Available online at https://li02.tci-thaijo.org/index.php/ssstj

Suan Sunandha Science and Technology Journal

A Risk Prediction Model of Road Accidents During Long Holiday in Thailand Using Ensemble Learning with Decision Tree Approach

Paranya Palwisut*

Department of Data Science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University,
85 Malaiman Road, Mueang Nakhon Pathom, Nakhon Pathom 73000, Thailand
***Corresponding author e-mail: paranya@npru.ac.th**

Received: *21 February 2023* / **Revised:** *11 May 2023* / **Accepted:** *13 June 2023*

Abstract

The rate of injury and death from traffic accidents during the New Year and Songkran Festival each year has high and are continuously on the increase. The researchers, therefore, has decided to study and develop a model for predicting the road accident risk during the holiday season with ensemble learning based on decision tree approach. The aim is to help reduce accidents and loss of life caused by road accidents. The dataset used in this research is traffic accidents resulting in injury and death data during the long holiday from 2008 to 2015 from hospitals across the country, accumulatively recorded by the National Institute for Emergency Medicine. This research compared the efficiency of data classification to find the best ensemble model for predicting traffic accident risk. The methods studied included Adaptive Boosting (AdaBoost), and Random Forest, and the decision tree techniques used in the experiment were J48, ID3, and CART. The results of experiment and comparisons of classification efficiency showed that the Random Forest algorithm with J48 decision tree was the most efficient model, providing an accuracy of up to 93.3%.

Keywords: Road accident, Decision tree, Ensemble learning

1. Introduction

Traffic accidents are a major problem in Thailand. There are a large number of victims of traffic accidents each year in the form of injury, disability, and fatality. Moreover, the incidence tends to become more increased. Road accidents are the leading cause of death in Thailand, particularly during long holiday. For the New Year and Songkran Festival, there are 7 dangerous days. This refers to the period when the statistic of road accidents, the number of casualties and injury has peaked to the highest point due to the highest rate of road travel (Sonwongsa, Pinpoo, & Wongkhae, 2016). According to the report of injuries and deaths from traffic accidents during the New Year and Songkran Festival, it was found that most of the accident victims were drivers and accompanying passengers, resulting in an increasing number of deaths. During the new year festival of 2022 (Accident Prevention Network, 2022), there were 2,707 road accidents, which killed 333 people and injured 2,672 others, from statistics found that there are still many accidents every year. Currently, various techniques have been used for road traffic accident data analysis, such as data mining algorithms. This process involves exploring large amounts of data to find patterns and relationships hidden in the dataset, with techniques and methods for predicting, classifying, and managing the data (Esenturk, Turley, Wallace, Khastgir, & Jennings, 2022; Parathasarathy, Soumya,Das, Saravanakumar, & Merjora, 2019). Because of the exponential rising number of road accidents on the New Year and Songkran Festival, the researchers decided to study and construct the risk prediction model of road accidents during the holiday season with ensemble learning method using decision tree approach. The researcher studied the theories and literatures related to this research as follows:

Suan Sunandha Science and Technology Journal

1.1 Decision Tree Technique

Decision Tree (Njoku, 2019) is a technique that presents outcomes in a tree-like graph. The data are partitioned based on their features travestying the tree structure to terminal class. The tree model is the collection of nodes. Each node denotes a test on a particular feature. Each branch denotes the possible value of the tested feature, while each leaf at the bottom of the decision tree denotes class labels, which is the outcome of prediction. The node at the top most of the tree is called root node. The structure of decision tree is shown in Figure 1.

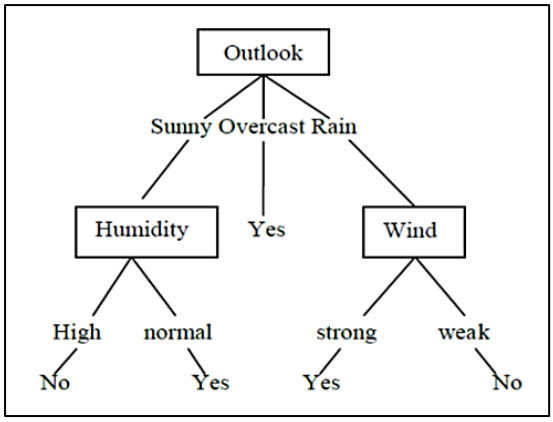

Figure 1. Decision Tree Structure (Njoku, 2019).

