Journal of Engineering Science and Technology
Special Issue on ACSAT, October (2021) 1 - 14


# A DATA MINING APPROACH TO ANALYSE CRASH INJURY SEVERITY LEVEL

ANGELA SIEW HOONG LEE[1,*], LING SZE YAP[1],
HUI NA CHUA[1] , YEH CHING LOW[1], MAIZATUL AKMAR ISMAIL[2]

[1]School of Engineering and Technology, Sunway University, 47500 Selangor, Malaysia
[2]Faculty of Computer Science and Information Technology, University of Malaya, 50603 KL, Malaysia
*Corresponding Author: angelal@sunway.edu.my

## Abstract

Existing research on road accidents has primarily focused on the relationship between a set of risk factors (speeding, intoxicated driving, or poor road conditions) and its impact on accidents. Currently, there is a lack of research on how many risk variables contributing to road accidents, such as human, vehicle, environmental, road, geographical, and crash type-related factors, have resulted in road crashes and their severity levels. Understanding the impact of these factors on traffic accidents allows authorities and researchers to acquire a better understanding of which factors demand more attention in order to reduce traffic accidents. As a result, the goal of this research is to identify and categorise the aspects based on the risk factor characteristics, and then further classify the aspects depending on the severity level of injury in an accident using a decision tree model. The model was created using crash data from Kaggle on a city of America, from 2004 to 2017. Several risk factors (human factor, vehicle factor) were identified to have an influence on the four-level (no injury, minor and moderate injuries, major injury, killed) of injury severity. The prediction accuracy of the model is measured at 65.78% based on the test data set.

Keywords: Data mining, Decision tree, Injury severity level, Traffic accidents.

## 1. Introduction

Road traffic accidents have long been a global problem that poses a daily danger to the economy and social order. According to the World Health Organization in 2021, it is estimated that 1.3 million people dying each year due to road traffic crashes [1]. It is also stated in the report that crash injuries were the primary fatality cause for kids and young adults aged 5 - 29 years. As a result, the United Nations General Assembly has determined to achieve a 50% reduction in the global number of fatality and injuries due to road traffic crashes [1]. Furthermore, with the proportionally increasing worldwide population and increased motorization, road safety preventative measures have become more important than ever.

Essential procedures and actions are required in order to address the aforementioned issue and reduce the number of deaths and injuries. However, there is a demand to recognize the contributing factors in causing the traffic accidents for the development of effective road safety precautions. Past traffic accidents data have been utilized by the researchers in finding out the underlying causes of such accidents, which aids in forming the right solutions to reduce or prevent this issue in time to come.

In recent years, a lot of historic road accident analysis has been done utilising various data mining techniques in the hopes of deriving useful insights from the data. Example of techniques includes Decision Tree (Classification and Regression Tree), Classification Techniques (Naïve Bayes, Clustering), Artificial Neural Network (ANN), and a combination of cluster analysis and statistical modelling using the Negative Binomial distribution [2-10]. Decision Tree is notably useful in analysing traffic accident data due to its non-parametric nature that does not require pre-defined relationship between the predictors and the target.

Previous studies are mostly focused on examining the relationship between a fixed risk factor (speeding, drunk driving or road condition) and its effect on road accidents [2-4]. Therefore, this paper aims to discover the determinants - from the human factor, vehicle factor, environmental factor, road factor, geographical factor, types of crash, and other factors of which have led to road crashes and its severity level, using classification and regression trees (CART) technique. This technique was used to explain several independent factors in terms of a continuous and categorical dependent variable which means the independent variables might be categorical or continuous.

The section that follows provides an overview of the literature on the various model strategies and approaches used in forecasting injury severity levels in traffic crashes. Section 3 describes the data and variables used in this study. The methodology of the study is illustrated in Section 4. Section 5 displays the results and discussion of the findings. Lastly, the study is concluded in Section 6.

