BAREKENG: Journal of Mathematics and Its Applications
December 2026 Volume 20 Issue 4 Page 3271-3282
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss4pp3271-3282

PERFORMANCE COMPARISON OF MISSFOREST AND MICE IN HANDLING MISSING CATEGORICAL DATA

Nurhidayah [1*], **Kusman Sadik** [2], **Aji Hamim Wigena** [3]

[1,2,3]*Statistics and Data Science Program, School of Data Science, Mathematics, and Informatics, IPB University*
Jln. Dramaga, IPB Dramaga Campus, Bogor,16680, Indonesia

*Corresponding author's e-mail: *nurhidayah@apps.ipb.ac.id.*

Article Info

Article History:

Received: 11st October 2025
Revised: 30th April 2026
Accepted: 31st May 2026
Available online: 24th August 2026

Keywords:

Categorical Data;
Imputation;
Missing Data.

ABSTRACT

Missing data represent a common challenge in statistical modeling and can substantially reduce the performance of classification algorithms. This study examines the impact of missing values on the performance of the XGBoost model by considering different proportions of missingness (50% and 75%) and various combinations of affected variables under the Missing Completely at Random (MCAR) and Missing at Random (MAR) mechanisms. Two imputation methods, MissForest and Multiple imputation by chained equations (MICE), were compared with a baseline model without imputation. The analysis of variance revealed that the interaction between imputation method, missing data proportion, and variable combinations had a significant effect on accuracy and specificity, while sensitivity remained relatively stable across scenarios. Tukey tests confirmed that MissForest consistently outperformed the other approaches, producing the highest accuracy and specificity, especially at 50% missingness with three variables affected. Moreover, the evaluation of categorical distributions before and after imputation indicated that MissForest better preserved category balance compared to MICE. These findings highlight that the performance of imputation methods strongly depends on the characteristics of missing data. Overall, MissForest demonstrated clear superiority in handling missing categorical data, maintaining distributional integrity while enhancing the classification performance of the XGBoost model. This study advances statistical learning by giving empirical evidence and practical tips for choosing strong imputation strategies in categorical datasets. This improves the reliability of predictive modeling when data is missing.

How to cite this article:

Nurhidayah, K. Sadik, and A.H. Wigena, "PERFORMACE COMPARISON OF MISSFOREST AND MICE IN HANDLING MISSING CATEGORICAL DATA", *BAREKENG: J. Math. & App.,* vol. 20, no. 4, pp. 3271-3282, Dec. 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journale-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*
Research Article · Open Access

1. INTRODUCTION

Categorical data is one of the most widely used types of data in research and surveys, especially in the fields of social sciences, health, economics, and information technology. This data represents variables in the form of categories [1], such as gender, employment status, or education level. Categorical data plays a crucial role because many real problems cannot be measured numerically but rather through specific classifications. Therefore, the quality of the analysis is highly dependent on the completeness of the categorical data obtained [2].

However, in reality, both categorical and numerical data often experience missing values [3]. This condition can occur due to various factors such as recording errors, respondents not completing surveys, or limitations in the data collection system. The presence of missing data can reduce the quality of analysis, cause bias, and reduce the validity of research results if not handled properly [4]. Handling missing values in categorical data is more challenging than in numerical data, as it cannot be addressed using simple methods [5].

According to [6], missing data can occur under three mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). In the MCAR mechanism, the probability of a value being missing is independent of both the variable itself and other variables in the dataset. In the MAR mechanism, the probability of missingness depends on other observed variables but not on the missing value itself. Meanwhile, in the MNAR mechanism, the probability of missingness depends directly on the missing value itself. Understanding these missing data mechanisms is essential for selecting appropriate handling methods, as different mechanisms may require different imputation strategies and can affect model performance [7].

There are two approaches to handle missing values, namely deletion and imputation. Deletion often results in the loss of information and reduces the sample size, which can decrease model performance [8]. Therefore, imputation is more widely used because it can estimate missing values based on information from other variables. Various imputation methods have been developed, ranging from simple approaches to those based on machine learning algorithms.

In recent years, machine learning-based imputation methods have been increasingly developed. MissForest is a non-parametric imputation method that uses the Random Forest algorithm iteratively to predict missing values, both in numerical and categorical data, without assuming a specific distribution [9]. Meanwhile, MICE (Multiple Imputation by Chained Equations) can be extended by incorporating Random Forest as the predictive model within each conditional imputation step. This approach allows for the identification of non-linear patterns and interactions between variables that are more complex than those found in classical regression-based MICE methods.

