IJTE, Vol. 12, No. 2, Autumn 2024, pp. 1719-1745
Research Paper

Developing a Framework for Selecting an Appropriate Model based on the Ensemble Learning

Alireza Mahpour[1,*], Mostafa Shafaati[2]

Received: 2024/01/27 *Accepted: 2024/10/08*

Abstract

We present a framework for selecting the optimal ensemble learning model based on 143310 crash observations with five classes. For non-ensemble models, we use five common models. 26 ensemble learning models are derived from these five models. We suggest Diff2 and Diff3 measures for choosing the right model. The diff2 is the difference between observations classified incorrectly as class 1 and incorrectly classified as class 3, 4, or 5. In Diff3, we compare observations misclassified as class 1 or 2 with observations misclassified as class 4 or 5. We select the best model based on the following criteria: for class 1, the largest R1, for class 2, the largest "Diff2", for class 3, a negative "Diff3", and for classes 4 and 5, the highest "F1-score". The paper ranks 31 models based on its criteria. There are five ranking series. By comparing these rankings, we can determine, for example, whether the 3rd best model for class 1 corresponds to the best model for class 2. For each model, 5 "Ranks" are determined. Relationships between the ranks were then evaluated. Rank1 and Rank2, Rank3 and 5 have a relatively strong relationship. A negative and relatively strong correlation exists between Rankings 2 and 3, as well as Rankings 2 and 5.

Keywords: Crash Severity Prediction, Machine Learning Model, Ensemble Voting Classifier, Imbalanced Multi-Class Classification

[*] Corresponding author. E-mail: a_mahpour@sbu.ac.ir
[1] Faculty of Civil, Water and Environmental Engineering, Shahid Beheshti University, Tehran, Iran
[2] Faculty of Civil and Environmental Engineering, Tarbiat Modares University, Tehran, Iran

1. Introduction

In accordance with the World Health Organization, 1.35 million people die each year as a result of road accidents, and 50 million suffer severe injuries as a result of them (WHO, 2018). As a result, it may be desirable for transport safety researchers to strive to reduce the severity of crashes (Islam et al., 2023). The analysis of crash data has been greatly aided by the use of statistical and machine learning (ML) models. It has been widely accepted that regression models are a valuable tool in the analysis of crash data (Haeri, et al., 2024; Eluru et al., 2008; Sahebi et al., 2015; Mannering et al., 2016; Vajari et al., 2020; Liu, 2022; Ahmed et al., 2023; Kabli et al., 2023; Feknssa et al., 2022; Mahpour et al., 2023). Statistical models are used in other area of transportation safety such as using surrogate safety measures for assessing pedestrian safety (Shafaati, and Boroujerdian, 2020). Also, decision making strategies are used to investigate transportation safety as well (Mahpour et al., 2021). The use of statistical models, however, is limited by the substantial reliance on statistical assumptions. Contrary to statistical assumptions, ML models do not require statistical assumptions before they are built (Santos et al., 2022). In this regard, (Mokhtarimousavi et al., 2020) believe that, there is a keen interest in the use of machine learning models for crash prediction in recent years. Ensemble learning is one of the most powerful ML models. It has been demonstrated that ensemble methods can perform better predictively than any of their constituent learning algorithms alone (Rokach, 2010). In addition, certain ensemble learning techniques, such as XGBoost, have the advantage of reducing computational effort and avoiding overfitting issues (T. Chen & Guestrin, 2016). The use of ensemble models has not received enough attention in terms of road safety, particularly the analysis of crash injury severity (Jamal et al., 2021). Thus, it appears that focusing on ensemble learning and examining its different aspects, particularly when dealing with highly imbalanced crash severity data, is of utmost importance, which is the primary concern of this paper. The remainder of the paper is as follows: in section 2, we review the literature regarding the use of ensemble learning in the prediction of crash severity. The methodology and proposed framework for selecting the best model are described in section 3. Section 4 presents results and discusses them, and section 5 concludes our study and offers several suggestions for possible future studies.

2. Literature Review

However, there are a number of papers that utilize one of the different ensemble learning techniques and compare it with other machine learning models. Multiple papers have focused on the use of Random Forests (RF) as an ensemble learning technique. In Hong Kong, Krishnaveni and Hemalatha (2011) found that among RF, Nave Bayes (NB), AdaBoost, Part Rule, and Decision Tree (DT), RF was the most effective method. According to Umer et al. (2020), their results indicate that the RF is the best model after trying several models, including several types of models. Other models include the Voting Classifier based on Logistic Regression (LR) and Stochastic Gradient Descent, the AdaBoost Classifier, the Extra Tree Classifier, and the Gradient Boosting Machine. In their 2022 paper, Bokaba et al. compare K-Nearest Neighbor (K-NN), LR, NB, AdaBoost, and Support Vector Machines (SVM). Based on five evaluation metrics, the RF was found to be the most effective. Singh et al., (2018) utilized multinomial logit, DT, and RF in their study. In comparison to the other two models, RF performed better. A comparison was made between RF and Multinomial Logit by Wang and Kim (2019). The RF was found to outperform the multinomial logit model. In line with other crash severity analyses, this study

uses data that is imbalanced with three classes in which fatal crashes only account for 0.47% of all observations. As an example, Wahab et al., (2019) investigated the performance of J48 DeT, RF, Instance-Based Learning, and a multinomial logit model in predicting the severity of motorcycle crashes in Ghana. Based on the results, it was determined that the RF performed better than the others in terms of accuracy. The Multinomial Logit model and the RF were used by Gan et al., (2020) to predict crash severity on highways with and without traffic hazards. Upon comparing these two models, it was found that the RF had a greater degree of accuracy. The results of Chen et al. (2020) showed that the RF is more effective than the LR and Regression Tree. According to Vajari et al., (2020), six machine learning models, including the RF, were tested for performance. In addition to Multinomial LR, NB, DT, RF, and SVM, they also examined Multilayer Perceptrons. The RF produced the most accurate predictions once again. RF has been proven to be more effective compared with artificial neural networks and deep neural networks by Lee et al., (2019). A comparison of K-NN, LR, NB, AdaBoost, SVM, and RF was carried out by Bokaba et al., (2022). Using five evaluation metrics, the RF was found to be superior to all the other approaches. According to Yang et al., (2023), the RF is more accurate than SVM and two types of neural networks. It should be noted, however, that some studies have demonstrated that ensemble models are not superior to other models. In order to address the problem of imbalanced datasets, Fiorentini et al. (2020) used Random Under Sampling the Majority Class (RUMC) and then utilized RF, K-NN, and LR. The K-NN model had the best true positive rate for the minority class (which included fatal accidents and injuries). According to (Ijaz et al., 2021), DJ was the most accurate with an overall accuracy of 83.7 % in predicting the severity of crashes involving Wheeled Motorized Rickshaws in Pakistan. According to Azhar et al., 2022, DT and RF were compared in crashes involving heavy vehicles. A slightly better accuracy was achieved by the RF classifier, according to the study. In their research, (Chakraborty et al., 2021) found that Deep Neural Net Classifier is more effective than RF when it comes to the rarest classification of the data. (Razapour et al., 2021) have shown that RF predicts motorcycle injury severity better than SVM, Logistic Regression, and MARS.

Also, other ensemble learning techniques have been evaluated. For example, for crash injury prediction, Jeong et al. (2018), compared LR, DT, Neural Network (NN), RF, AdaBoost, NB, and Gradient boosting models. It was found that decision trees combined with bagging outperformed the others. In another study about crash severity prediction, Tang et al. According to (2019), a two-layer stacking model with RFs outperformed other models such as Adaptive Boosting, Gradient Boosting, SVM, Multilayer Perceptron, and an RF all together. Mokoatle et al. (2019) analyzed 1,525 road crashes and used Multivariate Logistic Regression and Extreme Gradient Boosting Trees to model the injury severity of the crash participants. The Extreme Gradient Boosting Trees model achieved better performance than the Multivariate Logistic Regression model. Goswamy et al. (2023) compared XGBoost and Random Parameter Discrete Outcome Models (RPDOM) for nighttime pedestrian crashes in locations with and without a certain facility. It was found that the XGBoost model's accuracy was 97% while for the RPDOM, the accuracy was 73.8%. Zhang et al. (2023) compared the performance of Ordered Forest (ORF) and RF in predicting injury severity in single-cycle crashes in the UK. The results showed that ORF outperformed the RF. Additionally, (Abdulazeez et al., 2023) have found that Adaboost is better than other machine learning models at predicting crash severity in child occupants in the UAE.

There are some papers about the application of ensemble learning in other areas of transport safety analysis. For example, there are papers in

which it was found that XGBoost outperforms other machine learning algorithms such as LR, NN, SVM, BN, and gradient boosting in predicting the likelihood of traffic crashes (Mousa et al., 2019; Schlögl et al., 2019). In their work, Parsa et al. Showed that using the XGBoost technique for feature analysis and real-time crash detection resulted in a detection rate of 79 % and an accuracy of 99% (Parsa et al., 2020). Guo et al. Analyzed crash injury severity factors in traffic crashes involving elderly pedestrians (age >65 years) with the emerging XGBoost model (Guo et al., 2021). To identify the variables affecting crash severity of automated vehicles, (Chan et al., 2020) employed XGBoost and Classification and Regression Tree (CART) models. After comparing the models to each other, the former outperformed the latter. According to (Pradhan and Sameen, 2020), CGBoost performed well in crash severity prediction. In their research, (Ma et al., 2019) compared the performance of XGBoost and grid analysis with five traditional algorithms (RF, LR, MLP, SVM, and RF) to analyze factors contributing to fatal crashes in Los Angeles. XGBoost, with a modelling accuracy of 86.73%, outperformed other methods.

