sustainability

MDPI

Article

Classification of Driver Injury Severity for Accidents Involving Heavy Vehicles with Decision Tree and Random Forest

Aziemah Azhar [1], Noratiqah Mohd Ariff [2,*], Mohd Aftar Abu Bakar [2] and Azzuhana Roslan [3]

[1] Vehicle Safety and Biomechanics Research Centre (VSB), Malaysian Institute of Road Safety Research (MIROS), Kajang 43000, Selangor, Malaysia; aziemah@miros.gov.my
[2] Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, UKM Bangi, Bandar Baru Bangi 43600, Selangor, Malaysia; aftar@ukm.edu.my
[3] Crash Data Operational & Management Unit (CRADOM), Malaysian Institute of Road Safety Research (MIROS), Kajang 43000, Selangor, Malaysia; azzuhana@miros.gov.my
* Correspondence: tqah@ukm.edu.my

Abstract: Accidents involving heavy vehicles are of significant concern as it poses a higher risk of fatality to both heavy vehicle drivers and other road users. This study is carried out based on the heavy vehicle crash data of 2014, extracted from the MIROS Road Accident and Analysis and Database System (M-ROADS). The main objective of this study is to identify significant variables associated with categories of injury severity as well as classify and predict heavy vehicle drivers' injury severity in Malaysia using the classification and regression tree (CART) and random forest (RF) methods. Both CART and RF found that types of collision, driver errors, number of vehicles involved, driver's age, lighting condition and types of heavy vehicle are significant factors in predicting the severity of heavy vehicle drivers' injuries. Both models are comparable, but the RF classifier achieved slightly better accuracy. This study implies that the variables associated with categories of injury severity can be referred by road safety practitioners to plan for the best measures needed in reducing road fatalities, especially among heavy vehicle drivers.

Keywords: classification and regression tree; driver injury severity; heavy vehicles accident; machine learning; random forest

Citation: Azhar, A.; Ariff, N.M.; Bakar, M.A.A.; Roslan, A. Classification of Driver Injury Severity for Accidents Involving Heavy Vehicles with Decision Tree and Random Forest. *Sustainability* **2022**, *14*, 4101. https://doi.org/10.3390/su14074101

Academic Editors: Juneyoung Park, Yina Wu and Hochul Park

Received: 28 February 2022
Accepted: 25 March 2022
Published: 30 March 2022

1. Introduction

Increasing demand for road transportation solutions to deliver goods or for personal travel has been well-recognized as one of the impacts of economic development. However, along with the growth of the economy, the demand for motor vehicles especially heavy vehicles has raised another significant concern: road accidents. The risk of injury and fatality caused by heavy vehicle accidents is worrying as it may also involve other road users. Based on the Statistical Report of Road Accident in Malaysia 2018, there were 44,243 total buses and lorries reported to be involved in road crashes resulting in 192 casualties among bus and lorry drivers or passengers [1].

Many studies had discussed and explained the injury severity among heavy vehicle drivers by either human, vehicle or environmental settings. This research either focused on specific types of heavy vehicles such as trucks [2–14] and buses [15,16] or considered all types of heavy vehicles in the data [17–20].

Since heavy vehicles might affect other road users in a crash, some researchers scoped their study to crashes between heavy vehicles and other specific road users. For instance, Samerei et al. [15] explored factors affecting the injury severity of pedestrians who were involved in a collision with buses. On the other hand, Zhou et al. [14] focused their study on factors affecting passenger car and truck driver injury severity. Pahukula et al. [4] explored factors related to large trucks involved in crashes by time of day where they proved that different time settings had a different set of crash determinants. Likewise, Uddin, and

Huynh [13] also demonstrated a significant difference between factors affecting injury severity for several lighting conditions. Meanwhile, Chen et al. [8] discovered that road grade, driver age and seatbelt usage are among the factors affecting truck drivers' injury severity for crashes that occurred in rural areas.

In terms of methodology, analysis regarding injury severity in road traffic crashes had been conducted extensively by using a statistical method such as different types of regression models [21–24]. However, over the years, technology advancement specifically in artificial intelligence has led to numerous studies being conducted by using machine learning. Predictions can be made by calibrating a machine learning model, and various models have been studied and applied to injury severity-related data. Chen et al. [25] used the classification and regression tree (CART) model to identify the significant variables related to driver injury severity for rollover crashes. These variables were then used as inputs in the support vector machine (SVM) model. They pointed out that seatbelt usage is the factor associated most with injury severity. In determining the causes of fatal accidents in urban areas among inexperienced drivers, Moral-García et al. [26] developed a decision tree and found that most of the fatal injury cases were due to collisions with pedestrians or speeding. Predicting the injury severity of a specific road user was conducted by Wahab and Jiang [27,28] who developed a CART model to predict the injury severity among motorcycle crashes. They found that CART outperformed rule induction and multi-layer perceptron models based on a 10-fold cross-validation approach. Meanwhile, Haynes et al. [29] studied the factors of traffic accidents in the UK using various machine learning methods and found out that random forest (RF) provides the highest accuracy score in classifying accident severity. As for studies concerning heavy vehicles, Behnood and Mannering [9] used logit models to observe variation and temporal instability for factors affecting injury severity. Chang and Chien [6] developed a CART model for predicting injury severity among truck drivers and found that CART models provide useful insights and are good alternatives for analysing the injury severity of truck accidents.

Different road traffic cultures have different effects on road safety performance [30–32]. Since Malaysia has its own road traffic culture, different variables may be the main factors for causing injury among heavy vehicle drivers when involved in accidents. To the extent of the authors' knowledge, research on identifying variables associated with different levels of injury severity specifically among heavy vehicle drivers in the Malaysian context using the machine learning method is limited.

Traffic injury severity ranges from no injury at all to slightly or severely injured and lastly to fatally injured which makes it appropriate to be treated as an ordinal outcome. Taking this into consideration, this study explores ordinal logistic regression, and assumptions underlying this model were tested beforehand. A test of parallel lines was conducted to validate the assumptions of proportional odds where no difference in the coefficients between models was the null hypothesis. To proceed with this regression, a non-significant result needs to be achieved. However, the result of this test was significant with a p-value of <0.001 which indicates that the assumption of proportional odds was violated thus this analysis could not proceed and machine learning, as a data-driven method, is chosen to overcome this constraint. Machine learning does not generally require such assumptions hence they are advantageous in traffic injury severity analyses.

Hence, the main objective of this study is to (1) identify significant variables associated with categories of injury severity as well as (2) classify and predict heavy vehicle driver's injury severity in Malaysia using machine learning techniques: CART and RF. Variable importance measures were used to identify the most important predictors in classifying the injury severity. This ranking of variables can be a potential area for further specific research in Malaysia.

The remainder of the paper is organised as follows. In the next section, there is a brief introduction to Malaysia's road and traffic volume as well as the heavy vehicle crash dataset that has been used. The theoretical background of CART and RF is then provided. Following that, model results are discussed with the variable importance and model performance. In the final section, key findings are summarised.

2. Heavy Vehicle Crash Data

With reference to the Malaysia Road Transport Act 1987, motor vehicles can be divided into several classes based on their usage and load capacity. Table 1 shows the description of vehicles extracted from this Act and heavy vehicles may be classified into one of these classes. This study is carried out based on the heavy vehicle crash data of 2014, extracted from the MIROS Road Accident and Analysis and Database System (M-ROADS). M-ROADS is a database system developed by the Malaysian Institute of Road Safety Research (MIROS) storing the road crash data obtained from the Royal Malaysian Police (RMP).

Table 1. Classification of vehicles based on Malaysia Road Transport Act 1987.

Class	Description
Motor cars	Motor vehicles which are constructed to carry a load or passengers and the unladen weight of which does not exceed three thousand kilograms.
Motor cars heavy	Motor vehicles which are constructed to carry a load or passengers and the unladen weight of which exceeds three thousand kilograms.
Tractors heavy	Motor vehicles not constructed to carry any load (other than water, fuel, accumulators and other equipment and materials used for the purposes of propulsion, loose tools and loose equipment), the unladen weight of which exceeds five thousand kilograms.
Tractors light	The unladen weight of which does not exceed five thousand kilograms and which otherwise fall within the definition of "tractors heavy".
Mobile machinery heavy	Motor vehicles which are designed as self-contained machines, propelled by means of mechanism contained within themselves and the unladen weight of which exceeds five thousand kilograms and are capable of being used on roads.
Mobile machinery light	Motor vehicles, the unladen weight of which does not exceed five thousand kilograms and which otherwise fall within the definition of "mobile machinery heavy".
Trailers	Vehicles other than land implements drawn by a motor vehicle, whether or not part of the trailer is superimposed on the drawing vehicle.

