International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 1
ISSN: 2088-5334

Decision Tree Model for Non-Fatal Road Accident Injury

Fatin Ellisya Sapri[#], Nur Shuhada Nordin[#], Siti Maisarah Hasan[#], Wan Fairos Wan Yaacob[*], Syerina Azlin Md Nasir[*]

[#]*Faculty of Mathematical Science and Computer, Universiti Teknologi MARA, 15150 Kota Bharu, Kelantan, Malaysia*
E-mail: fatin_ellisya17@gmail.com,nurshuhada_wada93@yahoo.com,uchisaramai@gmail.com

[*]*Faculty of Mathematical Science and Computer, Universiti Teknologi MARA, 18500 Machang, Kelantan, Malaysia*
E-mail: wnfairos@kelantan.uitm.edu.my, syerina@kelantan.uitm.edu.my

***Abstract*— Non-fatal road accident injury has become a great concern as it is associated with injury and sometimes leads to the disability of the victims. Hence, this study aims to develop a model that explains the factors that contribute to non-fatal road accident injury severity. A sample data of 350 non-fatal road accident cases of the year 2016 were obtained from Kota Bharu District Police Headquarters, Kelantan. The explanatory variables include road geometry, collision type, accident time, accident causes, vehicle type, age, airbag, and gender. The predictive data mining techniques of decision tree model and multinomial logistic regression were used to model non-fatal road accident injury severity. Based on accuracy rate, decision tree with CART algorithm was found to be more accurate as compared to the logistic regression model. The factors that significantly contribute to non-fatal traffic crashes injury severity are accident cause, road geometry, vehicle type, age and collision type.**

***Keywords*— road accident injury severity; logistic regression; decision tree; CART; LR main**

I. INTRODUCTION

Road accident has become a major issue of concern worldwide. World Health Organization stated that road accident has resulted to 4.8 million of injury worldwide. There will be 1.25 million people killed by the year 2020 [1]. The mortality due to road accident is ranked third after the disease of circulatory system and cancer which are placed at first and second ranks respectively [2]. Road accident severities can be classified into several categories which are no injuries, slight injuries, and serious injuries.

Road accident has also become one of the contributing factors towards death for Malaysian. Malaysia is ranked at 20th in the road accident case leaving Namibia, Iran, and Thailand as the top three nominators respectively [3]. Focussing on a single district in Kelantan, Kota Bharu is ranked third for the city with the highest population in Kelantan. Based on the Department of Statistic Malaysia, the population in Kota Bharu has increased to 491237 in 2010. As the number of vehicles increases, the road accident occurrences are also expected to increase. The effectiveness of road safety campaigns has not yet been proven [4]. In Malaysia, specifically in Kelantan, the research regarding non-fatal injury severity on road accident is limited. Thus, to address the minister of transportation regarding the needs in reducing accident cases, more research must be done. Thus, this paper will develop a predictive model using a decision tree to examine factors that can cause road accident severity focusing on road geometry, collision type, accident time, accident causes, vehicle type, victim's age, airbag, types of injury and gender factor. The findings of the research are beneficial to all related parties including Road Transport Department, Royal Malaysian Police and Public Work Department.

A. Related Work

This section highlights the review on the previous study of the road accident models and factors. This chapter also discussed some of the mining models that are commonly used in dealing with non-fatal injury severities.

B. Factors Affecting Road Accident Severity

There are several factors found to be related to road accidents. The factors ranging from the collision type which includes collision from right-angle, sideswipe, rear-end, front and unknown. The other factors are the time of accident, accident cause, driver age and vehicle type. All these factors can be categorized into several groups such as human factors, vehicle factors, demographic factors, environmental factors and many more. Human factors such as driving license and safety belt are the determinants of accidents in Iran [5]. Seatbelts can help to prevent certain

injury and fatality towards the vehicle occupant [6]. Therefore, this study focuses on human factors and environmental factors.

The human factor can be divided into two categories which are human error and driver's characteristic. Human factors have been a part of the road safety problem [7]. The factors of age, gender and driving behaviour such as speeding contribute to road accident injury severity [8]-[10]. Age significantly influences road accident [11].

Another factor that contributes to road crash is road geometry, which is commonly being referred to as location and condition of the road. The factors of road accident such as lighting condition of the road and time of accidents (day/night) are some of the factors that contribute to a road accident. Junction, intersection, horizontal slope, and curve are significant factors that contribute to it [12]-[14]. The previous researchers reported that the possibility of a road accident to occur is high if the road is not straight.

