STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–10.
Published online in International Academic Press (www.IAPress.org)

The effect of the missing rate and its mechanism on the performance of the imputation methods on different real data sets

Mohammad Mehdi Saber[1], Sara Javadi[2,*], Mehrdad Taghipour[3], Mohamed S. Hamed[4,5], Abdussalam Aljadani[6], Mahmoud M. Mansour[7,5], Haitham M. Yousof[5]

[1]*Department of Statistics, Higher Education Center of Eghlid, Eghlid, Iran*
[2]*Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran*
[3]*Department of Statistics, Faculty of Sciences, University of Qom, Qom, Iran*
[4]*Department of Business Administration, Gulf Colleges, Saudi Arabia*
[5]*Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Benha University, Egypt*
[6]*Department of Management, College of Business Administration in Yanbu, Taibah University, Al-Madinah, Al-Munawarah 41411, Saudi Arabia*
[7]*Department of Management Information Systems, College of Business Administration in Yanbu, Taibah University, Madinah, Saudi Arabia*

Abstract The purpose of this paper is to explore the mechanisms of data missingness and evaluate various imputation techniques used to handle missing data. Missing data is a common issue in data analysis, and its treatment is crucial for accurate modeling and analysis. This paper assesses prevalent imputation methods, including mean imputation, median imputation, K-Nearest Neighbor imputation (KNN), Classification and Regression Trees (CART), and Random Forest (RF). These techniques were chosen for their widespread use and varying levels of complexity and accuracy. Simple methods like mean and median imputation are computationally efficient but may introduce bias, especially when the missingness is not random. In contrast, more advanced methods like KNN, CART, and RF offer better handling of complex missingness patterns by considering relationships among variables. This paper aims to provide guidance for data scientists and analysts in selecting the most appropriate imputation methods based on their data characteristics and analysis objectives. By understanding the strengths and weaknesses of each technique, practitioners can improve the quality and reliability of their analyses.

Keywords Imputation Methods, Missing Data, Multiple Imputation, Multiple Imputation by Chained Equations, Incomplete Data, K-Nearest Neighbor imputation, Random Forest; Single Imputation.

Mathematics Subject Classification: 62P10; 65C60; 60E05; 62G05

DOI: 10.19139/soic-2310-5070-2294

1. Introduction

Ensuring high-quality data is a top priority for scientists and researchers in data science and analysis. The output accuracy of machine learning algorithms is affected by various aspects, such as variable and algorithm selection, sampling and training methods, and appropriate testing and validation datasets [3]. Missing data introduces uncertainty into the analysis process and can negatively impact the accuracy of statistical estimators, leading to weakened statistical power and potentially misguided conclusions [23, 25].

The most widely accepted method for addressing missing data is missing data imputation, which involves estimating plausible values to replace the missing ones [17, 19]. This approach ensures that data analysis proceeds

*Correspondence to: Sara Javadi (Email: javadi.stat@gmail.com). Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

without losing valuable information due to missing values. The primary goal of imputing missing information is to reduce bias caused by missing data rather than excluding incomplete cases [5].

2. Methods of Handling Missing Data

According to [20], when the percentage of missing data is less than or equal to 25%, it is recommended to compare the results both before and after the imputation process.

2.1. Regression Imputation

This technique involves replacing missing values with predicted values generated by regression models using observed values of other variables. This method assumes a linear relationship between the variables, but in reality, such relationships are often nonlinear. Regression imputation preserves data structure but may introduce bias [22].

2.2. KNN Imputation

In this method, missing values are replaced by copying values from similar cases in the same dataset, where similarity is assessed using a distance function. In this method, missing values are replaced by estimating values derived from the k-nearest neighbors in the same dataset, where similarity between cases is determined using a distance function (e.g., Euclidean distance). The imputed value is typically calculated as the average (or weighted average) of the values from the k most similar cases.