In 2014, there were 476,196 total road crashes with 6674 fatalities. Out of this total number of crashes, 46,674 involved heavy vehicles where 250 drivers and occupants had a fatal injury. However, for this study's purposes, only driver's information regardless of the occupants were considered. Nevertheless, all types of buses and lorries are considered heavy vehicles in this study. M-ROADS comprised of variables related to crash details, road characteristics, environment, location, vehicle details and driver's information. Table 2 shows the complete variables used in this study.

Table 2. Input variables.

Attribute	Value
Number of vehicles involved	Single vehicle Two-vehicle Multiple vehicles
Type of first collision	Head-on Rear Right angle side Angular Sideswipe Forced Hitting objects on the road Hitting objects off road Hitting pedestrian Overturned Out of control

Table 2. *Cont.*

Attribute	Value
Road surface	Gravel Bricks Bitumen/tar Concrete Dirt
Road type	Straight road Road bends Roundabout Cross junctions Y/T junctions Staggered junctions Interchange
Quality of road surface	Flat Sink Hole Wavy
Vertical design of the road	Flat Slope
Road surface condition	Dry Wet Oily Sandy Under repair
Weather conditions	Clear Windy Foggy Rainy
Light conditions	Daylight Dawn/dusk Dark with lights Dark without lights
Road category	Expressway Federal roads State roads Municipal roads Others
Location	City Urban Built-up area Rural
Surrounding area	Residential Office Shopping area Industrial/construction Bridge/ foot bridge School Others
Heavy vehicle type	Express bus Stage bus Factory bus Mini bus Excursion bus School bus Articulated lorry Rigid lorry Pick-up lorry
Vehicle modification	Yes No
Driver's gender	Male Female

Table 2. *Cont.*

Attribute	Value
Driver's age	21–25 26–30 31–35 36–40 41–45 46–50 51–55 56–60 61–65 66–70 71–75
Driver's error	Careless at entrance/exit Overloading (passengers) Improper parking Drugs Careless driving Dangerous driving Dangerous turning Dangerous overtaking Tailgating Speeding Disobeying traffic light Other offences Not at fault
Drunk driving	Not suspected Positive Negative

RMP categorised crashes into four categories: crashes with fatal injury, severe injury, light injury and crashes that involved only damaged properties (no injury). This four-level injury was used as the response variable in this study. Descriptions [33] for each level of injury are elaborated in Table 3.

Table 3. Description for each level of injury severity.

Type of Injury	Description
Fatal	Any person who died within 30 days as result of an accident.
Severe injury	Any person who has injured as a result of an accident as referred to section 320 of the Penal Code which includes any of the following: • Emasculation • Permanent privation of the sight of either eye • Permanent privation of the hearing of either ear • Privation of any member or joint • Destruction or permanent impairing of the powers of any member or joint • Permanent disfiguration of the head or face • Fracture or dislocation of a bone • Any hurt which endangers life, or which causes the sufferer to be, during the space of ten days, in severe bodily pain, or unable to follow his ordinary pursuits
Slight injury	Any injury that does not fall under Fatal or Severely Injured category.
No injury	No injury.

Note: Based on Malaysia Statistical Report Road Accident 2014.

The frequency of each injury severity is presented in Table 4. Most heavy vehicle drivers had no injury (79.8%) while other drivers sustained either slight (9.8%), severe (4.3%) or fatal injury (4.3%).

Table 4. Frequency of each level of injury severity.

Level of Severity	Frequency (%)
Fatal	49 (5.3)
Severe injury	49 (5.3)
Slight injury	90 (9.7)
No injury	742 (79.8)
Total	930 (100)

3. Methodology

3.1. Decision Tree

The decision tree is a classification technique that recursively partitioning the instant space. Generally, the decision tree consists of nodes, edges, and leaves. The root node contains all the sample sets, which will split according to specific attributes into two or more subsets, known as the internal nodes. The internal nodes will continue splitting until it reaches the end of the tree, which is known as the leaf. The leaf will be assigned to one class representing the most common target value or category for that variable of interest.

There are many types of decision tree algorithms such as the CART [34], ID3 [35] and C4.5 [36]. In this study, the CART algorithm has been used to build the decision tree. The CART tree was constructed by splitting a node into two nodes repeatedly. Let Y be the categorical dependent variable with C classes, and X_i is the ith predictor variables which can be either ordinal, nominal or continuous. The decision tree was grown into a large tree until all the leaves were pure. Thus, all observations in a particular leaf were in the same category or class for that dependent variable. At each node, only one feature or predictor variable was considered as the condition for splitting.

The following steps were used on each node recursively:

For ordinal or continuous predictor variables, find the best split point which maximises the splitting criterion. Meanwhile, for the nominal predictor variable, find the best split based on the subset of categories for that variable which maximises the criterion.

From step 1, select the best predictor variable, X_i, which leads to the most significant reduction in node impurity. There are two standard measures of node impurity which are the Gini index, defined as

$$G = \sum_{k=1}^{C} \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad (1)$$

and the entropy, given by

$$D = \sum_{k=1}^{C} -\hat{p}_{mk} \log_2 \hat{p}_{mk} \quad (2)$$

where $\hat{p}_{mk}$ is the proportion of observation in the mth data subset that is from the kth class.

Both measures will have a small value if the node contains predominantly observations from the same class, which indicates homogeneity, while a higher value indicates a higher amount of disorder. In this study, the Gini index will be used to calculate the node impurity. The algorithm will only stop if the stopping criterion is satisfied, for example, until a node becomes pure, or until each leaf only contains less than the specified number of observations. The sum of the improvements (reductions in impurity) in the overall Gini index for every node in which a variable occurs is an important measure that will be used to identify the most important variables [37]. A flowchart of the CART algorithm is illustrated in Figure 1.

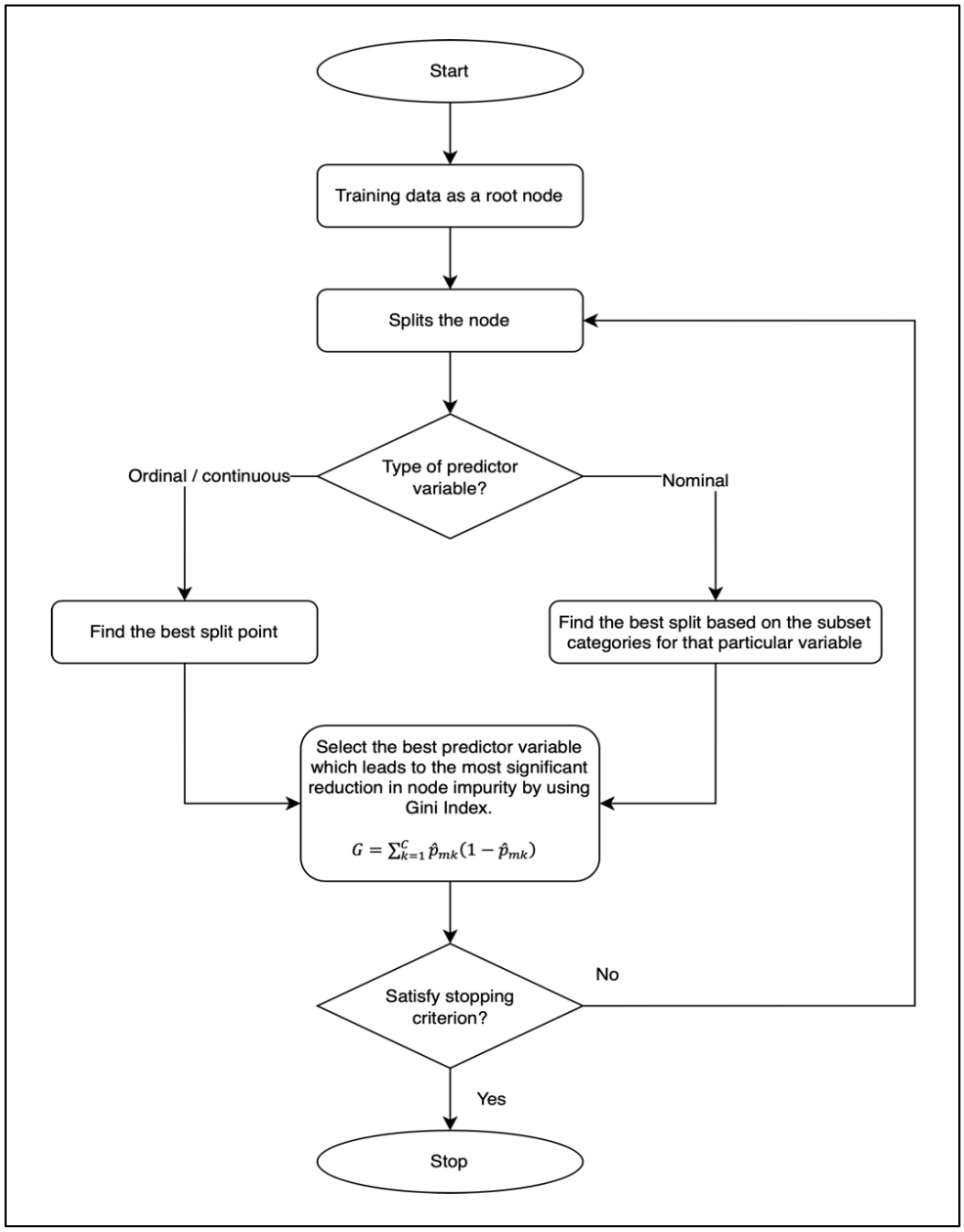

Figure 1. Flowchart of CART.

