Missing Data Handling

Dipalika Das [1*], Maya Nayak[2], Subhendu Kumar Pani[3]

[1] Dipalika Das, Asst. Prof., Trident Academy of Creative Technology, Research Scholar, Department of Computer Science and Engineering, Biju Patnaik University of Technology, Rourkela, Odisha, India.

[2] Dr. Maya Nayak, Professor, Trident Academy of Creative Technology (TACT), Bhubaneswar, Biju Patnaik University of Technology (BPUT) Rourkela, Odisha, India.

[3] Dr. Subhendu Kumar Pani, Professor, Krupajal Engineering College (KEC), Bhubaneswar, Biju Patnaik University of Technology (BPUT), Rourkela, Odisha, India.

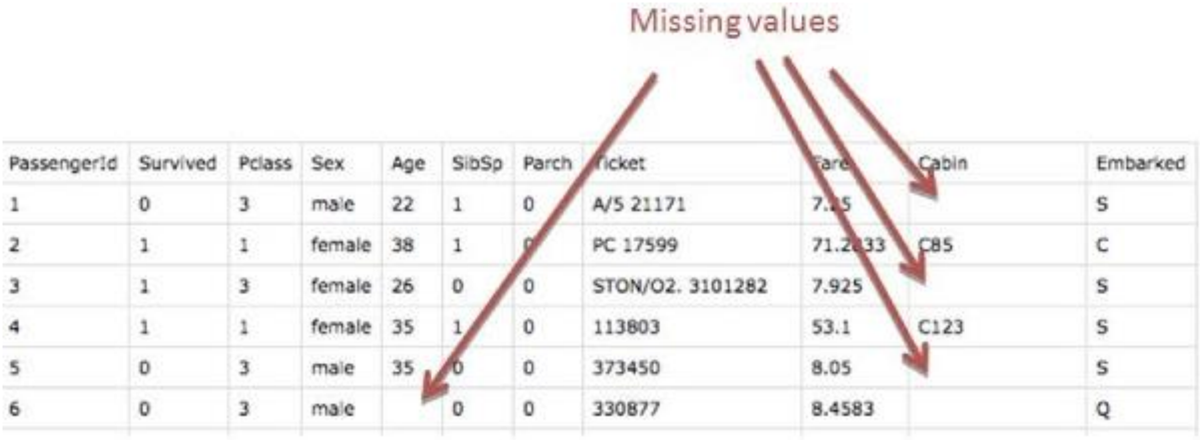
1. Introduction

Over the most recent few years, both the rate at which digital data is being produced and the rate at which computational power is being developed have accelerated dramatically. These enable the extraction of distinctive insights from enormous databases, commonly called "big data." Data analysts understand the challenges various industries face, including healthcare, banking, e-commerce, and finance. By uncovering valuable insights from large datasets, they help organizations navigate complexities and discover opportunities that can lead to better outcomes for businesses and the people they serve. [1, 2]. They emphasize the quality of the data they collect to ensure successful data analysis. The results of data analysis [22] include attribute selection, method selection, sampling approach, etc. Although it depends on many criteria, a critical dependence is on effectively managing missing values [3]. It is common practice to employ machine learning and data mining to derive predictions from massive datasets. Unless the data used to train the algorithms contain errors, these algorithms' predictions are typically accurate. Refining data for the purpose of teaching the system represents a fundamental component of the data analysis and mining phase. Normally, the term "data pre-processing" refers to the various steps involved in the data mining process. Practitioners often consider this phase the most challenging aspect of the overall procedure, as it lays the groundwork for effective data utilization. Recognizing its significance is essential for achieving successful outcomes in data-driven initiatives [4]. In many situations, a portion of data is either lost or entered erroneously by a person, leading to inaccurate forecasts. The presence of lost values [20] is the most significant issue when considering the data quality. When the dataset has missing values [21], it can significantly increase the cost of processing, distort the results, and put in trouble the researcher [5].

