APPENDIX A

LIST OF PUBLICATIONS

List of Publications

Mohamad, I., Jomnonkwao, S., & Ratanavaraha, V. (2022). Using a decision tree to compare rural versus highway motorcycle fatalities in Thailand. *Case Studies on Transport Policy*, *10*(4), 2165-2174. https://doi.org/https://doi.org/10.1016/j.cstp.2022.09.016 (IF : 3.043, Q1 in Scopus)

Case Studies on Transport Policy 10 (2022) 2165–2174

Contents lists available at ScienceDirect

Case Studies on Transport Policy

journal homepage: www.elsevier.com/locate/cstp

Using a decision tree to compare rural versus highway motorcycle fatalities in Thailand

Ittirit Mohamad [a], Sajjakaj Jomnonkwao [b,*], Vatanavongs Ratanavaraha [c]

[a] Program of Energy and Logistics Management Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand
[b] School of Transportation Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand
[c] School of Transportation Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand

ARTICLE INFO

Keywords:
Comparative analysis
Decision tree
Accidents
Big data
Transportation
Machine learning

ABSTRACT

Thailand ranks first in Asia and ninth in the world in term of road accident. As of 2020, the number of vehicles registered in Thailand was over 41 million, with motorcycles accounting for half of all vehicles. This study aimed to determine the cause of fatalities to reduce motorcycle accidents. The research entailed separating the accidents and fatalities into those occurring on highways (HWs) versus those occurring on rural roadways (RRs) and focused solely on rider at fault accidents to involve any confounding factors related to passengers or others involved. In Thailand, HWs have higher speed limits and allow more vehicle types than some RRs. Thailand's Department of Public Disaster Prevention and Mitigation recorded 115,154 motorcycle accidents from 2015 to 2020. Decision trees allow for processing large amounts of data to drill down into associations between the individual variables in a large data set; in this study, the tree also separated accidents into whether or not the driver was exceeding the speed limit. The model's performance for HWs, predicted misclassifications were found to be 28.3% (fatality to nonfatality) and a 44.5% (nonfatality to fatality) while predicted misclassification for RRs were 15.5% (fatality to nonfatality) and 60% (nonfatality to fatality). At all ages, the most fatalities were among male riders on dry straightaways in clear daytime weather; notably, however, on RRs, even when the rider was driving responsibly, fatalities were high at night on roads with no light. Following the presentation of the study findings, suggestions are made for ways the Thai government can improve the motorcycle accident and fatality statistics, including increasing the age limit for a motorcycle license, with engine size limits further divided according to age; proper enforcement of the existing rules will also improve the country's accident statistics. It will also be highly effective to improve road lighting, particularly on RRs.

1. Introduction

Thailand is one of the countries with a high rate of fatalities from road accidents, ranking first in Asia and ninth in the world. Thais are killed in traffic accidents at a rate of 32.7 per 100,000 persons (WHO, 2018). As of 2020, the total number of vehicles registered in Thailand was over 41 million, with motorcycles accounting for half of all vehicles (DLT, 2021).

Fig. 1 shows that the number of motorcycles registered in Thailand grew continually from 2015 to 2020, when there were 41,471,135 vehicles registered in the country, of which 21,396,980 were motorcycles; these were followed by lightweight 4-wheeled drive vehicles (10,446,505), mini-trucks (6,878,050), and others (2,749,600) (DLT, 2021). However, although motorcycles only account for half the vehicles in Thailand, motorcyclists account for the majority of road accident fatalities (Jomnonkwao et al., 2020). According to the Thailand Accident Research Center, in 2019, there were 4802 motorcycle fatalities, an average of 13.15 people per day, with most occurring among people aged 20 years and up, also known as the working age group. In terms of causes, 54 % of accidents were the fault of the motorcyclists, drivers were at fault in 40 % of the accidents, and the road and vehicle accounted for 4 % and 2 %, respectively (RSC, 2019). Worldwide research has identified elements that appear to be common to every accident. Thailand has six road types:

- A motorway is a HW designed for high mobility with low access based on limited entrances and exits to designated points, and no

* Corresponding author.
E mail address: sajjakaj@g.sut.ac.th (S. Jomnonkwao).

https://doi.org/10.1016/j.cstp.2022.09.016
Received 29 June 2022; Received in revised form 19 August 2022; Accepted 27 September 2022
Available online 30 September 2022

two-wheeled vehicles are permitted. Motorways are supervised by the Department of Highways.

- The national HWs link regions, provinces, and districts; they emphasize mobility, but access is not limited, and rules in general are less strict than those for motorways are. It is difficult to travel through cities, and the HWs were designed to bypass the cities; they are also supervised by the Department of Highways.
- Rural roadways (RRs) are located outside of municipalities and connect with the national HWs. They are supervised by the Department of Rural Roads.
- Municipal roadways provide the streets in municipalities and are maintained by the municipalities.
- Subdistrict roadways serve as the streets for those areas, and they are supervised by subdistrict organizations.
- Concession roadways are privately owned; the government grants concessions to private entities that are then responsible for supervising the roads.

For this study, all roads managed by the Department of Highways are designated HWs and rural and subdistrict roads are RRs. The motorcycle speed limits are lower on RRs—80 km/hr. for engines larger than 400 cc and 60 km/hr. for smaller cycles; for HWs, the limits are 90 km/hr. for engines larger than 400 cc and 70 km/hr. otherwise (DLT, 2021) and the study was distinctive in that it separated accidents and fatalities into those that occurred on highways (HWs) and those that occurred on rural roadways (RR), and it focused solely on rider-at-fault accidents in order to eliminate any confounding factors related to passengers or others involved.

