Automotive Experiences
Vol. 7 No. 2 (2024) pp. 236-251
p-ISSN: 2615-6202 e-ISSN: 2615-6636

Research Paper

The Road Safety: Utilising Machine Learning Approach for Predicting Fatality in Toll Road Accidents

Mutharuddin[1], **M. Rosyidi**[2], **Djoko Wahyu Karmiadji**[3,4], **Hastiya Annisa Fitri**[1], **Novi Irawati**[1], **Dwitya Harits Waskito**[1], **Tetty Sulastry**[1], **Subaryata**[1], **Sinung Nugroho**[1]

[1]Research Center for Transportation Technology, National Research and Innovation Agency, Indonesia
[2]Computation Research Center, National Research and Innovation Agency, Indonesia
[3]Research Center for Strength Structures Technology, National Research and Innovation, Indonesia
[4]Department of Mechanical Engineering, Pancasila University, Jakarta 12640, Indonesia

novi.irawati@brin.go.id

https://doi.org/10.31603/ae.11082

Published by Automotive Laboratory of Universitas Muhammadiyah Magelang

Article Info
Submitted: 23/02/2024
Revised: 13/06/2024
Accepted: 04/08/2024
Online first: 05/08/2024

Abstract

Road safety is one of the critical government transportation concerns, especially on the toll roads. With the increasing number of toll roads as part of infrastructure planning, road traffic accidents are significantly escalating. Developing a system that predicts accidents on toll roads will benefit to reduce the harm that is caused by traffic accidents. This study will propose a method for analysing toll road accidents in Indonesia using historical toll road accident data as a dataset to become a pattern to examine the frequency of accidents. This dataset consists of various parameters from three main factors that cause accidents: human, environmental, and road infrastructure factors. Machine learning technique will be mainly used to determine the most influencing factors by employing classifiers such as Logistic Regression (LR), Decision Tree (DT), Gaussian Naïve Bayes (GNB), and K-Nearest Neighbors (KNN) can construct the prediction model. Fourteen subfactors from the data were used to predict the future fatalities caused by accidents, which allowed the system to forecast the accident fatality. The results show accuracy performance on the test set with LR, DT, KNN, and GNB models, 85.3%, 79.4%, 87.1%, and 77.1%, respectively. The KNN Classifier model has the most minor error value of 0.6 compared to the other models. The study's findings will help analyse the causal factors involved in toll road accidents and could be utilised by road authorities to employ risk control options to mitigate the ramifications.

Keywords: Road safety; Logistic regression; Decision tree; Gaussian naive bayes; K-nearest neighbors

1. Introduction

Indonesia is a rapidly developing country that focuses on infrastructure development. Improving the infrastructure is of prime importance to improve the country's economic condition since it makes logistics faster. The development of road infrastructure is proliferating in Indonesia. The Indonesian government sets a target of 300% growth in the quantity and length of toll roads by 2030 compared to 2020. Unfortunately, the increasing number of highways has resulted in numerous road traffic accidents (RTA) [1]. The Indonesian Highway Authority reports the number of high-profile national RTA in 2022 (25138 cases). The RTA data shows that 86% of crashes are caused by driver-related factors or human error. A type of fatal accident caused by human error is rear-end collision accidents caused by speeding on the highway, representing 40% of traffic accidents [2].

The Cikopo-Palimanan (Cipali) highway is one of the busiest in Indonesia, with a length of 116 kilometres. The fatality rate on toll roads has a ratio of 0.30 per km. This figure is higher than that of regional and national roads, which have a fatality rate of 0.15 per km and 0.22 per km, respectively. From 2019 to 2021, there have been 1,000 traffic accidents on the Cipali Toll Road section, with 223 fatalities. Of these, 862 or 86.1 percent of accidents were caused by human factors, such as lack of anticipation, drowsiness, and driving over the speed limit. While the causes of traffic accidents caused by tire bursts were 127 cases, wheel disorders were as many as seven cases or a total of 13.6 percent [3]. With the relatively high number of fatalities on RTA on Cipali Highway, there is a demand for a highly advanced technique using advanced mathematics and computational analysis. With the high-level potential data, machine learning with various classifier techniques can be utilised to predict accident fatality. Numerous authors have utilised the Machine Learning (ML) classification technique to analyse data-driven RTA and will be explained in Section 2.

According to the explanation of several studies, accident analysis using the ML classifier technique for highways has a higher chance of predicting accident fatality. Therefore, this paper aims to formulate a fatality prediction of accident that occurs in Cipali Toll Road by using Machine Learning classifiers such as Logistics Regression, Decision Tree Classifiers, Gaussian Naïve-Bayes, and K-Neighbors Classifiers. Furthermore, this proposed ML prediction method is expected to serve as a warning for road users, road managers, and regulators to ensure safety in driving with the ultimate goal of achieving zero accidents.

2. Related Works

General studies related to toll road accidents have been conducted by several researchers about the period of accident from midnight to early morning, around 12:00 am to 05:59 am [4]–[7] affected by weather conditions [5], [7]–[9], drowsiness [10], fatigue [11], alcohol consumption, road darkness [12], intersection area [13]–[15], driver operating error [16], [17], vehicle error [18], road conditions [19], and minimum visibility [20]. There is a high correlation between time factor and number of fatalities [21]. Their research focuses on the correlation between many factors that impact fatality. Generally, they used statistical methods to analyse and present the results. Machine learning has been introduced to solve the problems [9], [10].

Machine learning classifiers have been utilised to evaluate traffic crashes in Sri Lanka by using Random Forest (RF), Decision Tree (DT), (XGB), and K-Nearest Neighbors (KNN) and compared with Logistic Regression (LR). Five factors have been studied from 279 data: road condition, location, weather, and lighting effect. The results show that the ML models proved more accurate than the LR Model [22]. Another study analyses and predicts accident severity in Bangladesh using DT, KNN, Naïve Bayes, and AdaBoost. The results conclude that AdaBoost had the best performance [23]. Similar studies have been conducted, and the results indicate that each ML classifier has advantages towards others depending on the dataset [24]–[27]. In order to sharpen and compare the ML classifier technique, the ML classifier was evaluated technique by using five evaluation metrics [28], [29]: accuracy, Root Mean Square Error (RMSE), precision, recall, and receiver operating characteristic curves.

Previous research has utilised the DT Classifier for accident analysis in highway roads [30]–[32]. Decision Tree was also used earlier for analysing road accidents [33], [34]. While [30] has used Gaussian Naïve Bayes for highway analysis. The K-Neighbors Classifier is also popular in analysing accidents in the highway sector [35], [36]. The study shows that the Neighbor method has the best results.

