LONTAR KOMPUTER VOL. 14, NO. 1 APRIL 2023 p-ISSN 2088-1541
DOI : 10.24843/LKJITI.2023.v14.i01.p04 e-ISSN 2541-5832
Accredited Sinta 2 by RISTEKDIKTI Decree No. 158/E/KPT/2021

The Use of XGBoost Algorithm to Analyse the Severity of Traffic Accident Victims

I Made Sukarsa[a1], Ni Kadek Dwi Rusjayanthi[a2], Made Srinitha Millinia Utami[a3], Ni Wayan Wisswani[b4]

[a]Department of Information Technology, Udayana University
Badung, Indonesia
[1]sukarsa@unud.ac.id (Corresponding author)
[2]dwi.rusjayanthi@unud.ac.id
[3] milliniautami@student.unud.ac.id
[b]Department of Information System Management, Bali State Polytechnic
Badung, Indonesia
[4]wisswani@pnb.ac.id

Abstract

Traffic accidents are still significant contributors to a fairly high death. Denpasar's resort police record every traffic accident in the form of a daily report. The stored data can generate valuable information to improve policies and propagate better traffic practices. This research utilizes the classification technique with the XGBoost, random forest algorithm, and SMOTE method. The study shows that the SMOTE technique can increase the model's accuracy. Using the classification method with the two algorithms produces factors that affect the severity of traffic accident victims with feature importance. The feature importance obtained using the XGBoost model by counting the weight value for testing using the original dataset, the dataset for the type of two-wheeled vehicle, and the dataset of the kind of vehicle other than two-wheeled indicate that the variables influencing the severity of victims in road accidents are the time of accident between 00.00-06.00, the type of vehicle motorcycle, the type of opponent vehicle truck and pickup car, the age of the driver between 16-25, sub-district road status and front – side type of accident.

Keywords: *Accident Factors, Data Mining, Random Forest, XGBoost*.

1. Introduction

Traffic accidents are still one of the contributors to a relatively high death rate in Indonesia [1]. Three people die every hour in traffic accidents on average [2]. The number of casualties in Indonesia and what happened in Denpasar City tends to increase yearly. Based on data from the Indonesian National Police published by the Central Statistic Agency, traffic accidents in 2019 recorded 116,411 cases in Indonesia, an increase of about 6% from the previous year, which recorded 109,215 cases. The high traffic accidents must be a concern and get effective handling.

The Police of Denpasar City record every traffic accident case daily. Data generally recorded on a traffic accident is where, when, and how the accident occurred. The data that is recorded and stored can produce valuable information that can be used to create or improve policies related to traffic accidents. Data mining is a method that can be utilized to obtain relevant information that was not previously available.

Data mining is a technique for finding and obtaining potentially valuable knowledge from huge amounts of data [3]. Data mining aims to extract hidden information from large data blocks [4]. One method that can be used in data mining is classification. Classification is work related to categorizing a particular group of items into targeted groups and mapping each set of variables to each target [5]. The XGBoost algorithm is one of the classification algorithms in development right now. Recently, the machine learning applications and Kaggle competition for structured and tabular data have been dominated by the XGBoost algorithm. Gradient-boosted decision trees are implemented using XGBoost to enhance efficiency and performance. The XGBoost algorithm is used for research purposes that require speed in execution and good model performance [6].

Several research conducted for data mining development, such as Noh et al.'s data mining research for traffic accidents, led to the development of a novel model for potential pedestrian risk event (PPRE) analysis. The system automatically recognizes vehicles and pedestrians, computes passes, and extracts frame-level behavioral data. They used video footage captured by road security cameras for their study. These occurrences are divided into six clusters using K-means clustering and a decision tree, and the groupings are then visualized and analyzed to see how they affect pedestrian danger at these crossings. The findings are presented as potential benefits and restrictions of the data received from the video in terms of identifying scenarios and locations with a high potential for pedestrian risk incidents [9].

Research related to the implementation of the XGBoost method was conducted by Sukarsa et al. to forecast or calculate the appropriate supply to optimize revenue from gourami sales. The research aims to estimate gourami supplies using transaction data with the XGBoost algorithm. Five XGBoost models with various properties, including lag, rolling window, mean encoding and mix, were made in this study. According to the findings, the mixed feature model has an accuracy of 97.54%, an MAE of 0.063, and a MAPE of 2.64% [10].

