Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 5, No. 1, March, 2024

Using Multilayer Perceptron Neural Network to Assess the Critical Factors of Traffic Accidents

Athapol Ruangkanjanases [1], Ornlatcha Sivarak [2]*, Zi-Jie Weng [3], Asif Khan [4, 5], Shih-Chih Chen [3]

[1] *Chulalongkorn Business School, Chulalongkorn University, Bangkok, Thailand.*

[2] *Mahidol University International College, Mahidol University, Nakhon Pathom, Thailand.*

[3] *Department of Information Management, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan.*

[4] *Southern Taiwan University of Science and Technology, Tainan, Taiwan.*

[5] *ANScientistify INC., Tainan, Taiwan.*

Received 15 September 2023; Revised 04 January 2024; Accepted 16 January 2024; Published 01 March 2024

Abstract

This study is based on the traffic accident data of Taoyuan City from the government's open data. The study compiled the data set of traffic accidents in Taiwan from 2012 to 2017, and six classifiers were applied to evaluate the effectiveness of traffic accident prediction with the number of injuries as the prediction target. In order to verify the classifier's stability, cross-validation was used to evaluate the model during the training process, and the multilayer perceptron neural network (MLPNN) classifier performed best in testing the dataset's accuracy and evaluating the model's best performance. Then, a boosting ensemble learning approach and a combination of traffic accident factors improve the experiment's performance. According to this experiment, the results show that this study uses the Pearson Chi-square feature selection method to select important traffic factor combinations, and the boosting method indeed helps improve the effectiveness of the construction of the traffic accident model. Finally, the experimental results of the NN-MLP model have a correct rate of 77% and AUC is 78.7%. In constructing the model, it was found that the degree of injury, the part of the vehicle hit, the type of accident, the leading cause, the type of vehicle, and the period of the accident were the main factors causing dangerous traffic accidents.

Keywords: Data Mining; Multilayer Perceptron Neural Network; Traffic Accident; Government Open Data; Feature Selection.

1. Introduction

Injuries caused by traffic accidents are a common problem worldwide, so it is necessary to avoid accidents and take precautionary measures. Traffic safety has always been an important issue for the government and the public, and the most effective way to improve traffic safety is to reduce the severity of accidents [1]. In recent years, road traffic analysis has focused on the risk factors for traffic severity and fatalities [2], but many of the major factors in traffic accidents have not yet been identified and analyzed. Traffic accidents are the costliest events in terms of human and financial

* Corresponding author: ornlatcha.siv@mahidol.ac.th

http://dx.doi.org/10.28991/HIJ-2024-05-01-012

resources. According to World Health Organization (WHO) statistics, traffic accidents are responsible for more than 1.25 million deaths worldwide each year, while the number of people who suffer non-fatal injuries is between 20 million and 50 million, many of whom are disabled as a result of their injuries. The report states that traffic accidents cause most countries to lose 3% of their gross domestic product (GDP) [3]. Taiwan's traffic accident data is calculated by the National Police Agency, Ministry of the Interior [4]. The number of fatalities in traffic accidents decreased from 2,040 in 2013 to 1,604 in 2016, but the number of traffic accidents is gradually increasing from 278,388 accidents and 373,568 injuries in 2013 to 305,556 accidents and 403,906 injuries in 2016. Deaths and injuries caused by car accidents have led to increased medical expenses, loss of productivity, reduced quality of life, and loss of personnel. Taiwan's institute of Transportation, Ministry of Transportation and Communications, estimated that the social cost of traffic accidents increased by approximately NT$444.5 billion, accounting for 3.3% of Taiwan's GDP in 2013 [5]. It is clear that traffic accidents have long been a major crisis in daily life and in the country.

In order to avoid traffic accidents, the NPA evaluates traffic accident data and analyzes the main factors causing the accident every year, and adopts publicity measures to reduce the accident rate in response to these factors. According to the NPA's data in 2016, the number of motor vehicles increased by 0.51% from 21,400,897 to 21,510,560 in 2015 [4]. At that time, because the government implemented a number of traffic safety policies, the number of Class A1 traffic accidents (resulting in injury or death within 24 hours) decreased by 1,555 from 1,639 (-5.13%) in 2015. From the data of the past years, Class A1 has a trend of decreasing year by year. However, summing Class A1 and A2 (causing injuries or more than 24 hours of death) showed an increase from 120,223 cases in 2003 to 305,556 cases in 2016, indicating that the government still needs to do more to prevent traffic cases. In the face of the yearly increase in the number of traffic accidents, the government has also summarized the following factors, including the most frequently occurring factors, location, and time of the accident [4]. The most common causes of accidents were failure to give way (19.54%), improper turn (14.1%), violation of signal control (8.33%), and failure to maintain safe driving distance and separation (7.26%). In terms of road types, traffic accidents are the most common on urban roads (75.66%). Observing the time of traffic accidents, it is found that daytime is higher than nighttime, and the two periods of 8–10 o'clock (10.80%) and 6–8 o'clock (10.42%) are the most. So, it is necessary to analyze past data on traffic accidents, and these analysis results can be used to make future safety promotion decisions to reduce traffic accidents and injuries. The above is based on descriptive statistics. However, the events that cause traffic accidents are caused by a variety of factors. Chen & Jovanis [6] argued that using traditional statistical techniques to analyze large-dimensional datasets may cause some problems and that statistical models have their own specific assumptions, the violation of which may lead to some erroneous results. This is a limitation of traditional statistical methods. But using data mining can unearth the hidden information in large-dimensional data sets [6]. Machine learning technology is also widely applied to traffic management problems, including group optimization algorithms and decision tree algorithms to find the best control between traffic signals on the road to solve traffic congestion problems [7, 8].

Past studies have explored factors affecting traffic accidents, including month, time, week, year, number of victims, weather, light, accident type, main factors of the accident, the severity of the injury, road signs, road characteristics, and road types [9-13]. In addition, some studies have shown that socioeconomic status is related to the occurrence of traffic accidents. For example, the GDP is directly proportional to the death rate from traffic accidents, but if the population using automobiles and motorcycles changes, different results will be obtained. Whether a city is prosperous is strongly correlated with the number of deaths caused by automobiles and motorcycles, so it is also an important factor. However, countries with different levels of wealth have different results. A study observed low- and middle-income countries and found that the degree of regional prosperity has a positive relationship with the fatalities of traffic accidents; in high-income countries, the fatalities of traffic accidents in the richest regions are relatively the lowest [14].