Vehicle type is also one of the factors that contribute to the occurrence of the road accident. The mass in the car plays the important role in reducing the driver's frontal crash [15]. Vehicle types can be categorized as small passenger car, large passenger car, pickup truck, taxi and others as the category for vehicle type. In 2014, the highest number of vehicles involved in road accident severity in Britain is car with 195576 cases, followed by motorcycle with 21378 cases and lastly other types of vehicles with 20146 cases [16].

C. Application of Data Mining Techniques in Road Accident

The literature on models developed for road accident can be seen ranging from logistic regression model to count data model. Ordinal logistic regression is used to analyse factors associated with higher level of injury severity [17]-[18]. Logistic regression is well suited for defining and verifying hypotheses about associations between a categorical response variable and one or more categorical or continuous predictor variables [19].

In other cases, panel count model is applied by developing the negative binomial model which is related to the effect of infrastructure [29] and demographic change on traffic-related fatalities and crashes [20]. This approach is used because the data that they have is in the form of time series and cross-sectional data. In addition, the study on road accident is always related to the models like Pooled Poisson, Fixed Effects Poisson and Fixed Effect Negative Binomial [21].

Besides count model, recently, the development of a predictive model on the basis of data mining technique has been widely used. The classification and regression tree (CART) model of the decision tree are used to find the most significant determinants influencing the injury severity in Iran [22]. The technique of data mining is also used as an approach on road accident in the study of road traffic accidents modelling [23]. The applications of Classification and Regression Trees (CART) and Multivariate adaptive Regression Splines (MARS) were developed in that study. Data mining appears as a useful tool to address the need for getting useful information such as hidden patterns from databases. Hence, this is the reason for them in applying data mining rule in the study. Decision tree which is developed through data mining can be used for predicting future decision making. Moreover, data mining technique was also used to link recorded road characteristics to accident severity in Ethiopia [24].

II. Material and Method

This section discusses the methodology being applied in this study. This section includes a description of data collection method. The target variables are coded into three categories: 1 = no injury, 2 = minor injury and 3 = serious injury. The sample data set was divided into two parts which are training part and validation part. The researchers have set 245 observations (70% of the total observations) for training sample, while validation sample is only 105 observations (30% of the total observations). SAS Enterprise Miner Workstation 7.1 was used to build logistic regression model and decision tree model.

A. Data Collection Method

In this study, the data was collected from Kota Bharu District Police Headquarters, Departments of Traffic regarding the accident occurrences that had been reported. The data contained 350 samples from the year 2016. The variables that the researchers had studied totally depended on previous studies and the availability of the variables from the traffic police. Only accident that has been reported in Kota Bharu was examined.

The researchers obtained the data on the 'Statistics of Accidents' which summarized the number of accidents according to the district, level of injuries, causes of accident and types of vehicle from January 2012 to September 2015. As the traffic department in Kota Bharu had not done any research on non-fatal road accident injury severity modeling, the variables that we need were not yet being summarized and it was still in the report form. Based on the accident's report, the researchers identified the important variables and coded all of the variables based on how the data have been categorized. Therefore, this study used the dependent variable of type of injury. There were 8 independent variables examined in this research which included road geometry, collision type, accident time, accident causes, vehicle type, victim's age, airbag, and gender. Table 1 shows the summary of the data description.

TABLE I
DESCRIPTION OF DATA

Variables	Description	Type	Categorical
Y	Types of Injury	Categorical	1 No Injury 2 Minor Injury 3 Serious Injury
X1	Road Geometry	Categorical	1 Straight 2 Bend 3 Roundabout 4 Junction 5 Unknown
X2	Collision type	Categorical	1 Right angle side 2 Side swipe 3 Rear end 4 Front 5 Unknown
X3	Accident time	Categorical	1 Midnight/early morning (0000-0559) 2 Peak hours (0600-0959; 1559-1959) 3 Non-Peak hours (1000-1600) 4 Evening (2000-2359)
X4	Accident causes	Categorical	1 Speed 2 Run red light 3 Follow to close 4 Overtake 5 Careless 6 No signal 7 Unknown
X5	Vehicle type	Categorical	1 Motorcycle 2 Car 3 MPV 4 Others
X6	Age	Integers (Years)	
X7	Airbag	Categorical	1 Yes 2 No
X8	Gender	Categorical	1 Male 2 Female