In statistical side, Discriminant analysis (DA) is a technique commonly used in Statistical Algorithms to classify a set of observations into predefined classes and LDA (linear Discriminant analysis) is a diagnostic method for detecting potentially influential observations. The usual assumptions relevant to discriminant analysis are linearity, normality, and homoscedasticity of within-group variances of independent variables. However, due to violations of these assumptions, discriminant analysis has been supplanted by LR (logistic Regression), which requires fewer assumptions, produces more robust results, and is easier to use and comprehend than discriminant analysis (Chen, 2012). A regression-type model is a CART model that predicts the value of continuous variables using a set of continuous or categorical predictor variables. For this study, we selected a decision tree (CART regression Tree) to drill down into a set of big data to identify the relevant variables and analyze the relationships among them. Decision tree mining is among the most popular machine learning techniques (Wu et al., 2007) for its comprehensibility and ease of interpretation. One of the primary advantages of a decision tree is the ability to derive decision rules; these rules can aid in identifying safety issues and developing performance metrics (Abellán et al., 2013).

2. Literature review

Previous studies on road accidents have been classified by group components that are suspected to be involved in every accident, according to international research.

2.1. Age and gender

Motorcycle accidents are more likely to occur among young people because they are less disciplined, are unfamiliar with traffic laws, and have less driving experience (Zhang and Fan, 2013). Men and women aged 20–39 years who ride motorcycles are more likely to be involved in major accidents, whereas when no motorcycle or cyclist is involved in the incident, the severity is likely to be minor(Ospina-Mateus et al., 2019). Jou et al. (2012) found that being older, male, and unlicensed; not wearing a helmet; riding after drinking; and driving heavy motorcycles (above 550 cc) were linked to higher motorcycle fatality rates. Additionally, rider age was the most important factor when the rider was not at fault (Champahom et al., 2019). Pakgohar et al. (2011) found that the majority of fatalities were among young persons who were in good health before the accident. Riders between the ages of 18 and 24 years have insufficient experience to make adjustments while driving including adjusting their speed to the road conditions (Bucsuházy et al., 2020).

2.2. Weather and road conditions

Research has established that external conditions such as fog, rain, and snow have a greater influence on road accidents than rider-related internal factors and that the drivers/riders are more likely than passengers to be injured or killed in an accident (El Abdallaoui et al., 2018). According to the findings, the most important and influential road accident variables are speed limit; weather conditions; road factors such as type, surface, and number of lanes; lighting conditions; and time of the accident. Factors that had less influence on accidents were gender, age, accident site, and vehicle type (Feng et al., 2020). Highway (HW) intersections have been identified as the most dangerous for all accidents (Kumar and Toshniwal, 2016),Malin et al. (2019). As noted above, however, there are still significant accidents on straightaways with no intersections, in part because riders disobey the speed limit and in part because of poor road conditions.

Fig. 1. Total number of vehicles and motorcycles registered in Thailand from 2015 to 2020.

2.3. Other important factors

Vehicle speed is the most critical determinant of an accident's severity (Al Mamlook et al., 2019), followed by factors such as speed limit, age, and road type (Rezapour et al., 2020). Travel at night increases the risk of an accident (Mphela, 2020) and increases the severity of any injuries, particularly when there is no light Shaheed et al. (2013), (Kim et al., 2013), (Jafari Anarkooli et al., 2017) and after midnight (Zhou and Chin, 2019). Xie and Huynh (2012) determined that the severity of injuries from accidents on dark roads decreases when riders are more cautious. Motorcycles are riskier in rural areas. Male riders, pillion riders, speeding, improper overtaking, and fatigue are all important factors that influence severe and fatal injuries (Se et al., 2021).

Additionally, the risk of motorcycle death increases for single-vehicle accidents that occur on nonurban roads at night, and the major factors that affect rear-end crashes are passenger characteristics and the rider's age, whereas side collisions are most commonly the result of lighting conditions and landscape (Anvari et al., 2017; Siskind et al., 2011). Focusing on driver factors, researchers discovered that high-speed driving, driving while intoxicated. And traffic violations all contributed to high rates of fatalities on RRs (Khorashadi et al., 2005). Researchers have used many tools in accident analysis, including measuring the accuracy between models or methods (Table 1), but few have studied the same model or method to compare two road types.

3. Methodology

The research begins with motorcycle accident data from Thailand's Department of Public Disaster Prevention and Mitigation, which counted 115,154 single-rider accidents between 2015 and 2020. Toward our study aim, the data was compiled on HR and RR motorcycle accident fatalities, developed a decision tree model, and measured its accuracy. Fig. 2 displays the steps in the study process, also listed below:

- After cleansing the data set, the initial dataset was validated for detecting and correcting missing and incompletely captured data as well as demonstrating the data's quality.
- Verified Dataset – Set the target to Fatal/non-fatal and partition the data in binary mode both HW and RR data set.
- Data Separation – Separated test and train data sets.
- Model Learning enables the model to learn from the test data set and then test with the remaining data set.
- Prediction and Model Measurement - To assess each model's prediction accuracy.

3.1. Data description

As noted earlier, the 2015–2020 motorcycle accident data from the Thailand Department of Public Disaster Prevention and Mitigation indicated 115,154 single-rider accidents, 61,866 on HWs and 53,288 on RRs (PDPM, 2020). Table 2 presents the categorical and descriptive statistics for the study data, which we divided into four categories: roadway characteristics, external factors involving the environment and weather conditions, internal factors involving driver behavior, and driver details. According to the descriptive data table, most accidents on both HWs and RRs were caused by being a male rider between 15 and 35 years of age exceeding the speed limit; most accidents occurred on dry surfaces and in clear weather, even when the driver stayed on the right side of the road.

Table 1
The Machine Learning Models Used in Extant Traffic Accident Studies.