Salahadin Seid Yassin and Pooja conducted research related to data mining using the K-Means and Random Forest methods to predict traffic accidents. Research shows that integrating clustering and classification can help to increase model accuracy and pinpoint the key contributing components directly from the acquired data. The accuracy obtained by adding new clusters and using the random forest classification algorithm is 99.86% [11].

Other supporting research related to traffic accidents has been successfully carried out by Comi et al. using data mining techniques to determine the significant factors that contribute to and the common trends in Rome's traffic accidents. The study implemented clustering approaches (K-Means Clustering and Kohonen Network) to analyze accident data from 2016 to 2019. The result shows that the kind of vehicle used is the most influential cause of accident severity [12]. Yuexu Zhao and Wai Deng also used the XGBoost algorithm to build a traffic accident prediction model. The study showed the model has good predictive accuracy and combines numerous models to highlight the variables significantly affecting outcomes. The ensemble model has advantages over elastic network regression, decision tree, and others. Accuracy, efficiency, and interpretability are balanced in ensemble learning techniques such as XGBoost [13]. Andri Irfan et al. developed a model that can predict accidents. The model was developed using Artificial Neural Network (ANN) and Support Vector Machine (SVM) data mining techniques to indicate and identify the factors that underlie accidents on toll roads in Indonesia. The study results show that the ANN method produces the best results for the model built compared to SVM and logistic regression [14].

The XGBoost algorithm has been used in previous studies and yielded promising results, such as predicting breast cancer. The research compared two classification algorithms, namely SVM and XGBoost. The number of data attributes is reduced using principal component analysis and clustering methods before classification. The results show that using XGBoost and PCA methods has a high accuracy rate of 97% [15]. Yu Jiang et al., through their research, used XGBoost to detect pedestrians. The built model combines XGBoost with genetic algorithms as tuning parameters as well as HOG and LBP features to describe pedestrians in tandem fusion. The result shows that the model can increase pedestrian detection accuracy on static images with an AUC value of 0.9913 [16]. Another study using the XGBoost algorithm for classification was conducted by Chen Wang et al. [17] and, respectively, for disease classification using the Parkinson dataset and text mining based on comments about taxes as big data. Both studies found that XGBoost is a good algorithm for classifying large amounts of data and data with unbalanced labels with the help of various additional features.

Based on the previous research, the method used to classify accident data and examine the most important variables influencing the severity of traffic accident victims in Denpasar City, this study uses the XGBoost and random forest algorithms. The XGBoost and random forest algorithms produce classification results with high accuracy from several previous studies [7][8] and have features for analyzing influencing factors or variables. To get a model with the highest level of accuracy, classification is done using a variety of test scenarios. By calculating the value of feature importance, the most effective model is used to discover the component that has the most impact on the severity of victims of traffic accidents. Supporting research for this research is taken from various journals and previous research. These studies have relevance that can be applied in this study.

2. Research Method

The proposed methodology performs preprocessing, builds models using several methods, and uses the best model to display the feature importance score to get features or variables that influence the class label, which is the severity level of the traffic accident victims.

2.1 Data

The resources taken in this study are traffic accident data in Denpasar, Indonesia, from 2020, with 546 lines of data. The traffic accident data used consists of several variables. Table 1 lists the variables.

Table 1. Variables List

Response Variable		Predictor Variables	
Variable Name	**Category**	**Variable Name**	**Category**
Victim severity	Material loss (Material), minor injury (LR), serious injury (LB), death (MD)	Type of accident	Front – back, front – side, pedestrian, front – front, side – side, pileup, solo, back – back
		Accident time	00.00-06.00, 06.00-12.00, 12.00-18.00, 18.00-00.00
		Road status	Sub-district, municipality, district, province, national
		Area	Shops, settlements, offices, mangroves, tourism, landscaping
		Gender	Male, Female
		Age	0-9, 10-15, 16-25, 26-30, 31-40, 41-50, 51-60, >60
		Last education	Primary school (SD), Junior High School (SMP), Senior High School (SMA), Vocational High School (SMK), College (Perguruan Tinggi)
		Driver license ownership	With SIM, no SIM
		Helmet use	Wearing a helmet, not wearing a helmet
		Vehicle type	Motorcycle, truck, pedestrian, car, pick up, bus, bicycle, heavy equipment car
		Opponent’s vehicle type	Motorcycle, truck, pedestrian, car, pick up, bus, bicycle, heavy equipment car

