Performance Analysis of Machine Learning Algorithms for Missing Value Imputation on Medical Dataset

Mrutyunjaya Behera, Pallav Kumar Bhardwaj, Dr. Niranjan Panigrahi

Parala Maharaja Engineering College, Berhampur,Odisha

ABSTRACT *Missing value imputation is an important pre-processing step in the field of data mining. It is highly essential because missing values in a dataset introduces ambiguities while processing these datasets. Many algorithms do not accept any missing data in their training datasets while others look over these values and perform analysis but the results could be biased. So prediction can only be done properly if the datasets are free of any ambiguities. To get rid of biased analysis and unwanted effects on analytical result, missing value imputation is desirable. In this paper, we have evaluated three state-of-the-art machine learning algorithms, namely, K-nearest neighbor, linear regression, and multiple regression on medical dataset to know their relative performance.*

KEYWORDS: Data Mining; Imputation, Machine Learning, K-Nearest Neighbors,Linear Regression, Multiple Regression

1 INTRODUCTION

Every second, vast amount of data is generated all over the world. Data mining takes care of this huge amount of generated data to retrieve some useful information. The medical and health industry generates huge amount of data from medical records and patient monitoring [1]. This huge data needs tools to analyse and these are nothing but data mining tools.

Knowledge discovery from data is most important aspect of data mining which involves various functions like data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge discovery. Fig. 1 explains the process through a systematic diagram [2]. Data mining tools could be used for improvement in various fields like future health care, market basket analysis, education, manufacturing engineering, customer relationship management, fraud detection, intrusion detection, customer segmentation, financial banking, corporate surveillance and research analysis. These areas are affected by improper analysis of data that why we need imputation methods.

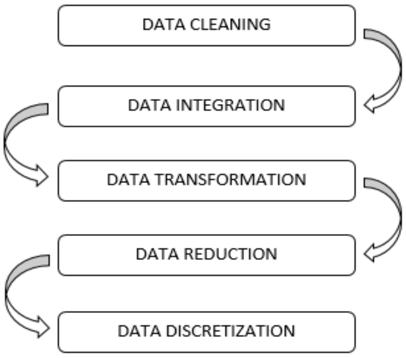

Fig. 1. Data Mining Task [2]

Many existing, industrial and research data sets contain missing values. They are introduced due to various reasons, such as manual data entry procedures, equipment errors and incorrect measurements. During data pre-processing which is a part of data cleaning process involves detecting incomplete, inaccurate, inconsistent and corrupt data. We use data imputation techniques to deal these problems.

The advantage of this approach is that the missing data treatment is independent of the learning algorithm used. This means if we use an algorithm for data imputation, then we are not bound to use only that algorithm for prediction purpose. Missing value problem is regarded as data completeness problem. There are 3 ways to deal with missing data [3]. The first way involves leaving out the missing data, second way can be to manually fill up the missing data and the third way involves imputing these missing data with a proper algorithms. Out of these three, the third way is one of the efficient method to deal with missing data. Imputation is done through various statistical and non-statistical approach. Statistical approach involves missing data imputation using mean, median or mode and some well-known methods in non-statistical approach are: linear regression, multiple regression, and K-nearest neighbour.

In statistics, missing data imputation means substitution of the missing data with a proper value that can generated in various ways. Missing data imputation creates various problems like increase in amount of bias, handling of data becomes hard and increase in inefficiency. However problem could be solved with the latest methods used in these processes. Imputation can be of 2 types single imputation and multiple imputation .Single imputation is good for small dataset whereas we prefer multiple imputation for large datasets. There are four general way to define missingness of the missing data MAR, MACR and NMAR.

Missing Completely at Random (MCAR)- this type missing means the probability of missing of any value is same as nay other values. This means the missing values are independent of other variables present. Missing at Random (MAR) -this means the missing values are dependent on the available information and estimation of these values an be done with the help of other values. Missing values depends on other predictors-Missing values are not missing at random if they depend on values that are not recorded and the missing values can be estimated using these values only. Missing values that depends on the others missing values- This missingness is also known as not missing at random(NMAR).The probability of missing value depends on the (potentially missing) values itself.