Previous studies have shown that modern imputation methods such as MissForest and MICE have advantages in handling missing data. A study conducted by [10] found that using MissForest as an imputation method for student graduation status data yielded significantly better results than median or mode imputation. Classification models built with Random Forest and XGBoost after MissForest imputation showed consistent improvements in prediction performance. Similarly, [11] compared various imputation techniques on health data and found that MissForest consistently produced the lowest prediction errors, followed by the MICE method. These results confirm that ensemble learning-based approaches are more effective in addressing missing data with complex patterns than simple approaches such as mode or KNN. In line with this, [12] shows that MICE and other decision tree-based variants are able to maintain parameter estimation accuracy and statistical inference validity in empirical studies, and are proven to be superior to the classic MICE method based on Predictive Mean Matching (PMM). In fact, further development by [13] combining MissForest with the Recursive Feature Elimination technique resulted in the RFE-MissForest method, which was able to significantly improve imputation accuracy in large-scale medical datasets with complex missing data patterns.

Despite the growing body of research on imputation methods, several gaps remain. Previous studies [10]-[13] primarily focused on improving prediction performance using imputation techniques, often on mixed-type or real-world datasets with relatively moderate missing proportions. However, limited attention has been given to scenarios involving purely categorical data with high proportions of missingness, particularly in the range of 50% to 75%. Furthermore, prior studies rarely examine the combined effects of

multiple experimental factors, such as imputation methods, missing data proportions, and combinations of affected variables, in a unified framework.

Therefore, this study aims to address these gaps by systematically evaluating the performance of MissForest and MICE using simulated categorical data under high rates of missingness. In addition, this study emphasizes the analysis of three-factor interactions to better understand how different experimental conditions jointly influence model performance. This approach offers a more comprehensive evaluation of imputation effectiveness, particularly in preserving categorical distributions and enhancing classification outcomes in the XGBoost model.

2. RESEARCH METHODS

2.1 Data

This study used simulated data, namely, artificially generated data created under controlled conditions by the researcher with predetermined characteristics, to examine imputation methods in handling missing data problems in categorical variables. The use of simulated data aimed to provide a controlled testing environment, allowing the evaluation of imputation methods to be conducted systematically and objectively.

Specifically, the simulation data consisted of five predictor variables and one binary response variable to evaluate the performance of the missForest and MICE imputation methods in handling missing data using the XGBoost algorithm. The simulation focused on datasets with categorical explanatory variables and aimed to assess the effectiveness of the two imputation techniques under various missing-data scenarios.

The explanatory variables were generated according to their categorical levels, where X_1 and X_2 are nominal variables with three levels, X_3 is nominal with four levels, X_4 is an ordinal variable with four levels, and X_5 is a nominal with the five levels. The response variable Y is binary, taking values 0 and 1. In this simulation, all five predictors variables were allowed to experience missing values in different combinations according to the predefined missing data scenarios. The detailed structure of the simulated variables is presented in Table 1.

Table 1. Variables of Simulation Data

Variables	Descriptions
Y	0 and 1
X_1	A, B, C
X_2	D, E, F
X_3	G, H, I, J
X_4	Low, medium, high
X_5	K, L, M, N, O

To examine the relationships between variables, Cramer's V was used (Fig. 1). The heatmap illustrates the strength of association among the categorical variables. The results indicate that X_5 has the strongest associations with the response variables Y (Cramer's V = 0.54) compared to other predictors, while the relationship between X_1, X_2, X_3, X_4 with Y are relatively weak, suggesting that these variables have no significant correlations with the response. In addition, a very strong relationship is observed between X_3 and X_5 (Creamer's V = 0.78), indicating a close association between these two predictors. Conversely, the associations among the other predictor variables are relatively weak with one another.

In the data generation process, variables X_1, X_2, X_4, and X_5 were generated independently, whereas X_3 was generated conditionally based on the value of $X_5$to represent the Missing at Random (MAR) mechanism. Consequently, when the combination of missing-data variables involved both X_3 and X_5, the missing-data mechanism was classified as MAR, since the probability of missingness in X_3 depended on the value of X_5. In contrast, when the missing-data combination did not involve these two variables, the mechanism was categorized as Missing Completely at Random (MCAR), as the missingness occurred randomly without depending on the values of other variables.