Traffic accidents are caused by four major factors, such as vehicles, roads, people, and irresistible events, and each contains various influencing factors. The conditions between the factors are not completely independent, and there is an interaction between the factors. Therefore, exploring the causes of traffic accidents is an important issue. Now, there have been many studies on the factors contributing to traffic accidents. For example, Abellán et al. [15] adopted decision trees to analyze the severity of traffic accidents. Their study discussed the type of crash, age of the perpetrator, weather, time of occurrence, light, age, shoulder width, number of people involved, road guardrails, and other factors. Then the research results show that whether to wear a seat belt, road shoulder driving, driver visibility, and lighting are the main factors affecting a collision's severity. In addition, they also found that the accident rate of motorcycle accidents on rural roads and in good weather was about 69.9%. If the above conditions are combined with the fact that the driver is male, the accident rate is about 68.5%. Therefore, several preventive strategies are proposed for motorcyclists, including strict monitoring of rural roads and lowering the maximum speed limit on rural roads, to reduce traffic accidents.

In addition, other scholars use different research methods and influencing factors to discuss the causes of traffic accidents. Kumar & Toshniwal [11] used data exploration to obtain factors including the number of injured, time, month, road type, severity of injury, type of accident, light, and surrounding environment of the area. They used data clustering and correlation rules to analyze the data, then grouped the data into six clusters and further explored them using the correlation method. They found that (1) the highest percentage of two-wheeled accidents occurred at road intersections and markets nearby (27.48%); (2) the highest rate of two-wheeled accidents occurred in non-highway areas; (3) two-wheeled accidents are more likely to occur near markets and between 4 p.m. and 8 p.m. Kumar & Toshniwal [12] used k-means and relevance rules to analyze factors such as a month, time, week, number of people involved, light, accident type, severity of injury, road type, road characteristics, age, and surrounding environment. They divided the accident frequency into three groups: high, medium, and low, and then further used correlation analysis to explore the reasons for the different frequencies of accidents. They found that in the high-frequency group, accidents were most likely to occur at intersections on high-speed roads, followed by multiple vehicle accidents on non-high-speed roads, agricultural roads, and road turns; in the medium-frequency group, pedestrians were likely to be struck at intersections in urban areas and on highways.

Castro & Kim [10] analyzed weather, light, injury severity, road conditions, vehicle oil replenishment, humidity, and vehicle operating conditions and applied a Bayesian network, multilevel perceptron, and J48 decision tree classifier. They found that light, vehicle operation, and road type were the most influential factors during classifier training. In addition, they found that the age of the vehicle and the weather did not significantly affect the injury's severity. Zeng et al. [13] used a Bayesian hierarchical logistic regression analysis to verify factors such as type of accident, severity of injuries, safety measures (seat belts), vehicle type, driver status, portion of the vehicle impacted, age of the vehicle, whether the location of the incident was near the workplace, age, and whether the driver had a record of violations. The results of this study showed that (1) older women and passengers who did not use safety equipment were most likely to be injured; (2) newer vehicles with low-speed conditions were less likely to be injured; and (3) head-on collisions or accidents on the road were the two most serious accidents. They believe that these findings will contribute to traffic safety education, regulations, and transportation facilities.

Alkheder et al. [9] adopted *k*-means Clustering and an artificial neural network classifier to analyze factors such as year, day, time, cause of accident, type of accident, gender, nationality, age, whether seat belts were used, the severity of injury, light, road surface condition, weather, and other variables. Their research results show that the use of artificial neural network models to analyze traffic accident data sets has a 74.6% correct probability of prediction. And they believe that the result can provide the transportation department of the United Arab Emirates with a reference for improving traffic safety.

This study summarizes the previous studies on the influencing factors of traffic accidents in Table 1. As a result of the above literature, this study plans to apply data mining technology to analyze the causes of traffic accidents. This study uses feature selection combined with the NN-MLP algorithm to analyze traffic accidents. The experimental design is based on the traffic accident factors and methods considered in previous traffic accident papers, and the results are analyzed using a combination of feature-selected factors and the NN-MLP algorithm. The data needed for this research uses machine learning, statistical methods, and optimization algorithms to mine complex and big data. Then, specific correlations and characteristics hidden in the dataset are analyzed. Finally, this study hopes that the results of this study can provide a reference for organizations or researchers to make decisions or forecasts. This study uses 70 complex factors, with the number of injuries as the main prediction target, to analyze traffic accidents by feature selection combined with the NN-MLP algorithm from open data in Taiwan. Among the factors, the six major influences on the number of injured persons in traffic accidents were found to be the most influential factors, including the initial impact site of the vehicle, vehicle type, time of day, type and type of accident, and the main cause. The actual application can be evaluated by referring to these factors or visualizing some combinations. For example, based on the visualization of three factors: vehicle type, main cause, and degree of injury, the type of motorcycle involved in a traffic accident is more likely to cause more than two injuries than a small passenger car. Therefore, we can strengthen the promotion of motorcycles in the distance limit between the front and back of the car or reduce the speed limit so that drivers can have a more defensive driving concept and reduce the medical burden and safety of the public.

This study uses feature selection combined with the NN-MLP algorithm to analyze traffic accidents. The experimental design is based on the traffic accident factors and methods considered in previous traffic accident papers, and the results are analyzed using a combination of feature-selected factors and the NN-MLP algorithm. The results of the data analysis in this study confirmed the following factors, for example, the six main influencing factors of traffic accidents with injuries, and the most influential factors were found to be the initial impact area of the vehicle, vehicle type, time of day, accident type and type, and the main cause. For example, the visualization of the three factors of vehicle type, main cause, and degree of injury, the occurrence of traffic accidents with motorcycles is more likely to cause more than two injuries than that of passenger cars, so we can strengthen the promotion of motorcycles in the distance limit between the front and rear of the car or reduce the speed and other restrictions so that drivers have a more defensive driving concept, reducing the medical burden and driving safety.

