

# On the Performance of Imputation Techniques for Missing Values on Healthcare Datasets

L.O. Joel <sup>\*,1</sup>, W. Doorsamy <sup>†,1</sup> and B.S. Paul <sup>‡,1</sup>

<sup>1</sup>Institute for Intelligent Systems, University of Johannesburg, South Africa

## Abstract

Missing values or data is one popular characteristic of real-world datasets, especially healthcare data. This could be frustrating when using machine learning algorithms on such datasets, simply because most machine learning models perform poorly in the presence of missing values. The aim of this study is to compare the performance of seven imputation techniques, namely Mean imputation, Median Imputation, Last Observation carried Forward (LOCF) imputation, K-Nearest Neighbor (KNN) imputation, Interpolation imputation, Missforest imputation, and Multiple imputation by Chained Equations (MICE), on three healthcare datasets. Some percentage of missing values - 10%, 15%, 20% and 25% - were introduced into the dataset, and the imputation techniques were employed to impute these missing values. The comparison of their performance was evaluated by using root mean squared error (RMSE) and mean absolute error (MAE). The results show that Missforest imputation performs the best followed by MICE imputation. Additionally, we try to determine whether it is better to perform feature selection before imputation or vice versa by using the following metrics - the recall, precision, f1-score and accuracy. Due to the fact that there are few literature on this and some debate on the subject among researchers, we hope that the results from this experiment will encourage data scientists and researchers to perform imputation first before feature selection when dealing with data containing missing values.

**Keywords:** Data, Missing Values, Techniques, Imputation, Healthcare

## 1 Introduction

Real-life datasets often contain some missing values or data, which pose a problem to data scientists and researchers working with them. The pattern of the missingness [1] of these

---

<sup>\*</sup>Corresponding Author: ljoel@uj.ac.za; oluwaseyejoel@gmail.com

<sup>†</sup>wdoorsamy@uj.ac.za

<sup>‡</sup>bspaul@uj.ac.za

missing values could be random, that is, missing completely at random (MCAR) or missing at random (MAR). It could also be non-random, that is, not missing at random (NMAR). Some of the reasons for these missing values could be due to errors in the equipment, inappropriate pattern of data capturing, faulty sampling, damages in the specimen used, respondents' irresponsible disposition to certain information or incorrect measurements. Hence, the need to find an appropriate technique in handling these missing values so as to obtain optimal results from the analysis of the data given.

This study compares the performance of seven imputation techniques, which are Mean imputation, Median Imputation, Last Observation carried Forward (LOCF) imputation, K-Nearest Neighbor (KNN) imputation, Interpolation imputation, Missforest imputation, and Multiple imputation by Chained Equations (MICE), on three healthcare datasets, which are the breast cancer [2], the heart disease [3] and the pima indian diabetes [4] datasets. Some percentage of missing values - 10%, 15%, 20% and 25% - were introduced into the datasets under the assumption of MCAR, and the imputation techniques were employed to impute these missing values. The comparison of their performance was done using two error evaluation metrics - root mean squared error (RMSE) and mean absolute error (MAE). While the evaluation metrics used to determine whether to perform selection before imputation or vice versa were the recall, precision, fi-score, and accuracy.

The rest of this paper is organised as follows. Section 2 talks about the datasets considered in the study and the percentage of the missing values introduced into the datasets. Section 3 gives the explanation of the missing data imputation techniques that will be examined in this study. Section 4 explains some details about feature selection and the context of it in this study. Section 5 describes the evaluation metrics - root mean squared error (RMSE), mean absolute percentage error (MAE), recall, precision, fi-score, and accuracy that will be used to evaluate the performance of the imputation methods. Section 6 gives the results and the discussion of the experiments. And lastly, the study ends with some concluding notes in Section 7.

## 2 Datasets

### 2.1 Breast Cancer Dataset

Breast cancer is the most common and leading cause of deaths in females in many countries of the world. The first common symptom of breast cancer is a growth or lump in the breast [5]. This lump can either be cancerous (malignant) or non-cancerous (benign), a doctor has to be consulted for appropriate diagnosis. The dataset for this breast cancer is taken from kaggle database [2]. Figure 1 shows the different features and their data type in the dataset [2]. While Figure 2 shows the distribution of the target feature, called "diagnosis",

in the dataset. "M" stands for malignant and "B" stands for benign.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 32 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   id                                           569 non-null    int64
1   diagnosis                                    569 non-null    object
2   radius_mean                                569 non-null    float64
3   texture_mean                               569 non-null    float64
4   perimeter_mean                             569 non-null    float64
5   area_mean                                  569 non-null    float64
6   smoothness_mean                            569 non-null    float64
7   compactness_mean                           569 non-null    float64
8   concavity_mean                             569 non-null    float64
9   concave points_mean                        569 non-null    float64
10  symmetry_mean                              569 non-null    float64
11  fractal dimension_mean                     569 non-null    float64
12  radius_se                                  569 non-null    float64
13  texture_se                                 569 non-null    float64
14  perimeter_se                               569 non-null    float64
15  area_se                                   569 non-null    float64
16  smoothness_se                             569 non-null    float64
17  compactness_se                            569 non-null    float64
18  concavity_se                              569 non-null    float64
19  concave points_se                         569 non-null    float64
20  symmetry_se                               569 non-null    float64
21  fractal dimension_se                      569 non-null    float64
22  radius_worst                              569 non-null    float64
23  texture_worst                             569 non-null    float64
24  perimeter_worst                           569 non-null    float64
25  area_worst                                569 non-null    float64
26  smoothness_worst                          569 non-null    float64
27  compactness_worst                         569 non-null    float64
28  concavity_worst                           569 non-null    float64
29  concave points_worst                      569 non-null    float64
30  symmetry_worst                            569 non-null    float64
31  fractal dimension_worst                   569 non-null    float64
dtypes: float64(30), int64(1), object(1)
memory usage: 142.4+ KB
```

Figure 1: The Breast Cancer Dataset Information

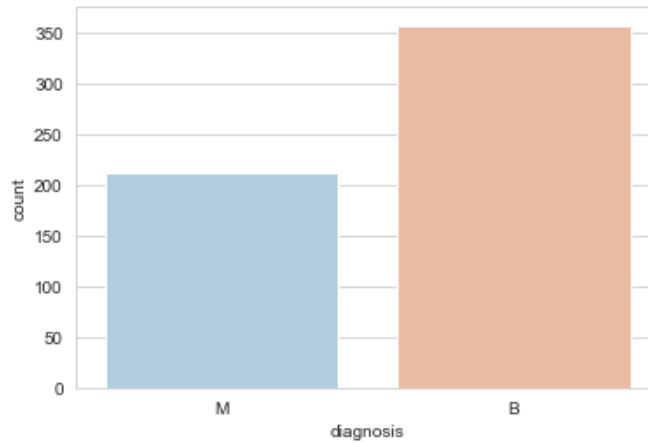


Figure 2: The Distribution of the Target Variable

## 2.2 Diabetes Mellitus Dataset

Diabetes mellitus, commonly called diabetes, is a disease that hinders the body from making enough insulin in order to move sugar from the blood into the cells that will make use of it for energy, thereby causing high blood sugar. This high blood sugar can cause damage to kidneys, eyes, nerves, and other organs in the body. Diabetes can be any of these three types: type 1, type 2, and gestational diabetes.

The diabetes dataset used in this study contains no missing values, but some percentages of missing values were later introduced into the dataset so as to evaluate the performance of the various imputation techniques. The dataset is taken from the popular kaggle database [4]. It contains 768 features (rows) and 9 columns, which include the target or dependent feature (called the Class variable), see Figure 3. The distribution of the class variable can be seen in Figure 4, where 1 represents the presence of diabetes and 0, otherwise.