B. Multinomial Logistic Regression

Logistic regression is a statistical method for analyzing a dataset with one or more independent variable. It allows the researcher in predicting a discrete outcome from a set of variables that may be continuous, discrete and dichotomous or a mix of any of these. The goal of logistic regression is to describe the relationship between a binary dependent variable and a set of explanatory variable. The binary dependent variable is a variable that consists of two possible outcomes which are success or failure and both are usually denoted as 0 and 1 respectively. Multinomial logistic regression is an extension of binary logistic regression. It deals with more than two categories of dependent variables. When considering this type of regression, the assumption of normality, linearity, and homoscedasticity are not to be taken into account [25]. In this study, the dependent variable is the type of injury which was divided into three naturally unordered elements; no injury, slight injury and serious injury. We have partitioned the data into a training set and validation set which included 70% and 30% respectively. The model is said to be significant if p-value for likelihood ratio Chi-Square is less than 0.05. For the variable to be significant, the p-value of Wald Chi-Square must also be less than 0.05. We have run seven types of the logistic regression model in order to choose the best predictive modelling among the logistic regression. The seven types of logistic regression are shown in Table 2 as follows:

TABLE III
DESCRIPTION OF LOGISTIC REGRESSION MODEL TYPES

Model Type	Description
i) LR Main ii) LR Inter iii) LR Poly	Includes in the model all sets of the variable used.
i) LR Main Inter ii) LR Main Poly iii) LR Inter Poly	All two factors interaction for class variable sets used were included in this part
i) LR Main Inter Poly	Poly Term includes in the model polynomial terms up to the degree specified for all interval variables used. Poly Degree: specifies the polynomial degree when the term was included in the model

In each of these seven types of logistic regression, the Misclassification Rate and Average Squared Error were compared in order to obtain the best model among them. If there was overfitting detected in both Misclassification Rate and Average Squared Error, the model was not considered as good and vice versa. Since it is difficult to decide which model is better to be applied to this data set, the researcher performs a model comparison based on another approach which is based on accuracy rate. From accuracy rate, the

model is said to be over fit and cannot be chosen as predictive modelling if there was a large gap between the training set and validation set. Hence, from the accuracy rate, the best model was obtained and odds ratio based on this best model was interpreted. The odds ratio is the odds of particular outcomes relative to the odds of another outcome. It is also a way of comparing whether the odds of a certain outcome is the same for two different groups. The odds ratio is one of a range statistics used to assess the risk of particular outcomes if a certain factor is present [26]. It is also a relative measure of risk, telling us how much more likely it is that someone who is exposed to the factor under study will develop as compared to someone who is not exposed. The odds of an event happening is the probability that the event will occur divided by the probability of an event not to occur.

The equation of odds ratio for the cross-sectional case can be like this:

$$\text{Odds for group 1} = \frac{PG1}{1 - PG1} \quad (1)$$

$$\text{Odds for group 2} = \frac{PG2}{1 - PG2} \quad (2)$$

where PG1 = Odds of the event of interest for group 1 and PG2 = Odds of the event of interest for group 2. Therefore;

Odds ratio = Odds ratio for group 1/Odds ratio for group 2 (3)

Interpretation of the Odds Ratio:
Interpretation of the Odds Ratio:

Odds ratio = 1; Indicates that there is no difference between the groups which means there is no association between the suggested exposure and the outcome.
Odds ratio > 1; Indicates that the odds of exposure are positively associated with adverse outcome compared to odds of not being exposed.
Odds ratio < 1; Indicates that the odds of exposure are negatively associated with the adverse outcome compared to the odds of not being exposed.

C. Decision Tree

A decision tree is a hierarchical collection of rules that describes how to divide a large collection of records into successively smaller groups of records [27]. Decision trees are one of the most powerful directed data mining techniques because it can be used on such a wide range of problems and they produce models that explain how they work. It is not only used for categorical target variable but also for continuous target variables although multiple regressions are more suitable for such variable. They also explained that decision will algorithms automatically determines the most important variables and sorts the observations into the correct output category when given a set of independent variables and a target variable.

There are five different types of decision tree model developed in this study. The models included Gini, Entropy, Logworth, CART and CHAID. The researchers go through the data cleaning to make sure no unwanted values exist in the data set. Next, the data was partitioned into a training set and validation set. The rational of partitioning the data is to evaluate the performance of the model. A model predicts the target value accurately and easily to generate a given set of training data. However, this model would only be able to predict the training data accurately. Hence, to overcome this issue, SAS Enterprise Miner is designed to use a validation data set which functions as gauging model performance. The partitioning set 70% of the total observations as training set and 30% of the total observations as a validation set.