Overfitting may occur when a tree is grown very deep as it tends to learn extremely irregular patterns. The training sets will be overfitted by having a low bias but with a very high variance. For reducing the variance, RF average multiple deep decision trees which are trained on different parts of the same training set [38]. Still, the decision tree is commonly used as it is comprehensible and a more detailed explanation can be made regarding predicting variables.

3.2. Random Forest (RF)

RF is an ensemble of decision trees, which involves the bootstrap aggregation technique, also known as bagging. Bagging is a procedure to reduce the variance of the supervised learning model which is usually suffered by the decision tree model. By taking B repeated samples from the same data set (bootstrap), there will be B decision trees that can be used for predictions. Since the response variable is categorical, the overall prediction will be the most commonly occurring category among the B predictions (aggregation).

However, if there is a crucial variable and several moderately important variables which heavily influence the model, most of the bagging trees will be almost similar, resulting in a highly correlated prediction. To decorrelate the trees, random forest only considered several feature (dependent) variables which are chosen randomly as split candidates each time while splitting the tree. This approach, called feature randomness, ensures other less important variables have a better chance of being considered resulting in decorrelation of the trees. The majority vote of all trees is reported as RF prediction. The clear flow of RF can be seen in Figure 2.

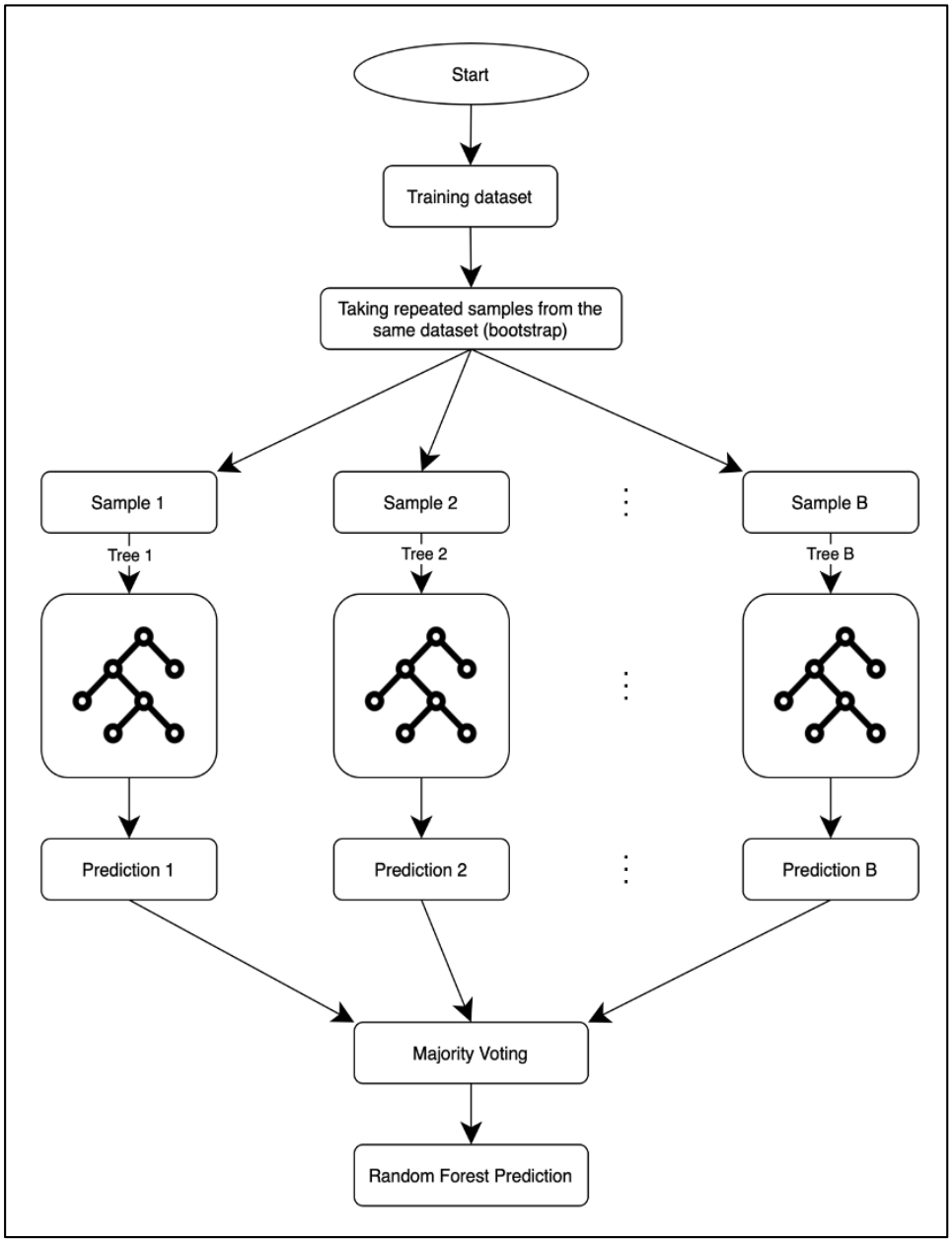

Figure 2. Flowchart of RF.

RF can also be used to rank the variable according to its importance. The calculation is similar to CART where node impurity decrease across every tree is added each time that variable is chosen to split a node. Then, the average is calculated by dividing the sum by the number of trees which can then be ranked. Apart from variable importance, Shapley additive explanations (SHAP) for RF was also used to visualize the impact of each feature on the response variable.

3.3. Model Evaluation

The confusion matrix and its associated performance metrics: classification accuracy, true-positive rate (TPR), false-positive rate (FPR), precision, recall, and F-measure, are the parameters used to evaluate the accuracies of the classifiers used in this study. Table 5 demonstrates the confusion matrix where the row and column indicate the observed and predicted class, respectively, while values in the diagonal represent correctly predicted instances. For multiclass classification, when a specific class is recognised as positive, the remaining class will be identified as negatives.

Table 5. Confusion matrix.

	Predicted Class	
Observed Class	**Yes**	**No**
Yes	True positive (TP)	False negative (FN)
No	False positive (FP)	True negative (TN)

TP and TN are the numbers of actual positives and actual negatives that were identified correctly. FP, on the other hand, is the number of negatives that were misclassified as positives, and conversely, TN is the number of positives that were misclassified as negatives. These values were then be used to calculate the performance measurement. TP rate (TPR) is the fraction of positives that are correctly identified while FP rate (FPR) is the fraction of negative that are incorrectly classified. Precision is a measure of correctly classified instances in all instances that are predicted as positives and recall has the same calculation as TPR. A low precision indicates many FP and low recall means several FN. The formula is expressed as below:

$$\text{TPR} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{F} - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

The area under the ROC curve (AUC) is a useful tool for evaluating the quality of class separation for classifiers. Nevertheless, in the multiclass setting, the performance of multiclass models can be visualised according to their one vs. all precision–recall curves (PRC). Since there is imbalanced distribution in the predictor class, PRC is preferable as it allows for an accurate and intuitive interpretation of classifier performance [39,40]. The performance of the minority class will also be achievable by constructing PRC [41]. Therefore, PRC is employed in this study to handle the imbalanced data of injury severity.

Figure 3 presents the methodology overview for this study. R software version 3.6.3 was used to build and evaluate the model. Table 6 describes the packages used for this study's purposes.

Table 6. R packages used to build and evaluate the model.