Some of the study results above are from studies conducted on highways in various countries such as the US, India, Bangladesh, and Sri Lanka. Meanwhile, some studies using ML for accident analysis in Indonesia include using ML to analyse the accident's severity. Three methods were used [37]: Random Forest, Gradient Boosting Machines, and Bagging Regression Tree. The results show that road-related features are most important in predicting the number of fatal accidents. Two studies related to ML regarding road accidents in Indonesia were conducted by [38] using multinomial logistic regression for categorising injury levels for pedestrians, and the heterogeneity of traffic in speeds and volumes was adopted in the study on fatality rates and accident rates [37] on inter-urban roads. Regarding Indonesian highways, the Zero Inflated Negative Binomial method has been utilised for designing highways according to the standard, and the data from non-highways have been used to comply with the standard [39].

3. Methodology

The general process follows the **Figure 1**. The dataset consists of historical data on toll road accidents, and the data class is represented by the fatality status (experiment dataset number 1) and fatality and major injury status (experiment dataset number 2). The attributes influence the decision of whether a fatality occurred or not. Based on the data condition, the data for each attribute was collected to get the suitable parameter for input in the learning process.

Data pre-processing is crucial before the dataset is divided into training and test sets. The dataset goes through a data pre-processing stage to produce a high-quality training set, minimising the model's error. A total of 1645 raw data records on Cipali toll road accidents (2018-2023) with 19 attributes were collected from the PT Astra Toll Cipali Indonesia information system following research ethics. Similarly, in applying supervised machine learning algorithms, the training set will significantly impact the model's performance after undergoing data pre-processing, which typically consists of stages like data cleaning, normalisation, transformation, feature extraction and selection.

Figure 1 shows that the pre-processing begins with data cleaning, integration, transformation, and defining input and output attributes. After getting I/O data, the learning process is executed through Python's Scikit-Learn, and the last step is to show the result by presenting errors using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

3.1. Pre-Processing Data

Data cleaning is sorting the raw data, much of the missing data has been removed. Duplication has been cleared, and some noise and inconsistency in the data have been cleaned. The process continued with the integration and transformation process of the data. A few random sample records which can represent the diversity of the raw data can be seen in **Table 1** and list of atributes before feature selection can be seen in **Table 2**.

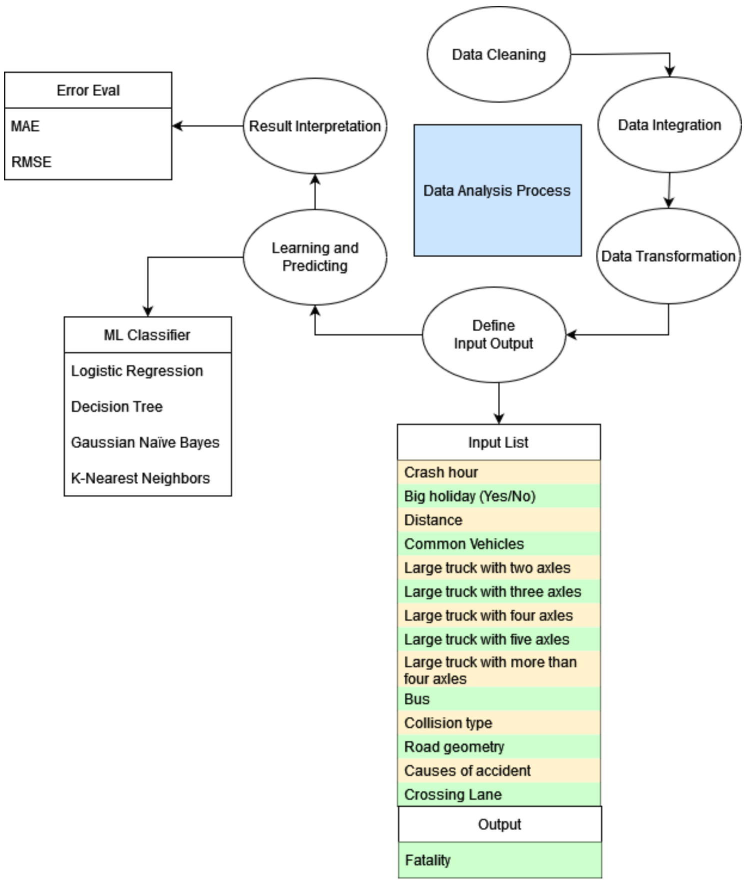

Figure 1. Diagram of research methodology

Table 1. Random sample records of the raw data

No.	Crash Hour	Crash Day	Date	Km	Lane A/B	CV	LT2A	LT3A	LT4A	LT5A	Bus	Collision Type	Weather	Road Geometry	Min-I	Ma-I	F	Causes of Accident	CL
1	0:40	Thursday	Jan 4, 2018	138.8	A	1						Single vehicle	Light Rain	Straight	2	1	1	Driver's Drowsiness and collided with the Public Street Lighting	No
2	4:10	Thursday	Jan 4, 2018	93.9	B	1						Multiple vehicle	Sunny	Straight	3	1	-	Driver's Drowsiness	No
3	6:51	Thursday	Jan 4, 2018	116.2	A			1			1	Multiple vehicle	Sunny	Straight	2	-	-	Driver's Drowsiness	No
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
517	0:45	Monday	Jun 16, 2019	150.9	B	3	1					Multiple vehicle	Sunny	Straight	21	11	12	Unexpectedly, a passenger sitting in the back attacked the bus driver caused the bus to swerve and change lanes, colliding with a minibus behind.	Yes
…	…	…	……	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
900	2:58	Monday	Aug 10, 2020	184.4	B	1	1					Multiple vehicle	Sunny	Curved	3	13	8	Driver's Drowsiness	Yes
…	…	…	……	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1645	13:45	Saturday	Dec 31, 2022	154	A	1						Single vehicle	Light Rain	Straight	1	-	-	The rear right tire burst	No

**Note:* CV: common vehicles; LT2A: large truck with two axles; LT3A: large truck with three axles; LT4A: large truck with four axles; LT5A: large truck with five axles; Min-I: minor injury; Ma-I: major injury; F: fatality; CL: crossing lane.*

As a detail, the source data of separate dates (day, month, and year) in the information system are integrated to create new attributes or predictor variables that are more relevant to the fatality decision. These new attributes include crash day, the potential for accidents to occur is from Friday to Saturday, big holiday, season, and trimesters. The accident day data is integrated with data from the Indonesian Meteorology, Climatology, and Geophysics Agency (BMKG) regarding the estimated wet and dry seasons according to the zoning of the Cipali toll road area.