Author	Methodology Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Gradient Boosting	K-Nearest Neighbor	K-Means	Multinomial Logistic Regression	Neural Network	Naïve Bayes	Random Forest	Support Vector Machine
Ospina-Mateus et al. (2021)	–	–	–	✓	–	✓	–	–	✓	✓	✓	✓
Harb et al. (2009)	–	–	–	✓	–	–	–	–	–	–	✓	–
Kuşkapan et al. (2021)	–	–	–	–	–	✓	–	–	–	✓	–	✓
Abellán et al. (2013)	–	–	–	✓	–	–	–	–	–	–	–	–
Mafi et al. (2018)	–	–	–	–	–	–	–	–	–	–	✓	–
Al Mamlook et al. (2019)	–	✓	✓	✓	–	✓	–	–	–	✓	✓	✓
Recal and Demirel (2021)	–	–	–	✓	✓	–	–	✓	✓	–	–	✓
Kumar and Toshniwal (2016)	✓	–	–	–	–	–	✓	–	–	–	–	–
Helen et al. (2019)	✓	–	–	–	–	–	✓	–	–	–	–	–
Feng et al. (2020)	✓	–	–	–	–	–	–	–	✓	–	–	–
Bhavsar et al. (2021)	✓	–	–	–	–	–	–	–	–	–	–	–
Bahiru et al. (2018)	–	–	–	✓	–	–		–	–	✓	–	–

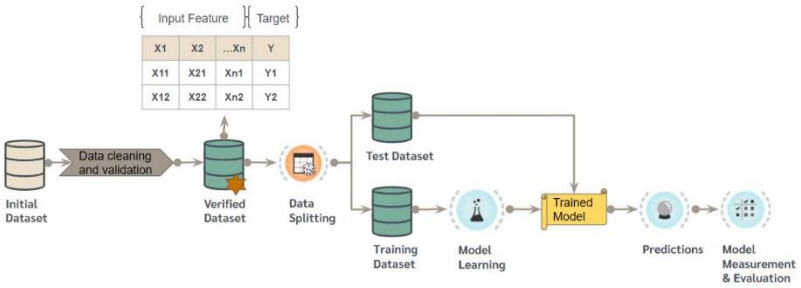

Fig. 2. The steps in the process for the study.

Table 2
The categorical variables and their descriptive statistics.

	HighWay				Non-HighWay			
	Fatality				Fatality			
	Yes		No		Yes		No	
Aciccent Event (Attribute)	Count	%	Count	%	Count	%	Count	%
RoadWay								
Dry Surface Road	18992	30.7%	40112	64.8%	8063	15.1%	43049	80.8%
Wet Surface	673	1.1%	2089	3.4%	277	0.5%	1899	3.6%
Straight Way	14171	22.9%	29389	47.5%	5619	10.5%	31698	59.5%
Not straight Way (Curve, Slope, Junction, etc)	5494	8.9%	12812	20.7%	2721	5.1%	13250	24.9%
Obstruction	343	0.6%	1301	2.1%	170	0.3%	1847	3.5%
Road condition	194	0.3%	886	1.4%	167	0.3%	1388	2.6%
Vehicle condition	226	0.4%	825	1.3%	126	0.2%	1402	2.6%
External Factor (Envi and Weather Con)								
Day Time (06.00-18.00)	9649	15.6%	24654	39.9%	4021	7.5%	26215	49.2%
Night with Light	5633	9.1%	11310	18.3%	2140	4.0%	9936	18.6%
Night without Light	4383	7.1%	6237	10.1%	2179	4.1%	8797	16.5%
Low visibility	1819	2.9%	5437	8.8%	659	1.2%	5523	10.4%
Clear Weather	17298	28.0%	35951	58.1%	7451	14.0%	39037	73.3%
Not Clear Weather (Rain, fog, etc)	2367	3.8%	6250	10.1%	889	1.7%	5911	11.1%
Internal Factor (Driver Behavior)								
Drunk	2410	3.9%	9521	15.4%	1357	2.5%	11065	20.8%
Over Speed limit	13524	21.9%	20018	32.4%	5888	11.0%	18129	34.0%
Break Through Traffic lights	185	0.3%	283	0.5%	50	0.1%	183	0.3%
Break Through Traffic Signs	289	0.5%	748	1.2%	88	0.2%	573	1.1%
Overtake	526	0.9%	702	1.1%	121	0.2%	576	1.1%
Use Mobile Phone	22	0.0%	171	0.3%	2	0.0%	185	0.3%
Short Cut off	4156	6.7%	9170	14.8%	1248	2.3%	10141	19.0%
Drug	3	0.0%	39	0.1%	3	0.0%	27	0.1%
Drive in opposite direction	385	0.6%	563	0.9%	58	0.1%	268	0.5%
Doze off	260	0.4%	536	0.9%	92	0.2%	369	0.7%
Overweight Carry	11	0.0%	28	0.0%	5	0.0%	37	0.1%
Cannot Conclude	676	1.1%	1686	2.7%	256	0.5%	1772	3.3%
Driver info								
Gender (male)	16921	27.4%	30998	50.1%	7236	13.6%	32500	61.0%
Gender (Female)	2744	4.4%	11203	18.1%	1104	2.1%	12448	23.4%
Youth 15-35	9894	16.0%	23204	37.5%	4010	7.5%	23277	43.7%
Adult 36-60	7111	11.5%	14830	24.0%	3205	6.0%	17126	32.1%
Senior 61-90+	2660	4.3%	4167	6.7%	1125	2.1%	4545	8.5%

**External factors are environment and weather conditions.

3.2. The decision tree

A decision tree is a predictor, h: X → Y, of the predecessors of an event x by spanning a tree from its root node to its leaves. For simplicity, we concentrated on the binary classification case, namely, Y = {0, 1}, but decision trees can be used for a range of prediction problems. Based on the division of the input space, the successor child is chosen at each node along the root-to-leaf path. Usually, the splitting is based on one of x's properties or a predefined set of splitting rules as follows:

- First, set the domain set: X is the accident event that needs to be labeled.

Set X to be binary {1,0}, and let Y be our possible labels.

Then, Y = {0, 1}, where 1 and 0 represent the possible options.

