CHAPTER 4
COMPARISON OF MACHINE LEARNING PREDICTABILITY PERFORMANCE: THE CASE OF MOTORCYCLE ACCIDENT IN THAILAND

4.1 Abstract

Every year in Thailand and other place around the globe, traffic accidents kill, injure, and kill, causing millions of deaths, injuries, and fatalities as well as billions of dollars in economic damages. Accurate models for predicting the severity of traffic accidents are essential for transportation systems. This inquiry attempt provides methods for selecting a collection of influential criteria and developing a model for categorizing the severity of injuries. Various supervised machine learning methods approaches are used such as, Decision Tree (DT), Support Vector Machine (SVM), Random Forests (RF), K-Nearest Neighbors (kNN), Neural Network (NN), Naive Bayes (NB) Logistic Regression (LR), and Gradient Boosting (GB).

The researcher included information about the incidence of motorcycle accidents in Thailand for 5 years; a total of 112,837 events. The data includes several factors causing accidents and focuses solely on the motorcycle rider who were the cause of accidents excluding the passengers.

Random Forest (RF) outperformed the other seven ML algorithms in predicting road accidents in 2-Class classification giving an average over class AUC of 0.768, CA of 0.777, Precision of 0.752, and recall of 0.777. Which top 5 features giving highest info gain from RF consist of Highway, Riding over speed limit, Day time, Night w/o light and gender. The ML is an effective tool for predicting accidents; the model performs well in terms of non-fatality prediction, but there is still room for improvement in fatality prediction, which can be interpreted to mean that the factors that cause fatal accidents may cause an accident but do not always result in a fatality.

4.1.1 Highlights

1) Random Forest (RF) outperformed the other seven ML algorithms in predicting motorcycle accidents

2) All ML models for this study found significant misclassifications of fatalities but a much better ability to predict nonfatalities

3) Most accidents occur on highway, Riding over speed limit, Day time, Night w/o light and gender

4.2 Introduction

Thailand has the highest number of road fatalities, ranking in the top ten worldwide. Thais die in traffic crashes at a rate of 32.7 for each 100,000 people. (WHO, 2018). Motorcycles comprise half of all 41 million registered vehicles. (DLT, 2021), potentially contributing to the highest number of fatalities from major accidents. As Jomnonkwao et al. (2020) observed, motorcyclists are responsible for most road fatalities., while prior studies showed different types of cars in difference country, such as rollover SUV/vans (Jafari Anarkooli et al., 2017), large truck (Huo et al., 2020; Jafari Anarkooli et al., 2017; Li et al., 2018), and pick-ups (Li et al., 2018). Despite government efforts to increase law enforcement for drink driving, speed control, engineering solutions for road safety, and so on, the number of road fatalities has remained consistent at 32-35 percent from 2015 to 2020 (PDPM, 2020). The current policy for reducing road fatalities appears to be ineffective. Machine learning is one of the new solutions for predicting and minimizing accidents that should be implemented.

Machine learning (ML) is the learning of computer algorithms that change automatically through education. It is pictured as the subset of artificial power which algorithms develop a science framework using sample information, called ‘preparation information’, in order to make prediction or conclusion without being explicitly programmed to do so. Why is ML comparison required for road accident prediction study? Because the ML model itself has different advantages and disadvantages, as shown in table 4.1 (Géron, 2019; Guido, 2017; Sarker, 2021; Wu et al., 2007). Understanding and analyzing the model is essential before deciding on a machine

learning algorithm. Before settling on a single machine learning algorithm, compare their accuracies on training and test sets.

Table 4.1 Comparison of the advantages and disadvantages of ML models

Model	PROs	CONs
SVM	- Performs admirably in higher dimensions - Outliers have less influence	- Does not perform well when classes overlap - Selecting appropriate hyperparameters is important
GB	- Running Quickly & interpret - High dimension of data are well handled	- Tuning parameters are difficult to find - If the parameter is not well tuned, it's easy to overfit
LR	- Simple and effective when the dataset can be divided linearly - Feature scaling is not required	- Inadequate performance on non-linear data - More commonly used for classification not regression - When used with high-dimensional datasets, it has the potential to overfit.
NN	- Can build extremely complex model for large datasets - Excellent for large data set w/ high dimension data - High prediction accuracy	- Scaling data and parameter selection are critical considerations - Time required for training

Table 4.1 Comparison of the advantages and disadvantages of ML models (Continued)

Model	PROs	CONs
Naïve Bayes	- Good performance with high-dimensional data - It only requires a small amount of training data to quickly estimate the required parameters.	- Training data should accurately represent the population
kNN	- It can be used for both classification and regression. - Can deal with multi-class problems - Very resistant to noisy training data	- Sensitive to outlier - Not handle well with missing data - The accuracy is determined by the quality of the data
Tree	- Missing value handling - Simple to explain and visualize - Can used for both the classification and regression tasks	- Prone to overfitting - Sensitive to data
RF	- Fits both categorical and continuous values well - Less overfitting - Useful to extract feature importance	- Not suitable for extremely high-dimensional sparse data

The primary objective of this research is to evaluate the performance of machine learning methods in classifying the severity of road accidents. (Fatality and Non-Fatality) and attempting to identify major factors in forecasting the severity of motorcycle accidents. The proposed study is unique in that it compared eight machine learning models in the same data set that only included motorcycle accidents involving the rider alone (no passenger or victim involved) in attempt to choose the model that predicts future accidents the most accurately.

4.3 Literature Review

According to international research, earlier studies on traffic accidents have been divided by group of factors that perceived to be related in traffic crashes.

4.3.1 Traffic Study Using Machine Learning

According to Table 4.2 shows a previous study that used several machine learning models to investigate traffic accidents; The studies compared prediction accuracy of ML models and discovered feature importance as it relates to road accidents.