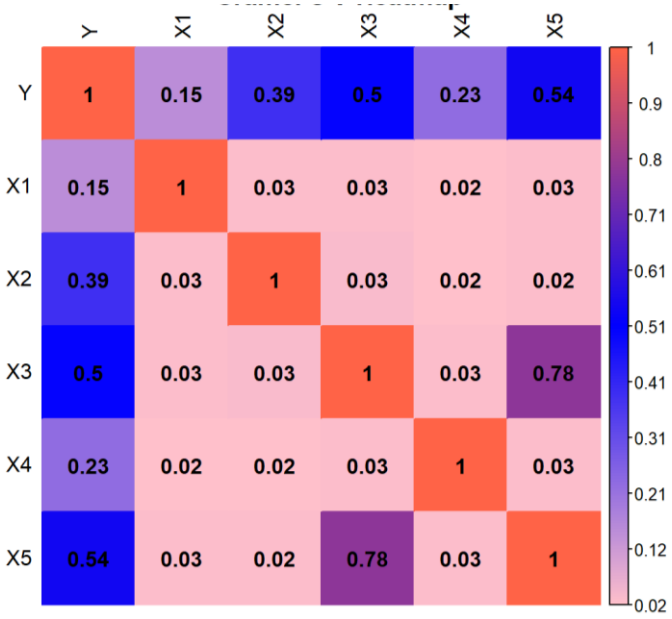

Figure 1. Cramer's V Heatmap

Thus, the missing-data scenarios in this simulation consisted of MCAR and MAR mechanisms. This is because many imputation algorithms, including missForest and MICE, are specifically designed to handle MCAR and MAR effectively without requiring additional assumptions. In contrast, the Missing Not at Random (MNAR) mechanism was not considered in this study because the probability of missingness depends directly on the missing value itself, making regression-based or relationship-based imputation approaches less effective [14].

The simulation was designed by considering two missing data proportions based on Table 3. Each scenario was analyzed using three treatments: no imputation (TP), MissForest, and MICE imputation, with 30 replications for each combination to evaluate model consistency and overall performance.

2.2 Methods

The following are the stages of research related to simulation data :

1. Generate simulation data consisting of five categorical predictor variables with a multinomial distribution with $n = 5000$.
2. All predictor variables are combined into a single data frame so that they can be modeled using logistic regression. They are then transformed into dummy variables. This is done to convert categorical variables into numerical variables without losing their categorical information
3. The regression coefficient is set subjectively to give weight to each category of the predictor variable; then a normally distributed error component is added. The resulting linear score equation (z) is written mathematically as follows:

$$z = 0.383X_{1B} - 1.935X_{1C} + 0.138X_{2E} - 2.985X_{2F} + 0.950X_{3H} - 2.349X_{3I} + 1.752X_{3J} + 2.0585X_{4Medium} - 0.023X_{4High} - 1.604X_{5L} - 1.445X_{5M} + 2.215X_{5N} + 0.122X_{5O}. \quad (1)$$

4. The z value is calculated based on the multiplication of the dummy variable with the regression coefficient plus the error. Next, the z is converted into probability (p) using the logistic function [15] :

$$P(Y = 1) = \frac{1}{1 + e^{-z}}. \quad (2)$$

To illustrate the classification process, Table 2 presents a sample value of the linear score z, probability P, and the resulting classified class Y.

Table 2. Summary of Statistical Values for Y

z	$P(Y=1)$	Y
0.462091717	0.6135103	1
0.444250490	0.3907286	1
-0.584758958	0.3578383	0
-1.805043593	0.1412382	0
⋮	⋮	⋮
3.150018658	0.9589095	1

Table 2 presents an example of the calculated z values, the response probabilities $P(Y=1)$, and the resulting binary response classes Y. The z values were obtained from the combination of explanatory variables through the Logistic regression model, while the probabilities $P(Y=1)$ were derived by transforming z using the sigmoid function. The table also shows a monotonically increasing relationship between z and $P(Y=1)$, indicating that higher z values correspond to a greater probability of the response $Y=1$.

5. Next, determine the threshold value using the median probability (quantile$(P, 0.5)$) generated from the logistic function to maintain a balance between classes 0 and 1. The calculation result is consistently around 0.3587, so it is used as a reference in the response variable classification process.

6. Binary response Y is formed based on the probability P, which is calculated from the linear score z using the logistic function. Classification is performed using the median probability value, which is 0.3587, as the threshold. This approach aims to maintain a balanced distribution between classes 0 and 1. The classification rule is defined as follows:

$$y = \begin{cases} 1, if\ P(Y=1) > 0.3587 \\ 0, if\ P(Y=1) \leq 0.3587 \end{cases}. \quad (3)$$

By using this threshold, the resulting response variable is balanced, consisting of approximately 50% class 0 and 50% class 1.

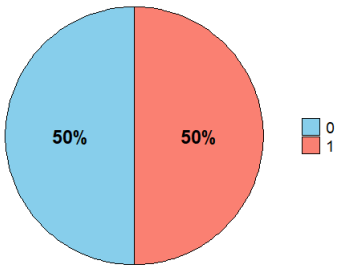

Figure 2. Distribution of Class Y

7. After the complete dataset is generated, the simulation design is adjusted based on two factors: the missing proportion and the missing pattern. Based on Table 2, these factors form eight scenarios that represent different conditions. The data loss process is generated using two mechanisms, namely MCAR, which is random, and MAR, which uses a single cut-off approach to create conditions where missing data is influenced by other variables [16].