The decision tree models developed are based on five different splitting criterion of Gini, Entropy, Log Worth, CART (Classification and Regression Trees) and CHAID (Chi-Square Automatic Interaction). The first model being discussed is decision tree based on Gini as a splitting criterion. Gini is one of the popular splitting criterions since it is also being used by biologist and ecologist studying population diversity. It gives the probability that two items chosen at random from a particular population are in the similar class. In addition, the measure of a node in the Gini is the sum of squares of the proportions of the classes in the node and a perfectly pure node has a Gini score of 1. Next is the decision tree based on Entropy reduction or also known as information gain as a splitting criterion. The information gain defines purity in a similar way as machine learning does. This means that if a leaf is entirely pure, then it is easy to describe the classes in the leaf. On the other hand, if a leaf is highly impure, then describing it is quite complicated. They also defined that the entropy for a node is the sum of all the target values in the node of the proportion of records with particular value multiplied by the base two logarithms of that proportion. Perfectly pure nodes of entropy have a lower score which is zero. The third model in the decision tree is Logworth. Logworth and the two previous splitting criteria apply to categorical targets. This suits the target variables in this study which is categorized as a nominal variable. The two researchers stated that Logworth computed the Chi-Square statistics of association between the binary targets and all potential splits of each competing input. So, the highest score would be the purest node. The last two models are CART (Classification and Regression Trees) and CHAID (Chi-Square Automatic Interaction Detector). Both models are similar in terms of pruning approach. Pruning is the process of cutting away leaves which are proven to lead to an unstable split. CART is one of the most important and popular data mining tools. It turned out to be a powerful method for dealing with prediction and classification problems when dealing with large variables and observations. Besides, CART is technically known as dichotomous recursive partitioning because parent nodes are always split exactly into two child nodes. The researcher also stated that CART is the most advanced decision-tree technology for data analysis, pre-processing and predictive modelling. Finally, CHAID algorithm is to test whether the distribution of validation set differs from the distribution of training set result. After testing on the training set and validation set, both sets are then compared based on confidence interval.

The model developed was evaluated based on several criteria. One of the model assessment methods is based on misclassification rate. Misclassification rate is the fraction of cases for which the wrong decision was made and the proportion misclassified is calculated for both training and validation data sets [28]. Predicted value is generated from

the proportion of cases with the primary outcome in each terminal leaf. So, the average squared errors are the deviation between prediction and actual result is squared and averaged across each leaf node. The last model assessment being measured is accuracy rate. Accuracy rate is the proportion of the total number of predictions that were correct. It can be simply calculated by adding the true positive rate and the true negative rate, then divided by the sum of true positive and negative rate and false positive and negative rate. All three model assessments which are misclassification rate, average squared error and accuracy rate were observed to indicate whether the model is underfitted or overfit. Underfit occurs when the model performed better than the validation set and performed poorly in the training set. This situation should be totally rejected. In contrary, overfit is a condition when a model is overly accommodating nuances of the random noise in the particular sample.

III. Results and Discussion

In this section, the results of predictive modelling using logistic regression and decision tree are presented.

A. Logistic Regression Result

Out of 350 road accident cases in Kota Bharu, the researchers have set 70% or 245 cases as the training set. While another 30% or 105 cases were set as a validation set before the analysis were run. Then, from the analysis [30], the researchers obtained the likelihood ratio chi-square of 98.1845 with p-value < 0.05 indicating that the logistic regression model is significant. Thus, this indicates that at least one independent variable is a significant predictor of road accident injury severity. Table 3 shows the Wald chi-square tests.

TABLE IIIII
P-value of Wald Chi-square

Independent Variables	P-Value
Age	0.17
Accident Causes	0.05
Accident Time	0.57
Airbag	0.77
Collision Type	0.28
Gender	0.47
Road Geometry	0.26
Vehicle type	0.00

From the results obtained, only 2 independent variables are significant or at least one category for each variable is significant which are accident causes and vehicle types because both are having a p-value < 0.05. In addition, the researchers had run 6 types of logistic regression model namely LR Main, LR Inter, LR Poly, LR Main Inter, LR Main Poly, LR Inter Poly and LR Main Inter Poly. The rational of running these 6 types of the logistic regression model is to obtain the best predictive modelling for the data set used. Table 4 is the summary of logistic regression results.