Studying previous research shows that a lot of efforts have been made to demonstrate the power of ensemble learning. However, when it comes to evaluating the performance of different combinations of models, finding relevant papers seems difficult. Also, when dealing with an imbalanced multiclass classification is the case, choosing the best combinations of models is a necessary task. This is because when our problem is crash severity analysis, facing highly imbalanced classes is inevitable. Therefore, this paper seeks answers to the following questions:

- In ensemble learning, how can simple models such as Decision Trees (T), Logit (L), and different types of Nave Bayes models such as Gaussian (G), Complement (C), and Bernoulli (B) result better?
- How can we choose the most appropriate model for an imbalanced multi-class classification problem?
- Which metrics may play more significant rules in determining the most appropriate models and are the existing metrics enough for choosing the most appropriate models?

To answer the questions of this paper, California crash data in year 2012 is used (Nujjetty et al., 2014). The dataset includes 143310 observations with 5 imbalanced injury levels. This paper solves a five-class classification problem. All possible combinations of models in voting ensemble settings and non-ensemble models are built. In total, 31 models are built, and their performance for each class is compared.

3. Methodology

The objective of this section is to provide an introduction to the machine learning classification algorithms used in this study. For classifying datasets, machine learning classifiers use supervised training algorithms. Their multidimensional data processing abilities, flexibility in deployment, versatility, and superior prediction capabilities can help them produce promising results. We implement three algorithms using Python analytic platform: Logistic Regression (L), Naive Bayes (NB), and Decision tree (T). The detailed methodology is presented in the following paragraphs. In machine learning, ensemble learning is the process of combining multiple individual models to increase prediction accuracy. Decision tree, neural networks, and SVM are all examples of different types of models that can be included in an ensemble.

3.1. Defining Non- Ensemble Models

In this subsection the non-ensemble models used in this paper to construct the ensemble models are introduced briefly.

3.1.1. Logistic Regression (L)

The logistic regression function is commonly used in logistic classification for classifying data. In fact, logit models are commonly used in

transportation planning and engineering (YousefAbadi et al., 2021).

3.1.2. Decision Tree (T)

Decision Tree is a supervised learning approach which is mostly used as a predictive model. This model is literally a tree with leaves and branches representing class labels and conjunctions of features that make the class labels appear (Studer et al., 2018).

3.1.3. Naïve Bayes

A Bayes theorem model is used in this model to determine the probability of an event in the presence of certain conditions (Zhang and Harry, 2004). According to Zhang and Harry (2004), NB algorithms assume that two features are conditionally independent. In the case of y being the class variable and xi being the features, then (Murty et al., 2011);

$$P(y|x_1,\dots,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y) \xrightarrow{yields} \hat{y} = argmax\, P(y)\prod_{i=1}^{n} P(x_i|y) \quad (1)$$

- **Gaussian Naïve Bayes (G)**

If it is assumed that features' likelihood is Gaussian, then the type of NB will be Gaussian. Equation 2 shows the probability function of this type of NB:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (2)$$

Using maximum likelihood, σ_y and μ_y can be estimated.

- **Bernoulli Naïve Bayes (B)**

Another NB method used in this paper is Bernoulli Naïve Bayes. This model works well especially when the features are binaries. Equation 3, shows the probability formula for B (Schütze et al., 2008; Metsis et al., 2006):

$$P(x_i|y) = P(x_i = 1|y)x_i + (1 - P(x_i = 1|y)(1 - x_i) \quad (3)$$

- **Complement Naïve Bayes (C)**

Rennie et al. (2003) describe a complement naive Bayes classifier. By using a Complement Naive Bayes classifier, the "severe assumptions" of the Multinomial Naive Bayes classifier may be corrected. Data sets that are imbalanced are particularly suited to this method.

It should be mentioned that the LR model output is a probability, and it can be used as a classifier by defining a cut-off point (Laskaris, 2015) In ensemble learning, the predictions of the individual models are combined based on a voting mechanism (Ryu et al., 2010).

3.2. SMOTE Resampling

As the name implies, Synthetic Minority Oversampling Technique (SMOTE), is a technique that is applied to imbalanced datasets (Chawla et al., 2002). As part of this technique, a minority class instance "a" is randomly selected and then its k closest minority class neighbors are identified. It is then determined whether one of the k nearest neighbors b should be chosen at random as the synthetic instance. In the feature space, "a" and "b" are then connected, forming a line segment. As a final step, synthetic instances are generated based on a convex combination of "a" and "b" (Weiss et al., 2013).

3.3. Evaluation Metrics

Due to the study objective, Precision, Recall, False Positive Rate (FPR), True Positive, False Positive, and Accuracy are used for evaluation. These evaluation metrics are described in the following paragraphs. For classification problems, and in this study's case, a confusion matrix has been developed. This matrix is composed of four possible scenarios: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Table 1. Four possible scenarios in the models

	Predicted Condition	
Actual Condition	**Positive**	**Negative**
Positive	True Positive (TP)	False Negative (FN)

	Predicted Condition	
Actual Condition	**Positive**	**Negative**
Negative	False Positive (FP)	True Negative (TN)

Considering table 1, the evaluation metrics needed for this study can be calculated using equation 4 and 5.

$$Recall\ (R) = \frac{TP}{TP + FN} \quad (4)$$

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

$$F1 = \frac{2 \times P \times R}{P + R} \quad (6)$$

3.4. Data Preprocessing, and Description

Missing values are not present in the data used in this paper, except for a few variables such as "Gender". To resolve this problem, the most frequent cases were used to fill in the missing observations. One common method of filling in missing values in a categorical feature is by using a mode (Kuhn and Johnson, 2013). Using mode to fill in missing values means we fill in missing values with the most probable case of that feature. The gender of 9320 observations is missing. While eliminating the observations with missing values is a solution, it will result in a relatively high loss of information (in this case, 6 percent of all observations have missing values in the variable "Gender"). As 'male' is the most frequent, the empty cells have been filled with male. In other words, the most probable case is "male", thus we fill in the missing value of the feature "Gender" with "male". It is important to note that the missing values occurred completely randomly in our dataset, with no relationship between the missing values and the observed values. Thus, the probability of missing a value is the same for all observations, regardless of their value. In cases where the missingness is random, a simple technique such as filling in the mode would be an acceptable solution. In some cases, however, we had to remove the variable because the number of missing values was too high. Weather 2 and Contrib2 were the only variables in the dataset that were affected by this issue. A total of 139422 observations were missing from Contrib2 and 137294 observations were missing from Weather2. These observations account for almost 97, and 95 percent of all observations respectively. It means that we have almost no information regarding those two features, therefore dropping them from the dataset seems logical. It should be mentioned that if the variable in the study was not a categorical one, other methods such as filling with the average could be used (Shafaati, and Saffarzadeh, 2023; Shafaati, and Saffarzadeh, 2024).

The dataset is heavily skewed by outliers. We employed the quantile method to address the problem of outliers. Outlier detection (also known as anomaly detection) is the process of finding data objects with behaviors very different from expectations. Such objects are called outliers or anomalies (Han et al., 2022). After that, the skewness of the continuous variables was determined to ensure that they fell within a range of -1 to +1. This approach to treating outliers was proposed by (Hubert, and Vandervieren, 2008). There is an option of deleting observations that contain outliers. In spite of this, deleting them would have resulted in a significant loss of information. The number of outliers for variables such as "MEDWID", which represents median width, would be 21902 even if the upper quantile is assumed to be 0.85 and the lower marginal, 0.15. About 15% of the observations would fall into this category. In light of one variable's outliers, we could not reasonably remove at least 15 percent of our observations. The best course of action would be to replace outliers. A normalization of the continuous variables was performed after outliers were replaced with upper and lower quantiles, and the categorical variables were converted into dummies. Utilizing an RF model, a feature selection analysis was conducted to select the most critical variables. There are 143310 observations in total. Data

was randomly divided into two groups, training (60%) and testing (40%) in Table 3. The training and test sets for each group are shown in Table 3. In addition, table 2 provides a brief description of the selected features.

Table 2. Description of variables prior to preprocessing

Variable	Definition	Mean	Min	Max
MILEPOST	Reference point where the crash occurred	20.04	0	186
LANEWID	Calculated average lane width of the roadway segment	40.97	3	89
SEG_LNG	Section length in miles.	0.18	0	6.43
MEDWID	Median width (in feet).	32.67	0	99
AADT	Calculated average AADT	111947.1	0	354772
LSHLDWID	Width of left shoulder of the roadway segment	4.82	0	26
PAV_WDL	Width of left paved shoulder of the roadway segment	4.53	0	26
SURF_WID	Width of traveled way of the roadway segment	37.12	0	83
DRV_AGE	the age of the driver of the vehicle involved in the crash.	117.41	0	998
LIGHT_A	The type/level of light that existed at the time of the crash. Daylight	-	0	1
LIGHT_C	The type/level of light that existed at the time of the crash. Dark - Street Lights	-	0	1
DRV_SEX_F	DRIVER GENDER (Female)	-	0	1
CONTRIB1_N	Violation or factor contributing to the crash. None Apparent	-	0	1

Table 3. Train and Test Set

	Class*	N**	Percent
Train	**1**	629	0.73%
	2	1865	2.17%
	3	8460	9.84%
	4	19309	22.46%
	5	55722	64.80%
	Class	**N**	**Percent**
Test	**1**	420	0.73%
	2	1243	2.17%
	3	5640	9.84%
	4	12873	22.46%
	5	37148	64.80%

*'1' Killed (Died No Later Than 30 Days after Collision), '2' Severe Injury, '3' Other Visible Injury, '4' Complaint of Pain, '5' Non-Injury (PDO) Crash

4. Results and Discussion

The evaluation results of all 31 models are indicated in table 4, and table A1. The algorithm used for all ensemble models is ensemble voting classifier. The results of confusion matrices can be seen in the appendix, in table A1.