Table 3. Simulation Data Design

Scenarios	Proportion of missing data	Pattern of missing data
1	75%	X_1, X_5
2		X_1, X_2, X_5
3		X_1, X_2, X_3, X_5
4		X_1, X_2, X_3, X_4, X_5

Scenarios	Proportion of missing data	Pattern of missing data
5	50%	X_1, X_5
6		X_1, X_2, X_5
7		X_1, X_2, X_3, X_5
8		X_1, X_2, X_3, X_4, X_5

8. After the eight scenarios were created, the missing values were imputed using MissForest and Multiple Imputation by Chained Equations (MICE). These imputation methods produced complete datasets for each scenario, which were subsequently used in the XGBoost modeling process. The dataset was divided into training and testing sets using an 80:20 ratio, resulting in 4,000 training samples and 1,000 testing samples. During model training, 10-fold cross-validation was applied to determine the optimal number of boosting rounds in the XGBoost model.
9. The performance of the XGBoost model was evaluated using a confusion matrix to compare actual and predicted values. Three evaluation metrics were used, namely accuracy, sensitivity, and specificity. Accuracy measures the proportion of correctly classified observations, sensitivity measures the model's ability to correctly identify positive cases, and specificity measures the model's ability to correctly identify negative cases. These evaluation metrics were selected because the primary objective of this study was to compare the impact of different imputation methods on classification performance [17]
10. Steps 7 through 9 were repeated for 30 iterations in each simulation scenario to ensure the robustness and reliability of the results, allowing for a comprehensive assessment of variability and stochastic effects in the imputation and modeling processes. Finally, analysis of variance (ANOVA) was conducted to evaluate the performance differences among the complete data model, MissForest, and MICE imputation models.

3. RESULTS AND DISCUSSION

Based on Fig. 2, X_5 was selected as the reference variable in the MAR mechanism, while other variables were subjected to missingness under MCAR. In the MCAR, missing data occur entirely at random and are not influenced by any variable, observations were randomly selected and replaced with missing data according to the specified missing data proportions. In contrast, the MAR introduces missingness influenced by another variable. Using a single cut-off approach, X_5 determined which data points were more likely to be missing, meaning the probability of missingness was related to the characteristics of X_5 rather than being completely random. In this simulation, scenarios that included both X_3 and X_5 were categorized as MAR, while all other combinations that did not include both variables were classified as MCAR. The detailed mechanism for each scenario is presented in Table 2.

In this study, missing data were handled using two imputation methods, MissForest and MICE using Random Forest. Both methods estimate missing values by leveraging patterns and relationships among variables in the dataset.

MissForest performs imputation directly and iteratively using the Random Forest algorithm. It was executed with 50 trees (ntree = 50) and a maximum of 5 iterations (maxiter = 5) to balance computational efficiency and imputation performance. The Random Forest model iteratively updates missing entries by fitting decision trees on the observed data, converging when the maximum number of iterations is reached.

MICE was implemented using the mice package in R, with the imputation method set to Random Forest by assigning the argument method = "rf" to all variables. The method was run with 3 imputed datasets and a maximum of five iterations. This approach sequentially models each variable with missing data using Random Forest, updating imputations until convergence.

The imputation performed can affect the distribution of categories in certain variables, especially when the amount of missing data is substantial. To illustrate this effect, the distribution of categories before and after imputation is presented in Table 3 below.

This comparison is crucial to evaluate whether the imputation methods successfully preserved the original data structure, particularly the distribution of categorical variables, or if they introduced significant shifts in category proportions. Maintaining the category distribution is essential to ensure that the imputed data remain representative of true underlying patterns, thereby preventing biases that could affect subsequent analyses and model performance.

Table 3. Frequency Distribution

Variables	Methods		
	TP	MissForest	MICE
X_1		Scenario 1	
A	856	2240	3418
B	244	1675	979
C	150	1085	603
X_5			
K	224	952	917
L	255	957	969
M	256	985	1067
N	267	1270	1042
O	248	836	1005
X_1		Scenario 8	
A	1728	2666	3470
B	477	1380	974
C	274	954	556
X_4			
Low	728	1165	1506
Medium	1252	2245	2554
High	477	1590	940

Table 3 shows an increase in the weight of each category level after imputation compared to the data without missing value handling (TP). In scenario 1, for X_1, category A experienced the largest increase using the MICE imputation method, followed by categories B and C. A similar pattern was observed in variable X_5, where all categories increased in weight, with category N showing the greatest increase for both MissForest and MICE. This demonstrates that imputation is capable of recovering missing data while simultaneously increasing the weight of the category, thus preventing information loss.

