CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 684-697
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Ensemble Bagging in Binary Logistic Regression for Transportation Mode Selection

Nuzulul Laili Nabila, Sobri Abusini, and Umu Sa'adah*

Department of Mathematics, Faculty of Science and Mathematics, Universitas Brawijaya, Malang , Indonesia

Abstract

This study examines train versus bus transportation mode choice on the Malang-Blitar route using binary logistic regression combined with ensemble bagging. Data from 100 respondents were analyzed using 80% for training and 20% for testing with k-fold cross-validation. Variables included travel cost differences, time, safety, comfort, and ease of access. Bagging was selected over other ensemble methods due to its effectiveness in reducing variance and overfitting with small datasets. Results showed the standard logistic regression achieved 85% accuracy on test data, while ensemble bagging with 200 replications improved accuracy to 90.83% (confidence interval: 90.379%-91.187%). McNemar's test confirmed statistically significant improvement ($p < 0.01$). Under equivalent conditions, 20.6% of respondents preferred trains while 79.4% chose buses. Ease of access emerged as the primary decision factor, outweighing cost and time considerations. The optimal replication number was 200; exceeding 300 replications decreased model performance. This research contributes an optimized ensemble methodology for transportation mode prediction in developing countries, demonstrating that accessibility infrastructure significantly influences passenger preferences over traditional economic factors.

Keywords: Binary Logistic Regression; Bagging; Bus; Train; Transportation Mode Selection.

1 Introduction

Mode selection is a critical element of urban transportation planning that requires a thorough understanding of user behavior. In the Indonesian local context, particularly the Malang-Blitar route in East Java, understanding transportation mode preferences becomes crucial for developing effective transportation systems. As modern transportation systems grow more complex, there is a need for advanced methods to predict and analyze people's choices of transportation modes. The research problem identified is the lack of accurate prediction models for transportation mode selection in Indonesia, particularly those combining ensemble learning methods with traditional statistical models [1]. In this context, combining ensemble learning techniques with traditional statistical models, such as binary logistic regression, has proven to be a promising solution.

Bagging (Bootstrap Aggregating) is an ensemble learning technique that improves prediction accuracy by combining multiple models. The theoretical justification for using bagging in this context is itsability to reduce model variance and improve prediction stability, especially on small datasetscommonly used in transportation research in Indonesia[2]. When applied to binary

*Corresponding author. E-mail: u.saadah@ub.ac.id

Submitted: February 25, 2025 Reviewed: June 09, 2025 Accepted: July 03, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.32241

logistic regression in the context of transportation, this technique helps improve the performance of predicting transportation mode selection [3], [4]. Binary logistic regression has long been used as a statistical method for binary classification tasks, such as predicting the choice between public and private transportation [5], [6].

The objectives of this research are: (1) To develop an ensemble bagging model for predicting trains bus transportation mode selection; (2) To determine the optimal number of replications in ensemble bagging; (3) To identify the main factors affecting transportation mode selection on the Malang-Blitar route.

The application of ensemble learning in transportation, particularly the bagging method, enables the integration of predictions from multiple models, including logistic regression, to enhance the accuracy and reliability of transportation mode choice predictions [7]–[9]. This method is particularly effective for modeling transportation mode choices, as machine learning models are utilized to predict and analyze these choices to enhance urban planning [10], [11]. Using machine learning methods alongside bagging in binary logistic regression not only enhances the model's predictive ability but also offers valuable insights into the factors that affect transportation mode selection. This can lead to the development of improved transportation systems [12], [13].

Recent advancements in machine learning have shown promising results. However, challenges persist in optimizing the integration of ensemble learning methods, such as bagging, with binary logistic regression models. It is essential to conduct further research to develop models that effectively blend these approaches for predicting transportation mode choices. This paper will evaluate the performance of these models by analysing accuracy confidence intervals derived from 30 ensemble bagging processes across three replication counts: 40, 80, and 200. The dependent variable will be coded as 1 for train travel and 0 for bus travel. The analysis will incorporate independent variables, including differences in travel costs, differences in travel time, differences in travel costs from the point of origin to the station or terminal, differences in travel costs from the station or terminal to the destination, security level, comfort level, and ease of travel level.

2 Methods

This study will leverage primary data gathered through interviews with train and bus passengers who have firsthand experience traveling on the Malang-Blitar route. The research population consists of all passengers who have used both train and bus transportation on the Malang-Blitar route. The sampling technique used is purposive sampling (non-probabilistic) with criteria of respondents who have used both modes of transportation in the last 6 months. The sample size of 100 respondents was determined based on Chochran formula for descriptive research with a 95% confidence level and 10% margin of error. Their experience will provide a rich understanding of the passenger experience and highlight key factors that influence travel decisions.

The variables that will be used are essential and will include: the dependent variable Y, which clearly represents the mode of passenger transportation, and the independent variable X_i, identifying the influential factors behind this choice. The variables used are presented in Table 1 below.