Table 4.2 Machine learning models in traffic accident study

Author	Methodology											
	Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Gradient Boosting	K-means	K-Nearest Neighbor	MultinomialLogistic	Neural Network	Naïve Bayes	Random Forest	Support Vector Machine
Ospina-Mateus et al. (2021)	-	-	-	✓	-	-	✓	-	✓	✓	✓	✓
Kumar and Toshniwal (2016)	✓	-	-	-	-	✓	-	-	-	-	-	-
Harb et al. (2009)	-	-	-	✓	-	-	-	-	-	-	✓	-
Mafi et al. (2018)	-	-	-	-	-	-	-	-	-	-	✓	-

Table 4.2 Machine learning models in traffic accident study (Continued)

Author	Methodology											
	Associated Rule	Bayesian Logistic	Cluster Analysis	Decision Tree	Gradient Boosting	K-means	K-Nearest Neighbor	MultinomialLogistic	Neural Network	Naïve Bayes	Random Forest	Support Vector Machine
Kuşkapan et al. (2021)	-	-	-	-	-	-	✓	-	-	✓	-	✓
Al Mamlook et al. (2019)	-	✓	✓	✓	-	-	✓	-	-	✓	✓	✓
Recal and Demirel (2021)	-	-	-	✓	✓	-	-	✓	✓	-	-	✓
Bahiru et al. (2018)	-	-	-	✓	-	-	-	-	-	✓	-	-
Ospina-Mateus et al. (2021)	-	-	-	✓	-	-	✓	-	✓	✓	✓	✓
Feng et al. (2020)	✓	-	-	-	-	-	-	✓	-	-	-	-
Helen et al. (2019)	✓	-	-	-	-	✓	-	-	-	-	-	-
(Santos et al., 2021)	-	-	-	✓	-	-	-	-	-	✓	✓	-
(Kim et al., 2021)	-	-	-	-	✓	-	-	-	✓	-	✓	-

4.3.2 Driver/Rider information

When the rider was not at wrong, Thailand cases discovered that the rider's age was the most significant element. (Champahom et al., 2019). Riders between the ages of 18 and 24 are lacking the driving experience, such as adjusting the speed to accommodate different roadways. (Bucsuházy et al., 2020). Motorcyclists aged 20-39 are more likely to engage in serious crashes, however when no motorbike or bicycle is engaged, the magnitude is likely to be mild, and men are involved in serious crashes than women. (Ospina-Mateus et al., 2019).

4.3.3 Roadway

Poor road conditions increase the likelihood of an accident, particularly on highways. (Malin et al., 2019). The characteristic of road, brightness, vehicle speed, and road conditions all have an impact on the frequency of accidents. (Feng et al., 2020). Roads that were dark or dim also played important roles in road accidents. (Shweta et al., 2021). Highway intersections have also been recognized as the ones most risky for all types of accidents. (Kumar & Toshniwal, 2016). Highways had an impact on injury severity outcomes in rural motorcycle crashed (Geedipally et al., 2011). Around 92.5% of the crashes on motorcycle occurred on dry roadway surfaces. (Shaheed et al., 2013)

4.3.4 Internal Factor (Driver/Rider Behavior)

Rider characteristics increase the risk of serious and fatal injury in motorcycle accidents (Cunto & Ferreira, 2017). Intoxicated drivers have a higher accident rate than other drivers. (Helen et al., 2019) and the most important factor of an injury is speeding. (Al Mamlook et al., 2019). Most alcohol-related accidents are caused by young people (35 years old) late at night. (John & Shaiba, 2022). Motorcycles are more dangerous in rural areas. Male riders, exceeding the posted speed limit, overtaking, and exhaustion are all contributing factors to serious and fatal injuries. (Mohamad et al., 2022)

4.3.5 External Factor (Environment & Weather Condition)

More than two-thirds of the two-vehicle motorcycle crashes reported occurred in clear weather (because clear weather encourages motorcycle riding), while one-quarter occurred in cloudy or partly cloudy conditions. Approximately 80% of the

crashes occurred during the day, while nearly one-fifth of the crashes occurred at night. These findings are most likely due to the increased exposure of motorcycles in daylight versus at night, as well as the higher associated crash risk. (Shaheed et al., 2013). Weather conditions, such as poor visibility, have a greater impact on traffic accidents than internal factors such as the driver. Sonal and Suman (2018). Driving at night increases the likelihood of a car accident. (Mphela, 2020). Traveling at night increases the likelihood of a car accident. (Mphela, 2020).

4.4 Methodology and Data

The framework of the article is known as data mining and ML techniques. (Figure 1), and it measures the performance of ML model predictability for fatalities and non-fatalities on road accident as following steps.

Initial Dataset – endorsed for identifying and fixing incomplete and imprecisely collected data, in addition to demonstrating data integrity after the data set has been purified.

Verified Dataset - Data partitioning to binary mode and set Fatal/non-fatal as target.

Data Splitting – Separated test and train data set to ratio 75:25.

Model Learning -Allows the machine to learn with 75% of the test dataset and later test with the leftover 25%.

Model Prediction and Measurement – To check the prediction accuracy of each model.

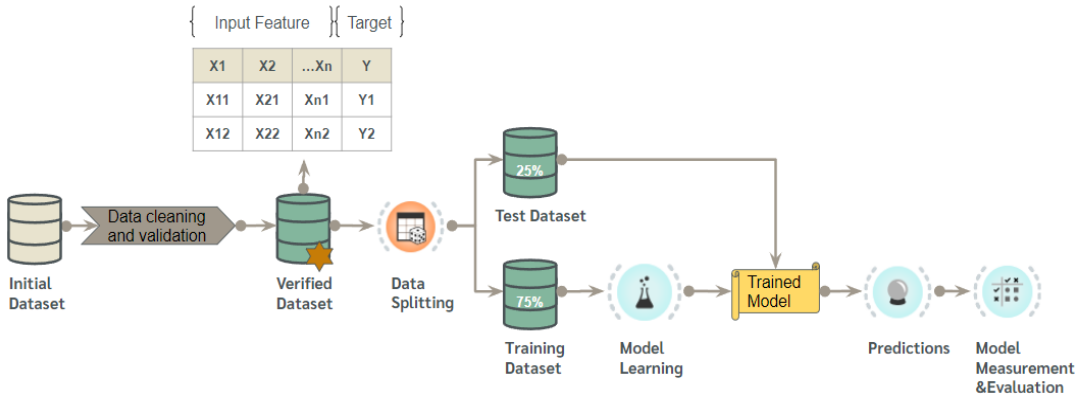

Figure 4.1 Machine learning Process flow

4.4.1 Data Description

The study begins with data on road accidents combined all kind of road, which totaled 112,837 incidents between 2015 and 2020 (PDPM, 2020), with the focus solely on rider who were primary caused in an accident (Table 4.3).