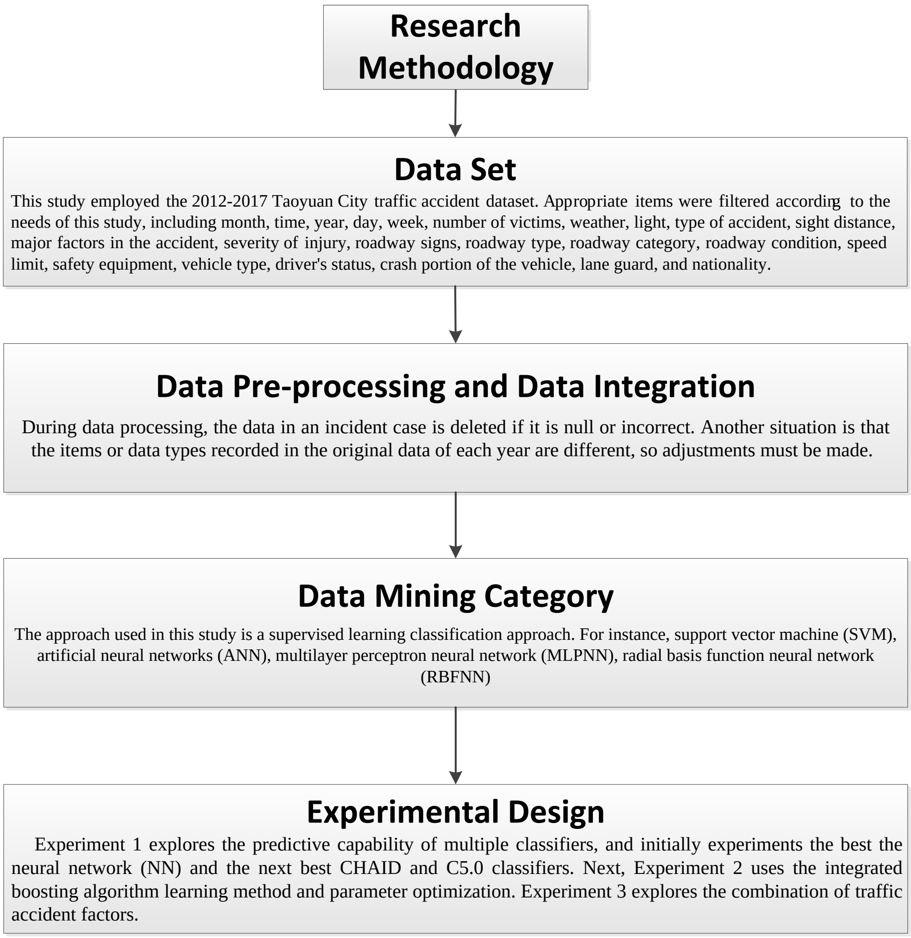
This research has some major contributions for transportation and traffic control agency practitioners and managers. This research designs a process for analyzing traffic accident data and constructs a prediction model for high-risk accidents. The results should be able to provide relevant government agencies to assess the key factors of traffic accidents, conduct safety advocacy for high-risk traffic factors, strengthen the importance of people's driving safety, and reduce the social costs caused by traffic incidents. The main research limitation of this study is that the data analysis was performed using commercial software. As a result, the model design and data presentation are relatively limited, which makes it difficult to adjust the details of the model computation process in terms of details or parameters. For future research, it is recommended to try to test the disease analysis model with more flexible algorithms. The rest of the paper is presented as follows. Section 2 explains the methods of this study, while Section 3 proposes the data analysis results and discussion. Finally, the conclusions and research limitations of this work are discussed in Section 4.

Table 1. Traffic accident impact factors

Scholars	Number of Factors	Factor Name
Abellan et al. [15]	20	Type of accident, age, weather, road guardrail, main factor of accident, week, lane width, light, month, number of injured, number of people involved, shoulder type, road width, road marking, gender, shoulder width, sight distance, time, car payment, severity of injury
Kumar et al. [11]	11	Number of injured persons, age, gender, time, month, light, road characteristics, road type, accident severity, area surrounding, type of accident
Kumar et al. [12]	13	Number of people involved, age, gender, type of accident, time, week, month, location (number), light, road characteristics, the severity of an accident, surroundings, type of road
Castro et al. [10]	8	Type of road, light, weather, humidity, operating condition of the car, oil used, age of the car, severity
Zeng et al. [13]	13	Degree of injury, age, gender, whether alcohol was consumed, safety equipment, driving violation record, age, vehicle type, speed rate, location of vehicle impact, location of the accident, whether near work, type of accident
Alkheder et al. [9]	16	Year, day, time, cause of the accident, type of accident, gender, nationality, age, seat belt, occurrence factor, severity of injury, light, road surface condition, speed limit, lane number, weather

2. Methods

This section introduces the dataset used in this study, including the redefinition of the data, the data exploration classifier used in the experiment, the experimental design framework, and the experimental evaluation method. Figure 1 represents the research methodology flowchart.



This study employed the 2012-2017 Taoyuan City traffic accident dataset. Appropriate items were filtered according to the needs of this study, including month, time, year, day, week, number of victims, weather, light, type of accident, sight distance, major factors in the accident, severity of injury, roadway signs, roadway type, roadway category, roadway condition, speed limit, safety equipment, vehicle type, driver's status, crash portion of the vehicle, lane guard, and nationality.

During data processing, the data in an incident case is deleted if it is null or incorrect. Another situation is that the items or data types recorded in the original data of each year are different, so adjustments must be made.

The approach used in this study is a supervised learning classification approach. For instance, support vector machine (SVM), artificial neural networks (ANN), multilayer perceptron neural network (MLPNN), radial basis function neural network (RBFNN)

Experiment 1 explores the predictive capability of multiple classifiers, and initially experiments the best the neural network (NN) and the next best CHAID and C5.0 classifiers. Next, Experiment 2 uses the integrated boosting algorithm learning method and parameter optimization. Experiment 3 explores the combination of traffic accident factors.

Figure 1. Research Methodology Flowchart

2.1. Data Set

The data for this study is the Taoyuan City traffic accident dataset, taken from Taiwan's government open data (https://data.tycg.gov.tw/). The data provided by the platform is from 2012 to 2017. Since the traffic accident investigation's content varies annually, the data was compiled based on the original survey form fields in 2012. The 70 factors used in this study are shown. However, the dataset has some problems, such as null values, incorrect case information, and different items recorded for each year. Thus, this study reorganizes these data. First, this study downloaded the 2012–2017 Taoyuan City traffic accident dataset from this platform and compiled it into a single dataset. Next, appropriate items were filtered according to the needs of this study, including month, time, year, day, week, number of victims, weather, light, type of accident, sight distance, major factors in the accident, the severity of injury, roadway signs, roadway type, roadway category, roadway condition, speed limit, safety equipment, vehicle type, driver's status, crash portion of the vehicle, lane guard, and nationality.

2.2. Data Pre-processing and Data Integration

During data processing, the data in an incident case is deleted if it is null or incorrect. Another situation is that the items or data types recorded in the original data for each year are different, so adjustments must be made. For example, the data for 2017 has the item of the week, but not in other years. Therefore, the data of the item of the week from other years must be extrapolated from the year, month, and day data.

Because of the disparity in the sample size of injuries, binary partitioning was applied following a previous study [15]. The types of speed limit items are complex, so this study adjusts the data in the speed limit items according to the classification of highway speed limits by Taiwan’s Directorate General of Highways (DGH), which is divided into seven types (30, 40, 50, 60, 80, 100, and 120). In this study, the rules for redefining the data content are organized in Table 2.