## 2. Literature review

### 2.1. Crash analysis using data mining

The growth of data mining has made it easier to find hidden patterns and relationships from historical data, which was previously difficult to achieve using traditional statistical methods. The use of data mining in road traffic study aids in the determination of accident causation as well as a thorough understanding of the

problem and potential remedies. Data mining focuses on finding new and unobserved insights that are explicit to the data, rather than updating existing hypotheses acquired from previous evidence-based statistical research [4]. In the broad term context of decision tree, there are different kind of techniques and algorithms available for building a decision tree [11-14] like classification and regression trees (CART), C4.5, which have been used in the previous studies to identify the probable factors which are likely to cause a road crash with the objective to regulate and eliminate those factors. Its hierarchical display allows easy interpretation [5].

## 2.2. Classification and Regression Trees (CART)

CART method is firstly introduced by Quinlan [15] back in 1984 and is the most common approach in building a decision tree for car crash analysis. It has been widely used by Chang and Chien [6], Kashani and Mohaymany [7], Kuhnert et al. [16] in analyzing the severity of traffic injury. There are many algorithms available when considering a decision tree construction. The distinction of the various algorithms is the splitting criteria used in each decision tree computation. The splitting criteria used in CART is the purity degree measurement - Gini Index. Data model built using CART is the outcome of a recursively partition of data and it will produce a binary large tree at the end. The tree will then be pruned to a smaller size to reduce its misclassification error rate. When it comes to validating the mode, CART uses a 10-fold cross-validation by default [16]. The dataset is divided into 10 subsets and 1 subset will be the validation data while the remaining 9 subsets will be used to train the model. The validation process will be repeated 10 times with each of the 10 subsets used once as validation data. After 10 rounds, the results will be averaged out to get a single final estimation rate. However, it is argued that the binary tree produced using CART might not be able to present the result for efficient interpretation.

## 2.3. C4.5

C4.5 is another popular method that is adopted in traffic accident severity analysis where it was first introduced by Quinlan in 1993 [15]. López et al. applied this algorithm in their research to identify the causes that affect the accident severity [5]. As mentioned above, CART uses Gini index as its splitting criteria, but C4.5 uses the Information Gain Ratio (IGR). The IGR of a variable X on a variable class Y can be expressed using Eq. (1):

$$IGR\ (C,Y) = \frac{IG(C,Y)}{H(Y)} \quad (1)$$

where $H(Y)$ is the entropy of Y. Furthermore, according to López et al. [5], C4.5 performed well with continuous variables as well as missing values. Once a tree is developed, hypothesis testing will be done to see whether it needs to be trimmed so that more branches can be split. Furthermore, unlike CART, this technique is not limited to creating binary trees.

## 2.4. Related work

Numerous researches have been conducted in the past in order to discover the significant factors that contribute to the severity of an accident. When it comes to road accident severity analysis, researchers used parametric modelling such as logit

models in the past. Al-ghamdi [8] used logistic regression in examining the contribution of various factors to crash severity. The results showed that the location (intersection, exit, road section, etc.) of which the accident happened and the accident cause (speeding, run red lights, not giving priority, etc.) has a detrimental effect to injury severity. Meanwhile, Loh [17] utilized multivariate logistic regression in investigating the contributing factors to which will cause driver fatalities. It was shown that the driver fatal risk would be reduced with the use of the seatbelts and the vehicle size was associated to the injury severity.

Another frequent method for determining the severity of an accident is ordered probit modelling. It evaluates the association between an ordered target variable and other predictors in the same way that logistic regression modelling does. Zajac and Ivan [18] adopted ordered probit modelling in evaluating the effect of the road and area type (downtown, rural areas, residential area, commercial area, etc.) on the pedestrian injury severity of whom was involved in a car crash. It was shown that pedestrian injuries were more common in rural regions and low-density residential regions. Renski et al. [19] observed that the speed limit had an influence on the severity of collision injuries using ordered probit. According to the findings, the link between speed limit and injury severity is directly proportionate [19].