Table 4.3 Categorical Attribute and descriptive statistics.

Accident Event (Attribute)	Fatality			
	Yes		No	
	Count	%	Count	%
Roadway				
Dry Surface Road	26609	23.6%	81459	72.2%
Wet Surface	920	0.8%	3849	3.4%
Straight Way	19460	17.2%	59907	53.1%
Not straight Way (Curve, Slope, Junction, etc.)	8069	7.2%	25401	22.5%
Obstruction	504	0.4%	3080	2.7%
Road condition	374	0.3%	2001	1.8%
Vehicle condition	298	0.3%	2116	1.9%
Highway	19325	17.1%	41403	36.7%
Non-Highway	8204	7.3%	43905	38.9%
External Factor (Environment & Weather Condition)				
Day Time (06.00-18.00)	13439	11.9%	49845	44.2%
Night with Light	7631	6.8%	20824	18.5%
Night without Light	6459	5.7%	14639	13.0%
Low visibility	2427	2.2%	10534	9.3%
Clear Weather	24362	21.6%	73456	65.1%
Not Clear Weather (Rain, fog, etc.)	3167	2.8%	11852	10.5%

Table 4.3 Categorical Attribute and descriptive statistics. (Continued)

Accident Event (Attribute)	Fatality			
	Yes		No	
	Count	%	Count	%
Internal Factor (Driver Behavior)				
Drunk	3670	3.3%	19928	17.7%
Over Speed limit	19121	16.9%	37338	33.1%
Break Through Traffic lights	228	0.2%	456	0.4%
Break Through Traffic Signs	373	0.3%	1299	1.2%
Overtake	641	0.6%	1238	1.1%
Use Mobile Phone	23	0.0%	348	0.3%
Short Cut off	5315	4.7%	19059	16.9%
Drug	6	0.0%	65	0.1%
Drive in opposite direction	432	0.4%	816	0.7%
Doze off	347	0.3%	881	0.8%
Overweight Carry	16	0.0%	64	0.1%
Cannot Conclude	932	0.8%	3458	3.1%
Driver info				
Gender (male)	23744	21.0%	62152	55.1%
Gender (Female)	3785	3.4%	23152	20.5%
Youth 15-35	13671	12.1%	45605	40.4%
Adult 36-60	10141	9.0%	31171	27.6%
Senior 61-90+	3717	3.3%	8532	7.6%

4.4.2 Methodology

To predict accident severity, eight ML algorithms were separately implemented on binary classification types (fatal/non-fatal) using Orange3.30 python base software (Demšar et al., 2013) to run. To determine the best-performing algorithm, the performance of predictive models was compared, and the most important variables were extracted from the eight models listed below.

4.4.2.1 Decision Tree (DT)_is a non-parametric supervised algorithm for learning that can be used for classification as well as regression tasks. Decision trees are a popular and powerful way of modeling decisions in machine learning. They are a type of tree data structure that begins with a root node and divides into two branches at each node. For classification and regression tasks, decision trees are used. They can be used to estimate values for a target variable or to predict the probability of an event based on features.

4.4.2.2 K-Nearest Neighbors- is an algorithm for non-parametric classification and pattern recognition (Sarker, 2021). It is frequently used to divide data into two or more classes. This algorithm begins by assigning each training instance in the dataset to the training set's most similar training instance. The distance between two instances is calculated using some measure of similarity, and it can then be any number, such as the Euclidean distance. The procedure is repeated until all data points are assigned to a class.

4.4.2.3 Support Vector Machine (SVM) - SVMs (Cortes & Vapnik, 1995) are supervised learning algorithms that can classify both continuous and categorical data. The SVM's goal is to find the best hyperplane in the input space that separates two classes of data. The hyperplane should minimize the distance between each class's nearest point.

4.4.2.4 Random Forests (RF) - Random Forest (Breiman, 2001) is a machine learning technique that is used for classification and regression. It can be used for both supervised and unsupervised learning. It has been shown to be more accurate than other algorithms in some cases. The models will then vote to determine which class is the most popular.

4.4.2.5 Neural Network (NN), - Neural network is a type of artificial intelligence that mimics the human brain. They enable computers to learn from examples and make predictions based on data patterns. A neural network is composed of many interconnected layers, each of which contains a number of nodes linked to the next layer. Weighted or unweighted links can exist between nodes in adjacent layers. In general, a node will have an output value for each input it receives from its connections with other nodes in the preceding layer, which will then be used as an input for its connections with other nodes in the subsequent layer. (Dongare et al., 2012; Géron, 2019)

4.4.2.6 Naive Bayes (NB) - It is a classification which employs probability principles to aid in estimation or predict likelihood of an event. (Webb, 2010)

4.4.2.7 Logistic Regression (LR) - is a technique in statistics and machine learning that predicts the probability of an event occurring. It is mainly used in predicting binary outcomes Logistic regression is often used for classification tasks where the dependent variable can take on any value from a discrete set of values. (Tolles & Meurer, 2016)

4.4.2.8 Gradient Boosting (GB) - Gradient Boosting choosing an optimization method by attempting to obtain each new Classifier instance. It improves its accuracy by learning from the cumulative error generated by the previous instance's prediction. (Géron, 2019)

According to table 4.4, the overall 112,837 incidents include 27 attributes from data collection that span coverage Roadway, environment, weather condition, driving behavior, driver data, and driver status. When two classification types were compared, binary models (Fatal & Non-fatal) outperformed the 3-Class model (Fatal, Major, Minor), which can be explained by the inability to effectively separate major and minor accidents in contrast, fatal accidents behaved similarly in both classification types. (Recal & Demirel, 2021).