This study does not use the item "year" because the data for this item is not complete in the dataset. In the 2017 dataset, the data type of this item for vehicle types was different from the previous years, so the data was converted to consistency. The 2012 dataset was missing data for the item "nationality," so this item was deleted in consideration of the integrity of the overall dataset.

Table 2. Data Content Redefinition

Factor Name	Factor Data Content
Injuries	Injuries less than one person = 0 (Dangerous traffic accidents); more than two people were injured = 1 (High-risk traffic accidents)
Speed Limit	30 = 0, 40 = 1, 50 = 2, 60 = 3, 80 = 4, 100 = 5, 120 = 6
Vehicle Type	Rickshaws, heavy motorcycles (250 to 500), heavy motorcycles (500 or more), compact/military vehicles, small light motorcycles, engineering vehicles/specially constructed vehicles, public motorbus, state-owned motor passenger carrier, private motorbus, private motor passenger carrier, personal-use bus, personal-use heavy truck, personal-use sedan, personal-use light truck.

2.3. Data Mining Category

Data mining is the process of digging out interesting or valuable information from a data set. The technology of mining is a combination of machine learning, optimization algorithms, and statistics. There are two types of data mining: supervised learning and unsupervised learning. The former means that the answer to the exact prediction target is known during the training process, and the prediction model for the problem is derived through iterative training; the latter means that there is no standard answer in the training data, so only the sample data of the discussion factor is input. After repeated training of unsupervised algorithms, the types of samples will be redefined.

The approach used in this study is a supervised learning classification approach. Related classification algorithms for supervised learning have been proposed and applied in various fields. For example, data mining and machine learning were applied to network security [16], the support vector machine (SVM) was applied to biological data classification [17], the support vector machine and the artificial neural network were applied to construction engineering development [18], the decision tree was applied to education data to predict student learning performance [19], and the nearest neighbor method was applied to WEB product recommendation [20].

Artificial Neural Networks (ANN) is a supervised learning algorithm that uses computers to simulate artificial neurons to imitate biological neural networks. Neurons are connected to each other and obtain external information to calculate and transmit the calculation results to the outside or other neurons. Each neuron has a weight. The weight is updated during model training to predict the category more accurately [21].

A. Multilayer Perceptron Neural Network (MLPNN)

MLPNN is a supervised learning approach [22, 23]. MLPNN consists of interconnected neurons or nodes modeled by a nonlinear mapping between input and output vectors. The nodes are connected by weights and output signals that are functions of the sum of the nodes' inputs, which are modified by simple nonlinear transmission or functions. MLPNN consists of a superposition of many simple nonlinear transmission functions so that the multilayer perceptron can approximate the extremely nonlinear functions, and the outputs of the nodes are scaled by the connected weights and fed forward as the input network of the next layer of nodes. Hence, the multilayer perceptron is called a feedforward neural network. Unlike other statistical techniques, MLPNNs do not require a priori assumptions about the data distribution; they can be trained to achieve smoother functions, model highly nonlinear functions, and train accurately on unexpected data.

B. Radial Basis Function Neural Network (RBFNN)

RBFNN is a feedforward neural network with different functions and network architectures. The focus is on the output layer, which is a combination of input layer RBF functions and neurons. This method deals with high-dimensional space and can compute suitable curves in multiple dimensions. Each neuron is trained by calculating the distance between the centroid of each neuron in the hidden layer and the input layer, and the Gaussian function is transformed to calculate the output value of each neuron (see Function 1). Then, the transformed neurons are computed by RBF units, and F is output (see Function 2) [24].

$$R_i(P) = exp\left[-\frac{\|P - C_i\|^2}{\sigma_i^2}\right] \tag{1}$$

$$y_i(P) = \sum_{i=1}^{u} R_i(P) \times \omega(j, i) \tag{2}$$

Based on the government's open data, this study adopts the supervised classification technique in data mining to construct a prediction model for high-risk (two or more injuries) traffic accidents. In this study, the factors and data sets for the training model were compiled based on the factors researched in the previous studies and considering the traffic accident data sets provided by the government (see Table 3). In addition, previous studies investigating traffic accidents have been able to determine the factors that influence traffic accidents, although the classifiers chosen are different. Therefore, this study was planned to test multiple classifiers to predict traffic accidents and evaluate the differences between classifiers.

Table 3. Factors for access to government open data

Obtainable factors	a	b	c	d	e	f
Month	✓	✓	✓			
Time	✓	✓	✓			✓
Year						×
Day						✓
Week	✓		✓			
Number of victims	✓	✓	✓			
Weather	✓			✓		✓
Light	✓	✓	✓	✓		✓
Type of accident	✓	✓	✓		✓	✓
Sight distance	✓					
Major factors of the accident	✓					✓
Severity of injury	✓	✓	✓	✓	✓	✓
Road signs	✓					
Road type		✓	✓			
Road category			✓			
Road condition				✓		✓
Speed limit						✓
Safety equipment					✓	✓
Vehicle type	✓				✓	
Driver status					✓	
Part of the vehicle hit					✓	
Lane guardrail	✓					
Nationality						×

Note 1: [a] Abellán et al. (2013), [b] Kumar et al. (2015), [c] Kumaret et al. (2016), [d] Castro et al. (2016), [e] Zeng et al. (2016), [f] Alkheder et al. (2017).

Note 2: "✓" indicates the factors to be considered; "×" indicates the factors not to be adopted.

2.4. Experimental Design and Evaluation Methods

2.4.1. Experimental Design

The original sample of traffic accidents was 174,665. After processing, 165,846 were left; the factors were filtered from 70 to 21. The modeling capabilities of IBM SPSS Modeler version 17.0 are used in this study, and the modeling classifiers in the software include various classifiers such as analogous neural networks, Bayesian networks, cardinality automatic interaction detection (CHAID), C5.0 decision trees, and support vector machines. Three experiments were conducted in this study. Experiment 1 explores the predictive capability of multiple classifiers and initially experiments with the best neural network (NN) and the next-best CHAID and C5.0 classifiers. Next, Experiment 2 uses the integrated boosting algorithm learning method and parameter optimization. Experiment 3 explores the combination of traffic accident factors.

The result shows that Pearson's chi-squared test and MLPNN boosting used in Experiment 3 are the best analysis methods in this study. The analysis process is as follows: first, Pearson's chi-squared test is used to filter out the factors that affect the training model; then, the classifier is used to build the model. The model is trained by k-fold cross-validation; 5-fold cross-validation is used to cut the training data into five equal data sets, and each experiment takes one as the test data set and the others as the training data sets. The model is trained for five rotations. Each sub-set is given an opportunity to test the task of the data set. After the rotation is completed, the evaluation value of the classifier is calculated [25].

