Performance Analysis of Machine Learning Algorithms for Missing Value Imputation

Nadzurah Zainal Abidin, Amelia Ritahani Ismail*
Department of Computer Science
Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia,
P.O Box 10, 50728 Kuala Lumpur, Malaysia

Nurul A. Emran
Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka (UTeM),
Hang Tuah Jaya, Durian Tunggal, Melaka, 76100 Malaysia

***Abstract*—Data mining requires a pre-processing task in which the data are prepared, cleaned, integrated, transformed, reduced and discretized for ensuring the quality. Missing values is a universal problem in many research domains that is commonly encountered in the data cleaning process. Missing values usually occur when a value of stored data absent for a variable of an observation. Missing values problem imposes undesirable effect on analysis results, especially when it leads to biased parameter estimates. Data imputation is a common way to deal with missing values where the missing value's substitutes are discovered through statistical or machine learning techniques. Nevertheless, examining the strengths (and limitations) of these techniques is important to aid understanding its characteristics. In this paper, the performance of three machine learning classifiers (K-Nearest Neighbors (KNN), Decision Tree, and Bayesian Networks) are compared in terms of data imputation accuracy. The results shows that among the three classifiers, Bayesian has the most promising performance.**

***Keywords*—Data Mining; Imputation; Machine Learning; K-Nearest Neighbors; Decision Tree; Bayesian Networks**

I. Introduction

Data mining is a modern approach to solve many complex and real world problems. This fairly self-explanatory term is a well-known and widely used process that evolves with new technologies. In data mining, data pre-processing is the most important step to ensure the quality of data and the results that leads to reliable decisions. According to Vivek, data pre-processing is the process of simple transformation of raw data into understandable format. Data pre-processing major activities include data cleaning, integration, transformation, data reduction and data discretization as shown in figure 1. One critical activity in data pre-processing is dealing with missing data. This process falls under the first stage of pre-processing data, which is data cleaning. This first stage of data pre-processing is concerned about detecting incomplete, inaccurate, inconsistent and corrupt data, and applying techniques to modify or to delete this spurious data [1]. Pyle proposed in his book Data Pre-preparation for Data Mining that major tasks in data cleaning are to impute missing data, remove outliers and resolve inconsistencies. In fact, in data quality, missing values has been recognized as one form of data completeness problem [2].

In certain observation of interest, missing data can be defined as the absence of data value for a variable. Missing data is commonly described as major issue in most scientific research domains that may originate from such mishandling samples, low signal-to-noise ratio, measurement error, non-response or deleted aberrant value [1]. Nevertheless, as claimed, missing data can also introduce the element of uncertainty in analyzing data. Previous researchers have proposed several ways in handling missing values. The simplest technique is to ignore the missing values [3]. This technique is usually adopted when to a missing class label. Nevertheless, the technique is not appropriate and effective in the case where the percentage of missing values differ significantly. The next technique is to manually fill in the missing value, which will only introduce tedious and infeasible results. Somasumdaram and Nedunchezhian claimed that the third technique used in dealing with missing values is using a global constant (such as 'unknown') to fill up the missing values in data sets. Even though this technique use global constant value to substitute the missing value, it treats all data sets as the same. As a results, a considerable amount of distortions will be introduced in the data sets of concerned. In addition, if similar global constant such as 'unknown' is used, the data is still implicitly incomplete, as the value represents a variation of 'NULL' that denotes missing especially in database community. The final technique is data imputation, that relies on observed data sets to predict missing values [4] (Fig. 1).

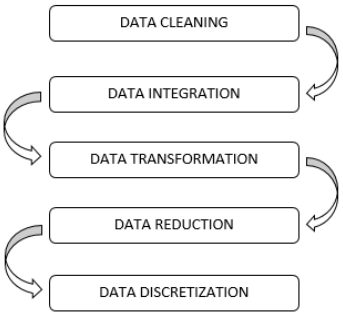

Fig. 1. Data Mining Task (Vivek Agarwal, 2015).