Table 1: Dependent and Independent Variables

	Variable		Data Characteristics	
Dependent	Y	Transportation Mode	$Y =$	$\begin{cases} 0, & \text{Bus} \\ 1, & \text{Train} \end{cases}$
Independent	X_1	Difference in travel costs	$X_1 =$	$\begin{cases} \text{Positive}, & \text{higher train costs} \\ \text{Negative}, & \text{lower train costs} \end{cases}$
	X_2	Difference in travel time	$X_2 =$	$\begin{cases} \text{Positive}, & \text{longer train time} \\ \text{Negative}, & \text{shorter train time} \end{cases}$
	X_3	Difference in travel costs from origin to station/terminal	$X_3 =$	$\begin{cases} \text{Positive}, & \text{higher train costs} \\ \text{Negative}, & \text{lower train costs} \end{cases}$

Continued on next page

Table 1 – continued from previous page

Variable		Data Characteristics	
X_4	Difference in travel costs from station/terminal to destination	$X_4 =$	Positive, higher train costs Negative, lower train costs
X_5	Comfort level	$X_5 =$	1, Train is more comfortable 2, Same 3, Bus is more comfortable
X_6	Security level	$X_6 =$	1, Train is safer 2, Same 3, Bus is safer
X_7	Ease of travel	$X_7 =$	1, Train is easier 2, Same 3, Bus is easier

The evaluation of security is based on how safe respondents feel while using transportation services. For the comfort variable, several key aspects are considered, including ergonomic seating, cleanliness of facilities, and the overall atmosphere of the mode of transportation being used. The convenience variable focuses on how accessible transportation services are to passengers. This includes ease of access to the station or terminal, the efficiency of the ticket purchasing system, and the availability of flexible departure schedules that meet user needs.

This methodology section outlines the comprehensive approach employed in this study, beginning with the fundamental statistical framework and progressing through advanced ensemble techniques. The following subsections detail each component of our analytical framework, starting with the theoretical foundation of binary logistic regression.

2.1 Binary Logistic Regression

Binary logistic regression describes the relationship between several predictor variables X_1, X_2, ..., X_k and a binary response variable Y. A binary response variable means that Y can have a value of 1 when a certain characteristic is present and a value of 0 when that characteristic is absent [14]. Logistic regression model equation is:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}. \tag{1}$$

with logit function $g(x)$:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}. \tag{2}$$

Equation (1) and (2) can be simplified to:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}}. \tag{3}$$

2.2 Parameter Estimation

Maximum Likelihood Estimation (MLE) is a statistical technique employed to estimate unknown parameters by maximizing a specific function. For example, consider random variables $x_1, x_2, ..., x_n$ that have a probability density function (pdf) $f(x_i;\theta)$, which includes a parameter θ. This relationship allows us to express the pdf as a likelihood function, yielding precise estimates that enhance our comprehension of the data [15]:

$$L(\beta) = \prod_{i=1}^{n} f(x_i;\beta). \tag{4}$$

From equation (4) we can write the log likelihood equation for binary logistic regression as:

$$\ell(\beta) = ln\ L(\beta) = \sum_{i=1}^{n} y_i \left(\sum_{j=0}^{k} \beta_j x_{ij} \right) - \sum_{i=1}^{n} ln \left(1 + exp \left(\sum_{j=0}^{k} \beta_j x_{ij} \right) \right). \tag{5}$$

Equation (5) can be differentiated against $\beta = \beta_0, \beta_1, \beta_2, ..., \beta_k$ and can be differentiated against. so that it is obtained:

$$\sum_{i=1}^{n} y_i x_{ia} - \sum_{i=1}^{n} x_{ia} \pi(x_i) = 0, \quad a = 0, 1, 2, 3, ..., k. \tag{6}$$

2.3 Hypotesis Testing

The conducted hypothesis test is identical to the hypothesis test used in standard logistic regression. Testing is performed both simultaneously and partially. The simultaneous test employs the G-test statistic, while the partial test utilizes the Wald test.

2.3.1 G-test statistics (Simultaneous test)

The Likelihood Ratio test assesses the significance of parameters collectively [16]. The hypotheses for the Likelihood Ratio test are as follows:

H_0 : $\beta_1 = \beta_2 = ... = \beta_k$, This indicates that there is no influence between the independent variable and the dependent variable.

H_1 : at least one $\beta_j \neq 0,\ j = 0, 1, 2, ..., k$, This indicates that at least one independent variable affects the dependent variable.

Gtest formula:

$$G^2 = -2ln \left(\frac{\text{log likelihood without independent variable}}{\text{log likelihood with independent variable}} \right) \tag{7}$$

At the designated significance level α, the null hypothesis H_0 will be rejected if the G^2 value surpasses the threshold of $G^2 > \chi^2_{(0.05,2)}$ or if the p-value is below the alpha level. This conclusion indicates that the predictor variables, both individually and collectively, significantly influence the response variable.

2.3.2 Wald test (Partial test)

The Wald test evaluates the significance of parameter coefficients in a model. The hypothesis tested by the Wald test is as follows [17]:

H_0 : $\beta_j = 0$, This indicates the lack of impact of the j^{th} independent variable on the dependent variable.

H_1 : $\beta_j \neq 0,\ j = 0, 1, 2, ..., k$, This indicates that the j^{th} independent variable affects the dependent variable.

Wald test formula:

$$W_j = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right)^2 \tag{8}$$

H_0 will be rejected if the value is greater than $\chi^2_{(0.05,2)}$ or if the p-value is less than α, concluding that β_j is significant; in other words, variable X partially influences the response variable.