From Table 4, it clearly shows that LR Inter, LR Main Inter, LR Inter Poly and LR Main Inter Poly have the greatest over fit since all of them had a large rate on training set but poorly performed on the validation set. These are actually resulting in a large gap between the training set and validation set. Hence, these four models are not good to be chosen as a predictive modelling for this set of data. Therefore, the researchers concluded that LR Main is the best model for logistic regression since it had the highest percentage of accuracy as compared to LR Poly and LR Main Poly even though there is slightly overfitting on that particular model. In Table 5, the value of odds ratio and its interpretations were summarized based on the significant category for each significant variables.

TABLE IVV
Summary of Logistic Regression Result

Model		Accuracy Rate (%)
LR Main	Training Validation	68.72 58.88
LR Inter	Training Validation	92.18 57.00
LR Poly	Training Validation	56.00 52.34
LR Main Inter	Training Validation	93.00 60.74
LR Main Poly	Training Validation	66.26 57.00
LR Inter Poly	Training Validation	93.42 56.07
LR Main Inter Poly	Training Validation	93.83 60.75

TABLE V
Table of Odds Ratio

Variables	Odds Ratio	Interpretation
Acc_cause3 3	3.408	The odds of being categorized as a serious injury for following too close are 2.408 times higher than unknown.
Acc_cause3 2	4.063	The odds of being categorized as a minor injury for following too close are 3.063 times higher than unknown.
Acc_cause4 3	0.191	The odds of being categorized as a serious injury for overtaking is 0.809 times lower than unknown.
Acc_cause4 2	0.157	The odds of being categorized as a minor injury for overtake is 0.843 times lower than unknown.
Acc_cause5 3	3.191	The odds of being categorized as a serious injury for carelessness are 2.191 times higher than unknown.
Acc_cause5 2	4.645	The odds of being categorized as a minor injury for carelessness are 3.645 times higher than unknown.
Veh_type1 3	4.671	The odds of being categorized as a serious injury for motorcycle are 3.671 times higher than the other types of vehicle.
Veh_type1 2	11.374	The odds of being categorized as a minor injury for motorcycle are 10.374 times higher than the other types of vehicle.
Veh_type2 3	0.493	The odds of being categorized as a serious injury for car is 0.507 times lower than the other types of vehicle.

B. Decision Tree

There are five types of decision tree algorithms used in this study which are the Gini, Entropy, Logworth, CART and CHAID. All the five types of decision tree model were compared based on the Average Squared Error and Misclassification Rate first. From the evaluation, LOGWORTH was under fit based on Misclassification Rate and CART was under fit based on Average Squared Error. Hence, both of the decision tree models were excluded from being chosen as the best predictive model. Therefore, GINI, ENTROPY and CART were evaluated using accuracy rate in order to find the best predictive modelling among them. Table 6 summarizes the accuracy rate for the three decision tree models applied on the training and validation sample.

TABLE VI
SUMMARY OF DECISION TREE RESULTS

Model		**Accuracy Rate (%)**
Gini	Training Validation	69.78 59.62
Entropy	Training Validation	58.27 51.92
CART	Training Validation	70.50 64.42

Hence, the researcher can conclude that the best predictive model to predict the types of road accident injury severity in Kota Bharu is CART. The CART model is still acceptable even though there is slightly over fitting. CART also found that two variables are significant which are accident causes and vehicle types. Fig. 1 is the output of the decision tree for CART. The output shows that the most important variable is accident cause. Besides the above output, decision tree also obtained the English rule. The English rule is too complicated to be displayed. Hence, the only interpretations of it are presented in Table 7.

TABLE VII
ENGLISH RULES FOR CART

1. The road accident occurrence will lead to no injury when the accident cause is overtaking.
2. The road accident occurrence will lead to serious injury when the road geometry is bending and the accident cause is either speed/run red light/follow too close/careless.
3. The road accident occurrence will lead to minor injury when the vehicle type is either car/MPV/others, the road geometry is either straight/junction, the age of victim is less than 21.5 years and the accident cause is either speed/run red light/follow too close/careless.
4. The road accident occurrence will lead to minor injury when the vehicle type is motorcycle, the road geometry is either straight/junction, the collision type is right angle side or the accident cause is either speed/run red light/follow too close/careless.
5. The road accident occurrence will lead to serious injury when the vehicle type is either car/MPV/others, the road geometry is either straight/junction, the collision type is either right angle side/side swipe, the age of victim is greater than or equal to 21.5 years and the accident cause is either speed/run red light/follow too close/careless.
6. The road accident occurrence will lead to minor injury when the vehicle type is motorcycle, the road geometry is either straight/junction, the collision type is either side swipe/front, the age of victim is less than to 36.5 years and the accident cause is either speed/run red light/follow too close/careless.
7. The road accident occurrence will lead to serious injury when the vehicle type is either car/MPV/others, the road geometry is junction, the collision type is either rear end/front, the age of victim is greater than or equal to 21.5 years and the accident cause is either speed/run red light/follow too close/careless.
8. The road accident occurrence will lead to serious injury when the vehicle type is either car/MPV/others, the road geometry is straight, the collision type is either rear end/front, the age of victim is greater than or equal to 21.5 years and the accident cause is either speed/run red light/follow too close/careless.
9. The road accident occurrence will lead to serious injury when the vehicle type is motorcycle, the road geometry is either straight/junction, the collision type is either side swipe/front, the age of victim is greater than or equal to 36.5 years and the accident cause is either speed/run red light/follow too close/careless.
10. The road accident occurrence will lead to serious injury when the vehicle type is motorcycle, the road geometry is either straight/junction, the collision type is either side swipe/front, the age of victim is greater than or equal to 41.5 years and the accident cause is either speed/run red light/follow too close/careless.