2.3. Multiple Imputation

In Multiple Imputation, instead of substituting a single value for each missing value, several plausible values are credited to account for uncertainty surrounding the actual values. This method generates m complete data sets, each containing both observed and imputed values. The process involves three steps:

1. Replacing missing data with multiple plausible values to create m complete data sets.
2. Analyzing each of the data sets separately.
3. Combining the results (see [21]).

2.4. CART in MICE

The CART method for imputation is a tree-based approach that does not require the specification of an imputation model. The CART algorithm essentially creates a binary decision rule by utilizing a variable to split the data into two nodes at each step. This process is aimed at reducing the variance of the outcome within each node, ultimately optimizing the decision-making process [2]. The tree grows by further dividing the data until it reaches a stopping point set by specific parameters. Imputations are then carried out by assigning new subjects to terminal nodes and sampling the outcomes in those nodes [2].

In our study, we utilized the `mice` package to implement Classification and Regression Trees (CART), with default tree-based tuning parameters including a complexity parameter of (E) and a minimum requirement of five observations in each terminal node [6, 5, 28]. For a more comprehensive understanding of the CART algorithm in MICE, see [4].

2.5. RF in MICE

The RF imputation method involves building multiple regression trees and randomly drawing imputations from potential values within each tree. To implement RF, variation is introduced by using bootstrap samples of the original data along with random input selection to construct each tree [1, 14]. The random input selection limits the variables used for node splitting to a random subset of all variables. For a detailed explanation of the RF algorithm in MICE, refer to [6].

In our study, we utilized the `mice` package to apply RF in MICE [24], creating 10 trees in each random forest as per the default setting in the `mice` function [5]. The number of predictors considered for node splitting is $p/3$ rounded down to the nearest integer by default [24], where p represents the total number of predictors [5, 28].

3. Materials and Methods

The dataset utilized in this study was sourced from the UCI Machine Learning Repository [15], with detailed descriptions provided in Table 1.

Initially, the four separate datasets in Table 1 were retrieved from the UCI repository. Then different percentages (10, 20, 30, 40, and 50%) of missing values were introduced to each original dataset. These simulated missing values were then imputed using different imputation techniques. Mean and median imputation involved calculating the average and middle value of the incomplete variable, respectively. KNN imputation was conducted using the `VIM` package in R, as detailed by [13].

For more complex multiple imputation approaches, such as classification and regression trees and random forest, the `mice` package in R was employed, with references provided by [5, 26, 29, 18]. Following imputation, the performance of each method was analyzed. The performance of imputation methods was evaluated using the mean normalized RMSE (NRMSE) metric. NRMSE was utilized due to variations in scales among different variables in the dataset. The mean NRMSE was calculated for each variable in the dataset and served as an assessment of the overall performance of the imputation methods. R was employed for data manipulation, imputation, and performance analysis of the various imputation techniques.

Table 1. Detailed Description of the Data Utilized in the Research Analysis

No	Dataset	Dataset Description	No. of Instances	No. of Attributes
1	Wine Dataset	These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars	178	13
3*2	32.5cmHeart Failure Clinical Records Data Set	This dataset contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features.	299	13
3*3	32.5cmHepatitis C Virus (HCV) for Egyptian Patients Data Set	Egyptian patients who underwent treatment dosages for HCV about 18 months. Discretization should be applied based on expert recommendations; there is an attached file show how	1385	29
3*4	32.5cmConcrete Compressive Strength Data Set	Concrete is vital in civil engineering because of its structural strength. Factors such as age and composition of constituents affect its compressive strength and complicate its analysis. The compressive strength of concrete is a highly non-linear function of age and constituent materials.	1030	9

