International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *12 August 2023* ***Revised:*** *12 October 2023* ***Accepted:*** *22 October 2023*

Imputation Techniques in Machine Learning – A Survey

Y. Angeline Christobel[1], R. Jaya Suji[2], J. Jeya A Celin[3]
[1]Dean, School of Computational Studies
Hindustan College of Arts & Science
Chennai-603103
angelinechristobel5@gmail.com
[2]Assistant Professor, Department of Computer Science
Hindustan College of Arts & Science
Chennai-603103
jayasuji1981@gmail.com
[3]Professor, Department of Information Technology
Kalasalingam Academy of Research and Education
Krishnankoil-626126
jjeyacelin@gmail.com

Abstract—Machine learning plays a pivotal role in data analysis and information extraction. However, one common challenge encountered in this process is dealing with missing values. Missing data can find its way into datasets for a variety of reasons. It can result from errors during data collection and management, intentional omissions, or even human errors. It's important to note that most machine learning models are not designed to handle missing values directly. Consequently, it becomes essential to perform data imputation before feeding the data into a machine learning model. Multiple techniques are available for imputing missing values, and the choice of technique should be made judiciously, considering various parameters. An inappropriate choice can disrupt the overall distribution of data values and subsequently impact the model's performance. In this paper, various imputation methods, including Mean, Median, K-nearest neighbors (KNN)-based imputation, Linear Regression, Miss Forest, and MICE are examined.
Keywords- Missing data, Imputation, Machine learning.

I. INTRODUCTION

The issue of missing values is a widespread challenge encountered across various domains that work with data. It can give rise to a range of problems, including reduced performance, complications in data analysis, and biased results stemming from disparities between missing and complete data. Additionally, the severity of the missing data problem depends on several factors, including the extent of missing data, the missing data pattern, the fundamental mechanism behind data absence, and the underlying mechanism driving the data's missingness. Many studies have been conducted for imputing missing values. Anil Jadhav et al. conducted a comparison of seven data imputation techniques, which included mean imputation, median imputation, kNN imputation, predictive mean matching, Bayesian Linear Regression (norm), Linear Regression, non-Bayesian (norm.nob), and random sample imputation. The findings of their analysis revealed that the kNN imputation method exhibited superior performance when compared to the other methods [1]. According to the findings of Ahmad R Alsaber et al. in [2], the Missing at Random (MAR) technique demonstrated the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Furthermore, the Multiple Imputation (MI) method, specifically employing the missForest approach, exhibited a high level of accuracy in estimating missing values. Pooja Rani et al. [3], evaluated four imputation techniques—k-nearest neighbor (KNN), multivariate imputation by chained equations (MICE), mean imputation, and mode imputation—utilizing four different classifiers, namely Naive Bayes (NB), support vector machine (SVM), logistic regression (LR), and random forest (RF). The objective of their study was to compare the root mean square error (RMSE) of these classifiers and identify the most effective imputation method. The results indicate that the MICE imputation method outperformed the other imputation techniques in this context. Vikesh Kumar Gond et al. [4] discussed imputation techniques and compared the merits and drawbacks. Emmanuel et al. [5], introduced and assessed two distinct methods: the k-nearest neighbor approach and an iterative imputation technique known as "missForest," which harnesses the power of the random forest algorithm. They conducted an evaluation using two datasets—the classic Iris dataset and a novel dataset concerning power plant fan performance. In these datasets, they intentionally introduced missing values at rates ranging from 5% to 20%. Their findings demonstrates that both missForest and the k-nearest neighbor method are proficient at effectively handling missing values. Md. Kamrul Hasan et al. [6], have curated a total of

