22**(A), 66–80 (2000)

Talib, R.J., Mohd, F.M., Sutiman, K., Ramlan, K.: Kemalangan Jalan Raya: Analisis Data Membabitkan Pengguna Motosikal. Jurnal Teknologi **38**(b), 1–14 (2013)

Wang, D., Liu, Q., Liang, M., Zhang, Y., Cong, H.: Road traffic accident severity analysis; a census-based study in China. J. Saf. Res. **70**, 135–147 (2019)

Zhang, X.F., Fan, L.: A decision tree approach for traffic accident analysis of Saskachewan highways. In: 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) (2013)

Rusli, R., Haque, M.M., Saifuzzaman, M., King, M.: Crash severity along rural mountainous highways in Malaysia: an application of a combined decision tree and logistic regression model. Traffic Injury Prev. **19**(7), 741–748 (2018). https://doi.org/10.1080/15389588.2018.1482537

Azhar, A., Mohd Ariff, N., Abu Bakar, M.A., Roslan, A.: Classification of driver injury severity for accidents involving heavy vehicles with decision tree and random forest. Sustainability **14**(2022). https://doi.org/10.3390/su14074101

Ramli, M.Z.: Development of accident prediction model by using artificial neural network (ANN). In: UTHM Master Thesis (2011)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.