```

RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
#   Column              Non-Null Count  Dtype
---  -
0   NoP                  768 non-null    int64
1   PGC                  768 non-null    int64
2   DBP                  768 non-null    int64
3   TSFT                 768 non-null    int64
4   2HSI                 768 non-null    int64
5   BMI                  768 non-null    float64
6   DPF                  768 non-null    float64
7   Age                  768 non-null    int64
8   Class variable       768 non-null    int64
dtypes: float64(2), int64(7)

```

Figure 3: The Diabetes Dataset Information

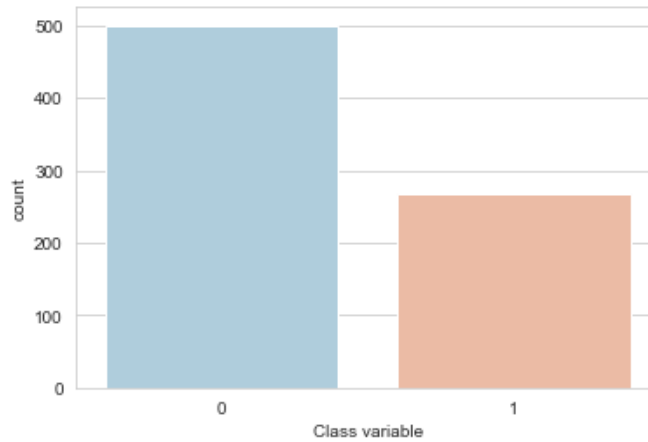


Figure 4: The Distribution of the Target Variable

### 2.3 Heart Disease Dataset

Heart disease can be referred to as any adverse condition affecting the heart. There are different types of heart disease [6] but the most common type is the coronary artery disease. This is when the arteries that supply blood to the heart is clogged, which in turn reduces blood supply, oxygen and nutrients needed for the proper functioning of the heart. It is the leading cause of death in the United States of America [7]. This makes heart disease a major concern in healthcare and any missing values in the dataset could adversely affect the outcome of any machine learning algorithm employed in its prediction.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
#   Column      Non-Null Count  Dtype
---  -
0   age         303 non-null   int64
1   sex         303 non-null   int64
2   cp          303 non-null   int64
3   trestbps    303 non-null   int64
4   chol        303 non-null   int64
5   fbs         303 non-null   int64
6   restecg     303 non-null   int64
7   thalach     303 non-null   int64
8   exang       303 non-null   int64
9   oldpeak     303 non-null   float64
10  slope       303 non-null   int64
11  ca          303 non-null   int64
12  thal        303 non-null   int64
13  target      303 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Figure 5: The Heart Disease Dataset Information

The heart disease dataset is taken from the popular dataset database [3]. It contains 303 features (rows) and 14 columns, which include the target or dependent feature, see Figure 5. In the distribution of the target variable in Figure 6, 0 represents presence of heart disease while 1 stands for the absence of it.

## 3 Missing Data Imputation Techniques

This section discusses some selected imputation techniques that will be used in this study. Additional notes on the selected techniques and/or other techniques used in handling missing values could be found in the following literature [8, 9, 10, 11, 12].

**Mean Imputation:** It is also called mean substitution. This method is very popular among researchers for missing data imputations. It replaces a missing variable in a feature

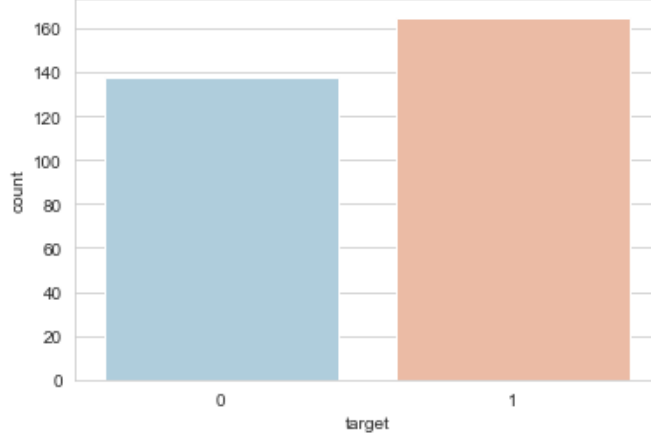


Figure 6: The Distribution of the Target Variable

with the mean of the non-missing variables in the same feature [13, 14]. While it is easy to understand and compute, it leads to underestimation of standard errors.

**Median Imputation:** Median imputation replaces all the occurrences of missing values with the middle value of the non-missing variables in the same feature [10]. It is suitable for continuous and discrete numerical variables only. Although, it is easy to implement, but it could cause distortion in the variable distribution and variance.

**Last Observation Carried Forward Imputation:** Last Observation Carried Forward (LOCF) imputation method, which is commonly used in the analysis of clinical results when the dataset are longitudinal, fills in the missing values of an independent feature with the last non-missing observation of the feature [10, 15]. Hence, this method works on the assumption that the response at the last observed value remains constant.

