Predicting Missing Data Using Multiple Imputation by Chained Process in Obesity Dataset

Gopika Venu, A. Sai Gnanika* , B. Rajani, A. Yasmeen, K. Kiranmai

Department of Computer Science & Technology, Madanapalle Institute of Technology & Science, Angallu, AP, India

* Corresponding author

doi: https://doi.org/10.21467/proceedings.178.3

ABSTRACT

This paper showcases the effectiveness of multiple imputation (MI) using a chained process (MIC) in imputing missing values in an obesity dataset, highlighting its superiority over single imputation methods due to its computational complexity and lack of familiarity. MIC is implemented and its performance is compared to basic statistical imputation techniques. The results show MIC provides lower error (MSE/RMSE) on numeric variables and higher accuracy on categorical variables versus statistical methods. MIC handles both numeric and categorical missing data well, provided column variables are correlated. By providing a template for applying MIC, this project aims to encourage the use of MI and promote awareness of its benefits over a single imputation for missing data problems in medical research.

Keywords: Mean Squared Error, Root Mean Squared Error, Multiple Imputation by Chained Process, Analysis, Statistical Imputation

I. INTRODUCTION

The problem of missing data is a widespread challenge in machine learning and statistics, especially in fields such as social sciences and computational biology. In social sciences, surveys often suffer from incomplete responses, while in computational biology, experiments such as gene expression analysis can produce missing values due to various reasons like technical errors or limitations in data acquisition methods. Effectively handling missing data is critical because most statistical models and machine learning algorithms depend on complete datasets to generate valid results. In this context, it becomes important to determine the extent of missing data and its underlying causes. To analyze the missing data problem, a common approach involves using a data matrix YYY, which contains the observed values, and a missing indicator matrix MMM, where the entries of the matrix are 1 if the corresponding value in YYY is missing and 0 if the value is observed. The goal is to understand the nature of the missing data and how to handle it in a way that minimizes bias and maintains the integrity of the analysis. One common approach for handling missing data is **Complete Case Analysis (CC)**, where only the observations with no missing values are used in the analysis. This approach assumes that the data are either MCAR or that the missingness mechanism does not introduce significant bias. The main advantage of this approach is its simplicity: by discarding rows with missing values, it ensures that only complete records are included in the analysis, making computations straightforward. However, **Complete Case Analysis** has limitations, especially when the missingness rate is high or when the data are MAR or MNAR. Discarding rows with missing values can lead to biased results if the missingness is not completely random. For instance, if the missing data are more likely to occur in certain subgroups (such as those with certain characteristics), CC can result in a sample

Proceedings DOI: 10.21467/proceedings.178; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-984081-8-1

that is not representative of the population, leading to invalid conclusions. Handling missing data is a critical task in data analysis, particularly in domains like social sciences and computational biology where missing data can arise for a variety of reasons. Understanding the missing data mechanisms—MCAR, MAR, and MNAR—is essential for choosing the right approach to address the issue. While **Complete Case Analysis** can be a viable option when the missingness is MCAR, it becomes problematic when the missing data mechanism is MAR or MNAR or when the missing rate is high. In these cases, more advanced methods such as **Multiple Imputation** or **Maximum Likelihood Estimation** should be considered. Properly addressing the missing data problem ensures more accurate and reliable results in statistical analyses and machine learning models, leading to more robust conclusions.