Table 4.4 Total 28 Attributes with setting description

Attribute Name	Attribute Description
Roadway	
Highway	1 - Yes, 0-Otherwise
Dry Surface Road	1 - Yes, 0-Otherwise
Straight Way	1 - Yes, 0-Otherwise
Obstruction	1 - Yes, 0-Otherwise
Road condition	1 - Yes, 0-Otherwise
Vehicle condition	1 - Yes, 0-Otherwise
External Factor (Environment and Weather Condition)	
Day Time (06.00-18.00)	1 - Yes, 0-Otherwise
Night with Light	1 - Yes, 0-Otherwise
Night without Light	1 - Yes, 0-Otherwise
Low visibility	1 - Yes, 0-Otherwise
Clear Weather	1 - Yes, 0-Otherwise
Internal Factor (Driver Behavior)	
Drunk	1 - Yes, 0-Otherwise
Over Speed limit	1 - Yes, 0-Otherwise
Break Through Traffic lights	1 - Yes, 0-Otherwise
Break Through Traffic Signs	1 - Yes, 0-Otherwise

Table 4.4 Total 28 Attributes with setting description (Continued)

Attribute Name	Attribute Description
Overtake	1 - Yes, 0-Otherwise
Use Mobile Phone	1 - Yes, 0-Otherwise
Short Cut off	1 - Yes, 0-Otherwise
Drug	1 - Yes, 0-Otherwise
Drive in opposite direction	1 - Yes, 0-Otherwise
Doze off	1 - Yes, 0-Otherwise
Overweight Carry	1 - Yes, 0-Otherwise
Cannot Conclude	1 - Yes, 0-Otherwise
Driver info	
Gender	1- Male, 0-Otherwise
Youth 15-35	1 - Yes, 0-Otherwise
Adult 36-60	1 - Yes, 0-Otherwise
Senior 61-90+	1 - Yes, 0-Otherwise
Driver Status	
Fatality (Death)	1 – Yes, 0-Otherwise

4.4.3 Performance Measurement

A confusion matrix could be used to evaluate the performance of the machine learning method. The measurements are true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

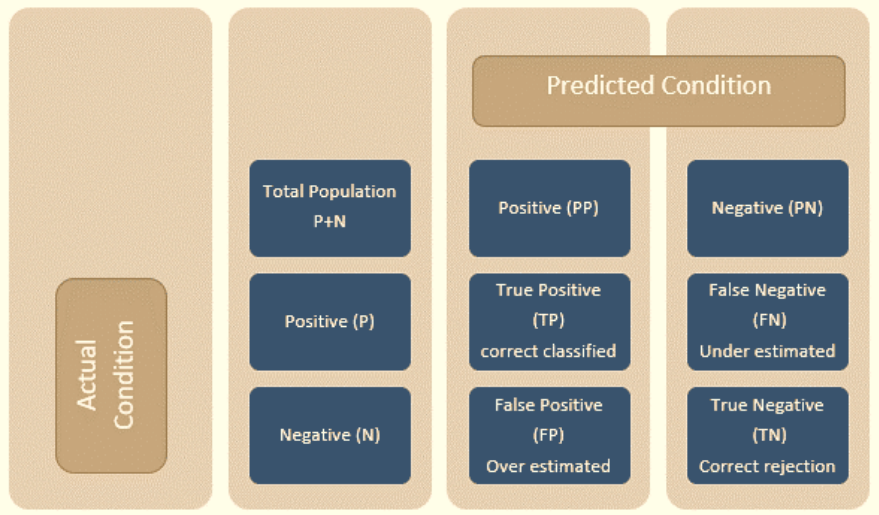

Figure 4.2 Confusion matrix diagram

Precision is a statistic that evaluates how accurate the classifier findings are. This statistic may be expressed as follows and figure 2:

$$Precision = \frac{\mathrm{TP}}{TP+FP} \tag{4.1}$$

Recall: Sensitivity (recall) Sensitivity (recall) represents the proportion of the positive class that was correctly identified.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{4.2}$$

True Negative rate also called specificity:

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TP+FN}} \tag{4.3}$$

False Positive Rate shows us how much of the negative class was misclassified by the classifier.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}} = 1 - \ \mathrm{TNR\ (Specificity)} \tag{4.4}$$

Accuracy: The ratio of correctly classification

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (4.5)$$

Information Gain is a measure of the entropy value.

$$Information\ Gain = \ Entropy\ (Initial) - \sum_{i=1}^{N} Pi\ \log_2 Pi$$

Where P is possibility of event of N (4.6)

4.5 Results

4.5.1 Model Result

According to table 4.5, the most common feature that causes motorcycle accidents for all eight models is the Highway and Gender which is ranked by top score (8 out of 8), Day time (7 out of 8) Over Speed limit (7 out 8), Night w/o light (5 out of 8), Drunk (2 out of 8), Night with light/Break through traffic light signal and Straight way (1 out of 8)

Table 4.5 Info. Gain Ranking by model

Model	Ranking		info. Gain
SVM	1	Break through Traffic light signal	0.009
	2	Highway	0.001
	3	Straight way	0.001
	4	Night with light	0.001
	5	Gender	0.000
Gradient Boosting	1	Highway	0.042
	2	Over Speed limit	0.038
	3	Day	0.037
	4	Night w/o light	0.029
	5	Gender	0.017

Table 4.5 Info. Gain Ranking by model (Continued)

Model	Ranking		info. Gain
Logistic Regression	1	Over Speed limit	0.058
	2	Highway	0.049
	3	Night w/o light	0.024
	4	Gender	0.023
	5	Day	0.019
Neural Network	1	Highway	0.046
	2	Over Speed limit	0.036
	3	Day	0.029
	4	Night w/o light	0.027
	5	Gender	0.021
Naïve Bayes	1	Over Speed limit	0.089
	2	Highway	0.075
	3	Day	0.042
	4	Gender	0.034
	5	Drunk	0.027
kNN	1	Over Speed limit	0.067
	2	Highway	0.047
	3	Gender	0.031
	4	Day	0.028
	5	Drunk	0.021
Tree	1	Night w/o light	0.043
	2	Highway	0.032
	3	Over Speed limit	0.030
	4	Day	0.013
	5	Gender	0.008

Table 4.5 Info. Gain Ranking by model (Continued)

Model	Ranking		info. Gain
Random Forest	1	Highway	0.041
	2	Over Speed limit	0.035
	3	Day	0.031
	4	Night w/o light	0.027
	5	Gender	0.015

4.5.2 Evaluation Results

Evaluation result from models in Table 4.6 & performance measurement avg over class in Figure 4.3 using sampling 75:25 (test data: train data) has high classification accuracy (Eq 5), Recall (Eq 2) and precision (Eq 1) to predict non-Fatality but low for fatality on recall.