2.4.2. Evaluation Method

Studies exploring the factors that influence traffic accidents point out that traffic accidents have an impact on social costs [26]. The consumption of healthcare resources is also an issue that has been focused on [27]. Therefore, the number of injuries caused by traffic accidents is also considered a subject to be discussed in this study.

The classifier evaluation is based on the Confusion Matrix. When the classifier predicts a dangerous accident (Positive), and it is actually a dangerous accident (True), this situation is called True Positive (TP); when it is predicted to be a high-risk accident (Negative), and it is actually a dangerous accident (True), this situation Called True Negative (TN); when the predicted dangerous accident (Positive) and the actual high-risk accident (False), this situation is called Flase Positive (FP); when the predicted high-risk accident (Negative) and the actual high-risk accident Incident (False), this condition is called Flase Negative (FN) (see Table 4).

Table 4. Injuries' confusion matrix

		Prediction results	
		Injuries less than 1 person	**More than 2 people injured**
Actual Results	Injuries less than 1 person	TP	FN
	More than 2 people injured	FP	TN

The evaluation of the research model uses accuracy, precision, recall, F-measure, and area under curve (AUC). Accuracy refers to the proportion of traffic accidents that the classifier can correctly predict whether a traffic accident is a dangerous traffic accident (the number of injured is less than 1) or a high-risk traffic accident (the number of injured is more than 2) (see function 3). Precision refers to the proportion of samples correctly predicting dangerous accidents to the total number of samples classified as correct (see function 4). Recall refers to the proportion of samples correctly predicting dangerous accidents to be classified as Positive (see function 5). F-Measure refers to the weighted average of Precision and Recall. In general, precision is high and recall is low. Therefore, if the two values are better, refer to F-Measure (see Function 6) [28, 29]. The AUC is used to evaluate classifiers and is one of the metrics to avoid classifier misclassification. The standard of AUC is 0.5, so a good classification model should be greater than 0.5. When the value is closer to 1, it means that the prediction of the model is more perfect [30].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

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

3. Results and Discussion

This section presents the experimental results. In this study, three experiments are designed, Experiment 1 uses the appropriate classifier for comparison, Experiment 2 compares the classifier that is better analyzed in Experiment 1 for model optimization, and Experiment 3 analyzes the prediction of various combinations of traffic accident factors based on the best classifier in Experiment 2. Experiment 3 analyzes the prediction of various combinations of traffic accident factors based on the best classifier in Experiment 2 and finally discusses the results of Experiment 3.

3.1. Comparison of Classifiers

The experimental results are shown in Table 5. The selected classifiers achieved more than 74% predictive power for traffic accidents. The model performance of the classifiers was evaluated to be above 0.7. In the model training datasets of this study, the top three classifiers in terms of accuracy are SVM RBF, KNN, and C5.0, and it is most common for the model to have a high accuracy rate during training and a large difference between the actual test set and the dataset. Then, we evaluate the Precision, Recall, and F-measure indicators, which show that the difference between C5.0 and C5.0 is very high. We then evaluated the Precision, Recall, and F-measure indices, which showed that C5.0 and CHAID had the highest prediction accuracy. The ratio of target prediction distribution is (73:27), so we use the AUC index to verify the imbalance of the model on the data. Finally, NN-MLP was the best classification, with a prediction accuracy of 76.4% and the AUC of 0.779, indicating good predictability for traffic accidents [30]. Therefore, it is proved that the experimental results meet the evaluation criteria.

Table 5. Comparison of Classifier Performance

	Accuracy	Precision	Recall	F-measure	AUC
KNN train	0.804	0.809	0.958	0.878	0.846
KNN test	0.741	0.771	0.917	0.838	0.705
C5.0 train	0.789	0.819	0.914	0.864	0.787
C5.0 test	0.766	0.804	0.899	0.848	0.762
SVM RBF train	0.991	0.993	0.995	0.994	0.999
SVM RBF test	0.747	0.824	0.830	0.827	0.747
Bayesian network train	0.771	0.811	0.896	0.851	0.791
Bayesian network test	0.754	0.798	0.888	0.841	0.763
CHAID train	0.767	0.804	0.903	0.850	0.785
CHAID test	0.764	0.800	0.902	0.848	0.778
MLPNN train	0.770	0.812	0.891	0.850	0.786
MLPNN test	0.764	0.808	0.888	0.846	0.779

3.2. Parametric Optimization Combined with Integrated Algorithms

In the previous section, the parameters taken by the classifier are set to the default values in the software and do not optimize the classifier's performance. Therefore, in Experiment 2, the best classifier NN and the second-best classifier CHAID, C5.0, are adjusted using integrated learning and parameters. The integration method uses the Boosting algorithm, which adjusts the weights of the training sample based on the previous model. Each time the weights of the constructed model are adjusted, the classification error rate decreases, and the best classification accuracy rate is obtained by repeatedly constructing the model. The experimental results are shown in Table 6. NN-MLP boosting has a small improvement in the accuracy of the data prediction ability (+0.3%) and the classifier performance has also increased (+0.3%), but replacing NN with another learning method is worse than expected. This experiment shows that the boosting is effective in improving this study's prediction performance and that NN-MLP is the best classifier for traffic accident prediction compared to other methods in this study.

Table 6. Ensemble learning and parameter adjustment

	Accuracy	Precision	Recall	F-measure	AUC
CHAID boosting train	0.774	0.810	0.902	0.854	0.788
CHAID boosting test	0.760	0.800	0.895	0.845	0.779
C5.0 boosting train	0.814	0.835	0.930	0.880	0.810
C5.0 boosting test	0.767	0.805	0.897	0.849	0.773
NN-RBF boosting train	0.732	0.732	1.000	0.845	0.698
NN-RBF boosting test	0.730	0.730	1.000	0.844	0.693
MLPNN boosting train	0.775	0.810	0.905	0.855	0.791
MLPNN boosting test	0.767	0.804	0.900	0.849	0.782

3.3. NN-MLP Traffic Accident Factor Combination Analysis

In the above section, Experiment 3 uses the NN-MLP boosting classifier to investigate the effect of different combinations of factors on the prediction. Many scholars have analyzed various combinations of factors that may influence traffic accidents. Abellán et al. [15] analyzed the factors affecting traffic accidents, including month, time, week, number of victims, weather, light, type of accident, sight distance, major factors of the accident, severity of the injury, road signs, vehicle type, and lane guardrail. The factors considered by these scholars are most similar to the data set items used in this study, so this study refers to and adopts their factors. In addition, a feature selection filter was used to select the data set factors, and the less influential factors (e.g., day, drinking status, distance to view) were removed. Finally, 18 factors were considered, including month, hour, day of the week, injury, weather, light, road type, speed limit, road type, road condition, diverging facilities, fast and slow lane interval, accident category and type, main cause, injury level, protection, equipment, vehicle crash site initial, and vehicle type. The results of Experiment 3 are shown in Table 7 [31]. The experimental evaluation shows that the NN-MLP boosting classifier analyzes the predictive performance of the four different factor combination data sets, and the factor data set with the feature selection method has the best predictive effect.