A big holiday is a period during which there is a mass exodus and widespread holidays to celebrate Eid al-Fitr, Christmas, and New Year. Indonesia is a unique country with the status of

Table 2. List of atributes before feature selection

No.	Attributes Name	N Instance	Instance	Data type transformation (Categorical)
1.	Crash hour	4	12:00 - 05:59 am; 06:00 - 11:59 am; 12:00 - 05:59 pm; 06:00 - 11:59 pm.	Ratio to nominal.
2.	Crash day	3	Weekday; End of weekday; Weekend.	String (date) to nominal.
3.	Friday – Sunday	2	No; Yes.	String (date) to nominal.
4.	Big holiday	2	No; Yes.	String (date) to nominal.
5.	Season	2	Wet season; Dry season.	String (date) to nominal.
6.	Trimesters	3	1st trimesters; 2nd trimesters; 3rd trimesters.	String (date) to nominal.
7.	Distance	6	72 < x ≤ 110.35 / West to East; 110.35 < x ≤ 174.05 / West to East; 174.05 < x ≤ 188.85 / West to East; 72 < x ≤ 110.35 / East to West; 110.35 < x ≤ 174.05 / East to West; 174.05 < x ≤ 188.85 / East to West.	Interval (km) to nominal.
8.	Lane	2	West to East; East to West.	Nominal to nominal.
9.	Common vehicle	2	No; Yes.	Interval to nominal.
10.	Large truck with two axles	2	No; Yes.	Interval to nominal.
11.	Large truck with three axles	2	No; Yes.	Interval to nominal.
12.	Large truck with four axles	2	No; Yes.	Interval to nominal.
13.	Large truck with five axles	2	No; Yes.	Interval to nominal.
14.	Large truck with more than four axles	2	No; Yes.	Interval to nominal.
15.	Bus	2	No; Yes.	Interval to nominal.
16.	Collision type	7	Fixed object collision; Run off road with or without collision; Collision with pedestrian or animal; Angle or side collision; Rear-end collision; Head-on collision; Chain reaction accidents.	String (report description) to nominal.
17.	Rollover	2	No; Yes.	String (report description) to nominal.
18.	Weather	5	Sunny; Light rain; Rain; Heavy rain; Overcast.	Nominal to nominal.
19.	Road geometry	2	Straight; Curved.	Nominal to nominal.
20.	Causes of accident	4	Vehicle fault; Humans: lack of anticipation; Humans: reckless driving; Humans: drowsiness.	String (report description) to nominal.
21.	Crossing lane	2	No; Yes.	Nominal to nominal.

having the largest Muslim population in the world. Approximately 87-90% of Indonesia's total population is Muslim. Unlike Christmas and New Year events, Eid al-Fitr follows the Hijri calendar or has a dynamic schedule in the Gregorian calendar. Almost everyone undertakes long and often monotonous journeys to their hometowns or temporarily migrates from urban to rural areas, using toll facilities. This aspect deserves attention when hypothesising the likelihood of fatalities. These differences have the potential to result in varying interpretations of accident data occurring during big holidays in predominantly Muslim countries compared to research findings in predominantly non-Muslim countries. Infrequently exposed attributes are integrated to maximise the impact of predictor variables.

Table 2 shows the attributes related to the involvement of large trucks with four axles, five axles, and more (except buses) in toll road accidents are combined into a single attribute named large trucks with more than three axles. Descriptions of accident reports indicating vehicle rollovers are integrated into an attribute called rollover. The categorical data classification method was applied as a limitation of the research. This method transformed all records in the newly created attributes and existing numerical (interval or ratio) attributes into categorical (nominal) forms. For example, records in the "crash hour" attribute, initially on an interval scale, were converted into categories: 12:00 - 05:59 am, 06:00 - 11:59 am, 12:00 - 05:59 pm, 06:00 - 11:59 pm (4 instances). Records in the "crash day" attribute were compactly converted into three categories, weekday, end of weekday, and weekend, to assess predictions of fatalities and significant injuries based on the day of the accident.

The last step is feature selection with the chi-square approach. In this process, attributes or input variables that function as independent variables must satisfy the hypothesis of association with the output variable or the attribute representing the decision of fatality and significant injury (dependent variable) to strengthen the predictive model.

3.2. *Learning and Predicting*

As mentioned, process learning and predicting go through a machine learning process using a famous classifier, Python's Scikit-Learn. Supervised learning is employed in this study, where the algorithms to be compared include Logistic Regression, Decision Tree Classifier, Gaussian Naïve Bayes, and K Nearest Neighbors Classifiers.

3.2.1. *Logistic Regression*

Logistic Regression (LR) is supervised machine learning in generalised linear model algorithms. It is a classification algorithm widely used for building predictive models that utilise probabilities and can be seen as a linear regression model with an associated cost function called the sigmoid or logistic function. This function maps predicted class values to the probability values between 0 and 1. The equation logistic regression follows:

$$g\big(E(y)\big) = \alpha + \beta x1 + \gamma x2 \tag{1}$$

where $g(E)$ is the link function, $E(y)$ is the expectation of the predicted variable, and $\alpha + \beta x1 + \beta x2$ are the predictors.

3.2.2. *Decision Tree*

The decision tree builds classification or regression models as a tree structure. It breaks down a dataset into smaller subsets with an increase in the depth of the tree. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). The leaf node (e.g., Play) represents a classification or decision. The root node is the topmost decision node in a tree corresponding to the best predictor. Decision trees can handle both categorical and numerical data. In this research, the decision tree follows the following steps;

a. Select the top from the fourteen attributes of the Cipali Toll Dataset as the root node.
b. Each iteration of the algorithm iterates through the very unused attribute of the set attribute and calculates the **Entropy (H)** and **Information gain (IG)** of this attribute.
c. Then, select the attribute that has the smallest entropy or most significant information gain.
d. The selected attribute splits the set attribute to produce a subset of the data.
e. The algorithm continues to recur on each subset, considering only attributes never selected before.

The entropy *E(H)* measures the randomness of the information defined by Equation 2.

$$E(H) = \sum_{i=1} -P_i log_2 P_i \tag{2}$$

H represents the current state of the input attributes, P_i is the probability of the selected following attribute for any event of state *H*. The information gain is computed as Equation 3.

$$Entropy\ (B) = \sum_{j=1}^{K} entropy\ (j, after) \tag{3}$$

B is the dataset before splitting, *K* is the number of subsets generated, and (*j*, after) is the *j*-th subset after splitting.