Packages	**Description**	**Reference**
rpart	Building CART	[42]
caret	Data splitting/confusion matrix table	[43]
E1071	Confusion matrix table	[44]
randomForest	Building random forest	[45]
ROCR	Precision–recall graph	[46]

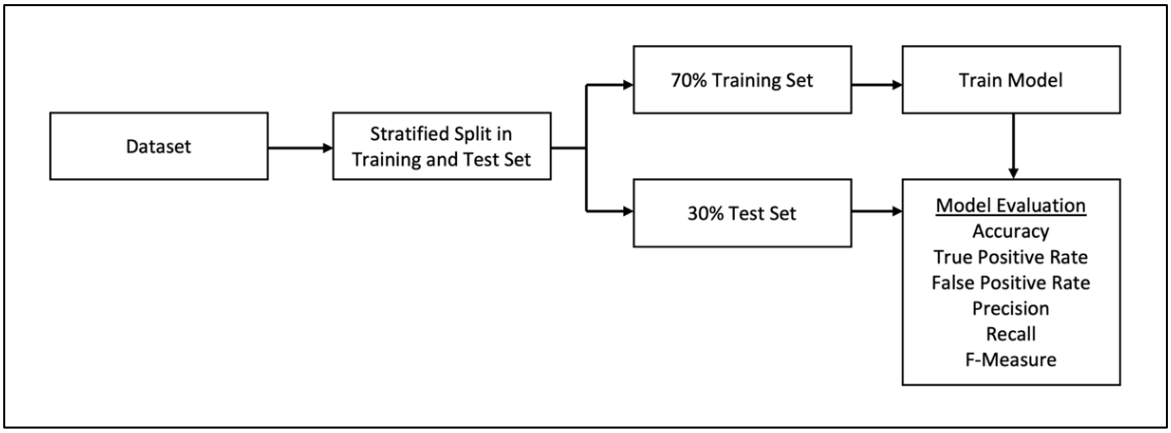

Figure 3. Methodology overview.

4. Results

4.1. CART

The dataset was partitioned into 70% of training data and 30% of testing data. There were 18 predictor variables that were used together with drivers' level of injury to identify the vital patterns and information regarding heavy vehicle crashes. The decision tree was built by using the Gini index splitting criterion, and Figure 4 represents the classification tree. There are 8 terminal nodes, and it can be seen that the type of the first collision, driver's error, number of vehicles involved, road surface type, driver's age, source of light and vehicle type are the main variables used to split the tree. This also indicates that these variables are important in classifying injury severity among heavy vehicle drivers who were involved in a road crash.

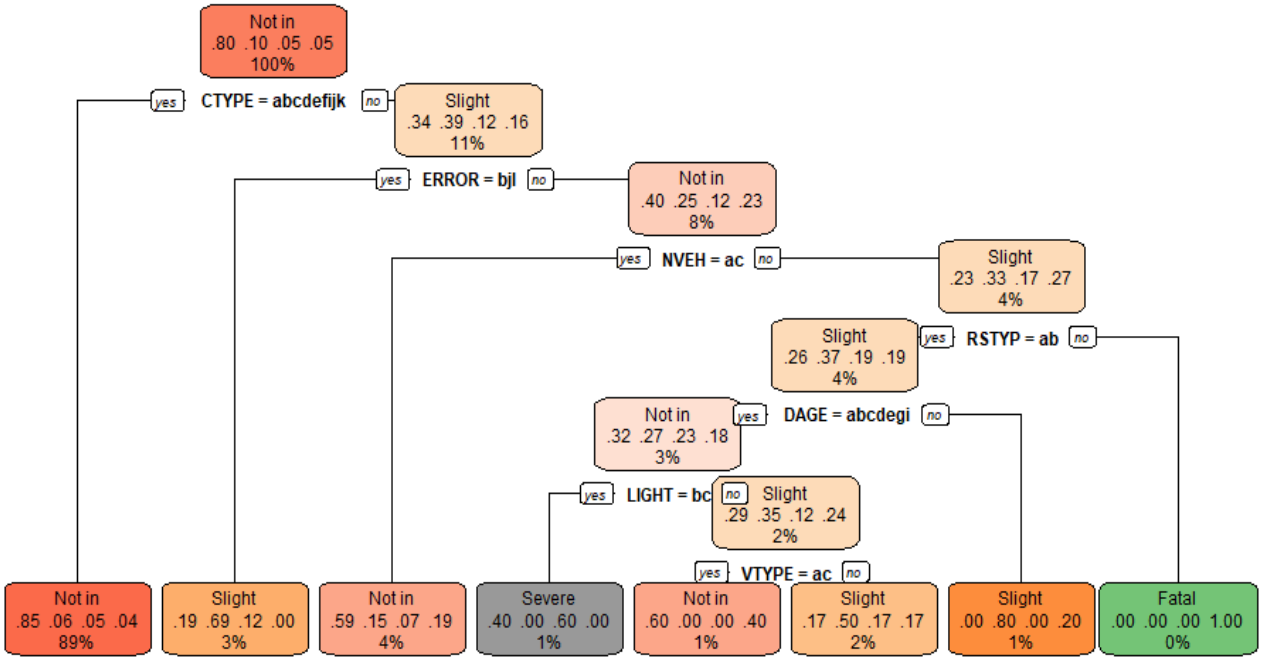

Figure 4. The output of decision tree. CTYPE: angular (a), forced (b), head-on (c), hitting objects off road (d), hitting objects on road (e), hitting pedestrian (f), rear (i), right angle side (j), sideswipe (k) ERROR: careless driving (b), other offences (j), speeding (l) NVEH: multiple vehicles (a), two-vehicle (c) RSTYP: bitumen/tar (a), bricks (b) DAGE: 21–25 y/o (a), 26–30 y/o (b), 31–35 y/o (c), 36–40 y/o (d), 41–45 y/o (e), 51–55 y/o (g), 61–65 y/o (i) LIGHT: dark without lights (b), dawn/dusk (c) VTYPE: articulated lorry (a), express bus (c).

For every node in the decision tree, three lines of information are displayed. The first line indicates the level of injury with the highest number of observations, i.e., not injured, slightly injured, severely injured, or fatal. The middle line consists of four values. Each value corresponds to the probability of each level of injury that can be calculated by dividing the number of observations for that injury level by the total observations in that node. For example, the first node in the decision tree shows a middle line displaying (0.80 0.10 0.05 0.05). This implies that the likelihood of heavy vehicle drivers who are involved in a crash to survive with "Not Injured" or being "Slightly", "Severely" and "Fatally injured" is 80%, 10%, 5% and 5%, respectively. Meanwhile, the percentage in the last line of every node shows the percentage of observations from the whole data set that resides in the node.

The first split is based on the type of the first collision, which implies that this variable is the most crucial in classifying injury due to crashes among heavy vehicle drivers. This decision tree groups nine types of collisions described as angular, forced, head-on, hitting objects off-road and on-road, hitting pedestrian, rear, right angle side and sideswipe into a terminal node which contains 89% of the total observations. As shown by the terminal node on the leftmost branch of the tree, if these collision types occurred during a crash, the tree shows that the drivers will likely experience no injury (85.4%). The other two collision types, which are out of control and overturned, form another node with the remaining 11% of observations. Based on this node, the probability of having injuries is higher if the crash involved out-of-control and overturned vehicles (39% slightly injured, 12% severely injured, and 16% fatally injured). This node is partitioned further based on the driver's error variable.

By considering the driver's error, another terminal node has been formed with 3.4% of the total observations. Based on this terminal node, 38.6% and 11.5% of the heavy vehicle drivers who are involved in a crash most likely will suffer from a slight and severe injury respectively if they committed careless driving, other unspecified offences, or speeding. Other types of error will mostly cause no injury to the drivers (40.4%), and respective observations are partitioned again by referring to the number of vehicles involved.

The partitioning concerning the number of vehicles involved in a crash had resulted in the formation of another terminal node which comprised 3.6% of the total observations. This terminal node suggests that in a heavy vehicle crash involving two or more vehicles, the driver had a higher chance of surviving with no injury (59.3%). Meanwhile, drivers who were involved in a single-vehicle crash had a greater chance of being injured (33% slightly injured, 17% severely injured, 27% fatal) compared to two-vehicle and multiple vehicle crashes.

Crashes involving only one vehicle are grouped and split further based on the road surface type. The tree shows that 37.0% of the heavy vehicle drivers who are involved in crashes along roads with bitumen/tar and bricks surface will be slightly injured. This group of drivers is partitioned further by taking the driver's age into account. Nevertheless, crashes along other road surface types (concrete, gravel, dirt) form another terminal node at the rightmost branch of the tree. Even though less than 1% of total observations are in this node, it is crucial to note that 100% of the drivers who are involved in crashes along the road with concrete, gravel or dirt surface will be fatally injured.