As mentioned before, in this study, two classes, namely class 1, and 2 are more important to be predicted correctly while this is less critical for Classes 3, 4, and 5. In spite of this, it would not be logical to increase the number of true predictions for classes 1 and 2 while increasing the number of false predictions for classes 5, 4, and 3. Since, if a model incorrectly incorporates a large number of class 5, 4, and 3 into class 1 and 2, it may lead us to either the wrong or expensive countermeasures for reducing the risk of fatal and severe injury crashes in unnecessary locations. A variety of single and ensemble learning models are discussed in light of this reasoning. A high R1 and R2 value is desirable in the first consideration. According to table 5, when Bernoulli (B) and Complement (C) are used together in an ensemble learning model, R1 is the highest. A model that incorporates Tree and Complement, with an R1 of 0.652, is the second-best model. According to the table, and figure 1, the lowest R1 belongs to decision tree (T) when used alone. However, when this model is accompanied by Complement Naïve in an ensemble learning model, the model with the second largest R1 is obtained. It should also be noted that while T (decision tree) has the lowest R1 among all non-ensemble models, Complement has the highest. A combination of these two models is superior

to both the strong single model (Complement) and the weak one (Tree). An ensemble model demonstrates its power in this instance. Thus, even the weakest model can enhance the power of the overall model when utilized in an ensemble setting. There is also an important point to take into consideration that for all models with the R1 greater than 0.6, one of the Naive Bayes models is present, and the most appropriate model is composed of two types of Nave Bayes models (Bernoulli and Complement). This is an evidence of the power of Naïve models for solving imbalanced multiclass classification problems. A top priority is given to identifying the class that contains the least number of members. Furthermore, using non-ensemble models in an ensemble framework enhances their prediction power except for the complement (C).

It is also desirable to have a high recall for class 2 (R2) since it represents "severe injury". The results in table 4 indicate, however, that in nearly all models, R2 is low (usually less than 0.200). In the case of this paper's problem, it may appear that a model with the highest R2 is the best choice, however, greater caution may be needed, so that, in the event a class-two observation is misclassified, it is better that the observation is falsely categorized as class 1, not class 3, 4, or 5. As a result, if a "severe-injury" crash is mistakenly predicted to be a PDO crash, wrong policies and countermeasures may be chosen, which may not contribute to the improvement of safety. Hence, we need to ensure that the number of misclassification observations for class 1 is greater than that for classes 3, 4, and 5. Therefore, we introduce a new measure called "Diff2" which is calculated using the confusion matrices in table “A” presented in the appendix. This measure represents the difference between the number of class 2 observations that are falsely predicted as classes 1 and the number of class 2 observations that are falsely predicted as classes 3, 4 and 5. Models with a larger "Diff2" perform better for class 2 than those with a smaller one. Table 5 shows that GT based on Diff2 is the most appropriate model for class 2 observations, whereas we would have selected BLT (Bernoulli + Logit + Tree) based on R2 measure. In this regard, Diff2 makes a significant difference in our choice for class 2. Since the same logic can be applied to class 3 observations, we introduce "Diff3" as well. “Diff3” refers to the difference between the number of class 3 observations that were incorrectly classified as either class 1 or class 2 and the number of class 3 observations that were incorrectly classified as either class 4 or class 5. However, our preference for Diff3 differs from that for Diff2. For Diff2, the highest number should be assigned to the most appropriate model, whereas for Diff3, the lowest number should be assigned to the most appropriate model. The reason is because class 3 observations are not “severe injuries”, but are “other visible” injuries, it appears that if the model misclassifies class 3 observations, placing them in class 4 or 5 would be more beneficial than placing them in class 1 or 2.

In this regard, models that include Naive Bayes are more effective than the others. Based on the results of table 5, GT and CT are the top models in the Rank3 column. Each "Rank" ranks the models according to their performance in predicting a particular class. For example, if Rank3 for a particular model is 6, it means that the model is the 6th best in terms of predicting class 3 observations.

The results of this study indicate that if ensemble learning models are chosen to be trained, including Naive Bayes models can be useful in predicting minority classes in the case of imbalanced multiclass classification. According to the earlier explanations, both classes 1 and 2 are very important. It is therefore highly recommended that a model be developed that prioritizes the prediction of these classes. Despite this, prioritizing minority classes at the expense of extremely low recall values for majority classes may lead to substantial difficulties at the end of the process. If, for

example, our ensemble model incorrectly predicts too many observations from class 5 as class 1 in its prediction results, policy makers may end up with highly expensive countermeasures at a time when "doing nothing" might be the best course of action. In this regard, it is crucial to evaluate how the model performs when dealing with majority-classes as well. It is for this reason that CT (Complement + Tree) may not be the most suitable model. In light of the reasoning outlined above, it appears that the following recommendations may be worth considering in order to select the most appropriate model:

- It is preferable to use the model with the highest R1 value for class 1 crashes ('fatal crashes').
- For class 2 crashes with severe injuries, the model with the largest "Diff2" is preferred.
- For class 3, where visible injuries are considered, a negative "Diff3" is preferred, and since these injuries are not expected to be severe, a highly negative "Diff3" can be used.
- It would be preferable to use a model with the highest value of "F1-score" for the classes 4 and 5 that are "complaint of pain" and "PDO" crashes, respectively.

Based on these four points, we are able to select the most appropriate model from all 31 trained models presented in this paper. According to Table 5, all 31 models are sorted according to the a to d recommendations. According to the framework proposed by this research, Rank1, 2, 3, 4, and 5 indicate the models' rankings for classes 1 to 5 for classes 1 to 5. Prior to providing more details, we would like to point out that, due to the absence of some important independent variables, the models built in this research may not be the best available. It is evident that accuracy, precision, and recall could have been increased if there had been unlimited access to all possible independent variables. Nevertheless, since this paper is primarily a methodological one, we must consider two points before proceeding. First, we attempted to construct the best ensemble models possible. The second point concerns the purpose of this paper. This paper aims at developing ensemble learning models for imbalanced multiclass classification. Furthermore, it offers a logical approach to selecting the appropriate model out of all the possible combinations. As a matter of fact, the paper presents a logical way of making trade-offs whenever faced with an imbalanced multiclass classification problem in which the most critical classes in terms of predictions are the minority classes with a small number of members. In light of table 4, as well as table 5, it would be easier to select the model that is most suitable.

As can be seen from table 5, models 1 or 2 in Rank1 correspond to models 4 and 2 in Rank2. It is equivalent to model 1 in Rank 4 if we select model 2 in Rank 1. In spite of this, this model represents the lowest-performing models in Ranks 3 and 5. In table 5, comparing the Rank columns may provide insight into imbalanced multiclass classification problems using ensemble learning. Firstly, models that are best at predicting minority classes in multiclass classification problems are the worst at predicting majority classes. An example would be that the first best model for class 1 (Rank 1) is the 25th best model for class 5 (Rank 5). The second best model for class 2 (Rank 2) is also the second poorest model (30th model) for class 5 (Rank 5). Figs. 1 to 10 provide an overview of the relationship between the most appropriate models for different classes. In Figure 1, the relationship between the best models for classes 1 and 2 is illustrated. Figure 2 illustrates the relationship between the best models for classes 1 and 2. According to table 5, the figures have been plotted. Figure 1 illustrates that in most cases, the best models for class 1 are also acceptable for class 2. According to figures and tables 5, the best models for class 5 correspond to the least suitable ones for class 1. Based on the explanations provided above, model 2 might be the best option for class 1. The best choice may be model 10 for class 1 if we wish to make

fair predictions for all classes without prioritizing any class. The model corresponds to model 15 for class 2, 5 for class 3, 11 for class 4, and 12 for class 5.

Figures 1 to 5 illustrate the comparison of the models' performance visually. In accordance with our proposed framework, to evaluate models' performances in relation to Class1, R1 can be considered a reliable measure. As illustrated in figure 1, BC (Bernoulli + Complement) achieves the highest R1, followed by CT and GT. In most of the models, the R1 score falls between 0.5 and 0.6. There are 6 models with R1s higher than 0.6, all consisting of at least one of the Naïve Bayes models. Also, there are only 4 models in which the R1 score is less than 0.5 with 2 of them less than 0.4. These four weak models are T (decision tree), B (Bernoulli), G (Gaussian), and L (Logit). The important point here is that the Complement (C) model alone outperforms multiple models including BCGLT which is the ensemble model that consists of all models used in this study. This shows the power of Complement Naïve Bayes model when predicting the minority class in a multiclass classification model.