**K-Nearest Neighbor Imputation:** The K-nearest neighbor (KNN) imputation method imputes missing values in a feature by finding the observations in the dataset closest to the observation which contains the missing values and averages these nearby points to substitute the missing values. The KNN imputation method appears to be robust and effective in missing values imputations [9, 11, 16]. KNN’s configuration often requires selecting a distance measure (such as the Hamming, Euclidean, or the Manhattan distance) and the number of neighbors,  $k$ , that will be used to predict each missing value.

**Interpolation Imputation:** This method fills the missing values with incrementing or decrementing values by performing linear, quadratic or cubic interpolation imputation on

the dataset containing the missing values [17, 18, 19]. The author in [18] compared the performance of linear interpolation imputation method with mean imputation for estimating the missing values in environmental data and found the linear interpolation method to perform better than the mean method in the three evaluating metrics used.

**Missforest Imputation:** Missforest is an imputation algorithm that uses random forest for the imputation of missing data [20, 21]. Missforest imputation first fills the missing values using the mean or mode, then it fits a random forest algorithm on the observed data in order to predict the missing data. This process is performed iteratively until a stopping criterion is met or a maximum number of iterations is attained. These multiple iterations allows the random forest algorithm to improve on the quality of the trained data for the final imputation.

**Multiple Imputation by Chained Equations:** The Mean, LOCF, KNN, LinR, and SR methods described above only create a single value for imputing each missing value. However, multiple imputation method creates multiple values for the imputation of a missing value in order to have different plausible imputed datasets [22, 23, 24]. It allows for the reflection of sampling variability which is lacking in the single imputation methods. One of the commonly used multiple imputation algorithms, among others, is the multiple imputation by chained equations (MICE).

Python programming language is employed for these imputation techniques listed above with the following two python packages - `imputena` [25] and `missingpy` [26]. The packages permit the automated, as well as the customized treatment of missing values in any given dataset.

## 4 Feature Selection

A given dataset often contains a plethora of features. However, in some cases or most cases, not all the features are useful in building a predictive machine learning model. Also, wrong selection of the features might make the prediction results worse. Hence, the need for right feature selection to be done in order to build an optimal machine learning model.

Feature selection is necessary in ML so as to reduce the curse of dimensionality and to build a model that is simple and explainable. It aims to choose a subset of the features in a given dataset, known as the relevant or best features, by removing irrelevant and redundant ones [27]. This can be done in three broad categories, namely (1) Filter Method, (2) Wrapper Method and (3) Embedded Method. See the following literature [28, 29, 30, 31].

This study employs the sequential forward selection (SFS) algorithm to select a subset of the features in each dataset that are most relevant to the each problem. SFS is a fam-

ily of greedy search algorithm that eliminates or adds features based on a given classifier performance metric.

## 5 Evaluation Metrics

This section discusses the metrics used for the evaluation of the performance of the imputation methods. The metrics are the RMSE, MAE, recall, precision, f1-score, and accuracy. A brief explanation of each is given below.