In scenario 8, the distribution pattern of X_1 remained consistent with the previous scenario, with category A again dominating and showing a significant weight increase under MICE. Conversely, category C showed a smaller increase, with a more balanced distribution produced by MissForest compared to MICE. Meanwhile, in variable X_4, the medium category experienced the largest increase in weight, followed by the low category, which also increased substantially.

Overall, these findings indicate that both MissForest and MICE effectively increase the weight of category levels with missing values. However, there is a tendency difference: MICE tends to amplify the weights of dominant categories more, while MissForest tends to produce a more balanced distribution, including minor and extreme categories.

The MissForest method, which describes the random forest algorithm for performing iterative imputation, shows that the distribution of categories after imputation tends to be more spread out and proportional compared to the MICE method. This is in line with [18] findings, which show that MissForest is effective in handling mixed data (numerical and categorical) and is able to maintain data structure without increasing the dominance of the majority class. With its ability to handle non-linear relationships and interactions between variables, MissForest maintains category diversity even when a significant amount of data is missing. In contrast, the MICE method, despite using random forest as the base model in its chained equations approach, shows a stronger tendency toward majority categories. This is in line with the findings of [19], who stated that multiple imputation using chained models sometimes reinforces the dominant class, so that the distribution of categories can become less balanced after imputation. Overall, the comparison between the two methods indicates that MissForest is better able to maintain the balance of category distribution, while MICE tends to reinforce the dominance of the majority class.

The differences in category distribution generated by these two imputation methods not only affect data representation but also have a direct impact on predictive model performance. Therefore, to assess the extent to which each imputation method affects the XGBoost model, model evaluation was conducted across eight scenarios using fixed predefined parameters (max depth = 6, eta = 0.1, nrounds = 100, subsample = 0.8, and colsample_bytree = 0.8). Hyperparameter tuning was not performed to ensure that the observed performance differences could be attributed to the imputation methods rather than model parameter optimization.

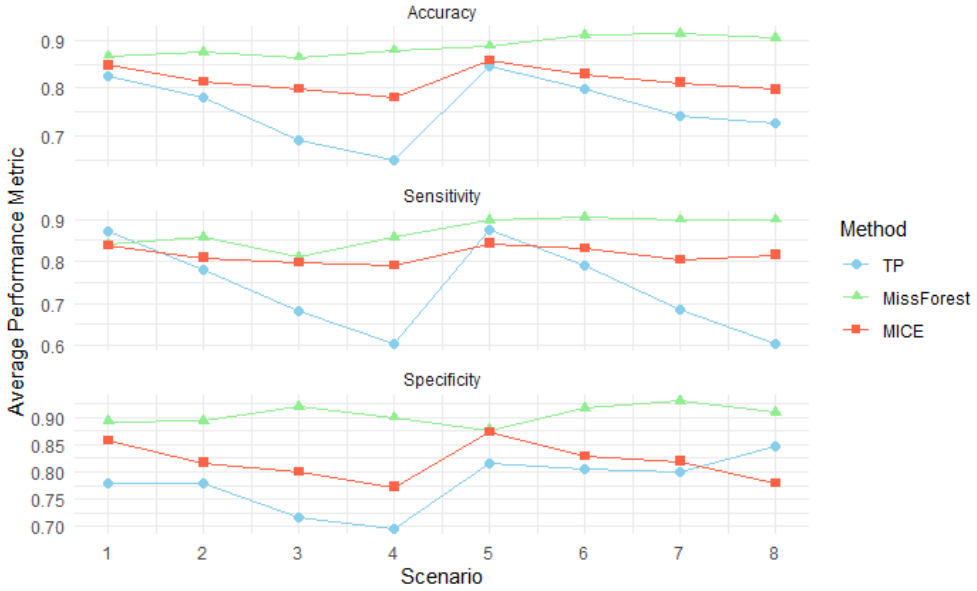

Figure 3. Average Performance of XGBoost

Fig. 3 illustrates the comparative patterns of accuracy, sensitivity, and specificity of the XGBoost model across eight scenarios, each repeated 30 times, employing three different approaches: without handling (TP), with MissForest and MICE Imputation. Consistently, MissForest (denoted by green triangles) demonstrates superior performance across all three metrics relative to other methods. The accuracy achieved by MissForest remains relatively stable, approximating 0.9 across all scenarios, whereas MICE exhibits moderately high but slightly lower performance. In contrast, the model without missing data imputation (TP) showed substantially lower performance, with a pronounced decline observed at 75% missing data proportion, particularly in terms of sensitivity.

