Thailand Statistician
January 2025; 23(1): 129-143
http://statassoc.or.th
Contributed paper

Impact of COVID-19 Pandemic on Road Traffic Accident Severity in Thailand: An Application of K-Nearest Neighbor Algorithm with Feature Selection Techniques

Teerawat Simmachan [a, b], Sangdao Wongsai [a, c], Rattana Lerdsuwansri [a] and Pichit Boonkrong [d]*

[a] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand.
[b] Thammasat University Research Unit in Statistical Theory and Applications, Thammasat University, Pathum Thani, Thailand.
[c] Thammasat University Research Unit in Data Learning, Thammasat University, Pathum Thani, Thailand.
[d] College of Biomedical Engineering, Rangsit University, Pathum Thani, Thailand.
*Corresponding author; e-mail: pichit.bk@rsu.ac.th

Received: 10 June 2024
Revised: 7 August 2024
Accepted: 21 August 2024

Abstract

This study aims to develop road crash severity classifiers utilizing available government data from Thailand, specifically focusing on the period of the COVID-19 outbreak and possible factors. Three primary machine learning algorithms including logistic regression, random forest, and K-Nearest Neighbor (KNN) were utilized. Focusing on factors affecting accident severity, the feature importance was analyzed by stepwise, mean decrease in accuracy and mean decrease in impurity selection techniques. Customizing the three ML models and three feature selection techniques, nine different predictive models were built and evaluated based on accuracy, precision, recall, and F1-score. The results indicated that KNN with feature selections outperform candidate models, particularly KNN-MDA and KNN-MDI for pre-pandemic and during pandemic periods, respectively. Among the eight features, vehicle type was the most important factor causing a higher number of fatal accidents, followed by region, crash type, weather, and time of incidence. That is, motorcycle riders and pedestrians are especially susceptible. Therefore, this study can aid practitioners in formulating effective management policies to enhance road safety.

Keywords: Classifiers, feature selection, KNN, machine learning, random forest, road safety.

1. Introduction

According to the 2018 World Health Organization road safety global situation report (World Health Organization 2018), it was found that road traffic deaths of the population worldwide are approximately 1.35 million people per year or 3,700 people per day. To reduce the number of fatal

accidents, more understanding on relevant factors should be implemented accordingly with proper education, traffic policy and law.

Road accidents are a major global problem. Thailand has one of the world's greatest rates of road traffic accidents and its road traffic fatality rate was approximately 22.7 per 100,000 inhabitants in 2018. Besides, Thailand's road traffic fatality rate ranks ninth in the world with approximately 32.7 fatalities per 100,000 population or an average of 22,491 deaths per year (60 people per day). Thus, road accidents have become a major public health concern in Thailand. Factors expected to affect the number of road accidents include the number of vehicles on the road increasing rapidly, environmental, driving vehicle and driving behavior factors. During the outbreak of COVID-19, the lives and driving behavior of people around the world changed. To reduce the spread of COVID-19, the Thai government has launched public policies such as lockdown of public places, quarantine of infected cases and social distancing so that people worked from home and students had online class. As a results, there was a decrease in road accidents in Thailand during the COVID-19 outbreak. However, Thailand remains the country with the highest death toll in Asia. Predicting road traffic accident (RTA) severity is crucial for preventing injuries, fatalities, and economic losses. K-Nearest Neighbors (KNN) classifiers, known for their simplicity and interpretability, have emerged as a promising tool for this task. Several studies have explored the effectiveness of KNNs for RTA severity prediction. Vaiyapuri and Gupta (2021) achieved an accuracy of 85% in predicting accident severity in India, demonstrating KNN's potential. Fiorentini and Losa (2020) further improved the true positive rate in Great Britain by incorporating resampling techniques with KNN. However, KNN's performance can be sensitive to the choice of k and feature selection. Zhang et al. (2022) tackled this by employing hybrid approaches, combining KNN with an intrinsic wrapper-based feature selection approach for better accuracy. Despite their advantages, KNNs can suffer from the curse of dimensionality and computational inefficiency with large datasets. Unlike LR and RF models, it is noted that KNN does not have its own feature selection mechanism (Guyon and Elisseeff 2003; Hastie et al. 2009; Pedregosa et al. 2011). All features are treated as equally important without any built-in method to prioritize or select specific features. Mansoor et al. (2020) addressed this by incorporating dimensionality reduction techniques like principal component analysis (PCA). Additionally, Bokaba et al. (2022) used KNN adjusting missing values to improve performance for RTA severity prediction. Regarding the level of measurement of features utilized in this study, all of them are qualitative so that this work was motivated to integrate the KNN algorithm with the more suitable feature selection techniques.

Recognizing the severity of road accidents in Thailand, particularly during COVID-19, this study utilizes mathematical, statistical models, and machine learning techniques to investigate crucial factors influencing accident severity. The aim is to develop an effective algorithm based on KNN for predicting accident severity, ultimately reducing injuries, fatalities, and economic losses.

2. Machine Learning Framework

Unveiling the severity of road accidents is crucial for proactive safety measures. This study employs a data-driven approach, data description firstly paints the initial picture, exploring accident factors like causes of accident, crash types, region where the incident occurred, vehicle type, etc. Ensuring the quality and readiness of data for further analysis, data pre-processing was meticulously performed. Then, classification algorithms were designed as machine learning from the data, identifying patterns that predict accident severity. Selecting the best algorithm and model architecture for predicting road traffic accident severity in Thailand, model evaluation metrics were finally evaluated.