Table 4.6 evaluation result from models

Model	Target Class	AUC	CA	Precision	Recall
SVM	Avg over	0.602	0.756	0.695	0.756
	Fatality	0.602	0.756	0.505	0.004
	Non-Fatality	0.602	0.756	0.756	0.999
kNN	Avg over	0.700	0.749	0.728	0.749
	Fatality	0.700	0.749	0.481	0.350
	Non-Fatality	0.700	0.749	0.807	0.878
Naïve Bayes	Avg over	0.724	0.762	0.726	0.762
	Fatality	0.724	0.762	0.526	0.232
	Non-Fatality	0.724	0.762	0.790	0.933
Logistic Regression	Avg over	0.729	0.764	0.725	0.764
	Fatality	0.729	0.764	0.556	0.162
	Non-Fatality	0.729	0.764	0.780	0.958

Table 4.6 evaluation result from models (Continued)

Model	Target Class	AUC	CA	Precision	Recall
Tree	Avg over	0.743	0.769	0.738	0.769
	Fatality	0.743	0.769	0.624	0.135
	Non-Fatality	0.743	0.769	0.777	0.974
Neural Network	Avg over	0.745	0.770	0.738	0.770
	Fatality	0.745	0.770	0.603	0.168
	Non-Fatality	0.745	0.770	0.782	0.964
Gradient Boosting	Avg over	0.751	0.771	0.742	0.771
	Fatality	0.751	0.771	0.616	0.168
	Non-Fatality	0.751	0.771	0.783	0.966
Random Forest	Avg over	0.768	0.777	0.752	0.777
	Fatality	0.768	0.777	0.640	0.198
	Non-Fatality	0.768	0.777	0.788	0.964

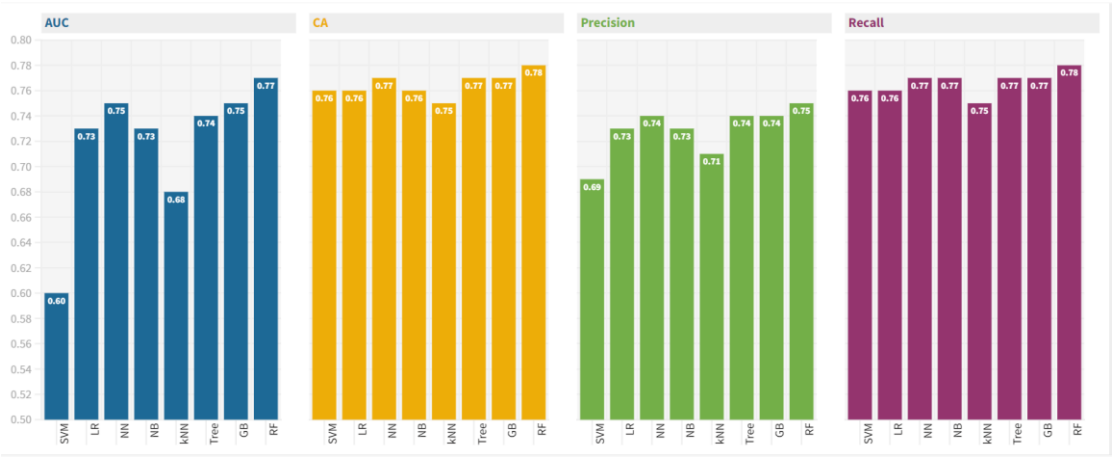

Figure 4.3 Performance Measurement models

Once the AUC (Area under the curve) of the ROC (Receiver Characteristic Operator) is between 0.5 and 1, there is a good chance that the classifier will be able to distinguish between positive and negative class values as Figure 4.4 since the classifier recognizes TF and TN (Eq 3) more than FN and FP (Eq 4).

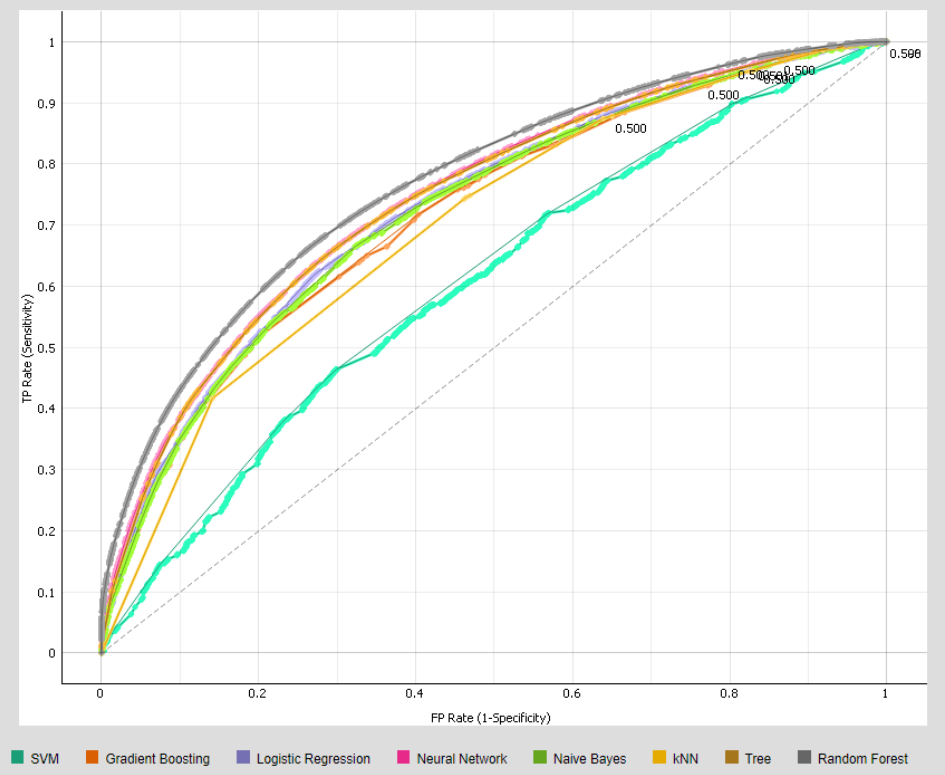

Figure 4.4 Model-specific ROC plot for predicting non-fatality.