There are several ways of imputing data but the oldest one used to be just throwing away the data. This is a primary approach cannot be considered nowadays. This methods have is further divided in various ways 1.Complete case analysis- This is a direct approach for missing values .It involves excluding any values which is missing at the time of input or output. 2. Available case analysis-In this method we study different aspect of problems with different subsets of data. Available-case analysis arises when we remove a set of variables because of their missing rates.

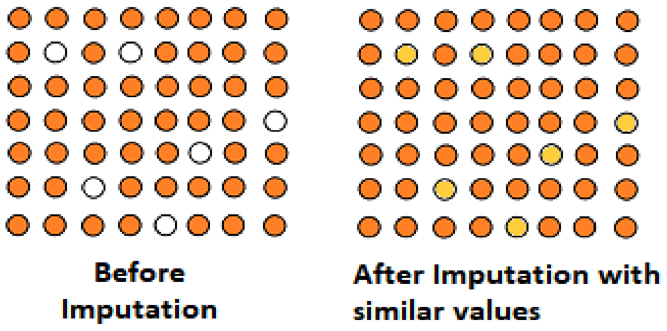

Fig. 2: Schematic Diagram for Imputation on datasets

Medical datasets required a better pre-processing than any other datasets because it is concerned human life. Medical datasets can have attributes which can differ in type of values and also sometimes a proper knowledge of medical domain is required to process these datasets[4][5][6]. This has encouraged us to evaluate some state-of-the-art imputation algorithms on medical dataset to know their relative performance. Fig. 2 represents the datasets before and after imputation.

2 LITERATURE REVIEW

Data imputation is an important issue since data generation is a never ending process in the coming era of big data. In this section, we have presented approaches used in the literature for data imputation.

2.1 STATISTICAL APPROACH FOR HANDLING MISSING DATA

This approach is based on statistical view of the datasets [5]. This is quite old in nature than the machine learning approach but it has been in use for some decades. Machine learning approach is designed because statistical approach is quite inefficient in nature and not good for large datasets.

LIST-WISE DELETION

In imputation most traditional theory used is throwing away the data. This means we will omit the missing values. And further continue to analyse data without these missing values. This is an old technique and of no use in the recent times because this huge amount of data cannot be imputed using this technique. So this technique is called list-wise deletion and falls under statistical technique. Handling missing values with this technique is a default technique this means when we cannot employ any technique this will be the last option we should go. Even if we use this method it should be limited to a small amount of missing data because if it is implemented in large datasets it will lead to biased result. It has more limitations like it could only be applied to missing values completely at random (MCAR) which rarely occurs. Apart from this if we delete all the missing values then we could lose a critical amount of data. So, the it concludes that we cannot use it because of its demerits.

PAIR-WISE DELETION

Another statistical way of handling missing data is pair-wise deletion. Pair-wise deletion eliminates information when a data point is missing and that is needed for calculation of particular assumption. If data is missing in any other area then existing values are used in the statistical testing. Pair-wise deletion preserves more information than list-wise deletion. This method further adds to some problems such as the parameters of the model will work on different sets of data with different values.

MEAN SUBSTITUTION

In a mean substitution, the mean value of a variable is used in place of the missing data value for that collected data. The theoretical background of the mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal distribution. Mean may produce a great bias in large datasets or in the datasets where there is great inequality in the distribution of missing values.

MULTIPLE IMPUTATIONS

Multiple imputations is another statistical approach for missing value imputation. Rather than filling the missing data with list of plausible values s that represent the uncertainty about the right value to impute. It was developed Rubin in 1987.These imputed data values are then analysed by the same standard procedures that we used foe the complete datasets.

Multiple imputations is performed in 3 steps: First, the data values are imputed m times to generate m datasets. Second, the m complete datasets are then analysed using the standard procedures. Third, the results from m datasets are then combined. There can different ways for multiple imputations which depends on missing pattern in the datasets.

MAXIMUM LIKELIHOOD

Likelihood: In Maximum Likelihood is implied, the assumption used is the observed data is from a multivariate normal likelihood function to a linear model. After the parameters are estimated using available data then missing data is estimated using these parameters. This is another statistical approach for imputation. If there are missing but relatively complete data, the statistics explaining the relationships among the variables may be computed using the maximum likelihood method. That means, the missing data may be estimated by using the conditional distribution of the other variables.