2.1. Data description

The RTA data utilized in this study consisted of accident open data on the highway network in Thailand (Open Government Data of Thailand 2020) from 2019 to 2021. The data was obtained from the Department of Highways, Ministry of Transport. The network has a total distance of 52,097 kilometers, reaching all provinces of Thailand. To combine machine learning methodologies with transportation data to provide a novel perspective on the effects of a global health crisis on national infrastructure, this study focuses on two distinct periods: 2019, representing the pre-pandemic period, and 2020-2021, encompassing the COVID-19 pandemic.

Table 1 Variable definitions and descriptive statistics of RTAs in Thailand for 2019-2021

Variable	Description	2019: Pre-Pandemic Period RTAs (%)	2020-2021: Pandemic Period RTAs (%)
Accident type (ACT)	0: Non-fatal accident	14,017 (86.34)	14,686 (89.16)
	1: Fatal accident	2,217 (13.66)	1,785 (10.84)
Cause of accident (COA)	1: Caused by a person	14,665 (90.34)	15,034 (91.28)
	2: Natural causes	1,354 (8.34)	1,340 (8.14)
	3: Others	215 (1.32)	97 (0.58)
Crash type (CRT)	1: Collision	15,674 (96.55)	16,056 (97.48)
	2: Overturned	547 (3.45)	415 (2.52)
Region of incidence (REG)	1: North	3,748 (23.09)	3,995 (24.25)
	2: Central	5,207 (32.07)	4,956 (30.09)
	3: East	1,950 (12.01)	1,908 (11.58)
	4: Northeast	3,102 (19.11)	3,269 (19.85)
	5: South	2,227 (13.72)	2,343(14.23)
Road type (ROT)	1: Rural Road	1,372 (8.45)	52 (0.32)
	2: National highway	14,862 (91.55)	16,419 (99.68)
Vehicle type (VHT)	1: Two-wheeler	3,516 (21.66)	2,959 (17.96)
	2: Three-wheeler	4,763 (29.34)	4,235 (25.71)
	3: Four-wheeler	5,689 (35.04)	6,564 (39.85)
	4: Others	2,266(13.96)	2,713 (16.47)
Road Section (ROS)	1: Straight with slope	114 (0.70)	119 (0.72)
	2: Straight	12,407 (76.43)	13,081 (79.42)
	3: Curve with slope	801(4.93)	751 (4.56)
	4: Curve	2,285 (14.08)	2,142 (13.01)
	5: Others	627 (3.86)	378 (2.29)
Weather (WEA)	1: Clear	14,014 (86.33)	13,969 (84.81)
	2: Others	2,220 (13.67)	2,502 (15.19)
Time of incidence (TIM)	1: Day	9,130 (56.24)	9,554 (58.00)
	2: Night	7,104 (43.76)	6,917 (42.00)

2.2. Data pre-processing

The data were divided into two situations: RTAs before and during the COVID-19 epidemic. RTAs for 2019 and 2020-2021 were defined before and during the COVID-19 pandemic, respectively. One of the study's key goals is to investigate factors influencing the severity of road accidents (fatal and non-fatal) in Thailand under two scenarios. Table 1 shows relevant variable definitions and their descriptive statistics. The accident type is the binary response variable, and the other eight categorical variables are the independent variables or features. It is noted that even after two years (2020-2021), COVID-19 pandemic RTAs are slightly lower than before the pandemic. There are 9 variables in this as shown in Table 1. The accident type is defined as response variable (Y) with binary values, i.e., 0 implies non-fatal accident while 1 denotes fatal accident. The other variables are predictors or features used to create predictive models.

2.2.1 Data cleansing and transformation

Prior to incorporating data into predictive models, meticulous data preparation is indispensable. This crucial stage entails guaranteeing data conformity to processing specifications of machine learning models and meticulously addressing missing values. A critical step in the machine learning pipeline is data preprocessing, specifically data transformation. This process refines raw data into a suitable format for analysis, enhancing model interpretability, mitigating outlier influence, and optimizing computational efficiency. In the context of categorical features, dummy variable encoding was employed. This technique addresses the limitations of certain models that require numerical inputs by converting qualitative features into binary vectors of 0s and 1s, effectively representing the presence or absence of each category. By undertaking this comprehensive data cleansing and transformation, the effectiveness and generalizability of subsequent model inferences are demonstrably enhanced.

2.2.2 Examining multicollinearity

The proposed classification models, particularly the LR model, necessitated an assessment of multicollinearity among the predictor variables. This was due to the model's sensitivity to highly correlated features, which can inflate standard errors and hinder interpretability. Therefore, feature correlations were meticulously examined. Correlations exceeding 0.70 were deemed problematic, prompting the removal of the corresponding features from further analysis. For evaluating the relationships between categorical features, Cramer's V coefficient was employed.

2.2.3 Handling imbalanced data

When applying classification, "imbalanced data" means that there are different numbers of observations for each group (class) of a response variable. Since imbalanced datasets negatively affect classification accuracy, handling imbalance in data can be done through undersampling so that the performance of ML model can also be improved (Hasanin et al. 2019; Miao and Zhu 2022; Tyagi and Mittal 2019). Regarding the dataset in this study, the road accident dataset is imbalanced, i.e., the majority class (non-fatal accident) has a lead over the minority class (fatal accident). ML models have a tendency to favor the majority class, which leads to unreliable performance (Kotb and Ming 2021; Simmachan et al., 2023). There are several methods to balance classes, e.g., data-based and algorithm-based focusing on data modification or learning algorithm modification (Mathew 2022). In this study, a stratified random sampling under the data-based methods was implemented to handle the imbalance issue (Moon et al. 2019; Simmachan et al. 2023). To appraise the model performance, the training and testing sets were created from the two datasets, separately. Balancing the RTA dataset for the free-pandemic period (2019), there were 14,017 non-fatal cases and 2,217 fatal cases. The training set was established by randomly selecting the arbitrary 3,000 observations using stratified random sampling. After randomly selecting 1,500 observations from 14,017 non-fatal cases, 1,500