II. RELATED WORKS

Angelina Hammon [1] used methods of Imputation algorithm. The paper discusses a new statistical method for handling missing data in surveys, specifically when values are missing in a way that depends on the missing values themselves (called MNAR). Hammon's research focuses on multiple imputations of ordinal missing not at random (MNAR) data, a challenging aspect of missing data analysis. By developing methods specifically tailored for ordinal MNAR data, Hammon's work improves the accuracy of analyses in fields like survey research, psychological assessments, and medical studies. This work also enhances the validity of research findings in scenarios where missingness is related to unobserved values. Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan [2] conducted a comparative study on data imputation methods for numeric datasets. They used techniques like KNN, missForest, and Phylopars, as well as predictive mean matching, Bayesian linear regression, non-Bayesian linear regression, and random sample. The study aimed to evaluate the strengths and weaknesses of these methods across different types of numeric data, providing guidance for researchers and data scientists in selecting the most suitable strategy for their specific datasets and analytical goals. Cong Li and Xupeng Ren [3] have developed a machine learning-based framework (MMDIF) to handle missing values in meteorological data. The framework uses Linear regression Algorithm to impute missing values across multiple meteorological variables. This innovative approach addresses the complexities of environmental data, where interconnected variables and complex patterns are common. The MMDIF has potential applications in improving weather forecasting, climate modeling, and other meteorological analyses requiring complete and accurate datasets. D. Cenitta, R Vijaya Arjunan, and Prema K [4] conducted a study using Logistic Algorithm and KNN to impute missing values in medical datasets. They used fuzzy-rough sets and predicted probabilities to generate probability machine learning. The team aimed to improve the completeness and reliability of medical datasets, which are crucial for accurate diagnoses, treatment planning, and medical research. Their work has potential implications for enhancing data-driven decision-making in healthcare settings. Donia Smaali Bouhlila and Fethi Sellaouti's [5] research uses logistic regression methods, multivariate normal models, and chained equations for multiple imputations in time series analyses. They apply this technique to economic or social science research, addressing temporal dependencies in missing values. Their case study demonstrates the application of this technique in specific contexts, contributing to the development of more accurate forecasting models and trend analyses in fields where time-series data is crucial. Elizabeth A. Stuart, Azur, Frangakis, and Leaf's[6] research uses a Linear regression algorithm and a case study to demonstrate the use of multiple imputations by chained equations (MICE) for handling large epidemiological datasets. The study aims to improve the quality of imputed data in public health research by showcasing the practical implementation of MICE in a real-world scenario. This approach contributes to improving the reliability and validity of findings from large-scale health studies that often encounter missing data challenges.

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

Hae-Ran Kim, Ho Young Soh, Myeong-Taek Kwak, and Soon-Hee Han [7] conducted a study using machine learning and field observations to predict Chlorophyll-a concentration in Korea's Coastal Zone. The team used the K-Nearest Neighbors (KNN) Algorithm and multiple iterations to refine their predictions, creating a robust model for seawater analysis. This work is valuable for marine ecologists and environmental scientists studying coastal ecosystems, as accurate Chlorophyll-a predictions are crucial for assessing water quality and marine productivity. Janus Christian Jakobsen, Christian Gluun, Jørn Wetterslev, and Per Winkel[8] conducted a study using Single imputation and the KNN algorithm to identify missing data in clinical research. They identified and categorized missing data using mechanisms like complete at random, random at random, and missing not at random. The study provides insights into missing data patterns in medical studies, contributing to the development of more effective strategies for handling incomplete datasets in clinical trials and other medical research contexts.

Konstantinos psychosis, Loukas ilias, Christos Thanos, and Dimitris Askounis [9]conducted a study using techniques like Sample, KNN, and Missforest to develop a Deep Auto-Encoder (DAE) method with KNN pre-imputation. They improved the architecture and training process of their imputation model, resulting in an XGBoost model that detected suicidal ideation with 85% accuracy using textual features. This work highlights the potential of sophisticated imputation techniques in improving the quality and utility of electronic health records, especially in mental health assessment. The study highlights the importance of addressing missing value imputation in electronic health records. Nadimi-Shahraki, Mohammadi, Zamani, Gandomi, and Gandom [10] conducted a study using the KNN Algorithm to tackle challenges in implementing multiple imputations on large datasets. They focused on selecting suitable models and managing computational limitations, particularly in handling complex missing data scenarios. Their research contributes to the development of more efficient imputation strategies for big data analytics, with potential applications in healthcare, finance, and social sciences. Nwamaka Okafor's[11] research uses methods like K-Nearest Neighbour, Missforest, and Neural Network With Random Weights to fill in missing values in IoT sensor data. They propose variational autoencoder imputation to improve the reliability and accuracy of sensor-based monitoring systems in fields like environmental monitoring, industrial IoT, and smart city applications. The study addresses challenges posed by real-time data collection in IoT environments, such as sensor malfunctions or communication errors, and contributes to the development of more accurate sensor-based monitoring systems.