From the parameter estimates of the logistic regression model that have been obtained, a Goodness of Fit test is then carried out. The Goodness of Fit test is used to measure how well the model can describe the response variable. The Goodness of Fit test is a test carried out to determine whether there is a difference between predictions and observation results (the model is appropriate or not) [18].

2.4 Model Validation

The model validation procedure uses k-fold cross-validation techniques with k=5 to ensure generalization of results [19]. Data is divided into 80% for training and 20% for testing. The bagging process is repeated 30 times for each number of replications (40, 80, 200, 300, 400, 500) to obtain robust confidence intervals. Evaluation metrics are calculated using the test set, not the training set.

2.5 Bagging

Ensemble methods capitalize on the strengths of multiple weak models by combining their outputs to achieve significantly improved and more robust results [20]. Ensemble techniques offer the power to achieve remarkably accurate predictions, making them a highly effective choice for improving forecast reliability [21]. The concept of an ensemble involves integrating multiple models that collectively tackle the same problem. This approach aims to enhance accuracy and improve overall performance in predictive outcomes.

Bagging, short for Bootstrap Aggregating, is an algorithm known for its excellent performance and ease of implementation. It involves creating multiple "n-bags" of data, referred to as bootstrap samples. These bags are formed by randomly drawing from the original training dataset, which has a total size of n. In detail, M bootstrap samples, $T_1, T_2, ..., T_M$, are generated from the original training set T. Each bootstrap sample T_i, which is also of size n, is then used to training data. To make predictions on new observations, the model takes an average of parameter estimations of each bootstrap samples. This process helps improve predictive accuracy by aggregating the results from multiple classifiers [22]. The process of bagging is depicted in the Figure 1.

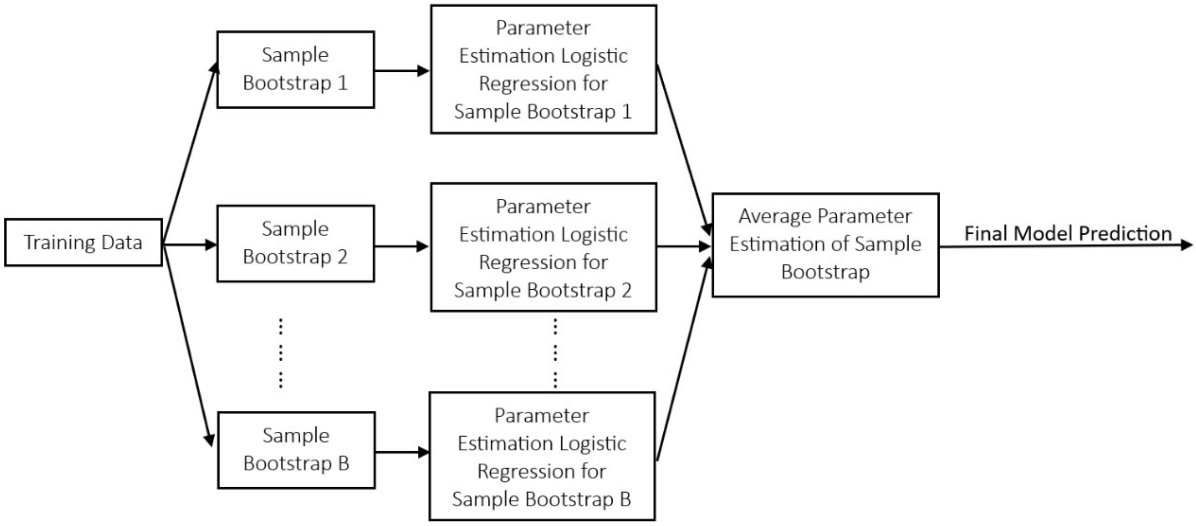

Figure 1: Bagging Logistic Regression Process

Bagging, or Bootstrap Aggregating, is an ensemble learning technique used to improve the stability and accuracy of machine learning algorithms, including logistic regression. Here is a step-by-step process of how bagging is applied to logistic regression:

1. Data Sampling
   Bagging begins with creating multiple subsets of the original dataset through random sampling with replacement. This means some data points may appear multiple times in a subset, while others may not appear at all [23], [24].
2. Model Training
   Each subset is used to train a separate logistic regression model. These models are

considered "weak learners" because they are trained on different samples of the data, which introduces variability [25].

3. Model Aggregation
   The predictions from all the logistic regression models are aggregated to form a final prediction. This aggregation can be done by averaging the parameter estimations [24].
4. Performance Evaluation
   The ensemble model's performance is evaluated using metrics such as accuracy, recall, and specificity. Bagging often results in improved performance by reducing overfitting and increasing the robustness of the model [24], [25].
5. Comparison and Validation
   The performance of the bagging logistic regression model is compared with other models or methods, such as single logistic regression or other ensemble methods, to validate its effectiveness [25].

2.6 Classification Accurancy

The classification error rate, commonly referred to as the Apparent Error Rate (APER), is used to evaluate how effectively a classification procedure determines group membership. Conversely, it can also be expressed as the Correct Error Rate, which measures the level of classification accuracy. Table 2 shows the form of the confusion matrix.

Table 2: Confusion Matrix

Observation results	Estimate	
	y_1	y_2
y_1	n_{11}	n_{12}
y_2	n_{21}	n_{22}

$$APER(\%) = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} \tag{9}$$

To determine the accuracy value, subtract the Apparent Error Rate (APER) from 1. This will provide the error value based on the previously explained APER calculation [26].