The wine dataset focuses on the chemical analysis of wines from three different grape cultivars grown in the same Italian region, offering a relatively small sample size (178 instances) with 13 attributes. The data is likely useful for studying how the grape variety impacts wine characteristics and composition. The relatively modest number of instances suggests that the research might be focused on in-depth analysis rather than large-scale predictive modeling. The heart failure clinical records data set with 299 patient profiles, this dataset provides a clinical context, offering 13 attributes related to heart failure. It represents a medical domain dataset that can potentially be used to predict patient outcomes, assess risks, and examine the relationship between clinical features and heart failure

prognosis. Despite its larger sample size compared to the wine dataset, the number of attributes remains consistent. The HCV for Egyptian patients contains medical records for 1385 Egyptian patients, making it the largest in the table. With 29 attributes, the dataset likely provides a deeper level of detail, reflecting the complexity of treatment regimens for Hepatitis C. The mention of discretization based on expert recommendations suggests the need for domain expertise in preparing the data for analysis, which highlights the importance of medical context in working with such datasets. The strength data set contains1030 instances in this dataset relate to concrete, a critical material in civil engineering, with 9 attributes that help model its compressive strength. The dataset's focus on non-linear relationships between factors such as age and material composition make it an important resource for understanding concrete performance, potentially informing structural engineering practices.

4. Real Data Analysis

We will now assess the effectiveness of the five imputation methods, mean, median, KNN, CART within MICE, and RF within MICE, that were introduced in the preceding sections. Among these classification methods, three are categorized as single imputation methods: mean, median, and KNN. The remaining two methods, CART and RF, belong to the category of multiple assignment methods.

Single imputation methods, such as mean, median, and KNN, replace missing values with a single value without accounting for the variability or uncertainty inherent in the imputation process. For instance, the mean imputation method replaces all missing values in a variable with the mean of the observed values, while the median imputation method uses the median instead. These approaches are straightforward and computationally efficient but may lead to biased estimates and an underestimation of variance, as they do not reflect the natural variability in the data. Similarly, KNN imputation identifies the k nearest neighbors based on a distance metric and replaces missing values with the average (or mode, in the case of categorical variables) of these neighbors. While KNN is more sophisticated than mean or median imputation, it still provides a single estimate for each missing value, thereby ignoring the uncertainty associated with the imputation process.

In contrast, multiple imputation methods like CART and RF within MICE account for the uncertainty in the imputation process by generating multiple plausible values for each missing data point. CART, when used within the MICE framework, employs classification and regression trees to predict missing values based on relationships between variables. This approach leverages the hierarchical structure of decision trees to model complex interactions and non-linear relationships in the data. On the other hand, RF within MICE extends this concept by utilizing random forests, which are ensembles of decision trees. Random forests improve upon single decision trees by reducing overfitting and increasing predictive accuracy through bootstrap aggregation and feature randomness. Both CART and RF within MICE generate multiple imputed datasets, each reflecting a different possible realization of the missing data, thereby providing a more comprehensive representation of the uncertainty in the imputation process.

For a more in-depth understanding of these multiple imputation methods, particularly their implementation within the MICE framework. By evaluating these methods across different scenarios and datasets, researchers can better appreciate their strengths and limitations, ultimately guiding the selection of the most appropriate imputation strategy for a given problem. To assess the efficiency of various imputation techniques, we measure the normalized root-mean-square error (NRMSE) for each variable in the dataset by utilizing the following formula:

$$\text{NRMSE} = \sqrt{\frac{\text{mean}\left((\text{original value} - \text{imputed value})^2\right)}{\max(\text{original value}) - \min(\text{original value})}}$$

where the original value represents the actual data point, and the imputed value is the estimated value following imputation. Next, the Mean of NRMSE is calculated using the following formula:

$$\text{Mean NRMSE} = \frac{1}{n}\sum_{i=1}^{n}\text{NRMSE},$$

where 'n' represents the total number of variables included in the dataset. A reduction in the mean NRMSE value suggests that the imputation methods performed better.