Back to the centre part of the tree, the driver's age variable has been used to partition the observations further where drivers of 21 to 45, 51 to 55 and 61 to 65 years old are grouped. This group of drivers is then partitioned again based on the light condition during a crash. On the other hand, age groups of 46 to 50, 56 to 60 and 66 to 75 years old are grouped and form a terminal node. This terminal node illustrates that 80% of the heavy vehicle drivers of this age will probably have a slight injury while another 20% will be fatally injured when involved in a crash.

Regarding the variable light condition during crashes, a terminal node is formed at the centre part of the tree where crashes that occurred on a dark road without lights or during dawn/dusk are grouped. Undoubtedly, heavy vehicle drivers who are involved in crashes with these environments had a higher chance to be severely injured (60%). In comparison,

the remaining 40% will probably survive without any injury. Dark environments may reduce visibility and increase the risk of a crash. Another partitioning based on vehicle type is conducted to the group of drivers who are involved in crashes on a dark road with lights or during daylight and resulted in the formation of two terminal nodes. The first terminal node indicates that the majority of articulated lorry and express bus drivers who were involved in a crash had a greater chance of surviving with no injury (60%). However, it is also important to point out that another 40% of these drivers will most likely be fatally injured. The second terminal node is formed by grouping other types of heavy vehicles and half of the observations in this node had a chance to be slightly injured.

Figure 5 explains the importance of each variable for data classification by the CART model. Collision type was the main factor used by this model in classifying the data. Apart from that, the top ten most important variables were the driver's error, age, number of vehicles involved, road surface type, vehicle type and lighting condition.

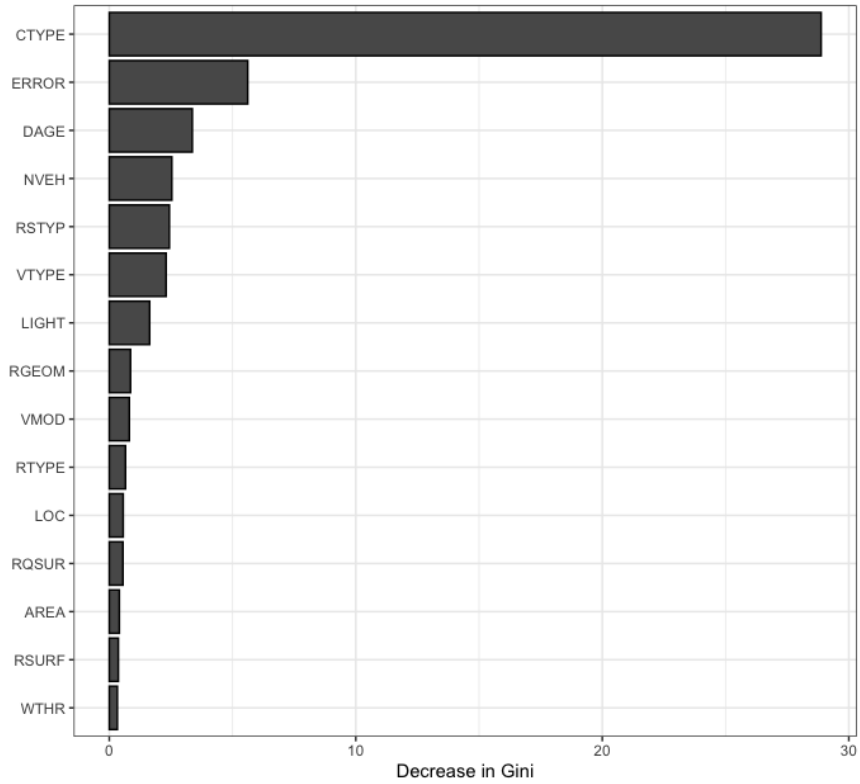

Figure 5. Variable importance plot for CART.

4.2. Random Forest (RF)

RF for the training data was built with 500 different trees and four predictive variables considered at each split within a tree. A different subset of the data will be involved in building these trees. For each iteration, out-of-bag error will be used to assess the model, where the tree was tested with testing data set which is not involved in the tree building process. The out-of-bag error, also known as the average of errors of all these interactions is 19.3%.

The variable importance plot in Figure 6 explains the importance of each variable for data classification by RF model. The rank of predictor variables is based on the mean decrease Gini, also known as the mean decrease accuracy, which measures each variable contribution to the purity on each node in a tree. As the decision tree, collision type was the main factor used by the model in classifying the data. Nevertheless, driver's age, vehicle type, road type, number of vehicles involved, driver's error and lighting condition were also being identified as the other important variables by this model. Regarding the top ten important variables for both classifiers, CART and RF are comparable in determining factors that contribute to driver injuries in accidents involving heavy vehicles.

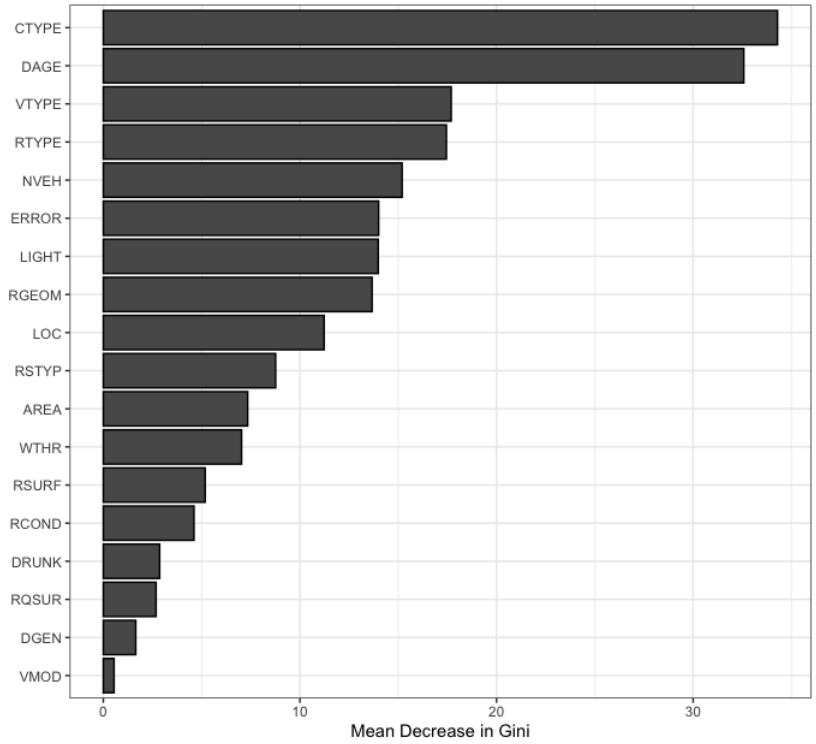

Figure 6. Variable importance plot for random forest.

To explain more on the impact of each explanatory variable towards the level of injury severity, SHAP for RF was carried out and the visualisation is presented in Figure 7. This summary plot combines the feature importance with feature effects. Based on the *y*-axis, features on the top are the most important features to the model output while the *x*-axis represents the SHAP value for each feature. The red colour represents the higher value of a feature while the blue represents the lower value of a feature. Collision type was the most important feature where the crashes related to a higher value of this feature (sideswipe) are less likely to suffer in a fatal injury. Both ends of factor driver's age are also worth mentioning where older drivers are more likely to suffer fatal injury compared to the young drivers.

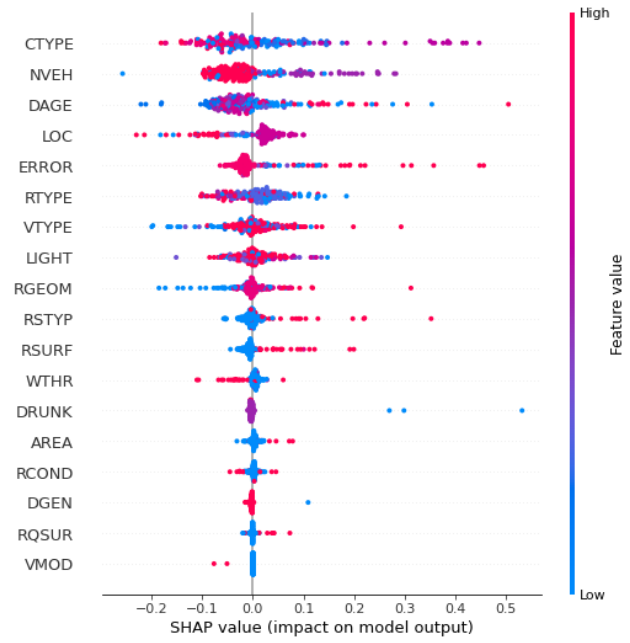

Figure 7. SHAP for RF.

