International Journal of Scientific Research & Engineering Trends
Volume 8, Issue 4, July-Aug-2022, ISSN (Online): 2395-566X

Performance Analysis of Missing Data Imputation Methods

Harmanpreet Singh
Department of Computer Science and Engineering
Punjabi University,
Patiala, India
gillharman1307@gmail.com

Amrit Kaur
Department of Computer Science and Engineering
Punjabi University,
Patiala, India
amrit.tiet@gmail.com

Harpreet Kaur
Department of Computer Science and Engineering
Punjabi University,
Patiala, India
khasria.harpreet@gmail.com

Abstract- Missing value can cause bias and makes the dataset not represent the actual situation. The selection of methods for handling missing values is important because it will affect the estimated value generated. This study aims to introduce basic concepts of missing data to a non-statistical audience, list and compare some of the most popular approaches for handling missing data in practice and provide guidelines and recommendations for dealing with missing data in scientific research. In this paper, we are going to compare mainly four imputation methods to handle missing values- K-Nearest Neighbor Imputation (KNNI), MICE (Multiple Imputation by Chained Equations) using PMM (Predictive Mean Matching) method, Multiple Imputations using Chained Random Forests and Likelihood via Expectation-Maximization algorithm. The difference in the way these methods work causes the estimation results to be different. Performance of the data imputation methods wasanalyzed using Normalized Root Mean Square Error (NRMSE) method. The results suggest that RF and KNN are effective but much slower than the EM and MICE though the difference in efficiency is negligible for limited iterations.

Keywords **–Missing Values, Imputation methods, MCAR, MAR, MNAR, Bayesian, KNN, Multiple Imputations, MICE, PMM, Maximum Likelihood, Expectation-Maximization, Random Forest, Chained Equations, NRMSE.**

I.INTRODUCTION

Incomplete or missing data is most common problem among all the challenges faced by data researchers all over the world. Data mining simply refers to mine the data for useful insights that were previously unknown. The quality of data used for mining plays a significant role for desirable outcomes.Missing values are the qualities of attributes related to data that are missing from the data set which sometimes influence the mined information critically. The problem of learning from inconsistent or incomplete data becomes fundamental in the field of the data mining[1].Missing data in data set presents various threats to data mining that it may significantly contribute for increase in the computational cost, skewed results. Ineffective handling of missing values can reduce the statistical power of the model, cause bias in estimation of parameters or can complicate the study which further leads to invalid conclusions[2], [3].

The best solution to missing data is to avoid it in the first place and thus to prevent the missing values in data by enhancing the data collection methodology. But it is not as simple as it in theory, for real problems it can comprises the redesign of whole data collection process from scratch. Effective way to deal with the missing data requires an essential task to observe first i.e., to determine the pattern and type of the missing instances which can be the deciding factor for selection of the technique used to handle the missing data[4].Abasic way to handle missing data is simply ignore the records that contain the missing values but it can produce biased result if the size of sample is limited or a significant quantity of data is missing. To minimize the effect of the loss of data several imputation techniques are used. These techniques can be statistical such as mean mode substitution or can be in the form of machine learning algorithms like KNN or K-means etc. Maximum likelihood and expected maximization are the other some model-based techniques. This article explains mechanism of missing data and a comparisonbetween the fourimputation techniques such as KNN. MICE, Random Forest and EM for these mechanisms i.e., MAR, MCAR and MNAR etc.This study aims to provide valuable insights into the performance of some popular methods of imputation under realistic conditions.

II.LITERATURE SURVEY

Missing Data is one of the most frequent data quality problems. Reasons for missing data can be many such data might be accidentally not recorded, loss of data in transmission, intentionally left blank by the users or integration errors. Researchers from several areas have been adding to the arsenal of techniques used to impute missing data throughout the course of the last few decades.Theoretical foundation for the missing data imputation laid firstly by Statisticians by describing

missingness mechanisms and proposed different approaches to handle missing values[5], [6], [7].
Missing values can be handled by various methods.In which, some of the traditional methods are such as deletion (pairwise or listwise) and the other one is imputing the missing value with the predicted value. There exist other advanced model-based methods like multiple imputation, maximum likelihood and Bayesian simulations etc. One other class is also mentioned by Rubin & Little i.e., Weighting Procedures[8] used to reduce bias in complete case analysis for nonresponsive and noncoverage[9].