In recent years, academics have moved away from parametric modelling and toward non-parametric methodologies combined with data mining approaches for studying traffic accidents. Researchers can use these approaches to detect trends and categorise elements that influence different levels of injury severity [7]. López et al. [5] conducted research to find out the contributing factors to accident severity, where three methods - CART, ID3, and C4.5 were used in the study to compare their accuracy. Among the three, ID3 has the lowest accuracy rate, whereas CART and C4.5 accuracy is ranging close. In addition, the Receiver Operating Characteristics (ROC) chart revealed that CART is the best model of all. The downside of CART method is that it produces a binary tree which grouped some variables under one branch in which will increase the node support, but it is challenging to analyse the effect of the variables individually. On the other hand, C4.5 creates more branches but not all the rules generated can meet the minimum support threshold. According to the findings, male drivers are more likely to cause a fatal or badly wounded accident, and the probability increases when pedestrians are involved.

The accuracy of three data mining approaches - neural network, logistic regression, and decision tree - in identifying factors that impact the severity of traffic collisions in Korea was compared by Sohn and Shin [9]. They first categorized the severity into 3 levels (i) death or major injury (ii) minor injury (iii) property damage. A preliminary variable selection was carried out using chi-square tests of independence before the modelling. The 22 variables selected were used to train the first decision tree that is used for further variable selection. The 5 possible factors shown in the first tree were used to model the second decision tree and the following neural network and logistic regression. The trees with 22 input variables and 5 input variables have an accuracy rate of 56.1% and 56.3% respectively which have no significant difference. Thus, to improve the accuracy rate, Sohn and Shin [9] further grouped the severity level into 2 categories, namely (i) death or injury, and (ii) property damage. The same parameters were applied again to train a new neural network and logistic regression models. Lastly, the accuracy rate of the three different models were put together, and they did not differ much from each other.

Due to its simpler form, the decision tree came out as the easiest to comprehend of the three models.

Kwon et al. [10] performed a study on identifying the risks factor using 2 different classification algorithms - Naïve Bayes classifier and Decision Tree. The Naïve Bayes classifier is the simpler version of the Bayes classifier. As opposed to the Bayes classifier which requires heavy computational costs, the Naïve Bayes allows a less expensive training of the model, easier interpretation, and under the assumption that there is no multicollinearity in the data [20]. The performance of both the methods were compared against the binary logistic regression. While most of the studies focused on determining the factors, this study intended to identify the dependencies among the factors. Kwon et al. [10] categorized the severity level into (i) death of injured (ii) property damage, and 70% of the dataset was used to train the model while the remaining of it was used for validation. A ROC chart was applied to evaluate the performance of the models. It was shown in the chart that the logistic regression model was underperform than both the Naïve Bayes and Decision Tree models. The Decision Tree model that considers of the variable's dependencies enhances the performance as compared to the Naïve Bayes.

## 3. Data Description

The data employed in this study is an archive data collection of 170358 records, which is in line with the aforementioned research goals. The data contains areas and information about every crash incident reported to the police in one of the states in America from 2004 to 2018. But the focus of this paper is not the location but the algorithm we used in analysing the severity level of crash injury.

This data encompasses attributes related to demographic details, vehicles, road conditions, environmental circumstances, the time and day of the accident, and injury severity. The data was selected due to its comprehensiveness and wide dimension of variables at disposal, which is useful to determine the possible risk factors that might contribute to the injury severity. There are several major categories that can be derived from the data when examining the possible factors influencing the injury severity, namely Human Factor, Vehicle Factor, Environmental factor, Road Factor, Geographical Factor, Types of Crash, and Other Factors.

### 3.1. Human factor

The human component may be seen from two perspectives: the driver's perspective and the non-perspective. driver's (passengers, pedestrians). Driver factor can be further divided into two elements:

i. driver's characteristics

ii. driver's behaviour.

In this data set, there is only one item recorded for driver's characteristics, which is the age of the driver. The driver age is divided into 8 classes: Age 16, 17, 18, 19, 20, 50-64, 65-74, 75. Driver's behaviour refers to the driver's conduct and condition during the journey of driving. There are several fields available in this data set, namely distracted driver indicator, drinking driver indicator, underage drinking driver indicator, unlicensed driver indicator, driver using cell phone indicator, fatigue or asleep indicator, drug use indicator, aggressive driving

indicator, speeding indicator, tailgating indicator, driver running red light indicator, and driver running stop sign indicator.