4.3. Model Performance

The performance for each model was assessed and compared in determining the best model for injury severity prediction. Table 7 displays the confusion matrixes between the fitted result and the actual observation for both CART and RF classifiers. The diagonal of these matrixes represents the number of correctly classified instances for each class of injury severity while the off-diagonal entries represent the misclassified instances. The number of heavy vehicle drivers who suffered from injury (slight injury, severe injury and fatal) but were misclassified as not injured has become of great interest. As shown in the table, CART had misclassified 12 drivers with a slight injury, seven drivers with severe injury and four drivers with fatal injury as not injured. On the other hand, the RF classifier seems to classify instances approximately similar to CART. Table 8 presents the overall estimation results for the test datasets. CART classifier achieved an accuracy of 81.5% while RF classifier achieved a slightly better accuracy with 82.1% correctly classified instances.

Table 7. Confusion matrix for the test dataset.

		Predicted Class			
Classifier	**Observed Class**	**Not Injured**	**Slightly Injured**	**Severely Injured**	**Fatal**
CART	Not injured	143	5	0	0
	Slightly injured	12	5	1	0
	Severely injured	7	1	1	0
	Fatal	4	2	2	1
RF	Not injured	145	3	0	0
	Slightly injured	13	4	0	1
	Severely injured	7	1	1	0
	Fatal	5	3	0	1

Table 8. Performance of the classifier.

Classifier		***n* (%)**
CART	Correctly classified instances	150 (81.5)
	Incorrectly classified instances	34 (18.5)
	Total	184 (100.0)
RF	Correctly classified instances	151 (82.1)
	Incorrectly classified instances	33 (17.9)
	Total	184 (100.0)

Table 9 summarises the estimation accuracies for each driver injury severity by using the test dataset. TPR for CART classifier are ranging from 0.111 for "Severely Injured" and "Fatal" to 0.966 for "Not Injured" with a weighted average of 0.815. These results demonstrate that CART can classify 96.6% of instances with no injury correctly. However, a low TPR rate for all levels of injury and fatal class must be paid attention to since predicting injuries and fatalities are the main interest for this study. RF had a similar distribution of TPR across the different classes of injury severity which indicates that both classifiers had a better performance in predicting no injury than other types of severity since no injury class dominated the training dataset.

By considering both recall and precision of the test dataset, F-measure is computed where a value approaching 1 reflects perfect precision and recall. In contrast, a value near 0 means low precision and recall. Concerning each class of injury, "Not injured" had the highest calculated F-measure for both classifiers while other classes of injury had a very low F-measure. Recall and precision values of the test dataset were also used to calculate the area under the curve for each classifier by plotting the precision–recall curves (PRC) as illustrated in Figures 8 and 9. The plots indicate that no injury cases can be predicted very well, while other levels of injuries are harder to classify for both CART and RF.

Table 9. Estimation accuracies for the test dataset of the heavy vehicle accident injury severity.

Classifier	Class	TPR/Recall	FPR	Precision	F-Measure
CART	Not injured	0.966	0.639	0.861	0.910
	Slightly injured	0.278	0.048	0.385	0.323
	Severely injured	0.111	0.017	0.250	0.154
	Fatal	0.111	0.000	1.000	0.200
	Weighted	0.815	0.520	0.791	0.781
RF	Not injured	0.980	0.694	0.853	0.912
	Slightly injured	0.222	0.042	0.364	0.276
	Severely injured	0.111	0.000	1.000	0.200
	Fatal	0.111	0.006	0.500	0.182
	Weighted	0.821	0.563	0.795	0.779

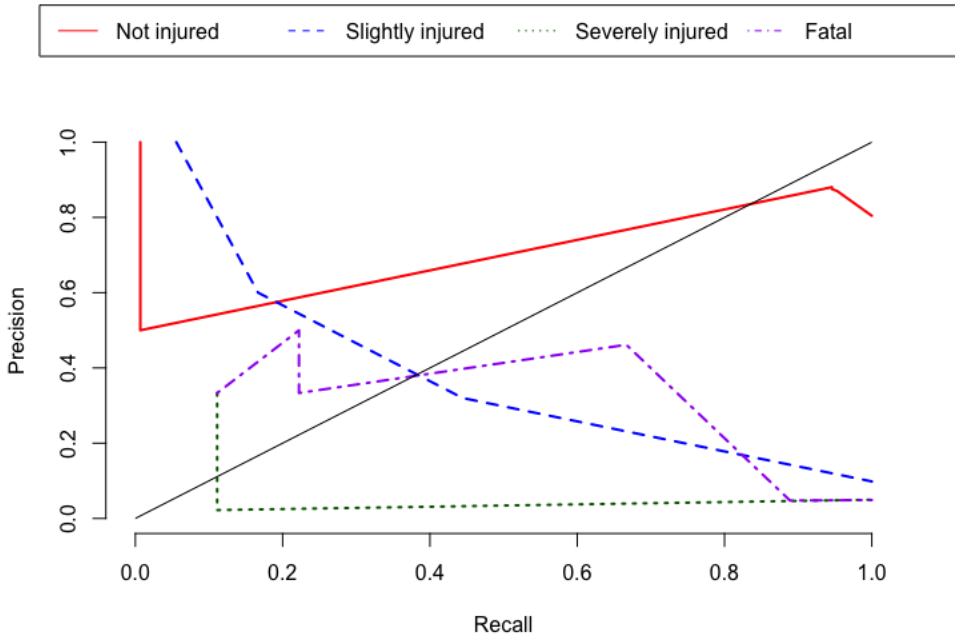

Figure 8. PRC for the classification of heavy vehicle accident injury severity using CART.

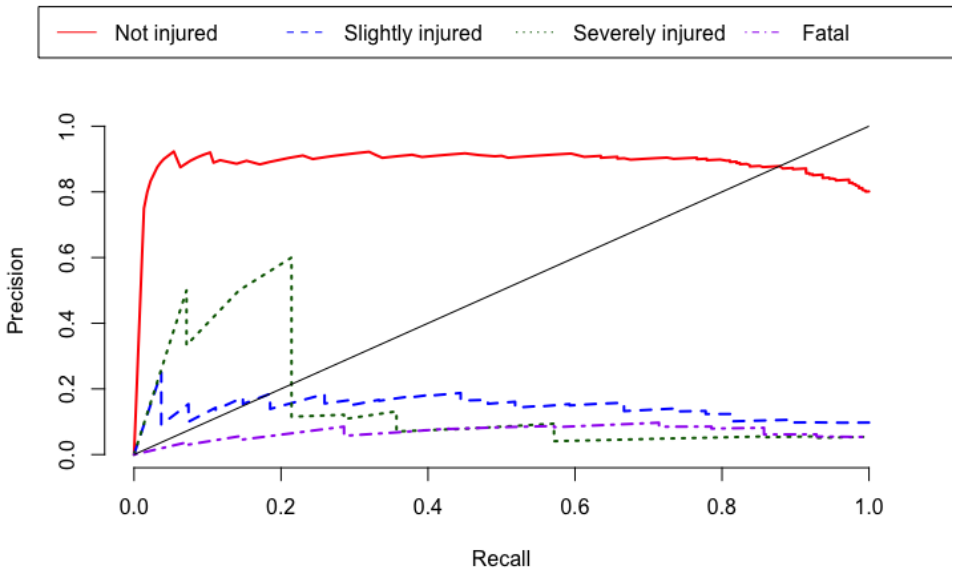

Figure 9. PRC for the classification of heavy vehicle accident injury severity using RF.

5. Discussion

This study constructed CART and RF to classify and predict traffic injury severity among heavy vehicle drivers in Malaysia. Type of first collision, driver's error, number of vehicles involved, road surface type, driver's age, a light condition during crash and vehicle type were found important in classifying injury severity among these drivers.

Based on the first split of CART, collision type is the most crucial variable in classifying injury, similar to the findings in previous studies [6,13,26,47]. The probability of having injuries is higher if the crash involved out-of-control and overturned vehicles. Specifically, overturned had been discussed in previous studies as one of the crash characteristics that may increase the risk of having injuries [7,48]. The driver of overturned heavy vehicles during road crashes may suffer from a more severe injury due to the vehicle size itself.

Other than that, heavy vehicle drivers who are involved in a crash most likely will suffer from a slight and severe injury if they committed careless driving, other unspecified offences, or speeding. Undeniably, driving with a higher speed or more than the posted speed limit has been proved to increase the chances of having a more severe injury which is consistent with findings in [7,20].