The decision tree measures the impurity of each features or variables as follows:

Gini Index is a value that indicates what features or variables should be used for classification algorithm J48 and CART.

$$Gini(t_i) = 1 - \sum_{i=1}^{N}[p(t_i)]^2 \qquad (1)$$

Entropy is the degree of uncertainty in the dataset which is classified based on the identification of ID3 algorithm,

$$Entropy(t_i) = 1 - \sum_{i=1}^{N}[p(t_i)]log_2 p(t_i) \qquad (2)$$

where

t_i is the feature selected to measure the entropy

$P(t_i)$ is proportion of number of group i members and the total number of sample group

Each algorithm gives different outcomes. The decision tree algorithms employed in this research are as follows:

- J48 or C 4.5 is an algorithm to create a tree model from a set of training data. The accuracy value of each data feature is used as a criterion to classify the data into subsets based on entropy values. The feature with the highest normalized information gain is chosen to make the decision (Panigrahi & Borah, 2018).

- Iterative Dichotomiser 3 or ID3 is an algorithm to generate a decision tree based on information gain. The data are split into subsets based on the entropy or the information gain of each feature. The decision tree is built according to selected features in order of gain value from high to low (Ogheneovo & Nlerum, 2020).

- Classification and Regression Trees (CART) is an algorithm to build a binary decision tree consisting of one or two branches for each node. This technique divides records of training data into subsets given the same target value (Zacharis, 2018).

1.2 Ensemble learning

Ensemble learning is a process by which multiple independent classifiers are combined or voted for decision making to boost the classification efficiency, as shown in Figure 2.

The following steps are performed:

(1) Generate new data from the original data

(2) Build classifiers from the generated data

(3) Combine the classifiers to find out the answer

There are many ensemble techniques available, but the most common methods used in this research are boosting and bagging.

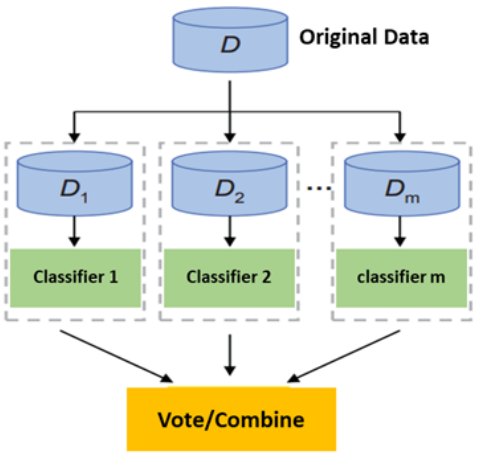

Figure 2. Basic structure of ensemble learning.

Suan Sunandha Science and Technology Journal

1.2.1 Boosting Method is a general ensemble method to build a model by weighing training samples. Boosting focuses on finding errors that arise from the learning process, called "weak learning". Finally in the last step, it combines the classifiers based on the mean weight and votes for a single accurate learner.

The adaptive boosting or AdaBoost algorithm proposed by Freund and Schapire in 1997 is one of the most widely used. Initially, all weights of each instance in the training datasets are set equally. On each round the weights of correct instances are lessen, while more weight is given to incorrectly classified instances to increase a chance of distribution in the next round (Tanha, Abdi, Samadi, Razzaghi, & Asadpour, 2020). Boosting Method is shown in Figure 3.

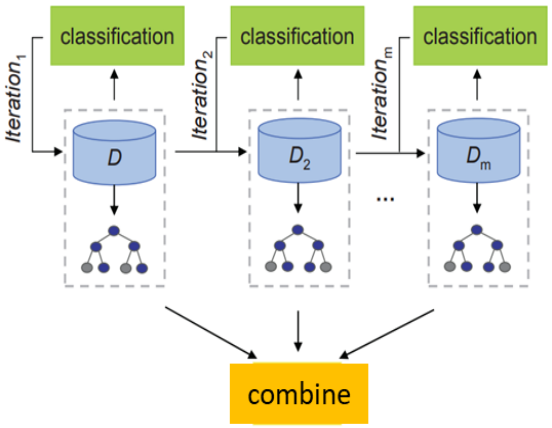

Figure 3. Boosting method (Yang, Yang, Zhou, & Zomaya, 2010).