Tables 2–5 display the average values of NRMSE for all datasets at various levels of missing data ratio. The calculations are based on different imputation techniques. In the following tables, each row represents a specific method used, while each column displays the corresponding percentage of missing data. The bold value signifies the lowest Mean NRMSE, indicating the most effective imputation method for that particular dataset. The data demonstrates a correlation between the proportion of missing values and the average NRMSE, showing that as the former rises, so too does the latter.

Table 2. Mean NRMSE for Wine dataset.

Method Used for 2*Data Imputation	**Percent of Imputed Data**				
	10	**20**	**30**	**40**	**50**
Mean Imputation	0.330121	0.47373	0.57983	0.675876	0.761662
Median Imputation	0.337515	0.481549	0.58874	0.690914	0.78079
KNN Imputation	**0.202615**	**0.29687**	**0.366345**	**0.429708**	**0.487329**
RF	0.318207	0.467305	0.578709	0.683289	0.768588
CART	0.267979	0.397142	0.48866	0.575026	0.661217

Table 3. Mean NRMSE for Heart dataset.

Method Used for 2*Data Imputation	**Percent of Imputed Data**				
	10	**20**	**30**	**40**	**50**
Mean Imputation	3.0634	4.40625	5.51389	6.279098	6.977441
Median Imputation	3.073493	4.427702	5.53078	6.303146	6.999989
KNN Imputation	3.200358	4.643237	5.769631	6.578572	7.349563
RF	4.18258	6.05988	7.637643	8.828734	9.857851
CART	4.421592	6.30822	7.727376	8.846299	10.07321

Table 4. Mean NRMSE for HCV dataset.

Method Used for 2*Data Imputation	**Percent of Imputed Data**				
	10	**20**	**30**	**40**	**50**
Mean Imputation	20.32316	28.90146	35.36952	40.8536	45.72216
Median Imputation	20.45093	29.07397	35.5861	41.10005	46.05424
KNN Imputation	**20.11328**	**28.65974**	**35.07199**	**40.51636**	**45.41055**
RF	25.86595	36.93773	45.19835	52.24019	58.51024
CART	25.91126	36.82787	45.07719	52.18786	58.30611

Table 5. Mean NRMSE for Concrete dataset.

Method Used for 2*Data Imputation	**Percent of Imputed Data**				
	10	**20**	**30**	**40**	**50**
Mean Imputation	1.067511	1.509434	1.847977	2.13922	2.396204
Median Imputation	1.149679	1.62484	1.989871	2.303988	2.581315
KNN Imputation	0.437119	0.642189	0.811931	0.980262	1.154321
RF	0.733182	1.085787	1.384627	1.655033	1.939051
CART	0.527952	0.787808	1.013739	1.238703	1.47805

To evaluate the consistency of the results for each imputation method concerning different levels of missing data, we have compared each technique based on Mean NRMSE values The average NRMSE values are arranged in ascending order to achieve this goal, with the lowest value assigned the highest rank. Tables 6–10 present the rankings of imputation methods for different percentages of missing data in four distinct datasets. The tables demonstrate the performance of various imputation methods on specific datasets with a particular level of missing data, enabling a comparison of imputation method consistency across datasets with the same amount of missing data. In each table, the final row presents Kendall's test statistics, which evaluate the agreement among the ranks of imputation methods across varied datasets. Still, the proportion of missing data remains constant. The results indicate that the performance of imputation methods is influenced by the data structure and variables being imputed.

Table 6. Rank of imputation method for 20% missing data for each dataset

Imputation Method	**Wine**	**Heart**	**HCV**	**Concrete**
Missing Percentage of Data: 20%				
Mean Imputation	4	1	2	4
Median Imputation	5	2	3	5
KNN Imputation	1	3	1	1
RF	3	4	5	3
CART	2	5	4	2
Kendall's Statistics:	W = 0.010, Chi-sq = 0.150, p-value = 0.985			