2.2 Research Flow

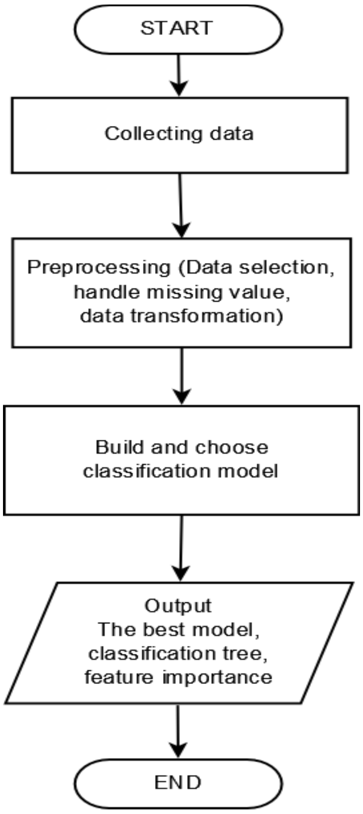

START
Collecting data
Preprocessing (Data selection, handle missing value, data transfomation)
Build and choose classification model
Output
The best model, classification tree, feature importance
END

Figure 1. Research Flow

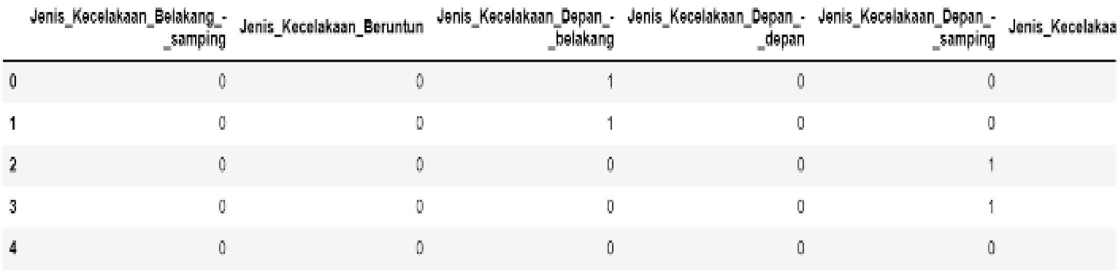
Data collection is the initial step in the research process. The data collected is in the form of detailed data on traffic accidents. The next stage is preprocessing. Data is selected at this stage, filling in missing values and converting categorical data to numeric. Data that is ready will be used for the modeling process. Modeling is carried out with various test scenarios using several methods, such as XGBoost and random forest for classification, RandomSearchCV, GridSearchCV for hyperparameter tuning, and the SMOTE method for resampling data.

2.3 Preprocessing

The processes involved in this stage include loading data, selecting data, filling in blank data (missing values), and converting categorical data into numeric. The loading data is the stage to display traffic accident data in Excel (.xlsx), which is stored as a data frame. The data selection stage is the stage for selecting columns from the data used for testing. The selected data can be used for testing and affects the label class. Variables were selected based on similar previous studies and weighted using the Chi-square test. The Chi-square test evaluates the correlation between the independent and dependent variables by removing features that are most likely class-independent and unnecessary for classification [18]. Based on feature selection, several columns on the dataset are deleted: number, place of accident, accident class, status, and occupation.

Furthermore, the missing value is then filled in. Filling in the missing value is a step to fill in the empty or NaN data value. Filling in the missing value is intended so that data with empty values is not deleted so as not to reduce the number of rows of data, and the data can be used for testing. Filling in missing values or data with empty values is done by filling in the mode value from all data in that column. This is done because the data is in the categorical form [19].