Similar to the challenges presented by traditional data analysis [9] activities, data loss is a significant issue in data analytics. Effectively managing large volumes of data requires several key prerequisites, including access to high-quality data. However, missing values lead to a decline in data quality. In the majority of cases, big data will have several different kinds of measurement mistakes, as well as outliers and values that are missing. Preparing data and conducting analysis is made more difficult due to these challenges. Extracting low-dimensional structures from high-dimensional data is crucial in big data analysis, and this task is often necessary. In significant data contexts, traditional statistical methods for imputing missing data usually perform poorly because of noise and the data's high dimensionality. In the presence of missing values within a dataset, implementing various machine learning algorithms, including neural networks and support vector machines, becomes impractical. It is essential to address these missing values before proceeding with analysis to ensure the effectiveness of the algorithms. [6]. Ignoring the observation that has missing values is a straightforward solution to this problem. There usually isn't a significant issue when there aren't many observations with missing data. Nevertheless, doing so results in a considerable loss of information [7] since it removes many observations with absent values. In addition, statistical power and efficiency are diminished [8]. Consequently, it is crucial to effectively utilize dependable imputation methods to address missing data. Maintaining database integrity by accounting for missing data is essential for big data analytics and small-scale data mining operations. This is because complete datasets can be used to make more accurate predictions.

2. Missing Value

Values or data absent for certain variables in the provided dataset are known as “missing data,” this term defines what is meant by the term "missing data." There are certain instances of missing data within the Titanic dataset, which will be exemplified in the following section. It is noticed that some of the numbers in the columns labeled "Age" and "Cabin" are absent from the table. In most cases, the value of a missing variable is represented by the symbol NaN. Not a Number is what this abbreviation stands for.


Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Fig.1 Missing value

For several reasons, some values might not be present; the absence of data within a dataset significantly influences the strategies employed to address missing data. It is vital to comprehend the underlying causes of data loss. Presented below is a list of several common causes:

- Inadequate maintenance may result in the corruption of previously stored data.
- In some fields, observations may not be retained for various reasons. Human error can lead to incorrect recording of values.
- Intentionally, the user did not supply the values.

- A nonresponsive item means the individual may have declined to answer.

Managing the missing values properly is crucial.

- Many machine learning methods won't function if the dataset contains values that are missing
- The statistical analysis may become imprecise due to missing data.

3. Missing Values: Types

Missing values can be of the following types:

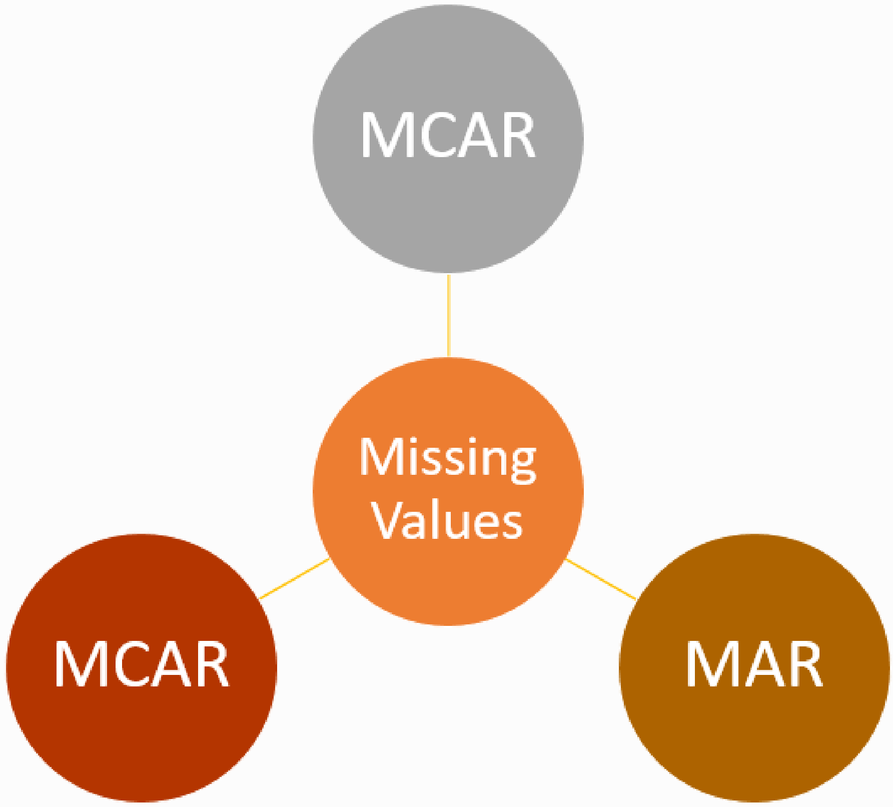


Fig 2. Categories of Missing Values in Datasets

a. Missing Completely at Random (MCAR)

When using MCAR, the likelihood of losing a perception is that some data will remain the same, regardless of the observation. In this instance, the missing data in the provided dataset has no bearing on alternate values, either observed or unobserved (the information that's not recorded). This is because missing data do not correspond to the other values. That is to say, missing numbers are unrelated to the different data in any way. There does not appear to be any pattern.