3 Results and Discussion

The results presented in this section demonstrate the effectiveness of ensemble bagging in binary logistic regression for transportation mode selection. Our analysis proceeds systematically from exploratory data analysis to model development and validation. We begin by examining the fundamental characteristics of our dataset to establish a solid foundation for subsequent modeling and interpretation.

3.1 Descriptive Statistic

Descriptive statistics provide a detailed overview of the key characteristics of the data through measurable and organized summary statistics, allowing researchers to identify patterns, trends, and variations within the dataset.

Table 3 provides a thorough summary of statistics for four continuous data variables, offering a detailed overview of their distribution characteristics. The analysis reveals that all four variables exhibit a wide range of values, indicating significant variability. The first and second variables show a consistent negative trend, as both their mean and median values are negative. In contrast, the last two variables display a different pattern, with a median of zero but differing means,

Table 3: Summary Statistic of Continuous Data

Summary	Variable			
	X_1	X_2	X_3	X_4
Minimum	-28000	-100	-35000	-35000
1st Quartile	-23000	-60	-5000	-8000
Median	-20000	-40	0	0
Mean	-18990	-39.87	2760	-500
3rd Quartile	-15000	-30	9250	5000
Maximum	0	50	200000	60000

which suggests asymmetry in the data distribution. Additionally, the relatively large interquartile ranges (IQR) of the third and fourth variables further confirm the high variability within the middle segment of the data.

Table 4: Summary Statistic of Discrete Data

Summary	Variable		
	X_5	X_6	X_7
Train	79	78	55
Same	18	22	21
Bus	3	0	24

Table 4 shows how people prefer different modes of transportation based on comfort, safety, and ease of use. Most respondents prefer trains, though levels of preference vary. Trains rated high for comfort and safety, with 79 and 78. However, only 55 chose trains for ease of use, indicating some may find them less accessible. In the same conditions, the three variables showed relatively stable consistency, with frequencies ranging from 18 to 22. This suggests that many see little difference between trains and buses. For buses, only 3 respondents rated them for comfort, and none chose them for safety. However, 24 preferred buses for ease of use, suggesting they may be more accessible or have better service.

Among the 100 respondents who traveled by train and bus on the Malang-Blitar route, 80 preferred trains while 20 preferred buses. From the data, analysis of respondent characteristics shows gender distribution and variations in travel purposes that represent passenger mobility patterns.

Figure 2: Respondent Profile Based on Gender

According to Figure 2, among the 80 respondents who preferred trains, 27 were male and 53 were female. In contrast, out of the 20 respondents who favored buses, 7 were male and 13 were female. This data shows that female respondents outnumber male respondents among both train and bus users.

According to Figure 3, the study identifies three main categories of travel purposes among the 100 respondents. Out of the total, 54 respondent trips for educational reasons, 17 trip for

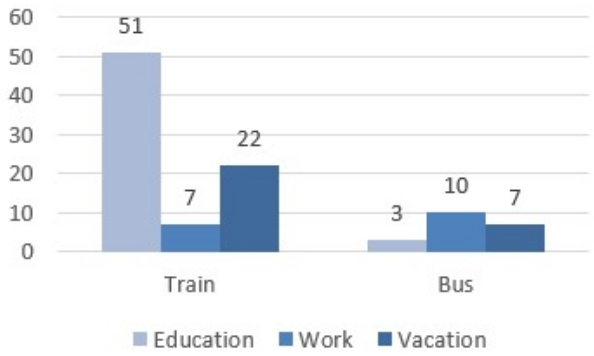

Figure 3: Respondent Profile Based on Travel Purpose

work, and 29 trip for vacation.

Among the 54 respondents who trip for educational purposes, 51 chose to travel by train, while 3 opted for buses. For vacation trips, 22 respondents preferred trains, and 7 chose buses. Overall, the majority of respondents selected trains over buses for both educational and vacation travel. However, when it came to work-related travel, the preference shifted. Out of the 17 respondents traveling for work, 10 chose buses while only 7 selected trains.

Based on the data in Figure 3, we can develop additional analysis by considering the age variable and how it may influence the choice of transport mode. If we integrate the age variable into the existing analysis, some interesting patterns may emerge. The younger age group (18-25 years) is most likely to dominate the educational trip category (54 respondents) with a strong preference for train (51 out of 54) for educational trips. Potential reasons for this preference include more affordable ticket prices. The older age group (26-45 years) may represent the majority of work trips (17 respondents). Work trips show a preference for buses (10 out of 17). Potential reasons include bus routes that may be more convenient to the office/business location, bus schedules that are more convenient to work hours or transportation incentives from the company. Meanwhile, vacation trip purposes cover all ages from children to the elderly (29 respondents) with a preference for train for leisure (22 out of 29). Potential reasons include higher comfort, faster travel time, and a more enjoyable travel experience.

3.2 Parameter Estimation

The next step involves estimating the parameters of the binary logistic regression model, which includes seven independent variables: $X_1, X_2, X_3, X_4, X_5, X_6$ dan X_7, along with the dependent variable Y.