1.2.2 Bagging Method, or Bootstrap Aggregating, is an effective ensemble algorithm introduced by Breiman, L. It is usually applied together with decision tree methods. The samples of each dataset are randomly drawn to generate several different models. Finally, majority vote is conducted to find the final model which is the best answer.

Random Forest is one of the most popular ensembles learning algorithms and also a notable improvement of bagging; it constructs an extensive collection of de-correlated trees and averages them. It creates multiple decision tree models to support decision making and votes to choose the best outcome. However, Random Forest adds a function of random feature selection from the sample set. This can reduce the correlation between features (Njoku, 2019). The method is shown in Figure 4.

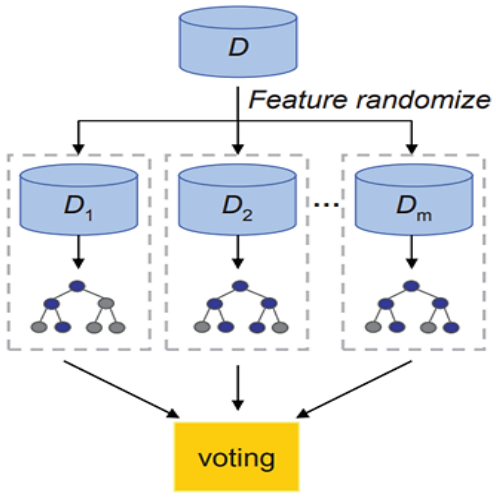

Figure 4. Random forest method (Yang et al., 2010).

1.3 Model efficiency measurement

1.3.1 The efficiency of model is measured by confusion matrix, which summarizes the number of correct and incorrect classified data, as shown in Figure 5.

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

(a)

		Predicted Class			
		C_1	C_2	...	C_N
Actual Class	C_1	$C_{1,1}$	FP	...	$C_{1,N}$
	C_2	FN	TP	...	FN
	...	...	...	...	...
	C_N	$C_{N,1}$	FP	...	$C_{N,N}$

(b)

Figure 5. Confusion matrix examples.
(a) Binary classification problem confusion matrix.
(b) Multiclass classification problem confusion matrix (Markoulidakis et al., 2021).

where
TP is correct prediction value of the target group
FP is incorrect prediction of the target group
TN is correct prediction value of other groups
FN is incorrect prediction value of other groups

Suan Sunandha Science and Technology Journal

To measure the effectiveness of the proposed method, the following details are described: (Markoulidakis et al., 2021)

(1) Accuracy is a value that indicates how close of the predictive value to the actual value, as shown in Equation (3).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

(2) Precision is the ratio of correctly predicted data to the total number of retrieval data. It also can be defined as the value denoted information retrieval and how predictive data is likely correct, as explained in Equation (4).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

(3) Recall is ratio of data that meet the requirements and are retrieved to all observed data, or the value that represents the retrieved data mostly responding to the user's needs, as in Equation (5).

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

(4) F-Measure is the harmonic mean of precision and recall and can be expressed in Equation (6).

$$F\text{-}Measure = \frac{2*Precision * Recall}{Precision + Recall} \quad (6)$$

1.3.2 Mean Absolute Error (MAE) is a measure of the average size of the mistakes in a collection of predictions, without taking their direction into account. It is measured as the average absolute difference between the predicted values and the actual values, as in Equation (7).

$$MAE = (1/n) * \Sigma|y\_i - \hat{y}\_i| \quad (7)$$

where

n is the number of observations in the dataset.

y_i is the true value.

ŷ_i is the predicted value.

1.4 Relevant research

Zamzuri and Qi (2022) studied the decision tree model to classify the severity levels of traffic accidents in Malaysia. This study aims to identify the main factors that drive the occurrence of road accidents in Malaysia. The Classification and Regression Tree (CART) and Chi-square Automatic Interaction Detector (CHAID) techniques are used to identify the effects of factors in this study. The performances of the two classification models are compared based on prediction accuracy and model reliability. It is found that CHAID performs slightly better than CART and offers richer information in terms of influential factors and decision rules.