Unbelted indication, alcohol usage indicator, and illicit drug use indicator are examples of indications that include both drivers and passengers. Pedestrian indicator which denotes the presence of pedestrian during the crash is also considered human factor. The total number of people involved in an accident (person_count) was also considered.

## 3.2. Vehicle factor

Types of vehicles taken into consideration in this study such as

i. automobile,
ii. school bus,
iii. motorcycle, bus,
iv. small truck,
v. heavy truck,
vi. SUV.

## 3.3. Environmental factor

These are the several environmental factors employed in this study:

i. illumination (daylight, dark - no streetlights, dark - streetlights, dusk, dawn, dark - unknown roadway lighting, other),
ii. weather (no adverse condition, rain, sleet, snow, fog, rain and fog, sleet and fog),
iii. hour, day of week, month, and year of the crash.

## 3.4. Road factor

There are several road factors under consideration in this research:

i. road condition (dry, wet, sand/ mud/ dirt/ oil/ gravel, snow covered, slush, ice, ice patches, water - standing or moving, other),
ii. crash's relativity to the road (on roadway, shoulder, median, roadside - off trafficway; on vehicle area, outside trafficway - in area not meant for vehicles, in parking lane, gore - intersection of ramp and highway),
iii. intersection type (mid-block, four-way intersection, "T" intersection, "Y" intersection, traffic circle or round about, multi-leg intersection, on ramp, off ramp, crossover railroad crossing, other),
iv. lane closure indicator, non-intersection indicator, intersection indicator, signalized intersection indicator.

## 3.5. Geographical factor

There are four fields relating to the location of the crash Population distribution of the crash location, whether it is a

i. rural area,

ii. small urban area,

iii. urbanized area,

iv. highly urbanized area.

Other fields are school zone indicator and work zone indicator, of which whether the crash occurred in a school zone or work zone.

### 3.6. Types of crash

The collision type is captured as it might also influence the severity level of accident. There are nine groups of collision collected in this data: non collision, rear-end, head-on, rear-to-rear, angle, same direction sideswipe, opposite direction sideswipe, hit fixed object, hit pedestrian, overturned, run-off road. There are different indicators available if the vehicles involved were to hit on different objects - hit deer indicator, hit tree shrub indicator, hit embankment indicator, hit pole indicator, hit guide rail indicator, hit barrier indicator, hit parked vehicle, and hit bridge indicator.

### 3.7. Other factors

Other factors include the vehicles condition after the crash occurred, such as at least one vehicle with fire damage indicator, and vehicle failure indicator.

## 4. Research Method

### 4.1. Data understanding

This data set contains 158 variables. Each record depicts a crash incident occurred was reported to the police. Information such as injury severity, fatalities, details about the vehicles involved, the geographic features of locations, and elements that may have been instrumental in causing the accident.

### 4.2. Data preparation

This data cleaning step was performed using R before importing the data to SAS Enterprise Miner.

i. Imputation. The missing values that were discovered in this data set were imputed with the mode value of each variable as the variables are nominal data type.

ii. Values removal. Records with severity level of the crash of 8 and 9 were removed from the data set because 8 represents “unknown severity” and 9 represents “unknown”. This serves to narrow the scope of the study as severity level is the target of the analysis.

iii. Variable selection (Table 1). Variables such as street name that are of no use to the study were removed. There are numerous of unique streets to be taken account into which will cause the curse of dimensionality. Duplicated variables such as pedestrian indicator and count of pedestrian were detected, and pedestrian count was removed to avoid any redundancy. Variables that

are associated with one another were also resolved by removing one of the variables. For example, minor injury indicator and severity level (which consists of minor injury) are associated with one another. Therefore, minor injury indicator will be removed from the data set.