were randomly drawn from 2,217 fatal cases. The testing set was created by the rest of observations, and it contained 12,517 non-fatal and 717 fatal cases, totaling 13,234. Similarly for the pandemic period (2020-21), 1,500 observations in each class were randomly picked to form the training set, and the testing set were sequentially created by the remaining observations. Finally, the testing set included a total of 13,471 observations with 13,186 non-fatal and 285 fatal cases.

2.3. Machine learning algorithms

Applying machine learning algorithms in conjunction with feature selection techniques, this study aims to uncover insights into the impact of the pandemic on Thailand's highway system and potentially identify patterns or anomalies in road usage and infrastructure demands. Unveiling the hidden features of traffic accidents, this study explores three machine learning models including LR, RF, and KNN. Delving into traffic incidence, the key features were identified and their impact was modeled. The three machine learning models are described as follows.

2.3.1 Logistic regression (LR)

The LR is a statistical model used to represent a categorical response variable, which can be influenced by either numerical or categorical independent variables or predictors, and it is commonly used for classification tasks (Cessie and Houwelingen 1992). The LR model is a member of generalized linear models (GLMs), and it uses the logistic function (sigmoid function) to convert the linear combination of predictors into a probability between 0 and 1 (Agresti 2013; Harrell 2015; Shu et al. 2018). In cases where the response variable has two possible values (0 and 1), the model is called a binary LR, and it forecasts the probability that an observation belongs to a specific category using one or more predictors. In our case, the response variable Y represents the accident type; where $y=1$ signifies a fatal accident and $y=0$ indicates a non-fatal accident corresponding with the eight predictors or features described in Section 2.1. The LR model can be expressed as

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{X}_i'\boldsymbol{\beta}\ , \tag{1}$$

where π_i is the interest event probability, denoted by $P(Y=1)$, or the likelihood of the accident resulting in a fatality; $\mathbf{X}_i'$ is the vector of predictors; and $\boldsymbol{\beta}=[\beta_0,\beta_1,...,\beta_p]^T$ is the vector of coefficients. The following is a direct transformation of (1) in terms of π_i;

$$\pi_i = \frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i'\boldsymbol{\beta})}.$$

2.3.2 Random forest (RF)

The RF algorithm was proposed by Breiman (2001), and it was also developed by Liaw and Wiener (2002) and Cutler et al. (2007). The RF algorithm, one of the most popular tree-based methods, consists of multiple individual decision trees operating collectively as an ensemble. In the initial step, *n* bootstrap samples were produced from the training set to create *n* decision trees. Then, at each note in each bootstrap sample, $m=\sqrt{p}$ predictors or features were randomly chosen from total number of features (p) based on feature split criteria or feature importance scores (James et al. 2021). Each tree generates a class prediction, and the RF model selects the class with the highest number of votes as the final outcome (James et al. 2021; Kowshalya and Nandhini 2018; Simmachan et al. 2023). The feature split criteria are related to a classification error rate (E). This error rate is simply the fraction of training observations in that region that are not in the most common class. It can be written as

$$E = 1 - \max_k \left(\hat{p}_{mk} \right),$$

where $\hat{p}_{mk}$ denotes the proportion of k^{th}-class training observations in the m^{th} region. Nevertheless, the error rate is not sufficiently sensitive for tree-growing, hence two other measurements are preferable (James et al. 2021). The first measurement is Gini index (G), and it can be defined as

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk} \right).$$

The Gini index measures the total variance across the K classes, and it takes on a small value if all of the $\hat{p}_{mk}$'s are close to zero or one. Consequently, this index is referred to as a measure of node purity. A low value suggests that a node consists primarily of observations from a single class. Another alternative measure of the error rate is cross-entropy (C), and it can be expressed as

$$C = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

Because the value of $\hat{p}_{mk}$ lines between zero and one, it follows that $0 \le -\hat{p}_{mk} \log \hat{p}_{mk}$, and the cross-entropy takes on a value is close to zero if the $\hat{p}_{mk}$'s are all near zero or one. Therefore, the Gini index and the cross-entropy are quite similar numerically.

2.3.3 K-nearest neighbor (KNN)

The KNN algorithm anticipates adjacent similar data points. Thus, it classifies an observation by examining the closest k data points using a distance approach and a member's classification by its neighbors' majority vote (Prasasti et al. 2020). Euclidean distance utilized to quantify similarity (Wang et al. 2007; Boonkrong and Simmachan 2016; Farghaly at al. 2023) can be evaluated by

$$D(p_i, q_j) = \sqrt{\sum_{k=1}^{n} \left(p_{ik} - q_{jk} \right)^2},$$

where p and q are subjects to be compared with n characteristics.

2.4. Feature selection techniques

Choosing relevant and useful features from a bigger set is feature selection. The approach saves computing time, improves model performance, avoids overfitting, and improves data pattern representation, especially in high-dimensional datasets (Kotb and Ming 2021; Vanishkorn and Supanich 2022). In this study, four sets of features were determined as follows:

2.4.1 Full model

In this scenario, the predictive models were created using all features, or they were formulated without feature selection. All features were immediately inserted into LR, RF, and KNN algorithms. Totally, there are three experimental scenarios from this setting as shown in Table 2.