The study by Peter C. Austin et al.[12] utilized Logistic regression methods to handle missing data in cardiology studies. They highlighted the importance of selecting appropriate variables, determining the number of imputations, handling derived variables, and using predictive mean matching. The tutorial provided by the authors bridges the gap between statistical theory and practical application, enhancing the quality and reliability of data analyses. The authors emphasize the significance of accurate patient data in treatment decisions and outcomes research. Yang Liu and Anindya De[13] conducted a comprehensive epidemiologic study in Namibia, using fully conditional specification multiple imputation (FCS MI) to identify missing data in national blood utilization patterns. Their work demonstrated the potential of advanced imputation techniques in large-scale epidemiological studies, particularly in resource-limited settings. This research has implications for improving health policy decisions and resource allocation in developing countries, as it showcases the potential of sophisticated imputation methods in addressing missing data challenges. Zhang et al.[14] developed a workflow and machine learning-based multiple imputation method to address missing data in the Health and Aging Brain Study-Health Disparities Alzheimer's disease study. Their approach, focusing on the Missforest Algorithm, demonstrated its effectiveness in managing missing data in complex medical research contexts. This work is particularly

valuable for researchers in neurodegenerative diseases, where comprehensive and accurate datasets are crucial for understanding disease progression and developing effective interventions. The method is particularly useful in large-scale health studies. Zuraira Libasin, Ahmad Zia Ul-Saufie, and Hasfazilah Ahmat, Wan Nur Shaziayani's [15] research examines single and multiple imputation methods used in air pollution data to address missing values. The study focuses on the environmental sector and provides valuable insights for researchers dealing with incomplete datasets. The findings contribute to improving the accuracy and reliability of air quality assessments, which are crucial for public health and environmental policy decisions.

III. SYSTEM MODEL

Design is an essential engineering concept that outlines the intended structure or product, while system design converts requirements into software representations, providing quality evaluations for software.