**Table 1. Types of variables selected.**

| Factors | Types of Variables Selected |
|---|---|
| **Human Factor** | 1. Driver characteristics<br>2. Driver behaviour<br>3. Passenger behaviour<br>4. Presence of pedestrian |
| **Vehicle Factor** | 1. Types of vehicles |
| **Environmental Factor** | 1. Illumination<br>2. Weather<br>3. Date and time of the crash |
| **Road Factor** | 1. Road condition<br>2. Crash's relativity to the road<br>3. Intersection type |
| **Geographical Factor** | 1. Crash population distribution<br>2. School zone<br>3. Work zone |
| **Types of Crash** | 1. Collision type<br>2. Objects hit during the crash |
| **Other Factors** | 1. Vehicle's condition after the crash<br>2. Vehicle failure indicator |

**Discretization.** The values of Hour in which the crash happened are classified into six brackets (00:00 - 06:00, 07:00 - 08:00, 09:00 - 11:00, 12:00 - 16:00, 17:00 - 19:00, 20:00 - 23:00) to decrease the data from a wide selection of numeric values to a subgroup of categorical values. This will aid in the data modelling as the model will be able to learn faster and accuracy will be enhanced.

In the original dataset, the injury severity level (target variable) is classified into seven levels - not injured, killed, major injury, moderate injury, minor injury, injury/ unknown severity, unknown.

The unknown class and injury with unknown severity are removed from the study to reduce the noise. This brings the target variable to five levels, which is highly explanatory for clearly distinguishing severity levels.

However, the distribution of the crash records is highly disproportionate among the levels. For instance, the percentage of fatality is relatively low with only 0.58% reported in comparison to the percentage of no injury of 48.83%, where approximately half of the records are placed as no injury.

As a result, the data analysis findings from using these five levels of goal are likely to be substantially skewed in comparison to the distribution [15]. As a result, the overall model's or each group's misclassification rate is likely to grow.

As a result, mild and moderate injuries will be grouped together to increase model accuracy, resulting in 4 levels of injury levels shown in Table 2.

**Table 2. Injury severity levels definition and used in this study.**

| Raw Data | Definition | Percentage | This Study |
|---|---|---|---|
| **No injury** | No one is injured in the crash. | 48.83% | No injury |
| **Minor injury** | Possible injury. Injury can be treated by first-aid application. | 22.95% | Minor and moderate injuries |
| **Moderate injury** | Non-incapacitating injury which includes bruises, abrasions, swelling. Injury which requires medical treatment or hospitalization. | 8.66% | |
| **Major injury** | Incapacitating injury which includes bleeding wounds and distorted members. Injury which requires transport of the patient from the scene. | 1.84% | Major injury |
| **Killed** | Death of one or more person within 30 days of accident. | 0.58% | Killed |
| **Injury/ Unknown severity** | Injury with unknown severity. | 13.56% | |

## 4.3. Decision tree

In this work, a non-parametric tree-based approach called decision tree is used to assess and forecast the severity degree of a crash's damage since decision tree allows for multi-level classes of the target variable. When predicting a nominal target variable, a decision tree is known as a classification tree, and when predicting a continuous target variable, it is known as a regression tree. The CART method is a classification technique that uses Gini's impurity index as a splitting criterion to create a decision tree. Therefore, decision tree was used in this process.

The establishment of decision tree process involves three steps, which are tree induction, tree pruning, and optimum tree selection [7]. A decision tree is first constructed by having a root node, which consists of all the data, placed at the top of the tree. The root node is then separated into two child nodes based on an attribute that improves the purity the most. This implies that the proportion of data in each node carries the highest instance of a single category. The child nodes derived from the parent nodes are ought to has higher homogeneity than the parent nodes. This process of tree induction is ceased when all the data in each node have the greatest purity. A maximal tree is produced at the end of the tree induction, which consists of a target class in each terminal node. A maximal tree generally has a considerable number of terminal nodes and is inadequate in classification and prediction capabilities. Hence, an optimum tree is usually deduced to prevent an overfit model, which lack in indicative samples.

Because the goal of this study is to forecast the severity of a traffic accident's injuries, a classification tree was created using the target variable, injury severity level, as an ordinal variable. The Gini index was utilised as the splitting criterion in

this study, and it was utilised to identify the factors that increased the purity index of the succeeding leaf nodes. The purpose of tree pruning is to obtain an optimum tree by removing the insignificant leaf nodes from a maximal tree. The optimum tree is determined with respect to the complexity and the misclassification rate of the testing data. Therefore, the data will be separated into training data, which serves to model the decision tree and validation data, which serves to assess the performance of the model by measuring the misclassification rate. Training data set will comprise 70% of the whole data set while the remaining of 30% data accounts for validation. The data will be partitioned by adopting the stratified sampling method. This method is selected to enhance the precision of classification as it will preserve the representation of data within each level of injury.