191 articles published between 2010 and August 2021 for review, employing the widely recognized Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology. Within this review, they condense these articles, outlining pertinent definitions, theories, and analyses, all of which are pivotal in constructing a precise decision-making framework. Additionally, they delve into the evaluation metrics utilized for MVI techniques and delve into their applicability within Machine Learning-based classification models. Thomas, T et.al [7], research findings reveal that the most frequently suggested Multi-View Imputation (MVI) methods are clustering and instance-based algorithms. Among the various evaluation metrics utilized in these studies, the Percentage of Correct Prediction (PCP) and Root Mean Square Error (RMSE) stand out as the most commonly employed measures. Sebastian Jäger et. Al[8], performed an extensive series of experiments involving a diverse array of datasets characterized by heterogeneous data and realistic missing data scenarios. These experiments entail a comparison between cutting-edge deep learning methodologies and conventional Machine Learning (ML) imputation techniques. The evaluation of each imputation method encompasses an assessment of the imputation quality as well as the influence of imputation on the performance of downstream ML tasks. Youngdoo Son et al. [9], assessed the accuracy of height estimations derived from anthropometric data using three distinct imputation methods. Across various levels of missing data, the support vector machine consistently exhibited the highest accuracy. In a study by Carol M. Musil et al. [10], a thorough evaluation of five different strategies for handling missing data was carried out. These strategies included listwise deletion, mean substitution, simple regression, regression with an error term, and the Expectation Maximization (EM) algorithm. The researchers assessed the influence of each method on descriptive statistics and correlation coefficients. Their analysis encompassed both the imputed data subset (n = 96) and the complete sample (n = 492) when the imputed data were incorporated into the analysis. While the study identified limitations in all the methods investigated, the findings clearly suggest that mean substitution was the least effective approach. In contrast, both regression with an error term and the EM algorithm showcased superior performance, yielding estimates that closely approximated the characteristics of the original variables.

II. MISSING DATA MECHANISM

Missing data mechanism refers to the process or underlying reason for the absence of certain data values in a dataset. Understanding the missing data mechanism is crucial in statistical analysis and data imputation, as it helps determine the appropriate imputation methods and the potential impact of missing data on study outcomes. Broadly, missing data can be categorized into three main groups:

Missing Completely at Random (MCAR)

In this scenario, the absence of data does not correlate with any discernible factors, whether known or unknown, within the dataset. For instance, values for certain cells are randomly omitted without any systematic pattern or reason.

Missing at Random (MAR)

In this case, the absence of a data point in a cell is related to some known value within the same sample. For instance, in real estate data, a fraudulent builder might choose not to disclose their office zip code, which is a known factor influencing the missingness.

Missing not at Random (MNAR)

In MNAR situations, the missing value in a cell is dependent on the variable itself. For instance, individuals with higher income levels may be less inclined to share their salary information, creating a non-random pattern of missing data.

The choice of the most suitable imputation technique depends on the nature of the missing data and the distribution of the feature under consideration.

A. IMPUTATION METHODS

Imputation is the process of making educated estimations and substituting missing values within a dataset. In situations where errors are identified in the dataset, and these errors are deemed to be irrelevant or uncorrectable, these erroneous values are marked as missing and subsequently replaced with a reasonable estimate. Conversely, when data is originally missing within the dataset, imputation is employed to generate a comprehensive dataset for analytical purposes.

Some popular methods for data imputation are discussed below:

Mean Imputation

Mean imputation is a simple and frequently used method for managing missing data in statistical analysis and data processing. This method entails substituting missing values within a dataset with the mean (or average) value derived from the observed data for that specific variable.

Median Imputation

Median imputation, sometimes referred to as median substitution, is a method employed to address missing values within a specific variable. It entails substituting the missing values with the median value computed from the available non-missing cases of that variable. Similar to mean imputation, median imputation is particularly effective when dealing with data that is missing completely at random

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *12 August 2023* ***Revised:*** *12 October 2023* ***Accepted:*** *22 October 2023*

(MCAR). It offers the advantage of being straightforward to implement and provides a quick way to obtain complete datasets.

Linear Regression Imputation

Linear regression imputation is a method used to handle missing data by estimating missing values based on linear relationships between the variable with missing data and other variables in the dataset.

Step1: Identify the variable with missing data (the dependent variable) and the other variables in the dataset that are used as predictors in the imputation process.

Step 2: Build a linear regression model where the variable with missing data is the dependent variable, and the other variables are independent variables (predictors). This model is estimated using the observed data.

Step 3: Use the estimated linear regression model to predict the missing values for the dependent variable based on the values of the predictor variables.

KNN Imputation

K-Nearest Neighbors (KNN) Imputation is a method that leverages the k-Nearest Neighbors algorithm to replace missing values in a dataset. It operates by calculating the mean value from the 'n_neighbors' closest data points found in the training set and using this mean to impute the missing values. By default, KNN Imputation employs the Euclidean distance metric for this purpose. It is essential to note that KNN Imputation is a distance-based imputation technique, and it necessitates data normalization. Failure to normalize the data can lead to biased replacements for the missing values due to differences in the scales of the data.