2.2 MACHINE LEARNING APPROACH OF HANDLING MISSING DATA

This section briefly reviews the state-of-the-art machine learning approaches used on the medical dataset in this paper.

LINEAR REGRESSION

Linear Regression is another common predictive model used to impute missing values. It is a widely used supervised learning method. It is used to predicts the value of an outcome variables " Y " based on one or more input variables "X". The main aim is to establish a linear relationship between the independent variable i.e., predictor variable(s) and dependent variable i.e., response variable so that we can use this formula to estimate the value of the response Y, when only the predictor X values are known.

The mathematical equation can be generalized in the form:

$$Y = B + B_1 X \qquad (1)$$

Where B is the intercept and B1 is the slope. Collectively both are called regression coefficient.

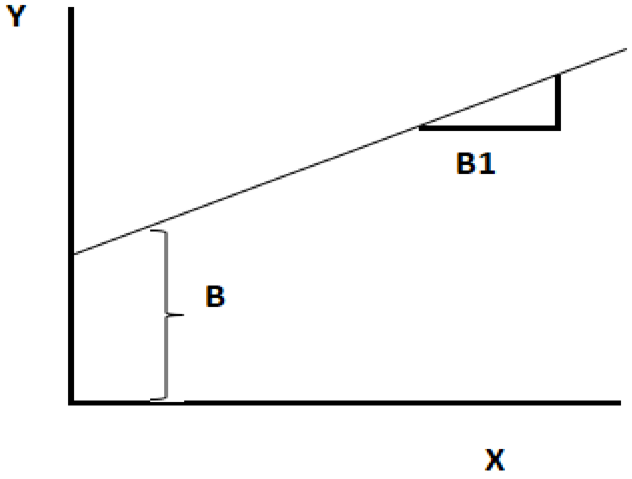

Fig 3: Schematic Diagram of Linear Regression

One of the challenging problems while performing Linear Regression is the number of attributes. As the number of attributes, increases, the handling of the dataset also increases. This is due to the increasing attribute combination which makes the problem analysis NP-Hard. The complexity shows non-linearity. Feature selection are usually applied to handle when the datasets are very large.

Linear Regression Algorithm

Step 1: Place the dataset in your records.

Step 2: Find the correlation between the attribute.

Step 3: Use the function symnum to find the most correlated attribute (attribute with most * are called as most significant and usually used as a predictor variable).

Step 4: Call a function and check where the NA is present and categorized them as 0 (non missing value) & 1 (missing value) and assign to an another attribute let's say I.

Step 5: Put lm function to find the intercept and coefficient between the y and x.

Step 6: Place the intercept and coefficient value in linear regression formula to impute all the missing values.

Step 7: Print the data.

Drawbacks:

The linear regression model will not be helpful for large dataset for that we have to use other technique to impute the missing values. It is also becomes very difficult when there are many attributes.

MULTIPLE REGRESSION

Another machine learning technique used for data imputation is Multiple Regression. It is an extension of linear regression into relationship between more than two variables. In simple linear regression we have one predictor variable and one response variable, but in multiple regression we have more than one predictor variable and one response variable.

The general form of multiple regression is :

$$Y= a + b_1x_1 + b_2x_2 + b_3x_3 +\ldots\ldots\ldots b_{nxn} \qquad (2)$$

Where y is response variable, a, b1 , b2 , b3, … bn is the coefficients and x1 ,x2 , x2 , xn are the predictor variables.

It uses the same algorithm as simple linear regression, only in a lm function you can add more number of attribute (predictor variable) .

It is the algorithm commonly used because of the simplicity of imputation. In this regression we can select as many attribute as we required but there should be a functional dependency between the attribute. It generally works poor with larger datasets.

MEAN , MEDIAN AND MODE IMPUTATION

This is one of the frequently used methods. It consists of replacing the missing value for a given attribute by the mean or median or mode of all known values of that attribute.

The mathematical formula of Mean ($\overline{\ }$) is as follows:

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{\mathbf{I}} \qquad (3)$$

K-NEAREST NEIGHBOUR

KNN algorithm is one of the basic machine learning algorithm. It is a straightforward algorithm. KNN algorithm is regarded as a better algorithm than random forest, CART and logistic regression in terms of easy to operate, predictive power and calculation time. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance.

Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j.

$$\text{Euclidean Distance}(x, xi) = \sqrt{\sum (x_j - x_{ij})^2} \qquad (4)$$

Other popular measures are

1. Hamming Distance: Calculate the distance between binary vectors

2. Manhattan Distance: Calculate the distance between real vectors using the sum of their absolute difference.

3. Minkowski Distance: Generalization of Euclidean and Manhattan distance.

If value of k is very small then it will susceptible to over-fitting and sensitive to noise points.If k is very large then the this may cover all data points that are located far away from its neighbors .Here we take value of K between 5-20 as it regarded as a safer values for prediction.Fig. 4 explains that how the nearest neighbour of a data points changes with changing value of K.

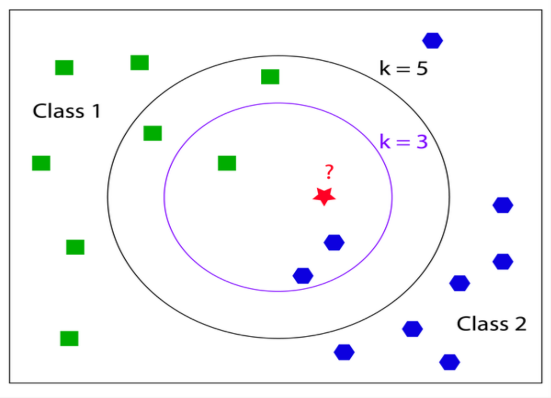

Fig. 4: K-Nearest Neighbour Method

Algorithm:

1. First we read the datasets and simultaneously add NA to every missing value in the datasets. Because it will be easier for the algorithm to calculate the nos of missing value.

2. We can see the summary of the datasets to know many missing values are there.

3. Now we will run the knn function on the dataset and save it to another datasets.

4. If we see the summary of the new dataset the we can see there is no NAs left and hence it means all the values are imputed.

3 EXPERIMENTAL SETUP

Fig. 5 represents the flow in which we have performed the analysis. The flow of analysis is same as in [5] but, different datasets are tested by different machine learning algorithms. The first step involves acquiring medical datasets from UCI repository, healthdata.gov and few other sources. After the medical datasets are available, we have checked for availability of missing values. If missing values are present then only further steps will be performed. Data cleaning will be done and then transformation is performed on the datasets so that imputation will be done. Interpretation involves representation of the data in an understandable and meaningful manner. In this experiment, we have obtained the datasets with no missing values. So we intentionally created the missing values to check it against the actual values.

We have used the R programming language and the R studio application for our experiment. R Studio is a free and open source integrated development environment for R. The reason behind using R is that it is widely used and popular among data scientists and data analysts. Its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others.

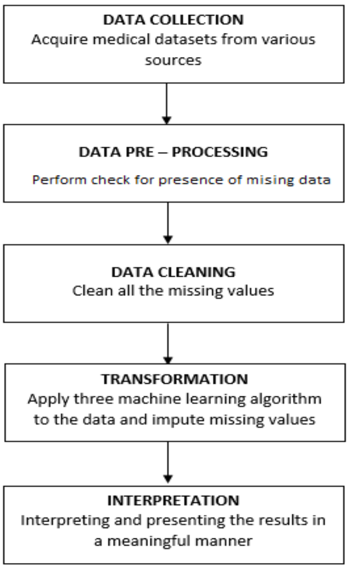
DATA COLLECTION
Acquire medical datasets from various sources
DATA PRE – PROCESSING
Perform check for presence of mising data
DATA CLEANING
Clean all the missing values
TRANSFORMATION
Apply three machine learning algorithm to the data and impute missing values
INTERPRETATION
Interpreting and presenting the results in a meaningful manner

Fig 5 : Experiment Flow [5]

4 EVALUATION METRICS

The machine learning algorithms are evaluated using three metrics [5]: Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as given in equations 5, 6 and 7.