In figure 2, the models' performances are compared based on their prediction powers for class 2. As mentioned earlier, this comparison is based on the measure "Diff2", and the higher "Diff2" is, the better the models' performances. As illustrated in figure 2, the maximum value of "Diff2" belongs to GT (Gaussian + Tree). According to figure 2, when used alone, both G and T result in poor and even negative "Diff2", but in an ensemble model, they have the best performance for class 2 observations based on the proposed framework of this paper. The only models with their "Diff2" being more than 300 are GT, and CT which are the most desirable models. BT, BL, GL, BG, and CG are the second best group of models with their "Diff2" being between 200 and 300. Also, the poorest and the second poorest models are B (Bernoulli), and T (Tree) respectively, but BT is one of the most powerful models, showing the great power of ensemble learning models again. In addition, BCGLT assembled all models with a positive "Diff2", but its "Diff2" is not big enough to compete with the other exemplary models shown in figure 2. Again, it means that mixing all models in an ensemble setting doesn't guarantee an acceptable outcome.

Figure 3, illustrates the comparison of the models' performances for class 3 based on our proposed method for this class which is based on the measure "Diff3". As explained earlier, the most suitable model for class 3 is the one with the highest negative "Diff3" value. However, it should be mentioned that the most critical classes to predict correctly are class 1, and class 2. Therefore, selecting the most appropriate model for class 3, 4, and 5 should not cost selecting the poor model for class 1, and 2. As illustrated in figure 3, the most suitable models for class 3 are mostly non-ensemble ones. There is only one ensemble model with negative "Diff3", and that is BCG which includes only Naïve models (Bernoulli + Complement + Gaussian).

Figures 4, and 5 illustrate the comparisons of the models' performances for classes 4, and 5 respectively. The models are compared based on F1 scores. F1 is a measure of whether a model is appropriate for classes 4, and 5 predictions. According to figures 4 the most suitable models for class 4 are T, GLT, and CT. The performance of the other models in this class is more or less the same. In fact, in most models, the F1 score is between 0.25, and 0.3 which means that choosing not the most suitable models based on class 4's performance will not make significant differences in the end. About class 5, as illustrated in figure 5, and shown in table 4, in most models (21 out of 31), the F1 score is between 0.4, and 0.5. As explained earlier, and shown in figure 1, and table 4, the best models for class 1, are BC, CT, GT, CG, BG, and CL while these models except the CL, do not have appropriate performances for class 5. Obtaining such a result is expected because in a highly imbalanced dataset like ours, the best

model for the minority class is often the worst or one of the worst models for the majority class. This is the case with the problem in this paper too. To address this problem, we use different combinations of models in an ensemble setting. We come up with a framework along with two new measures in hopes we can reach an acceptable solution.

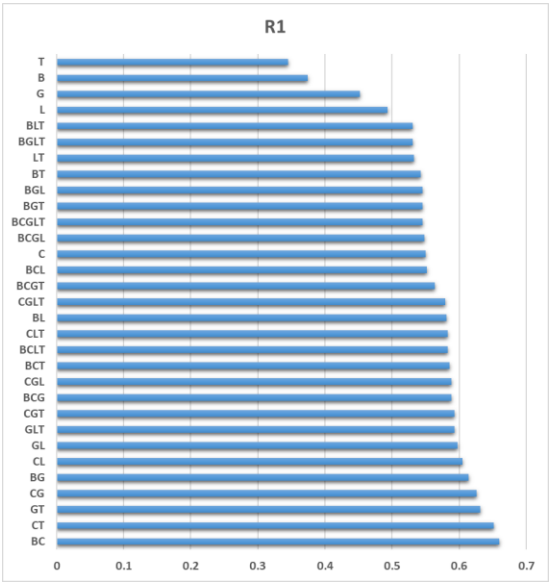

Figure 1. Visual comparison of the performance of different models for class 1

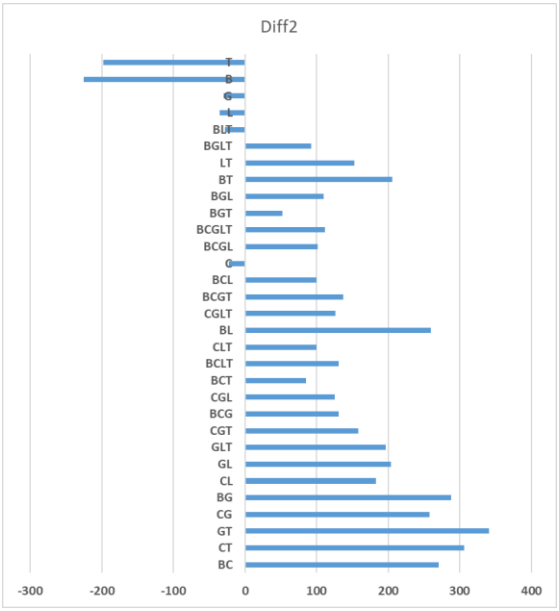

Figure 2. Visual comparison of the performance of different models for class 2

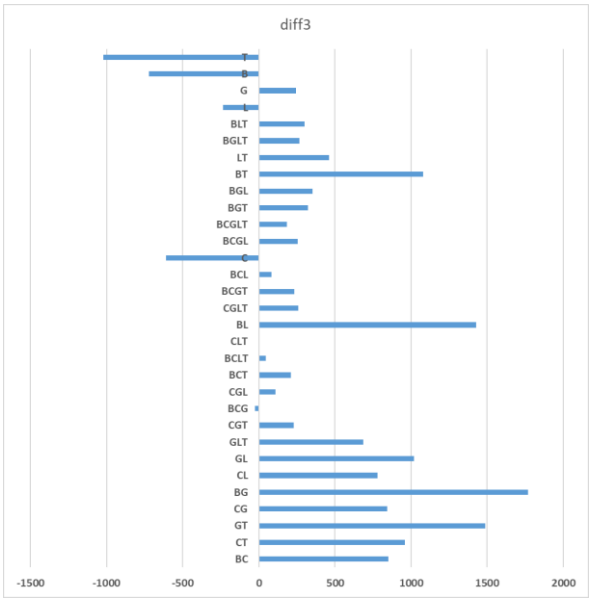

Figure 3. Visual comparison of the performance of different models for class 3

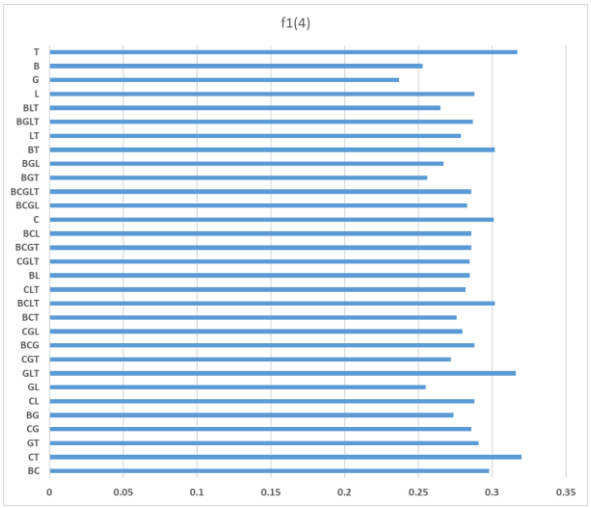

Figure 4. Visual comparison of the performance of different models for class 4

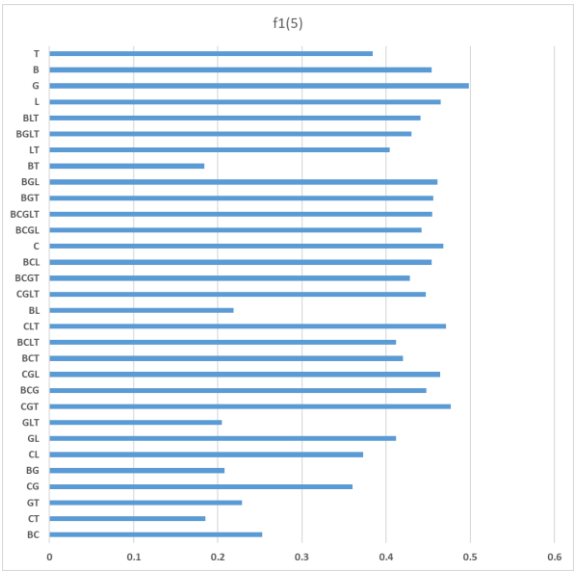

Figure 5. Visual comparison of the performance of different models for class 5

To provide a better understanding of our proposed method and also the performance of the models for each class in relation to other classes, we present table 5, and figures 6- 15. Table 5 ranks each model based on its performance in each class. For example, the first model in the table which is BC (Bernoulli + Complement) is the 1st best model for class 1, 4th best for class 2, 25th best for class 3, 7th best for class 4, and 25th best for class 5. Figures 6- 15 help the reader better understand the relationship between the ranks and each other.