### Root Mean Square Error

The root mean square error (RMSE) represents the quadratic mean of the differences between the imputed and observed data. It is shown in Equation (1) and it is one of the most commonly used metrics in the literature [10, 32, 33]. The value of RMSE is always non-negative and a lower value is better than an higher value.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n \left( X_i^{\text{obs}} - X_i^{\text{imputed}} \right)^2}. \quad (1)$$

### Mean Absolute Error

The mean absolute error (MAE) is the mean absolute difference between the actual and the imputed data. It is also one of the most commonly used metrics in the literature [34]. The formula for MAE is shown in Equation (2). It has an advantage of the absolute value used in the formula, and a lower value is preferable to a larger value. It is robust to outliers [35].

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |X_i^{\text{obs}} - X_i^{\text{imputed}}| \quad (2)$$

### Recall

Recall, also called sensitivity or True positive rate, is the percentage of the total relevant outcome correctly predicted or classified by the algorithm. It gives the measure of how accurately our model is able to identify those patients that have the disease (either has malignant breast cancer or diabetic or heart disease). We need to predict as many of them as possible, hence a high recall value is needed. That is, we need a low value of false negative. The formula for recall is given in Equation (3).

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3)$$



## Precision

Precision, which is also called Positive Predictive Value (PPV), is defined as the fraction of positive predictions that are actually correct. The formula is given in Equation (4). This means when precision is improved, typically recall will be reduced and vice versa.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (4)$$

## F1-Score

F1-score is the harmonic mean of recall and precision. Harmonic mean is used, instead of arithmetic or geometric mean, because it equalizes the weights of the recall and precision. F1-score shows the predictive power of the classification algorithm or model. A Higher F1-score value shows a higher predictive power of the model. The formula for F1-score is shown in Equation (5).

$$\text{F1-score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (5)$$

## Accuracy

Accuracy is the fraction of the total number of predictions that were correctly predicted. In simple terms, it is the fraction of the predictions that the model got correctly. The formula for accuracy is given in Equation (6).

$$\text{Accuracy} = \frac{\text{Correctly Predicted}}{\text{Total Predictions}} \quad (6)$$

## 6 Results and Discussion

First, it is observed that the target feature shown in the diabetes dataset (Figure 4) seems imbalanced. To avoid bias by the classification algorithm, oversampling method was used for this, which brings the total observations in the dataset to 1000. It was 768 before the sampling method was applied. The other two datasets - breast cancer and heart disease - do not need to be balanced. Also, for the breast cancer disease dataset, the target feature, called "diagnosis", was encoded into 1 and 0 using the "LabelEncoder" transformer. Hence, in Figure 2, 'M' is encoded to 1 and 'B' is encoded to 0. The other two datasets - diabetes and heart disease - do not require this process. The missing values were imputed using the imputation methods. For each percentage of missing values introduced into the dataset, we perform imputations using the various methods.