From the evaluation model result, all models perform well for non-fatality prediction, with high precision and recall, but not for fatality prediction, with only fair precision and low recall. The best model for prediction is Random Forest, that has an average AUC of 0.768, CA of 0.777, Precision of 0.752, and recall of 0.777 follow by GB avg AUC of 0.751, CA of 0.771, Precision of 0.742, and recall 0. 771.Neural network avg AUC of 0.745, CA of 0.770, Precision of 0.738, and recall 0.770. Tree avg AUC of 0.743, CA of 0.769, Precision of 0.738, and recall 0.769. Logistic Regression avg AUC of 0.729, CA of 0.764, Precision of 0.725, and recall 0.764. Naïve Bayes avg AUC of 0.724, CA of 0.762, Precision of 0.726, and recall 0. 762.kNN avg AUC of 0.700, CA of 0.749, Precision of 0.728, and recall 0. 749. SVM avg AUC of 0.602, CA of 0.756, Precision of 0.695, and recall 0. 756.

Table 4.7 Confusion Metrix for each model

Confusion Metrix			Proportion of actual		Proportion of predicted		
					Predicted		
Model			Non-Fatality	Fatality	Non-Fatality	Fatality	Total
SVM	Actual	Non-Fatality	99.90%	0.10%	75.60%	49.50%	85308
		Fatality	99.60%	0.40%	24.40%	50.50%	27529
		Total	112637	200	112637	200	112837
Gradient Boosting		Non-Fatality	96.6%	3.4%	78.3%	38.4%	85308
		Fatality	83.2%	16.8%	21.7%	61.1%	27529
		Total	105328	7509	105328	7509	112837
Logistics Regression		Non-Fatality	95.80%	4.20%	78.00%	44.40%	85308
		Fatality	83.80%	16.20%	22.00%	55.60%	27529
		Total	104838	7999	104838	7999	112837
Neural Network		Non-Fatality	96.40%	3.60%	78.20%	39.70%	85308
		Fatality	83.20%	16.80%	21.80%	60.30%	27529
		Total	105142	7695	105142	7695	112837
Naïve Bayes		Non-Fatality	93.90%	6.70%	79.00%	47.40%	85308
		Fatality	76.80%	23.20%	21.00%	52.60%	27529
		Total	100723	12114	100723	12114	112837
kNN		Non-Fatality	87.80%	12.20%	80.70%	51.90%	85308
		Fatality	65.00%	35.00%	19.30%	48.10%	27529
		Total	92797	20040	92797	20040	112837
Tree		Non-Fatality	97.40%	2.60%	77.70%	37.60%	85308
		Fatality	86.50%	13.50%	22.30%	62.40%	27529
		Total	106887	5950	106887	5950	112837

Table 4.7 Confusion Metrix for each model (Continued)

Confusion Metrix		Proportion of actual		Proportion of predicted		
		Predicted				
Model		Non-Fatality	Fatality	Non-Fatality	Fatality	Total
Random Forest	Non-Fatality	96.50%	3.50%	78.80%	35.70%	85308
	Fatality	80.60%	19.40%	21.20%	64.30%	27529
	Total	104519	8318	104519	8318	112837

The confusion matrix in table 4.7 Random Forest is best in class, revealing 80.6% classification errors of actual F (fatalities) to NF (nonfatalities) and 3.5% of NF to F. There was a misclassifying of 21.2% for F to NF and 35.7% for NF to F for the predicted values. The worst model (SVM) misclassifies 99.6% of actual F to NF accidents and 0.1% of NF to F accidents. There was a misclassifying of 24.4% for F to NF and 49.5% for NF to F for predicted values. About information gain (Eq 6) from the model's prediction which top 5 attributes (features) from RF model consist of Highway, Over speed limit, Day time, Night w/o light and Gender. According to feature finding interpretation, all those features play a significant role in motorcycle accidents.

4.6 Conclusion & Discussion

This study aimed to focus on motorcycle crashes by focusing on the rider who was involved in the accident as well as the environmental factors that contribute to fatal crashes. To begin, identify the machine learning model that is best suited for predicting road accidents with high accuracy, as well as factors that are contributing to an increase in fatality accidents. A nonparametric analysis was performed on accident data from 2015 to 2020 to assess the significance of other elements that affect target factors such as rider information, road condition, weather condition, and rider behaviors. All eight models discovered significant classification errors of fatalities but a much better ability to predict nonfatalities. Random Forest outperformed the other seven ML algorithms in predicting road accidents in 2-Class classification giving an

average over class AUC of 0.768, CA of 0.777, Precision of 0.752, and recall of 0.777. Fatality AUC of 0.768, CA of 0.777, Precision of 0.640 and recall 0.198. Non-fatality AUC of 0.768, CA of 0.777, Precision of 0.788 and recall of 0.964. Which top 5 features which giving highest info gain consist of Highway, Riding over speed limit, Day time, Night w/o light and gender.

In many accident-related fields, ML models have been used to predict accidents. Kim et al. (2021) used to predict accidents at a Korean Container Port; the best model was chosen by comparing the accuracy, precision, recall, and F1 score of different models. The results show that a deep neural network model and a gradient boosting model outperform all other performance metrics, but the data set for port accidents is limited. While Santos et al. (2021) was discovered that a decision tree can detect the most important factors describing the severity of a road accident. Furthermore, the predictive model results indicate that the RF model could be a useful tool for forecasting accident hotspots, which is consistent with this study's finding that RF is the most accurate one to predict road accidents. For two-class prediction when compared to other methods, SVM and GB are the best in class (Recal & Demirel, 2021) while our research discovered that GB is the second runner after RF. Because of the complexities and wide range of factors involved in traffic accidents. The comparison analysis aids in determining which models outperformed and provided a useful prediction with the least amount of error, and which will be implemented. Different models work better for different data. Naive Bayes works well when features are highly independent. SVM is useful when there are too many features, and the dataset is medium in size. If the dependent and independent variables have a linear relationship, linear regression, logistic regression, and SVM are appropriate. kNN can be used with small data sets where the relationship between the dependent and independent variables is unknown. As a result, before deciding on which ML algorithm to use, the data must first be understood and analyzed or compared their accuracies on training and test sets.