Table 4. The evaluation results of the models

MN	Model	P1	P2	P3	P4	P5	R1	R2	R3	R4	R5	F1(1)	F1(2)	F1(3)	F1(4)	F1(5)	A
1	BC	0.013	0.018	0.122	0.258	0.678	0.66	0.166	0.037	0.353	0.155	0.025	0.031	0.057	0.298	0.253	0.191
2	CT	0.013	0.029	0.122	0.247	0.710	0.652	0.130	0.021	0.453	0.106	0.025	0.047	0.035	0.320	0.185	0.180
3	GT	0.013	0.028	0.114	0.243	0.718	0.631	0.150	0.071	0.364	0.136	0.025	0.047	0.087	0.291	0.229	0.185
4	CG	0.014	0.025	0.118	0.257	0.708	0.626	0.125	0.069	0.323	0.241	0.027	0.042	0.087	0.286	0.360	0.243
5	BG	0.013	0.020	0.118	0.252	0.688	0.614	0.192	0.086	0.300	0.123	0.026	0.037	0.100	0.274	0.208	0.164
6	CL	0.014	0.030	0.123	0.261	0.704	0.605	0.156	0.089	0.320	0.253	0.028	0.050	0.104	0.288	0.373	0.253
7	GL	0.015	0.028	0.111	0.258	0.701	0.598	0.163	0.100	0.252	0.291	0.029	0.048	0.105	0.255	0.412	0.263
8	CGT	0.015	0.031	0.120	0.262	0.698	0.593	0.133	0.039	0.282	0.363	0.029	0.050	0.059	0.272	0.477	0.309
9	GLT	0.014	0.038	0.104	0.244	0.723	0.593	0.163	0.071	0.449	0.119	0.027	0.062	0.084	0.316	0.205	0.193
10	BCG	0.014	0.032	0.123	0.265	0.694	0.588	0.129	0.043	0.314	0.331	0.028	0.051	0.064	0.288	0.448	0.296
11	CGL	0.015	0.032	0.119	0.264	0.701	0.588	0.135	0.059	0.298	0.347	0.030	0.051	0.079	0.280	0.464	0.305
12	BCT	0.014	0.025	0.148	0.258	0.682	0.586	0.150	0.037	0.297	0.304	0.028	0.043	0.059	0.276	0.420	0.274
13	BCLT	0.014	0.039	0.124	0.260	0.695	0.583	0.150	0.024	0.360	0.293	0.027	0.062	0.041	0.302	0.412	0.281
14	CLT	0.015	0.038	0.116	0.262	0.699	0.583	0.158	0.045	0.306	0.355	0.030	0.061	0.065	0.282	0.471	0.311
15	BL	0.011	0.030	0.120	0.256	0.694	0.581	0.195	0.092	0.320	0.130	0.022	0.051	0.104	0.285	0.219	0.174
16	CGLT	0.015	0.034	0.133	0.261	0.702	0.579	0.160	0.055	0.315	0.328	0.029	0.056	0.078	0.285	0.447	0.296
17	BCGT	0.013	0.032	0.143	0.260	0.687	0.564	0.143	0.020	0.316	0.311	0.025	0.053	0.035	0.286	0.428	0.282
18	BCL	0.014	0.035	0.128	0.266	0.692	0.552	0.150	0.040	0.311	0.337	0.027	0.056	0.060	0.286	0.454	0.300
19	C	0.015	0.029	0.128	0.263	0.686	0.550	0.114	0.013	0.352	0.355	0.029	0.046	0.024	0.301	0.468	0.317
20	BCGL	0.014	0.032	0.132	0.264	0.694	0.548	0.152	0.062	0.304	0.324	0.027	0.053	0.084	0.283	0.442	0.292
21	BCGLT	0.014	0.036	0.130	0.264	0.699	0.545	0.167	0.051	0.311	0.337	0.027	0.060	0.074	0.286	0.455	0.301
22	BGL	0.015	0.030	0.116	0.264	0.699	0.545	0.170	0.070	0.271	0.344	0.029	0.052	0.087	0.267	0.461	0.298
23	BGT	0.016	0.026	0.141	0.251	0.685	0.545	0.192	0.041	0.261	0.342	0.030	0.046	0.064	0.256	0.456	0.293
24	BT	0.010	0.033	0.108	0.251	0.679	0.543	0.202	0.044	0.379	0.107	0.020	0.057	0.063	0.302	0.184	0.167
25	LT	0.011	0.037	0.117	0.265	0.682	0.533	0.166	0.024	0.294	0.287	0.021	0.060	0.040	0.279	0.404	0.262
26	BGLT	0.013	0.041	0.130	0.259	0.698	0.531	0.171	0.059	0.321	0.310	0.025	0.066	0.081	0.287	0.430	0.286
27	BLT	0.017	0.028	0.118	0.254	0.687	0.531	0.237	0.049	0.278	0.325	0.034	0.050	0.069	0.265	0.441	0.287
28	L	0.017	0.039	0.117	0.261	0.701	0.493	0.180	0.101	0.320	0.348	0.032	0.064	0.108	0.288	0.465	0.315
29	G	0.015	0.030	0.122	0.263	0.693	0.452	0.187	0.111	0.216	0.389	0.028	0.052	0.116	0.237	0.498	0.318
30	B	0.010	0.033	0.126	0.261	0.650	0.374	0.163	0.068	0.249	0.349	0.019	0.054	0.089	0.253	0.454	0.294

MN	Model	P1	P2	P3	P4	P5	R1	R2	R3	R4	R5	F1(1)	F1(2)	F1(3)	F1(4)	F1(5)	A
31	T	0.012	0.054	0.138	0.239	0.687	0.345	0.200	0.017	0.468	0.266	0.023	0.085	0.030	0.317	0.384	0.286

P(i) = Precision for class i/ R(i) = Recall for class i/ F1(i) = F1 for class i/ A= Accuracy

Table 5. Models' ranking based on the framework proposed in this research

Models	R1	Diff2	diff3	f1(4)	f1(5)	Rank1	Rank2	Rank3	Rank4	Rank5
BC	0.66	271	849	0.298	0.253	1	4	25	7	25
CT	0.652	306	961	0.320	0.185	2	2	26	1	30
GT	0.631	341	1486	0.291	0.229	3	1	30	8	26
CG	0.626	258	845	0.286	0.360	4	6	24	16	24
BG	0.614	288	1767	0.274	0.208	5	3	31	24	28
CL	0.605	183	779	0.288	0.373	6	10	23	10	23
GL	0.598	204	1021	0.255	0.412	7	8	27	29	20
GLT	0.593	197	685	0.316	0.205	8	9	22	3	29
CGT	0.593	158	228	0.272	0.477	9	11	12	25	2
BCG	0.588	131	-27	0.288	0.448	10	15	5	11	12
CGL	0.588	125	110	0.28	0.464	11	17	9	21	6
BCT	0.586	85	209	0.276	0.42	12	24	11	23	18
BCLT	0.583	131	47	0.302	0.412	13	14	7	4	19
CLT	0.583	100	-1	0.282	0.471	14	22	6	20	3
BL	0.581	260	1429	0.285	0.219	15	5	29	18	27
CGLT	0.579	126	261	0.285	0.447	16	16	16	17	13
BCGT	0.564	137	233	0.286	0.428	17	13	13	14	17
BCL	0.552	100	85	0.286	0.454	18	21	8	15	11
C	0.55	-23	-611	0.301	0.468	19	26	3	6	4
BCGL	0.548	102	257	0.283	0.442	20	20	15	19	14
BCGLT	0.545	112	185	0.286	0.455	21	18	10	13	9
BGT	0.545	52	321	0.256	0.456	22	25	19	28	8
BGL	0.545	110	352	0.267	0.461	23	19	20	26	7
BT	0.543	206	1080	0.302	0.184	24	7	28	5	31
LT	0.533	153	463	0.279	0.404	25	12	21	22	21
BGLT	0.531	92	266	0.287	0.430	26	23	17	12	16
BLT	0.531	-28	299	0.265	0.441	27	27	18	27	15
L	0.493	-35	-234	0.288	0.465	28	29	4	9	5
G	0.452	-30	243	0.237	0.498	29	28	14	31	1
B	0.374	-225	-720	0.253	0.454	30	31	2	30	10
T	0.345	-198	-1023	0.317	0.384	31	30	1	2	22

The figures 6 to 15 illustrate the relationship between the Ranks described in table 5. A trend line has also been drawn for all plots. It can be seen from figure 6 that Rank1 and Rank2 have a rather strong direct relationship. Accordingly, the most appropriate model for class 1 roughly favors the most appropriate model for class 2. Figure 7 illustrates that Rank1 and Rank3 have an inverse relationship, but not a significant one. Rank1 and 4, Rank2 and 4, Rank3 and 4 do not appear to have any relationship. It is possible to establish a reverse relationship between 1 and 3, 1 and 5, and 4 and 5, but it is relatively weak. Rank2 and 3 as well as Rank2 and 5 have a reversed and relatively strong relationship. Furthermore, there is a relatively strong direct relationship between Rank 3 and 5.