Table 7. Traffic accident factor combinatorial analysis

	Accuracy	Precision	Recall	F-measure	AUC
Full Column of Literature Train	0.775	0.810	0.905	0.855	0.791
Full Column of Literature Test	0.767	0.804	0.900	0.849	0.782
Abellán et al. [15] Train	0.767	0.802	0.906	0.850	0.784
Abellán et al. [15] Test	0.762	0.797	0.905	0.848	0.778
Feature Selection Train	0.779	0.815	0.903	0.857	0.797
Feature Selection Test	0.770	0.808	0.898	0.850	0.787
Full Column Train	0.752	0.785	0.910	0.843	0.771
Full Column Test	0.755	0.789	0.911	0.845	0.764

3.4. Traffic Accident Factors to Explore

With the progress of the times, the city is booming, and the pace of the people is becoming faster. Taiwanese consider convenience and speed, and the majority of commuters in Taiwan choose private transportation. The number of vehicles is increasing every year, which also raises the chance of traffic accidents. Most traffic accidents can be prevented in advance. However, the main factors that cause traffic accidents are unpredictable. Traffic accidents are caused by a variety of factors. If high-risk traffic accidents can be predicted, the government can prevent accidents and promote traffic safety in advance. Scholars have suggested factors that may lead to traffic accidents in the past, but they have considered different factors and their results have varied. At present, there is no unified definition of relevant indicators of traffic accidents that can provide transportation agencies for policy development and advocacy. However, traffic accidents increase social costs, so it is important to find out the factors that affect traffic accidents.

Taiwan has been promoting the disclosure of government information since 2011. There are many traffic-related datasets accumulated. Hence, this study obtains traffic accident datasets from the government open data platform, applies data mining to analyze these datasets, and constructs a high-risk traffic accident prediction model. It is hoped that the results of this study can provide a reference for government-related units to conduct safety promotion and reduce the occurrence of traffic accidents.

After the literature review, the factors that may cause traffic accidents were compiled and then compared with the data provided by the government platform, and finally, 21 factors were concluded to construct a high-risk traffic accident prediction model. Next, this study uses various data classifiers to construct the model. The preliminary experimental results show that the CHAID Tree, C5.0 Tree, and NN classifier have the best prediction results, so these three classifiers are used for Boosting integrated learning and parameter tuning to improve the model accuracy. The results of experiment 2 show that the classifier has at least 73% accuracy and an AUC of at least 0.693. Among them, the classifier with the best predictability is NN-MLP boosting, with 76.7% accuracy and 0.782 AUC. This means that the predictability of the traffic accident model is good. Experiment 3 investigates the combination of traffic accident factors. The results of this experiment use feature selection to screen factors, which is the best predictive method, with 77.0% accuracy and 0.787 AUC.

From the experimental results of the best model, the first approach to NN-MLP boosting (FS NN-MLP boosting), it was observed that the most important factors influencing the number of injuries in a traffic accident were the part of the vehicle initially hit, the vehicle type, the time, the degree of injury, the accident category and type, and the major cause. Observing the experimental results, the best model is FS NN-MLP boosting, and the results show that the factors affecting the number of traffic accident injuries include the initial impact of the vehicle, the vehicle type, the time, the degree of injury, the type and type of accident, and the main cause.

The results of the important factors affecting the accidents are explained as follows. For original heavy-duty motorcycles, the main cause of the accident is the failure to pay attention to the state of the front of the vehicle (23), which caused the most injuries, followed by the inability to clarify the factors (43), and the third is the failure to give way to the vehicle (6). In the case of personal-use sedans, the main cause of the accident is the failure to let the car in accordance with the regulations (6) lead to the largest number of injuries, followed by the failure to let the car in accordance with the regulations (6), and the third is the inability to clarify the factors (43). Observing the factor of the degree of injury, the party involved in the original heavy-duty motorcycle accident is bound to be injured (degree of injury, 2). In addition, the number of injuries caused by original heavy-duty motorcycle accidents is more than twice that of personal-use sedans.

The results of this study have made some contributions to predicting the occurrence of traffic accidents in Taiwan. The study's results can be compared to a previous study conducted by Ardakani et al. [32]. The study conducted by Ardakani et al. [32] proposed a predictive model based on different road accident data, including the intensity of the accident, the number of cars, and accidents. Hence, this research used a pre-processing framework to eradicate the meaningless data. Ardakani et al.’s [32] study used multinomial logistic regression, random forest, naïve Bayes and decision trees as the classification methodologies. According to the findings of the data-driven study by Ardakani et al. [32], three algorithms produced the results with an accuracy of 60 to 80 percent. The casualties and intensity of accidents were higher than 80 percent; the number of cars was less than 64 percent. Furthermore, according to the records of the study conducted by Ardakani et al. [32], the minor accidents ratio was more than 90 percent.

In addition, a previous study conducted by Ding et al. [33] proposed that the intensity and number of bus road accidents increase yearly. Ding et al.’s [33] study employed data from the public transportation company of Chongqing Liangjiang; the data was related to road accidents, service and driver safety traffic mistakes from January 2022 to June 2022. Ding et al.’s [33] study used SVM, XGBoost, BP neural network, extra trees and gradient boosting tree as the prediction framework to investigate traffic safety violations. Furthermore, twenty-seven-point nine percent of traffic accidents were impacted by safety violations, whereas vehicle safety operations accounted for twenty percent of traffic accidents. Finally, vehicle service violations accounted for sixteen-point five percent of traffic accidents.

Additionally, according to another recent study by Shunshun et al. [34], urbanization has become one of the major causes of traffic accidents and urban safety. Hence, Shunshun et al.’s [34] study used the traffic accident prediction model from the research published in English journals and provided a detailed literature on the traffic accident prediction methodologies. In addition, Shunshun et al.’s [34] study further provides in-depth descriptions related to conventional statistical methods, neural networks, machine learning, data mining, and time series analysis. The methodologies are efficient for analyzing the road accident factors, prevention factors to ensure urban safety, building and enhancing prediction models. Finally, the limitations and future research opportunities for the development of traffic prediction models were discussed.