Table 5: Parameter Estimation

Variable	**Parameter Estimation**	p-**value**
Constant	-11.850000	0.0307
X_1	-0.000200	0.1685
X_2	-0.156900	0.0140
X_3	-0.000055	0.0678
X_4	-0.000200	0.0194
X_{5k}	5.193000	0.0153
X_{5b}	-10.650000	0.9973
X_{6k}	5.950000	0.0188
X_{7k}	2.442000	0.1861
X_{7b}	-8.105000	0.0111

X_6, there is only X_{6k} and no X_{6b}, because out of 100 respondents, no one chose buses as safer than trains.

3.3 Hypothesis Testing

Based on Table 5, it is clear that the significant independent variables include X_2, X_4, X_{5k}, X_{6k} and X_{7b}, all of which have p-values less than $\alpha = 0.05$. Conversely, the variables X_1, X_3, X_{5b} and X_{7k} do not show significance. Since some variables are not significant, a re-estimation will be conducted after eliminating the variable with the highest p-value. This process will be repeated three times, as shown in Table 6 below.

Table 6: Parameter Re-estimation

	2nd iteration		3rd iteration		4th iteration	
Variable	**Estimation**	***p*-value**	**Estimation**	***p*-value**	**Estimation**	***p*-value**
Constant	-11.8700	0.023	-9.4950	0.038	-7.9137	0.035
X_1	**-0.0002**	**0.167**	-0.0002	**0.235**	**-0.0001**	**0.298**
X_2	-0.1571	0.014	-0.1359	0.010	-0.1156	0.008
X_3	**-0.00005**	**0.067**	**-0.00005**	**0.057**	–	–
X_4	-0.0002	0.019	-0.0002	0.017	-0.0002	0.013
X_{5k}	5.2010	0.015	5.6700	0.008	5.1275	0.007
X_{5b}	–	–	–	–	–	–
X_{6k}	5.9580	0.018	5.5290	0.014	4.7935	0.010
X_{7k}	**2.4440**	**0.186**	–	–	–	–
X_{7b}	-8.1180	0.011	-8.3750	0.005	-7.2363	0.003

In the 4^{th} iteration, the *p-value* for variable X_1 remains above 0.05. However, X_1 will still be included in the model because are retained in the model based on theoretical justification from transportation literature showing that travel cost differences remain an important factor in transportation mode selection[27]. Therefore, X_1 is regarded as significant, albeit with the least influence.

The statistical value of the G^2 test is 74.2632738, indicating that G^2 exceeds $G^2 > \chi^2_{(0.05,2)}$. Consequently, it can be concluded that the simultaneous testing of the transportation mode selection model, comparing trains and buses using binary logistic regression, is significant at the 95% confidence level. In other words, we reject H_0, which implies that at least one significant parameter exists, necessitating further partial testing. So the binary logistic regression model can be written:

$$\pi(x) = \frac{e^{-7.9137-0.0001X_1-0.1156X_2-0.0002X_4+5.1275X_{5k}+4.7938X_{6k}-7.2363X_{7b}}}{1+e^{-7.9137-0.0001X_1-0.1156X_2-0.0002X_4+5.1275X_{5k}+4.7938X_{6k}-7.2363X_{7b}}}$$

3.4 Model Accuracy Test

Model accuracy testing is performed using the test set (20% of total data = 20 observations) toensure objective evaluation. Table 7 shows the confusion matrix based on the test set data.

Table 7: Confusion Matrix

Observation results	Estimate	
	Bus	Train
Bus	3	1
Train	2	14

Based on Table 7, from 20 test set data, the model correctly predicts 17 observations (3 bus + 14train). The model accuracy on the test set is:

$$APER = \frac{1+2}{20} = \frac{3}{20} = 0.15 = 15\%$$

To determine the level of accuracy, you can use 1- APER. So:

$$model\ accuracy = 1 - 0.15 = 0.85 = 85\%$$

This means the model can predict with 85% accuracy on unseen data.

3.5 Bagging Binary Logistic Regression

The ensemble bagging implementation uses a systematic validation procedure with k-fold cross-validation. The dataset is divided into 80% (80 data) for training and 20% (20 data) for testing. The bagging process is repeated 30 times for each replication scenario to obtain robust confidence intervals.

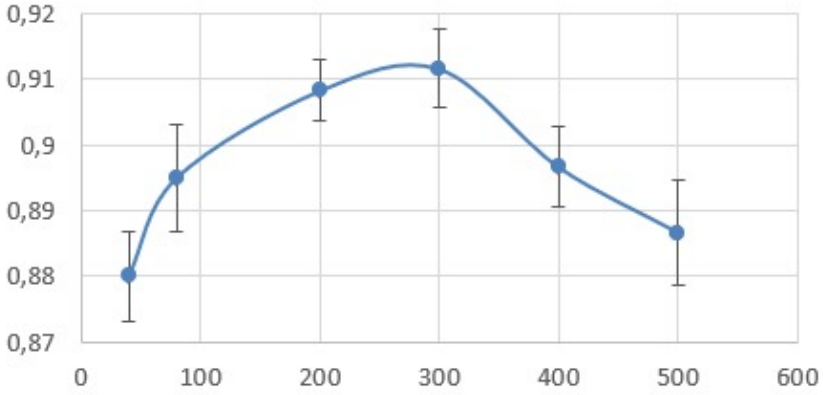

Figure 4: Confidence Interval Accuracy of Bagging Binary Logistic Regression Model with 80% Confidence Level

The experimental results at Figure 4 show an interesting pattern in the relationship between the number of replications and model accuracy. The standard binary logistic regression model produces a baseline accuracy of 85% on the test set. When ensemble bagging is applied with 40 replications,there is a significant increase in accuracy with a confidence interval of 87.325% - 88.674% (average 88%). Increasing the number of replications to 80 produces even better performance, with a confidence interval of 88.707% - 90.292% (average 89.5%).