## 6.1 Performance of the Imputation Methods

### 6.1.1 Breast Cancer

Figure 7 gives the RMSE and MAE of the missing data handling technique imputations for 10%, 15%, 20% and 25% missing values respectively. The results given in the figure shows that Missforest algorithm has the lowest errors for both RMSE and MAE. This demonstrates that it performs best compare to other imputation methods used. Following the Missforest algorithm, is the MICE algorithm, which has lowest errors among the remaining imputation methods excluding the Missforest algorithm. The next algorithm with the lowest errors following Missforest and MICE is the KNN algorithm, for both RMSE and MAE. The LOCF method has the highest errors for both RMSE and MAE. Hence, it performs worst than the others for data missing imputation on the breast cancer data. The order of performance of the missing data imputation methods (see Figure 7), from the lowest error to the highest error, is Missforest, MICE, KNN, Median, Mean, Interpolation, and LOCF. This order is the same for both RMSE and MAE.

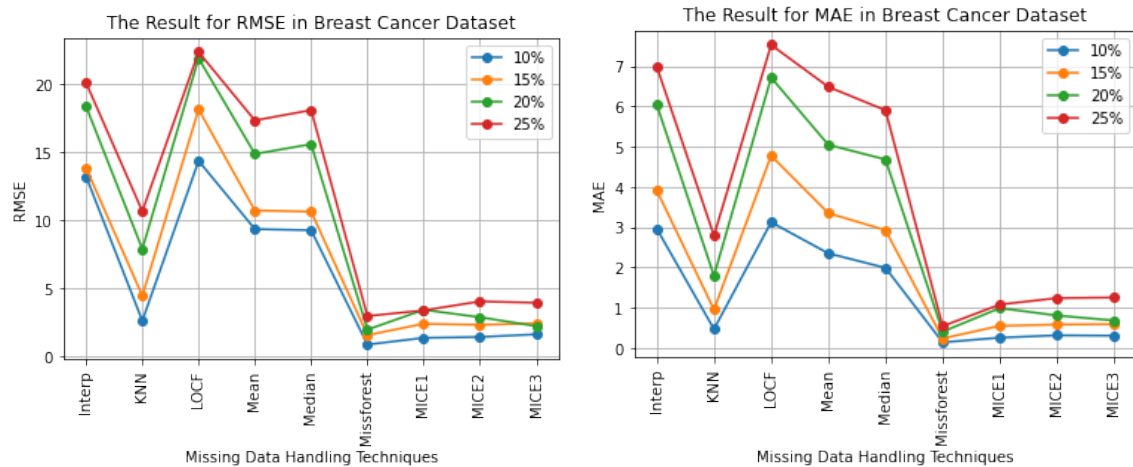


Figure 7: The Errors for each of the Missing Data Handling Technique Imputations in Breast Cancer Dataset: RMSE (left side) and MAE (right side).

The order of performance for each imputation method is the same for all percentages - 10%, 15%, 20% and 25% - of the missing values. For instance, the order of performance for 10% missing values in the RMSE (breast cancer dataset) is Missforest, MICE, KNN, Median, Mean, Interp, and LOCF. This order is also the same for 15%, 20% and 25% missing values. This scenario is played out for both the results in RMSE and MAE on breast cancer dataset.

### 6.1.2 Diabetes

The results of the imputation errors on diabetes dataset is shown in Figure 8, which gives the RMSE and MAE of the imputations for 10%, 15%, 20% and 25% missing values respectively. Here, the Missforest algorithm also has the lowest errors for both RMSE and MAE. Thus, it is the best performing algorithm on the diabetes dataset. The next imputation method with the lowest errors, aside the Missforest algorithm, is the KNN method. The KNN imputation method outperforms all other remaining imputation methods including the MICE, making it the second best. The third performing imputation method is the MICE. Again, the method with the highest errors for both RMSE and MAE is the LOCF imputation method. The order of performance for the missing data imputation methods (see Figure 8), from the one with the lowest RMSE to the one with the highest RMSE is Missforest, KNN, MICE, Mean, Median, Interpolation, and LOCF. While the order for MAE is Missforest, KNN, MICE, Median, Mean, Interpolation and LOCF.

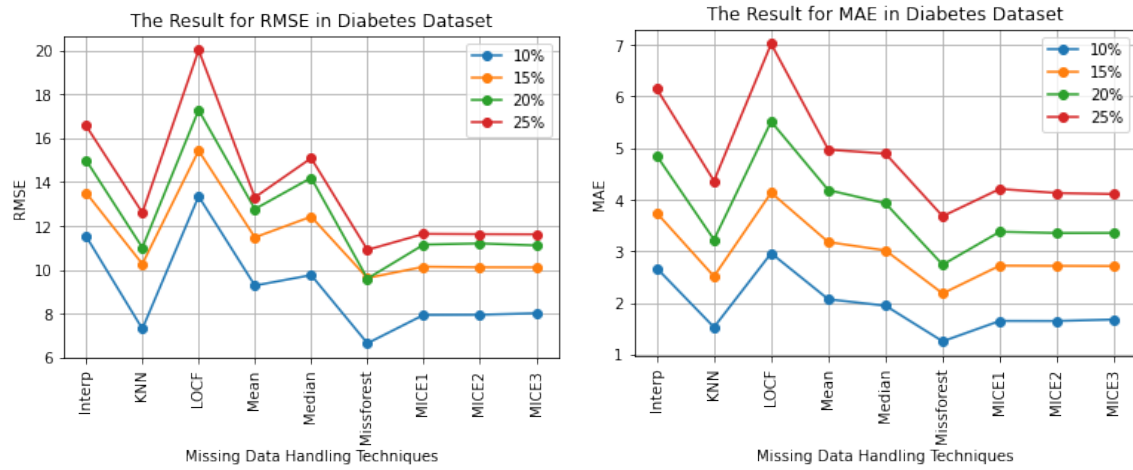


Figure 8: The Errors for each of the Missing Data Handling Technique Imputations in Diabetes Dataset: RMSE (left side) and MAE (right side).

As described for breast cancer dataset, the order of performances of each of the imputation methods in the diabetes dataset is also maintained for all the percentages - 10%, 15%, 20% and 25% - of the missing values for the results in RMSE and MAE.

### 6.1.3 Heart Disease

The results for the imputations in heart disease dataset is shown in Figure 9. The RMSE and MAE for 10%, 15%, 20% and 25% missing values is slightly different from what was seen in the previous two datasets. Although the Missforest imputation method still maintains the

lowest error for RMSE but has around the same error values with MICE for MAE. The errors in KNN imputation method is higher than both Mean and Median imputations, which is not the case in the breast cancer and diabetes datasets. The order, from the lowest error to the highest error for the RMSE is: Missforest, MICE, Median/Mean, KNN, Interpolation, and LOCF. While the order, from the lowest error to the highest error for the MAE is: Missforest/MICE, Median, Mean, KNN/Interpolation, LOCF.

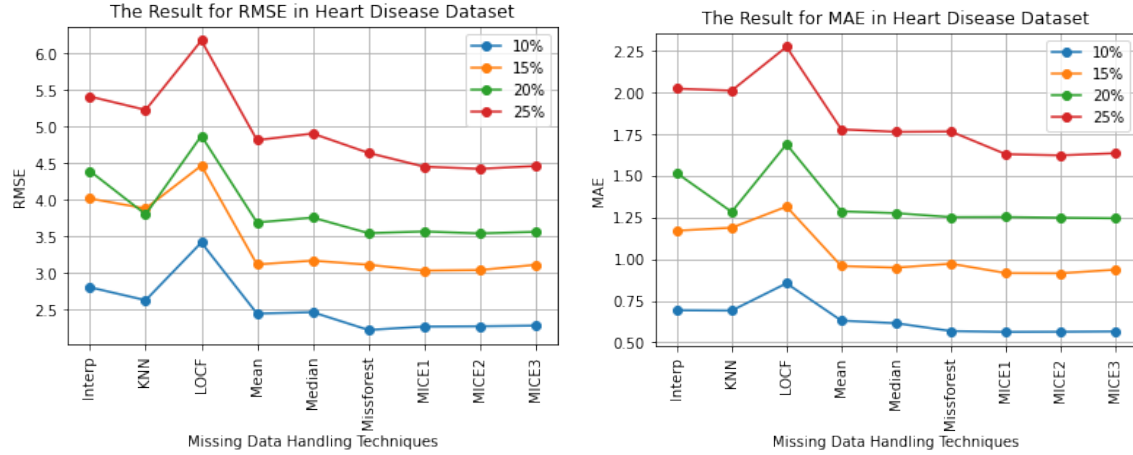


Figure 9: The Errors for each of the Missing Data Handling Technique Imputations in Heart Disease Dataset: RMSE (left side) and MAE (right side).