Moreover, according to a similar study conducted by Almanie et al. [35], road accidents are deemed a serious issue and are considered to be an intense issue impacting social well-being. Hence, a reduction in road accidents can be an essential demand to ensure public safety. Almanie et al.’s [35] study employed two data mining models built on naïve bayes and decision trees. The data was classified based on the road's features, the accident's timing, and the weather conditions. The accident data was collected from 2016 to 2021 in Virginia. Almanie et al.’s [35] study also used ANOVA, paired observations, and visual tests as three quantitative analyses. Finally, Almanie et al.’s [35] study implied that decision trees were found to be the most accurate based on the experimental results.

Finally, another parallel study conducted by Khabiri et al. [36] followed a somewhat similar approach. Khabiri et al.’s [36] study contained huge amounts of data to process. Hence, their study used data mining approaches like decision trees to gain information on ensuring road safety and the factors necessary to avoid traffic accidents. The main purpose of Khabiri et al.’s [36] study was to employ this tool to control and reduce traffic issues with the aid of data mining. According to the results of Khabiri et al.’s [36] study, there was an increase in traffic assistance by 41. Furthermore, there was no significant difference in traffic accidents. Hence, the smart relay station strategies are grouped with the smart transportation system. The smart transportation tool contains supervision, implementation, and service tools, including traffic improvement and assistance.

4. Implications of the Study

First, this research designs a process for analyzing traffic accident data and constructs a prediction model for high-risk accidents. The results should be able to provide relevant government agencies to assess the key factors of traffic accidents, conduct safety advocacy for high-risk traffic factors, strengthen the importance of people's driving safety, and reduce the social costs caused by traffic incidents.

Second, the experimental results show that the problem of traffic accidents can be effectively analyzed. This research process uses data downloaded from the government open data platform and applies data mining methods to construct a predictive traffic accident model. In the process, this research overcomes various difficulties in data collection and building a data platform. Then, through experiments, it was found that the predictability of SVM was much higher than that of other classifiers. In addition, considering that the classifier may be overtrained, the cross-validation method is introduced to confirm that the prediction result of SVM is indeed overfitting, and it is also found that the NN classifier is the algorithm with the highest stability and the best predictability.

Third, this study compiles the factors discussed in previous studies, collects a dataset of traffic accidents in Taiwan from the government open data platform, and then uses feature selection to extract the important factors that cause traffic accidents and build a training dataset. From 70 factors, 17 factors that have a significant impact on the occurrence of high-risk accidents were selected, including month, hour, week, weather, light, road category, speed limit, road type, road condition, diverging facilities, fast and slow lane interval, accident category and type, primary cause, injury level, protective equipment, initial vehicle impact area, and vehicle type. Applying the feature selection method to establish the important factor combination data set has significantly improved the predictability, analysis time, and efficiency of the factors compared to without using this method. In addition, this study combined the factors discussed in the literature with the factors in the Taiwan traffic accident data set. There are 13 factors that appear most frequently in these two sources. Therefore, this study established a literature factor combination dataset, and the results showed that the predictability of this factor combination dataset was poor.

Finally, this study analyzes the main influences that cause traffic accidents. The best predictive model, FS NN-MLP boosting, was used to explore the factors with the experimental results. The factors used for NN modeling include the initial impact site of the vehicle, vehicle type, time, injury level, accident type, and the main cause, and these six factors are illustrated graphically. Among the factors of accident type, side-impact and vehicle-to-vehicle accidents have been the most frequent types of traffic accidents over the years. The most frequent time periods for accidents were 7-8 a.m. and 17–18 p.m. From the perspective of vehicle types, major accidents, and injuries, drivers of original heavy-duty motorcycles fail to pay attention to the conditions in front of the vehicle, resulting in injuries that are more than twice those of personal-use sedans.

5. Conclusion

This research employed the government's open traffic accident data of Taoyuan City from 2012 to 2017. This study used cross-validation to evaluate the model during the training process, and the multilayer perceptron neural network (MLPNN) classifier was employed to test the accuracy and performance of the model. Furthermore, a boosting ensemble learning approach and a combination of traffic accident factors enhanced the experiment's performance. This research designed a process for analyzing traffic accident data and constructed a prediction model for high-risk accidents. The experimental results showed that the problem of traffic accidents can be effectively analyzed. The research process used data downloaded from the government open data platform and applied data mining methods to construct a predictive traffic accident model. The study compiled the factors discussed in previous studies and collected a dataset of traffic accidents in Taiwan from the government open data platform, and then used feature selection to extract the important factors that caused traffic accidents and built a training dataset. This study also analyzed the main influences that cause traffic accidents. This study indicated that the degree of injury, the part of the vehicle hit, the type of accident, the leading cause, the type of vehicle, and the period of the accident were the main factors causing dangerous traffic accidents.

5.1. Research Limitations and Future Research Directions

The main research limitation of this study is that the data analysis was performed using commercial software. As a result, the model design and data presentation are relatively limited, which makes it difficult to adjust the details of the model computation process in terms of details or parameters. For future research, it is recommended to try to test the disease analysis model with more flexible algorithms. Hence, it will enable future researchers to offer different insights and present the data visually with the help of a diverse set of graphs and charts.

6. Declarations

6.1. Author Contributions

Conceptualization, A.R., Z.W., and S.C.; methodology, A.R., O.S., and Z.W.; validation, A.K. and S.C.; formal analysis, Z.W.; resources, S.C.; data curation, Z.W.; writing—original draft preparation, A.R., O.S., Z.W., A.K., and S.C.; writing—review and editing, A.R., O.S., Z.W., A.K., and S.C.; visualization, A.K. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. 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.

7. References

[1] Qiu, C., Wang, C., Fang, B., & Zuo, X. (2014). A multiobjective particle swarm optimization-based partial classification for accident severity analysis. Applied Artificial Intelligence, 28(6), 555-576. doi:10.1080/08839514.2014.923166.

[2] Ernstberger, A., Joeris, A., Daigl, M., Kiss, M., Angerpointner, K., Nerlich, M., & Schmucker, U. (2015). Decrease of morbidity in road traffic accidents in a high income country - An analysis of 24,405 accidents in a 21 year period. Injury, 46, S135–S143. doi:10.1016/S0020-1383(15)30033-4.

[3] Torayan, T., Laych, K., Peden, M., Kurg, E., Heimann, A., Senisse, A., ... & Chowdhury, S. (2015). Global Statue Report on Road Safety 2015. World Health Organization: Geneva, Switzerland.