2.4.2 Stepwise procedure

The binary LR via a stepwise selection method was utilized to investigate the important features. Forward selection and backward elimination are combined in stepwise selection (Mauša 2012). This procedure assesses whether the existing features in the model should be eliminated. The Akaike Information Criterion (AIC) improvement is used to determine model feature inclusion and exclusion since the AIC is a common model selection tool. The AIC is popular since it evaluates both model quality of fit and complexity. Additionally, it may compare non-nested models with different parameters.

2.4.3 Mean decrease in accuracy (MDA)

The MDA is an indicator used to determine the importance of features for RF model (Breiman 2001). This method assesses feature importance by estimating the mean decrease in out-of-bag accuracy for each tree (Baek et al. 2008). The objective is to determine the importance of a feature by measuring how much the model's accuracy declines when its values are randomly shuffled, breaking any significant relationship between the feature and the response (Moon et al. 2019). Greater accuracy decline indicates greater feature importance (Breiman 2001; Liaw and Wiener 2002).

2.4.4 Mean decrease in impurity (MDI)

Another common indicator for evaluating variable importance for RF model is known as MDI. The Gini index described in Section 2.3.2 was used. At each node of a decision tree, the algorithm chooses the split that minimizes impurity (Tan et al. 2006; Baek et al. 2008; Moon et al. 2019). MDI measures a feature's overall impurity decrease across all forest trees. What it does calculates how much each feature helps to tree node purity. Features that reduce impurity most are considered more important (Breiman 2001; Liaw and Wiener 2002). To explore and optimize predictive performance, this study investigates the combined application of three distinct ML algorithms and four established feature selection techniques. As shown in Table 2, these methods were evaluated under various experimental scenarios to identify the most effective combination for improving predictive performance.

Table 2 Description of classification algorithms used for predicting RTA severity

Algorithm	Description
LR-Full	Logistic regression model without feature selection.
LR-Step	Logistic regression model with stepwise selection of features.
RF-Full	Random forest model without feature selection.
RF-MDA	Random forest model using features obtained from MDA index.
RF-MDI	Random forest model using features obtained from MDI index.
KNN-Full	K-Nearest neighbor model without feature selection.
KNN-Step	K-Nearest neighbor model using features obtained from stepwise.
KNN-MDA	K-Nearest neighbor model using features obtained from MDA index.
KNN-MDI	K-Nearest neighbor model using features obtained from MDI index.

2.5. Model evaluation

In machine learning, evaluating model performance is crucial. This study delves into the confusion matrix, its role in understanding predictions. By exploring these metrics, valuable insights into a classifier's strengths and weaknesses are obtained, enabling informed decision-making for real-world applications.

2.5.1 Confusion matrix

A confusion matrix is commonly employed as an assessment metric in classification tasks (Simmachan et al. 2022; Yilmaz and Demirhan 2023; Akarajarasroj et al. 2023; Na Bangchang et al., 2023). In binary classification problems dealing with road accident severity, the positive class represents fatal accidents, and the negative class depicts non-fatal accidents. The matrix shows predicted and actual outcome counts, and four distinct outcomes are employed. TP reflects the number of fatal accidents correctly identified by the model. TN denotes the number of non-fatal accidents correctly identified by the model. FP refers to the number of non-fatal accidents that are mistakenly classified as fatal accidents. FN presents the number of fatal accidents that are mistakenly classified as non-fatal accidents.

2.5.2 Evaluation metrics

The effectiveness of the proposed approach can be assessed using widely employed evaluation criteria including accuracy, precision, recall and F1-score. The higher performance of model can be mimicked through the high value of each metric. Each metric is computed as follows.

- Accuracy, a popular metric, represents the overall predictive performance of the models, and it indicates overall correct prediction.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{7}$$

- Precision indicates the ratio of true positive predictions to the total number of positive predictions made by the models.

$$\text{precision} = \frac{TP}{TP + FP}. \tag{8}$$

- Recall, also called sensitivity, represents the correct rate of predicting the positive class.

$$\text{recall} = \frac{TP}{TP + FN}. \tag{9}$$

- F1-score is computed by precision and recall simultaneously and is the harmonic mean of both metrics. It presents the model's overall performance of the positive class.

$$\text{F1-score} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}. \tag{10}$$

The four evaluation metrics were used to measure how effective the models given in Table 2 have performed.

3. Numerical Results

The investigation was conducted on collinearity, feature importance scores, and model performance across nine ML classifiers to predict RTA severity. Numerical results indicate varying levels of collinearity and feature impact depending on the algorithms, underscoring the importance of evaluating multiple classifiers to optimize predictive capability. More specific and interesting results are illustrated in the following subsections.