- Training set $S = ((X1, Y1)\cdots(Xn, Yn))$ is a limited number of pairings in $X \times Y$, that is, a list of labeled domain points. This is the information to which the learner has access.
- For the output, the learner is asked to generate a prediction rule, h: X → Y. This function is also referred to as a prediction, hypothesis, or classifier. The predictor can forecast new domain elements (Ben-David, 2014).

Decision trees comprise three parts: decision nodes, branches, and leaf nodes. Each decision node in the structure displays the variable, and each branch displays one variable value based on decision rules; the leaf nodes display the expected values of the target variables (Song and Ying, 2015). We used Orange 3.30 software (Demšar et al., 2013) to run the CART decision tree set classification to stop at when majority reach 95 % and limit of maximum tree depth is 7. Data was divided into two flows, HW and RR, and extracted 27 binary categorical variables that were most relevant to the 115,154 single-rider motorcycle accidents in Thailand from 2015 to 2020; the variables were set as binary (1 or 0) to facilitate interpretation and classification. Table 3 presents the 27 most relevant variables related to single-rider accidents under the following factors: roadway factors, external (environment, weather) and internal (driver behaviors) factors, driver data, and driver status.

3.3. Performance measurement

To assess the performance of the supervised machine learning decision tree in this study, we used tests data to validation how well the model performed with a confusion matrix with the following components: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). For example, TP shows the number of positive values projected to be positive, whereas FP predicts an accident as fatal when it was not; recall and precision were also measured. These strategies are especially useful for unbalanced data sets, in which one answer category accounts for the bulk of the responses. Precision refers to the accuracy of the classifier findings, expressed as eq (1) and shown in Fig. 3:

$$Precision = \frac{\mathrm{TP}}{TP + FP} \tag{1}$$

Recall, or sensitivity, gives the proportion of the positive class that was correctly classified, expressed as eq (2) :

$$Recall = \frac{\mathrm{TP}}{TP + FN} \tag{2}$$

The TN rate (TNR), also called specificity, is computed as eq (3) :

$$TNR = \frac{\mathrm{TN}}{TP + FN} \tag{3}$$

The FP rate (FPR) shows how often the classifier misclassified the negative class and is computed as eq (4):

$$FPR = \frac{\mathrm{FP}}{TN + FP} = 1 - TNR(Specificity) \tag{4}$$

The ratio of correct classifications reflects the data accuracy and is calculated as eq (5):

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{TP + TN + FP + FN} \tag{5}$$

Table 3
The Measurement Categories for the 27 Identified Motorcycle Accident Variables.

Factor and Variables	Measurement
Roadway	
Dry road	1 – Yes, 0-Otherwise
Straight road	1 – Yes, 0-Otherwise
Obstruction	1 – Yes, 0-Otherwise
Road conditions	1 – Yes, 0-Otherwise
Vehicle conditions	1 – Yes, 0-Otherwise
External Factors (Environment and Weather Conditions)	
Day Time (06.00–18.00)	1 – Yes, 0-Otherwise
Night with light	1 – Yes, 0-Otherwise
Night without light	1 – Yes, 0-Otherwise
Low visibility	1 – Yes, 0-Otherwise
Clear weather	1 – Yes, 0-Otherwise
Internal Factors (Driver Behaviors)	
Drunk	1 – Yes, 0-Otherwise
Over speed limit	1 – Yes, 0-Otherwise
Ran a traffic light	1 – Yes, 0-Otherwise
Ran a traffic sign	1 – Yes, 0-Otherwise
Passing (overtaking)	1 – Yes, 0-Otherwise
Used a mobile phone	1 – Yes, 0-Otherwise
Short cutoff	1 – Yes, 0-Otherwise
Used drugs	1 – Yes, 0-Otherwise
Drove in opposite direction	1 – Yes, 0-Otherwise
Dozed off	1 – Yes, 0-Otherwise
Overweight cargo	1 – Yes, 0-Otherwise
Inconclusive	1 – Yes, 0-Otherwise
Driver Data	
Gender	1- Male, 0-Otherwise
Youth 15–35	1 – Yes, 0-Otherwise
Adult 36–60	1 – Yes, 0-Otherwise
Senior 61–90+	1 – Yes, 0-Otherwise
Driver Status	
Fatality	1 – Yes, 0-Otherwise

4. Results

Fig. 4 presents plots of the HW and RR likelihoods of fatalities over 24 h.

HWs

1 = fatality: μ =13.47, σ = 6.34
0 = nonfatality: μ =13.71, σ = 6.31

RRs

1 = fatality: μ =13.14, σ = 7.11
0 = nonfatality: μ =13.89, σ = 6.29

There is a higher probability of a fatality on a HW than on a RR: HW, 0.3–0.4 and RR, 0.25–0.15. However, both RRs and HWs have higher fatality rates at night (00.00–07.00). Both the HW and the RR decision trees were set to target rider fatalities, and they identified the main causes. The tree node was divided into whether the rider was following or exceeding the speed limit.

- HW Fatalities (Fig. 5):
  - o Rider exceeds the speed limit (40.3 %: 13,524/33,542): male (44.7 %: 11,757/26,312), age 15–35 years (35 %: 2,411/6,891), age 61–90 years (48.2 %: 1,808/4,152), straight road (43.5 %: 1,808/4,152), daytime (39.3 %: 5,364/13,653), clear weather (63.5 %: 1,933/3,048).
  - o Rider does not exceed the speed limit (21.7 %: 6,141/28,324): male (23.9 %: 5,164/21,607), age 15–35 years (22.7 %: 1,471/6,477), age 61–90 years (43.1 %: 447/1,037), drunk (17.9 %: 1,459/8,165), daytime (17.8 %: 581/3,257), night w/o light (25.3 %: 311/1,230), clear weather (19.1 %: 1,214/6,367), short cutoff (35 %: 1,008/2,881). On HWs, fatalities occurred most commonly among male riders who were between the ages of 15 and 35 years and were exceeding the speed limit on a straight road in clear

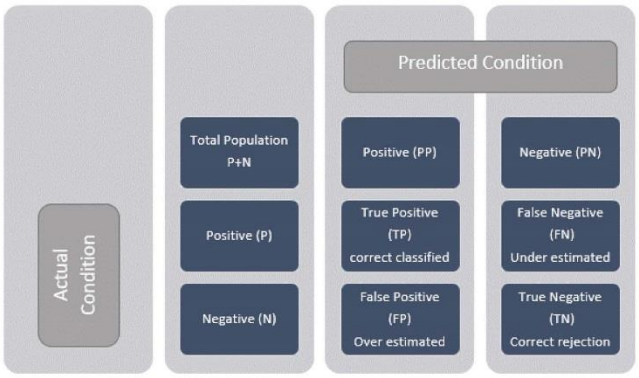

Fig. 3. Diagram of the confusion matrix.