The order of performances of each of the imputation methods in the heart disease dataset is only maintained for 10%, 15%, and 20% of missing values in RMSE and for 10% and 20% of missing values in MAE. In RMSE, the order of the performances for 10%, 15% and 20% of missing values is Missforest, MICE, Mean, Median, KNN, Interp, and LOCF. However, for 25% of missing values, the order of performance is MICE, Missforest, Mean, Median, KNN, Interp, and LOCF. Also, in MAE, the order of performances for 10% and 20% of missing values is Missforest, MICE, Median, Mean, KNN, Interp, and LOCF. While the order of performances for 15% of missing values is MICE, Median, Mean, Missforest, Interp, KNN and LOCF. And the order of performance for 25% of missing values is MICE, Missforest, Median, Mean, KNN, Interp and LOCF.

## 6.2 Feature Selection Before Imputation or Vice Versa

The best two imputation methods from the previous subsection were selected to determine whether it is better to do feature selection before imputation or to do imputation before feature selection. These two methods are Missforest and MICE. Random Forest algorithm was used for the classification while the best two methods mentioned earlier were used for

imputing the missing values. The experiment was done on the following percentages - 15% and 20% - of the missing values. And the recall, precision, f1-score and accuracy were used to evaluate the performances. Firstly, the results for breast cancer dataset, Figure 10, showed

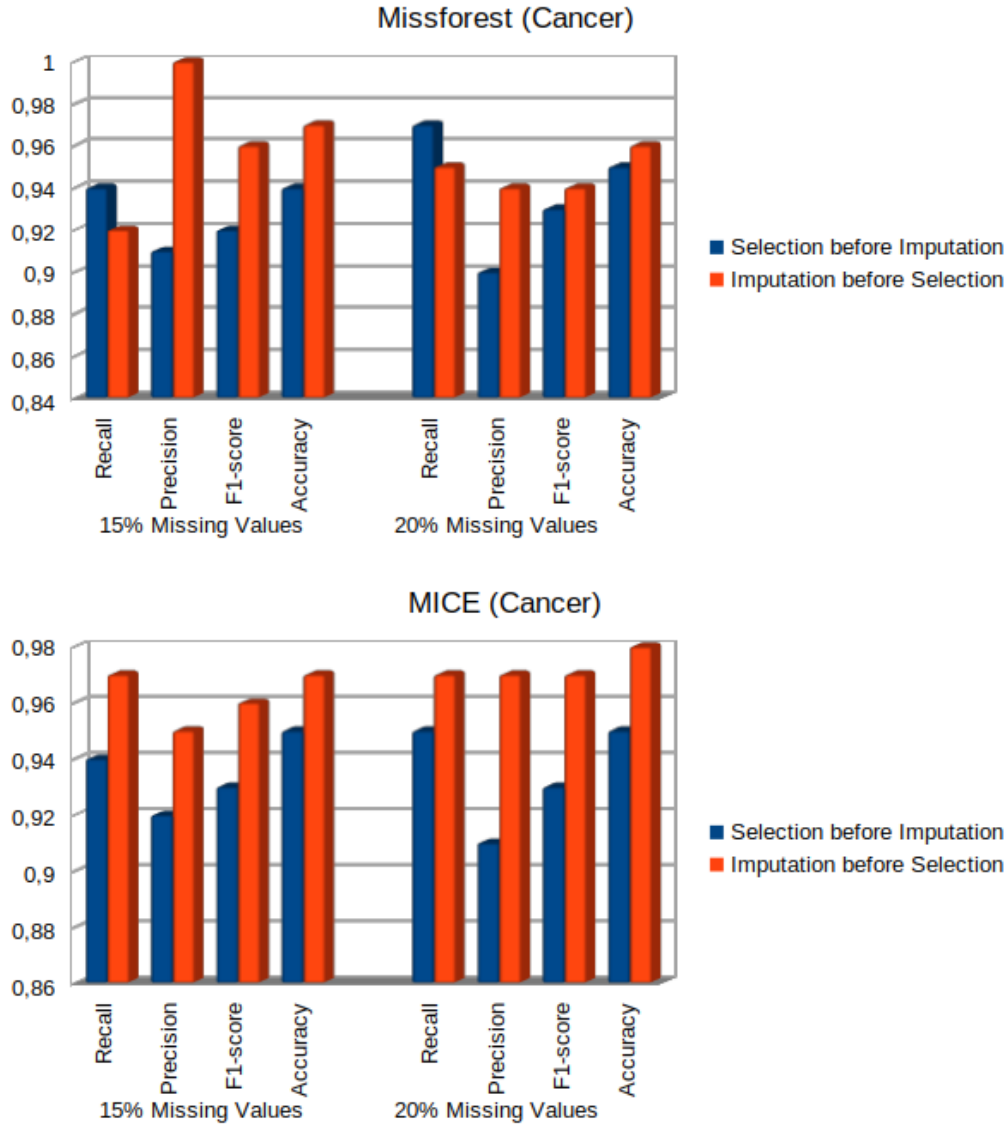


Figure 10: The performance of the Classification Algorithm for Feature Selection before Imputation Versus Imputation Before Feature Selection (Breast Cancer): Missforest (up) and MICE (down)

that the performance of the classification algorithm when the imputation is done before feature selection is better. However, it is observed that the recall score of the Missforest

imputation method (see the figure at the top of Figure 10) for both 15% and 20%, suggests otherwise. There is a higher value for recall when feature selection is done before imputation. Hence, the performance for Missforest (Breast Cancer) classification when feature selection is performed before imputation can be rated 1/4 for both 15% and 20% of missing values.

While the results of the performance of Missforest (Breast Cancer) when imputation is done before feature selection can be rated 3/4 for both 15% and 20% of missing values. On the other hand, the results for MICE (Breast Cancer) classification for feature selection before imputation can be rated 0/4 while that of imputation before feature selection can be rated 4/4 for both 15% and 20% of missing values. Hence, for the breast cancer dataset, we can conclude, from the experiment, that it is better to perform imputation before feature selection.

Secondly, the results for the diabetes dataset (Figure 11) also showed that the performance of the classification algorithm is better when the imputation of missing values is done before feature selection in a given dataset. In Missforest (Diabetes) classification, the performance when imputation is done before feature selection can be rated 4/4 for both 15% and 20% missing values. While the performance when feature selection is performed before imputation can be rated 0/4 in both missing percentages.

Also, in MICE (Diabetes) classification results, the performance when imputation is done before feature selection can be rated 4/4 for both 15% and 20% missing values. While the converse process also gives 0/4 in both missing percentages. Hence, the results suggest that it is better to perform imputation before the feature selection step in working with any given dataset. Lastly, in Figure 12, the classification performance on heart disease dataset for imputing the missing values before feature selection is, again, observed to be better than when feature selection is done before imputation. However, the recall score for 15% missing values in both Missforest (Heart Disease) and MICE (Heart Disease) showed otherwise. However, this is just one out of the four metrics used.