3.2.3. Gaussian Naïve Bayes

Naïve Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes' classifier is a fast, accurate, and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. The naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, whether a loan applicant is desirable depends on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, which is why it is considered naïve, as shown in Equation 4. This assumption is called class conditional independence.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{4}$$

Where: *P(h)*: the probability of hypothesis h being actual (regardless of the data). This is known as the prior probability of h; *P(D)*: the probability of the data (regardless of the hypothesis). This is known as the prior probability; *P(h|D)*: the probability of hypothesis *h* given the data *D*. This is known as posterior probability; *P(D|h)*: the probability of data *d* given that the hypothesis *h* was true. This is known as posterior probability.

3.2.4. K-Nearest Neighbors

K-nearest neighbors (KNN) Classifiers are supervised machine learning algorithms that can be used for regression and classification tasks. A supervised machine learning algorithm depends on labelled input data, which the algorithm learns and uses its learned knowledge to produce accurate outputs when unlabeled data is inputted. The use of KNN is to make predictions on the test data set based on the training data's characteristics (labelled data). The method used to make these predictions is by calculating the distance between the test data and training data, assuming that similar characteristics or attributes of the data points exist within proximity. It allows us to identify and assign the new data category while considering its characteristics based on learned data points from the training data. The KNN algorithm will learn these characteristics of the new data point, and based on its proximity to other data points, it will be categorised.

4. Results and Discussion

4.1. Results of the Study

According to **Table 3**, attributes that impact fatality related to the accidents in toll road from experiment are crash hour, crash day, Friday-Sunday (end of weekday to weekend), significant holiday, season, trimesters in a year, distance, lane, exist of typical vehicle, presence of large truck with two axles, three axles, and more than three axles, presence of bus, collision type, possibility of rollover, weather, road geometry, causes of accident, and possibility of crossing lane.

Based on the results of the Chi-Square test, presented in **Table 3**, there are 13 attributes in data set 1 that are associated with fatality status, which is indicated by p-value <0.05, including crash hours, big holiday, distance, standard vehicles, large truck with two axles, three axles and more than three axles, bus, collision type, road geometry, causes of accident and crossing lane. These attributes will be used as input of data set 1 for prediction modelling with random sampling, test-training splitting method. Meanwhile, the attributes not associated with the fatality status are crash day, Friday–Sunday, season, trimesters, lane, rollover, and weather, which are discarded. 13 attributes are not similar in data set 2, which are associated with the fatality and significant injury status, which is indicated by p-value < 0.05, including crash hours, crash day, Friday-Sunday, big holiday, common vehicles, large trucks with two axles, three axles and more than three axles, bus, collision type, road geometry, causes of accidents and crossing lane. These attributes will be used as input data set 2 for prediction modeling with random sampling, test-training splitting method. Meanwhile, the attributes not associated with fatality are season, trimesters, distance, lane, rollover, and weather.

A total of 1530 sample accident records containing 14 selected attributes from dataset 1 and dataset 2 have undergone effective data pre-processing as shown in **Table 4** and **Table 5**. These records were filtered to create a new dataset for applying the random sampling method to divide them into training and test sets. Among the four supervised machine learning algorithms, KNN Classifier and LR exhibited the lowest MAE values for test set 1 (Fatality Yes/No), measuring 16.1 and 16.5, respectively. In model evaluation, a smaller number of MAE is considered better.

MAE calculates the average of the absolute differences between predicted and actual values. A more petite MAE indicates that the model's predictions are closer to the actual values, demonstrating better predictive performance.

Conversely, for test set 2, all MAE values are considerably higher when compared to applying algorithms to test set 1 as shown in **Figure 2**. This implies that using class labels with statuses of fatality and significant injury may not be the most suitable choice due to their poor performance, as reflected by the relatively high MAE values. Subsequently, cross-validation was employed to enhance the model's performance for dataset 1, and all four models were evaluated using more considerable parameters, including Accuracy, RMSE, Precision, and Recall.

Figure 3 shows the accuracy performance on the test set with LR, DT Classifier, KNN Classifier, and Gaussian Naive Bayes models being 85.3%, 79.4%, 87.1%, and 77.1%, respectively. The KNN Classifier model has the smallest RMSE value (0.6) compared to the other models as depicted in **Figure 3**. LR, DT Classifier, and Gaussian Naive Bayes have higher RMSE values of 0.62, 0.67, and 0.69, respectively, in line with the accuracy performance, suggesting that the use of Gaussian Naive Bayes tends to be less accurate and potentially overfitting in predicting the target status of whether fatal accidents occur on toll roads compared to DT Classifier and LR. The KNN Classifier demonstrates the perfect Precision performance, 2.5 times higher than the logistic regression model in the second highest position with a 0.4 or 40% value. Unlike accuracy and RSME

Table 3. Feature selection - attributes relationship with data classes of two data set

Attributes	**Chi-Square Test Result**	
	Asymptotic Significance (2-sided)	
	Fatality (Yes / No)	**Fatality** and **Major Injury** (Yes/No)
Crash hour	**0.01**	***
Crash day	0.56	**0.04**
Friday–Sunday	0.06	**0.02**
Big holiday	**0.03**	**0.04**
Season	0.39	0.07
Trimesters	0.31	0.19
Distance	**0.03**	0.16
Lane	0.77	0.67
Common vehicles	***	***
Large truck with two axles	***	***
Large truck with three axles	***	***
Large truck with four axles	**0.01**	0.07
Large truck with five axles	**	**
Large truck with more than four axles	***	**
Bus	**	**0.02**
Collision type	***	***
Rollover	0.33	0.19
Weather	0.48	0.38
Road geometry	***	**0.02**
Causes of accident	***	**
Crossing lane	***	**0.02**

*Note: significance values are presented as ** = p <.01, *** = p <.001.*

Table 4. Data Set 1

No. data set	No. Raw data	Crash hour	Big holiday	Distance	CV	LT2A	LT3A	LT4A	LT5A	LTM4A	Bus	Collision type	Road geometry	Causes of accident	Crossing Lane	Fatality
1	1	12:00 - 05:59 am	Yes	110.35 < x ≤ 174.05 / West to East	Yes	No	No	No	No	No	No	Fixed object collision	Straight	Humans: drowsiness	No	Yes
2	2	12:00 - 05:59 am	Yes	72 < x ≤ 110.35 / East to West	Yes	No	No	No	No	No	No	Rear-end collision	Straight	Humans: drowsiness	No	No
3	3	06:00 - 11:59 am	Yes	110.35 < x ≤ 174.05 / West to East	No	No	Yes	No	No	No	Yes	Angle or side collision	Straight	Humans: drowsiness	No	No
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
460	517	12:00 - 05:59 am	No	110.35 < x ≤ 174.05 / East to West	Yes	Yes	No	No	No	No	No	Chain reaction accidents	Straight	Humans: lack of anticipation	Yes	Yes
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
836	900	12:00 - 05:59 am	No	174.05 < x ≤ 188.85 / East to West	Yes	Yes	No	No	No	No	No	Head-on collision	Curved	Humans: drowsiness	Yes	Yes
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1530	1645	12:00 - 05:59 pm	Yes	110.35 < x ≤ 174.05 / West to East	Yes	No	No	No	No	No	No	Fixed object collision	Straight	Vehicle fault	No	No

