Comparative Analysis of Missing Data Imputation Techniques for Classification Tasks

Guilherme Rodrigues, Lucas Aparicio, Rúben Oliveira

December 31, 2025

Abstract

This report presents a comparison of five missing data imputation techniques (mean, median, K-nearest neighbors, MICE, and MissForest) applied to two classification datasets. We evaluate the performance of three classifiers across multiple imputation strategies using metrics including accuracy and F1-score. Our findings demonstrate that the choice of imputation technique significantly influences classifier performance, with implications for preprocessing pipelines in machine learning applications. MissForest, MICE and KNN were the techniques that performed the best, and there was no correlation found between the RMSE of the imputation and classification performance.

1 Introduction

Missing data is a ubiquitous challenge in machine learning and data analysis. Real-world datasets frequently contain incomplete records due to data collection failures, measurement errors, or intentional non-response. Improper handling of missing values can lead to biased estimates and compromised model performance. The purpose of this project is to investigate how different imputation strategies affect the predictive accuracy of classification models.

We evaluate five imputation techniques across two distinct classification problems and three different classifiers. Besides, we also introduce artificial missing values in the Wine Quality dataset, and compare the imputed values to the real ones, to confirm if a better "estimation" of the missing values always leads to better classifier performance.

This report begins with a theoretical introduction of missing data types and the datasets, imputation techniques and classifiers we use. Then the project methodology is described and finally the results are presented and analyzed.

2 Missing Data Analysis

2.1 Definition and Types of Missing Data

Missing data refers to the absence of values in a dataset where observations are expected. Understanding the mechanism behind missing data is crucial for selecting the appropriate handling strategies. Missing data can be categorized in three types:

Missing Completely at Random (MCAR) occurs when the probability of missingness is independent of both observed and unobserved data. This is the least common but most benign scenario as it does not introduce bias in parameter estimates.

Missing at Random (MAR) indicates that the probability of missingness depends on observed data but not on the unobserved values themselves. Many real-world scenarios approximate MAR, making it a common assumption in practical applications.

Missing Not at Random (MNAR) represents the most problematic case, where missingness depends on the unobserved values themselves. This mechanism can introduce substantial bias and requires careful handling techniques.

2.2 Why Imputation is Important

Deletion of missing values is a simple approach but leads to loss of information and potential bias, especially when missingness is not random. Imputation replaces missing values with estimated values based on available information, preserving sample size and leveraging relationships between variables. Proper imputation enables researchers to conduct more powerful analyses and reduces bias compared to deletion methods. In classification tasks specifically, imputation helps maintain balanced datasets and improves model training efficiency by utilizing all available information.

3 Used Imputation Techniques

3.1 Simple Methods: Mean and Median Imputation

Mean imputation replaces missing values with the average of observed values in that feature. This method is computationally efficient and maintains the global mean of the variable. However, it artificially reduces variance and can distort distributional properties, potentially leading to underestimated standard errors.

Median imputation similarly replaces missing values with the median of observed data. For ordinal data, median imputation is often preferable to mean imputation as it preserves the ordinal nature and is more robust to outliers. Both methods ignore relationships between variables.

3.2 K-Nearest Neighbors (KNN) Imputation

KNN imputation identifies the K nearest neighbors of an observation with missing values based on available features and imputes using the mean or median of those neighbors' values. This method preserves local data structure and relationships between variables. The choice of K influences bias-variance tradeoff; smaller K values capture local patterns while larger K values provide more stable estimates. KNN is more computationally intensive than simple methods but generally produces more realistic imputations by leveraging multivariate information.

3.3 MICE: Multivariate Imputation by Chained Equations

MICE is an iterative imputation method that handles multivariate missing data by imputing each variable with missing values using a flexible series of regression models, creating multiple complete datasets, each with different plausible imputations, then pooling the results for robust analysis, preserving relationships between variables, and accounting for imputation uncertainty.The algorithm cycles through variables with missing data, fitting regression models that predict missing values based on observed data in other variables. After multiple iterations (typically 10-50), the algorithm reaches convergence and produces a single completed dataset.

MICE is theoretically grounded in the principle of imputing under the assumption of MAR or MCAR. It naturally incorporates relationships between variables. The iterative nature allows the method to leverage information from all variables, making it particularly suitable for datasets with complex multivariate patterns.