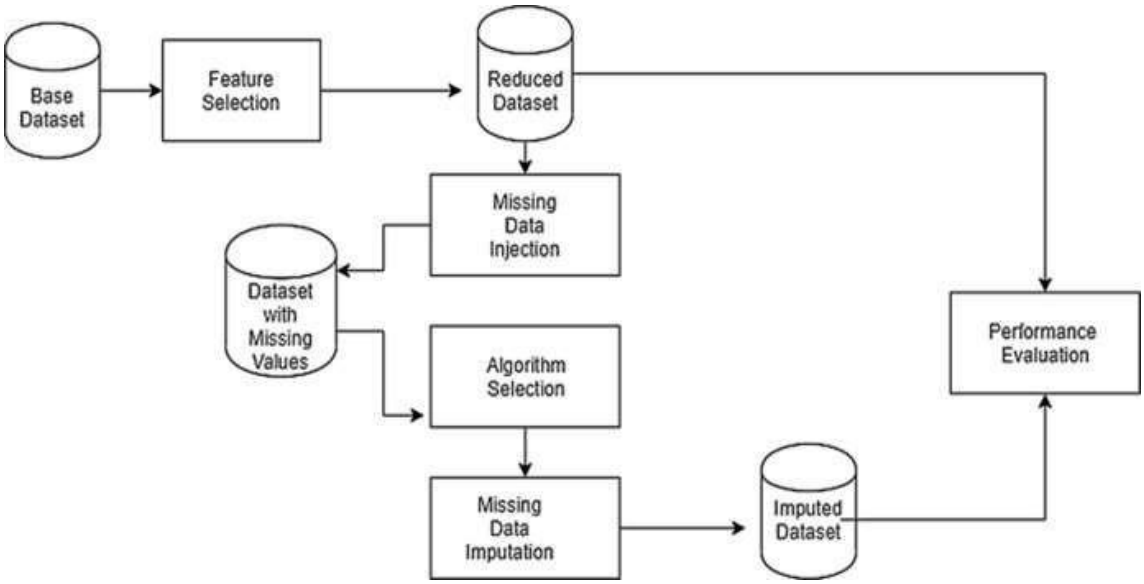

Figure 1: Architecture Diagram

The obesity dataset, which contains 2111 records and 17 attributes, serves as the base dataset for analysis. These attributes likely encompass factors such as gender, age, height, weight, eating habits, physical activity levels, and potentially family history of obesity, providing a comprehensive view of the factors contributing to obesity. However, the dataset may include missing values, a common issue in real-world data collection, particularly in health-related studies, and the quality and completeness of this base dataset significantly influence subsequent steps and the effectiveness of the imputation process. Feature selection is an essential step in data preparation, where techniques like correlation analysis, statistical tests, and machine learning methods are used to identify the most relevant features for predicting obesity levels. The goal is to reduce the dataset's dimensionality by eliminating irrelevant or redundant features while retaining those that offer valuable insights. This process enhances the efficiency of subsequent analyses and improves the accuracy of imputation and modeling. After feature selection, the dataset is reduced to include only the selected features, making it more focused and manageable for obesity analysis. This step helps ensure efficient imputation and analysis by avoiding unnecessary complexity and irrelevant variables, while still retaining crucial information. If the reduced dataset does not have enough missing values for evaluating imputation algorithms, missing data can be artificially injected to simulate different scenarios and test the robustness of the imputation methods. This optional step allows for controlled experiments, enabling the assessment of imputation algorithms under various missingness conditions, which helps to evaluate their reliability and

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

effectiveness. Once the dataset is ready, the next step is algorithm selection, where the appropriate imputation algorithm is chosen. The Multiple Imputation by Chained Equations (MICE) algorithm is typically the primary focus, as it generates multiple imputed datasets by iteratively cycling through incomplete variables and filling in missing values based on other available variables. Other imputation techniques, such as mean/median imputation or k-nearest neighbors, may also be considered for comparison purposes. The imputation process involves applying the selected algorithm to the dataset with missing values, iterating over variables until convergence or a predefined number of iterations is reached. In the case of MICE, each variable with missing values is imputed using other variables as predictors in an iterative process, cycling through all variables with missing data multiple times to generate plausible estimates of missing values. The output of this process is a single imputed dataset or multiple imputed datasets, each representing a complete version of the original dataset without missing values. These imputed datasets are then used for further analysis and modeling, such as obesity prediction or classification. The final step involves evaluating the performance of the imputation algorithm, which is done by assessing metrics like error rates, accuracy, F1-score, and Rubin's rules for combining multiple imputed datasets. The performance of the chosen imputation algorithm is compared to other methods like mean/median imputation or k-nearest neighbors, and the effectiveness of the imputed values is assessed by comparing them against known values or by evaluating the impact of the imputation on subsequent analyses. Common evaluation metrics include Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for numeric variables, as well as accuracy and F1-score for categorical variables. Rubin's rules are often applied to combine the results of multiple imputed datasets, helping to ensure that the final imputation method is both accurate and reliable. The goal is to identify the best imputation technique for the obesity dataset to ensure valid and trustworthy predictions and analyses.

IV. METHODOLOGY

A) Obesity Dataset With Attributes And Class Variables

The dataset, consisting of 2111 records, uses 17 attributes like gender, age, food habits, and activity to estimate obesity levels based on eating habits and physical condition. It categorizes individuals from insufficient weight to obesity type 3. The base dataset, which includes various attributes related to physical characteristics, lifestyle habits, and health indicators, is crucial for subsequent analysis. However, the initial dataset may have missing values, a common challenge in real-world data collection, affecting the effectiveness of the imputation process and the reliability of subsequent analyses.

B) Import Libraries And Exploratory Data Analysis

The text outlines the process of importing key libraries like pandas, numpy, matplotlib, seaborn, and sklearn, loading datasets, converting columns to categories, conducting exploratory analysis, and visualizing attribute relationships. The next step is feature selection, which involves examining all variables within the dataset to identify the most relevant ones for the study's objectives. Techniques used include correlation analysis, statistical tests, and advanced machine learning methods. The goal is to reduce dimensionality by eliminating irrelevant or redundant features while retaining those that provide valuable information for understanding and predicting obesity levels.

C) Correlation Check

Pandas is used to create a correlation matrix, visualize it as a heatmap, identify high correlations for relationship identification, and predict missing values by identifying variables.