It is also crucial to note that 100% of the drivers involved in crashes along the road with concrete, gravel or dirt surface will be fatally injured. Adanu et al. [49] labelled dirt, sand and gravel as a group of 'unpaved roads' and pointed out a contradictory finding where crashes on these roads are not likely to be fatal.

Another important finding of this study is based on the driver's age. Older drivers tend to be fatally injured when involved in a crash. On the contrary, several studies [4,50] found out that the younger heavy vehicle drivers are more prone to injuries due to their little experience and reckless behaviour on the road. This difference might be due to a variety of ways in categorising drivers' age, and RMP's definition of age categories is being referred for this study. In addition, deteriorating health status among older drivers may influence injury severity.

Lighting condition was also one of the contributing factors. Undoubtedly, heavy vehicle drivers who are involved in crashes on a dark road without lights or during dawn/dusk had a higher chance of being severely injured. This is in line with findings by Zheng et al. [5] and Zhu et al. [51]. Dark environments may reduce visibility and increase the risk of a crash.

The performance for each model was assessed and compared in determining the best model for injury severity prediction. RF achieved better accuracy than CART in this study similar to what had been proven by Wahab and Jiang [27]. However, a low TPR rate for all levels of injury and fatal class must be paid attention to since predicting injuries and fatalities are the main interest for this study. CART and RF had a similar distribution of TPR across the different classes of injury severity. This indicates that both classifiers had a better performance in predicting drivers with no injury since they dominated the training dataset. F-measure and PRC also showed the same findings which reflect that injured or fatal drivers were difficult to be classified.

Findings discussed earlier do have some implications for the safety of heavy vehicle drivers. The variables associated with categories of injury severity can be referred by road safety practitioners or employers to plan for the best measures needed in reducing road fatalities, especially among these groups of drivers. For example, employers may develop or organize training courses that are related to the variables found such as training that focus on defensive driving, emergency response and vehicle operational. Employers may also record feedback from the drivers regarding risky road conditions to ensure drivers' safety. In 2010, the Malaysia Department of Occupational Safety and Health (DOSH) has formulated an Occupational Safety and Health Industry Code of Practice for Road Transport Activities. This Code of Practice provides employers with information on driver management, vehicle management, and journey and risk management. Employers shall implement the recommended guidelines in order to ensure the best performance of their company's safety management.

However, some study limitations should be addressed in future studies. Dataset for this study was extracted from the 2014 road accident database in Malaysia. This one-year dataset is limited in terms of specific predictor values which may lead to bias in estimating the parameters. To enhance estimation accuracy and result transferability, more extensive datasets, particularly crash datasets covering many years with adequate qualifying information, are preferred. Methodologically, even though PRC as one of the tools in evaluating the models had been utilised to overcome the data imbalanced issue, it is still not sufficient to predict the minority class of injury severity. Other techniques for imbalanced data such as data pre-processing or cost-sensitive learning could be employed in future studies along with PRC to enhance the model reliability.

6. Conclusions

Determining factors related to injury severity among heavy vehicles driver is important in road safety management, especially among logistics and transportation companies. Based on the 2014 crash dataset for heavy vehicles, this study applied CART and RF to predict and classify the injury severity among drivers. The findings revealed that the RF had better accuracy than CART in classifying the data. However, the TPR rate for all injury classes is low. TPR for RF classifier is almost similar to the TPR for CART, which indicates that both classifiers had a better performance in predicting drivers with no injury than other types of severity since no injury class dominated the training dataset. Thus, future research shall consider treating data imbalance to increase model performance.

This study has also pointed out important factors that are related to different classes of injury severity. Both CART and RF determined collision type as the main factor in determining injury severity among heavy vehicle drivers who were involved in a crash. Other than that, driver's error, number of vehicles involved, road surface type, driver's age, source of light and vehicle type were also identified as factors influencing injury severity. This finding may serve as a basis for future study which may focus deeply on a specific factor contributing to traffic injury among heavy vehicle drivers in Malaysia. Nevertheless, findings by this study may be different from other crashes involving light vehicles such as cars or motorcycles since vehicle size and crash configuration may affect the crash severity. Other countries with other settings than Malaysia may also have different factors that contribute to injury severity. Therefore, the method used by this study can be extended to the case of other types of vehicles or to the case of other countries.

Author Contributions: Conceptualization, N.M.A., M.A.A.B. and A.A.; methodology, N.M.A., M.A.A.B. and A.A.; validation, N.M.A., M.A.A.B. and A.R.; formal analysis, A.A.; investigation, A.A.; resources, A.R.; data curation, A.A. and A.R.; writing—original draft preparation, N.M.A., M.A.A.B. and A.A.; writing—review and editing, N.M.A., M.A.A.B. and A.R.; visualization, A.A.; supervision, N.M.A., M.A.A.B. and A.R.; funding acquisition, N.M.A., M.A.A.B. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universiti Kebangsaan Malaysia with the grant number GUP-2019-048 and by the Ministry of Higher Education (MOHE) with the grant number FRGS/1/2019/STG06/UKM/02/4.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Malaysian Institute of Road Safety Research (MIROS) and are available from MIROS upon request.

Acknowledgments: The authors would like to thank the Malaysian Institute of Road Safety Research (MIROS) for providing the data used in the study.