Nedjmedine and Tahar (2022) studied the decision tree model to analyze road accidents in Algeria. With the enormous number of death and injuries, this problem pushes governments to create solutions to reduce those statistics. Then, the decision tree model compares with similar works using accuracy as a performance evaluation metric. This work can help government and traffic safety entities to improve road safety and minimize the number of accidents.

Chen and Chen (2020) studied statistical and nonparametric data mining techniques for road accidents, namely, logistic regression (LR), classification and regression tree (CART), and random forest (RF), to compare their prediction capability, identify the significant variables (identified by LR) and important variables (identified by CART or RF) that are strongly correlated with road accident severity, and distinguish the variables that have significant positive influence on prediction performance. In this study, three prediction performance evaluation measures, accuracy, sensitivity, and specificity.

Boonraksa and Thongkam (2018) studied the effectiveness of models in predicting road accidents in Khon Kaen Province, Thailand. Five modeling techniques were used, which included Linear Regression (LR), Artificial Neural Network (ANN), Sequential Minimal Optimization for Regression (SMOReg), Support Vector Machine Regress (SVMR), and Gussian Process (GP). The predictive efficiency of the models was measured with mean absolute error (MAE) and root mean square error (RMSE). The results indicated that SVMR technique is effective in building a predictive model of road accidents with the lowest error value, compared to LR, ANN, SMOreg, and Gussian Process models.

Taamneh, Alkheder, and Taamneh (2017) studied data mining techniques for traffic accidents modeling and predictions in the United Arab Emirates. Four classification algorithms were

Suan Sunandha Science and Technology Journal

employed, included: Decision Tree (J48), Rule Induction (PART), Naïve Bayes (NB), and Multilayer Perceptron (MLP). The results showed that the overall accuracy of The J48, PART, and MLP classifiers in predicting the severity of severity injury resulting from traffic accidents, using 10-fold cross-validation was similar. The results revealed that the 18-30-year-old age group was most vulnerable to traffic accidents. Drivers were more frequently involved in traffic accidents than passengers and pedestrians. Male drivers are more involved in traffic accidents than female drivers.

According to recent researches, data mining techniques, especially decision tree, are widely used to analyze data and given satisfactory results. Therefore, the researchers adopted the principle of decision tree to build a risk prediction model of road accidents during the New Year and Songkran Holiday.

2. Materials and Methods

In this study, the risk prediction model of road accidents during long holiday was constructed with ensemble learning using decision trees as a fundamental algorithm. The overall work process is shown in Figure 6.

2.1 Data collection

Data of road accidents during long holiday were complied. The dataset used in this study is the record of injury and death from road accidents during the New Year and Songkran Holiday from 2008 to 2015 from hospitals across the country, collected by the National Institute for Emergency Medicine. Examples of road accident data, as shown in Figure 7.

2.2 Data preparation

The selection of inputs is the most important aspect of creating a useful prediction, as it represents all of the knowledge that is available to the model to base the prediction. The 13 features were selected for use in model construction from the total number of 18 features, which have removed unwanted features such as hospital code, hospital name, transporting the injured, number of days of treatment, and province name. Unwanted data is duplicate or irrelevant data. This redundant data should be removed as it is of no use and will only increase the amount of data and the time to train the model. The Feature 13 result of the accident was employed to classify the dataset. There are 417,122 complete records. Ignoring the tuple was applied to the records with missing values. Dataset descriptions are shown in Table 1.