[4] Chan, Y. S., Chen, C. S., Huang, L., & Peng, Y. I. (2017). Sanction changes and drunk-driving injuries/deaths in Taiwan. Accident Analysis and Prevention, 107, 102–109. doi:10.1016/j.aap.2017.07.025.

[5] MOTC. (2024). Freeway Bureau: Traffic Volume Survey. Ministry of Transportation and Communications, New Taipei City, Taiwan.

[6] Chen, W. H., & Jovanis, P. P. (2000). Method for identifying factors contributing to driver-injury severity in traffic crashes. Transportation Research Record, 1717, 1–9. doi:10.3141/1717-01.

[7] Sadollah, A., Gao, K., Zhang, Y., Zhang, Y., & Su, R. (2019). Management of traffic congestion in adaptive traffic signals using a novel classification-based approach. Engineering Optimization, 51(9), 1509–1528. doi:10.1080/0305215X.2018.1525708.

[8] Nallaperuma, S., Jalili, S., Keedwell, E., Dawn, A., & Oakes-Ash, L. (2020). Optimisation of Signal Timings in a Road Network. Springer Proceedings in Complexity, 257–268. doi:10.1007/978-3-030-40943-2_22.

[9] Alkheder, S., Taamneh, M., & Taamneh, S. (2017). Severity Prediction of Traffic Accident Using an Artificial Neural Network. Journal of Forecasting, 36(1), 100–108. doi:10.1002/for.2425.

[10] Castro, Y., & Kim, Y. J. (2016). Data mining on road safety: Factor assessment on vehicle accidents using classification models. International Journal of Crashworthiness, 21(2), 104–111. doi:10.1080/13588265.2015.1122278.

[11] Kumar, S., & Toshniwal, D. (2015). A data mining framework to analyze road accident data. Journal of Big Data, 2(1), 26. doi:10.1186/s40537-015-0035-y.

[12] Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. Journal of Modern Transportation, 24(1), 62–72. doi:10.1007/s40534-016-0095-5.

[13] Zeng, Q., Wen, H., & Huang, H. (2016). The interactive effect on injury severity of driver-vehicle units in two-vehicle crashes. Journal of Safety Research, 59, 105–111. doi:10.1016/j.jsr.2016.10.005.

[14] Van Beeck, E. F., Borsboom, G. J. J., & Mackenbach, J. P. (2000). Economic development and traffic accident mortality in the industrialized world, 1962-1990. International Journal of Epidemiology, 29(3), 503–509. doi:10.1093/intjepid/29.3.503.

[15] Abellán, J., López, G., & De Oña, J. (2013). Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications, 40(15), 6047–6054. doi:10.1016/j.eswa.2013.05.027.

[16] Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys and Tutorials, 18(2), 1153–1176. doi:10.1109/COMST.2015.2494502.

[17] Zou, Q., Zeng, J., Cao, L., & Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing, 173, 346–354. doi:10.1016/j.neucom.2014.12.123.

[18] Yu, Z., Haghighat, F., & Fung, B. C. M. (2016). Advances and challenges in building engineering and data mining applications for energy-efficient communities. Sustainable Cities and Society, 25, 33–38. doi:10.1016/j.scs.2015.12.001.

[19] Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, 2(2), 43–47. doi:10.13189/wjcat.2014.020203.

[20] Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Applied Computing and Informatics, 12(1), 90–108. doi:10.1016/j.aci.2014.10.001.

[21] Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of Pharmaceutical and Biomedical Analysis, 22(5), 717–727. doi:10.1016/S0731-7085(99)00272-1.

[22] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. Atmospheric Environment, 32(14–15), 2627–2636. doi:10.1016/S1352-2310(97)00447-0.

[23] West, D. (2000). Neural network credit scoring models. Computers and Operations Research, 27(11–12), 1131–1152. doi:10.1016/S0305-0548(99)00149-5.

[24] Er, M. J., Wu, S., Lu, J., & Toh, H. L. (2002). Face recognition with radial basis function (RBF) neural networks. IEEE Transactions on Neural Networks, 13(3), 697–710. doi:10.1109/TNN.2002.1000134.

[25] Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. Bioinformatics, 21(15), 3301–3307. doi:10.1093/bioinformatics/bti499.

[26] Wijnen, W., & Stipdonk, H. (2016). Social costs of road crashes: An international analysis. Accident Analysis and Prevention, 94, 97–106. doi:10.1016/j.aap.2016.05.005.

[27] Bambach, M. R., & Mitchell, R. J. (2015). Estimating the human recovery costs of seriously injured road crash casualties. Accident Analysis and Prevention, 85, 177–185. doi:10.1016/j.aap.2015.09.013.

[28] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259–268. doi:10.1016/j.eswa.2014.07.040.

[29] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45(4), 427–437. doi:10.1016/j.ipm.2009.03.002.

[30] Maas, A. I. R., Hukkelhoven, C. W. P. M., Marshall, L. F., & Steyerberg, E. W. (2005). Prediction of outcome in traumatic brain injury with computed tomographic characteristics: A comparison between the computed tomographic classification and combinations of computed tomographic predictors. Neurosurgery, 57(6), 1173–1181. doi:10.1227/01.NEU.0000186013.63046.6B.

[31] Guha, R., Ghosh, M., Mutsuddi, S., Sarkar, R., & Mirjalili, S. (2020). Embedded chaotic whale survival algorithm for filter–wrapper feature selection. Soft Computing, 24(17), 12821–12843. doi:10.1007/s00500-020-05183-1.

[32] Ardakani, S. P., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. Sustainability (Switzerland), 15(7), 5939. doi:10.3390/su15075939.

[33] Ding, T., Zhang, L., Xi, J., Li, Y., Zheng, L., & Zhang, K. (2023). Bus Fleet Accident Prediction Based on Violation Data: Considering the Binding Nature of Safety Violations and Service Violations. Sustainability (Switzerland), 15(4), 3520. doi:10.3390/su15043520.

[34] Wang, S., Changshun, Y., & Yong, S. (2023). A Review of Road Traffic Accident Prediction Methods. American Journal of Management Science and Engineering. doi:10.11648/j.ajmse.20230803.12.

[35] Almanie, T. (2023). Quantitative Study of Traffic Accident Prediction Models: A Case Study of Virginia Accidents. The International Journal of Advanced Networking and Applications, 14(05), 5582–5589. doi:10.35444/ijana.2023.14501.

[36] Khabiri, M. M., Ghahfarokhi, F. M., Sarfaraz, S., & Anaie, H. M. (2022). Application of Data Mining Algorithm To Investigate the Effect of Intelligent Transportation Systems on Road Accidents Reduction By Decision Tree. Communications - Scientific Letters of the University of Žilina, 24(2), F36–F45. doi:10.26552/com.C.2022.2.F36-F45.