The most fundamental technique employed by earlier scholars is mean imputation. The mean of the attribute's non-missing values is used to fill in the gap left by the missing value. As it does not preserve the correlation among the features, most of the researchers not consider mean substitution as a good solution[10].

k-nearest neighbour (KNN) imputation is another effective approach to impute missing data byidentifying the k-nearest neighbours, within the dataset which are the most similar to the missing record among all by using the Euclidean distance (number of k is specified by the user).
Maximum Likelihood, also referred to as FIML (Full-Information Maximum Likelihood), Multiple Imputations, and EM Algorithm are some of the contemporary missing data approaches[11].Multiple Imputation acknowledges the uncertainty by generating m number of complete data sets, each one with unique estimate of the missing values. The m complete data sets are then analysed individually using standard statistical procedures and these sets are pooled together yield a single dataset[12].

In Maximum Likelihood, the parameters are estimated by maximizing the likelihood of the observed data. In cases where simple solution to solutions to the maximising of a probability are not feasible, an iterative method is used to determine the estimate of a parameter (such as the means and covariance matrix), also known as the EM algorithm suggested by Dempster, Laird, and Rubin[13].The theory of the EM method for missing data analysis is presented by Little and Rubin[14] and Schafer[15], both of whom assume multivariate normal data[16].

Each method has its own limitations and advantages. To select an efficient method sometime can be quite problematic as it depends on the size, missingness and type of the data. In addition, researchers often have difficultyto determine the proper method to handle the missing values. Numerous studies have been conducted to compare the effectiveness of various techniques and algorithms to conclude the optimum solution to missing data. In comparison, KNN is widely used traditional method for single imputation[17]. Paul argues that in case of software availability, the likelihood estimation is preferable over the multiple imputation [18].

MISSING DATA MECHANISMS

According to Rubin[5], there are three mechanismsunder which missing data can occur: missing at random(MAR), missing completely at random (MCAR), andmissing not at random (MNAR). To understand missingdata mechanisms[12], we partition the data matrix Y intotwo parts: the observed part (Y_{obs}) and the missing part(Y_{mis}). Hence, Y = (Y_{obs},Y_{mis}).
Missing at Random (MAR)
Rubin defined MARto be a condition in which the probability that data aremissing depends only on the observed Y_{obs}, but not onthe missing Y_{mis}, after controlling for Y_{obs}.The MAR values are those random missing values where missing pattern depend or can be observed via some known variable [19].
$P(R|Y, \xi) = P(R|Y_{obs}, Y_{mis}, \xi) = (R|Y_{obs}, \xi)$ (1)
Missing Completely at Random (MCAR)
MCAR is a special case of MAR. It is a missing datacondition in which the likelihood of missingness dependsneither on the observed data Y_{obs}, nor on themissing data Y_{mis}.The missing value said to MCAR if the missing data items are independent of both the observable and non-observable variables or parameters that means missing entirely at random i.e., maximum level of randomness [20].
$P(R|Y, \xi) = P(R|Yobs, Ymis, \xi) = (R|\xi)$ (2)
Missing Not at Random (MNAR)
It occurswhen the probability of missing depends on the missingvalue itself.This mechanism occurs when values that are missing depend upon the other missing values so that researchers cannot use the available data for approximation of missing data, also described as non-ignorable case [21].

III. METHODOLOGY

This section describes procedure followed for analyzing performance of theimputation methods. This study is to analyze performance of imputationmethods that includes: **1.Single Imputation Methods**-K-Nearest Neighbor Imputation (KNNI) (b) Multiple Imputation Methods: MICE (Multiple Imputation by Chained Equations) using PMM (Predictive Mean Matching) method, Multiple Imputations using Chained Random Forests and c) Imputation using Likelihood using Expectation-Maximization Algorithm.The criteria used to compare these methods are the Normalized RootMean Squared Error (NRMSE). For the purpose of this analysis, wehave used R and RStudio as a tool for data imputation, data manipulation,and analyzing performance of different imputation methods andeach imputation method is computed in base R by using VIM, missRanger, mice, norm and Metrics etc. R packages.
2.KNN (K-Nearest Neighbor)
KNN is a method based on finding the nearest neighbor in terms of Euclidian distance.The k nearest neighbor

imputation method values from similar records in the given dataset is used to substitute the missing values[22]. The value to be imputed is calculated by finding the nearest values in the training set and the average of these nearest points is used to fill in the missing values[23]. The similarity for the nearest points is usually obtained using a distance function[24]. The Euclidean distance function is popularly used to calculate distance. The k-NN defines a set of k nearest neighbors and then for substitution of themissing observations for a given variable, average of its neighbors' observed values is used.