As a summary the research method of this study is depicted in a flowchart (Fig. 1):

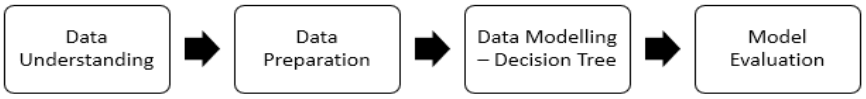


**Fig. 1. Methodology flowchart.**

## 5. Results and Discussion

### 5.1. Variable importance

Due to the large number of fields accessible in the data set, Table 3 displays the important variables in the decision tree by calculating the relative significance of variables for the model, which is useful in analysing the variables. Individual factors are utilised as the foundation for assessing the many categories that influence injury severity. It was discovered that the human element, vehicle factor, and different sorts of crash all have a role. Under human factor, pedestrian indicator, the use of seat belt, and total number of people involved are crucial variables in the model. Based on the findings, if pedestrian was involved in the accident, there was a higher probability of injury sustained as compared to no injury sustained. It was discovered that this restraint and safety system could lower the odds of an occupant being injured. Moreover, the presence of motorcycle and bicycle in a crash, which is categorized as vehicle factor, will tend to cause the injury to be more severe, or even result in death. The types of crash (overturned indicator, collision type, fire damage indicator) are also salient in studying the injury effect of the crash. Other factors such as vehicle damage due to fire are considered important determinant.

**Table 3. Variable importance of each variable.**

| Variable Name | Importance | Validation Importance |
|---|---|---|
| **PEDESTRIAN** | 1.0000 | 1.0000 |
| **MOTORCYCLE** | 0.7428 | 0.6996 |
| **UNBELTED** | 0.7237 | 0.6553 |
| **BICYCLE** | 0.5230 | 0.4767 |
| **PERSON_COUNT** | 0.1900 | 0.1743 |
| **OVERTURNED** | 0.1533 | 0.1636 |
| **COLLISION_TYPE** | 0.0775 | 0.0807 |
| **FIRE_IN_VEHICLE** | 0.0600 | 0.0376 |

## 5.2. Decision tree

The most optimum classification tree was generated by SAS Enterprise Miner, which consists of twelve leaf nodes. It reveals that vehicle factor and human factor are significant in segmenting all levels of injury severity in a traffic crash.

The root node, which is the first node of the tree, is split into two child nodes based upon the vehicle factor, namely the motorcycle indicator. This implies that the primary variable to divide and predict the injury severity in an accident is whether motorcycle is involved in the crash. On the left branch of the tree, eight leaf nodes are identified. Among all these eight leaf nodes, there are two nodes correspond to no injury, and another six nodes correspond to minor injury to moderate injury. On the contrary, the right branch has three leaf nodes, where all nodes belong to minor to moderate injury. There is no rule to predict major injury or fatality in this decision tree model.

A few node rules are developed to better interpret the decision tree results. In classifying and predicting non-injury traffic accident, vehicle factor, human factor, and other factors are pronounced. The variables that are notable for vehicle factor includes motorcycle indicator and bicycle indicator, whereas variables for human factor encompass pedestrian indicator, unbelted indicator and total count of people involved in an accident, and other factors are overturned indicator. For instance, the accident is to occur with the conditions of the absence of motorcycle and bicycle, the use of seat belt by occupants, no pedestrian is being hit, the total number of people involved in the crash is less than or equal to four, and the crash does not result in vehicle being overturned. This finding stresses the importance of utilizing the safety seat belts as it secures the body of a driver or passengers in the original position, which minimize the risk of injury. Besides, the use of bicycle or motorcycle as a means of transportation is more susceptible to cause injuries to the riders in a collision due to the riders are completely exposed to the impact of other vehicles or its surrounding. This rule has an accuracy rate of 69.96% in predicting no injury.