**Note: CV: common vehicles; LT2A: large truck with two axles; LT3A: large truck with three axles; LT4A: large truck with four axles; LT5A: large truck with five axles; LTM4A: large truck with more than four axles.*

Table 5. Data Set 2

No. data set	No. Raw data	Crash hour	CDWBH	F-S	BH	CV	LT2A	LT3A	LT4A	LT5A	LTM4A	Bus	Collision type	Road geometry	Causes of accident	Crossing Lane	F&MI
1	1	12:00 - 05:59 am	Weekday	No	Yes	Yes	No	No	No	No	No	No	Fixed object collision	Straight	Humans: drowsiness	No	Yes
2	2	12:00 - 05:59 am	Weekday	No	Yes	Yes	No	No	No	No	No	No	Rear-end collision	Straight	Humans: drowsiness	No	Yes
3	3	06:00 - 11:59 am	Weekday	No	Yes	No	No	Yes	No	No	No	Yes	Angle or side collision	Straight	Humans: drowsiness	No	No
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
460	517	12:00 - 05:59 am	Weekday	No	No	Yes	Yes	No	No	No	No	No	Chain reaction accidents	Straight	Humans: lack of anticipation	Yes	Yes
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
836	900	12:00 - 05:59 am	Weekday	No	No	Yes	Yes	No	No	No	No	No	Head-on collision	Curved	Humans: drowsiness	Yes	Yes
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1530	1645	12:00 - 05:59 pm	Weekend	Yes	Yes	Yes	No	No	No	No	No	No	Fixed object collision	Straight	Vehicle fault	No	No

**Note: CDWBH: crash day without big holiday; F-S: riday-Sinday; BH: big holiday; CV: common vehicles; LT2A: large truck with two axles; LT3A: large truck with three axles; LT4A: large truck with four axles; LT5A: large truck with five axles; LTM4A: large truck with more than four axles; F&MI: fatality and major injury.*