Table 7. Rank of imputation method for 20% missing data for each dataset

Imputation Method	**Wine**	**Heart**	**HCV**	**Concrete**
Missing Percentage of Data: 20%				
Mean Imputation	4	1	2	4
Median Imputation	5	2	3	5
KNN Imputation	1	3	1	1
RF	3	4	5	3
CART	2	5	4	2
Kendall's Statistics:	W = 0.010, Chi-sq = 0.150, p-value = 0.985			

Table 8. Rank of imputation method for 30% missing data for each dataset

Imputation Method	**Wine**	**Heart**	**HCV**	**Concrete**
Missing Percentage of Data: 30%				
Mean Imputation	4	1	2	4
Median Imputation	5	2	3	5
KNN Imputation	1	3	1	1
RF	3	4	5	3
CART	2	5	4	2
Kendall's Statistics:	W = 0.883, Chi-sq = 0.220, p-value = 0.001			

Table 9. Rank of imputation method for 40% missing data for each dataset

Imputation Method	**Wine**	**Heart**	**HCV**	**Concrete**
Missing Percentage of Data: 40%				
Mean Imputation	3	1	2	4
Median Imputation	5	2	3	5
KNN Imputation	1	3	1	1
RF	4	4	5	3
CART	2	5	4	2
Kendall's Statistics:	W = 0.051, Chi-sq = 0.618, p-value = 0.892			

Table 10. Rank of imputation method for 50% missing data for each dataset

Imputation Method	**Wine**	**Heart**	**HCV**	**Concrete**
Missing Percentage of Data: 50%				
Mean Imputation	3	1	2	4
Median Imputation	5	2	3	5
KNN Imputation	1	3	1	1
RF	4	4	5	3
CART	2	5	4	2
Kendall's Statistics:	W = 0.005, Chi-sq = 0.070, p-value = 0.995			

Tables11–14 present the ranking of each imputation technique for a specific dataset across different proportions of missing values. The purpose of this analysis is to assess the consistency in the performance of each imputation technique as the rate of missing data varies. The final row in each table includes Kendall's te

st statistics, which evaluate the agreement among the ranks of imputation techniques for a given dataset with varying levels of missing data.

To assess the reliability of each imputation approach, we have developed the following null and alternative hypotheses. Kendall's W test statistics are used to test the hypotheses related to the degree of concordance in rankings. The coefficient of concordance, denoted as W, ranges from 0 to 1. A value of 0 suggests no agreement in ranking, whereas a value of 1 signifies perfect agreement. The statistical significance of Kendall's W can be determined using a chi-square test with $n-1$ degrees of freedom.

For example, in Table11 (Wine dataset), the W statistic is 0.883 with a p-value of 0.001, suggesting strong agreement in rankings across the different proportions of missing data. Similarly, Table 12 (Heart dataset) shows a perfect agreement with W = 1 and a p-value of 4.99E-4. This indicates that the ranking of imputation methods remains stable regardless of the proportion of missing data for these datasets.

Table 13 (HCV dataset) and Table 14 (Concrete dataset) show similar results, with W values close to 1 and significant p-values, further supporting the robustness of the rankings across datasets. The consistent high Kendall's W statistics across all datasets point to a strong consensus in the performance of the imputation methods. Despite the overall high consistency in rankings, it is important to note that the performance of the imputation methods varied across different datasets. This variation is likely due to factors such as the structure of the data, variable types, and specific study design. For instance, when datasets included time-dependent variables, imputation methods were generally less effective, highlighting the need for specialized techniques for handling such data.

Table 11. Rank of imputation method for Wine dataset for different percentages of imputed data.