Missing Completely At Random (MCAR) data can have missing values due to human error, equipment failure, sample loss, or inaccuracies in the data recording process. Take, for instance, the situation at a library where some books have been overdue for some time. Within the computer framework, there are lost values for some of the overdue books. It's possible that the cause was a mistake made by man, like when the library staff forgot to enter the values. Therefore, the lost values of late books don't have a connection to any of the other variables or data in the system. Due to the unusual nature of the occurrence, no assumptions should be made about it. Statistical analysis can maintain objectivity when such data are used, which is a significant advantage.

b. Randomly Missing (MAR)

Missing data at random (MAR) refers to scenarios in which the absence of specific values can be attributed to other variables with complete information. This concept underscores the importance of considering the relationships among variables when analyzing datasets that contain missing values. Put differently, you can use the variables about which you have full knowledge to explain the missing values. In this specific case, none of the observations lack data. The missing values in the data only occur in specific subsamples and generally follow a particular pattern.

For example, the survey data shows that while everyone responded with "Gender," many of those who identified as "female" were missing "Age" information. This is because most women are reluctant to disclose their age. Thus, the observed value or data alone determines the likelihood of missing data. The variables “gender” and “Age” are connected in this instance.

The "Gender" variable is an explanatory factor for the missing values observed in the "Age" variable; however, it is essential to note that the missing values cannot be predicted. Let's say a library polls its patrons on overdue books. In addition to asking about respondents' genders, the survey also asks how many books are past due. Assume that fewer men than women answered the question and that women make up the majority of poll participants. So, there has to be another element involved, and that is gender.

This clarifies the reason for the data gap. In this specific instance, bias may result from the statistical analysis. An estimate of the parameters free from bias can only be obtained by modeling the lost information.

c. Missing Not at Random (MNAR)

The collected data helps us assess whether missing values are appropriate. When missing data exhibits a specific pattern or structure that other observable data cannot explain, it is classified as Missing-Not-At-Random (MNAR). If the instances of missing data do not qualify as Missing-At-Random (MAR) or Missing-Completely-At-Random (MCAR), they can be categorized as MNAR. People may be reluctant to divulge the necessary information, which could lead to this happening. It's possible that confident respondents won't answer specific questions in the survey you're conducting. To help clarify, let's consider a hypothetical poll conducted by a library, which asks respondents for the library's name and the number of overdue books they have. In this scenario, most people with no overdue books will likely participate in the survey. Conversely, those with more overdue books are less inclined to respond to the poll. Therefore, in this particular scenario, the number of books missing overdue values is contingent on the individuals with a more significant number of overdue books. One such illustration would be the likelihood that people with lower earnings will withhold some information while responding to a survey or questionnaire. In the case of MNAR, the statistical analysis may also produce biased results.

4. Understanding Patterns in Missing Values

The missing data model illustrates the data structure regarding available and unavailable values. It is not to be mistaken with a missing data mechanism specifying possible correlations between data and an individual's propensity for missing values. Instead, this is to be distinguished from such a mechanism. Patterns are used to characterize the locations of the holes in the data, while mechanisms are used to explain why the values are absent.