3.MICE

In contrast to the single imputation methods, which substitute the missing observation with a single value, multipleimputations replace the missing value with multiple values. The method, which is due to, is now becoming themost used approach in handling missing values[25]. The multiple imputations use many imputed values to substitutethe missing observations. Using this method, each set of generated imputed values is used to substitute the missing values and obtain a completedata set. Therefore, for any given number of multiple imputations, there is the same number of sets of values tosubstitute the missing observations, resulting in having the same number of complete data sets.

The multiple data setsare analyzed afterward to test the statistical method of interest. The multiple results are then combined to obtain asingle best result[26]. Multivariate imputation by chained equations (MICE), sometimes called "fully conditional specification" or "sequential regression multiple imputation" has emerged in the statistical literature as one principled method of addressing missing data[27]. MICE is a particular multiple imputation[28]. MICE operates under the assumption that given the variables used in the imputation procedure, the missing data are Missing At Random (MAR), which means that the probability that a value is missing depends only on observed values and not on unobserved values[6]. Implementing MICE when data are not MAR could result in biased estimates.

Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression[29]. Random Forest is a popular form of the Decision trees approach, which is a stack of decision trees through bagging which combines multiple random predictors to aggregate predictions and the prediction rule is based on the majority vote or average over all trees[30].

Maximum Likelihood & EM Algorithm

In maximum likelihood parameters are estimated which are likely to be in the resulted data. Likelihood for complete data case for all variables and some of the attributes get computed separately and maximized to estimate parameters in ML. ML is considered as an alternative to Posterior draw based on Bayesian Posterior Distributions.

Expectation-Maximization (EM) is a type of the maximum likelihood technique that can be used to generate a new data set by imputing all the missing values, estimated using maximum likelihood methods[3]. For imputation the missing data in an incomplete dataset Expectation-Maximization Imputation (EMI) algorithm combines the mean and covariance matrix of the dataset[10]. It firstly computes the derived matrix that includes the mean and the covariance values of the dataset, which contains missing values then imputes the missing values through the mean and covariance matrix[31], [32].

Dataset

In this study, we used a publicly available dataset ''Melbourne Housing Market'' from Kaggle Inc.[33]which provides data scientistswith a huge amount of data to train and practice their machine learning models. The author of this post used the original dataset published by Domain.com.au. The dataset includes Address, Type of Real estate, Suburb, Method of Selling, Rooms, Price, Real Estate Agent, Date of Sale and distance from Central Business District (CBD) etc. First, dataset is cleaned to zero missing value. After that we can add some missing values (20% as MAR) by ourself so that we can compare the imputed data with the original one for performance and accuracy. The dataset contains 21 variables and 8887 number observation rows.

Evaluation Criteria

To evaluate the performance of imputation methods there exists different ways such as accuracy, MAE (mean absolute error), and RMSE(root mean square error), SSE (Sum Squared Error) and MSE (Mean Square Error). However, RMSE is one of the most effectiveand widely used performance metric in the imputation. However, we have preferred Normalized version of RMSE (i.e., NRMSE), for the reason that variables have different scales. We have used the difference of maximum and minimum value to normalize the RMSE.Oncethe error for each variable in the dataset is calculated then Mean of these metrics iscalculated for the dataset and is used as a measure the performance ofthe imputation methods.The imputation method that has asmaller Error is said to be more accurate than the estimation method thathas a more significant Error[34], [35], [36].
The formula for calculating NRMSE is given below

$$\text{NRMSE} = \frac{\sqrt{\text{mean}\left(\text{Observed Value-imputed value}\right)^2}}{O_{max} - O_{min}} \quad (3)$$

IV.RESULTS & DISCUSSION

Comparative analysis was conducted to compare the KNNI, MICE, Random Forest and EM methods.In the analysis process, the values obtained from the simulations on the dataset were averaged andcombined into a comparison table as in
Table .Table Iis a table of the average results of imputation with the MAR, MCAR and MNAR mechanism. The MARmechanism is a situation where missing values occur randomly and can be observed with the help of known variables. The MAR mechanismusually occurs due to non-response to a survey.

Table 1 NRMSE for MICE KNN RF & EM

Imputation Method	NRMSE (20%MAR)	NRMSE (20%MCAR)	NRMSE (20%MNAR)
MICE	1.0153846	0.9461538	0.9615385
KNN	0.9923077	0.8692308	0.7846154
RF	0.7384615	0.6538462	0.6846154
EM	1.0384615	0.9538462	0.9615385

Based in the Table I, for each mechanism i.e., MAR, MCAR and MNAR at 20% of missing data Random Forest (RF) imputations perform slightly better than the other three methods. We found that Mean NRMSE islowest for RF imputation method and so we can conclude that RF imputation method outperforms the other methods yet the slowest among all. KNN comes in second in terms ofNRMSE and EM performs the lowest (also shown in
Fig. 1as graphical). But these results areapplicable only to numeric and ordered variables of the dataset and one must always consider that thereis no universal method to perform best in every situation.