Miss Forest

Miss Forest is widely considered one of the most effective imputation methods, particularly when precision is paramount. This iterative imputation technique harnesses the power of the Random Forest algorithm to achieve highly accurate data imputation.

Step 1: Initially, the missing values are addressed by replacing them with either the mean value of the respective columns for continuous data or the most frequent value for categorical data.

Step 2: The dataset is subsequently divided into two segments: the training data, which includes the observed variables, and the missing data, which is set aside for prediction. Both of these sets are input into the Random Forest algorithm. The algorithm proceeds to predict and impute the missing data in the relevant locations. After this imputation process is finalized, one iteration of the process is concluded.

Step 3: The above step is repeated iteratively until a stopping condition is met. This iterative approach ensures that the algorithm continually refines its imputations based on improved data quality in each subsequent iteration. The process continues until a predefined stopping criterion is satisfied. Typically, it takes around 5-6 iterations to attribute the data accurately.

MICE (Multiple Imputation By Chained Equation)

he MICE algorithm, short for Multiple Imputation By Chained Equation, is a versatile imputation technique that can employ various predictive models. In this specific instance, the MICE method utilizes LightGBM for prediction.

The MICE imputation process involves the following steps:

1. Determine the number of iterations (k) and generate k copies of the original dataset.
2. In each iteration, for every column containing missing values, a temporary substitution of those missing values is made using an estimated value, typically the 'mean,' derived from the non-missing values within that specific column. By the end of this step, all the missing values in the dataset should be effectively replaced, completing the imputation process for that iteration.
3. In the specific iteration, for the column you wish to impute (e.g., column A), revert the imputed values back to missing.
4. Construct a regression model aimed at forecasting the missing values within the chosen column, such as column A, by employing other columns (e.g., B and C) as predictive variables. This model exclusively incorporates rows where column A has non-missing values. Column A is designated as the response variable, while columns B and C serve as predictor variables. The model's purpose is to predict and replace the missing values in column A based on the relationships established with columns B and C.
5. Repeat steps 2-4 for other columns, such as B and C, in the same iteration.

Each complete cycle of predictions for columns A, B, and C constitutes one iteration. Perform these iterations for the predefined value of k. As each iteration progresses, the temporary predictions for each column improve iteratively. This continuous refinement of predictions from one iteration to the next gives rise to the term 'chained.' At the conclusion of the kth iteration, the latest predictions for each variable become the final imputations.

III. COMPARITIVE ANALYSIS

When dealing with various imputation techniques to address missing data, each method comes with its own advantages and disadvantages. The choice of the most suitable technique depends on our specific needs. To summarize the insights

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *12 August 2023* ***Revised:*** *12 October 2023* ***Accepted:*** *22 October 2023*

from the above methods discussed, TABLE 1 is created, highlighting their strengths and weaknesses.

TABLE 1: Comparison of imputations

Methods	Strengths	Weaknesses
Mean	Straightforward and easy to implement. It preserves the overall structure and distribution of the data, making it suitable for certain types of analysis.	Mean imputation can lead to a significant loss of variability in the dataset, as all missing values are replaced with the same value (the mean).
Median	Median imputation is robust to outliers in the data.	Median imputation is robust to outliers in the data. In cases where data are not missing completely at random (MCAR) but are missing at random (MAR), median imputation can be less biased than mean imputation since it is less influenced by extreme values.
KNN	KNN imputation can be applied to both numerical and categorical data. When data are missing not completely at random (MAR), KNN imputation can reduce bias compared to simple imputation methods like mean or median imputation.	KNN imputation can be computationally demanding, especially with large datasets. Calculating distances between data points for imputation can lead to increased processing time and resource usage.
Linear Regression	It is effective in preserving the linear structure and patterns in the data. It is well-suited for imputing missing values in numerical variables where the assumption of linearity is reasonable.	Linear regression assumes a linear relationship between the dependent and independent variables. If the true relationship is non-linear, this method may provide inaccurate imputations.
Miss Forest	Miss Forest can effectively handle datasets with a combination of numerical and categorical variables. Miss Forest is robust to outliers, making it suitable for datasets with extreme values or skewed distributions.	Miss Forest can be computationally intensive, especially with large datasets. Building multiple decision trees for each variable with missing values can be time-consuming.
MICE	MICE is a versatile imputation method that can handle various types of data, including numerical, categorical, and mixed data. It can also accommodate data with complex patterns and dependencies.	MICE is computationally intensive, when dealing with large datasets or much iteration. Building multiple regression models for each variable in each iteration can increase processing time and resource usage.