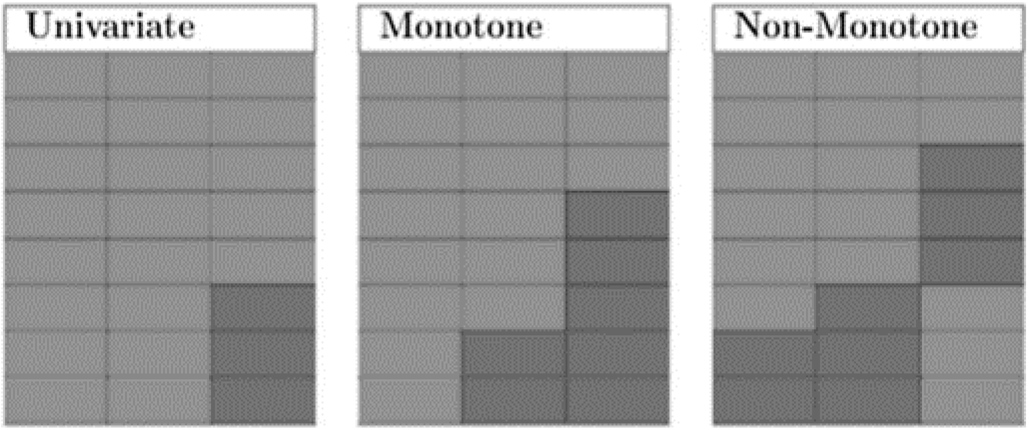


Fig. 3 Missing data patterns are represented.

Univariate: A lost information design is univariate when a single attribute has lost information. This design is uncommon in various sectors and disciplines, leading to several exploratory findings [4].

Monotone: The concept of lost information design is classified as monotone when the various factors can be systematically organized. This design is often associated with a longitudinal approach, in which participants may drop out of the study and fail to return. This information design is effortlessly discernible and less demanding than bargaining with designs among the lost values. [10].

Non-monotone: This is typically a lost information design in which the unavailability of one attribute value does not influence the loss of other alternate factors [11].

5. Approaches to Missing Data Handling

A. Deletion

When carrying out analysis using this method, every entry that does not include all of the required values is eliminated and thrown away. Because there is no need to attempt to evaluate value, deletion is generally seen as the strategy with the least amount of complexity. The research conducted by Small and Rubin [12] underscores the critical importance of managing deletions in data analysis, as these may introduce bias, especially in scenarios where missing data is not missing at random. Two primary approaches for addressing missing data are pairwise deletion and listwise deletion. Both of these methods are explained in detail here.

a) List-specific or Case-specific deletion

When using list-specific removal, it considers every case missing with one or more values and deletes them. When assessing data, listwise deletion has evolved to become the option selected by default in the vast majority of statistical software packages. On the other hand, listwise analysis produces biased conclusions [12], presuming that the data are not MCAR. The list-wise deletion approach, however, might be a good choice if the data samples are big enough and it can be shown that the MCAR assumption has been met. If the MCAR assumption was unmet or the sample size was insignificant, the list-specific removal procedure is considered the only optimal course of action. When there are many cases to discard, and the list-specific removal procedure is adopted, some essential information may be lost. This is particularly true when the number of cases to be deleted is large.

Benefits:

- Suitable for statistical tests (e.g., SEM, multi-level regression, etc.)
- Both the parameter estimations and their standard errors are objective in the case of MCAR.

Cons:

- It will provide higher standard errors than other, more advanced techniques that will be covered later.
- Biased estimations may result from your listwise deletion if the data is MAR instead of MCAR.
- Listwise deletion may not always produce better estimates than complex techniques like regression analysis.

b. Pairwise deletion

The usage of pairwise deletion is a strategy that can be implemented to reduce the amount of information lost when carrying out deletions in a list-wise fashion. The reason for this is that pairwise deletion is done in a way that minimizes any possible losses that list-wise deletion can cause. Values are deemed to be missing exclusively when a particular data point is necessary to ascertain their status. [14]. Pairwise removal has the disadvantage of potentially producing an inter-correlation matrix that lacks positive definiteness. This lack of positivity in the matrix may prohibit additional analysis, such as the estimation of coefficients, from being performed. Last but not least, it is well known that pairwise deletion also produces minimal bias results when applied to MCAR or MAR data [13].

Advantages:

1. Pairwise deletion will produce unbiased, consistent estimates in big samples if MCAR is the actual cause for missing data.
2. When there is little association between the variables, pairwise deletion works better than listwise deletion.
3. When there are substantial correlations between the variables, listwise deletion works better than pairwise deletion.

Cons:

1. Pairwise deletion will produce skewed estimates if the data method is MAR.
2. Calculating coefficient estimates in small samples is often impossible because the covariance matrix may not be positive definite.