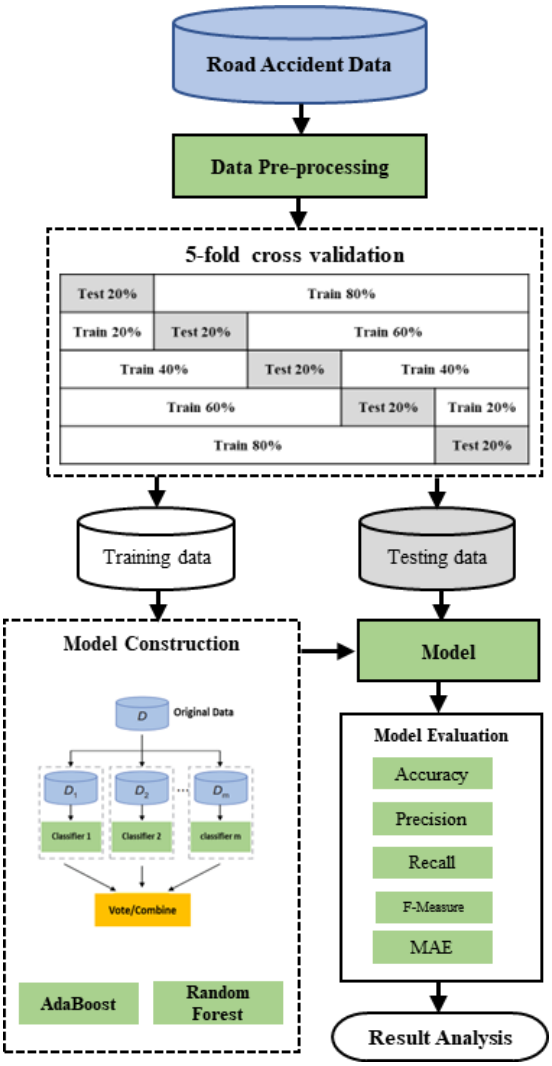

Figure 6. Research procedure.

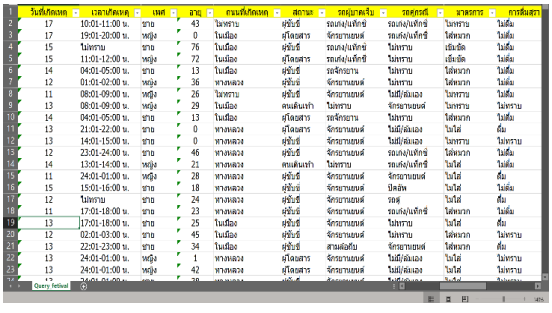

Figure 7. Examples of road accident data.

Suan Sunandha Science and Technology Journal

Table 1. Feature description.

No.	Feature Name	Description
1	Holiday	Type of holiday 1 = New Year 2 = Songkran
2	Province	Province code where the accidents happened 10 = Bangkok 11 = Samutprakarn : 97 = Buengkan
3	Date	In what day in 7-dangerous days that the accidents happened 1 = Day 1 2 = Day 2 3 = Day 3 : 7 = Day 7
4	Time	Time when the accidents happened 1 = 00.01-01.00 2 = 01.01-02.00 : 24 = 23.01-24.00
5	Sex	1. Female 2. Male
6	Age	1. <=10 2. 11-20 3. 21-30 4. 31-40 5. 41-50 6. 51-60 7. 61-70 8. 71-80 9. >80
7	Roadacc	Types of road where the accidents happened 1 = City road 2 = Rural road 3 = Highway 4 = No information
8	Status	1 = Driver 2 = Passenger 3 = Pedestrian
9	Injured_car	1 = None/Falling 2 = Motorcycle 3 = Pick-up 4 = Private car/Taxi 5 = four-wheel passenger car 6 = Big bus 7 = Bicycle 8 = Van 9 = Truck 10 = Motor-tricycle 11= Tricycle 12 = Other
10	Parties_car	1 = None/Falling 2 = Motorcycle 3 = Pick-up 4 = Private bar/Taxi 5 = four-wheel passenger car 6 = Big bus 7 = Bicycle 8 = Van 9 = Truck 10 = Motor-tricycle 11= Tricycle 12 = Other
11	Protection	1 = Wearing seatbelts 2 = Wearing helmets 3 = Not-wearing seatbelts or helmets 4 = No information
12	Alcohol	1 = Yes 2 = No 3 = No information
13	Result	Accident severity 1 = Injury/Recovery 2 = Death

2.3 Model construction

In this research, the models constructed based on AdaBoost and Random Forest using decision tree techniques such as J48, ID3, and CART, each with a different 5-fold cross validation partitioning of road accident dataset were compared to the prediction efficiency.

The random variables are selected to create differences in the training sets, consisting of N number of constructed models. The randomly generated data set is called Bootstrap. The probability of teaching instances can be explained in Equation (8).

$$n_i = 1 - \left(1 - \frac{1}{m}\right)^m \quad (8)$$

where

n_i is a randomly generated set of data

m is the total number of sample data in the training set

In this section, experiments were performed to determine the parameters and optimize the parameters for models. The minimum number of experiments was 20 to evaluate the parameter optimization of models. The details of model construction are explained as follows:

2.3.1 Adaptive Boosting Method (AdaBoost)

The parameters in the experiment were:

Number of Model: 10, 30, 50, 80, 100

According to the experiments, the number of models with the highest efficiency of each decision tree technique is presented in Table 2.