Data imputation is defined as a technique of replacing missing data with substituted values [5]. Selection of imputation method usually determined by the mechanism of how the values are missing. Rubin has described the three missing values

mechanism as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). MCAR is to describe a situation where the missing values is not correlate to a certain value which assumes to be obtained or to an observed responses [5]. MAR data is the situation where the likelihood of missing value instance mostly depends on the known values instead on the real value of the missing data itself [6]. While MNAR describes a situation when the propensity of a missing value in a class instance is to depend on the value of that variable.

In the literature, various data imputation techniques have been introduced, Statistical and Machine Learning techniques have been used in various application contexts of data imputation as we shall see in the next section. Even though the conventional, statistical technique has been adopted for decades, the machine learning-based data imputation techniques are becoming popular in handling missing values especially in large data sets. In the next section, description of statistical and machine learning techniques (classifiers) used for data imputation will be given. Section III covers the evaluation methods for the comparison of three classifiers namely KNN, Decision Trees, and Bayesian Networks. These classifiers will further be measured by evaluating three parameters: Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Square Mean Error (RMSE) in section III. This is followed by Section IV for the results and discussion. Finally, Section V concludes this paper.

A. Literature Review

Data imputation theory is an emerging topic in statistics and machine learning. In this paper, we aimed to explore the characteristics of the techniques.

B. Statistical Approach of Handling Missing Values

1) Listwise Deletion: In imputing missing values, the most traditional theory used is by throwing away data. By this way, we omit records with missing values and continue to analyze the remaining data [4]. This technique is reputably known as listwise deletion, and falls under one of the statistical techniques. Handling missing values with listwise deletion is a default option in most statistical analysis. However, this approach is only pertinent to be used if there is only limited number of missing values, as otherwise it will eventually lead to biased analysis. Another limitation with listwise deletion, it is only relevant when missing values are completely at random (MCAR) which unfortunately rarely happens in reality [5]. Apart from that, one might risk loss of critical information if all missing values are deleted. Ultimately, this approach leads to bias parameters and estimates.

2) Pairwise Deletion: Another known statistical method of handling missing data is pairwise deletion. One researcher [5] claimed that pairwise deletion technique gets rid of information on a particular information data to test if a particular assumption is missing. This statistical testing will be adapted to the observed data if there are missing value elsewhere in the dataset. A disadvantage of pairwise deletion is the tendency to produce a standard of errors that are either underestimated or overestimated [7]. Besides, pairwise deletion is not able to compare analyses as sample dataset different each time.

Marina Soley-Bori mentioned that the two improved approaches that have been proposed to handle missing values are multiple imputation and maximum likelihood [8].

3) Multiple Imputation: In multiple imputation, a new technique of treating missing values is introduced, where it imputes missing values with a set of acceptable values that may contain uncertainty to the original values, instead of replacing a single data to all missing attributes [6].

This approach usually begins with a prediction of the existing data from another variable and then replaced the missing values with the predicted values [6]. A full set of plausible values is the results of the imputed data set. Nevertheless, it has been reported that the downside of this method is different uncertainty values may be yielded for the same data set used for imputation [9].

4) Maximum Likelihood: In Maximum Likelihood is implied, the assumption used is the observed data is from a multivariate normal likelihood function to a linear model. According to researchers [10], the equation of maximum likelihood estimation for incomplete data set are:

$$y \epsilon R^n$$

$$z \epsilon R^1$$

$$(y, z) \epsilon R^n + 1$$

where y is observed data, z is missing data and (y,z) are the complete data.

This technique behaves by estimating the observed data using existing data and estimate missing values with respect to the estimated parameters. The limitation of this approach are it requires specialized software, which may be challenging and time-consuming.

Imputation supposed to produce a complete data set in order to improve its usefulness. However, the statistical techniques described so far still suffer from loss of information. This will eventually lead to invalid conclusions and biased parameters. Therefore, in the next section, alternative way of imputation for missing values using machine learning techniques (or also called as classifiers) will be presented.