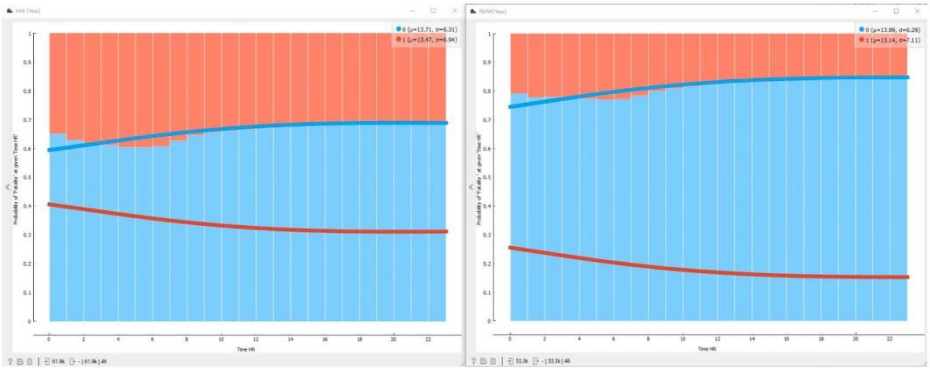

Fig. 4. HW and RR fatality probabilities at different times of the day.

weather during the day. For riders aged 61 to 90 years, the fatalities occurred most often at night, without lights, under rider intoxication, and with short cutoffs.

- RR Fatalities (Fig. 6):
  - o Rider exceeds the speed limit (24.5 %: 5,888/24,017): male (28.3 %: 5,132/18,131), age 15–35 years (20.2 %: 950/4,696), age 36–60 years (26.2 %: 952/3,639), daytime (24.3 %: 2,343/9,645), night w/o light (40.1 %: 974/2,427), short cutoff (55.8 %: 115/206), clear weather (36.1 %: 2,114/5,863).
  - o Rider does not exceed the speed limit: male (9.7 %: 2,104/21,605), night w/o light (11.2 %: 567/5,057).

Table 4 presents the final-level leaf sets for HW and RR motorcycle accident fatalities according to whether or not the driver was exceeding the speed limit. For instance, for RR fatalities, the common factors in both age groups were being male and riding over the speed limit during the day; at night on RRs, male riders who died were most often speeding and making short cutoffs at night with no light. When drivers were not speeding, most fatalities occurred among males at night with no light.

Short cutoffs were among the most common causes of fatalities on both HWs and RRs, but excess speed was also a factor only on RRs. Gender was the most significant variable in fatalities: Most fatalities were among male riders on both HWs and RRs irrespective of the rider speed limit, consistent with earlier findings that men who ride motorcycles are more likely to be involved in serious accidents Ospina-Mateus et al. (2019).

4.1. Evaluation results

Table 5 presents the cross-validated data results with 20 folds. The table presents the area under the receiver-operating curve (AUC) and the classification accuracy (CA; Equation (5)), recall (Equation (2)), and precision (Equation 1). CA is high for HWs and RRs. When $0.5 < AUC < 1$, there is a good possibility that the classifier will be able to differentiate between positive and negative class values. This is because the classifier is better able to recognize TP and TN (Equation (3)) than FN and FP (Eq. (4)).

The HW confusion matrix in Fig. 7 shows misclassifications in 77.9 % of actual fatalities to nonfatality accidents and 8.3 % of nonfatalities to fatalities. For the predicted values, we identified a misclassification of 28.3 % for accident fatalities to nonfatalities and 44.5 % for nonfatalities to fatalities. Fig. 8 shows the RR confusion matrix, reflecting

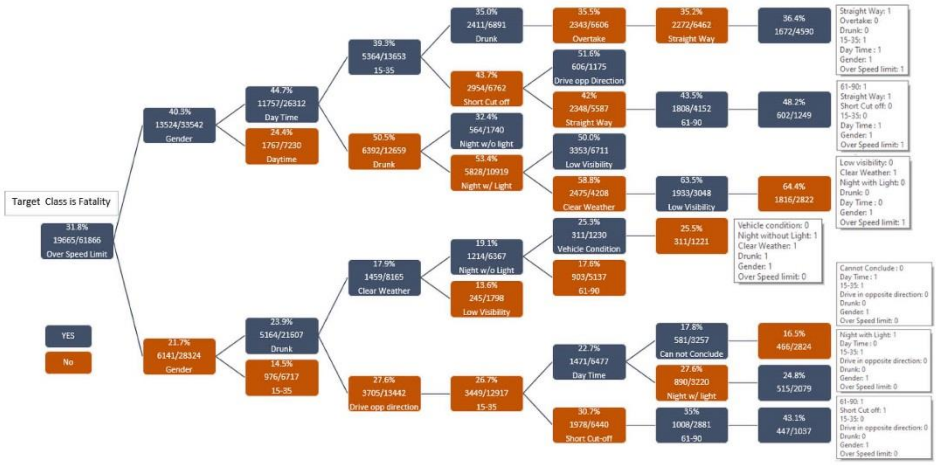

Fig. 5. The HW tree model.