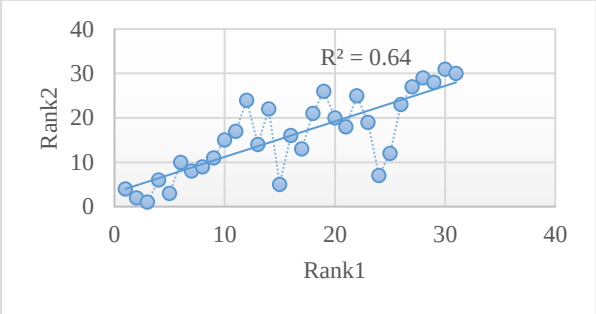

Figure 6. The relation between Rank 1 and Rank 2

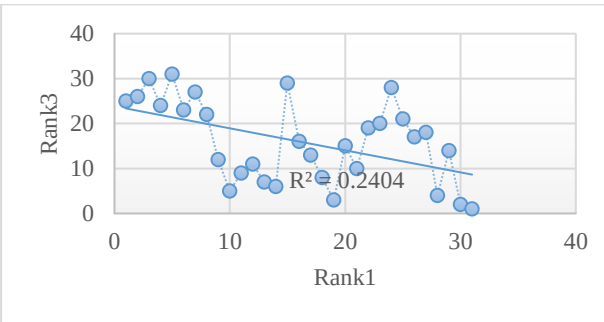

Figure 7. The relation between Rank 1 and Rank 3

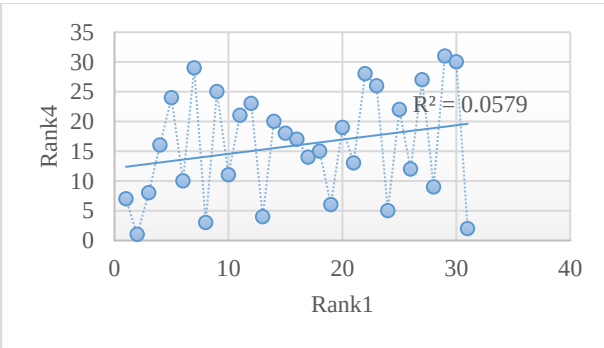

Figure 8. The relation between Rank 1 and Rank 4

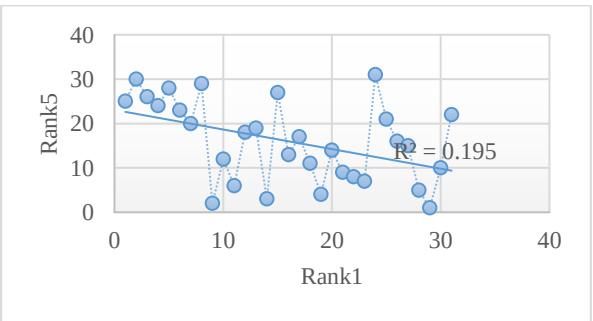

Figure 9. The relation between Rank 1 and Rank 5

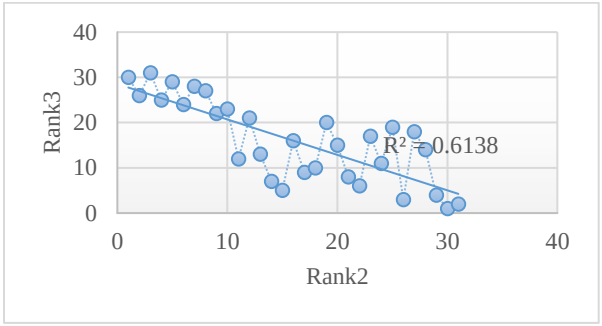

Figure 10. The relation between Rank 2 and Rank 3

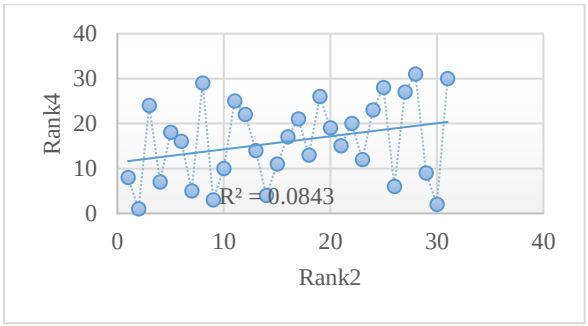

Figure 11. The relation between Rank 2 and Rank 4

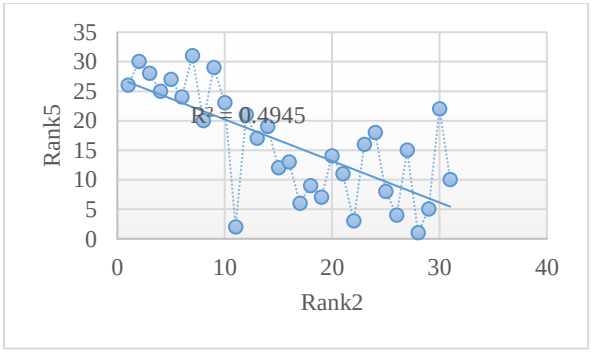

Figure 12. The relation between Rank 2 and Rank 5

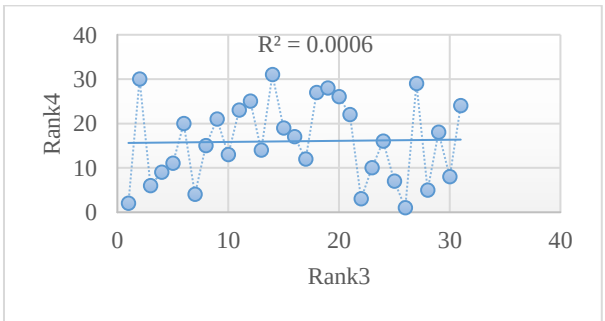

Figure 13. The relation between Rank 3 and Rank 4

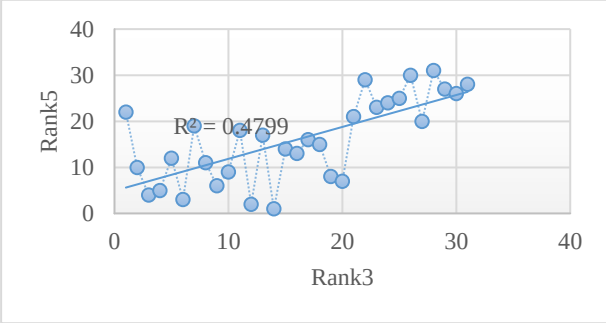

Figure 14. The relation between Rank 3 and Rank 5

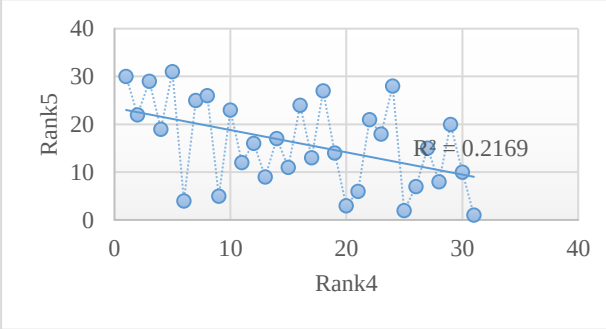

Figure 15. The relation between Rank 4 and Rank 5

5. Conclusion and Future Work

This paper examines the application of voting ensemble machine learning models. Specifically, a multiclass classification problem with five classes was examined in this research. In order to construct non-ensemble models, five common models were used. Then, using different combinations of these five models, 26 ensemble learning models were constructed. SMOTE sampling was used to resolve the imbalanced classification issue. In addition, an analysis of feature selection using RF model was performed in order to select the most relevant features. The most important class to predict in this paper was the first (fatal crashes), which accounted for 0.73% of all observations. The least important class was the fifth (PDO crashes), which accounted for 64.8% of all observations. As a general rule, the larger the number of a specific class, the larger is its share in the dataset, and the less important the class in terms of having accurate predictions. We aim to provide a solution to the problem of choosing a suitable model for a multiclass classification problem with highly imbalanced classes that exhibit the characteristics outlined above. To achieve these objectives, two measures have been introduced in this paper: Diff2 and Diff3. "Diff2" refers to the difference between class 2 observations that were incorrectly classified as class 1 and class 2 observations that were incorrectly classified as class 3, 4, or 5. Furthermore, "Diff3" represents the difference between the number of class 3 observations that are misclassified as class 1 or 2 and the number of class 3 observations that are misclassified as class 4 or 5. A proposal for selecting the best model is based upon the following criteria: for class 1, the model with the largest R1, for class 2, the model with the largest "Diff2", for class 3, the smallest "Diff3" (negative values are preferable), and for classes 4 and 5, the model with the highest "F1-score" for each of these classes. Furthermore, for some cases (for class 1), the decision tree model is the worst, but when it is used with a Naïve Bayes model such as Complement in an ensemble learning model, it is one of the best models. The results of this study demonstrate the effectiveness of ensemble machine learning models in predicting the minority class in a multiclass classification problem that is highly imbalanced. The decision tree alone resulted in a weaker prediction of the class with the smallest number of members in this study. Complement gave the best results in non-ensemble models, and when both of these models are combined into an ensemble learning model, the best model is obtained, which means that even the weakest model can contribute to improving the performance of the most powerful model when used in an ensemble setting. The 31 models were ranked based on the criteria set forth in this paper for each class. As a result, five ranking series were derived. By examining these rankings, it may be possible to determine, for example, whether the 3rd best model for class 1 corresponds to the best model

for class 2, class 3, etc. Accordingly, a "Rank" was determined for each model in each class. Each model was assigned five different ranks (Rank1 for class1, Rank2 for class2, etc.) and then the relationships between the ranks were evaluated. A relatively strong direct relationship was found between Rank1 and Rank2, Rank3 and 5. In addition, there is a reverse and relatively strong correlation between Rankings 2 and 3, as well as Rankings 2 and 5.

Using the framework and logic of this paper, practitioners will be able to tackle the problem of dealing with imbalanced multiclass classification problems, not only in the field of traffic safety but also in other areas such as medical and health problems, and the applications of machine learning in these fields as well. Traffic safety practitioners may find it useful to utilize the proposed method in this paper in order to develop appropriate countermeasures that are economically and technically justified. Through the use of our proposed method, one will be able to achieve a balance between predicting minorities (fatal and severe injuries) and predicting majority classes that do not result in fatalities or severe injuries. Only focusing on the minority classes would result in countermeasures that would not be economically viable, and ignoring the minorities would result in countermeasures that would not decrease the risk of death or severe damage. Also, according to our findings, to achieve acceptable performance in imbalanced multiclass classification problems, using Naïve Bayes models in combination with more practical and common models such as Logit, and Decision Tree can enhance the power of prediction especially when predicting the minority classes are of more concern. It is important that practitioners take into consideration this point as well.