Imputation

Imputation is a procedure that involves filling in gaps in data with specific projected values to make up for missing data. In most cases, researchers refer to the data set containing non-missing values to forecast appropriate replacements for the missing values [15]. In the forthcoming sections, this document will analyze several imputation methods frequently employed in the existing body of research.

a. Simple imputation

The simple imputation method involves filling in lost values against each value by imputing them with a quantitatively viable property or a subjective quality shared by all the values for which there are data. Simple imputation for missing data can use various methods, such as the available values' mode, mean, or median. Because of their ease of use and versatility, simple imputation methods are utilized in most research endeavors [16]. This is due mainly to simple imputation methods being straightforward to reference. On the other hand, when used to high-dimensional data sets, straightforward imputation algorithms can potentially give biased or unrealistic conclusions. This method performs poorly, so it is unsuitable for data sets. In addition, the emergence of big data presents a new challenge for data analysis.

b. Single Imputation: Estimated Values through Regression

In this process, missing values are addressed using predictions generated from a regression model or another comparable modeling approach if applicable. This methodology is predicated on the assumption that the patterns observed in the complete dataset will be mirrored in the missing data. It is crucial to acknowledge that this approach may fail to account for the potential differences in relationship patterns between cases with missing and complete data. This oversight could, in turn, introduce a risk of bias in the study.

Additionally, the imputed values tend to have lower variance than the actual observed values because the predicted values do not consider error variance. Introducing simulated "residuals" to reflect this missing error variance can enhance the accuracy of these imputed values. However, subsequent analyses that treat imputed values the same as correctly observed values do not adequately recognize that imputed values are essentially estimates and, therefore, should not be treated as equivalent to known values.

Consequently, such studies often produce standard errors, confidence intervals, and p-values significantly lower than necessary. Despite the drawbacks of these straightforward techniques, they are used in almost every study in some capacity. For example, they are used to complete occasionally absent items on a rating scale with many items. Currently, a range of evidence-based strategies are being investigated [17].

c. Hot-deck imputation

Addressing missing values, commonly called hot-deck imputation, entails comparing missing values with other existing values within the dataset, utilizing various attributes that possess complete values. There exist several iterations of this procedure; however, the variant that considers the inherent variability in missing data identifies a set of cases designated as the donor pool. This donor pool includes identical instances, except for data on some factors. A single case is selected randomly from among those in the donor pool. After that, the data from one of the cases that was chosen at random is used to replace the missing value. Alternatively, instead of selecting a single donor from a group of potential donors, you might decide to replace the donor who is nearest to you [19]. The technique does not consider the varying amounts of information that may be lost. Two hot deck imputation method variants are the Weighted Random Hot Deck and the Weighted Sequential Hot Deck. In the Weighted Random Hot Deck approach, there is no limit to the number of times a donor can be selected; donors are randomly chosen from the donor pool. In contrast, the Weighted Sequential Hot Deck restricts how often a donor can be selected, ensuring that multiple distinct recipients are not paired with the same donor.

Because it produces rectangular data [18], the hot-deck approach is quite familiar among all single-imputation-methods. This is because secondary data analyzers can use the data produced by this method. Furthermore, the approach replaces a missing value without depending on model fitting and avoids user inconsistencies. Compared to strategies that use parametric methods like regression imputation, this approach may be less affected by model specifications. Additionally, the approach lessens bias in regions where respondents did not provide information. The method is often applied in research, but despite its prevalence, the theoretical foundation supporting it is not as robust as that of other imputation strategies.

Sullivan et al. [10] introduced a hot-deck imputation method capitalizing on information from fully observed variables. This methodology enabled a thorough examination of the consequences of data unavailability, specifically across the spectrum of Missing At Random (MAR) to Missing Not At Random (MNAR) situations. The implementation of this method facilitated an assessment of the effectiveness of various imputation strategies. Through simulation, we analyzed both the bias and coverage associated with the estimations generated by the proposed approach. In addition, the findings demonstrated that the methodology functioned most effectively when the outcome was connected with all of the wholly observed values.

The expectation-maximization (EM) technique is an iterative approach to address the challenge of missing values in numerical datasets. This method enhances data analysis by effectively estimating the unknown values through a systematic process. This methodology employs a systematic approach encompassing three essential stages: Impute, Estimate, and Iterate until convergence is achieved. This process effectively resolves the challenges associated with incomplete data. The steps that make up each cycle are called expectation and maximization, respectively. In the expectation maximization approach, the current values maximize the likelihood of all available information, while the expectation step calculates missing values based on observed data. [21] This difference is essential since maximization is often used in place of expectation.