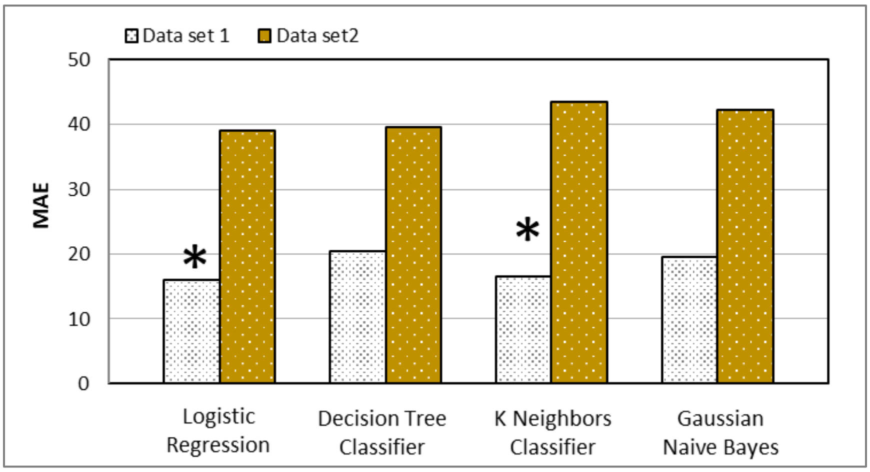

Figure 2. MAE of the model with random sampling

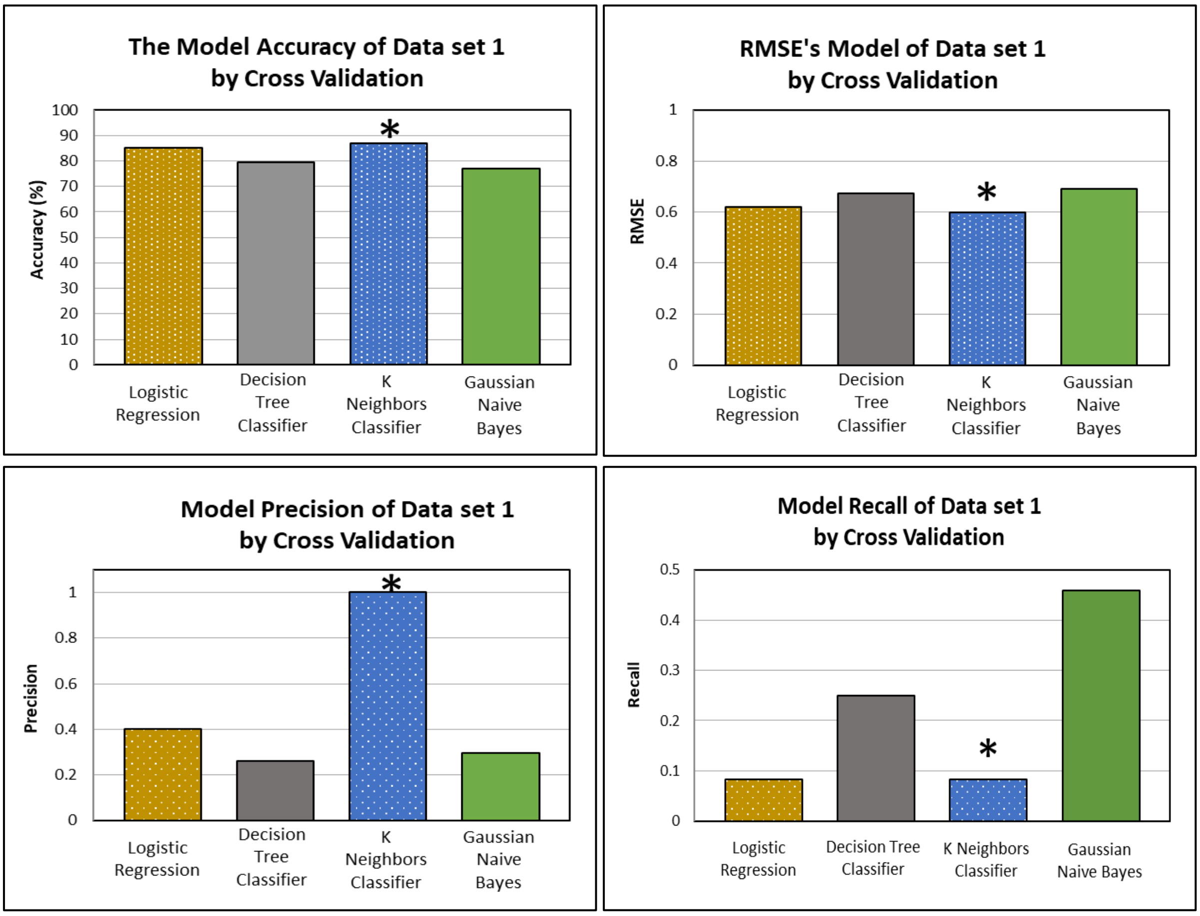

Figure 3. Histogram evaluation of each model of classifier

the DT Classifier tends to be weaker or gives many false optimistic predictions among the truly relevant results compared to Gaussian Naïve Bayes with values of 0.28 and 0.30, respectively. High precision is essential to minimise the number of false predictions of fatalities, although the consideration and recall (sensitivity) to get a complete picture of the classification model's performance cannot be separated. The KNN Classifier and LR provide the best recall performance with the same value, 0.083, while the DT Classifier and Gaussian Naïve Bayes tend to miss many cases of positive fatal occurrences that should have been there with recall values of 0.292 and 0.458, respectively as shown in Figure 3.

Figure 4, which illustrates the contribution of features causing fatalities as found in this research paper, shows that for the category of accident

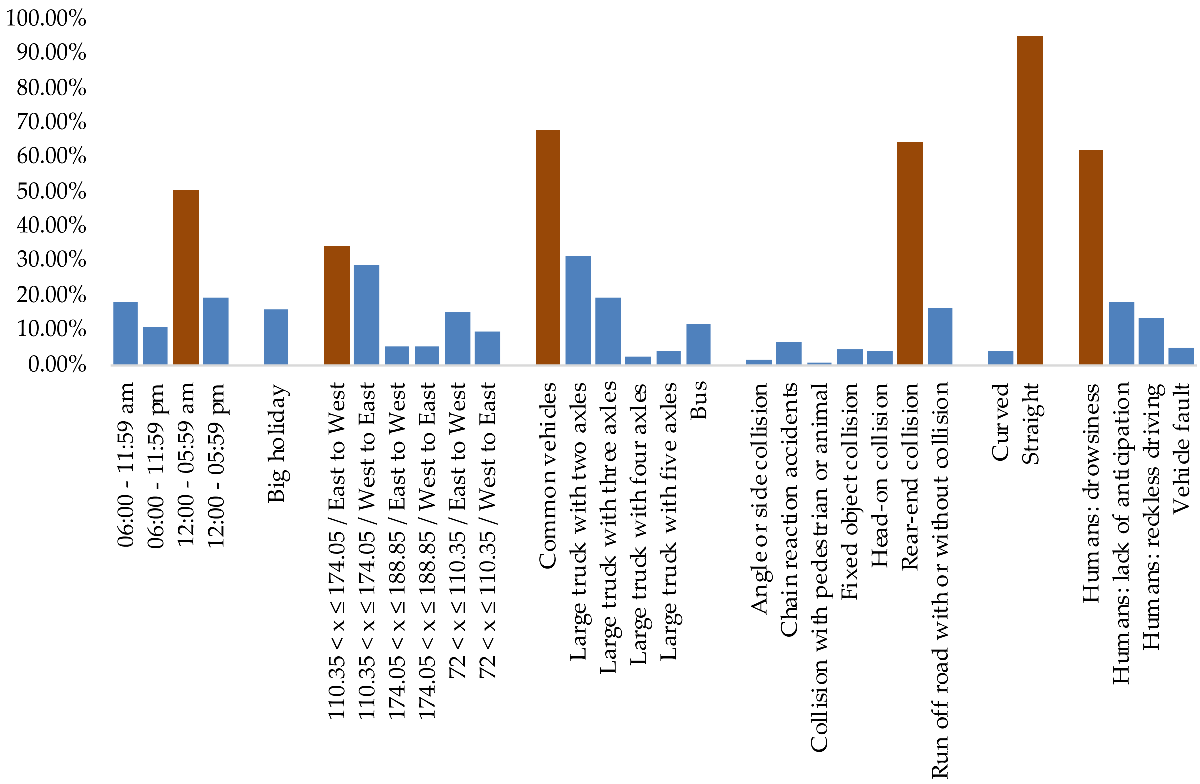

Figure 4. The contribution of features causing fatalities

timing, the highest number of accidents occurs between 12:00 AM and 5:59 AM. The segment with the highest number of accidents is the east to west route between km 110.35 and km 174.05. The most frequent type of accident involves common vehicles, specifically rear-end collisions, occurring on straight roads and caused by human drowsiness.

4.2. *Discussions*

The findings of this research are consistent with the previous study of [4], indicating that the period from midnight to early morning, specifically from 12:00 am to 05:59 am, results in the highest number of accidents leading to fatalities on toll roads compared to other time segments. Specific kilometres of driving on toll roads or certain toll road segments also influence the occurrence of fatalities in toll road accidents in Indonesia. Human factors such as drowsiness, lack of anticipation, and reckless driving significantly contribute to the heightened risk of fatal accidents occurring on toll roads. Among these factors, drowsiness in drivers emerges as an especially perilous element, capable of directly contributing to fatal incidents on toll roads.

From the perspective of collision types, rear-end accidents have the most significant contribution to the risk of fatalities on toll roads, which is similar to [4]. The percentage of accident cases that resulted in fatalities from rear-end collisions in this study reached 65%, which is significantly higher when compared to the factors contributing to run-off road with or without collision, chain reaction accidents, fixed object collision, head-on collision, angle or side collision, and collision with pedestrian or animal. The findings of this study also support previous research by [40] indicating that the interaction of large vehicles during toll road accidents is significantly associated with fatalities. A higher proportion of large vehicles significantly influences the crash rate because a more significant number of large vehicles introduces heterogeneity into the traffic flow, as they typically travel at slower speeds and have reduced manoeuvrability compared to regular small cars. A 1% rise in the proportion of large vehicles leads to a 7.6% increase in the crash incident rate.

The unique outcome of this study identifies that, statistically, factors such as season and weather are not associated with the occurrence of fatalities on toll roads in Indonesia, which differs from previous studies conducted in other Asian countries, including China, Sri Lanka, and Pakistan [4], [22], [40]. Another interesting finding is that the occurrence of rollovers during accidents, with or without collisions, also does not affect the incidence of fatalities on toll roads according to the model prediction. Also, this study finds that the fatality rates in toll roads can be affected on unusual days, such as during big holiday seasons.