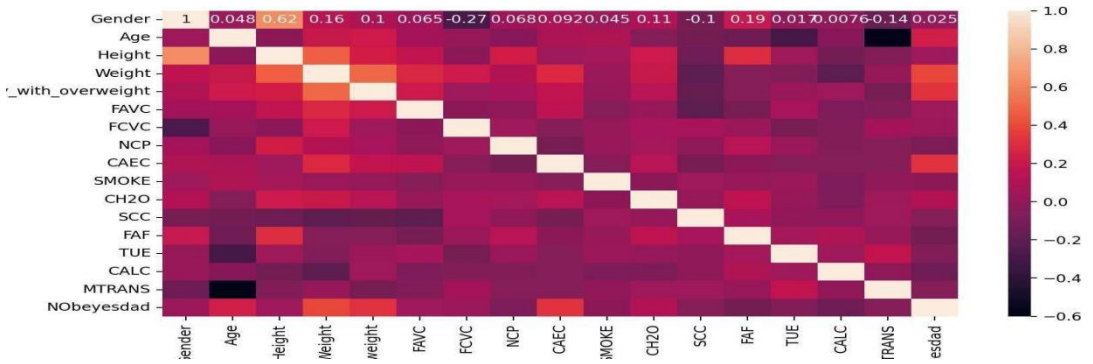

Figure 2:Correlation Check between the variables

D) Statistical Imputation

Mean and mode imputation replace missing numeric and categorical variables, while a simple substitution method establishes a baseline for performance comparison against MIC. The process of missing data imputation in MICE involves a complex iterative procedure. The algorithm starts with simple imputation for all missing values, such as imputing the mean or drawing random samples. It then forgets the imputed values for one variable and regresses it on all other variables in the imputation model. The missing values are replaced with predictions from this regression. This process is repeated for each variable with missing data until convergence or a specified number of iterations is completed.

E) MICE Imputation

The model uses the IterativeImputer from the fancyimpute library to address missing data by incorporating other variables, cycling through incomplete variables, rounding imputed values, and leveraging correlations. The imputation process generates imputed datasets, which can be multiple complete datasets for MICE, representing a plausible estimate of the full dataset without missing values, or a single imputed dataset for simpler methods. These imputed datasets serve as the basis for further analysis and modeling related to obesity prediction or classification.

F) Performance and Accuracy Check

The final step in the methodology is performance evaluation, where the effectiveness of imputation techniques is assessed. This involves comparing imputed values against known values or assessing the impact of imputation on subsequent analyses. Metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are used for numeric variables, while accuracy and F1-score are used for categorical variables. Rubin's rules are often applied for multiple imputation methods like MICE, providing a comprehensive assessment of imputation effectiveness. Comparisons between different imputation techniques are also made.

FLOW DIAGRAM

The flow diagram demonstrates a multi-step process for addressing missing data values using mean imputation and regression modeling as an iterative method.

Step 1: Simple mean imputation is a technique where missing values are temporarily replaced with the mean of available values.

Step 2: The placeholder value is used to set the mean value for one variable to missing.

Step 3: The regression model is created by combining observed variable values with the re-introduced missing value as the target and using other variables as predictors.

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

Step 4: The missing value is predicted using the regression model developed in Step 3.