MAE measures the average difference between imputed values and true values as in the following equation:

$$\mathrm{MAE}= \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} |y_i - \hat{y}_i | \qquad (5)$$

While MSE is equal to the sum of variance and squared of the predictions of missing values, defined as:

$$\mathrm{MSE}=\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(y_i - \tilde{y}_i)^{2} \qquad (6)$$

RMSE calculates the difference between predicted (imputed) and actual values. Basically, it represents the sample of differences in standard deviation as follows:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(X_i^abs-X_i^imputed)^2}{n}} \qquad (7)$$

5 RESULT ANALYSIS

The performance of linear regression, multiple regression and KNN on the medical datasets are shown in Fig. 6,7 and 8 respectively. The figures shown below are obtained from behavioral analyzer implicitly present in R studio. To have a unfiform analysis, the above mentioned evaluation metrics are used which is described below.

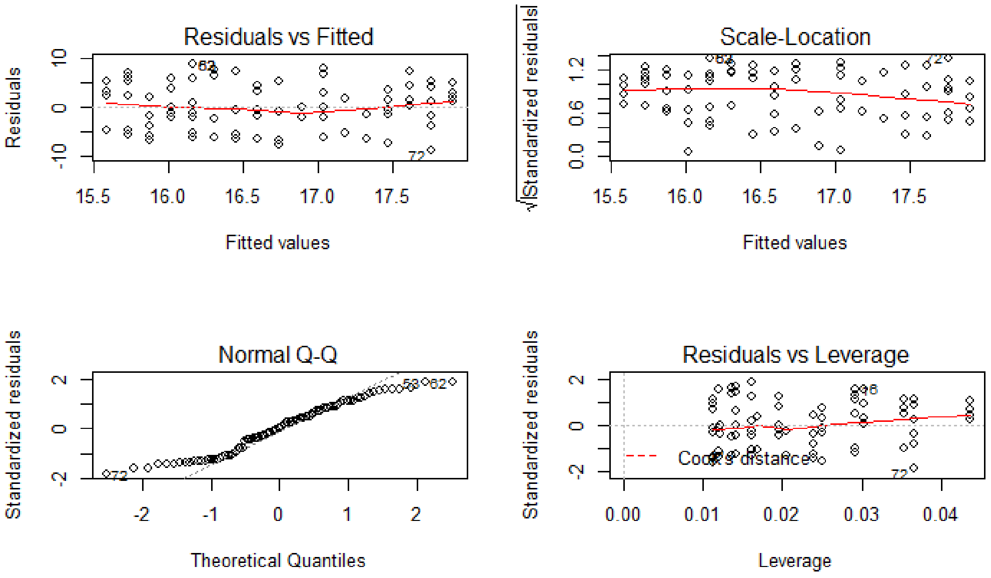

Fig. 6: Behavoiur of linear regression

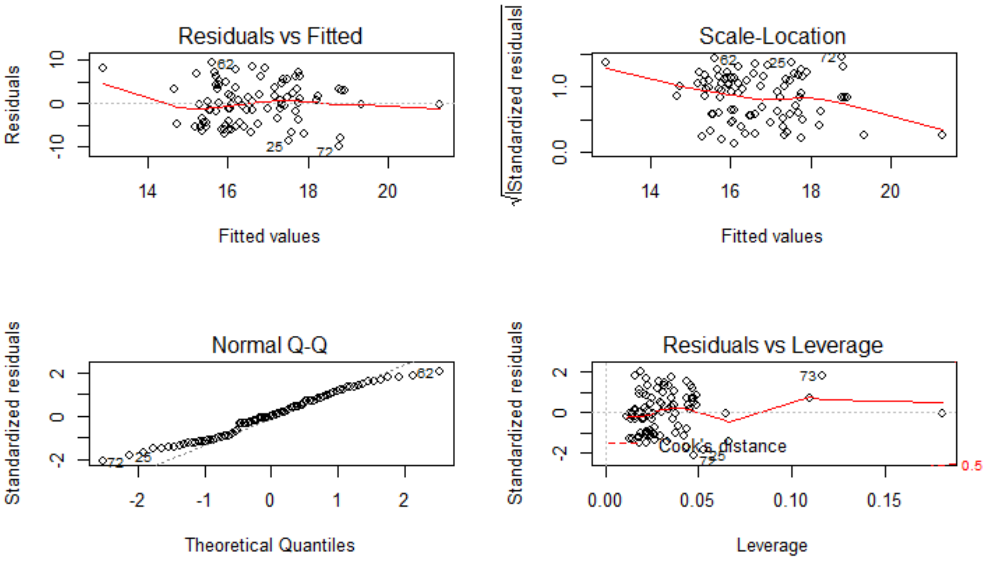

Fig. 7: Behavoiur of multiple regression

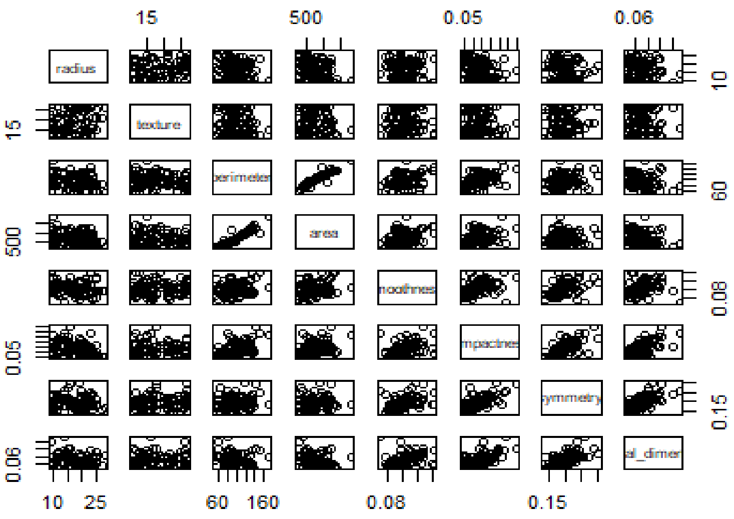

Fig. 8: Behavoiur of K-Nearest Neighbour

The optimization of the different algorithm is achieved by setting up respective net parameters.The dependence of the MAE and RMSE on the number of samples used as input for training the linear regression, multiple regression and KNN. It is also observed that when there are many number of attributes its becoming very difficult to find the functional dependent between them while solving through linear regression. The performance of the imputation, both in its training and prediction mode, clearly deteriorates as the number of input vector decreases, resulting in an imputation of lesser quality. For instance, when the incomplete vectors/total vectors ratio increased from 13% to 35%, the MAE increased 65% and the RMSE increased over 42%. This decrease in the ability of the datasets with increasing loss of information meaning that imputation of data sets with large proportions of missing values might lead to erroneous results. If we compare between these three algorithms then KNN is the best method to impute the missing values as it can easily replace the missing values with nearest value.