With respect to minor to moderate injury prediction, the road factor (intersection type) is crucial besides vehicle factor (motorcycle and bicycle indicator) and human factor (pedestrian indicator and total number of people involved in the collision). There are two distinct storylines that could be established for this prediction. Firstly, when bicycle is involved in the accident, then there is a high likelihood of 91.24% that the occupants will suffer from minor to moderate injury. This is because the impact arose from the collision might cause injuries to the body. The second rule is that the crash involves pedestrian on a four-way intersection or "T" intersection and having more than eight people involved in the accident. This rule has a predicting accuracy rate of 86.68%.

In relation to segmenting and predicting the occupants with major injury or killed, the accuracy rate is comparatively low due to the limited records accessible. However, a rule that appertains to both major injury and fatality could be constructed if the accuracy of both the major injury and fatality are computed together. This scenario is associated with the vehicle factor, human factor, road factor, and types of crash. In this instance, an occupant or driver could sustain major injury or death if the collision involves pedestrian on a closed lane. This rule gives a 32.48%. Besides, an accident will also have a higher probability to cause major injury or death when there is a fixed object collision, or an angle collision, or head-

on collision, or sideswipe collision, or pedestrian is hit and there is motorcycle involved in the crash. This rule will render a 27.55% accuracy rate.

In summary as shown in Table 4, it was discovered that the vehicle factor, human factor, road factor, types of crash, and other factors are significant in predicting the injury severity level of an accident. This model is not satisfactory in associating environmental factor and geographical factor to the injury severity level of a crash.

**Table 4. Factors associating with each injury severity levels.**

| Factors | No Injury | Minor and Moderate Injuries | Major injury | Killed |
|---|---|---|---|---|
| **Human** | ✓ | ✓ | ✓ | ✓ |
| **Vehicle** | ✓ | ✓ | ✓ | ✓ |
| **Environmental** | | | | |
| **Road** | | ✓ | ✓ | ✓ |
| **Geographical** | | | | |
| **Types of Crash** | | | ✓ | ✓ |
| **Others** | ✓ | | | |

## 5.3. Performance of the model

In this study, the model's accuracy refers to the percentage of data that are correctly classified and forecasted. In training data, the decision tree model had a 66.24 percent accuracy, and in testing data, it had a 65.78 percent accuracy. Despite this, the model's ability to classify each severity category, particularly significant injury and death, is very unsatisfactory due to the disproportionality of the data.

A confusion matrix and its measures are computed in Tables 5 and 6 respectively to measure the accuracy rate of each injury severity level. This model has the highest precision in predicting no injury, indicating that there is 95.71% that the no injury prediction appears to be accurate in the actual sample. However, the precision rate for the other levels is rather inadequate. The recall rates of no injury and, minor and moderate injuries are 0.6658 and 0.6466 respectively. This denotes that 66.58% is correctly predicted out of all the no injury sample, and 64.66% is correctly predicted out of all the minor and moderate injuries sample.

**Table 5. Precision and recall rates of each injury severity level.**

| Actual | Precision | Recall |
|---|---|---|
| **No injury** | 0.9571 | 0.6658 |
| **Minor and moderate injuries** | 0.2095 | 0.6466 |
| **Major injury** | 0 | 0 |
| **Killed** | 0 | 0 |

## 6. Conclusion

To illustrate the empirical link between possible risk variables (human factor, vehicle factor, road factor, geography factor, environmental factor, types of crash, other variables) and injury severity levels in the site, the decision tree model was used. The results revealed that human factor, vehicle factor, road factor, types of crash, and other factors are contributing determinants to the injury severity level of crash, whereas environmental factors and geographical factors are least significant in this model. However, the accuracy of the model, which is 65.78%, is less satisfactory due

to the highly skewed nature of distribution of the records. This model suggests a less useful/functional in grouping and predicting major injury and fatality.

Nonetheless, this decision tree model is a valuable tool for determining the risk variables that are linked to the degree of a crash's injury. It may be easily understood and interpreted because to its graphical depiction of findings. It also tells a tale about the underlying link discovered between the explanatory factors and the severity of the injury. Significant variables could be ascertained easily as decision tree model is a pragmatic approach in dealing with plentiful data of numerous explanatory variables.