The ML is effective tools in accident predicting, the model performs well in terms of non-fatality prediction, but fatality prediction still has room to improve, which can be interpreted to mean that the factors that cause fatal accidents may cause a

major accident but do not always result in a fatality. Specific feature selection may be required before entering the model to predict fatality since fatality has been a quite random feature involve, such as riding experience, rider health fit, and even the time duration from the accident area to hospital. Over in terms of information gained from the model, speed limit is the key runner for road accident (M. Yu et al., 2020), (Osman et al., 2018), (Krull et al., 2000) and a drunk cyclist is also a potential accident risk. Nonetheless, more education and stricter enforcement for intoxicated motorcycle riders may necessitate more research. For internal factors such as gender, age, accident location, and vehicle type were observed (Bahiru et al., 2018). Those were discovered to have an impact on the severity of road accidents, even though being male is still one of the leading causes of highway fatalities since gender and highway road were seen as a key factor in our study as well same as Ospina-Mateus et al. (2019) observed that men are more likely to be involved in serious accidents. As motorcycle accidents are heavily influenced by factors like as speed limit, age, Highway functional class, and speed compliance. (Rezapour et al., 2020). Ospina-Mateus et al. (2019) observed that men are more likely to be involved in serious accidents. According to previous research, intoxicated drivers have a higher accident rate. (Krull et al., 2000), (Xie & Huynh, 2012), (Kim et al., 2013), (Wu et al., 2016),(Zhou & Chin, 2019), (John & Shaiba, 2019), (Helen et al., 2019) ,(Champahom et al., 2020). The severity of the injury will increase as the rider's age below 25(Behnood & Mannering, 2017) , (Li, Ci, et al., 2019). Policymakers can use the prediction model with the most recent data set to see if there any factors changed or after the laws are implemented. Authorities should consider proposed laws to regulate speed limits and drunk riders more serious than before.

4.7 Limitations and Future Studies

It was predicted in an acceptable level (above 50% accuracy) by the model, and there is still room to improve the model and adjust the parameters in future studies to increase accuracy. The study used accident data from 2015 to 2020, but the covid-19 pandemic has spread so far in the last two years (2019-2020) that the government has ordered the country to close down and prohibit travel between

provinces, especially between the hours of 22.00 and 04.00. People are also hesitant to travel only to the separated zone, implying that they have not traveled far from home. Finally, the figures for 2019-2020 may not accurately reflect the country's true number of accidents and fatalities.

4.8 Reference

Al Mamlook, R. E., Ali, A., Hasan, R. A., & Mohamed Kazim, H. A. (2019). Machine Learning to Predict the Freeway Traffic Accidents-Based Driving Simulation. Proceedings of the IEEE National Aerospace Electronics Conference, NAECON,

Bahiru, T. K., Kumar Singh, D., & Tessfaw, E. A. (2018). Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity. Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018,

Behnood, A., & Mannering, F. (2017). The effect of passengers on driver-injury severities in single-vehicle crashes: A random parameters heterogeneity-in-means approach. *Analytic Methods in Accident Research*, *14*, 41-53. https://doi.org/https://doi.org/10.1016/j.amar.2017.04.001

Breiman, L. (2001). Mach Learn.

Bucsuházy, K., Matuchová, E., Zůvala, R., Moravcová, P., Kostíková, M., & Mikulec, R. (2020). Human factors contributing to the road traffic accident occurrence. Transportation Research Procedia,

Champahom, T., Jomnonkwao, S., Chatpattananan, V., Karoonsoontawong, A., & Ratanavaraha, V. (2019). Analysis of Rear-End Crash on Thai Highway: Decision Tree Approach. *Journal of Advanced Transportation*, *2019*, 1-13. https://doi.org/10.1155/2019/2568978

Champahom, T., Jomnonkwao, S., Watthanaklang, D., Karoonsoontawong, A., Chatpattananan, V., & Ratanavaraha, V. (2020). Applying hierarchical logistic models to compare urban and rural roadway modeling of severity of rear-end vehicular crashes. *Accident Analysis & Prevention*, *141*, 105537. https://doi.org/ https://doi.org/10.1016/j.aap.2020.105537

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Cunto, F. J. C., & Ferreira, S. (2017). An analysis of the injury severity of motorcycle crashes in Brazil using mixed ordered response models. *Journal of Transportation Safety & Security*, *9*(sup1), 33-46. https://doi.org/10.1080/19439962.2016.1162891

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., Zupan, B. (2013). Orange: Data mining toolbox in python [Article]. *Journal of Machine Learning Research*, *14*, 2349-2353. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84885599052&partnerID=40&md5=75d2df52a0c46b5ab58ab08e1576114e

DLT. (2021). Department of Land Transportation. *https://www.dlt.go.th/th/public-news/view.php?_did=2806*.

Dongare, A., Kharde, R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189-194.

Feng, M., Zheng, J., Ren, J., & Xi, Y. (2020). Association Rule Mining for Road Traffic Accident Analysis: A Case Study from UK. In *Advances in Brain Inspired Cognitive Systems* (pp. 520-529). https://doi.org/10.1007/978-3-030-39431-8_50

Geedipally, S. R., Turner, P. A., & Patil, S. (2011). Analysis of Motorcycle Crashes in Texas with Multinomial Logit Model. *Transportation Research Record*, *2265*(1), 62-69. https://doi.org/10.3141/2265-07

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow. http://oreilly.com/catalog/errata.csp?isbn=9781492032649

Guido, A. C. M. S. (2017). Introduction to machinelearning with python. http://oreilly.com/catalog/errata.csp?isbn=9781449369415 (Third Release) (O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.)

Harb, R., Yan, X., Radwan, E., & Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests [Article]. *Accident Analysis and Prevention*, *41*(1), 98-107. https://doi.org/10.1016/j.aap.2008.09.009

Helen, W. R., Almelu, N., & Nivethitha, S. (2019). Mining Road Accident Data Based on Diverted Attention of Drivers. Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018,

Huo, X., Leng, J., Hou, Q., & Yang, H. (2020). A Correlated Random Parameters Model with Heterogeneity in Means to Account for Unobserved Heterogeneity in Crash Frequency Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *2674*, 036119812092221. https://doi.org/10.1177/0361198120922212

Jafari Anarkooli, A., Hosseinpour, M., & Kardar, A. (2017). Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model. *Accident Analysis & Prevention*, *106*, 399-410. https://doi.org/10.1016/j.aap.2017.07.008

John, M., & Shaiba, H. (2019). Apriori-Based Algorithm for Dubai Road Accident Analysis. Procedia Computer Science,

John, M., & Shaiba, H. (2022). Analysis of Road Accidents Using Data Mining Paradigm. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 68, pp. 215-223).