3.4 MissForest: Iterative Imputation via Random Forests

MissForest is a non-parametric imputation method that uses the random forest algorithm iteratively. It treats imputation as a prediction problem, fitting random forests on observed data to predict missing values. Like MICE, MissForest iterates until convergence, refining imputation estimates with each round.

The algorithm begins by imputing missing values with variable means, then iteratively updates these estimates using random forest predictions. Its non-parametric nature makes it robust to complex, non-linear relationships between variables. MissForest typically outperforms parametric methods on datasets with intricate variable interactions, though it requires higher computational resources.

4 Datasets

4.1 UCI Heart Disease Dataset

The UCI Heart Disease dataset contains medical records from 921 patients with 13 clinical and demographic features including age, sex, chest pain type, blood pressure, cholesterol levels, maximum heart rate, and electrocardiographic measurements. The target variable is the presence and level (from 0 to 4) of heart disease. This dataset naturally contains missing values in several features, making it ideal for evaluating imputation techniques on real-world incomplete data. After analysing the relation between missing values, we concluded that the missing values in the slope, ca and thal attributes are heavily correlated and so the missin data can be considered MAR.

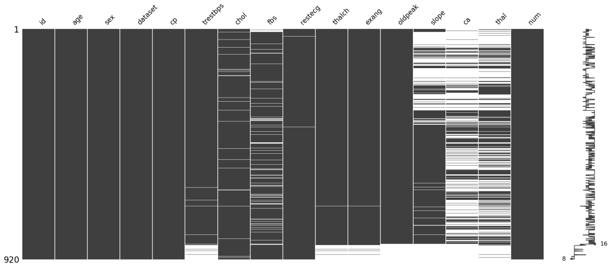

Figure 1: Missing values analysis

4.2 Wine Quality Dataset

The Wine Quality dataset comprises 1144 samples of wine with 11 physicochemical features (e.g., fixed acidity, volatile acidity, sulfur dioxide levels, alcohol content) and a quality rating on an ordinal scale from 1 to 10. Unlike the Heart Disease dataset, Wine Quality is naturally complete, making it suitable for controlled experiments where we artificially introduce missing values. The ordinal classification target (wine quality ratings) aligns with our multi-class classification objectives. By randomly removing data and then comparing imputation results against the original values, we can assess imputation accuracy and check if a better ”prediction” of missing values necessarily leads to better accuracy in classification.

5 Classifiers

We use three classifiers for evaluation: logistic regression, which provides a linear baseline; random forest, which handles non-linear relationships and feature interactions robustly; and XGBoost, which has the advantage of built-in missing value handling for better comparison metrics.

6 Methodology

6.1 Data Preparation and Analysis

For the heart disease dataset, we did an exploratory data analysis to characterize existing missing values, concluding that they are MAR. We also needed to encode categorical values. As for the Wine dataset, we needed to artificially create missing data at two percentages: 10%, and 20%. Missing values were randomly introduced across all features to simulate MCAR conditions. This controlled approach allows us to compare imputed values against true values, providing an objective measure of imputation accuracy.

6.2 Imputation Application

Each dataset undergoes imputation using five techniques:

- Mean imputation
- Median imputation
- KNN imputation (with k=5 and k=10)
- MICE (10 iterations and mean as the initial strategy)
- MissForest (20 iterations and 50 estimators)

For each imputation technique, we generate a complete dataset suitable for classifier training.

6.3 Classification and Evaluation

Three classifiers are trained and evaluated on imputed training data with cross validation using standard classification metrics (accuracy, which measures the proportion of correct predictions among all predictions, providing an overall success rate and f1-score, which is the harmonic mean of precision and recall, providing a balanced metric). Besides, we also did a "no handling case" where the rows with missing values are simply removed, and took advantage from the fact that XGBoost can handle missing values to to the classification before imputation, in order to have a more diverse baseline. We did 10 iterations of this and present the average of the results, in order to account for statistical variance. For the Wine Quality dataset, we additionally compare imputation quality by computing the mean squared error between imputed and original values. For the Heart Disease dataset, we tested a different implementation of MICE (MICERF) with Random Forest Regressor as the estimator instead of Bayesian Regressor.

7 Results and Analysis

7.1 Imputation Quality Assessment

For the Wine Quality dataset, imputation accuracy is assessed by comparing imputed values against original values. The table below present the RMSE for each technique across the different missing value percentages. As we can see, in both missing values percentages, MissForest is the one performing better.