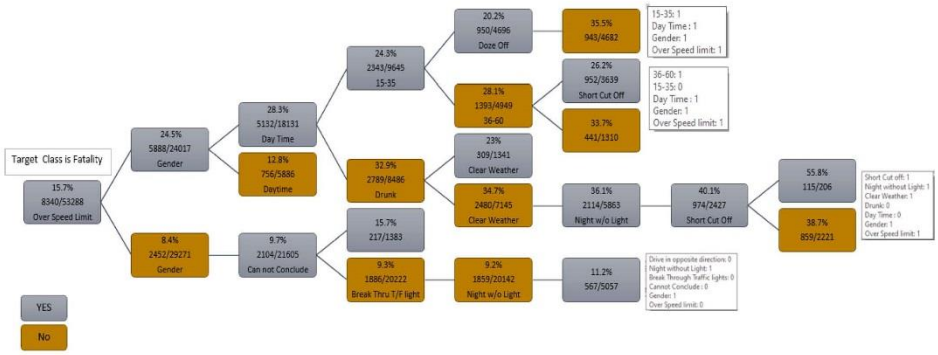

Fig. 6. The RR tree model.

misclassifications of 97.9 % (fatality to nonfatality) and 0.6 % (nonfatality to fatality) for the actual cases and of 15.5 % (fatality to nonfatality) and 60 % (nonfatality to fatality). That is, the decision tree for this study shows significant misclassifications of fatalities for both HWs and RRs but much better ability with nonfatalities.

5. Conclusion and discussion

The aim of this research was to design a decision tree to identify individual contributors to motorcycle accident fatalities among riders in Thailand, with a focus on single-rider crashes. Contributing variables included roadway features along with external and internal (driver-related) factors. In addition, using accident data from 2015 to 2020, we performed a nonparametric analysis to determine the importance of factors that influence target variables, such as road and weather conditions, speeding, being on a straightaway (with no intersections), gender, and substance use.

The decision tree concluded that the most common causes of fatality on both HWs and RRs were being a male rider and exceeding the speed limit, with the other variables showing differing levels of importance (Fig. 9). HWs have more contributors to fatalities than RRs; for instance, accidents were common on HWs when riders had been drinking, especially at night with no light. HWs also have much heavier traffic and a wider variety of vehicles traveling at higher speeds than RRs, facilitating accidents that can cause serious injury and death. Age was another significant contributor to motorcycle accident fatalities on both types of roads, although notably, only HW fatalities extended to riders up to the age of 90 years; the highest age in RR fatalities was 60 years. One interesting observation here was that RR fatalities generally occurred at night with no street lighting whether or not a rider was speeding. Road accidents have many contributing factors, but speeding is a key area of concern for the severity of road accident injuries (Yu et al., 2020; Osman et al., 2018; Krull et al., 2000). Thus, we propose that Thailand must properly post speed restrictions and support the enforcement of compliance within these limits. Short-cutoff riding was another key predictor of motorcycle fatalities Bahiru et al. (2018). Although male

Table 4
The Final HW and RR Sets by Rider Speed.

	HW Fatalities (Fig. 4):	RR Fatalities (Fig. 5):
Rider exceeds the speed limit	Set 1: (15–35, Day time, Male, Straight way) 36.4 % (1,672/4,590) Set 2: (61–90, Day time, Male, Straight way) 48.2 % (602/1,249) Set 3 (Clear Weather, Male) 64.4 % (1,816/2,822)	Set 1: (15–35, Day time, Male) 35.5 % (943/4,682) Set 2: (36–60, Day time, Male) 26.2 % (952/3,639) Set 3: (Clear weather, short cutoff, Night without light, Male) 55.8 % (115/206)
Rider *does not exceed* the speed limit	Set 1: (Clear Weather, Male, Drunk) 25.5 % (311/1,221) Set 2: (Daytime,15–35, Male) 16.5 % (466/2,824) Set 3: (Night with light,15–35, Male) 24.8 % (515/2,079) Set 4: (Short cutoff,61–90, Male) 43.1 % (477/1,037)	Set 1: (Night without light, Male) 11.2 % (567/5,057)

Table 5
The Model Evaluation Results.

Model	Road	Target Class	AUC	CA	Precision	Recall
Tree	HW	Avg over	0.685	0.696	0.665	0.696
		Fatality	0.686	0.696	0.555	0.221
		Nonfatality	0.686	0.696	0.717	0.917
	RR	Avg Over	0.706	0.842	0.776	0.842
		Fatality	0.703	0.842	0.400	0.021
		Nonfatality	0.703	0.842	0.845	0.994

gender was a primary factor in HW motorcycle fatalities in this study, we found less influence from age, accident location, and vehicle type. However, substance use was a factor in many accidents, and we propose more education on the dangers of riding while intoxicated in addition to tighter enforcement of penalties for infractions.

In Thailand, persons of all ages ride motorcycles, although the majority of riders are between the ages of 15 (bike capacity of no more than 110 cc) and 35 years. Looking at all four accident scenarios in the research, three featured young people, consistent with Zhang and Fan (2013), in which accidents were more likely among young people (25 years old), who are less disciplined and unfamiliar with traffic laws and have less driving experience. Policymakers might consider raising the minimum age for obtaining a motorcycle license to at least 18 years or imposing further restrictions on engine size dependent on rider age. We also identified road lighting as a considerable factor in motorcycle accidents, particularly on RRs but in fact in all accident categories except for speeding-related deaths on HWs. Therefore, we propose that better lighting be installed wherever possible on Thai roadways, particularly in rural areas.

6. Limitations and future studies

This study's model showed acceptable (above 50 %) accuracy, but there is room for improvement; adjusting the parameters in a future study could increase the accuracy. Additionally, we used accident data from 2015 to 2020, but during the last two years, 2019 and 2020, the circumstances in Thailand as well as around the world changed drastically overnight because of the COVID-19 pandemic. Governments worldwide locked down and ordered people to stay indoors, and Thailand limited travel between provinces, particularly between the hours of 22.00 and 04.00. Because mobility was so limited during 2019 and 2020, the overall findings for those years might not accurately reflect what would have been the country's true numbers of accidents and fatalities.