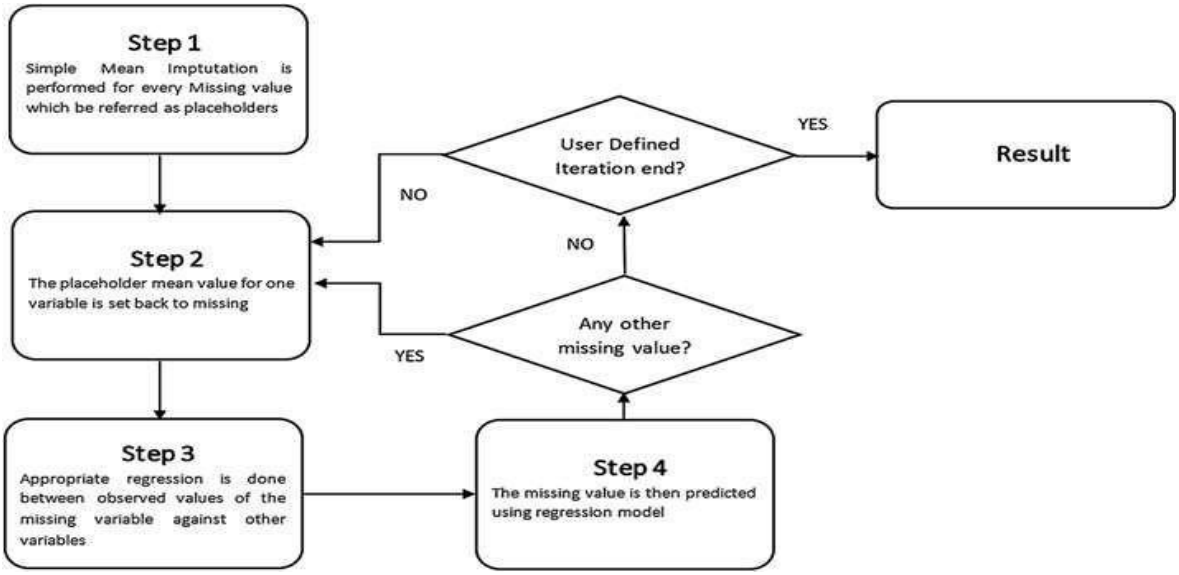

Figure 3: flow diagram of MIC Imputation

The process checks if any missing values are present and repeats steps 2 through 4 for the next variable. If no missing values are found, it checks if the user-defined iteration end condition is met. If not, it continues with another variable with a placeholder mean value. Once the iteration end condition is met, the result with all missing values is obtained. This iterative approach improves the quality of imputations compared to simple mean imputation alone.

V. RESULT ANALYSIS

Linear Regression:

Linear regression is a statistical technique that predicts a continuous target variable through a linear relationship between input features and output variables. It helps understand the relationship between variables by identifying the strength and direction of the relationship. Once trained, linear regression models can predict target variables for new data instances based on input features, making them interpretable. The output includes weights assigned to each input feature, and the constant term represents the predicted value when all input features are zero. The model's performance is assessed using metrics like MSE, RMSE, and R-squared.

Logistic Regression:

Logistic regression is a statistical method for binary classification, predicting the likelihood of an instance belonging to a positive class, offering benefits like classification, probability estimation, and interpretability. The output of a logistic regression model includes the logistic equation represents the weights assigned to each input feature. The logistic equation's constant term and predicted probabilities are utilized to determine the probability of a positive class in new data instances. The model can categorize instances into positive or negative classes by setting a threshold of 0.5 to predicted probabilities. The model's classification performance is evaluated based on accuracy, precision, recall, F1-score, and AUC-ROC.

Ridge Regression:

Ridge regression is a linear regression method that uses a regularization term to address multicollinearity and overfitting issues. It is used when input features are highly correlated or there are more features than observations. It reduces the impact of correlated features by shrinking coefficients toward zero and adding a penalty term to the cost function but with additional components and the model coefficients are decreasing towards zero when compared to ordinary least squares regression. The constant term in a linear equation. The model predicts target variable values for new data instances using input features and learned coefficients. The model's training and test data performance is evaluated using metrics such as MSE, RMSE, and R-squared.

Table 1: Comparison of MSE, RMSE, Accuracy for MIC & Statistical Imputation

	Statistical Imputation	MIC
MSE Height	0.0025	0.0010
MSE Age	13.0656	6.9281
RMSE Height	0.0505	0.0322
RMSE Age	3.6146	2.6321
Accuracy	93	95

The result above shows that imputation with MICE is constant. Produces lower MSE and RMSE values than the Statistical method for columns' Height and Age. As is known, the lower the MSE and RMSE values, the imputation value is closer to the actual value. Then the calculation of the accuracy value in Family. History also shows a better accuracy score which is reaching 96%.MICE successfully imputed well for all types of columns namely numeric (Height, Age) and categorical (Family History).

VI.CONCLUSION:

This project demonstrates the effectiveness of Multiple Imputation by Chained Equations (MICE) for handling missing data in an obesity dataset. The results show that MICE outperforms basic statistical imputation techniques like mean/mode imputation for both numeric and categorical variables. For numeric variables, MICE achieved lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) compared to statistical imputation. Specifically, for the "Height" variable, MICE had an MSE of 0.0010 and RMSE of 0.0322, versus 0.0025 and 0.0505 for statistical imputation. Similar improvements were seen for the "Age" variable, with MICE achieving an MSE of 6.9281 and RMSE of 2.6321, compared to 13.0656 and 3.6146 for statistical imputation. The study demonstrates that Multi-Input Correlational Imputation (MICE) outperforms single imputation methods in handling missing data in medical and healthcare research. MICE achieves 95% accuracy for categorical variables, surpassing statistical imputation's 93% accuracy. This is due to its ability to leverage correlations between variables, despite its higher computational complexity. The study encourages the adoption of multiple imputation techniques and highlights their benefits in handling missing data problems. The results highlight MICE's superiority in handling missing data, provided the variables are correlated. Future research could explore MICE's performance on diverse datasets and compare it with other advanced imputation techniques.