C. Machine Learning Approach of Handling Missing Data

Machine learning approach has revolutionized the world with various algorithms to aid data analysis. However, in data imputation, machine learning is in its infancy, and thus offers many research opportunities. In this paper, we focus on four machine learning techniques that have been proposed in data imputation. These techniques are as follows:

1) Decision Tree: Decision tree is another common predictive model used to impute missing values. Decision tree has introduced imputation techniques to the missing values that allows validation of the imputed values against the actual values. This technique begins by splitting the leaves of a tree until running out of questions.

A decision tree has two kinds of nodes. First, this approach tackles imputation by determining each leaf node that has a class label with a majority vote of training examples reached the leaf. Besides, each internal node should represent a question on features that will be branching out according to the answers as Fig. 5 [11] (Fig. 2).

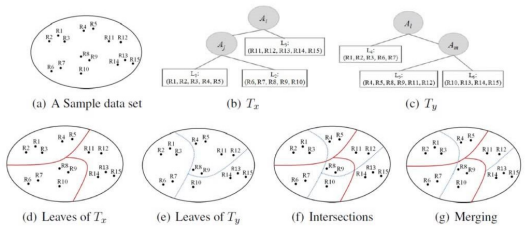

Fig. 2. Basic Concept of Decision Tree (Rahman and Islam, 2013).

$$H(D) = -\sum_{i=1}^{k} P(C_i|D) log_k(P(C_i|D)) \quad (1)$$

The equation assumes that all trees are equally split through the dataset.

As claimed, the transparency of decision tree has made it as the most frequent algorithm used in data mining approach [12]. Nevertheless, the researchers explained that the root in decision tree algorithm should illustrate a question with multiple answers. For imputation purposes, each answer should generate a set of questions that help to determine the data and make the final decision based on it. The final result of decision tree should indicate the possibility of all scenario of decision and outcome.

Despite all benefits mentioned, one researcher claimed that main drawback of decision tree is the computational cost such as running time and trees to construct different test samples [13].

2) K-Nearest Neighbor (KNN): K-nearest neighbors (KNN) is the most straightforward algorithm in imputing missing values. Besides, this algorithm has been used to solve many predictive problems.

In order to impute a value of a variable, K-nearest neighbors (KNN) defines a set of nearest neighbor for a sample and substitutes the missing data by calculating the average of non-missing values to its neighbors [6]. Nearest neighbors is measured as the closest values based on the Euclidean distance as follows.

$$D(a,b) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2} \quad (2)$$

As KNN imputes missing values based on its neighbor, it may introduce an uncertain analysis in relation to the value of *k*. If *k* is too small for a big dataset, the classifier may be susceptible to over-fitting and sensitive to noise points. On the other hand, if *k* is too large, this may cover all data points that are located far away from its neighbors. The decision will eventually lead to bias as it covers a greater instance space.

As to the matters mentioned in relation to *k*, the best choice of *k* influence t o m ake a b etter d ecision a nd a nalysis. One researcher [14] claimed that the most suitable value for *k* can be obtained through a formula of 1/*k* as shown in Fig. 3 with regards on the size of dataset and percentage of missing values.

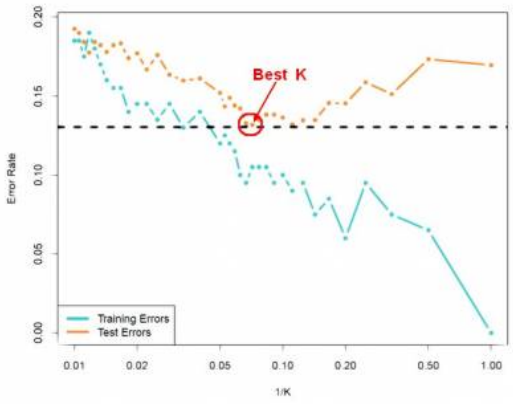

Fig. 3. Best K-Value (Gerardnico, 2017).