	COA	ROS	VHT	ROT	CRT	WEA	REG	TIM
COA	1.000	0.182	0.145	0.110	0.134	0.069	0.047	0.087
ROS	0.182	1.000	0.106	0.263	0.209	0.197	0.116	0.073
VHT	0.145	0.106	1.000	0.292	0.123	0.152	0.085	0.032
ROT	0.110	0.263	0.292	1.000	0.243	0.025	0.043	0.007
CRT	0.134	0.209	0.123	0.243	1.000	0.020	0.020	0.005
WEA	0.069	0.197	0.152	0.025	0.020	1.000	0.111	0.001
REG	0.047	0.116	0.085	0.043	0.020	0.111	1.000	0.077
TIM	0.087	0.073	0.032	0.007	0.005	0.001	0.077	1.000

	COA	ROS	VHT	ROT	CRT	WEA	REG	TIM
COA	1.000	0.226	0.168	0.036	0.185	0.073	0.047	0.080
ROS	0.226	1.000	0.104	0.029	0.375	0.156	0.120	0.071
VHT	0.168	0.104	1.000	0.035	0.082	0.171	0.086	0.029
ROT	0.036	0.029	0.035	1.000	0.088	0.006	0.022	0.011
CRT	0.185	0.375	0.082	0.088	1.000	0.043	0.023	0.000
WEA	0.073	0.156	0.171	0.006	0.043	1.000	0.121	0.012
REG	0.047	0.120	0.086	0.022	0.023	0.121	1.000	0.060
TIM	0.080	0.071	0.029	0.011	0.000	0.012	0.060	1.000

Figure 1 The Cramer's V coefficients examining collinearity among categorical features: pre-pandemic period (left); during pandemic period (right)

3.1. Investigation on Collinearity

Firstly, the collinearity among the 8 categorical features needed to be examined to meet the LR

assumption that predictor variables are not highly correlated (Hosmer et al. 2013; Tabachnick and Fidell 2013; Field 2017; Schreiber-Gregory 2018). Collinear categorical variables provide redundant information, causing inflated standard errors and instability in severity prediction models. Identifying and removing correlated categorical features improves model fit, accuracy, and generalizability for classifying Thailand traffic accident severity. The Cramer's V coefficient was utilized in screening the collinearity among the 8 categorical features during pre-pandemic and pandemic periods. The correlation matrices presented in Figure 1 indicate that all calculated Cramer's V values were positive and remained below 0.5. This observation suggests that multicollinearity among the eight categorical features was not a significant concern.

3.2. Feature Importance Score

Prioritizing the most important to least important variables, the stepwise, MDA and MDI selection techniques were applied. Figure 2 exhibits the feature importance scores based on the three selection techniques and their accuracy from the learning set. For pre-pandemic period, VHT, TIM, WEA, REG, CRT and COA were included and ranked by their Wald statistic values. During the outbreak of COVID-19, VHT, TIM, REG, WEA, ROT and CRT were selected into the LR model. The features ROS and ROT were removed from the LR model for pre-pandemic period while COA and ROS were removed during the pandemic period. The MDA and MDI variable importance plots show similar rankings on the most important features. Obviously, VHT presented its highest score among all 8 features, i.e., the vehicle type is considered as the most important feature towards minimizing the error rate in classification.

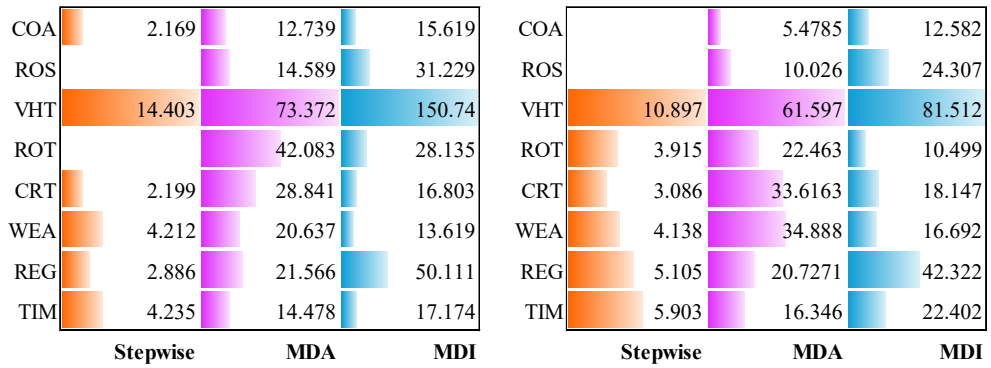

Figure 2 Feature importance scores based on stepwise, MDA and MDI selections: pre-pandemic period (left); during pandemic period (right)

3.3. Model performance

Regarding the evaluation metrics in Table 3, it was found that the KNN-MDA and KNN-MDI models presented their best performances in predicting road traffic accident severity in pre-pandemic and pandemic periods. Focusing on the fatal accidents, accuracy and F1-Score were able to summarize the overall performance, i.e., the KNN-MDA model presented the highest accuracy of 81.69% and F1-Score of 89.95% in pre-pandemic period. Similarly, the KNN-MDI model demonstrated the highest accuracy of 80.71% and F1-Score of 89.19% during the period of COVID-19 outbreak. Considering LR and RF models separately, the LR-Full and RF-MDI models were outstanding in both periods. To reduce noise, boost accuracy, and fasten learning, the feature importance was considered for removing irrelevant and redundant data. Table 3 has verified that the

combination of KNN algorithm and the feature important scores contributed the best performance among the 9 scenarios.

Table 3 Performance of nine sub-models

Model	2019: Pre-Pandemic Period				2020-2021: Pandemic Period			
	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LR-Full	74.48	75.27	97.11	84.81	63.96	63.95	98.82	77.65
LR-Step	74.04	74.88	97.02	84.52	61.29	61.25	98.72	75.59
RF-Full	76.22	76.96	**97.36**	85.97	63.25	63.32	98.65	77.14
RF-MDA	73.48	74.06	97.26	84.09	64.08	64.21	98.60	77.78
RF-MDI	78.11	79.18	97.16	87.25	64.92	65.08	98.61	78.41
KNN-Full	79.11	80.28	97.16	87.92	80.23	80.81	98.78	88.89
KNN-Step	76.62	77.50	97.24	86.25	78.14	78.73	98.67	87.58
KNN-MDA	**81.69**	**83.25**	96.96	**89.59**	78.14	78.73	98.67	87.58
KNN-MDI	78.35	79.45	97.16	87.41	**80.71**	**81.28**	**98.80**	**89.19**

4. Discussion

This study has successfully demonstrated how machine leaning approaches and feature selection were embedded for predicting the road traffic accident severity in Thailand during pre-pandemic and pandemic periods. Three main models including LR, RF and KNN were implemented jointly with three feature selection methods including stepwise, MDA and MDI. Based on the numerical results, there are three key aspects to be discussed in the following subsections.