Time taken to impute the data is another main factor that must be considered. Despite the NRMSE is lowest in case of RF but MICE imputation is second fastest to Expectation-Maximization (EM). EM is fastest among others. KNN performs average both in time and efficiency. Form this study we can also conclude that in case of large datasets, it would be more difficult to iterate with methods like RF and KNN and study will undoubtedly affect the performance of these methods.

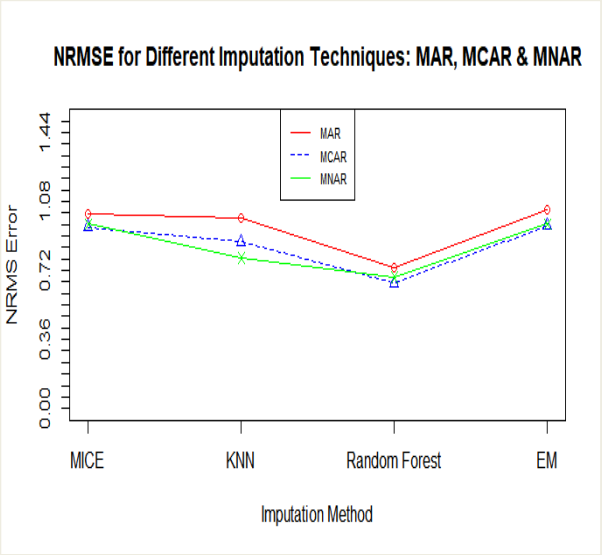

Fig. 1. NRMSE Comparison for MICE KNN RF & EM (20% missingness)

V.CONCLUSION& FUTURE SCOPE

Among the conventional methods this study aims to establish the most reliable method ofimputing missing data. In addition,our sole intention is to provide a general conclusion that is independent fromthe application domain. This study shows that there are various ways that can be used to handle the missing data, the factor we have to consider here if we can afford the time over efficiency that means size and complexity of the dataset affects the performance of these methods.

In conclusion, the study show RF is better option in terms of efficiency but overall MICE and EM imputation is more reliable in terms of time and data size compares to RF and KNN as the difference in error percentage is can be manageable. Researchers can further improve the accuracy by increasing iterations or implementing some domain knowledge in imputation model.One of the limitations of this study is that results are only valid in case of numeric or ordered variable, the nominal and the text data is present but not taken into consideration while evaluation. In future, we willextend our study by including other evaluation metrics and evaluate these methods by increasing the missingness percentage at higher level.

REFERENCES

[1] M. M. Marek´smieja, Ł. Struski, J. Tabor, B. Zielí, and P. Spurek, “Processing of missing data by neural networks.”
[2] S. I. Khan and A. S. M. L. Hoque, “SICE: an improved missing data imputation technique,” J. Big Data, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-