Table 1: Imputation Quality (RMSE) on Wine Quality Dataset

Imputation Technique	10% Missing	20% Missing
Mean	8.216	9.664
Median	7.260	9.543
KNN (K=5)	8.086	9.926
KNN (K=10)	7.220	8.889
MICE	8.068	12.142
MissForest	5.641	7.014

7.2 Classification Performance Comparison

The tables below present the classification metrics across the used datasets.

Table 2: Heart Disease Dataset - Accuracy

Method	Logistic Reg.	Random Forest	XGBoost
No Handling	0.5857	0.5381	0.5101
XGBoost Native	–	–	0.5475
Mean	0.5924	0.5788	0.5448
Median	0.5829	0.5775	0.5435
KNN (k=5)	0.5856	0.5570	0.5597
KNN (k=10)	0.5910	0.5774	0.5516
MissForest	0.6304	0.6671	0.6604
MICE	0.5992	0.5883	0.5828
MICERF	0.6290	0.6522	0.6658

Table 3: Heart Disease Dataset - F1-Score

Method	Logistic Reg.	Random Forest	XGBoost
No Handling	0.5512	0.4843	0.4899
XGBoost Native	–	–	0.5250
Mean	0.5600	0.5452	0.5256
Median	0.5502	0.5525	0.5279
KNN (k=5)	0.5544	0.5276	0.5376
KNN (k=10)	0.5574	0.5500	0.5339
MissForest	0.6080	0.6465	0.6494
MICE	0.5678	0.5617	0.5706
MICERF	0.6027	0.627	0.0.6548

Table 4: Wine Quality 10% Missing - Accuracy

Method	Logistic Reg.	Random Forest	XGBoost
No Handling	0.5270	0.5439	0.5528
XGBoost Native	–	–	0.5632
Mean	0.5656	0.5739	0.5561
Median	0.5679	0.5775	0.5516
KNN (k=5)	0.6389	0.6681	0.6527
KNN (k=10)	0.6285	0.6700	0.6515
MissForest	0.6079	0.6340	0.6268
MICE	0.5942	0.5983	0.5740

Table 5: Wine Quality 10% Missing - F1-Score

Method	Logistic Reg.	Random Forest	XGBoost
No Handling	0.4995	0.5097	0.5248
XGBoost Native	–	–	0.5474
Mean	0.5451	0.5503	0.5413
Median	0.5462	0.5540	0.5369
KNN (k=5)	0.6207	0.6481	0.6380
KNN (k=10)	0.6094	0.6491	0.6364
MissForest	0.5891	0.6143	0.6137
MICE	0.5747	0.5753	0.5576

Table 6: Wine Quality 20% Missing - Accuracy

Method	**Logistic Reg.**	**Random Forest**	**XGBoost**
No Handling	0.4349	0.5159	0.3231
XGBoost Native	–	–	0.5415
Mean	0.5556	0.5491	0.5418
Median	0.5533	0.5552	0.5414
KNN (k=5)	0.6540	0.6884	0.6793
KNN (k=10)	0.6603	0.7032	0.7004
MissForest	0.5997	0.6368	0.6240
MICE	0.5882	0.5846	0.5637

Table 7: Wine Quality 20% Missing - F1-Score

Method	**Logistic Reg.**	**Random Forest**	**XGBoost**
No Handling	0.3963	0.4473	0.2613
XGBoost Native	–	–	0.5256
Mean	0.5338	0.5241	0.5263
Median	0.5304	0.5306	0.5267
KNN (k=5)	0.6368	0.6687	0.6651
RM KNN (k=10)	0.6421	0.6829	0.6851
MissForest	0.5808	0.6175	0.6109
MICE	0.5680	0.5615	0.5484

In the images below we present the RMSE density and compare the classifier's accuracy with the RMSE for each imputation technique. We can see that a lower RMSE does not necessarily lead to better classification results, meaning there is no correlation between the two.