Multiple strategies employing expectation minimization have been suggested to effectively address the challenges of missing data within the study. Rubin et al. analyzed how to manage missing data.[17]. Using a complete information set, they studied the effects of the feeding method on medicated animals and the animals that were not given the drug. The expectation-maximization algorithm was employed and systematically compared with several alternative methodologies, including list-wise deletion, the Bayesian approach, and mean substitution regression—this comparative analysis aimed to evaluate the efficacy and robustness of each method in handling missing data. List-wise deletion was shown to be the method with the least success overall. The authors concluded that the Expectation Maximization approach analyses the particular kind of data they effectively employ. In any case, the fact that real-life datasets were used in the study might have caused the findings to be unique to the dataset's peculiarities and the sampling process, or they may be reflective of predictions based on hypothetical situations.

In various studies, an expectation maximization approach was employed for imputation to effectively address the challenges associated with training Gaussian models on high-dimensional datasets that contain missing values. [23]. This was done to address the difficulty of the Gaussian mixture model. Subsequently, the imputed datasets were evaluated using classification models, significantly improving the basic missing value imputation method. However, implementing the projected maximization strategy incurred substantial expenses due to matrix calculations.

d. Multiple Imputation

The key issue not addressed by a single imputation from a regression model was that the essential variability and uncertainty of the imputed records were not incorporated into the analysis phase. This was the primary unsolved challenge in the usage of single imputation. The method of multiple imputation can be utilized to accomplish this goal. This is a relatively new and active area of research that is still developing. It is too soon to classify it as an all-encompassing generalization. Nonetheless, because the challenging imputation aspect may be separated from the analysis phase, specialized staff and software can be used to create imputed data sets, which can subsequently be examined by comparatively inexperienced users using ordinary tools. This is made possible by isolating the challenging imputation from the analysis. This is particularly useful for datasets with low to moderate missing data, mainly when the missing data is structured, and the data sets are intended for public release [17].

Three requirements must be met before multiple imputations may be considered successful. First, the imputations must accurately reflect all the variables' relationships.

Although the challenging aspect of imputation can be separated from the analysis phase, specialized staff and software can be utilized to create imputed data sets. These sets can then be examined by less experienced users using standard tools. This is made possible by isolating the challenging imputation from the analysis, which would call for applying specialized approaches. In the third step of the process, multiple data sets are compiled to convey to the analysis stage the degree of uncertainty associated with the values that have been imputed. When applying various imputation methods to distinct datasets, the observed values will remain unchanged; however, the resulting imputed values may differ significantly. The ultimate result is not a single file with some missing data but many files, each with all the necessary information. When only a few pieces of information are missing, it is generally recommended to use five imputation replicates. As the amount of missing data increases and uncertainty grows, more repetitions may be necessary.

6. Adapted Imputation Technique for Missing Information

Fig. Depicts the pseudo-code of the improved imputation method's main routine. The dataset with lost values is the input needed for this. Initially, all attributes were subjected to the Spearman correlation test. Next, identify and pair the variables with the highest correlation, i.e., variable X representing the attribute with missing values and variable Y representing the attribute that will be utilized to impute the missing values.

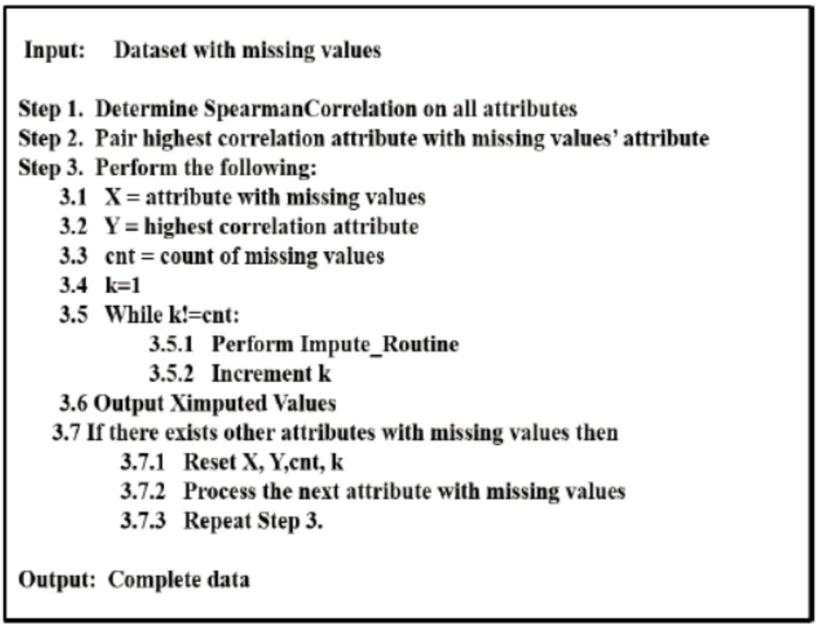
The formula for Spearman correlation:

$$(R) = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Where n is the instance, and d is the rank difference.

The Pearson correlation follows, with X and Y representing the variables and n representing the instance, as seen below.

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{n}(\sum Y^2 - \frac{(\sum Y)^2}{n})}}$$



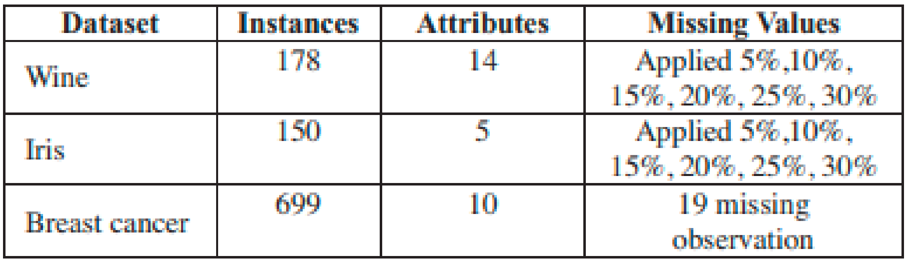
```
Input:    Dataset with missing values

Step 1.  Determine SpearmanCorrelation on all attributes
Step 2.  Pair highest correlation attribute with missing values' attribute
Step 3.  Perform the following:
      3.1   X = attribute with missing values
      3.2   Y = highest correlation attribute
      3.3   cnt = count of missing values
      3.4   k=1
      3.5   While k!=cnt:
                  3.5.1   Perform Impute_Routine
                  3.5.2   Increment k
      3.6 Output Ximputed Values
     3.7 If there exists other attributes with missing values then
               3.7.1   Reset X, Y,cnt, k
               3.7.2   Process the next attribute with missing values
               3.7.3   Repeat Step 3.

Output:  Complete data
```

Fig.4. Modified imputation method.

Table 1 Dataset used



Dataset	Instances	Attributes	Missing Values
Wine	178	14	Applied 5%,10%, 15%, 20%, 25%, 30%
Iris	150	5	Applied 5%,10%, 15%, 20%, 25%, 30%
Breast cancer	699	10	19 missing observation

Fig. 4 displays the accuracy performance results for each classifier. The combined average precision, accuracy, and ROC values were calculated using modified imputation, and the results are shown in Table II. After imputation utilizing the three performance indicators, total performance improved. This generally indicates that classification performance can be enhanced by applying the improved imputation procedure.