The quality of the predictive model built relies heavily on the quality of the dataset and the input factors involved. It was concluded that the diversity of climate, topography, population profiles, and driver characteristics across different countries inevitably impacts varying research conclusions. It is also emphasized that, besides relying on empirical data, it is crucial in research to carry out data pre-processing, including data reduction or feature selection. This process helps quantify the relationships among presented factors affecting the decision of fatalities effectively.

5. Conclusion

Toll road accidents are one of the most prevalent incidents that can cause detrimental effects towards fatality. Consequently, a novel methodology for analysing accidents with enormous data is required. Machine Learning is considered one of the most promising data analysis methods. In this research, four machine learning classifier methods, Logistic Regression, Decision Tree, Gaussian Naïve Bayes, and K-Nearest Neighbors, were utilised to predict the fatality with the input "fatality" and "non-fatality". Four of the Machine Learning Classifiers show excellent results for analysing and predicting toll datasets to aid in preventing road accident fatalities, with accuracy between 60-90%, error MSE 10-20%, and RSME 0.6-0.8. Data training and testing have been manually selected using cross-validation to increase the accuracy of four classifiers. The final result shows that the K-Nearest Neighbors classifier can predict well for this dataset, with an accuracy of around 87.1 % with RMSE 0.60 and a high precision value. The findings of this study will be beneficial for the knowledge of data science, especially for the analysis of causal factors involved in toll road accidents.

Acknowledgements

This research data was supported by PT.Astra Toll Cipali, The authors declare no conflict of interest. The supporters had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish.

Author's Declaration

Authors' contributions and responsibilities

The authors made substantial contributions to the conception and design of the study. The authors took responsibility for data analysis, interpretation and discussion of results. The authors read and approved the final manuscript.

Funding

This research received no external funding.

Availability of data and materials

All data are available from the authors.

Competing interests

The authors declare no competing interest.

Additional information

No additional information from the authors.

References

[1] M. of Transportation, "National Development in the 2020-2024 RPJP," Jakarta, 2019.

[2] BPS-Statistics Indonesia, "Land Transportation Statistics," Jakarta, 2021.

[3] Kompas, "List of accidents on Cipali Toll Road throughout 2023," 2023. .

[4] A. Iqbal, Z. U. Rehman, S. Ali, K. Ullah, and U. Ghani, "Road traffic accident analysis and identification of black spot locations on highway," *Civil Engineering Journal*, vol. 6, no. 12, pp. 2448–2456, 2020, doi: 10.28991/cej-2020-03091629.

[5] H. Hanafi, F. Rusgiyarto, and R. Pratama, "Analisis Tingkat Keselamatan Jalan Tol Berdasarkan Metode Pembobotan Korlantas (Studi Kasus: Jalan Tol Cipularang)," *Jurnal Teknik: Media Pengembangan Ilmu dan Aplikasi Teknik*, vol. 18, no. 2, p. 49, 2020, doi: 10.26874/jt.vol18no2.106.

[6] S. Plainis, I. J. Murray, and I. G. Pallikaris, "Road traffic casualties: understanding the night-time death toll," *Injury prevention*, vol. 12, no. 2, pp. 125–138, 2006, doi: 10.1136/ip.2005.011056.

[7] T. Åkerstedt, G. Kecklund, and L. G. Hörte, "Night driving, season, and the risk of highway accidents," *Sleep*, vol. 24, no. 4, pp. 401–406, 2001, doi: 10.1093/sleep/24.4.401.

[8] K. H. Abdullah, "Road Safety Intervention : Publication Trends and Future Research Directions," *International Journal of Road Safety*, vol. 2, no. 1, pp. 10–18, 2021.

[9] A. J. Ghandour, H. Hammoud, and S. Al-Hajj, "Analyzing factors associated with fatal road crashes: A machine learning approach," *International Journal of Environmental Research and Public Health*, vol. 17, no. 11, 2020, doi: 10.3390/ijerph17114111.

[10] A. Tavakoli Kashani, M. Rakhshani Moghadam, and S. Amirifar, "Factors affecting driver injury severity in fatigue and drowsiness accidents: a data mining framework.," *Journal of injury & violence research*, vol. 14, no. 1, pp. 75–88, 2022, doi: 10.5249/jivr.v14i1.1679.

[11] M. L. S. Zainy, G. B. Pratama, R. R. Kurnianto, and H. Iridiastadi, "Fatigue Among Indonesian Commercial Vehicle Drivers: A Study Examining Changes in Subjective Responses and Ocular Indicators," *International Journal of Technology*, vol. 14, no. 5, pp. 1039–1048, 2023, doi: 10.14716/ijtech.v14i5.4856.

[12] N. Md Yusof *et al.*, "Effect of Road Darkness on Young Driver Behaviour when Approaching Parked or Slow-moving Vehicles in Malaysia," *Automotive Experiences*, vol. 6, no. 2, pp. 216–233, May 2023, doi: 10.31603/ae.8206.

[13] E. Yong *et al.*, "Investigation of the Vehicle Driving Trajectory During Turning at Intersectional Roads Using Deep Learning Model," *Automotive Experiences*, vol. 7, no. 1, pp. 63–76, Apr. 2024, doi: 10.31603/ae.10649.

[14] W. A. Al Bargi, M. M. Rohani, B. D. Daniel, N. A. Khalifaa, M. I. M. Masirin, and J. Kironde, "Estimating of Critical Gaps at Uncontrolled Intersections under Heterogeneous Traffic Conditions," *Automotive Experiences*, vol. 6, no. 2, pp. 429–437, 2023, doi: 10.31603/ae.9406.

[15] A. I. Petrov and A. V Pistsov, "Training and Applying Artificial Neural Networks in Traffic Light Control: Improving the Management and Safety of Road Traffic in Tyumen (Russia)," *Automotive Experiences*, vol. 6, no. 3, pp. 528–550, 2023, doi: 10.31603/ae.10025.

[16] A. Sudiarno, A. M. D. Ma'arij, I. P. Tama, A. Larasati, and D. Hardiningtyas, "Analyzing Cognitive Load Measurements of the Truck Drivers to Determine Transportation Routes and Improve Safety Driving: A Review Study," *Automotive Experiences*, vol. 6, no. 1, pp. 149–161, Apr. 2023, doi: 10.31603/ae.8301.

[17] D. H. Waskito *et al.*, "Analysing the Impact of Human Error on the Severity of Truck Accidents through HFACS and Bayesian Network Models," *Safety*, vol. 10, no. 1, p. 8, 2024, doi: https://doi.org/10.3390/safety10010008.