With respect to sensitivity, MissForest maintains a consistently high and stable value, ranging approximately between 0.8 and 0.9, indicating robust detection of the positive class. Meanwhile, MICE experiences a modest decrease in certain scenarios, and the TP model drops most sharply, falling below 0.7. This suggests that the model's capacity to detect positive cases is better.

For specificity, MissForest again shows the best and most consistent performance, with values above 0.9, meaning the model effectively recognizes negative cases. In comparison, MICE and TP methods show more fluctuations and lower values.

Next, to evaluate the effect of the imputation method on the statistical performance of the XGBoost model, an analysis of variance was performed with three main factors, namely method (A), proportion of missing data (B), and combination of variables with missing values (C). The results of the ANOVA test are shown in Table 4.

Table 4. ANOVA of Accuracy, Sensitivity, and Specificity

Factor	*p*-Value		
	Accuracy	Sensitivity	Specificity
A	0.000000	0.000000	0.000000
B	0.000000	0.000000	0.000000
C	0.000000	0.000000	0.000000
A × B	0.000145	0.000013	0.000000
A × C	0.000000	0.000000	0.000000
B × C	0.006067	0.850	0.000018

Factor	p-Value		
	Accuracy	Sensitivity	Specificity
A × B × C	0.000838	0.502	0.000000

Table 4 shows that the three-way interaction between A, B, and C has a significant effect on accuracy and specificity, but not on sensitivity. This indicates that model performance, particularly in terms of accuracy and the ability to correctly recognize negative classes (specificity), is strongly influenced by the simultaneous combination of these three factors. In other words, the joint effect of the imputation methods, the proportion of missing data, and the combination of missing variables is complex and can not be interpreted separately.

Conversely, the three-way interaction was not significant for sensitivity, suggesting that the combined influence of the three factors does not jointly affect the model's ability to detect positive classes. Therefore, the analysis of sensitivity was further examined through two-way interactions (A×B and A×C), which were found to be significant, indicating that the model's detection capability varies mainly due to pairwise interactions between the factors rather than their combined effect.

Accordingly, accuracy and specificity were analyzed based on the three-way interaction, while sensitivity was analyzed using two-way interactions. Therefore, a post-hoc Tukey HSD test was conducted on the significant interactions to identify the rank and differences among treatment combinations based on the values of accuracy, sensitivity, and specificity.

Based on Tukey test results, MissForest imputation consistently provides the highest accuracy and specificity across all scenarios, with scenario 3 (50% missing data proportion with the combination of missing data in the variable X_1, X_2, X_3, X_5) showing the best performance. At the 50% missing data proportion, MissForest achieves an average accuracy of 91.46% and specificity of 92.9%, substantially outperforming both MICE and TP. The scenario without handling (TP) at 75% missing data proportion yields the lowest performance. The MICE method generally ranks in the middle, indicating that its performance is inferior to MissForest but better than that without imputation (TP).

As the missing data proportion increases to 75%, accuracy decreases across all models. However, MissForest maintained its superiority with a relatively smaller decline compared to MICE. Meanwhile, MICE performs fairly steadily under Missing Completely at Random (MCAR) conditions but shows a sharper decrease under Missing at Random (MAR) conditions. Conversely, the model without any missing data treatment consistently yields the lowest accuracy and specificity, especially in scenarios with high missing data proportions involving multiple variables.

The sensitivity results showed that MissForest with 50% missing data proportion achieved the highest average sensitivity (0.9009), while TP with 75% missing data yielded the lowest (0.7340). In general, sensitivity decreased as the missing data proportion increased, and MissForest continued to outperform MICE. In the interaction between method and variable combination, the highest sensitivity was observed in MissForest with the combination of $X_1, X_2,$ and X_5 (0.8815), whereas the lowest was found in TP for the combination $X_1, X_2, X_3, X_4,$ and X_5 (0.6031). These findings indicate that MissForest consistently produces higher sensitivity, while the absence of imputation results in a decline in the model's ability to detect positive classes.

Missing data can significantly affect the performance of classification models, as incomplete information may lead to suboptimal performance. XGBoost is incorporated in the Attribute (MIA) approach, which utilizes a default direction procedure during tree splitting [20]. In the process, XGBoost automatically learns the optimal branch (left or right) for observations with missing data, thereby minimizing information loss [21]. However, this built-in mechanism is not always optimal for handling missing data; data imputation is still required to improve model performance [22], particularly for categorical data with a high proportion of missingness.

These findings underscore that the effectiveness of imputation methods is strongly influenced by the interaction between the type of method, the proportion of missing data, and the variables affected by missingness. Overall, in this simulation study, MissForest proved to be more adaptive to both MCAR and MAR patterns, whereas MICE with the random forest function tends to be stable only under MCAR conditions. The absence of missing data handling in the XGBoost model results in a notable decline in model performance. Therefore, despite XGBoost's internal MIA mechanism, applying missing data imputation during data preprocessing remains essential to ensure stable and optimal XGBoost performance.