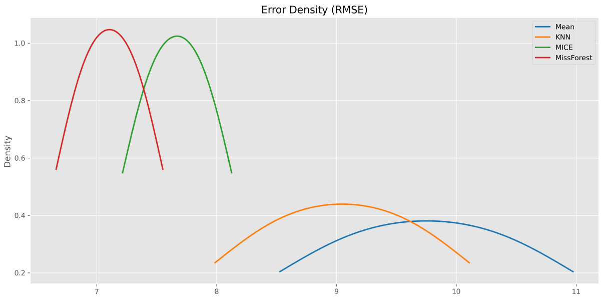

Figure 2: RMSE density

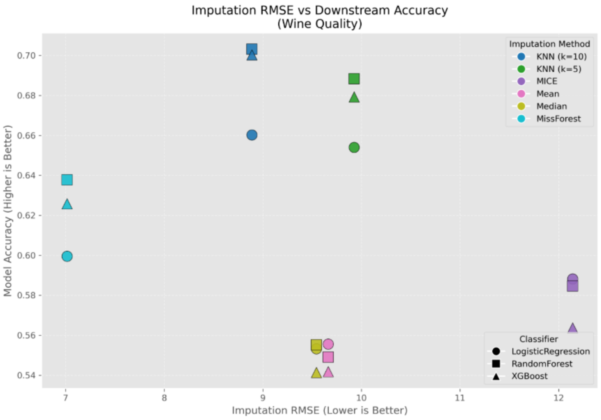

Figure 3: RMSE vs accuracy

7.3 Results Discussion

In the heart disease dataset, MissForest and MICERF are the techniques which perform better across all classifiers. All the others have similar accuracies, with MICE being slightly above. The no handling result shows that after imputation the classification is almost always better, as deleting rows with missing values leads to the loss of information. Also, even tough XGBoost can handle missing values natively, imputation also leads to better classification results.

As for the Wine dataset, the results differ a lot and KNN seems to perform better (both for 10 and 20 percent missing values), being followed by Missforest. MICE seems to perform slightly better than the mean/median, but it is not a much significant difference. As for the relation with the RMSE of the imputation, MissForest was the one performing the best but that did not lead to better classification results. The same goes for MICE, which has the worst RMSE but leads to a better classification than mean/median. We can conclude that "estimating" the correct imputed values does not improve imputation performance.

This results also show that the choice between imputation techniques need to take into account several factors, as in this case more complex and computationally intensive techniques like MICE and MissForest did not always lead to significantly better results than just using the mean or the median.

The differences in the results of both datasets may be because of the nature of the missing values. In the heart disease dataset, the missing values are MAR and are correlated between each other, so MICE and KNN better estimate the relation between the different attributes. As for the artificially introduced values in the wine dataset, these are completely random. Also we need to take into account that we only used two relatively simple and small datasets. This relations may not hold for larger datasets with different types of data. For example, MICE is generally considered one of the best imputation techniques, but the choice of the estimator plays a critical role, and with the Bayes Linear Regressor as a estimator the results where not much better than simpler techniques for the heart disease dataset.

8 Conclusion

This study compared five imputation techniques across Heart Disease and Wine Quality datasets, revealing critical insights for missing data preprocessing in classification pipelines and showing that the chosen imputation technique significantly affects classifier performance. KNN, MICE and Random Forest imputation were the best performers. MICE results with the Bayes Regressor were lower than expected maybe because of the nature of the datasets. Possible improvements could be extending the list and variety of the datasets and perform imputation during Cross-Validation in order to avoid data leakage.

References

[1] L. O. Joel, W. Doorsamy, and B. S. Paul, "A comparative study of imputation techniques for missing values in healthcare diagnostic datasets," *Journal Name*, 2023.

[2] Tahani Aljuaid and Sreela Sasi, "Proper imputation techniques for missing values in data sets," *IEEE Conference Proceedings*, 2021.

[3] Peter C. Austin, Ian R. White, Douglas S. Lee, and Stef van Buuren, "Missing data in clinical research: A tutorial on multiple imputation," *Journal of Clinical Epidemiology*, vol. 63, no. 1, pp. 1–11, 2010.

[4] Rolmez, "Handling missing values strategies and practice," *Kaggle Notebook*, 2024. `https://www.kaggle.com/code/rolmez/handling-missing-values-strategies-and-practice`.

Author Contributions and Acknowledgments

Guilherme Rodrigues (33.3%) report writing, missing values analysys, debugging.

Lucas Aparicio (33.3%) software implementation and validation.

Rúben Oliveira (33.3%) MICE implementation, techniques and evalution research.