[18] I. Ansori *et al.*, "Enhancing Brake System Evaluation in Periodic Testing of Goods Transport Vehicles through FTA-FMEA Risk Analysis," *Automotive Experiences*, vol. 6, no. 2, pp. 320–335, Aug. 2023, doi: 10.31603/ae.8394.

[19] D. W. Karmiadji *et al.*, "Theoretical Experiments on Road Profile Data Analysis using Filter Combinations," *Automotive Experiences*, vol. 6, no. 3, pp. 584–598, 2023, doi: 10.31603/ae.9901.

[20] F. Valent, F. Schiava, C. Savonitto, T. Gallo, S. Brusaferro, and F. Barbone, "Risk factors for fatal road traffic accidents in Udine, Italy," *Accident Analysis and Prevention*, vol. 34, no. 1, pp. 71–84, 2002, doi: 10.1016/S0001-4575(00)00104-4.

[21] N. Verzosa and R. Miles, "Severity of road crashes involving pedestrians in Metro Manila, Philippines," *Accident Analysis and Prevention*, vol. 94, pp. 216–226, 2016, doi: 10.1016/j.aap.2016.06.006.

[22] J. P. S. S. Madushani, R. M. K. Sandamal, D. P. P. Meddage, H. R. Pasindu, and P. I. A. Gomes, "Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers," *Transportation Engineering*, vol.

13, no. April, p. 100190, 2023, doi: 10.1016/j.treng.2023.100190.

[23] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das, and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," *2019 7th International Conference on Smart Computing and Communications, ICSCC 2019*, pp. 1–5, 2019, doi: 10.1109/ICSCC.2019.8843640.

[24] G. Mahendra and R. H. Roopashree, "Prediction of Road Accidents in the Different States of India using Machine Learning Algorithms," *2023 IEEE International Conference on Integrated Circuits and Communication Systems, ICICACS 2023*, pp. 1–6, 2023, doi: 10.1109/ICICACS57338.2023.10099519.

[25] A. K. Goel, K. Khan, A. Kushwaha, V. Srivastava, S. Malik, and A. Singh, "A Machine Learning Approach to Analyze Road Accidents," *2022 IEEE International Conference on Blockchain and Distributed Systems Security, ICBDS 2022*, pp. 1–5, 2022, doi: 10.1109/ICBDS53701.2022.9935867.

[26] S. Soleimani, M. Leitner, and J. Codjoe, "Applying machine learning, text mining, and spatial analysis techniques to develop a highway-railroad grade crossing consolidation model," *Accident Analysis and Prevention*, vol. 152, no. January, p. 105985, 2021, doi: 10.1016/j.aap.2021.105985.

[27] P. A. Nandurge and N. V. Dharwadkar, "Analyzing road accident data using machine learning paradigms," *Proceedings of the International Conference on IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2017*, pp. 604–610, 2017, doi: 10.1109/I-SMAC.2017.8058251.

[28] T. Bokaba, W. Doorsamy, and B. S. Paul, "Comparative study of machine learning classifiers for modelling road traffic accidents," *Applied Sciences*, vol. 12, no. 2, p. 828, 2022.

[29] R. E. Al Mamlook, A. Ali, R. A. Hasan, and H. A. Mohamed Kazim, "Machine Learning to Predict the Freeway Traffic Accidents-Based Driving Simulation," *Proceedings of the IEEE National Aerospace Electronics Conference, NAECON*, vol. 2019-July, pp. 630–634, 2019, doi: 10.1109/NAECON46414.2019.9058268.

[30] O. H. Kwon, W. Rhee, and Y. Yoon, "Application of classification algorithms for analysis of road safety risk factor dependencies," *Accident Analysis and Prevention*, vol. 75, pp. 1–15, 2015, doi: 10.1016/j.aap.2014.11.005.

[31] J. De Oña, G. López, R. Mujalli, and F. J. Calvo, "Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks," *Accident Analysis and Prevention*, vol. 51, pp. 1–10, 2013, doi: 10.1016/j.aap.2012.10.016.

[32] S. Y. Sohn and S. H. Lee, "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea," *Safety Science*, vol. 41, no. 1, pp. 1–14, 2003, doi: 10.1016/S0925-7535(01)00032-7.

[33] D. D. Clarke, R. Forsyth, and R. Wright, "Machine learning in road accident research: Decision trees describing road accidents during cross-flow turns," *Ergonomics*, vol. 41, no. 7, pp. 1060–1079, 1998, doi: 10.1080/001401398186603.

[34] O. Nedjmedine and M. Tahar, "Analysis of road accident factors using Decision Tree Algorithm: a case of study Algeria," *ISIA 2022 - International Symposium on Informatics and its Applications, Proceedings*, pp. 1–6, 2022, doi: 10.1109/ISIA55826.2022.9993530.

[35] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis and Prevention*, vol. 108, no. August, pp. 27–36, 2017, doi: 10.1016/j.aap.2017.08.008.

[36] W. Lu, J. Liu, X. Fu, J. Yang, and S. Jones, "Integrating machine learning into path analysis for quantifying behavioral pathways in bicycle-motor vehicle crashes," *Accident Analysis and Prevention*, vol. 168, no. February, p. 106622, 2022, doi: 10.1016/j.aap.2022.106622.

[37] M. L. Siregar, T. Tjahjono, and N. Yusuf, "Predicting the Segment-Based Effects of Heterogeneous Traffic and Road Geometric Features on Fatal Accidents," *International Journal of Technology*, vol. 13, no. 1, pp. 92–102, 2022, doi: 10.14716/ijtech.v13i1.4450.

[38] M. L. Siregar, R. Jachrizal Sumabrata, A. Kusuma, O. B. Samosir, and S. N. Rudrokasworo, "Analyzing driving environment factors in pedestrian crashes injury levels in Jakarta and the surrounding cities," *Journal of Applied Engineering Science*, vol. 17, no. 4, pp. 482–489, 2019, doi: 10.5937/jaes17-22121.

[39] A. Rizaldi, V. Dixit, A. Pande, and R. A. Junirman, "Predicting casualty-accident count by highway design standards compliance," *International Journal of Transportation Science and Technology*, vol. 6, no. 3, pp. 174–183, 2017, doi: 10.1016/j.ijtst.2017.07.005.

[40] J. Zhang, X. Chen, and Y. Tu, "Environmental and Traffic Effects on Incident Frequency Occurred on Urban Expressways," *Procedia - Social and Behavioral Sciences*, vol. 96, no. Cictp, pp. 1366–1377, 2013, doi: 10.1016/j.sbspro.2013.08.155.