4. CONCLUSION

Based on the results of the study indicate that the factors of methods, proportion, and combination of variables with missing data have a significant effect on the performance of the XGBoost model. The model performance is significantly influenced both by the main effects of each factor and by interactions among them. In addition, the imputation method was found to alter the level distribution of categorical variables, suggesting that maintaining distribution stability across levels is important to ensure the consistency and performance of the XGBoost model.

Overall, the MissForest imputation method proved to be the most effective in maintaining the performance of the XGBoost model on simulated data. MissForest produced consistent and adaptive results under MCAR. Without any missing data handling, the model performance declined significantly, indicating that the imputation process is an essential step to preserve the performance of the XGBoost model.

Author Contributions

Nurhidayah: Investigation, Visualization, Writing- Original Draft, Writing – Review and Editing. Kusman Sadik: Conceptualization, Supervision, Methodology, Validation. Aji Hamim Wigena: Conceptualization, Supervision, Methodology, Validation. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research was supported by the Graduate Program in Statistics and Data Science, IPB University.

Acknowledgment

The authors gratefully acknowledge the support from the Indonesia Endowment Fund for Education (LPDP) and the Graduate Program in Statistics and Data Science, IPB University.

Declarations

The authors state that there are no conflicts of interest related to this study.

Declaration of Generative AI and AI-assisted Technologies

Generative AI (e.g, ChatGPT) was used exclusively for language editing and stylistic improvement. The authors take full responsibility for the content and confirm that all analyses, results, and interpretations are their own. The final manuscript was additionally reviewed for linguistic accuracy by an English-language expert.

REFERENCES

[1] A. C. Miola, “COMPARING CATEGORICAL VARIABLES IN CLINICAL AND EXPERIMENTAL STUDIES,” vol. 7301, 2022.

[2] N. Nezami, P. Haghighat, D. Gándara, and H. Anahideh, “ASSESSING DISPARITIES IN PREDICTIVE MODELING OUTCOMES FOR COLLEGE STUDENT SUCCESS: THE IMPACT OF IMPUTATION TECHNIQUES ON MODEL PERFORMANCE AND FAIRNESS,” *Educ. Sci.*, vol. 14, no. 2, p. 136, Jan. 2024. doi: https://doi.org/10.3390/educsci14020136.

[3] I. M. Sumertajaya, E. Rohaeti, A. H. Wigena, and K. Sadik, “VECTOR AUTOREGRESSIVE-MOVING AVERAGE IMPUTATION ALGORITHM FOR HANDLING MISSING DATA IN MULTIVARIATE TIME SERIES,” *IAENG Int. J. Comput. Sci.*, vol. 50, no. 2, 2023, [Online]. Available: https://www.researchgate.net/profile/I-Made-Sumertajaya/publication/371322705_Vector_Autoregressive-Moving_Average_Imputation_Algorithm_for_Handling_Missing_Data_in_Multivariate_Time_Series/links/647f2db979a722376513992f/Vector-Autoregressive-Moving-Average-Imputation-Algorithm-for-Handling-Missing-Data-in-Multivariate-Time-Series.pdf

[4] O. Akande, F. Li, and J. Reiter, “AN EMPIRICAL COMPARISON OF MULTIPLE IMPUTATION METHODS FOR CATEGORICAL DATA,” *Am. Stat.*, vol. 71, no. 2, pp. 162–170, 2017. doi: https://doi.org/10.1080/00031305.2016.1277158.

[5] Y. Sun, J. Li, Y. Xu, T. Zhang, and X. Wang, “DEEP LEARNING VERSUS CONVENTIONAL METHODS FOR MISSING DATA IMPUTATION: A REVIEW AND COMPARATIVE STUDY,” *Expert Syst. Appl.*, vol. 227, no. August 2022, 2023.