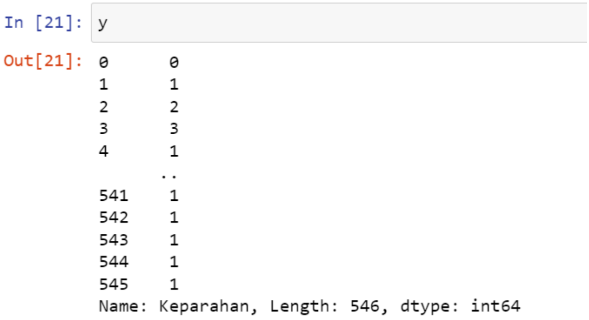
The next stage is to change the data in the categorical form into data in numerical form. The purpose of changing the data is to be processed using XGBoostClassifier, which only accepts input in the form of numerical data.



	Jenis_Kecelakaan_Belakang_-_samping	Jenis_Kecelakaan_Beruntun	Jenis_Kecelakaan_Depan_-_belakang	Jenis_Kecelakaan_Depan_-_depan	Jenis_Kecelakaan_Depan_-_samping	Jenis_Kecelakaa
0	0	0	1	0	0	
1	0	0	1	0	0	
2	0	0	0	0	1	
3	0	0	0	0	1	
4	0	0	0	0	0	

Figure 2. Data X in numerical form

Figure 2 is the result of converting X data, which contains predictor variables in categorical form to numeric. The data is converted to numeric by One-hot-encode using the get_dummies library [20]. Each data category is transformed into a column, and a value of 1 in each row indicates that the category belongs to each data row.



```
In [21]: y

Out[21]: 0      0
         1      1
         2      2
         3      3
         4      1
               ..
         541    1
         542    1
         543    1
         544    1
         545    1
         Name: Keparahan, Length: 546, dtype: int64
```

Figure 3. Data y in numerical form

Figure 3 results from converting data y from categorical to numerical form. Data y, the target class, is changed by mapping for each data. Each value in the array represents a class of labels that will be predicted using the model.

2.4 Build Model

The model is built with several test scenarios with various methods and train and test data splits.

2.4.1 XGBoost

The first model is built with the XGBoost algorithm without hyperparameter tuning. One of the ensemble learning algorithms and a boosting algorithm is XGBoost [21]. The application of ensemble learning trains many models (weak learners) to solve problems and is used to make predictions. Ensemble learning helps reduce the difference between actual and predicted values. Based on initial training data, the sequential ensemble process known as boosting creates weak learners. Weak learners are further developed and correct errors in previously weak learners. To create the final prediction model, all weak learners are combined across numerous iterations [22].

2.4.2 RandomSearch CV

The second model is built with XGBoost and RandomSearch CV for hyperparameter tuning. The RandomSearch CV algorithm will look for possible combinations of parameters inputted randomly to produce the best combination.

Table 2. Parameter Setting for RandomSearch CV Model

Parameter	Value
learning_rate	[0.001, 0.1, 0.1, 0.25, 0.5, 0.4]
max_depth	[1, 2, 3, 4, 5, 6]
max_features	[1, 2, 3, 4, 5, 6]
n_estimators	[20, 40, 50, 70, 100]

Table 2 shows the parameters used to build and run the second model. The second model was built using RandomSearchCV. The parameter's value will be chosen randomly to get the best result.

Table 3. XGBoost Parameter Function

Parameter	Function
learning_rate	Step size shrinkage was used in the update to prevent overfitting*. Range* parameter dari 0 sampai 1. Default = 0.3.
max_depth	Maximum depth of a tree. Increasing this value will make the model more complex and likely to overfit—default = 6.
max_features	Determines the number of features considered when searching for the best split.
n_estimators	Determine the number of models built.

Table 3 shows the parameters used for setting the built model. The parameter details described include the name and the function of the parameter. The parameters used for testing the XGBoost model are learning_rate, max_depth, max_features, and n_estimators.

2.4.3 GridSearch CV

The third model was built using GridSearch CV. GridSearch CV aims to perform validation for more than one model and each hyperparameter automatically and systematically. The parameters used for development and testing using the third model are shown in Table 4.