CRediT authorship contribution statement

Ittirit Mohamad: Conceptualization, Data curation, Formal analysis, Methodology, Software. **Sajjakaj JomnonKwao:** Validation, Writing – review & editing. **Vatanavongs Ratanavaraha:** Visualization, Supervision.

Confusion matrix for Tree (showing proportion of actual)

		Predicted 0	Predicted 1	Σ
Actual	0	91.7 %	8.3 %	42201
	1	77.9 %	22.1 %	19665
	Σ	54031	7835	61866

Confusion matrix for Tree (showing proportion of predicted)

		Predicted 0	Predicted 1	Σ
Actual	0	71.7 %	44.5 %	42201
	1	28.3 %	55.5 %	19665
	Σ	54031	7835	61866

Fig. 7. The confusion matrix actual and predicted results for HWs.

Confusion matrix for Tree (showing proportion of actual)

		Predicted 0	Predicted 1	Σ
Actual	0	99.4 %	0.6 %	44948
	1	97.9 %	2.1 %	8340
	Σ	52861	427	53288

Confusion matrix for Tree (showing proportion of predicted)

		Predicted 0	Predicted 1	Σ
Actual	0	84.5 %	60.0 %	44948
	1	15.5 %	40.0 %	8340
	Σ	52861	427	53288

Fig. 8. The confusion matrix actual and predicted results for RRs.

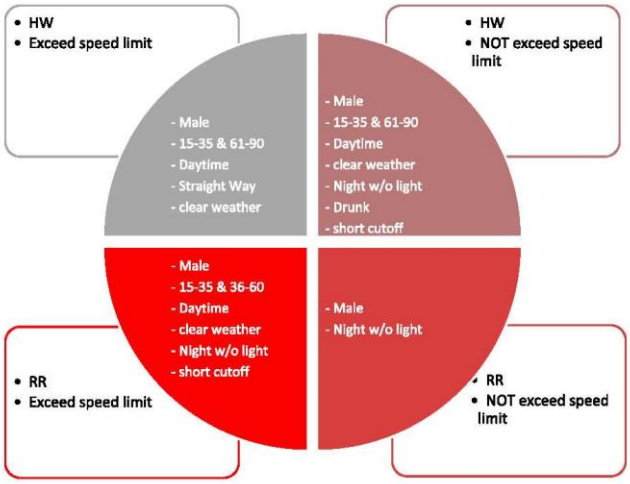

Fig. 9. Key accident factors: HWs versus RRs.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abellán, J., López, G., de Oña, J., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Syst. Appl. 40 (15), 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027.

Al Mamlook, R.E., Ali, A., Hasan, R.A., Mohamed Kazim, H.A., 2019. Machine Learning to Predict the Freeway Traffic Accidents-Based Driving Simulation. Proceedings of the IEEE National Aerospace Electronics Conference.

Anvari, M.B., Tavakoli Kashani, A., Rabieyan, R., 2017. Identifying the most important factors in the at-fault probability of motorcyclists by data mining, based on classification tree models. Int. J. Civil Eng. 15 (4), 653–662. https://doi.org/10.1007/s40999-017-0180-0.

Bahiru, T.K., Kumar Singh, D., Tessfaw, E.A., 2018. Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity. Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018.

Ben-David, S.S.-S.A.S. (2014). <understanding-machine-learning-theory-algorithms.pdf>. Cambridge University Press. http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning.

Bhavsar, R., Amin, A., Zala, L., 2021. Development of model for road crashes and identification of accident spots [Article]. Int. J. Intell. Transp. Syst. Res. 19 (1), 99–111. https://doi.org/10.1007/s13177-020-00228-z.

Bucsuházy, K., Matuchová, E., Zůvala, R., Moravcová, P., Kostíková, M., Mikulec, R., 2020. Human factors contributing to the road traffic accident occurrence. Transportation Research Procedia.

Champahom, T., Jomnonkwao, S., Chatpattananan, V., Karoonsoontawong, A., Ratanavaraha, V., 2019. Analysis of rear-end crash on Thai highway: decision tree approach. J. Adv. Transp. 2019, 1–13. https://doi.org/10.1155/2019/2568978.

Chen, M.-Y., 2012. Comparing traditional statistics, decision tree classification and support vector machine techniques for financial bankruptcy prediction. Intell. Autom. Soft Comput. 18 (1), 65–73. https://doi.org/10.1080/10798587.2012.10643227.

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., Zupan, B., 2013. Orange: Data mining toolbox in python [Article]. J. Mach. Learn. Res. 14, 2349–2353. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84885599052&partnerID=40&md5=75d2df52a0c46b5ab58ab08e1576114e.

DLT, 2021. Department of Land Transportation. https://www.dlt.go.th/th/public-news/view.php?_did=2806.

El Abdallaoui, H.E.A., El Fazziki, A., Ennaji, F.Z., Sadgal, M., 2018. Decision Support System for the Analysis of Traffic Accident Big Data. Proceedings - 14th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2018.

Feng, M., Zheng, J., Ren, J., Xi, Y., 2020. Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK. In Advances in Brain Inspired Cognitive Systems (pp. 520-529). 10.1007/978-3-030-39431-8_50.

Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests [Article]. Accid. Anal. Prev. 41 (1), 98–107. https://doi.org/10.1016/j.aap.2008.09.009.

Helen, W.R., Almetu, N., Nivethitha, S., 2019. Mining Road Accident Data Based on Diverted Attention of Drivers. Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018.

Jafari Anarkooli, A., Hosseinpour, M., Kardar, A., 2017. Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model. Accid. Anal. Prev. 106, 399–410. https://doi.org/10.1016/j.aap.2017.07.008.