The TABLE 2 shows the data characteristics and missing data mechanisms for different imputation methods.

TABLE 2: Data characteristics and missing data mechanisms of imputation methods.

Imputation Method	Data characteristics	Missing Data Mechanisms
Mean	Continuous and numerical data	MCAR or MAR
Median	Continuous and numerical data	MCAR or MAR
KNN	Continuous, numerical, or categorical data	MCAR or MAR
Linear Regression	Continuous and numerical data	MCAR or MAR, linear relationships
MissForest	Complex datasets with mixed data types	MCAR, MAR, and MNAR
MICE	Mixed data (both continuous and categorical)	Complex datasets with MCAR or MAR

The TABLE 2 lists the imputation methods being compared. Data characteristics describe the types of datasets or for which each imputation method is typically suitable. The suitability of imputation methods can also depend on the quality and characteristics of the dataset, as well as the specific analysis or modeling goals. Missing data mechanisms helps in selecting appropriate imputation methods.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *12 August 2023* ***Revised:*** *12 October 2023* ***Accepted:*** *22 October 2023*

IV. CONCLUSION

The choice of imputation method should be guided by the nature of the data, the missing data mechanism, and the specific research objectives. There is no one-size-fits-all solution, and it is often beneficial to perform sensitivity analyses with multiple methods. For datasets with simple missing patterns, Mean, Median, or Linear Regression imputation may be sufficient and straightforward. For datasets with more complex or mixed data types, consider KNN, MICE, or MissForest for more robust imputation. When dealing with high-dimensional, complex datasets, MissForest and MICE often provide the most reliable results. The best imputation method should be chosen based on a thorough understanding of data and the goals of analysis.

REFERENCES

[1] Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan, "Comparison of performance of Data Imputation Methods for Numeric Dataset", APPLIED ARTIFICIAL INTELLIGENCE 2019, VOL. 33, NO. 10, 913–933 https://doi.org/10.1080/08839514.2019.1637138

[2] Ahmad R Alsaber , Jiazhu Pan , Adeeba Al-Hurban, "Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012 to 2018)", Int J Environ Res Public Health, 2021 Feb 2;18(3):1333 PMCID: PMC7908071, DOI: 10.3390/ijerph18031333

[3] Pooja Rani, Rajneesh Kumar Gujral, Anurag Jain, "Multistage Model for Accurate Prediction of Missing Values Using Imputation Methods in Heart Disease Dataset", February 2021, Innovative Data Communication Technologies and Application (pp.637-653)
DOI :10.1007/978-981-15-9651-3_53

[4] Vikesh Kumar Gond; Aditya Dubey; Akhtar Rasool, "A Survey of Machine Learning-Based Approaches for Missing Value Imputation", 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), DOI: 10.1109/ICIRCA51532.2021.9544957

[5] Tlamelo Emmanuel , Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago and Oteng Tabona, "A survey on missing data in machine learning", J Big Data (2021) 8:140, https://doi.org/10.1186/s40537-021-00516-9

[6] Md.Kamrul Hasan, Ashraful Alam , Shidhartho Roy , Ai shwariya Dutta , MdTasnim Jawad , Sunanda Das, "Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)", Informatics in Medicine UnlockedVolume 27, 2021, 100799, https://doi.org/10.1016/j.imu.2021.100799

[7] Thomas, T. and Rajabi, E. (2021), "A systematic review of machine learning-based missing value imputation techniques", Data Technologies and Applications, Vol. 55 No. 4, pp. 558-585. https://doi.org/10.1108/DTA-12-2020-0298

[8] Sebastian Jäger,Arndt Allhorn, Felix Bießmann, "A Benchmark for Data Imputation Methods", Front. Big Data, 08 July 2021 Sec. Data Mining and Management Volume 4 – 2021, https://doi.org/10.3389/fdata.2021.693674

[9] Youngdoo Son, Wonjoon Kim, "Missing Value Imputation in Stature Estimation by Learning Algorithms Using Anthropometric Data: A Comparative Study", Appl. Sci. 2020, 10(14), 5020; https://doi.org/10.3390/app10145020

[10] Carol M Musil 1, Camille B Warner, Piyanee Klainin Yobas, Susan L Jones, "A comparison of imputation techniques for handling missing data", West J Nurs Res, 2002 Nov;24(7):815-29. PMID: 12428897, DOI: 10.1177/019394502762477004