Table 4. Parameter Setting for GridSearch CV Model

Parameter	Value
learning_rate	[0.001, 0.1, 0.1, 0.25, 0.5, 0.4]
max_depth	[1, 2, 3, 4, 5, 6]
max_features	[1, 2, 3, 4, 5, 6]
n_estimators	[20, 40, 50, 70, 100]

Table 4 shows the parameter details for the third model created with GridSearchCV. Furthermore, the previously created XGBoost model is run with a grid search. Parameter adjustment with grid search will try all possible combinations of parameters and find the best one.

2.4.4 SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is a well-known resampling method for dealing with class imbalance problems. Resampling is useful for balancing classes, clearing border areas, and enlarging minority class areas. The SMOTE technique helps to increase generalizability by creating new data rows of the minority class. The basis of this method is to perform interpolation among neighbor minority class instances. The method has the benefit of quick calculation speed and the successful provision of a balanced and accurate classification performance [23].

p-ISSN 2088-1541
e-ISSN 2541-5832

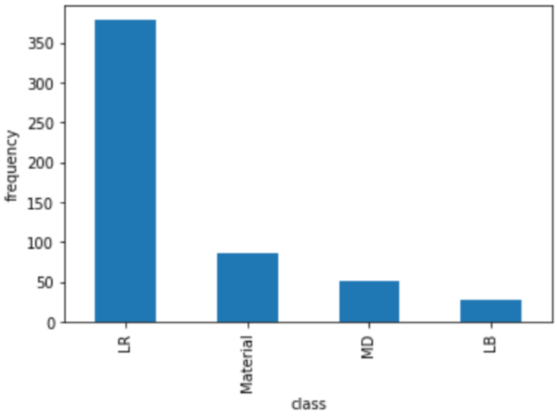


Figure 4. Class Distribution Graph of the Initial Dataset

The fourth model is built using the SMOTE method. Figure 4 shows the distribution of classes in the initial dataset. The frequency of the data that appears for each class is unbalanced, with the amount for each class being much different. In order to balance the dataset, the SMOTE method generates a new sample from the minority class.

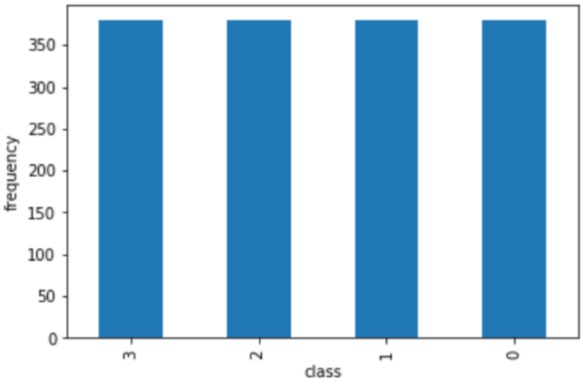


Figure 5. Class Distribution Graph of Balanced Dataset

The graph in Figure 5 shows the distribution of classes from the balanced dataset that has been resampled using the SMOTE method. The frequency of data for each class is relatively the same, in contrast to before resampling using SMOTE, as shown in Figure 4. The data generated by creating new samples for the minority class resulted in 1515 rows from 459 rows of the initial dataset.

2.4.5 Random Forest

Random forest is a classification and regression method in the form of decision trees. The random forest method uses the bootstrap aggregation (bagging) technique. By randomly selecting observations and attributes from the training set, bagging decreases high variance [24]. Each tree in the random forest model is a Classification and Regression Tree (CART), which uses reduced impurity in selecting a separator predictor from a randomly selected subset of all available predictor variables. Class determination is taken based on the majority of the votes from all trees formed. The following are the phases of modeling using the random forest algorithm [25].

a. Randomly select n training samples from the original dataset using Bootstrap.
b. k training sets are obtained once k rounds of extraction are completed.
c. Training k training sets for k decision tree models.
d. The average of each model’s prediction results is the final result for the charging load prediction problem.

The fifth model is built using RandomForestClassifier. RandomForestClassifier performs classification with a random forest algorithm.

Table 5. Parameter Setting for Random Forest Model

Parameter	Value
n_estimators	100
random_state	0
criterion	‘entropy’

Table 5 shows the parameters for running the fifth model with the random forest algorithm. The details and function of each parameter can be seen in Table 6.