In the future, one can focus on creating "super-ensemble" models. As far as we are aware, the term "super-ensemble" has not been used before in previous studies. Super-ensemble models can be defined as ensemble machine learning models that are composed of all possible non-ensemble models that can be combined in a variety of ways. In our biggest ensemble model, we employ 5 models. However, in a super ensemble model, we may employ as many models as we desire in different combinations and evaluate their performance using the proposed method of this paper in an imbalanced multiclass classification problem. Further, another possibility for future studies would be to extend our method to classification problems with more than five classes. The same problem as our paper may also be investigated using other non-ensemble models, such as Supplementary Vector Machines (SVM), or even Decision Random Forests (RF), as well as to determine whether the difference in tree numbers with different models in an ensemble setting can be significant.

6. Limitations

In this study, there are two main limitations. First of all, the original dataset contains a greater number of variables than what is used in this paper. Due to limited access to the data, we were not able to include all variables in our paper. In spite of this, it does not mean that having a complete dataset would have significantly affected our results. There is no doubt that our precision, or accuracy, would have been better than it is now, but since the focus of this paper is on methodological issues, proposing new measures, and approaches, and the dataset is only being used to explain our method, using this dataset would not be detrimental to our research.

A second limitation is that to summarize the paper and present our work in one paper, we did not use some other common and well-known machine learning models, such as SVM. In spite of the fact that the inclusion of other models in our ensemble models has nothing to do with the main purpose of the paper (which is to develop a new framework to solve imbalanced multiclass classification models), it can

nonetheless be considered a limitation in the paper.

7. Appendix

Table A. The confusion matrices of the models

Model	Class	1	2	3	4	5
L	1	207	81	27	41	64
	2	492	224	102	186	239
	3	1610	808	570	1322	1330
	4	2412	1272	1177	4124	3888
	5	7768	3322	3011	10120	12927
G	1	190	87	33	19	91
	2	490	233	114	117	289
	3	1626	1003	625	815	1571
	4	2646	1807	1202	2776	4442
	5	8061	4666	3140	6848	14433
C	1	231	54	2	60	73
	2	539	142	8	241	313
	3	1845	632	75	1545	1543
	4	3028	1053	164	4530	4098
	5	9778	2985	336	10862	13187
B	1	157	70	40	47	106
	2	408	202	91	167	375
	3	1546	721	386	1054	1933
	4	3157	1352	621	3170	4573
	5	10719	3826	1921	7723	12959
DT	1	145	95	10	100	70
	2	398	249	24	365	207
	3	1536	725	95	2164	1120
	4	2575	988	176	6030	3104
	5	7789	2572	385	16519	9883
BCGLT	1	229	75	13	40	63
	2	574	207	54	175	233
	3	1971	797	289	1268	1315
	4	3277	1265	541	3999	3791
	5	10306	3329	1326	9653	12534
BGLT	1	223	78	18	43	58
	2	561	213	60	197	212
	3	2044	744	330	1316	1206
	4	3435	1212	591	4133	3502
	5	10832	2978	1545	10272	11521
GLT	1	249	73	19	63	16
	2	619	202	65	299	58
	3	2238	724	401	1892	385
	4	3687	1212	954	5781	1239
	5	11499	3080	2430	15706	4433
BCLT	1	245	59	8	57	51
	2	594	186	28	223	212
	3	2097	678	137	1519	1209
	4	3581	1085	258	4635	3314
	5	11483	2709	671	11402	10883
CGLT	1	243	63	13	34	67
	2	585	199	54	180	225
	3	2011	783	313	1270	1263
	4	3259	1349	587	4052	3626
	5	10102	3491	1391	9978	12186
BCGT	1	237	65	7	48	63
	2	601	178	23	191	250
	3	2145	736	111	1289	1359
	4	3764	1245	185	4074	3605
	5	11843	3261	449	10039	11556
LT	1	224	72	6	52	66
	2	595	206	23	176	243
	3	2211	773	135	1214	1307
	4	4219	1240	276	3789	3349
	5	13384	3298	710	9093	10663

Model	Class	1	2	3	4	5
CT	1	274	59	6	64	17
	2	694	161	20	298	70
	3	2529	713	117	1874	407
	4	4378	1274	269	5834	1118
	5	13739	3383	548	15532	3946
BT	1	228	84	16	61	31
	2	599	251	46	259	88
	3	2352	883	250	1641	514
	4	4505	1717	536	4875	1240
	5	14532	4630	1458	12571	3957
GT	1	265	66	20	48	21
	2	699	186	56	230	72
	3	2533	831	398	1419	459
	4	4296	1588	865	4689	1435
	5	13008	4007	2140	12932	5061
BCGL	1	230	70	14	41	65
	2	578	189	61	162	253
	3	1991	784	347	1192	1326
	4	3354	1312	635	3914	3658
	5	10532	3477	1573	9527	12039
BCL	1	232	70	10	47	61
	2	578	187	37	172	269
	3	1995	756	223	1253	1413
	4	3412	1195	435	3999	3832
	5	10789	3194	1041	9587	12537
BL	1	244	89	21	45	21
	2	630	243	84	191	95
	3	2335	940	519	1328	518
	4	4321	1905	1035	4115	1497
	5	14215	5042	2683	10371	4837
GL	1	251	67	27	18	57
	2	622	203	95	125	198
	3	2119	929	565	939	1088
	4	3418	1724	1217	3241	3273
	5	10575	4358	3174	8215	10826
CL	1	254	62	24	36	44
	2	616	194	85	177	171
	3	2124	834	503	1262	917
	4	3511	1437	989	4118	2818
	5	11127	3953	2478	10178	9412
BG	1	258	70	30	34	28
	2	646	239	74	185	99
	3	2252	1209	485	1206	488
	4	3940	2651	974	3857	1451
	5	12445	7617	2535	9994	4557
CG	1	263	44	28	37	48
	2	673	155	51	200	164
	3	2268	780	389	1292	911
	4	3895	1463	795	4159	2561
	5	11980	3668	2046	10504	8950
BC	1	277	48	13	54	28
	2	685	144	31	236	147
	3	2382	757	211	1574	716
	4	4343	1722	410	4550	1848
	5	13694	5361	1068	11254	5771
BCT	1	246	59	10	45	60
	2	571	186	39	195	252
	3	2011	810	207	1284	1328
	4	3529	1571	339	3819	3615
	5	10851	4754	800	9466	11277
CLT	1	245	52	17	37	69
	2	573	197	42	178	253

Model	Class	1	2	3	4	5
	3	1945	747	255	1324	1369
	4	3289	1141	527	3935	3981
	5	9979	3066	1357	9566	13180
CGT	**1**	249	43	18	36	74
	2	618	165	28	171	261
	3	2074	749	222	1173	1422
	4	3531	1207	442	3628	4065
	5	10546	3163	1134	8831	13474
BLT	**1**	223	84	13	36	64
	2	460	295	55	181	252
	3	1593	1238	277	1199	1333
	4	2580	2302	572	3574	3845
	5	7945	6656	1421	9068	12058
BCG	**1**	247	52	10	42	69
	2	607	160	43	180	253
	3	2035	649	245	1341	1370
	4	3513	1114	485	4042	3719
	5	10915	3101	1213	9640	12279
BGT	**1**	229	68	15	38	70
	2	528	239	48	167	261
	3	1802	1062	233	1122	1421
	4	3010	2005	402	3362	4094
	5	9106	5673	950	8704	12715
CGL	**1**	247	44	20	40	69
	2	600	168	50	167	258
	3	2017	691	334	1261	1337
	4	3290	1252	666	3839	3826
	5	10115	3175	1729	9247	12882
BGL	**1**	229	66	20	33	72
	2	571	211	56	150	255
	3	1927	872	394	1115	1332
	4	3130	1603	809	3490	3841
	5	9606	4190	2123	8452	12777

8. References

- Abdulazeez, M.U., Khan, W. and Abdullah, K.A., 2023. Predicting child occupant crash injury severity in the United Arab Emirates using machine learning models for imbalanced dataset. IATSS Research, 47(2), pp.134-159.

- Ahmed, S.S., Corman, F. and Anastasopoulos, P.C., 2023. Accounting for unobserved heterogeneity and spatial instability in the analysis of crash injury-severity at highway-rail grade crossings: A random parameter with heterogeneity in the means and variances approach. Analytic methods in accident research, 37, p.100250.

- Azhar, A., Ariff, N.M., Bakar, M.A.A. and Roslan, A., 2022. Classification of driver injury severity for accidents involving heavy vehicles with decision tree and random forest. Sustainability, 14(7), p.4101.

- Bokaba, T., Doorsamy, W. and Paul, B.S., 2022. Comparative study of machine learning classifiers for modelling road traffic accidents. Applied Sciences, 12(2), p.828.

- Chakraborty, M., Gates, T. and Sinha, S., 2021. Causal Analysis and Classification of Traffic Crash Injury Severity Using Machine Learning Algorithms. arXiv preprint arXiv:2112.03407.

- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.

- Chen, M.M. and Chen, M.C., 2020. Modeling road accident severity with comparisons of logistic regression, decision tree and random forest. Information, 11(5), p.270.

- Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

- Eluru, N., Bhat, C.R. and Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis & Prevention, 40(3), pp.1033-1054.