Imputation Method	**10%**	**20%**	**30%**	**40%**	**50%**
Dataset Name: Wine					
Mean Imputation	4	4	4	3	3
Median Imputation	5	5	5	5	5
KNN Imputation	1	1	1	1	1
RF	3	3	3	4	4
CART	2	2	2	2	2
Kendall's Statistics:	W = 0.883, Chi-sq = 17.657, p-value = 0.001				

Table 12. Rank of the Performance of Imputation Methods on Heart Dataset Across Various Imputed Data Percentages

Imputation Method	**10%**	**20%**	**30%**	**40%**	**50%**
Dataset Name: Heart					
Mean Imputation	1	1	1	1	1
Median Imputation	2	2	2	2	2
KNN Imputation	3	3	3	3	3
RF	4	4	4	4	4
CART	5	5	5	5	5
Kendall's Statistics:	W = 1, Chi-sq = 20, p-value = 4.99E-4				

Based on the test statistics, it is evident that the W statistics (located in the last row of Tables 11–14) are close to 1, with p-values significant at a 5% level of significance, leading to the rejection of the null hypothesis in all cases. This suggests that the rank of the imputation method is not influenced by the proportion of missing values, indicating an agreement in rankings among different imputation methods. However, varying rankings of imputation methods across different datasets indicate that performance is contingent on factors such as the structure of observations, type of variables, and study design. Our study revealed that imputation methods were found to be less effective when applied to datasets containing time-dependent variables.

For datasets without these variables, the KNN imputation method showed the lowest Mean NRMSE, highlighting its superior performance compared to other methods. It is important to note that there is no one-size-fits-all approach to imputation methods, as performance can vary depending on the specific characteristics of the dataset. As a result, we recommend using a combination of KNN and Mean imputation methods to enhance imputation

Table 13. Rank of imputation method for HCV dataset for different percentages of imputed data.

Imputation Method	**10%**	**20%**	**30%**	**40%**	**50%**
Dataset Name: HCV					
Mean Imputation	2	2	2	2	2
Median Imputation	3	3	3	3	3
KNN Imputation	1	1	1	1	1
RF	4	5	5	5	5
CART	5	4	4	4	4
Kendall's Statistics:	W = 0.962, Chi-sq = 15.4, p-value = 0.004				

Table 14. Rank of imputation method for Concrete dataset for different percentages of imputed data

Imputation Method	**10%**	**20%**	**30%**	**40%**	**50%**
Dataset Name: Concrete					
Mean Imputation	4	4	4	4	4
Median Imputation	5	5	5	5	5
KNN Imputation	1	1	1	1	1
RF	3	3	3	3	3
CART	2	2	2	2	2
Kendall's Statistics:	W = 1, Chi-sq = 20, p-value = 4.99E-4				

accuracy by averaging imputed values from both methods, we can create a more accurate replacement for missing data, improving the overall reliability of the dataset.

The results presented in Tables 11–14 provide valuable insights into the performance of various imputation methods across multiple datasets and varying proportions of missing data. Our analysis consistently shows that the KNN imputation method ranks the highest across all datasets, regardless of the proportion of missing data. This suggests that KNN is a robust method that performs well in diverse scenarios, especially when dealing with datasets that lack time-dependent variables.

In contrast, mean imputation and median imputation generally rank lower, with the mean imputation method showing some improvement at higher missing data percentages, particularly in the Wine dataset. The CART and RF imputation methods exhibit moderate performance in most datasets, with CART slightly outperforming RF in the Wine and Concrete datasets. However, their rankings are often more variable compared to KNN, indicating that these methods may be more sensitive to the structure of the data. In general, the ranking of imputation methods remains consistent across different proportions of missing data, which is supported by Kendall's W statistics, with values approaching 1 in all cases. This strong agreement in rankings suggests that the proportion of missing data does not significantly impact the relative effectiveness of the methods, supporting the reliability of these rankings across varying conditions.