Table 6. Random Forest Parameter Function

Parameter	Function
n_estimators	The number of trees in the forest. Default = 100
random_state	Controls the randomness of the bootstrapping of the samples used when building trees. Default = None
criterion	The function is to measure the quality of a split. Default = Gini

Table 6 shows the details of the parameters used for setting the random forest model that was built. The parameter details described include the name and function of the parameter. The parameters used for testing the random forest model are n_estimators, random_state, and criterion.

3. Result and Analysis

The best model was selected from developing and testing several classification models. The classification model is made using two classification algorithms, namely XGBoost and random forest, with several additional methods, such as RandomSearchCV and GridSearchCV for hyperparameter tuning and the SMOTE method for resampling data from minority classes from datasets with unbalanced classes.

Table 7. Classification Model Test Results

Method	Train and Test Size	Accuracy
XGBoost	Train: 0.8	0.88
	Test: 0.2	0.80
	Train: 0.7	0.90
	Test: 0.3	0.81
	Train: 0.6	0.89
	Test: 0.4	0.81
XGBoost with RandomSearchCV	Train: 0.8	0.82
	Test: 0.2	0.80
	Train: 0.7	0.80
	Test: 0.3	0.81
	Train: 0.6	0.82
	Test: 0.4	0.81
XGBoost with GridSearchCV	Train: 0.8	0.80
	Test: 0.2	0.83
	Train: 0.7	0.80
	Test: 0.3	0.82
	Train: 0.6	0.80
	Test: 0.4	0.81

Method	Train and Test Size	Accuracy
XGBoost + SMOTE	Train: 0.8	0.99
	Test: 0.2	0.90
	Train: 0.7	0.97
	Test: 0.3	0.90
	Train: 0.6	0.98
	Test: 0.4	0.89
Random Forest	Train: 0.8	0.97
	Test: 0.2	0.79
	Train: 0.7	0.97
	Test: 0.3	0.81
	Train: 0.6	0.98
	Test: 0.4	0.79
Random Forest + SMOTE	Train: 0.8	0.99
	Test: 0.2	0.93
	Train: 0.7	0.99
	Test: 0.3	0.90
	Train: 0.6	0.99
	Test: 0.4	0.89

The test results employing all the constructed models are used to choose the model with the highest accuracy value and the lowest error value. The best model chosen is built with a combination of XGBoost and SMOTE methods, and the model with random forest and SMOTE with training and testing data distribution of 80% and 20%, respectively.

Feature importance is a technique to calculate how much influence features or variables have on the target class of the model built. Determination of the influential variables using the frequency of selection of these variables as a sorter to divide the data across all trees (F-score). The greater the frequency of the variable chosen, the greater the influence of the variable in the modeling. XGBoost provides a feature to display the value of feature importance directly.

Regarding the existence of irrelevant variables, if used for all types of vehicles, a test was carried out by dividing the data into data for two-wheeled vehicles and data for other than two-wheeled vehicles to get a better result of feature importance and in accordance with the original dataset. The feature importance test for data with the type of two-wheeled vehicle includes "Helmet Use" as one of the variables.

Table 8. Feature Importances

Dataset	Method	Feature Importance
Original with feature selection	XGBoost + SMOTE + GridSearchCV	1. Accident time 00.00-06.00 2. Age 16-25 3. Front–side type of accident 4. Motorcycle vehicle type 5. Accident time 06.00-12.00
Original with feature selection	XGBoost + GridSearchCV	1. Shops area 2. Accident time 00.00-06.00 3. Front–side type of accident 4. Drivers with SIM (driver's license) 5. Age 16-25
Data with the	XGBoost + GridSearchCV	1. Sub-district road status

Dataset	Method	Feature Importance
two-wheeled type of vehicle		2. Accident time 18.00-00.00 3. Opponent's vehicle type truck 4. Opponent's vehicle type pickup 5. District road status
Data with other than the two-wheeled type of vehicle	XGBoost + GridSearchCV	1. Opponent's vehicle type pickup 2. Pedestrian 3. Age more than 60 years old 4. Male gender 5. Sub-district road status

The feature importance graph generated from several tests shows that each test gets different results. It happens because the missing value is filled in by filling in the mode value of the entire data in the column, where the data for each test is different. The test also shows that not all data is suitable for resampling using the SMOTE method. Resampling is done by adding a new data line from the minority class so that the resulting feature importance value differs from the original data.