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

Declarations

Competing Interests

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

How to Cite

Gopika Venu, A. Sai Gnanika, B. Rajani, A. Yasmeen, K. Kiranmai (2025). Predicting Missing Data Using Multiple Imputation by Chained Process in Obesity Dataset. *AIJR Proceedings,* 19-27. https://doi.org/10.21467/proceedings.178.3

REFERENCES

[1] Angelina Hammon, "multiple imputations of ordinal missing not at random data, AStA Advances in Statistical Analysis"(2023)107 pp:671–692.

[2] Anil Jadhav, Dhanya Pramod & Krishnan Ramanathan,"comparison of the performance of data imputation methods for the numeric dataset", 2019, VOL. 33, NO.10,913–933.

[3] Cong Li, Xupeng Ren and Guohui Zhao, "Machine-learning-based imputation method for filling missing values in ground meteorological observation data algorithms" 2023,16, pp 422.

[4] D. Cenitta, R Vijaya Arjunan, Prema K V2, "Ischemic heart disease multiple imputation techniques using machine learning algorithm", Eng. Sci., 2022, 19,pp 262-272..

[5] Donia Smaali Bouhlila and Fethi Sellaouti," multiple imputations using chained equations for missing data in times: a case study", pp 1-33.

[6] Elizabeth A. Stuart, Melissa Azur, Constantine Frangakis, and Philip Leaf, "Multiple imputations with large data sets: a case study of the children's mental health initiative", Am J Epidemiol 2009;169 pp1133–1139.

[7] Hae-Ran Kim, Ho Young Soh, Myeong-Taek Kwak, and Soon-Hee Han, "Machine learning and multiple imputation approach to predict chlorophyll-a concentration in the coastal zone of Korea", Water 2022, 14, 1862.

[8] Janus Christian Jakobsen, Christian Gluun, Jørn Wetterslev, and Per Winkel, "Multiple imputations be used for handling missing data in randomized clinical trials", pp1186 12.

[9] Konstantinos Psychogyios,loukas ilias, Christos Manos, and Dimitris accounts, "missing value imputation methods for electronic health records", IEEE Access, vol 11, pp. 21562- 21574.

[10] Mohammad H. Nadimi-Shahraki, Saeed Mohammadi, Hoda Zamani, Mostafa Gandomi and Amir H. Gandomi, "A hybrid imputation method for multi-pattern missing data: a case study on type ii diabetes diagnosis electronics",2021, 10, 3167.

[11] Nwamaka, U. Okafor, Declan T. Delane, "Missing data imputation on IoT sensor networks: implications for on-site sensor calibration", IEEE SENSORS JOURNAL, VOL. 21, NO. 20, OCTOBER 15, 2021, pp 22833- 22845.

[12] Peter C. Austin, Ian R. White,d Douglas S. Lee, MD and Stef van Buuren, "Missing data in clinical research: a tutorial on multiple imputations", Canadian Journal of Cardiology 37 (2021)1322e1331.

[13] Yang Liu and Anindya De, multiple imputations by the fully conditional specification for dealing with missing data in a large epidemiologic study, Int J Stat Med Res. 2015,4(3):287–295. doi:10.6000/1929-6029.2015.04.03.7.

[14] Zhang, Melissa Petersen, Leigh Johnson, James Hall, Raymond F. Palmer, Sid E. O'Bryant, "A machine learning-based multiple imputation method for the health and aging brain study–health disparities", Informatics 2023,10, 77.

[15]Zuraira Libasin, Ahmad Zia Ul-Saufie and Hasfazilah Ahmat, Wan Nur Shaziayani, "Single and multiple imputation method to replace missing values in air pollution", pp 112-167.