Optimal performance was achieved with 200 replications, resulting in a confidence interval of 90.379% to 91.187%, with an average of 90.83%. McNemar's test was performed to compare the performance of the standard model with ensemble bagging, showing a statistically significant-difference ($\chi^2 = 8$, $p < 0.01$), confirming that the accuracy improvement is not due to chance [28].

Interestingly, increasing the number of replications beyond this point led to a decrease inperformance. For instance, with 300 replications, the accuracy remained high at an average of91.17%. However, when the replications were increased to 400 and 500, the accuracy consistentlydeclined, with averages dropping to 89.67% and 88.83%, respectively.

This pattern suggests there is an optimal "sweet spot" for the number of replications in ensemble bagging, which in this study is found to be 200 replications. With too few replications (such as 40), there may not be enough data to capture the variability effectively. Increasing the number of replications to 200 strikes an ideal balance between model variation and prediction stability. However, using too many replications (more than 300) can lead to excessive oversampling, which can actually reduce model performance.

Using 80% confidence intervals derived from 30 ensemble iterations for each replication scenario offers amore robust and reliable evaluation than relying on a single accuracy value. This approach allows us toconfidently conclude that ensemble bagging with 200 replications leads to a significant performanceimprovement compared to the standard binary logistic regression model, achieving an accuracy increase of 5.83 percentage points. Bagging with binary logistic regression model $(\pi(x)^{*(200)}$ can be written:

$$\pi(x)^{*(200)} = \frac{e^{-13.5057-0.0003X_1-0.1547X_2-0.0003X_4+7.9627X_{5k}+8.058X_{6k}-9.426X_{7b}}}{1+e^{-13.5057-0.0003X_1-0.1547X_2-0.0003X_4+7.9627X_{5k}+8.058X_{6k}-9.426X_{7b}}}$$

3.6 Model Intepretation

The analysis of transportation mode selection preferences under equivalent conditions can be conducted using a binary logistic regression bagging model graph. The equivalent conditions refer to scenarios where the differences in travel costs and travel times are both zero, and the levels of comfort, safety, and ease are the same for the two modes of transportation being compared. In this model, the variables $X_1, X_2, X_4, X_{5k}, X_{6k}$,and X_{7b} are key factors that influence passenger transportation mode selection. By applying the binary logistic regression bagging model ($\pi(x)$ with the difference in cost, represented as ΔC, can be calculated using the equation $-0.0001X_1 - 0.1156X_2 - 0.0002X_4 + 5.1275X_{5k} + 4.7938X_{6k} - 7.2363X_{7b}$, we can create a clear visualization of passenger decision-making patterns. The cost difference (ΔC) includes differences in travel costs, differences in travel time, differences in travel costs from the station or terminal to the destination, security level, comfort level, and ease of travel level.

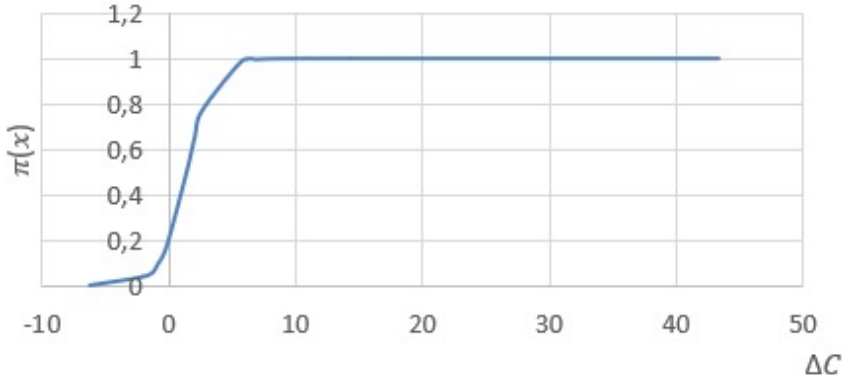

Figure 5: Bagging with Binary Logistic Regression Model Graph

Based on Figure 5, if the cost and time differences are zero and the comfort, safety, and ease levels of both transportation modes are same, then the value of ($\pi(x)$ is probability of passengers who would prefer the train when travel costs and travel times are both zero, and the levels of comfort, safety, and ease are the same. It can be concluded that 20.6% of people will choose the train, while the remaining 79.4% will choose the bus. The choice of transportation mode is primarily influenced by differences in cost and travel time. However, the variation in one-way ticket prices does not appear to be significant. Despite a notable price discrepancy, with bus tickets being considerably more expensive, there are still 20 individuals who prefer using the bus. Their preference largely stems from the convenience of accessing bus transportation. Although there are significant differences in cost and time, respondents indicated that convenience is the main factor in their decision to choose the bus. In terms of flexibility regarding travel times and ticket purchases, ease of access is the most influential aspect.