KNN is one of the algorithms commonly used because of the simplicity of imputation. However, this imputation technique requires scanning the entire dataset to find the k-nearest neighbors and thus it can be expensive and suffers poor performance especially for a large dataset [15].

3) Bayesian Network: Another machine learning technique used for data imputation is Bayesian networks. Bayesian networks are growing as the model of choice in resolving many problems. Bayesian capture probabilistic relationships between variables in a concise manner by enforcing conditional independence constraints [16]. Using Bayesian networks for imputation offers several advantages: 1) the ability to handle missing data models encodes dependencies among all variables, 2) it preserves the joint probability distribution of the variables which KNN methods do not promise. Unfortunately, Bayesian cannot afford to support a large size of dataset as it requires to learn a network and discretization of all data accurately. This process is usually required unless conditional probability of Bayesian are explicitly modeled and can be parameterized, which frequently with higher computational expense [17].

A particularly elegant way Bayesian handle missing data is as follows (assuming that *xj* has the missing values):

$$P(x_1...x_j...x_d|y) = P(x_1|y)...P(x_j|y)...P(x_d|y) \quad (3)$$

$$\sum_{xj} P(x_1...x_j...x_d|y) = \sum_{xj} P(x_1|y)...P(x_j|y)...P(x_d|y) \quad (4)$$

$$= P(x_1|y)... \sum_{xj} P(x_1|y)...P(x_d|y) \quad (5)$$

$$= P(x_1|y)...1...P(x_d|y) \quad (6)$$

$$(7)$$

The above equation shows that all prediction of missing values will eventually equal to 1. The Bayesian approach relies on the collection of data then calculating the probability that data is significantly related to the information that was extracted.

The key ingredient of Bayesian approach is treating missing data as added unknown quantities to be able to estimate a posterior distribution. A posterior distribution can be defined as the total knowledge of integration between prior distribution and likelihood function to a parameter after been observed [18]. Regardless, the Bayesian approach helps to easily adapt to include partially adapted observed cases as well as incorporate realistic assumptions for the reasons of missingness of datasets.

In the next section, details on how to evaluate the accuracy of the machine learning techniques described in this section will be provided.

II. Experimental Setup

This section attempts to establish the most appropriate classifiers in relation to the percentage of missing values in a dataset.

Fig. 4 shows the flow of experiment conducted. The first s tep i s w ith a cquiring m edical d ataset f rom data.gov.uk, Canada Open Data, UCI Machine Learning Repository and World Health Organization (WHO).

Second steps emphasize on calculating the percentage of missing values in all ten medical datasets. The objective of this activities is to analyze the most fitting classifier that suits with various percentage of missing values.

Before the real experiment phase begins, all missing values shall be cleaned to prevent problems caused by missing values when training a model [**?**]. For the purpose of this study, we artificially create missing values from a complete data to validate the imputed missing values against actuals. The validation is measured with MAE, MSE, and RMSE. The third step helps to identify what data need to be analyzed. In this phase also identify a different algorithm for developing the rules and classification techniques to concentrate on the missing information that you need. As claimed by Ian H. Witten, Eibe Frank and Mark A. Hall in their book Data Mining: Practical Machine Learning Tools and Techniques, second and third steps should cover the role of implementing processes and decision making that generate ultimately results.