C. Model Comparison

The comparison of LR Main and CART model is summarized in Table 8.

TABLE VIII
SUMMARY OF LR MAIN AND CART RESULT

Model		**Accuracy Rate (%)**
LR Main	Training Validation	68.72 58.88
CART	Training Validation	70.50 64.42

From the result presented in Table 8, CART was chosen as the best predictive modelling to predict the types of road accident injury because it gave the highest accuracy rate for training and validation set as compared to LR Main. CART is still acceptable even though there is slight overfitting.

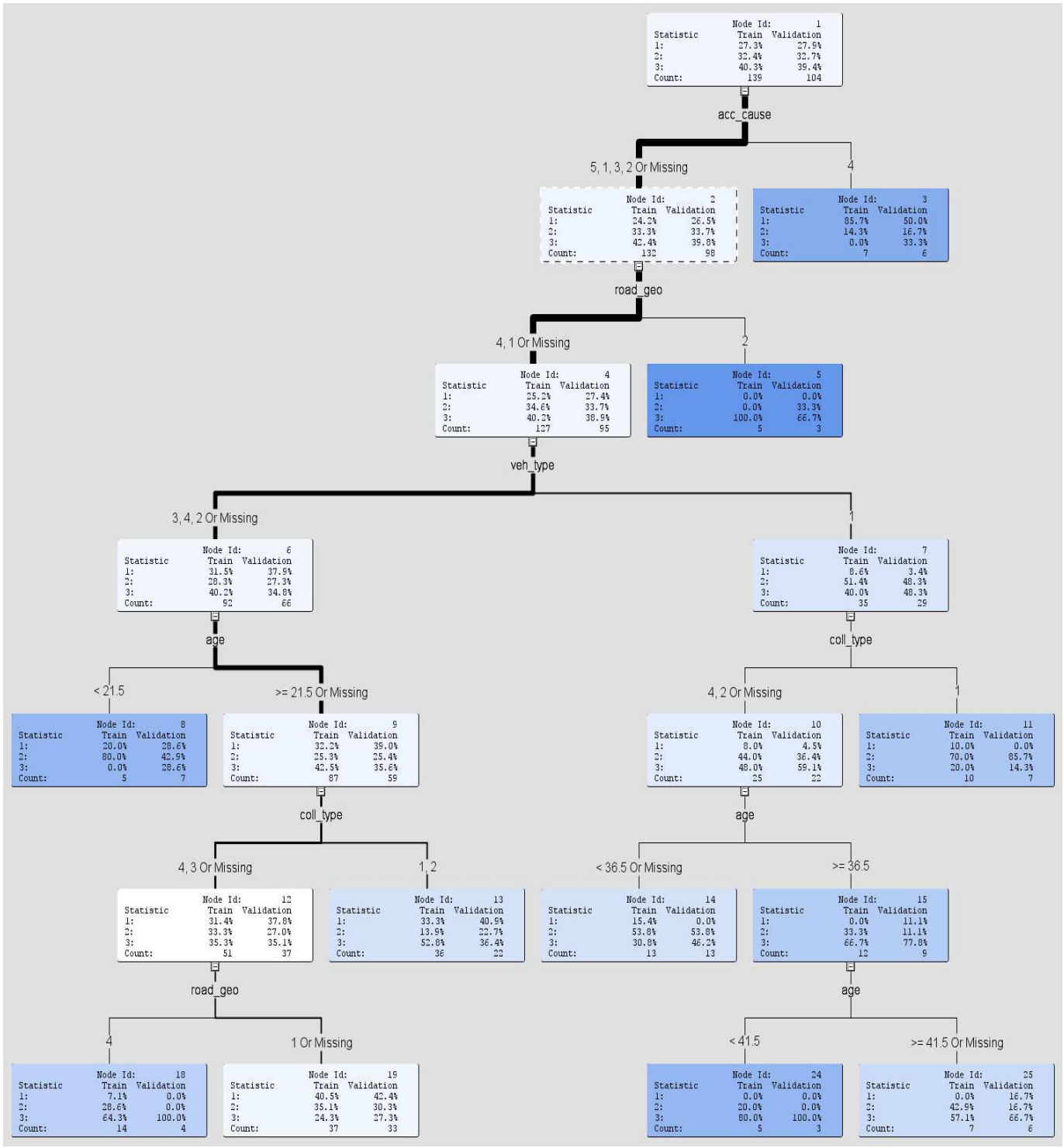

Fig. 1 Decision tree output

IV. CONCLUSION

As a conclusion, the result of this study shows that decision tree which is CART is the most suitable for modelling road accident injury severity. One big advantage of the decision tree is its transparent nature. The decision tree makes explicit possible alternatives and traces each alternative to its conclusion in a single view. This makes the comparison among those alternatives much easier. The result obtained from CART provided important information on how the accident cause, road geometry, vehicle type, age and collision type are related to types of injury when road accidents occur. The most important variable in predicting types of injury of road accident occurrence is accident cause. The accident causes included speed, run a red light, follow too close and careless while overtaking does not really contribute to the type of injury. Overtaking may give more impact on another issue such as contributing to fatality. So, in order to prevent and reduce the road accident from happening, modifying the road environment is recommended since this action can slow down the speed. The other action that can be taken into account is installing and adding more

bumps and roundabouts on the roads. There are some limitations in this study that analyses the police reports. First, it is time-consuming because the researchers need to summarize all the reports into a favorable data form. Besides, there is a restriction of having a large data set since the report forms are private and confidential. So, the researchers are only given a small portion of the data set. In addition, there are many important independent variables not being included in the report form such as road surface condition, weather, the speed of the vehicles and seatbelt use as well as alcohol and drug contribution to the road accident occurrence.