4 Conclusion

Significant variables in the choice of transportation mode include X_2, X_4, X_5, X_6 and X_7 (p-*value* < 0.05). Although variable X_1 (one-way ticket price difference) has a p-*value* > 0.05, this variable is retained in the model based on strong theoretical justification from transportation literatures how the importance of cost factors in mode selection.

The standard binary logistic regression model produces an accuracy of 85% on the test set, but with the application of ensemble bagging, the model accuracy increases significantly. The optimal number of replications was found at 200 replications with an average accuracy rate of 90.83% (confidence interval 90.379% - 91.187%), indicating an increase of 5.83 percentage points that is statistically significant based on McNemar's test ($p < 0.01$). The bagging method in ensemble learning is effective for improving model performance. However, utilizing excessive

replications (over 300) can actually decrease model accuracy, showing a trade-off between model complexity and performance.

performance. In the resulting model, under equivalent conditions (when all variables are set to zero), probability of selecting train transportation is 20.6%, while the probability of selecting the bus is 79.4%. This indicates a strong preference for bus transportation on the Malang-Blitar route, with ease of access as the main factor influencing decisions.

The contribution of this research to the body of knowledge is the development of an optimized ensemble bagging methodology for transportation mode selection prediction, particularly in the Indonesian context, as well as identification of the optimal number of replications to achieve maximum accuracy.

CRediT Authorship Contribution Statement

Nuzulul Laili Nabila: Conceptualization (lead), Methodology (lead), Investigation (lead), Formal analysis (lead), Data curation (supporting), Writing-original draft (lead). **Sobri Abusini:** Conceptualization (supporting), Methodology (supporting), Supervision (lead). **Umu Sa'adah:** Conceptualization (supporting), Data curation (lead), Methodology (supporting), Writing – review & editing (lead), Supervision (supporting).

Declaration of Generative AI and AI-assisted Technologies

The authors used Grammarly for English language enhancement, grammar checking, and sentence structure improvement throughout the manuscript preparation. This AI-assisted tool was utilized to ensure proper English grammar, spelling accuracy, and clarity of expression. All suggestions provided by Grammarly were carefully reviewed, evaluated, and selectively implemented by the authors to maintain the integrity and accuracy of the scientific content. The AI tool did not contribute to research design, methodology development, data collection, statistical analysis, or interpretation of results. All intellectual content, research findings, and conclusions remain entirely the work of the authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding and Acknowledgments

This research received no external funding. The authors would like to thank the Department of Mathematics, Faculty of Science and Mathematics, Universitas Brawijaya for providing the institutional support for this research. We also express our gratitude to all respondents who participated in this study by sharing their transportation experiences on the Malang-Blitar route.

Data and Code Availability Statement

The datasets generated and analyzed during the current study are not publicly available due to privacy considerations of survey respondents but are available from the corresponding author on reasonable request. The code used for ensemble bagging binary logistic regression analysis is available from the corresponding author upon reasonable request.

References

[1] R. Maulana and M. H. Yudhistira, "Socio-economic factors affecting the choice of transportation mode in jakarta metropolitan area," *Jurnal Pembangunan Wilayah dan Kota*, vol. 16, no. 4, pp. 1245–1252, 2020. DOI: 10.14710/pwk.v16i4.32222.

[2] C. B. et al, "A machine learning comparison of transportation mode changes from high-speed railway promotion in thailand," *Results in Engineering*, vol. 24, no. 103110, pp. 1–17, 2024. DOI: 10.1016/j.rineng.2024.103110.

[3] X. S. T. Cao and J. Wang, "A comparison of the effectiveness of techniques for predicting binary dependent variables," *2022 21st International Symposium on Communications and Information Technologies, ISCIT 2022*, pp. 160–165, 2022. DOI: 10.1109/ISCIT55906.2022.9931323.

[4] Q. J. X. Cai and W. Zhang, "Traffic flow prediction: A method using bagging-based ensemble learning model," *Science Journal of Applied Mathematics and Statistics,*, vol. 12, no. 5, pp. 72–79, 2024. DOI: 10.11648/j.sjams.20241205.11.

[5] M. W. A. Brenner and S. Amin, "Interpretable machine learning models for modal split prediction in transportation systems," *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, pp. 901–908, 2022. DOI: 10.1109/ITSC55140.2022.9921938.

[6] N. F. et al., "Travel mode choice modeling: Predictive efficacy between machine learning models and discrete choice model," *The Open Transportation Journal*, vol. 15, pp. 241–255, 2021. DOI: 10.2174/18744478021150102.

[7] Z. D. P. Wang and J. Tang, "Modelling intercity travel mode choice behavior based on the logistic regression stacking fusion algorithm," *7th IEEE International Conference on Transportation Information and Safety, ICTIS 2023*, pp. 2236–2242, 2023. DOI: 10.1109/ICTIS60134.2023.10243830.

[8] X. Y. X. Zhao and P. V. Hentenryck, "Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models," *Travel Behav Soc*, vol. 20, pp. 22–35, 2020. DOI: 10.1016/j.tbs.2020.02.003.

[9] T. H. J. Á. Martín-Baos and R. García-Ródenas, "A prediction and behavioural analysis of machine learning methods for modelling travel mode choice," *Transp Res Part C Emerg Technol*, vol. 156, pp. 1–44, 2021. DOI: 10.1016/j.trc.2023.104318.