6 CONCLUSION & FUTURE WORK

In data mining, missing values can potentially produce wrong final analysis. Handling missing values is a challenging and also an interesting area of research in medical data mininng. In this work, we have adapted different machine learning techniques to impute the missing values of attributes for the records of the medical dataset. With this extra effort, we can get a good quality data for better classification, analysis and decision support. Our future work will include testing of other machine learning algorithms on medical datasets, and optimizing the highest accuracy of a machine learning classifier to impute missing values.

REFERENCES

[1] Liu Y, Gopalakrishnan V.,:An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data. Data.;2(1):1-23(2017) doi: 10. 3390 / data2010008.

[2] Agarwal V. ,: Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis, International Journal of Computer Applications 131(4), 30– 36 (2015)

[3] Rahman M.M., Davis D.N.: Machine Learning-Based Missing Value Imputation Method for Clinical Datasets. In: Yang GC., Ao S., Gelman L. (eds) IAENG Transactions on Engineering Technologies. LNEE,Vol. 229,pp. 245-257 Springer, Dordrecht (2013)

[4] Mathura B. B., Mangathayaru N., Padmaja Ran B.,: An Approach to Find Missing Values in Medical Datasets. In Proceedings of the The International Conference on Engineering & MIS 2015 (ICEMIS '15). ACM, New York, NY, USA, , Article 70 , pp.1-7. DOI: http://dx.doi.org/10.1145/2832987.2833083 (2015)

[5] Abidin N.Z., Ismail A.R., Emran N.A,: Performance Analysis of Machine Learning Algorithms for Missing Value Imputation. International Journal of Advanced Computer Science and Applications 9(6), 442-447 (2018)

[6] Folguera L., Zupan J.,Cicerone D., Magallanes J.F.,: Self-organizing maps for imputation of missing data in incomplete data matrices, Chemometrics and Intelligent Laboratory Systems,143,pp.146-151(2015),ISSN01697439