Hence, the result obtained from this study might further be improved by collecting more data in the future study. Besides, more independent variables should be made available in the agencies that deal with road accident occurrence.

ACKNOWLEDGMENT

The researchers would like to thank Kota Bharu District Police Headquarters, Department of Traffic officers for their contribution of data for this study.

REFERENCES

[1] World Health Organization (WHO). (2015) 10 Facts on global road safety. [Online]. Available: http://www.salute.gov.it/imgs/C_17_pubblicazioni_1662_ulteriorialllegati_ulteriореallegato_0_alleg.pdf

[2] M. T. Obaidat and T. M. Ramadan, "Traffic accidents at hazardous locations of urban roads," *Jordan Journal of Civil Engineering*, vol. 6, pp. 436-447, Oct. 2012.

[3] A. R. A. Rahman. (2015) Malaysia ke-20 rekod kemalangan tertinggi. [Online]. Available: http://m.utusan.com.my/berita/nahas-bencana/malaysia-ke-20-tertinggi-catat-kemalangan-jalan-raya-1.158437

[4] W. F. W. Yaacob, W. Husin, and W. Zakiyatussariroh. (2005) Modelling Malaysian road accident deaths: An econometric approach. [Online]. Available: https://www.mysciencework.com/publication/show/4b06b048160fd6aa20b2c753be69370c

[5] A. Pakgohar, R. S. Tabrizi, M. Khalili, and A. Esmaeili, "The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach," *Procedia Computer Science*, vol. 3, pp. 764-769, Dec. 2001.

[6] S. Kulanthayan, T. H. Law, A. R. Raha, and U. R. Radin, "Seat belt use among car users in Malaysia," *IATSS Research*, vol. 28, pp. 19-25, Dec. 2004.

[7] P. V. Elslande, C. L. Naing, and R. Engel. (2008) Analyzing human factors in road accidents: TRACE WP5 summary report. [Online]. Available: http://www.trace-project.org/publication/archives/trace-wp5-d5-5-v2.pdf

[8] M. Bedard, G. H. Guyatt, M. J. Stones, and J. P. Hirdes, "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities," *Accident Analysis and Prevention*, vol. 34, pp. 717-727, Nov. 2002.

[9] G. Zhang, K. K. Yau, and G. Chen, "Risk factors associated with traffic violations and accident severity in China," *Accident Analysis and Prevention*, vol. 59, pp. 18-25, Oct. 2013.