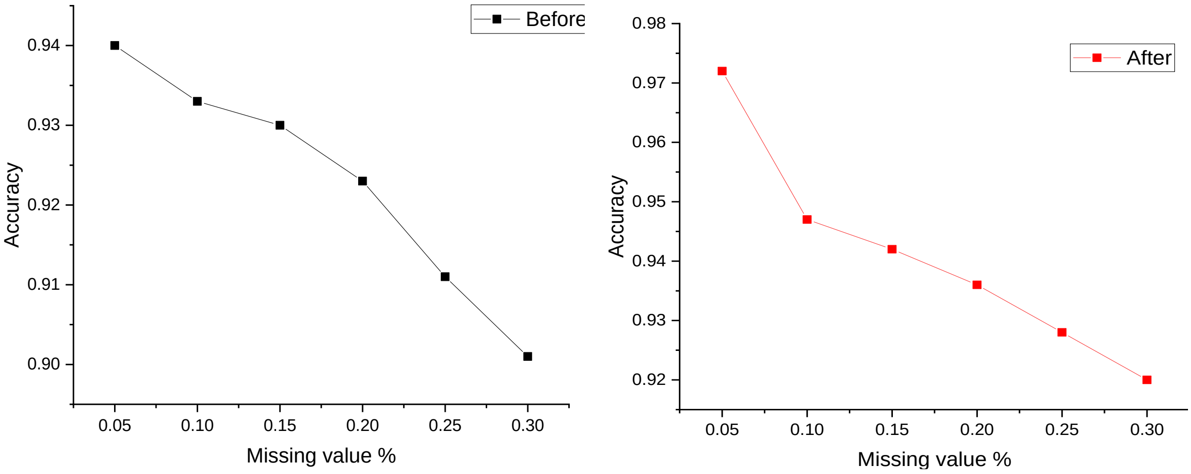


Fig 5. Performance of Classifier

Table 2: Overview of Performance

Method	Before Imputation	Following Imputation
GNB	0.94	0.96
KNN	0.95	0.971

This generally indicates that the improved imputation procedure can enhance classification performance. This information identified the classifier with the best performance after applying modified imputation.

References

1. Chen Y - C, Pattern graphs: A graphical approach to non-monotone missing data, arXiv. 2004.00744, v2, 2020.
2. Lin W-C, Tsai C-F. Missing value imputation: A review and analysis of the literature (2006 – 2017). Artificial Intelligence Review. 2020; Vol. 53, Issue 2, 1487 – 1509.
3. Rubin, D. B., & Little, R. J. A. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.

4. Demirtas H. Flexible imputation of missing data; J Stat Softw, 2018; Vol.85(4).
5. Lee C H, Yoon H- J. Medical big data: promise and challenges, Kiney Res Clin Pract. 2017; 36(1); 3-11.
6. Kwak S.K., Kim J.H. Statistical data preparation: management of missing values and outliers. Korean Journal of Anesthesiology. 2017; 70(4): 407-411.
7. Zhang Z. Missing values in considerable data exploration some introductory chops. Ann Transl Med. 2015; https//doi.org/10.3978/j.issn.2305-5839.2015.12.11.
8. Tsai C-W, Lai C- F, Chao H- C, Vasilakos A V; Big data analytics: a survey; J Big Data, 2015; 2(1), 21.
9. Williams, R. (2015). Overview of Missing Data: Traditional Methods. University of Notre Dame.
10. Sullivan D, Andridge R. A hot deck imputation procedure for multiply imputing non-ignorable missing data: the proxy pattern-mixture hot deck. Comput Stat Data Anal. 2015; 82:173–185.
11. Cheema JR. A review of methods for handling missing data in education research. Review of Educational Research. 2014; Volume 84(4): 487-508.
12. Slavakis K., Giannakis G., & Mateos G. (2014). Modeling and optimization for big data analytics: statistical learning tools for our data-deluge era. IEEE Signal Processing Magazine, 31(5), 18-31.
13. Delalleau O., Courville A., Bengio Y. Efficient EM training of Gaussian fusions with missing data. arXiv preprint arXiv 1209.0521. 2012.
14. Graham J. W., Cumsille P. E., & Fisk E. E. (2012). Methods for handling missing data. In Handbook of Psychology (2nd ed., Vol. 2).
15. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, Franco L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine. 2010; 50(2): 105-115.
16. Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey nonresponse. International Statistical Review, 78(1), 40-64.
17. Rubin L. H., Witkiewitz K., Andre J. S., Reilly S.; Methods for handling missing data in behavioral neuroscience: Don’t throw the baby rat out with the bathwater; Journal of Undergraduate Neuroscience Education; 2007; 5(2): A71-A77.
18. Donders ART, Van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to the imputation of missing values. J Clin Epidemiol. 2006; 59(1).
19. M. M. El-Masri, S. M. Fox-Wasylyshyn; Missing Data: An Introductory Conceptual Overview for the Novice Researcher; CJNR, 2005, 37(4), 156-171.
20. Brown ML, Kros J. Data Mining and the Impact of Missing Data. Industrial Management & Data Systems. 2003; 103(8): 611-621.
21. Allison, P.D. (2001). *Missing Data*. Sage Publications. 104 pages.
22. Rahm E., Do H. H.; Data Cleaning: Problems and Current Approaches. IEEE CSTC on Data Engineering Bulletin; 2000; 23(4).
23. Liu, C. (1995). Missing data imputation using the multivariate t distribution. Journal of Multivariate Analysis, 53(1), 139-158.