4.1. Features and their mechanisms

Considering the impact of feature selection, the performances of classifiers without and with feature selection methods were evaluated. However, the LR model with stepwise selection was initially utilized by classical statistics approach reflecting the effect of features. Due to the interventional policy against the outbreak of COVID-19 such as work from home, online class, quarantine, social distancing and lockdown of public places, the traffic volume was reduced so that the annual prevalence of fatal accidents was also lower. In fact, all 8 features can influence the accident severity either directly or indirectly. The major causes of fatal accident were commonly from humans such as distracted driving, speeding, and alcohol impairment. From the findings in this study, motorcycles and older vehicles were strongly associated with higher fatality rates in Thailand. The region of incidence was the second one as the urban areas often had higher fatalities due to higher traffic volume. The collision and overturn were the common crash types when the weather and time of incidence were physically hazardous. Rain, smokey and fog could increase crash risks due to reduced visibility. Weekends and nights often saw more fatal accidents due to fatigue and impaired driving. Considering road type, high-speed roads and rural roads with limited infrastructure had higher risks. Road sections including intersections, curves, and poorly maintained roads posed greater dangers. Furthermore, incorporating the festive time factor (Lerdsuwansri et al. 2022; Taveekal et al. 2023) and demographic factors (Phaphan et al. 2023) could enhance model performance.

4.2. Classifiers and their performance

According to the empirical results all models performed slightly better pre-pandemic than pandemic for all metrics except recall. In terms of recall, the developed models are comparable in both periods. These consequences might result from the two different situations in RTAs. The features in a normal situation (pre-pandemic period) can assist to refine predictive models. Nevertheless, during the epidemic period, the existing features are unable to capture response variable relationships effectively. In other words, there would be hidden features in an anomalous situation. The LR statistical model without feature selection exhibits a marginal superiority over those with feature selection. In RF models, RF-MDI outperforms the others, then RF-Full and RF-MDA. However, their performances are approximately the same. In comparison to LR models, RF models showed a bit better overall performance. On average, KNN models outperform the other nine models. Specifically, the models have the potential to significantly enhance their ability to make accurate predictions during the pandemic period compared to LR and RF models. The performance of the original KNN or KNN without feature selection yields satisfactory results. Moreover, KNN models with feature selection outperform in different scenarios; KNN-MDA and KNN-MDI are the best models during free-pandemic and pandemic periods, respectively. Out of the four KNN models, KNN-Step appears to be the least effective. Nonetheless, it is superior to LR and RF models, especially during the pandemic period. Interestingly, the KNN combined with feature selection strategies in our proposed models increase the original version's prediction performance.

4.3. Balancing methods and model behavior

Although the resampling techniques can enhance the performance of the LR and ML models, they may create an autocorrelation problem (Fiorentini and Losa 2020). To be precise, when the original dataset was resampled, data points lacked of independence. The presence of autocorrelation can lead to a bias in the estimated coefficients leading to poor predictive accuracy and tend to alter the distribution of the training data, resulting in an overfitting of the learning model (Wan and Zhu 2022). Particularly to our study, the stratified random sampling was utilized to randomly select the arbitrary observations for training set whilst the rest of observations were allocated in testing set. Subsequently, the stratified random downsampling through the experimental designs as mentioned in Table 2 can improve performance with imbalanced datasets such as the RTA data in Thailand. Literally, novel algorithms like adaptive random sampling and classification tree have shown promise in handling imbalanced data (Dong et al. 2018; Miao and Zhu 2022; Simmachan et al. 2023). Unforgettably, not all resampling algorithms can boost the model performances; on the other hand, they can lead to bias model and lower model performances (Cavus and Biecek 2024; Stando et al. 2024; Goorbergh et al. 2022). Thus, the empirical results corroborate the notion that the evaluation of balancing techniques and dataset analysis should transcend mere comparisons of model performance metrics.

4.4. Real-world implementation

To reduce the traffic accidents as well as fatality rate, the driver demographics, vehicle condition, road condition and infrastructures should be safe or have minimal risk. The age, experience, driving expertise, familiarity with the route and readiness of drivers directly effect on road safety and fellow travelers. Well-maintained vehicles with advanced safety features (airbags, ABS, etc.) can absorb impact, prevent crashes, and reduce injury severity, potentially lowering accident fatalities. Sharing the traffic route in Thailand, there are not only cars, but also motorbikes, Tuk-Tuks, trucks and trailers. Car drivers and passengers should always fasten seatbelts and motorcycle drivers must wear helmets. Regarding the road condition, potholes, poor signage, missing guardrails, and weak

infrastructure can increase driver error, limit maneuverability, and offer less protection in crashes, creating a recipe for tragedy on the road. Improving road safety for everyone, effective traffic enforcement acts as a shield against accidents and fatalities. Clear traffic laws and strong penalties can deter risky behavior like racing, speeding or drunk driving. Consistent enforcement reinforces this message while educational campaigns and soft power through mass media and social media (TikTok, Facebook, Instagram, etc.) build awareness and positive driving habits.