Several tests that have been carried out have shown that the most influential factors on the severity of traffic accident victims in Denpasar City are the time of the accident between 00.00-06.00, the type of vehicle motorcycle, the type of opponent vehicle, a truck and pickup car, the age of the driver between 16-25, sub-district road status and front-side accident type. The test results will be given to the policymakers and related institutions that can then be used to make and improve policies as a consideration in the design and implementation of traffic safety improvement programs and other matters related to handling traffic accidents, especially in Denpasar City.

Table 9. Suggestions for Applying Feature Importance

Variable(s) Name	Suggestion
Front–side type of accident	Continuous installation of road signs and markings tailored to the needs, installation of special signs to indicate the direction of the bend (chevron), improvement of intersection layout for traffic conflict management and providing adequate visibility for drivers, use of colored materials on the center line of the road, construction improvements the road so that there are no potholes and damage the normal slope of the road (collapse).
Motorcycle type of vehicle, age of driver 16-25 years old	Focusing on traffic safety programs associated with motorcycle users and on the age group of 16-25 years, which is equivalent to the age of students and college students.
Road status sub-district	Installing warning banners, adding traffic signs, and checking road conditions.
Accident time between 00.00-06.00	Installation of warning banners and adding traffic signs.

Table 9 provides examples of recommendations for the practical implementation of the analysis of factors affecting the severity of victims of traffic accidents sourced from the traffic accident module [31] that can be given related to handling traffic accidents based on the value of feature importance obtained. Feature importance displays the variables or factors that most influence the severity of traffic accident victims.

4. Conclusion

Model development and testing are carried out with several test scenarios to obtain a model with the best level of accuracy. The model with the best level of accuracy is used to find the value of feature importance in determining the severity of traffic accident victims. According to the test findings, the best model is the model with a combination of XGBoost and SMOTE methods with the distribution of training and testing data of 80% and 20%, respectively, with an accuracy value of 99% for training data and 90% for test data. Feature importance obtained using the XGBoost model by taking into account the weight value is carried out with several tests using the original dataset and the dataset that has been separated into data with the type of two-wheeled vehicle and data with the type of vehicle other than two-wheeled. Several tests that have been carried out have shown that the most influential factors in traffic accidents are the time of the accident between 00.00-06.00, the type of vehicle motorcycle, the type of opposing vehicle truck and pickup car, the age of the driver between 16-25, sub-district road status and front – side type of accident.

References

[1] I. F. Anshori and Y. Nuraini, “Pengelompokan Data Kecelakaan Lalu Lintas di Kota Tasikmalaya Menggunakan Algoritma K-Means,” *Jurnal Responsif: Riset Sains dan Informatika*, vol. 2, no. 1, pp. 118–127, 2020, doi: 10.51977/jti.v2i1.198.

[2] Marroli, “Rata-rata Tiga Orang Meninggal Setiap Jam Akibat Kecelakaan Jalan.” https://kominfo.go.id/index.php/content/detail/10368/rata-rata-tiga-orang-meninggal-setiap-jam-akibat-kecelakaan-jalan/0/artikel_gpr (accessed Mar. 08, 2022).

[3] J. Yang *et al.*, “Brief introduction of medical database and data mining technology in big data era,” *J Evid Based Med*, vol. 13, no. 1, pp. 57–69, 2020, doi: 10.1111/jebm.12373.

[4] R. R. Asaad and R. M. Abdulhakim, “The Concept of Data Mining and Knowledge Extraction Techniques,” *Qubahan Academic Journal*, vol. 1, no. 2, pp. 17–20, 2021, doi: 10.48161/qaj.v1n2a43.

[5] A. O. Adebayo and M. S. Chaubey, “Data Mining Classification Techniques on the Analysis of Student’s Performance,” *Global Scientific Journals*, vol. 7, no. 4, pp. 79–95, 2019, [Online]. Available: www.globalscientificjournal.com.

[6] J. Brownlee, *XGBoost With Python*. 2018.

[7] C. Zhang *et al.*, “Cause-aware failure detection using an interpretable XGBoost for optical networks,” *Optics Express*, vol. 29, no. 20, p. 31974, 2021, doi: 10.1364/oe.436293.