[10] E. O. D. W. H. Naseri and Z. Patterson, "Application of machine learning to child mode choice with a novel technique to optimize hyperparameters," *Int J Environ Res Public Health*, vol. 19, no. 24, pp. 1–19, 2022. DOI: 10.3390/ijerph192416844.

[11] H. Chen and Y. Cheng, "Travel mode choice prediction using imbalanced machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 3795–3808, 2023. DOI: 10.1109/TITS.2023.3237681.

[12] V. Kotenko, "Application of algorithmic models of machine learning to the freight transportation process," *Transport technologies*, vol. 2022, no. 2, pp. 10–21, 2022. DOI: 10.23939/tt2022.02.010.

[13] A. P. A. J. J. M. D'Cruz and V. S. Manju, "Mode choice analysis of school trips using random forest technique," *Archives of Transport*, vol. 63, no. 3, pp. 39–48, 2022. DOI: 10.5604/01.3001.0015.9175.

[14] S. L. D. W. . Hosmer and R. X. . Sturdivant, *Applied logistic regression*. Wiley, 2013.

[15] S. Burhan and A. K. Jaya, "Penaksiran parameter regresi linier logistik dengan metode maksimum likelihood lokal pada resiko kanker payudara di makassar," *Jurnal Matematika, Statistika, dan Komputasi*, vol. 14, no. 2, pp. 159–165, 2018. DOI: 10.20956/jmsk.v14i2.3556.

[16] R. Ramandhani and D. Safitri, "Metode bootstrap aggregating regresi logistik biner untuk ketepatan klasifikasi kesejahteraan rumah tangga di kota pati," *Jurnal Gaussian*, vol. 6, no. 1, pp. 121–130, 2017. DOI: 10.14710/j.gauss.6.1.121-130.

[17] L. Anisa and N. A. K. Rifai, "Analisis regresi logistik biner dengan metode penalized maximum likelihood pada penyakit covid-19 di rsud pringsewu," *Jurnal Riset Statistika*, pp. 129–136, 2022. DOI: 10.29313/jrs.v2i2.1425.

[18] D. B.-F. Z. Sanchez-Varela and M. A. Gomez-Solaeche, "Prediction of loss of position during dynamic positioning drilling operations using binary logistic regression modeling," *Journal of Marine Science and Engineering*, vol. 9, pp. 1–18, 2021. DOI: 10.3390/jmse9020139.

[19] S. A. R. L. A. Yates and B. W. Brook, "Cross validation for model selection: A review with examples from ecology," *Ecol Monogr*, vol. 93, no. 1, pp. 1–24, 2023. DOI: 10.1002/ecm.1557.

[20] L. M. Cendani and A. Wibowo, "Perbandingan metode ensemble learning pada klasifikasi penyakit diabetes," *Jurnal Masyarakat Informatika*, vol. 13, no. 1, pp. 33–44, 2022. DOI: 10.14710/jmasif.13.1.42912.

[21] R. F. A. Efendi and B. Rahayudi, "Ensemble adaboost in classification and regression trees to overcome class imbalance in credit status of bank customers," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 17, pp. 3428–3437, 2020.

[22] V. G. A. I. Marqués and J. S. Sánchez, "Exploring the behaviour of base classifiers in credit scoring ensembles," *Expert Syst Appl*, vol. 39, no. 11, pp. 10 244–10 250, 2012. DOI: 10.1016/j.eswa.2012.02.092.

[23] F. L. J. Wang and J. Liang, "Rss-bagging: Improving generalization through the fisher information of training data," *IEEE Trans Neural Netw Learn Syst*, vol. 36, no. 2, pp. 1974–1988, 2025. DOI: 10.1109/TNNLS.2023.3270559.

[24] T. W. S. Innassuraiya and I. T. Utami, "Analisis klasifikasi menggunakan metode regresi logistik biner dan bootstrap aggregating classification and regression trees (bagging cart) (studi kasus: Nasabah koperasi simpan pinjam dan pembiayaan syariah (kspps))," *Jurnal Gaussian*, vol. 11, no. 2, pp. 183–194, 2022. DOI: 10.14710/j.gauss.v11i2.35458.

[25] M. Raihan-Al-Masud and M. R. H. Mondal, "Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms," *PLoS One*, vol. 15, no. 2, pp. 1–21, 2020. DOI: 10.1371/journal.pone.0228422.

[26] D. L. W. Putri and S. Mariani, "Peningkatan ketepatan klasifikasi model regresi logistik biner dengan metode bagging (bootstrap aggregating)," *Indones. J. Math. Nat. Sci.*, vol. 44, no. 2, pp. 61–72, 2021. DOI: 10.15294/ijmns.v44i2.33144.

[27] K. Tanaka, "Institute of developing economies transport costs, distance, and time: Evidence from the japanese census of logistics," *IDE-JETRO*, no. 241, pp. 1–32, 2010.

[28] S. Z. S. Xu and F. Zeng, "Comparison of different approaches of machine learning methods with conventional approaches on container throughput forecasting," *Applied Sciences (Switzerland)*, vol. 12, no. 19, pp. 1–19, 2022. DOI: 10.3390/app12199730.