- Feknssa, N., Venkataraman, N., Shankar, V. and Ghebrab, T., 2023. Unobserved heterogeneity in ramp crashes due to alignment, interchange geometry and truck volume: Insights from a random parameter model. Analytic methods in accident research, 37, p.100254.

- Fiorentini, N. and Losa, M., 2020. Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures, 5(7), p.61.

- Gan, X. and Weng, J., 2020. Predicting Crash Injury Severity for the Highways Involving Traffic Hazards and Those Involving No Traffic Hazards. In CICTP 2020 (pp. 4195-4206).

- Goswamy, A., Abdel-Aty, M. and Islam, Z., 2023. Factors affecting injury severity at pedestrian crossing locations with Rectangular RAPID Flashing Beacons (RRFB) using XGBoost and random parameters discrete outcome models. Accident Analysis & Prevention, 181, p.106937.

- Guo, M., Yuan, Z., Janson, B., Peng, Y., Yang, Y. and Wang, W., 2021. Older pedestrian traffic crashes severity analysis based on an emerging machine learning XGBoost. Sustainability, 13(2), p.926.

- Haeri, S., Mahpour, A., Vafaeinejad, A., 2024, Forecasting urban travel demand with geo-AI: a combination of GIS and machine learning

techniques utilizing Uber data in New York City, Environmental Earth Sciences, In press.

- Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

- Ho, T.K., 1995, August. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

- Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. Computational statistics & data analysis, 52(12), 5186-5201.

- Ijaz, M., Zahid, M. and Jamal, A., 2021. A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. Accident Analysis & Prevention, 154, p.106094.

- Islam, A.M., Shirazi, M. and Lord, D., 2023. Grouped Random Parameters Negative Binomial-Lindley for accounting unobserved heterogeneity in crash data with preponderant zero observations. Analytic methods in accident research, 37, p.100255.

- Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H.M., Almoshaogeh, M., Farooq, D. and Ahmad, M., 2021. Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. International journal of injury control and safety promotion, 28(4), pp.408-427.

- Jeong, H., Jang, Y., Bowman, P.J. and Masoud, N., 2018. Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. Accident Analysis & Prevention, 120, pp.250-261.

- Kabli, A., Bhowmik, T. and Eluru, N., 2023. Exploring the temporal variability of the factors affecting driver injury severity by body region employing a hybrid econometric approach. Analytic methods in accident research, 37, p.100246.

- Krishnaveni, S. and Hemalatha, M., 2011. A perspective analysis of traffic accident using data mining techniques. International Journal of Computer Applications, 23(7), pp.40-48.

- Laskaris, R., 2015. Artificial Intelligence: a modern approach.

- Lee, J., Yoon, T., Kwon, S. and Lee, J., 2019. Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. Applied Sciences, 10(1), p.129.

- Liu, D.X., 2022. A spatial data statistical model of urban road traffic accidents. Advances in transportation studies, 1.

- Ma, J., Ding, Y., Cheng, J.C., Tan, Y., Gan, V.J. and Zhang, J., 2019. Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: a city management perspective. IEEE Access, 7, pp.148059-148072.

- Mahpour, A., Farzin, I., Izadi, A.R. and Ashouri, S., 2023. Expanding the VBN theory on succeeding the transportation demand management policies. Transportation Research Interdisciplinary Perspectives, 21, p.100903.

- Mahpour, A., Forsi, H., Vafaeenejad, A. and Saffarzadeh, A., 2022. An improvement on the topological map matching algorithm at junctions: a heuristic approach. International journal of transportation engineering, 9(4), pp.749-761.

- Mahpour, A., Shafaati, M., & Mohammadian Amiri, A. (2021). The effective factors on the safety culture of HAZMAT drivers. AUT Journal of Civil Engineering, 5(1), 69-78.

- Mannering, F.L., Shankar, V. and Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research, 11, pp.1-16.

- Metsis, V., Androutsopoulos, I. and Paliouras, G., 2006, July. Spam filtering with naive bayes-which naive bayes?. In CEAS (Vol. 17, pp. 28-69).

- Miqdady, T., de Oña, R. and de Oña, J., 2023. In search of severity dimensions of traffic conflicts for different simulated mixed fleets involving connected and autonomous vehicles. Journal of Advanced Transportation, 2023.

- Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A. and Hadi, M., 2020. Factors affecting injury severity in vehicle-pedestrian crashes: A day-of-week analysis using random parameter ordered response models and Artificial Neural Networks. International journal of transportation science and technology, 9(2), pp.100-115.

- Mokoatle, M., Vukosi Marivate, D. and Michael Esiefarienrhe Bukohwo, P., 2019, June. Predicting road traffic accident severity using accident report data in South Africa. In Proceedings of the 20th annual international conference on digital government research (pp. 11-17).

- Mousa, S.R., Bakhit, P.R. and Ishak, S., 2019. An extreme gradient boosting method for identifying the factors contributing to crash/near-crash events: a naturalistic driving study. Canadian Journal of Civil Engineering, 46(8), pp.712-721.

- Murty, M.N. and Devi, V.S., 2011. Pattern recognition: An algorithmic approach. Springer Science & Business Media.

- Nujjetty, A.P., Mohamedshah, Y.M. and Council, F.M., 2014. Highway safety information system: Guidebook for data files California. Washington, DC: Federal Highway Administration.

- Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S. and Mohammadian, A.K., 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Analysis & Prevention, 136, p.105405.

- Pradhan, B., Ibrahim Sameen, M., Pradhan, B. and Ibrahim Sameen, M., 2020. Predicting injury severity of road traffic accidents using a hybrid extreme gradient boosting and deep neural network approach. Laser Scanning Systems in Highway and Safety Assessment: Analysis of Highway Geometry and Safety Using LiDAR, pp.119-127.

- Rennie, J.D., Shih, L., Teevan, J. and Karger, D.R., 2003. Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 616-623).

- Rezapour, M., Farid, A., Nazneen, S. and Ksaibati, K., 2021. Using machine leaning techniques for evaluation of motorcycle injury severity. IATSS research, 45(3), pp.277-285.

- Rokach, L., 2010. Ensemble-based classifiers. Artificial intelligence review, 33, pp.1-39.

- Ryu, J.W., Kantardzic, M. and Walgampaya, C., 2010. Ensemble classifier based on misclassified streaming data. In Proc. of the 10th IASTED int. Conf. on artificial

intelligence and applications, austria (pp. 347-354).

- Sahebi, S., Mirbaha, B., Mahpour, A., & Noruzoliaee, M. H. (2015). Predicting pedestrian accidents in rural roads using ordered logit model.

- Santos, K., Dias, J.P. and Amado, C., 2022. A literature review of machine learning algorithms for crash injury severity prediction. Journal of safety research, 80, pp.254-269.

- Schlögl, M., Stütz, R., Laaha, G. and Melcher, M., 2019. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. Accident Analysis & Prevention, 127, pp.134-149.

- Schütze, H., Manning, C.D. and Raghavan, P., 2008. Introduction to information retrieval (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

- Shafaati, M., & Boroujerdian, A. M. (2020). Investigating the influential factors in changing the likelihood of involving pedestrians in dangerous situations. AUT Journal of Civil Engineering, 4(3), 357-366.

- Shafaati, M., & Saffarzadeh, M. (2023). In light of the automated fare collection data, how did the travel patterns of transit riders in Tehran change following COVID-19?. International Journal of Transportation Engineering.

- Shafaati, M., & Saffarzadeh, M. (2024). Does Crowding Have a More Complicated Effect on Public Transport Users with Respect to Perceived Travel Time?. Transportation Research Record, 03611981241230297.

- Singh, G., Sachdeva, S.N. and Pal, M., 2018. Comparison of three parametric and machine learning approaches for modeling accident severity on non-urban sections of Indian highways. Advances in transportation studies, 45.

- Studer, M., Struffolino, E. and Fasang, A.E., 2018. Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure. Sociological Methodology, 48(1), pp.103-135.

- Tang, J., Liang, J., Han, C., Li, Z. and Huang, H., 2019. Crash injury severity analysis using a two-layer Stacking framework. Accident Analysis & Prevention, 122, pp.226-238.

- Tayarani Yousefabadi, A., Mahpour, A., Farzin, I., & Mohammadian Amiri, A. (2021). The Assessment of the Change in the Share of Public Transportation by Applying Transportation Demand Management Policies. AUT Journal of Civil Engineering, 5(2), 199-212.

- Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Saher, N. and Madni, H.A., 2020. Comparison analysis of tree based and ensembled regression algorithms for traffic accident severity prediction. arXiv preprint arXiv:2010.14921.

- Vajari, M.A., Aghabayk, K., Sadeghian, M. and Shiwakoti, N., 2020. A multinomial logit model of motorcycle crash severity at Australian intersections. Journal of safety research, 73, pp.17-24.

- Wahab, L. and Jiang, H., 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS one, 14(4), p.e0214966.

- Weiss, G.M., 2013. Foundations of imbalanced learning. Imbalanced Learning: Foundations, Algorithms, and Applications, pp.13-41.

- Yang, J., Han, S. and Chen, Y., 2023. Prediction of Traffic Accident Severity Based on Random Forest. Journal of Advanced Transportation, 2023.

- Zhang, H., 2004. The optimality of naive Bayes. Aa, 1(2), p.3.

- Zhang, Y., Li, H. and Ren, G., 2023. Analyzing the injury severity in single-bicycle crashes: an application of the ordered forest with some practical guidance. Accident Analysis & Prevention, 189, p.107126.