doi: https://doi.org/10.1016/j.eswa.2023.120201.
[6] D. B. Little, Roderick J. A.; Rubin, *STATISTICAL ANALYSIS WITH MISSING DATA*. 2019. doi: https://doi.org/10.1002/9781119482260
[7] P. Ayokunle, J. Tapamo, and A. G. Honor, "EFFECTIVE AND EFFICIENT HANDLING OF MISSING DATA IN SUPERVISED MACHINE LEARNING," vol. 8, no. December 2024, pp. 361–373, 2025. doi: https://doi.org/10.1016/j.dsm.2024.12.002.
[8] D.-H. Lee, S.-E. Woo, M.-W. Jung, and T.-Y. Heo, "EVALUATION OF ODOR PREDICTION MODEL PERFORMANCE AND VARIABLE IMPORTANCE ACCORDING TO VARIOUS MISSING IMPUTATION METHODS," *Appl. Sci.*, vol. 12, no. 6, p. 2826, Mar. 2022. doi: https://doi.org/10.3390/app12062826.
[9] T. Shadbahr *et al.*, "THE IMPACT OF IMPUTATION QUALITY ON MACHINE LEARNING CLASSIFIERS FOR DATASETS WITH MISSING VALUES," pp. 1–15, 2023. doi: https://doi.org/10.1038/s43856-023-00356-z.
[10] I. Nirmala, H. Wijayanto, and K. A. Notodiputro, "PREDICTION OF UNDERGRADUATE STUDENT'S STUDY COMPLETION STATUS USING MISSFOREST IMPUTATION IN RANDOM FOREST AND XGBOOST MODELS," *ComTech Comput. Math. Eng. Appl.*, vol. 13, no. 1, pp. 53–62, Feb. 2022. doi: https://doi.org/10.21512/comtech.v13i1.7388.
[11] L. O. Joel, W. Doorsamy, and B. S. Paul, "ON THE PERFORMANCE OF IMPUTATION TECHNIQUES FOR MISSING VALUES ON HEALTHCARE DATASETS," pp. 1–20, 2024, [Online]. Available: http://arxiv.org/abs/2403.14687
[12] J. Schwerter, K. Gurtskaia, A. Romero, B. Zeyer-Gliozzo, and M. Pauly, "EVALUATING TREE-BASED IMPUTATION METHODS AS AN ALTERNATIVE TO MICE PMM FOR DRAWING INFERENCE IN EMPIRICAL STUDIES," 2024, [Online]. Available: http://arxiv.org/abs/2401.09602
[13] Y.-H. Hu, R.-Y. Wu, Y.-C. Lin, and T.-Y. Lin, "A NOVEL MISSFOREST-BASED MISSING VALUES IMPUTATION APPROACH WITH RECURSIVE FEATURE ELIMINATION IN MEDICAL APPLICATIONS," *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 269, Nov. 2024. doi: https://doi.org/10.1186/s12874-024-02392-2.
[14] P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, "MISSING DATA IN CLINICAL RESEARCH: A TUTORIAL ON MULTIPLE IMPUTATION," *Can. J. Cardiol.*, vol. 37, no. 9, pp. 1322–1331, Sep. 2021. doi: https://doi.org/10.1016/j.cjca.2020.11.010.
[15] D. Dey, S. Haque, M. Islam, U. I. Aishi, and S. S. Shammy, "THE PROPER APPLICATION OF LOGISTIC REGRESSION MODEL IN COMPLEX SURVEY DATA : A SYSTEMATIC REVIEW," *BMC Med. Res. Methodol.*, vol. 5, 2025. doi: https://doi.org/10.1186/s12874-024-02454-5.
[16] X. Zhang, "HOW TO GENERATE MISSING DATA FOR SIMULATION STUDIES," vol. 19, no. 2, pp. 100–122, 2023. doi: https://doi.org/10.20982/tqmp.19.2.p100
[17] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "SELECTING AND INTERPRETING MULTICLASS LOSS AND ACCURACY ASSESSMENT METRICS FOR CLASSIFICATIONS WITH CLASS IMBALANCE : GUIDANCE AND BEST PRACTICES," pp. 1–22, 2024. doi: https://doi.org/10.3390/rs16030533
[18] P. Buczak, J.-J. Chen, and M. Pauly, "ANALYZING THE EFFECT OF IMPUTATION ON CLASSIFICATION PERFORMANCE UNDER MCAR AND MAR MISSING MECHANISMS," *Entropy*, vol. 25, no. 3, p. 521, Mar. 2023. doi: https://doi.org/10.3390/e25030521.
[19] N. B. Mendoza *et al.*, "EVALUATING IMPUTATION METHODS TO IMPROVE PREDICTION ACCURACY FOR AN HIV STUDY IN UGANDA," pp. 1405–1420, 2024. doi: https://doi.org/10.3390/stats7040082
[20] M. Le Morvan and G. Varoquaux, "IMPUTATION FOR PREDICTION: BEWARE OF DIMINISHING RETURNS," *13th Int. Conf. Learn. Represent. ICLR 2025*, pp. 27597–27636, 2025.
[21] T. Chen and C. Guestrin, "XGBOOST," IN *PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: https://doi.org/10.1145/2939672.2939785.
[22] D. Aulia and R. Hendri, "XGBOOST IN HANDLING MISSING VALUES FOR LIFE INSURANCE RISK PREDICTION," *SN Appl. Sci.*, vol. 2, no. 8, pp. 1–10, 2020. doi: https://doi.org/10.1007/s42452-020-3128-y.