In both the Missforest (Heart Disease) and MICE (Heart Disease) classification, the performance when imputation is done before feature selection can be rated 3/4 for 15% missing values. While the performance when feature selection is performed before imputation can be rated 1/4 for 15% missing values. However, the classification performance rating in both Missforest (Heart Disease) and MICE (Heart Disease) for 20% when imputation is done before feature selection is 4/4 while the converse procedure is 0/4.

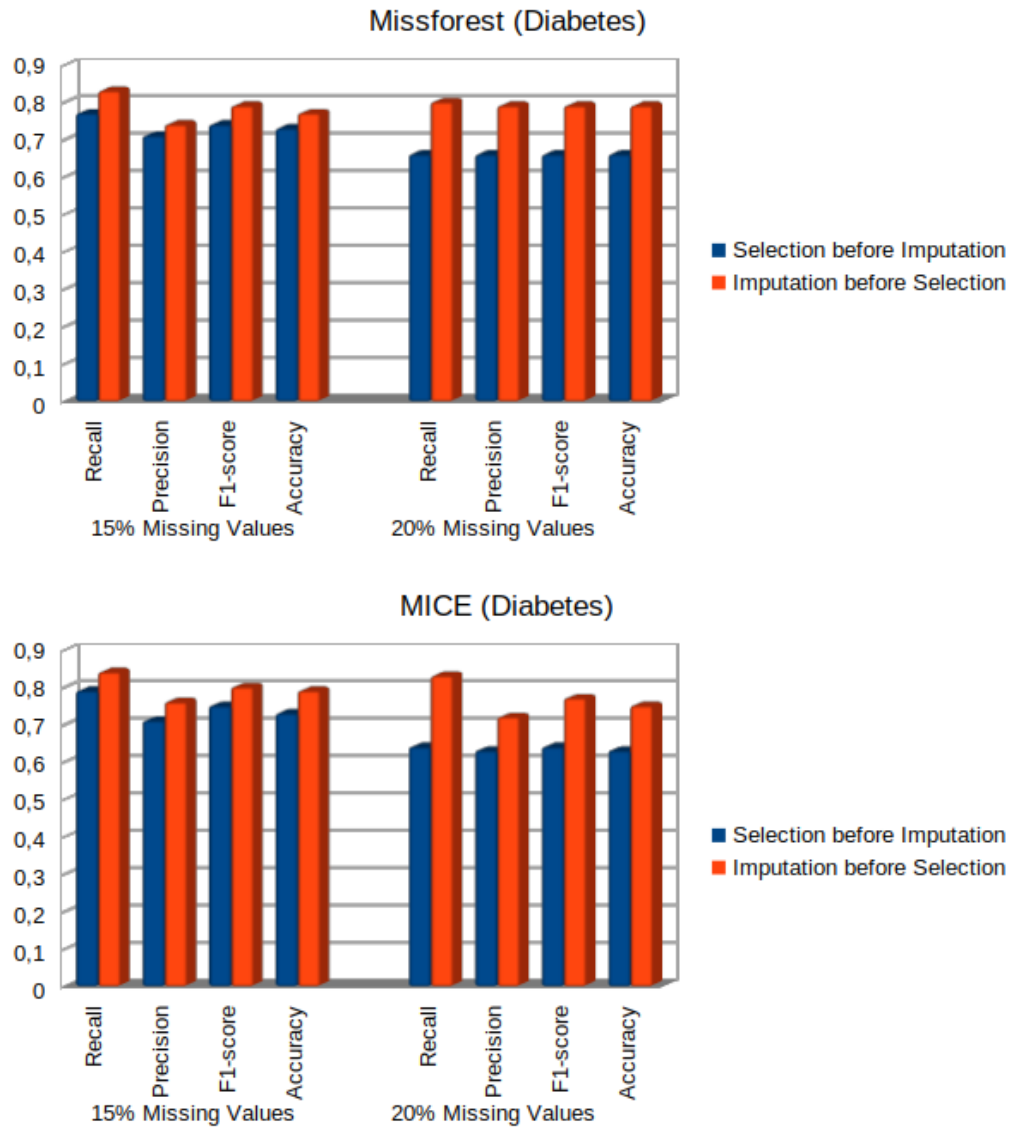


Figure 11: The performance of the Classification Algorithm for Feature Selection before Imputation Versus Imputation Before Feature Selection (Diabetes): Missforest (up) and MICE (down)

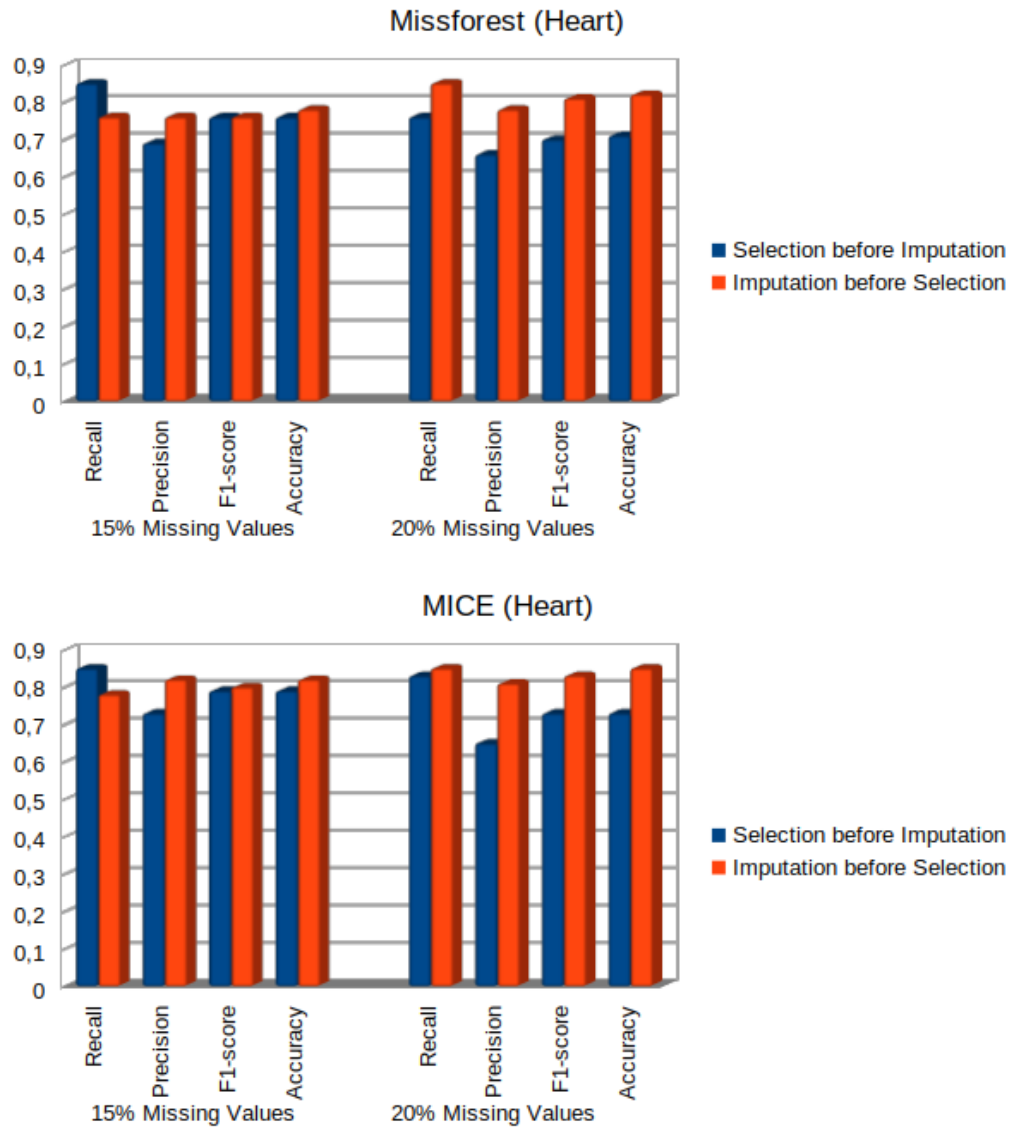


Figure 12: The performance of the Classification Algorithm for Feature Selection before Imputation Versus Imputation Before Feature Selection (Heart Disease): Missforest (up) and MICE (down)



## 7 Conclusion

This study aimed to achieve two things: (1) to evaluate the performance of seven missing values imputation methods on three healthcare datasets, namely the breast cancer, diabetes mellitus and heart disease datasets. (2) to determine whether it is better to impute missing values before performing feature selection on a given dataset or to perform feature selection on the dataset before imputing the missing values.

To achieve the first objective, the RMSE and MAE were used as evaluation metrics for the performances of the missing data handling techniques. Lower value of both RMSE and MAE demonstrates better performance of the methods. Missforest imputation method got the lowest error for both RMSE and MAE in most of the percentages of the missing values introduced into the three healthcare datasets. Hence, it performed the best among the imputation methods. Next in performance, is the MICE imputation. In a similar study carried out by Wu et al.[36], MICE was one of the two suggested best imputation methods that could perform better with small scale database.

For the second objective, random forest algorithm was used for the classification predictions and the metrics used were the recall, precision, f1-score, and accuracy. The experiments were conducted using the two best imputation methods - Missforest and MICE - from the results of the performances of the seven imputation methods used in the previous experiments. The results, from the second experiments, show that it is better to impute the missing values first in a given healthcare dataset before performing feature selection than to perform feature selection before imputation.

## References

- [1] Roderick J A Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., USA, 2002.
- [2] Kaggle. Breast cancer wisconsin (diagnostic) data set. <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>, 2016.
- [3] Kaggle. Heart disease uci. <https://www.kaggle.com/ronitf/heart-disease-uci>, 2018.
- [4] Kaggle. Pima indians diabetes database. <https://www.kaggle.com/uciml/pima-indians-diabetes-database>, 2016.
- [5] Adam Felman. What to know about breast cancer. <https://www.medicalnewstoday.com/articles/37136#symptoms>, 2021.

- [6] Adam Felman. Everything you need to know about heart disease. <https://www.medicalnewstoday.com/articles/237191>, 2021.
- [7] Sherry L Murphy, Jiaquan Xu, Kenneth D Kochanek, and Elizabeth Arias. Mortality in the united states, 2017. 2018.
- [8] Edith D De Leeuw, Joop J Hox, and Mark Huisman. Prevention and treatment of item nonresponse. *Journal of Official Statistics*, 19:153–176, 2003.
- [9] Yaohui Ding and Arun Ross. A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recognition*, 45(3):919–933, 2012.