5. Conclusions

There were nine sub- models in this experimental scenario and their performance were investigated. To avoid the collinearity problems among the eight categorical features, the Cramer's V coefficients were assessed, and it was found that the coefficients were less than 0.5 indicating no serious multicollinearity problems. The categorical features were converted to the dummy variables since ML algorithms can process only numerical inputs. Since the two datasets were imbalanced, a stratified random sampling was used for more precise outcomes. Each class was randomly selected from the whole dataset with equal observations to create the training sets; and the remaining observations were automatically formulated the testing sets. Then, the feature selection methods were utilized in the training set. Executing nine sub-models for classification problems, the KNN-MDA and KNN-MDI models presented their best performances. Since the KNN algorithm and the feature importance scores, MDA and MDI under RF algorithm are both nonparametric methods, their performance directly reflected their good combination. Ranking the common features influencing the traffic-related fatality, the vehicle type was ranked as number one, followed by region, crash type, weather and time of incidence. Therefore, this study has successfully achieved in applying KNN approach with feature selections for predicting the road traffic accident severity in Thailand during the COVID-19 pre-pandemic and pandemic periods. Implementing the findings from this study, the collaboration of the government, private and public sectors can create a safer environment, discouraging dangerous choices and encouraging responsible driving, ultimately leading to fewer and less severe accidents.

Acknowledgements

The authors would like to express gratitude to the anonymous referees for their informative suggestions to further improve the overall presentation of the manuscript. The authors gratefully acknowledge the financial support provided by the Faculty of Science and Technology, Contact No. SciGR 19/2566.

References

Agresti A. Categorical data analysis. 3rd ed. Hoboken, New Jersey: John Wiley and Sons; 2013.

Akarajarasroj T, Wattanapermpool O, Sapphaphab P, Rinthon O, Pechprasarn S, Boonkrong P. Feature selection in the classification of erythemato-squamous diseases using machine learning models and principal component analysis. BMEiCON 2023: Proceedings of the 15th Biomedical Engineering International Conference; 2023 Oct 28-31; Japan. Tokyo: IEEE; 2023. pp. 1-5.

Baek S, Moon H, Ahn H, Kodell RL, Lin CJ, Chen JJ. Identifying high-dimensional biomarkers for personalized medicine via variable importance ranking. J Biopharm Stat. 2008; 18(5): 853-868.

Bokaba T, Doorsamy W, Paul BS. Comparative study of machine learning classifiers for modelling road traffic accidents. Appl Sci. 2022; 12(2), https://doi.org/10.3390/app12020828.

Boonkrong P, Simmachan T. A Multigroup SEIR epidemic model with vaccination on heterogeneous network. Chiang Mai J Sci. 2016; 43(4): 897-903.

Breiman L. Random forests. Mach Learn. 2001; 45(1): 5-32.

Cessie SL, Houwelingen JCV. Ridge estimators in logistic regression. J R Stat Soc Ser C Appl Stat.1992; 41(1): 191-201.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. Random forests for classification in ecology. Ecology. 2007; 88(11): 2783-2792.

Dong LY, Wang YQ, Li YL, Zhu Q. Adaptive random sampling algorithm based on the balance maximization. J Northeast Univ Nat Sci, 2018; 39(6): 792-796.

Field A. Discovering statistics using IBM SPSS Statistics. 5th ed. Thousand Oaks, CA: Sage Publications; 2017.

Fiorentini N, Losa M. Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures. 2020; 5(7), https://doi.org/10.3390/infrastructures5070061.

Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003; 3: 1157-1182.

Harrell FE. Binary logistic regression. In: Regression Modeling Strategies. Springer Series in Statistics. Switzerland: Springer International Publishing; 2015. p. 219–274.

Hasanin T, Khoshgoftaar TM, Leevy JL, Seliya N. Examining characteristics of predictive models with imbalanced big data. J Big Data. 2019; 6(1): 1-21.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.

Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken, New Jersey: John Wiley and Sons; 2013.

James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2021.

Kotb MH, Ming R. Comparing SMOTE Family Techniques in Predicting Insurance Premium Defaulting using Machine Learning Models. Int J Adv Comput Sci Appl. 2021; 12(9): 621-629.

Kowshalya G, Nandhini M. Predicting fraudulent claims in automobile insurance. In ICICCT 2018: Proceedings of the Second International Conference on Inventive Communication and Computational Technologies (ICICCT); 2018 Apr 20-21; India. Coimbatore: IEEE; 2018. pp. 1338-1343.

Lerdsuwansri R, Phonsrirat C, Prawalwanna P, Wongsai N, Wongsai S, Simmachan T. Road traffic injuries in Thailand and their associated factors using Conway-Maxwell-Poisson regression model. Thai J Math. 2022; Special Issue (2022): IMT-GT International Conference on Mathematics, Statistics and Their Applications 2021: 240-249.

Liaw A, Wiener M. Classification and regression by random Forest. R News. 2002; 2(3): 18-22.

Mansoor U, Ratrout NT, Rahman SM, Assi K. Crash severity prediction using two-layer ensemble machine learning model for proactive emergency management. IEEE Access. 2020; 8: 210750-210762.

Mathew TE. Appositeness of Hoeffding tree models for breast cancer classification. J Curr Sci Technol. 2022; 12(3): 391-407.

Mauša G, Grbac TG, Bašić BD. Multivariate logistic regression prediction of fault-proneness in software modules. In MIPRO 2012: Proceedings of the 35th International Convention on Information and Communication Technology, Electronics and Microelectronics; 2012 May 21-25; Croatia. Opatija: IEEE; 2012. pp. 698-703.