Jomnonkwao, S., Uttra, S., Ratanavaraha, V., 2020. Forecasting road traffic deaths in Thailand: applications of time-series, curve estimation, multiple linear regression and path analysis models. *Sustainability* 12 (1). https://doi.org/10.3390/su12010395.

Jou, R.C., Yeh, T.H., Chen, R.S., 2012. Risk factors in motorcyclist fatalities in Taiwan. Traffic Inj Prev 13 (2), 155–162. https://doi.org/10.1080/15389588.2011.641166.

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accid. Anal. Prev. 37 (5), 910–921. https://doi.org/10.1016/j.aap.2005.04.009.

Kim, J.-K., Ulfarsson, G.F., Kim, S., Shankar, V.N., 2013. Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. Accid. Anal. Prev. 50, 1073–1081. https://doi.org/10.1016/j.aap.2012.08.011.

Krull, K.A., Khattak, A.J., Council, F.M., 2000. Injury effects of rollovers and events sequence in single-vehicle crashes. Transp. Res. Rec. 1717 (1), 46–54. https://doi.org/10.3141/1717-07.

Kumar, S., Toshniwal, D., 2016. A data mining approach to characterize road accident locations [Article]. J. Modern Transp. 24 (1), 62–72. https://doi.org/10.1007/s40534-016-0095-5.

Kuşkapan, E., Çodur, M.Y., Atalay, A., 2021. Speed violation analysis of heavy vehicles on highways using spatial analysis and machine learning algorithms [Article]. Accid. Anal. Prev. 155, 106098 https://doi.org/10.1016/j.aap.2021.106098.

Mafi, S., AbdelRazig, Y., Doezy, R., 2018. Machine learning methods to analyze injury severity of drivers from different age and gender groups. Transp. Res. Rec. 2672, 171–183.

Malin, F., Norros, I., Innamaa, S., 2019. Accident risk of road and weather conditions on different road types. Accid. Anal. Prev. 122, 181–188. https://doi.org/10.1016/j.aap.2018.10.014.

Mphela, T., 2020. Causes of road accidents in botswana: An econometric model [Article]. J. Transp. Supply Chain Manage., 14, 1-8, Article a509. 10.4102/jtscm.v14i0.509.

Osman, M., Mishra, S., Paleti, R., 2018. Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: accounting for unobserved heterogeneity and age group differences. Accid. Anal. Prev. 118 https://doi.org/10.1016/j.aap.2018.05.004.

Ospina-Mateus, H., Quintana Jiménez, L. A., López-Valdés, F. J., Morales-Londoño, N., Salas-Navarro, K., 2019. Using Data-Mining Techniques for the Prediction of the

Severity of Road Crashes in Cartagena, Colombia. In Communications in Computer and Information Science (Vol. 1052, pp. 309-320).
Ospina-Mateus, H., Quintana Jiménez, L.A., Lopez-Valdes, F.J., Berrio Garcia, S., Barrero, L.H., Sana, S.S., 2021. Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists [Article]. J. Ambient Intell. Hum. Comput. 12 (11), 10051–10072. https://doi.org/10.1007/s12652-020-02759-5.
Pakgohar, A., Tabrizi, R.S., Khalili, M., Esmaeili, A., 2011. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Comput. Sci. 3, 764–769. https://doi.org/10.1016/j.procs.2010.12.126.
PDPM, 2020. Thailand Department of Public Disaster Prevention and Mitigation. https://www.disaster.go.th/en/.
Recal, F., Demirel, T., 2021. Comparison of machine learning methods in predicting binary and multi-class occupational accident severity [Article]. J. Intell. Fuzzy Syst. 40 (6), 10981–10998. https://doi.org/10.3233/JIFS-202099.
Rezapour, M., Mehrara Molan, A., Ksaibati, K., 2020. Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. Int. J. Transp. Sci. Technol. 9 (2), 89–99. https://doi.org/10.1016/j.ijtst.2019.10.002.
RSC, T., 2019. Thailand Accident Research Center Thailand Accident Research Center https://www.thairsc.com/.
Se, C., Champahom, T., Jomnonkwao, S., Chaimuang, P., Ratanavaraha, V., 2021. Empirical comparison of the effects of urban and rural crashes on motorcyclist injury severities: A correlated random parameters ordered probit approach with heterogeneity in means. Accid. Anal. Prev. 161, 106352 https://doi.org/10.1016/j.aap.2021.106352.
Shaheed, M.S., Gkritza, K., Zhang, W., Hans, Z., 2013. A mixed logit analysis of two-vehicle crash severities involving a motorcycle. Accid. Anal. Prev. 61 https://doi.org/10.1016/j.aap.2013.05.028.
Siskind, V., Steinhardt, D., Sheehan, M., O'Connor, T., Hanks, H., 2011. Risk factors for fatal crashes in rural Australia. Accid. Anal. Prev. 43 (3), 1082–1088. https://doi.org/10.1016/j.aap.2010.12.016.
Song, Y.-Y., Ying, L., 2015. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry 27 (2), 130.
WHO. (2018). World Health Organization: Global status report on road safety 2018. https://extranet.who.int/roadsafety/death-on-the-roads/.
Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D., 2007. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14 (1), 1–37. https://doi.org/10.1007/s10115-007-0114-2.
Xie, Y., Huynh, N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accid. Anal. Prevent. 47, 36–44. https://doi.org/10.1016/j.aap.2011.12.012.
Yu, M., Zheng, C., Ma, C., 2020. Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. Anal. Methods Accid. Res. 27, 100126 https://doi.org/10.1016/j.amar.2020.100126.
Zhang, X.F., Fan, L., 2013. A decision tree approach for traffic accident analysis of Saskatchewan highways. Canadian Conference on Electrical and Computer Engineering.
Zhou, M., Chin, H.C., 2019. Factors affecting the injury severity of out-of-control single-vehicle crashes in Singapore. Accid. Anal. Prev. 124, 104–112. https://doi.org/10.1016/j.aap.2019.01.009.