- [10] Steven J Hadeed, Mary Kay O’Rourke, Jefferey L Burgess, Robin B Harris, and Robert A Canales. Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of The Total Environment*, page 139140, 2020.
- [11] Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933, 2019.
- [12] Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402, 2013.
- [13] Fredrick Ochieng’Odhiambo. Comparative study of various methods of handling missing data. *Mathematical Modelling and Applications*, 5(2):87, 2020.
- [14] Jaemun Sim, Jonathan Sangyun Lee, and Ohbyung Kwon. Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. *Mathematical problems in engineering*, 2015, 2015.
- [15] John M Lachin. Fallacies of last observation carried forward analyses. *Clinical trials*, 13(2):161–168, 2016.
- [16] Shichao Zhang. Nearest neighbor selection for iteratively knn imputation. *Journal of Systems and Software*, 85(11):2541–2552, 2012.
- [17] MN Noor, AS Yahaya, Nor Azam Ramli, and Abdullah Mohd Mustafa Al Bakri. *Filling missing data using interpolation methods: Study on the effect of fitting distribution*, volume 594. Trans Tech Publ, 2014.
- [18] Norazian Mohamed Noor, Mohd Mustafa Al Bakri Abdullah, Ahmad Shukri Yahaya, and Nor Azam Ramli. Comparison of linear interpolation method and mean method

to replace the missing values in environmental data set. In *Materials Science Forum*, volume 803, pages 278–281. Trans Tech Publ, 2015.

- [19] Kurt Kornelsen and Paulin Coulibaly. Comparison of interpolation, statistical, and data-driven methods for imputation of missing values in a distributed soil moisture dataset. *Journal of Hydrologic Engineering*, 19(1):26–43, 2014.
- [20] Shangzhi Hong and Henry S Lynn. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC medical research methodology*, 20(1):1–12, 2020.
- [21] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.
- [22] Jesper N Wulff and Linda Ejlskov. Multiple imputation by chained equations in praxis: Guidelines and review. *Electronic Journal of Business Research Methods*, 15(1), 2017.
- [23] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.
- [24] Maria Pampaka, Graeme Hutcheson, and Julian Williams. Handling missing data: analysis of a challenging data set using multiple imputation. *International Journal of Research & Method in Education*, 39(1):19–37, 2016.
- [25] Miguel Macarro. imputena 1.0. <https://pypi.org/project/imputena/>, 2020.
- [26] Ashim Bhattarai. missingpy 0.2.0. <https://pypi.org/project/missingpy/>, 2018.
- [27] Saurav Kaushik. Introduction to feature selection methods with an example (or how to select the right variables? <https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>, 2016.
- [28] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.
- [29] Siva Sankari Subbiah and Jayakumar Chinnappan. Opportunities and challenges of feature selection methods for high dimensional data: A review. *Ingénierie des Systèmes d’Information*, 26(1), 2021.
- [30] Vipin Kumar and Sonajharia Minz. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014.

- [31] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.
- [32] Peter Schmitt, Jonas Mandel, and Mickael Guedj. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, 6(1):1, 2015.
- [33] Yuebiao Li, Zhiheng Li, and Li Li. Missing traffic data: comparison of imputation methods. *IET Intelligent Transport Systems*, 8(1):51–57, 2014.
- [34] Alexei Botchkarev. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006*, 2018.
- [35] Christian Pascual. Tutorial: Understanding regression error metrics in python. <https://www.dataquest.io/blog/understanding-regression-error-metrics/>, 2018.
- [36] Xuetong Wu, Hadi Akbarzadeh Khorshidi, Uwe Aickelin, Zobaida Edib, and Michelle Peate. Imputation techniques on missing values in breast cancer treatment and fertility data. *Health Information Science and Systems*, 7(1):1–8, 2019.