Conflicts of Interest: The authors declare no conflict of interest in the analyses or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Royal Malaysian Police. *Statistical Report Road Accident Malaysia*; Royal Malaysian Police: Kuala Lumpur, Malaysia, 2018.
2. Islam, M.; Hernandez, S. Large Truck–Involved Crashes: Exploratory Injury Severity Analysis. *J. Transp. Eng.* **2013**, *139*, 596–604. [CrossRef]
3. Rezapour, M.; Ksaibati, K. Application of Multinomial and Ordinal Logistic Regression to Model Injury Severity of Truck Crashes, Using Violation and Crash Data. *J. Mod. Transp.* **2018**, *26*, 268–277. [CrossRef]
4. Pahukula, J.; Hernandez, S.; Unnikrishnan, A. A Time of Day Analysis of Crashes Involving Large Trucks in Urban Areas. *Accid. Anal. Prev.* **2015**, *75*, 155–163. [CrossRef]
5. Zheng, Z.; Lu, P.; Lantz, B. Commercial Truck Crash Injury Severity Analysis Using Gradient Boosting Data Mining Model. *J. Safety Res.* **2018**, *65*, 115–124. [CrossRef]
6. Chang, L.-Y.; Chien, J.-T. Analysis of Driver Injury Severity in Truck-Involved Accidents Using a Non-Parametric Classification Tree Model. *Saf. Sci.* **2013**, *51*, 17–22. [CrossRef]
7. Rahimi, E.; Shamshiripour, A.; Samimi, A.; Mohammadian, A.K. Investigating the Injury Severity of Single-Vehicle Truck Crashes in a Developing Country. *Accid. Anal. Prev.* **2020**, *137*, 105444. [CrossRef]
8. Chen, C.; Zhang, G.; Tian, Z.; Bogus, S.M.; Yang, Y. Hierarchical Bayesian Random Intercept Model-Based Cross-Level Interaction Decomposition for Truck Driver Injury Severity Investigations. *Accid. Anal. Prev.* **2015**, *85*, 186–198. [CrossRef]
9. Behnood, A.; Mannering, F. Time-of-Day Variations and Temporal Instability of Factors Affecting Injury Severities in Large-Truck Crashes. *Anal. Methods Accid. Res.* **2019**, *23*, 100102. [CrossRef]
10. Gudes, O.; Varhol, R.; Sun, Q.; Meuleners, L. Investigating Articulated Heavy-Vehicle Crashes in Western Australia Using a Spatial Approach. *Accid. Anal. Prev.* **2017**, *106*, 243–253. [CrossRef]
11. Dong, C.; Richards, S.H.; Huang, B.; Jiang, X. Identifying the Factors Contributing to the Severity of Truck-Involved Crashes. *Int. J. Inj. Contr. Saf. Promot.* **2015**, *22*, 116–126. [CrossRef]
12. Yu, M.; Ma, C.; Zheng, C.; Chen, Z.; Yang, T. Injury Severity of Truck-Involved Crashes in Work Zones on Rural and Urban Highways: Accounting for Unobserved Heterogeneity. *J. Transp. Saf. Secur.* **2020**, *14*, 83–110. [CrossRef]
13. Uddin, M.; Huynh, N. Truck-Involved Crashes Injury Severity Analysis for Different Lighting Conditions on Rural and Urban Roadways. *Accid. Anal. Prev.* **2017**, *108*, 44–55. [CrossRef] [PubMed]
14. Zhou, B.; Wang, X.; Zhang, S.; Li, Z.; Sun, S.; Shu, K.; Sun, Q. Comparing Factors Affecting Injury Severity of Passenger Car and Truck Drivers. *IEEE Access* **2020**, *8*, 153849–153861. [CrossRef]
15. Samerei, S.A.; Aghabayk, K.; Shiwakoti, N.; Karimi, S. Modelling Bus-Pedestrian Crash Severity in the State of Victoria, Australia. *Int. J. Inj. Contr. Saf. Promot.* **2021**, *28*, 233–242. [CrossRef] [PubMed]
16. Samerei, S.A.; Aghabayk, K.; Mohammadi, A.; Shiwakoti, N. Data Mining Approach to Model Bus Crash Severity in Australia. *J. Saf. Res.* **2021**, *76*, 73–82. [CrossRef]
17. Abrari Vajari, M.; Aghabayk, K.; Sadeghian, M.; Moridpour, S. Modelling the Injury Severity of Heavy Vehicle Crashes in Australia. *Iran. J. Sci. Technol. Trans. Civ. Eng.* **2021**, *46*, 1635–1644. [CrossRef]
18. Anderson, J.; Hernandez, S. Heavy-Vehicle Crash Rate Analysis: Comparison of Heterogeneity Methods Using Idaho Crash Data. *Transp. Res. Rec. J. Transp. Res. Board* **2017**, *2637*, 56–66. [CrossRef]
19. Meuleners, L.; Fraser, M.L.; Govorko, M.H.; Stevenson, M.R. Determinants of the Occupational Environment and Heavy Vehicle Crashes in Western Australia: A Case–Control Study. *Accid. Anal. Prev.* **2017**, *99*, 452–458. [CrossRef]
20. Anderson, J.; Hernandez, S. Roadway Classifications and the Accident Injury Severities of Heavy-Vehicle Drivers. *Anal. Methods Accid. Res.* **2017**, *15*, 17–28. [CrossRef]
21. Kardar, A.; Davoodi, S.R. A Generalized Ordered Probit Model for Analyzing Driver Injury Severity of Head-on Crashes on Two-Lane Rural Highways in Malaysia. *J. Transp. Saf. Secur.* **2020**, *12*, 1067–1082. [CrossRef]
22. Anarkooli, A.J.; Hosseinpour, M.; Kardar, A. Investigation of Factors Affecting the Injury Severity of Single-Vehicle Rollover Crashes: A Random-Effects Generalized Ordered Probit Model. *Accid. Anal. Prev.* **2017**, *106*, 399–410. [CrossRef] [PubMed]
23. Jiang, X.; Huang, B.; Zaretzki, R.L.; Richards, S.; Yan, X.; Zhang, H. Investigating the Influence of Curbs on Single-Vehicle Crash Injury Severity Utilizing Zero-Inflated Ordered Probit Models. *Accid. Anal. Prev.* **2013**, *57*, 55–66. [CrossRef] [PubMed]
24. Xie, Y.; Zhao, K.; Huynh, N. Analysis of Driver Injury Severity in Rural Single-Vehicle Crashes. *Accid. Anal. Prev.* **2012**, *47*, 36–44. [CrossRef] [PubMed]
25. Chen, C.; Zhang, G.; Qian, Z.; Tarefder, R.A.; Tian, Z. Investigating Driver Injury Severity Patterns in Rollover Crashes Using Support Vector Machine Models. *Accid. Anal. Prev.* **2016**, *90*, 128–139. [CrossRef] [PubMed]
26. Moral-García, S.; Castellano, J.; Mantas, C.; Montella, A.; Abellán, J. Decision Tree Ensemble Method for Analyzing Traffic Accidents of Novice Drivers in Urban Areas. *Entropy* **2019**, *21*, 360. [CrossRef]
27. Wahab, L.; Jiang, H. Severity Prediction of Motorcycle Crashes with Machine Learning Methods. *Int. J. Crashworthiness* **2020**, *25*, 485–492. [CrossRef]
28. Wahab, L.; Jiang, H. A Comparative Study on Machine Learning Based Algorithms for Prediction of Motorcycle Crash Severity. *PLoS ONE* **2019**, *14*, e0214966. [CrossRef]
29. Haynes, S.; Estin, P.C.; Lazarevski, S.; Soosay, M.; Kor, A.-L. Data Analytics: Factors of Traffic Accidents in the UK. In Proceedings of the 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT), Leeds, UK, 5–7 June 2019; pp. 120–126. [CrossRef]

30. Lund, I.O.; Rundmo, T. Cross-Cultural Comparisons of Traffic Safety, Risk Perception, Attitudes and Behaviour. *Saf. Sci.* **2009**, *47*, 547–553. [CrossRef]
31. Nordfjærn, T.; Şimşekoğlu, Ö.; Rundmo, T. Culture Related to Road Traffic Safety: A Comparison of Eight Countries Using Two Conceptualizations of Culture. *Accid. Anal. Prev.* **2014**, *62*, 319–328. [CrossRef]
32. Nordfjærn, T.; Şimşekoğlu, Ö.; Rundmo, T. A Comparison of Road Traffic Culture, Risk Assessment and Speeding Predictors between Norway and Turkey. *Risk Manag.* **2012**, *14*, 202–221. [CrossRef]
33. Royal Malaysian Police. *Statistical Report Road Accident Malaysia*; Royal Malaysian Police: Kuala Lumpur, Malaysia, 2014.
34. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: London, UK, 2017. [CrossRef]
35. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
36. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Langley, P., Ed.; Morgan Kaufmann Publishers: Burlington, MA, USA, 1993.
37. Therneau, T.; Atkinson, E.J. *An Introduction to Recursive Partitioning Using the RPART Routines*; Mayo Foundation: Rochester, MN, USA, 2019.
38. Hastie, T.; Tibshirani, R.; Jerome, F. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009. [CrossRef]
39. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
40. Davis, J.; Goadrich, M. The Relationship between Precision-Recall and ROC Curves. In Proceedings of the 23rd international conference on Machine learning—ICML '06, Pittsburgh, PA, USA, 25–29 June 2006; Volume 148, pp. 233–240. [CrossRef]
41. He, H.; Ma, Y. *Imbalanced Learning*; Wiley: Hoboken, NJ, USA, 2013. [CrossRef]
42. Therneau, T.; Atkinson, B. Recursive Partitioning and Regression Tree; R Package Version 4.1-15; 2019. Available online: https://CRAN.R-project.org/package=rpart (accessed on 20 January 2022).
43. Max, K. caret: Classification and Regression Training; R Package Version 6.0-86; 2020. Available online: https://CRAN.R-project.org/package=caret (accessed on 20 January 2022).
44. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, R package version 1.7-4; 2020. Available online: https://CRAN.R-project.org/package=e1071 (accessed on 20 January 2022).
45. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News.* **2002**, *2*, 18–22. Available online: https://CRAN.R-project.org/doc/Rnews/ (accessed on 20 January 2022).
46. Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: Visualizing classifier performance in R. *Bioinformatics.* **2020**, *21*, 7881. [CrossRef] [PubMed]
47. Wang, X.; Kim, S.H. Prediction and Factor Identification for Crash Severity: Comparison of Discrete Choice and Tree-Based Models. *Transp. Res. Rec.* **2019**, *2673*, 640–653. [CrossRef]
48. Yu, H.; Li, Z.; Zhang, G.; Liu, P. A Latent Class Approach for Driver Injury Severity Analysis in Highway Single Vehicle Crash Considering Unobserved Heterogeneity and Temporal Influence. *Anal. Methods Accid. Res.* **2019**, *24*, 100110. [CrossRef]
49. Adanu, E.K.; Riehle, I.; Odero, K.; Jones, S. An Analysis of Risk Factors Associated with Road Crash Severities in Namibia. Int. *J. Inj. Contr. Saf. Promot.* **2020**, *27*, 293–299. [CrossRef]
50. Lee, C.; Li, X. Predicting Driver Injury Severity in Single-Vehicle and Two-Vehicle Crashes with Boosted Regression Trees. *Transp. Res. Rec.* **2015**, *2514*, 138–148. [CrossRef]
51. Zhu, M.; Li, Y.; Wang, Y. Design and Experiment Verification of a Novel Analysis Framework for Recognition of Driver Injury Patterns: From a Multi-Class Classification Perspective. *Accid. Anal. Prev.* **2018**, *120*, 152–164. [CrossRef]