Table 2. Parameter of AdaBoost.

Algorithms	Number of Model
J48	80
ID3	100
CART	100

2.3.2 Random Forest

The parameters in the experiment were:

- Number of Model: 10, 30, 50, 80, 100
- Number of Feature: 2, 4, 6, 8,10

As in the experiments, the parameters with the highest efficiency of each decision tree technique are shown in Table 3.

Table 3. Parameter of Random Forest.

Algorithms	Number of Model	Number of Feature
J48	80	6
ID3	100	6
CART	80	8

3. Results and Discussions

In this research, the data were classified with ensemble learning to predict the risk of road accidents during holiday season. The efficiency measurement was conducted by comparing accuracy, recall, precision, f-measure, and MAE of the constructed models. The results are as follows:

3.1 AdaBoost modeling

The results of model construction from AdaBoost algorithm are explained in Table 4.

Table 4. Data analysis by AdaBoost.

	J48	ID3	CART
Accuracy	87.5%	83.3%	76.7%
Precision	0.871	0.843	0.769
Recall	0.875	0.833	0.767
F-Measure	0.872	0.831	0.767
MAE	0.090	0.108	0.156

According to Table 4, the J48 decision tree technique of the Number 80 parameter had the highest efficiency with accuracy of 87.5%, precision of 0.871, recall of 0.875, f-measure of 0.872, and MAE of 0.090.

3.2 Random forest modeling

The results of model construction from Random Forest are presents in Table 5.

Table 5. Data analysis by Random Forest.

	J48	ID3	CART
Accuracy	93.3%	82.5%	74.2%
Precision	0.934	0.868	0.734
Recall	0.933	0.850	0.742
F-Measure	0.931	0.833	0.733
MAE	0.042	0.117	0.181

According to Table 5, with the J48 decision tree, the Number 80 parameter and Number 6 Feature expressed the highest efficiency with accuracy of 93.3 %, precision of 0.934, recall of 0.933, f-measure of 0.931, and MAE of 0.042.

3.3 Comparison of accurate efficiency

In this research, accurate efficiency obtained from the experimental results of predictive model construction from AdaBoost and Random Forest techniques. The results are shown in Table 6 and Figure 8.

Table 6. Comparison of accuracy.

Algorithms	Boosting (AdaBoost)	Bagging (Random Forest)
J48	87.5%	93.3%
ID3	83.3%	82.5%
CART	76.7%	74.2%

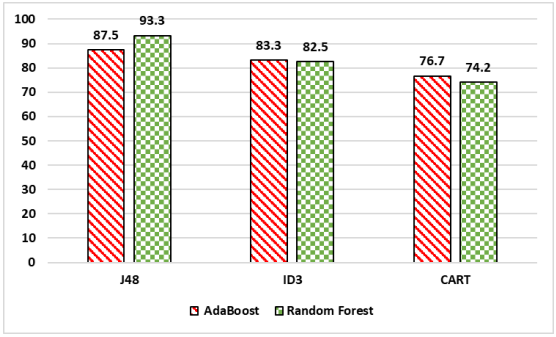

Figure 8. Results of accuracy comparison.

According to Table 6 and Figure 8, the J48 decision tree technique with Random Forest modeling has the highest efficiency with the

accuracy of 93.3%. Therefore, the model is suitable to employ for risk prediction of road accidents during holiday season.

4. Conclusions

This research aimed to propose a risk prediction model of road accidents during the holiday season based on ensemble learning using decision tree as a basic algorithm. The results revealed that:

The risk prediction model constructed from J48 algorithm with Random Forest Method had the highest prediction efficiency with accuracy of 93.3%. When comparing the accurate efficiency of the two techniques, Random Forest mostly expressed more accuracy than Boosting Method. It is because the Random Forest is added the random function to select data features for analysis, which reduces the correlation between each feature. The features were independent when building the decision trees. The As a result, the constructed decision trees are varied, small-structured, rapid processing, and highly efficient. In conclusion, the constructed model from Random Forest is suitable for predicting road accident risk during the next holiday season.