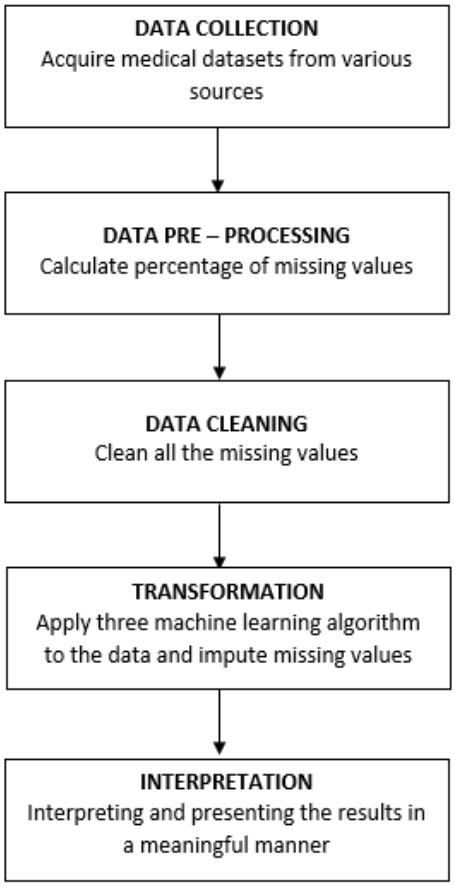

Fig. 4. Experiment Flow.

Next phase covers the identification of relevant values and information, substituting missing data with valid estimations. Besides, this phase should be able to define the appropriate approach to imputing missing values for the medical dataset. The performance of each approach is compared and results presented.

The final step is interpretation step where the results yielded are analyzed. The performance is gathered as an element to validate our hypothesis. In this step, the final results of data imputation is also compiled.

III. Evaluation Criteria

An experiment is conducted to demonstrate the performance of machine learning techniques where ten simulated datasets were acquired and publicly available at: data.gov.uk[1] and Canada Open Data portals[2], UCI Machine Learning Repository[3] and World Health Organization (WHO)[4].

Generally, there are many possible reasons clinical has the most missing values such as patient refusal to answer questions when it related to privacy issues, unable to understand questions given, patient migration, early successes of a treatment,

[1]https://data.gov.uk/
[2]https://open.canada.ca/en
[3]http://archive.ics.uci.edu/ml
[4]http://www.who.int/gho/en/

treatment or instrumental failures, adverse events and death of respondent due to accident or other reasons [16], [19].

All these real life datasets are medical datasets and has missing values due to several reasons mentioned. The percentage of missing value for each dataset are shown in Table I. Table I refers to information regarding the number of records and the amount of missing values (in percentage) are provided along with the data sets.

TABLE I: Summary of Datasets

Dataset	Records	Percentage of Missing Values
Admissions	192	1.56%
Alcohol	39	10.26%
Autism	229	1.4%
Body Mass Index (BMI)	864	1.7%
Drug	458	45.33%
Funerals	60	30%
Infection	1386	19.19%
KPI Health	730	57.22%
Mental Health	108,342	8.07%
Obesity	1458	13.31%

The three machine learning classifiers are evaluated using three criteria: Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid \tag{8}$$

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

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{9}$$

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i^abs - X_i^imputed)^2}{n}} \tag{10}$$

IV. Analysis of Results

This section presents the result of simulations done on the ten datasets with respect to accuracy and percentage missing values. Based on Table II below, the accuracy of each algorithm were compared using three parameters as mentioned in the previous section. These three parameters: MAE, MSE, and RMSE were estimates by observing the lowest values. All these three parameters are negatively-oriented scores, which concludes the lower results the better.

MAE, MSE, and RMSE are the most useful parameters to evaluate the performance of predicting methods and to measure forecast accuracy. Generally, all these parameters are measured on the error difference between the imputed values and actual values.