[10] S. Dissanayake and J. J. Lu, "Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes," *Accident Analysis and Prevention*, vol. 34, pp. 609-618, Sep. 2002.

[11] D. I. White, "An investigation of factors associated with traffic accident and casualty risk in Scotland," Phd thesis, Edinburgh Napier University, Scotland, 2002.

[12] M. Mohanty and A. Gupta, "Factors affecting road crash modeling," *Journal of Transport Literature*, vol. 9, pp. 15-19, Apr. 2015.

[13] Y. Wah, N. Nasaruddin, W. Voon, and M. Lazim, "Decision tree model for count data," in *Proc. WCE'12*, 2012, p. 1.

[14] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," *Accident Analysis and Prevention*, vol. 34, pp. 729-741, Nov. 2002.

[15] L. E. DPhil, "Causal influence of car mass and size on driver fatality risk," *American Journal of Public Health*, vol. 91, pp. 1076-1081, Jul. 2001.

[16] Department of Transport. (2015) Reported road casualties Great Britain: Annual report 2014. [Online]. Available: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2014

[17] M. Singleton, H. Qin, and J. Luan, "Factors associated with higher levels of injury severity in occupants of motor vehicles that were severely damaged in traffic crashes in Kentucky, 2000-2001," *Traffic Injury Prevention*, vol. 5, pp. 144-150, Jun. 2004.

[18] M. Khalili and A. Pakgohar, "Logistic regression approach in road defects impact on accident severity," *Journal of Emerging Technologies in Web Intelligence*, vol. 5, pp. 132-135, Jan. 2013.

[19] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The Journal of Educational Research*, vol. 96, pp. 3-14, Sep. 2002.

[20] R. B. Noland and L. Oh, "The effect of infrastructure and demographic change on traffic-related fatalities and crashes: A case study of Illinois county-level data," *Accident Analysis and Prevention*, vol. 36, pp. 525-532, Jul. 2004.

[21] W. F. W. Yaacob, M. A. Lazim, and Y. B. Wah, "Applying fixed effects panel count model to examine road accident occurrence," *Journal of Applied Sciences*, vol. 11, pp. 1185-1191, Jul. 2011.

[22] A. T. Kashani, A. Shariat-Mohaymany, and A. Ranjbari, "A data mining approach to identify key factors of traffic injury severity," *PROMET-Traffic and Transportation*, vol. 23, pp. 11-17, Jan. 2011.

[23] P. M. Kuhnert, K. A. Do, and R. McClure, "Combining non-parametric models with logistic regression: An application to motor vehicle injury data," *Computational Statistics and Data Analysis*, vol. 34, pp. 371-386, Sep. 2000.

[24] T. Beshah and S. Hill, "Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia," in *Proc. AAAI'10*, 2010, p. 1.

[25] A. A. L. L. Sandoval-Mejıa, and Y. E. Wang. (2016) Multinomial logistic regression. [Online]. Available: http://myweb.ttu.edu/pwestfal/ISQS5349/MLR_16.pdf.

[26] A. Westergren, S. Karlsson, P. Andersson, O. Ohlsson, and I. R. Hallberg, "Eating difficulties, need for assisted eating, nutritional status and pressure ulcers in patients admitted for stroke rehabilitation," *Journal of Clinical Nursing*, vol. 10, pp. 257-269, Mar. 2001.

[27] U. K. Rao and D. U. Shekhar, "Building customer relationship management through Data Mining and Data Warehousing (DMDW)," *International Journal of Research in Organizational Behavior and Human Resource Management*, vol. 4, pp. 125-134, 2016.

[28] V. S. Jotsov and E. Iliev, *Applications of Advanced Analytics Methods in SAS Enterprise Miner*, ser. Advances in Intelligent Systems and Computing. Switzerland: Springer International Publishing, 2015, vol. 323.

[29] I. M. Yassin, A. Zabidi, M. S. A. M. Ali, N. M. Tahir, H. A. Hassan, H. Z. Abidin, and Z. I. Rizman, "Binary particle swarm optimization structure selection of nonlinear autoregressive moving average with exogenous inputs (NARMAX) model of a flexible robot arm," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, pp. 630-637, Oct. 2016.

[30] M. N. M. Nor, R. Jailani, N. M. Tahir, I. M. Yassin, Z. I. Rizman, and R. Hidayat, "EMG signals analysis of BF and RF muscles in autism spectrum disorder (ASD) during walking," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 6, pp. 793-798, Oct. 2016.