However, the model proposed by the researchers has focused on model construction by ensemble classification using Boosting and Bagging Methods. There are many other methods have not been mentioned, e.g. Stacking, Voting, Random Subspace, and Hybrid Experts, etc. Furthermore, the application of new algorithms for data classification, such as Artificial Neural Network algorithm (deep learning), can be employed to enhance the accuracy; and the hybrid algorithms can be used to improve the efficiency of data classification.

References

Accident Prevention Network. (2022). *Traffic accident statistics report: New Year - Songkran Festival 2022*. Retrieved from http://www.accident.or.th/ index.php/2017-12-04-07-32-28/289-2565

Boonraksa, P., & Thongkam, J. (2018). Performance comparison of the road occurrence accidents prediction models using time series techniques. *Journal of Technology Management Rajabhat Maha Sarakham University*, *4*(2), 39-46.

Chen, M.-M., & Chen, M.-C. (2020). Modeling road accident severity with comparisons of logistic regression, decision tree and random forest. *Information*, *11*(5), 270. doi:10.3390/info11050270.

Esenturk, E., Turley, D., Wallace, A., Khastgir, S., & Jennings, P. (2022). A data mining approach for traffic accidents, pattern extraction and test scenario generation for autonomous vehicles. *International Journal of Transportation Science and Technology*. doi:10.1016/j.ijtst. 2022.10.002

Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., & Doulamis, N. (2021). Multiclass confusion matrix reduction method and its application on net promoter score classification problem. *Technologies*, *9*(4), 81. doi:10.3390/technologies9040081

Nedjmedine, O., & Tahar, M. (2022). Analysis of road accident factors using Decision Tree Algorithm: A case of study Algeria. *5th International Symposium on Informatics and its Applications (ISIA)*. M'sila, Algeria. doi:10.1109/ISIA55826.2022.9993530.

Njoku, O. C. (2019). *Decision trees and their application for classification and regression problems* (Master's thesis). MSU Graduate Theses. Retrieved from https://bearworks. missouristate.edu/theses/3406

Ogheneovo, E. E., & Nlerum, P. A. (2020). Iterative Dichotomizer 3 (ID3) decision tree: A machine learning algorithm for data classification and predictive analysis. *International Journal of Advanced Engineering Research and Science (IJAERS)*, *7*(4), 514-521. doi:10.22161/ijaers.74.60

Panigrahi, R., & Borah, S. (2018). Rank allocation to J48 group of decision tree classifiers using binary and multiclass intrusion detection datasets. *Procedia Computer Science*, *132*, 323-332. doi:10.1016/j.procs.2018.05.186

Parathasarathy, G., Soumya, T. R., Das, Y. J., Saravanakumar, J., & Merjora, A. A. (2019). Using hybrid data mining algorithm for analysing road accidents data set. *3rd International Conference on Computing and Communications Technologies (ICCCT)* (pp. 7-13). Chennai, India. doi:10.1109/ICCCT2.2019.8824860

Suan Sunandha Science and Technology Journal

Sonwongsa, R., Pinpoo, S., & Wongkhae, K. (2016). Severity analysis of car accidents during "7 dangerous days" of New Year and Songkran festival. *The 12th Mahasarakham University Research Conference* (pp. 39-47). Thailand.

Taamneh, M., Alkheder, S., & Taameh, S. (2017). Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *Journal of Transportation Safety & Security*, *9*(2), 146-166. doi:10.1080/19439962.2016.1152338

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: An experimental review. *Journal of Big Data*, *7*(1), 1-47. doi:10.1186/s40537-020-00349-y

Yang, P., Yang, Y. H., Zhou, B. B., & Zomaya, A. Y. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, *5*(4), 296-308. doi:10.2174/157489310794072508

Zacharis, N. Z. (2018). Classification and Regression Trees (CART) for predictive modeling in blended learning. *International Journal of Intelligent Systems and Applications*, *3*, 1-9. doi:10.5815/ijisa.2018.03.01

Zamzuri, Z. H., & Qi, K. Z. (2022). Classifying the severity levels of traffic accidents using decision trees. *Proceedings of the International Conference on Mathematical Sciences and Statistics 2022 (ICMSS 2022)*. doi:10.2991/978-94-6463-014-5_17