[8] P. Song and Y. Liu, “An xgboost algorithm for predicting purchasing behaviour on e-commerce platforms,” *Technical Gazette*, vol. 27, no. 5, pp. 1467–1471, 2020, doi: 10.17559/TV-20200808113807.

[9] B. Noh, W. No, J. Lee, and D. Lee, “Vision-Based Potential Pedestrian Risk Analysis on Unsignalized Crosswalk Using Data Mining Techniques,” *Applied Sciences*, vol. 10, no. 3, 2020, doi: 10.3390/app10031057.

[10] I. M. Sukarsa, N. N. Pandika Pinata, N. Kadek Dwi Rusjayanthi, and N. W. Wisswani, “Estimation of Gourami Supplies Using Gradient Boosting Decision Tree Method of XGBoost,” *TEM Journal*, vol. 10, no. 1, pp. 144–151, 2021, doi: 10.18421/TEM101-17.

[11] S. S. Yassin and Pooja, “Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach,” *SN Applied Sciences.*, vol. 2, no. 9, pp. 1–13, 2020, doi: 10.1007/s42452-020-3125-1.

[12] A. Comi, A. Polimeni, and C. Balsamo, “Road Accident Analysis with Data Mining Approach: evidence from Rome,” *Transportation Research Procedia*, vol. 62, no. Ewgt 2021, pp. 798–805, 2022, doi: 10.1016/j.trpro.2022.02.099.

[13] Y. Zhao and W. Deng, “Prediction in Traffic Accident Duration Based on Heterogeneous Ensemble Learning,” *Applied Artificial Intelligence.*, vol. 00, no. 00, pp. 1–24, 2022, doi: 10.1080/08839514.2021.2018643.

[14] A. Irfan, R. Al Rasyid, and S. Handayani, “Data mining applied for accident prediction model in Indonesia toll road,” *AIP Conference Proceedings*, vol. 1977, no. June 2018, 2018, doi: 10.1063/1.5043013.

[15] A. Jamal, A. Handayani, A. A. Septiandri, E. Ripmiatin, and Y. Effendi, “Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction,” *Lontar*

Komputer: Jurnal Ilmiah Teknologi Informasi, vol. 9, no. 3, p. 192, 2018, doi: 10.24843/lkjiti.2018.v09.i03.p08.

[16] Y. Jiang, G. Tong, H. Yin, and N. Xiong, "A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters," *IEEE Access*, vol. 7, pp. 118310–118321, 2019, doi: 10.1109/ACCESS.2019.2936454.

[17] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," *Pattern Recognition Letters*, vol. 136, pp. 190–197, 2020, doi: 10.1016/j.patrec.2020.05.035.

[18] J. Wang, J. Xu, C. Zhao, Y. Peng, and H. Wang, "An ensemble feature selection method for high-dimensional data based on sort aggregation," *Systems Science & Control Engineering.*, vol. 7, no. 2, pp. 32–39, 2019, doi: 10.1080/21642583.2019.1620658.

[19] J. Poulos and R. Valle, "Missing Data Imputation for Supervised Learning," *Applied Artificial Intelligence.*, vol. 32, no. 2, pp. 186–196, 2018, doi: 10.1080/08839514.2018.1448143.

[20] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00305-w.

[21] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[22] K. Liu, Z. Dai, R. Zhang, J. Zheng, J. Zhu, and X. Yang, "Prediction of the sulfate resistance for recycled aggregate concrete based on ensemble learning algorithms," *Construction and Building Materials*, vol. 317, no. November 2021, p. 125917, 2022, doi: 10.1016/j.conbuildmat.2021.125917.

[23] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification," *Pattern Recognition.*, vol. 124, 2022, doi: 10.1016/j.patcog.2021.108511.

[24] Y. Lu *et al.*, "The application of improved random forest algorithm on the prediction of electric vehicle charging load," *Energies*, vol. 11, no. 11, 2018, doi: 10.3390/en11113207.

[25] P. P. dan P. J. P. P. dan P. I. Wilayah, "Data Kecelakaan Lalu Lintas Tahun 2016," vol. 53, no. 9, 2016.