Jomnonkwao, S., Uttra, S., & Ratanavaraha, V. (2020). Forecasting Road Traffic Deaths in Thailand: Applications of Time-Series, Curve Estimation, Multiple Linear Regression, and Path Analysis Models. *Sustainability*, *12*(1). https://doi.org/10.3390/su12010395

Kim, J.-K., Ulfarsson, G. F., Kim, S., & Shankar, V. N. (2013). Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. *Accident Analysis & Prevention*, *50*, 1073-1081. https://doi.org/https://doi.org/10.1016/j.aap.2012.08.011

Kim, J. H., Kim, J., Lee, G., & Park, J. (2021). Machine Learning-Based Models for Accident Prediction at a Korean Container Port. *Sustainability*, *13*(16), 9137. https://www.mdpi.com/2071-1050/13/16/9137

Krull, K. A., Khattak, A. J., & Council, F. M. (2000). Injury Effects of Rollovers and Events Sequence in Single-Vehicle Crashes. *Transportation Research Record*, *1717*(1), 46-54. https://doi.org/10.3141/1717-07

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

Kuşkapan, E., Çodur, M. Y., & Atalay, A. (2021). Speed violation analysis of heavy vehicles on highways using spatial analysis and machine learning algorithms [Article]. *Accident Analysis and Prevention*, *155*, Article 106098. https://doi.org/10.1016/j.aap.2021.106098

Li, Z., Chen, C., Wu, Q., Zhang, G., Liu, C., Prevedouros, P. D., & Ma, D. T. (2018). Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. *Analytic Methods in Accident Research*, *20*, 1-14. https://doi.org/https://doi.org/10.1016/j.amar.2018.08.001

Li, Z., Ci, Y., Chen, C., Zhang, G., Wu, Q., Qian, Z., Ma, D. T. (2019). Investigation of driver injury severities in rural single-vehicle crashes under rain conditions using mixed logit and latent class models. *Accident Analysis & Prevention*, *124*, 219-229. https://doi.org/https://doi.org/10.1016/j.aap.2018.12.020

Mafi, S., AbdelRazig, Y., & Doczy, R. (2018). Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups. In *Transportation Research Record* (Vol. 2672, pp. 171-183).

Malin, F., Norros, I., & Innamaa, S. (2019). Accident risk of road and weather conditions on different road types. *Accid Anal Prev*, *122*, 181-188. https://doi.org/10.1016/j.aap.2018.10.014

Mohamad, I., Jomnonkwao, S., & Ratanavaraha, V. (2022). Using a decision tree to compare rural versus highway motorcycle fatalities in Thailand. *Case Studies on Transport Policy*, *10*(4), 2165-2174. https://doi.org/https://doi.org/10.1016/j.cstp.2022.09.016

Mphela, T. (2020). Causes of road accidents in botswana: An econometric model [Article]. *Journal of Transport and Supply Chain Management*, *14*, 1-8, Article a509. https://doi.org/10.4102/jtscm.v14i0.509

Osman, M., Mishra, S., & Paleti, R. (2018). Injury severity analysis of commercially-licensed drivers in single-vehicle crashes: Accounting for unobserved heterogeneity and age group differences. *Accident Analysis & Prevention*, *118*. https://doi.org/10.1016/j.aap.2018.05.004

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists [Article]. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051-10072. https://doi.org/10.1007/s12652-020-02759-5

Ospina-Mateus, H., Quintana Jiménez, L. A., López-Valdés, F. J., Morales-Londoño, N., & Salas-Navarro, K. (2019). Using Data-Mining Techniques for the Prediction of the Severity of Road Crashes in Cartagena, Colombia. In *Communications in Computer and Information Science* (Vol. 1052, pp. 309-320).

PDPM. (2020). Thailand Department of Public Disaster Prevention and Mitigation. *https://www.disaster.go.th/en/*

Recal, F., & Demirel, T. (2021). Comparison of machine learning methods in predicting binary and multi-class occupational accident severity [Article]. *Journal of Intelligent and Fuzzy Systems*, *40*(6), 10981-10998. https://doi.org/10.3233/JIFS-202099

Rezapour, M., Mehrara Molan, A., & Ksaibati, K. (2020). Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. *International Journal of Transportation Science and Technology*, *9*(2), 89-99. https://doi.org/10.1016/j.ijtst.2019.10.002

Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. *Computers*, *10*(12), 157. https://www.mdpi.com/2073-431X/10/12/157

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Shaheed, M. S., Gkritza, K., Zhang, W., & Hans, Z. (2013). A mixed logit analysis of two-vehicle crash seventies involving a motorcycle. *Accident; analysis and prevention*, *61*. https://doi.org/10.1016/j.aap.2013.05.028

Shweta, Yadav, J., Batra, K., & Goel, A. K. (2021). A Framework for Analyzing Road Accidents Using Machine Learning Paradigms. Journal of Physics: Conference Series,

Sonal, S., & Suman, S. (2018). A framework for analysis of road accidents. 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research, ICETIETR 2018,

Tolles, J., & Meurer, W. J. (2016). Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*, *316*(5), 533-534. https://doi.org/10.1001/jama.2016.7653

Webb, G. I. (2010). Naïve Bayes. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 713-714). Springer US. https://doi.org/10.1007/978-0-387-30164-8_576

WHO. (2018). World Health Organization: Global status report on road safety 2018. . *https://extranet.who.int/roadsafety/death-on-the-roads/*.

Wu, Q., Zhang, G., Zhu, X., Liu, X. C., & Tarefder, R. (2016). Analysis of driver injury severity in single-vehicle crashes on rural and urban roadways. *Accident Analysis & Prevention*, *94*, 35-45. https://doi.org/https://doi.org/10.1016/j.aap.2016.03.026

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

Xie, Y., & Huynh, N. (2012). Analysis of driver injury severity in rural single-vehicle crashes. *Accident; analysis and prevention*, *47*, 36-44. https://doi.org/10.1016/j.aap.2011.12.012

Yu, M., Zheng, C., & Ma, C. (2020). Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. *Analytic Methods in Accident Research*, *27*, 100126. https://doi.org/ https://doi.org/10.1016/j.amar.2020.100126

Zhou, M., & Chin, H. C. (2019). Factors affecting the injury severity of out-of-control single-vehicle crashes in Singapore. *Accident Analysis & Prevention*, *124*, 104-112. https://doi.org/10.1016/j.aap.2019.01.009