In the future, the multi-class target variable may be reduced down to two or three classes to minimise the rate of misclassification, and a comparison between the models may be established. The reduction of classes can be performed in which the target classes are limited to minor injury and at least serious injury, or minor injury, moderate to major injuries and fatality, or other possible combinations. Other non-parametric methods such as association rules and artificial neural network can also be adopted to examine the possible contributing factors that influence the level of injury severity.

## References

1. W. H. Organization (2021). Road traffic injuries. Retrieved August 16, from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
2. Bahiru, T.K.; Singh, D.K; and Tessfaw, E.A. (2018). Comparative study on data mining classification algorithms for predicting road traffic accident severity. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT).* 1655-1660.
3. Karimnezhad, A.; and Moradi, F. (2017). Road accident data analysis using Bayesian networks. *Transportation Letters*, 9(1), 12-19.
4. Raihan, A.; Hossain, M.; and Hasan, T. (2018). Data mining in road crash analysis: the context of developing countries. *International Journal of Injury Control and Safety Promotion*, 25(1), 41-52.
5. López, G.; Abellán, J.; and O˜na, J.D. (2013). Extracting decision rules from police accident reports through decision trees. *Accident Analysis and Prevention*, 50, 1151-1160.
6. Chang, L.; and Chien, J. (2013). Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. *Safety Science*, 51(1), 17-22.
7. Kashani, A.T.; and Mohaymany, A.S. (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. *Safety Science*, 49(10), 1314-1320.
8. Al-ghamdi, A.S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention*, 34(6), 729-741.
9. Sohn, S.Y.; and Shin, H. (2001). Pattern recognition for road traffic accident severity in Korea. *Ergonomics*, 44(1), 107-117.

10. Kwon, O.H.; Rhee, W.; and Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. *Accident Analysis and Prevention*, 75, 1-15.
11. Chan, K.W.; Lee, A.S.H.; and Zainol, Z. (2020). Profiling patterns in healthcare system: A preliminary study. *International Journal of Advanced Computer Science and Application*, 11(4), 661-668.
12. Lee, A.S.H.; Ng, C.; Zainol, Z.; and Chan, K.W. (2019). Decision tree: customer churn analysis for a loyalty program using data mining algorithm. *In: Berry M., Yap B., Mohamed A., Köppen M. (eds) Soft Computing in Data Science: 5th International Conference, SCDS 2019. Communications in Computer and Information Science*, 1100, 14-27.
13. N, K. -W. Chan,; A, S. H. Lee.; and Z, Zainol. (2021). Predicting employee health risks using classification ensemble model." *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pp. 52-58, doi: 10.1109/CAMP51653.2021.9498106.
14. Lim, T.M.; and Lee, A.S.H. (2017). Loyalty Card membership challenge: A study on membership churn and their spending behaviour. A*rchives of Business Research*, 5(6), 66-88.
15. Quinlan, J.R. (1993). *C4.5: Programs for machine learning* (1st ed.). Morgan Kaufmann Publishers.
16. Kuhnert, P.M.; Do, K.; and McClure, R. (2000). Combining non-parametric models with logistic regression: an application to motor vehicle injury data. *Computational Statistics & Data Analysis*, 34(3), 371-386.
17. Loh, W. (2011). Classification and regression trees. *WIREs Data Mining Knowledge Discovery*, 1, 14-23.
18. Zajac, S.S.; and Ivan, J.N. (2003). Factors influencing injury severity of motor vehicle - crossing pedestrian crashes in rural Connecticut. *Accident Analysis and Prevention*, 35(3), 369-379.
19. Renski, H.; Khattak, A.J.; and Council, F.M. (1999). Effect of speed limit increases on crash injury severity: Analysis of single-vehicle crashes on North Carolina Interstate highways. *Transportation Research Record*, 1665(1), 100-108.
20. Xhemali, D.; Hinde, C.J.; and Stone, R.G. (2009). Naïve bayes vs. decision trees vs. neural networks in the classification of training web pages. *International Journal of Computer Science Issues*, 4(1), 16-23.