020-00313-w.
[3] H. Kang, "The prevention and handling of the missing data," Korean Journal of Anesthesiology, vol. 64, no. 5. pp. 402–406, May 2013, doi: 10.4097/kjae.2013.64.5.402.
[4] M. Soley-Bori, M. Horn, J. Morgan, and K. Min Lee, "Dealing with missing data: Key assumptions and methods for applied analysis," 2013.
[5] D. B. Rubin, "Inference and missing data," 1976. [Online]. Available: http://biomet.oxfordjournals.org/.
[6] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," Psychol. Methods, vol. 7, no. 2, pp. 147–177, 2002, doi: 10.1037/1082-989X.7.2.147.
[7] S. Jäger, A. Allhorn, and F. Bießmann, "A Benchmark for Data Imputation Methods," Front. Big Data, vol. 4, no. July, pp. 1–16, 2021, doi: 10.3389/fdata.2021.693674.
[8] R. J. A. Little and D. B. Rubin, Statistical analysis with missing data, vol. 3rd editio. 2020.
[9] G. Kalton and I. Flores-Cervantes, "Weighting Methods," 2003. [Online]. Available: www.asc.org.uk.
[10] K. M. Fouad, M. M. Ismail, A. T. Azar, and M. M. Arafa, "Advanced methods for missing values imputation based on similarity learning," PeerJ Comput. Sci., vol. 7, pp. 1–38, 2021, doi: 10.7717/PEERJ-CS.619.
[11] J. W. Graham, "Missing data analysis: Making it work in the real world," Annu. Rev. Psychol., vol. 60, pp. 549–576, 2009, doi: 10.1146/annurev.psych.58.110405.085530.
[12] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers," 2013. [Online]. Available: http://www.springerplus.com/content/2/1/222.
[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. ofthe R. Stat. Soc. Ser. B, vol. 39, no. 1, pp. 1–38, 1977, doi: 10.1115/1.3424485.
[14] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data., 1st ed. New York: John Wiley & Sons, 1987.
[15] J. L. Schafer, Analysis of Incomplete Multivariate Data, 1st Editio. New York: Chapman and Hall/CRC, 1997.
[16] T. D. Pigott, "A Review of Methods for Missing Data," 2001.
[17] K. Manimekalai and A. Kavitha, "MISSING VALUE IMPUTATION AND NORMALIZATION TECHNIQUES IN MYOCARDIAL INFARCTION," ICTACT J. SOFT Comput., p. 3, 2018, doi: 10.21917/ijsc.2018.0230.
[18] P. D. Allison, "Handling Missing Data by Maximum Likelihood," 2012.
[19]Y. Fu, H. Liao, and L. Lv, "A comparative study of various methods of handling missing data in unsoda," Agric., vol. 11, no. 8, Aug. 2021, doi: 10.3390/agriculture11080727.
[20] S. Rawal, S. C. Gupta, and M. S. Singh, "Predicting Missing Values in a Dataset: Challenges and Approaches," 2017.
[21] F. V Nelwamondo, S. Mohamed, and T. Marwala, "Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques."
[22] M. B. Mohammed, H. S. Zulkafli, M. B. Adam, N. Ali, and I. A. Baba, "Comparison of five imputation methods in handling missing data in a continuous frequency table," AIP Conf. Proc., vol. 2355, no. May, 2021, doi: 10.1063/5.0053286.
[23] M. Kuhn and K. Johnson, Applied Predictive Modeling with Applications in R, vol. 26. 2013.
[24] R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," Int. Stat. Rev., vol. 78, no. 1, pp. 40–64, 2010, doi: 10.1111/j.1751-5823.2010.00103.x.
[25] D. B. Rubin, Multiple Imputation for Nonresponse in Surveys. New York, 1987.
[26] K. M. McKnight , Patrick E. McKnight, S. Sidani, and A. J. Figueredo, Missing Data: A Gentle Introduction. New York, 2007.
[27] M. J. AZUR and P. J. STUART, ELIZABETH A. FRANGAKIS, CONSTANTINE LEAF, "Multiple imputation by chained equations: what is it and how does it work?," Int. J. Methods Psychiatr. Res., vol. 20, no. 1, pp. 40–49, 2011, doi: 10.1002/mpr.
[28] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger, "A multivariate technique for multiply imputing missing values using a sequence of regression models," Surv. Methodol., vol. 27, no. November 2000, pp. 85–95, 2001.
[29] S. ER, "Understanding Random Forest," Analytics Vidhya, 2021. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=Random forest is a Supervised,average in case of regression.
[30] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, A survey on missing data in machine learning, vol. 8, no. 1. Springer International Publishing, 2021.
[31] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," Atmos. Environ., vol. 38, no. 18, pp. 2895–2907, 2004, doi: 10.1016/j.atmosenv.2004.02.026.
[32] T. Schneider, "Analysis of incomplete climate data: Estimation of Mean Values and covariance matrices and imputation of Missing values," J. Clim., vol. 14, no. 5, pp. 853–871, 2001, doi: 10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2.
[33] T. Pino, "Melbourne Housing Market," Kaggle Inc., 2016. https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market (accessed Jul. 15, 2022).
[34] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance of Data Imputation Methods for Numeric Dataset," Appl. Artif. Intell.,

vol. 33, no. 10, pp. 913–933, 2019, doi: 10.1080/08839514.2019.1637138.

[35] S. P. Mandel J, "A Comparison of Six Methods for Missing Data Imputation," J. Biom. Biostat., vol. 06, no. 01, pp. 1–6, 2015, doi: 10.4172/2155-6180.1000224.

[36] T. Rhaudatunnisa, N. Wilantika, T. Rhaudatunnisa, and N. Wilantika, "Performance Comparison of Hot-Deck Imputation , K-Nearest Neighbor Imputation , and Predictive Mean Matching in Missing Value Handling , Case Study : March 2019 SUSENAS Kor Dataset," no. March, pp. 753–770, 2019.