Moreover, the Kendall's W statistics, combined with significant p-values (e.g., p-value = 0.001 for the Wine dataset), allow us to confidently reject the null hypothesis, confirming that the rankings of imputation methods are statistically consistent. This finding highlights that certain imputation methods, particularly KNN, provide stable performance across different datasets, and the rankings are not likely influenced by the percentage of missing data. However, our analysis also reveals that the imputation methods are not universally effective across all types of datasets. For example, in datasets that include time-dependent variables, imputation methods, including KNN,

performed less effectively. This suggests that for such datasets, specialized imputation techniques may be required. The results also indicate that no single imputation method is universally the best. Therefore, a hybrid approach, combining the strengths of different methods such as KNN and Mean imputation, may provide more accurate and reliable imputation outcomes.

5. Conclusion

In this paper, we have provided a comprehensive review of the effect of missing data rates and mechanisms on the performance of various imputation methods, with a particular focus on the MICE-CART framework. Our analysis highlights the complexity of handling missing data and underscores the importance of selecting appropriate imputation strategies based on the nature of the missingness, MCAR, MAR, or Missing Not at Random (MNAR).

We found that while simpler imputation methods like mean and median imputation are computationally efficient, they often fail to capture underlying patterns in the data, leading to biased or imprecise imputations. In contrast, more sophisticated methods like K-Nearest Neighbors (KNN) and Classification and Regression Trees (CART), particularly when implemented through MICE, provide more accurate imputations in scenarios involving complex missingness patterns. The MICE-CART approach, combining the flexibility of multiple imputation with the predictive power of decision trees, emerged as one of the most reliable strategies, especially in cases of high missing data rates and non-random missingness. However, we also discussed the trade-offs involved in using these advanced methods, including computational complexity and the need for careful model tuning.

The effectiveness of imputation techniques, particularly MICE-CART, also depends on the characteristics of the dataset, such as size, variable types, and the missing data mechanism. Selecting an imputation method requires a nuanced understanding of the data's missingness mechanism and the computational resources available. We encourage data practitioners to consider the MICE-CART framework for more robust imputation in complex real-world datasets and to carefully assess the trade-offs between accuracy and complexity when choosing imputation strategies.

Below are some potential future points:

- Extend the study by comparing the performance of MICE-CART with other advanced imputation techniques such as deep learning-based methods as the KNN, or MICE with different algorithms (e.g., Random Forests, Bayesian Ridge Regression).
- Investigate how different characteristics of datasets (e.g., size, dimensionality, correlation structure, and distribution) influence the effectiveness of MICE-CART under varying missing rates and mechanisms.
- Explore the performance of MICE-CART in scenarios where missingness mechanisms are more complex, such as Missing Not At Random (MNAR) or a combination of Missing At Random (MAR) and MNAR. Develop strategies to improve its robustness in these cases.
- Assess the applicability of MICE-CART in real-time or streaming data environments where data arrives continuously and missing values need to be handled dynamically.
- Study the scalability of MICE-CART with large datasets and propose modifications to enhance computational efficiency without sacrificing accuracy.
- Examine how MICE-CART imputed data affects the performance of various machine learning models, including both traditional (e.g., logistic regression, support vector machines) and modern deep learning models (e.g., neural networks).
- Investigate hybrid approaches that combine MICE-CART with other techniques, such as ensemble learning or domain-specific knowledge, to improve imputation accuracy and reliability.
- Develop new evaluation metrics or refine existing ones to better assess the quality of imputed data in terms of preserving statistical properties, relationships between variables, and predictive performance.
- Conduct sensitivity analyses to determine how sensitive the results are to changes in parameters such as tree depth, number of iterations, or sample size within the MICE-CART framework.

- Tailor MICE-CART for specific domains like healthcare, finance, or environmental science, where missing data patterns and implications may vary significantly.
- Explore methods to quantify uncertainty in imputed values produced by MICE-CART and evaluate how this uncertainty propagates through subsequent analyses or predictions.