Miao J, Zhu W. Precision–recall curve (PRC) classification trees. Evol Intell. 2022; 15(3): 1545-1569.

Moon H, Pu Y, Ceglia C. A predictive modeling for detecting fraudulent automobile insurance

claims. Theor Econ Lett. 2019; 9(6): 1886-1900.

Bangchang KN, Wongsai S, Simmachan T. Application of data mining techniques in automobile insurance fraud detection. In ICoMS 2023: Proceedings of the 2023 6th International Conference on Mathematics and Statistics; 2023 Jul 14-16; Germany. Leipzig; 2023. pp. 48-55.

Open Government Data of Thailand. Road Accident Data Set [Internet]. 2020 [cited 2023 Jan 29]. Available from: https://data.go.th/dataset/gdpublish-number-of-road-accidents-in-the-country

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011; 12: 2825-2830.

Phaphan W, Sangnuch N, Piladaeng J. Comparison of the effectiveness of regression models for the number of road accident injuries. Sci Technol Asia. 2023; 28(4): 54-66.

Prasasti IMN, Dhini A, Laoh E. Automobile insurance fraud detection using supervised classifiers. In IWBIS 2020: Proceedings of the 5th International Workshop on Big Data and Information Security (IWBIS); 2020 Oct 17-18; Indonesia. Depok: Institute of Electrical and Electronics Engineers Inc; 2020. pp. 47-51.

Stando A., Cavus M., Biecek, P. The effect of balancing methods on model behavior in imbalanced classification problems. In LIDTA 2023: Proceedings of the 5th International Workshop on Learning with Imbalanced Domains: Theory and Applications; 2023 Sep 18; Turin. Italy: Proceedings of Machine Learning Research 241; 2023. pp. 16-30.

Schreiber-Gregory DN. Multicollinearity: what is it, why should we care, and how can it be controlled? In: SAS Global Forum 2018; 2018 Apr 8-10; Proceedings of the SAS® Global Forum 2018 Conference. Cary, NC: SAS Institute Inc.; 2018. Paper 1404-2017.

Shu J, Tang Y, Cui J, Yang R, Meng X, Cai Z, et al. Clear cell renal cell carcinoma: CT-based radiomics features for the prediction of Fuhrman grade. Eur J Radiol. 2018; 109: 8-12.

Simmachan T, Manopa W, Neamhom P, Poothong A, Phaphan W. Detecting fraudulent claims in automobile insurance policies by data mining techniques. Thail Stat. 2023; 21(3): 552-568.

Simmachan T, Wongsai N, Wongsai S, Lerdsuwansri R. Modeling road accident fatalities with underdispersion and zero-inflated counts. PLoS One. 2022; 17(11): e0269022.

Tabachnick BG, Fidell LS. Using multivariate statistics. 6th ed. Boston, MA: Pearson; 2013.

Tan PN, Steinbach M, Kumar V. Introduction to data mining. Boston, MA: Pearson Addison Wesley; 2006.

Taveekal P, Rajchanuwong P, Wongwiangjan R, Lerdsuwansri R, Intrakul J, Simmachan T, Wongsai S. Modelling road accident injuries and fatalities in Suratthani Province of Thailand using Conway-Maxwell-Poisson regression. Thail Stat. 2023; 21(3): 569-579.

Tyagi S, Mittal S. Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning. In Proceedings of ICRIC 2019: Recent innovations in computing. Lecture Notes in Electrical Engineering Proceedings of ICRIC 2019. Springer International Publishing. 2019: 209-221.

Vaiyapuri T, Gupta M. Traffic accident severity prediction and cognitive analysis using deep learning. Soft Comput. 2021: 1-13.

Goorbergh RVD, Smeden MV, Timmerman D, Calster BV. The Harm of Class Imbalance Corrections for Risk Prediction models: Illustration and Simulation Using Logistic Regression. J Am Med Inform Assoc. 2022; 29(9): 1525-1534.

Vanishkorn B, Supanich W. Crash severity classification prediction and factors affecting analysis of highwayaccidents. In ICAICTA2022: Proceedings of the 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA); 2022 September 28-29; Japan. Tokoname: IEEE; 2022. pp. 1-6.

Wan J, Zhu S. Cross-city crash severity analysis with cost-sensitive transfer learning algorithm. Expert Syst Appl. 2022; 208(4), https://doi.org/10.1016/j.eswa.2022.118129.

Wang J, Neskovic P, Cooper LN. Improving nearest neighbor rule with a simple adaptive distance measure. Pattern Recogn Lett. 2007; 28(2): 207-213.

World Health Organization. Global status report on road safety 2018. 2018 [cited 2024 Jan 20]. Available from: https://books.google.co.th/books?hl=th&lr=&id=uHOyDwAAQBAJ&oi=fnd&pg=PR6&dq=World+Health+Organization.+(2018).+Global+status+report+on+road+safety+218.&ots=2T-m0zreWW&sig=FpsjOkITsJO1WEHWSeJI6aCH5R0&redir_esc=y#v=onepage&q&f=false.

Yilmaz AE, Demirhan H. Weighted kappa measures for ordinal multi-class classification performance. Appl Soft Comput. 2023; 134, https://doi.org/10.1016/j.asoc.2023.110020.

Zhang S, Khattak A, Matara CM, Hussain A, Farooq A. Hybrid feature selection-based machine learning classification system for the prediction of injury severity in single and multiple-vehicle accidents. PLoS One. 2022; 17(2), https://doi.org/10.1371/journal.pone.0262941.