REFERENCES

1. L. Breiman, Random forests, Machine Learning, vol. 45, pp. 5–32, 2001.
2. L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and regression trees, CRC Press, 1984.
3. M. L. Brown and J. F. Kros, Data mining and the impact of missing data, Industrial Management & Data Systems, 2003.
4. L. F. Burgette and J. P. Reiter, Multiple imputation for missing data via sequential regression trees, American Journal of Epidemiology, vol. 172, pp. 1070–1076, 2010.
5. S. van Buuren and K. Groothuis-Oudshoorn, MICE: Multivariate imputation by chained equations in R, Journal of Statistical Software, vol. 1, pp. 1–68, 2010.
6. L. L. Doove, S. van Buuren, and E. Dusseldorp, Recursive partitioning for missing data imputation in the presence of interaction effects, Computational Statistics & Data Analysis, vol. 72, pp. 92–104, 2014.
7. J. W. Graham, Missing data analysis: Making it work in the real world, Annual Review of Psychology, vol. 60, pp. 549–576, 2009.
8. J. W. Graham, S. M. Hofer, S. I. Donaldson, D. P. MacKinnon, and J. L. Schafer, Analysis with missing data in prevention research, 1997.
9. J. W. Graham, S. M. Hofer, and D. P. MacKinnon, Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures, Multivariate Behavioral Research, vol. 31, pp. 197–218, 1996.
10. J. W. Graham, S. M. Hofer, and A. M. Piccinin, Analysis with missing data in drug prevention research, NIDA Research Monograph, vol. 142, p. 13, 1994.
11. J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques, Elsevier, 2011.
12. G. King, J. Honaker, A. Joseph, and K. Scheve, Analyzing incomplete political science data: An alternative algorithm for multiple imputation, American Political Science Review, pp. 49–69, 2001.
13. A. Kowarik and M. Templ, Imputation with the R Package VIM, Journal of Statistical Software, vol. 74, pp. 1–16, 2016.
14. A. Liaw and M. Wiener, Classification and regression by randomForest, R News, vol. 2, pp. 18–22, 2002.
15. M. Lichman, UCI machine learning repository, Irvine, CA, 2013.
16. R. J. Little and D. B. Rubin, Single imputation methods, Statistical Analysis with Missing Data, pp. 59–74, 2002.
17. R. J. Little and D. B. Rubin, Statistical analysis with missing data, John Wiley & Sons, 2019.
18. P. Royston and I. R. White, Multiple imputation by chained equations (MICE): Implementation in Stata, Journal of Statistical Software, vol. 45, pp. 1–20, 2011.
19. D. Rubin, Multiple imputation: A primer, Statistical Methods in Medical Research, vol. 8, pp. 3–15, 1999.
20. D. B. Rubin, Inference and missing data, Biometrika, vol. 63, pp. 581–592, 1976.
21. D. B. Rubin, Statistical analysis with missing data, Wiley, 1987.
22. J. L. Schafer and J. W. Graham, Missing data: Our view of the state of the art, Psychological Methods, vol. 7, pp. 147, 2002.
23. P. Schmitt, J. Mandel, and M. Guedj, A comparison of six methods for missing data imputation, Journal of Biometrics & Biostatistics, vol. 6, pp. 1, 2015.
24. A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study, American Journal of Epidemiology, vol. 179, pp. 764–774, 2014.
25. R. Somasundaram and R. Nedunchezhian, Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values, International Journal of Computer Applications, vol. 21, pp. 14–19, 2011.
26. J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls, BMJ, vol. 338, 2009.
27. J. Stevens, Outliers and influential data: Our review of the state of the art, Psychological Methods, vol. 7, pp. 147–177, 2002.
28. T. Therneau, B. Atkinson, and B. Ripley, part: Recursive Partitioning and Regression Trees, R Package, 2018.
29. I. R. White, P. Royston, and A. M. Wood, Multiple imputation using chained equations: Issues and guidance for practice, Statistics in Medicine, vol. 30, pp. 377–399, 2011.
30. J. Zhang and H. Aytug, Comparison of imputation methods for discriminant analysis with strategically hidden data, European Journal of Operational Research, vol. 255, pp. 522–530, 2016.