TABLE II: Results of Machine Learning Classifiers

Dataset	ML Classifier	MAE	MSE	RMSE
Admissions	KNN	5.823	4.924	7.017
	Decision Tree	4.314	2.289	4.784
	Bayesian Network	2.534	1.919	4.381
Alcohol	KNN	7.560	5.424	7.365
	Decision Tree	139.25	41955.2	204.829
	Bayesian Network	507.25	319865.25	565.566
Autism	KNN	2.1207	6.1548	2.4809
	Decision Tree	2.500	11.500	3.391
	Bayesian Network	0.5	1.0	1.0
Body Mass Index (BMI)	KNN	12.4323	346.4292	18.613
	Decision Tree	15.975	270.788	16.456
	Bayesian Network	12.416	418.579	4.150
Drug	KNN	10.691	172.65	13.140
	Decision Tree	11.925	201.057	14.179
	Bayesian Network	23.0377	642.887	25.355
Funerals	KNN	794.25	916747.2	957.47
	Decision Tree	815.965	1206574.0	1098.442
	Bayesian Network	817.49	1248121.39	1117.2
Infection	KNN	1.124	3.003	1.733
	Decision Tree	4.951	2539.14	50.389
	Bayesian Network	6.3534	136.744	11.694
KPI Health	KNN	9.410	3.603	1.898
	Decision Tree	1.253	2.116	4.599
	Bayesian Network	7.573	5.599	2.366
Mental Health	KNN	6.234	1.725	1.313
	Decision Tree	5.988	1.703	1.305
	Bayesian Network	1.039	3.349	1.830
Obesity	KNN	1.124	3.003	1.733
	Decision Tree	4.951	2539.14	50.389
	Bayesian Network	6.353	136.744	11.694

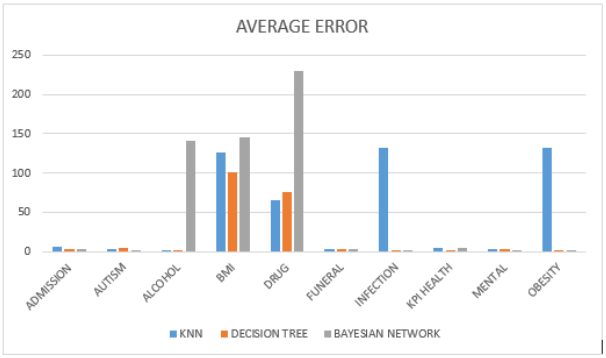

Fig. 5. Average Error for All Datasets.

In accordance with Table II, bayesian has consistently produced the lowest imputation error against all three parameters. This findings in II proves that Bayesian approach is the most appropriate machine learning classifier to impute missing data with regards to smaller sizes of the dataset, less than 20 percent. However, imputation with Bayesian network can be computationally expensive for larger datasets.

Besides, the result drawn from Table II concludes that: the second most standout machine learning classifier is decision tree. Although Bayesian network and decision tree have almost the same results, decision tree is best to apply for larger datasets with higher missing values to imputes.

Nonetheless, KNN also shows the lowest value of error accuracy in some datasets. Surprisingly, the datasets with KNN as the lowest value has a higher percentage of missing values, 30 percent and above. This demonstrates that although KNN consumes time searching through entire datasets, KNN performs better in imputing missing values regardless how big

the size of datasets. Nevertheless, the findings also show that KNN imputation method will never extrapolate outside the range of missing value.

To conclude, the experiments have proved that the proposed machine learning classifiers have a better approach of imputing missing values compared to statistical techniques.

V. Conclusion

In data mining, missing values can be a root cause to produce the wrong final analysis. Besides, in many research area, missing data is a universal problem that may influence the biased estimations and wrong conclusions. To overcome the negative impacts of missing values, a process called missing data imputation should be taken before proceeding to the next phase such as data mining. This paper evaluates three machine learning classifiers namely decision tree, KNN, and Bayesian network, to substitutes missing data and compare each accuracy. The result shows that, the Bayesian network has the lowest value for the three parameters which conclude that the best approach to imputing missing values. However, other factors also influence this error estimators such as percentages of missing values and sizes of datasets. Although Bayesian consistently shows the lowest values, the results are only significant for small sizes of the dataset with less than 20 percent missing values.

VI. Future Work

A future work for imputation in medical dataset must emphasis on optimizing the highest accuracy of a machine learning classifier to impute missing values. This optimization helps to boost machine learning performance for out-of-sample trained using the imputed dataset.

Acknowledgment

This research was supported by the IIUM Research Initiative Grants Scheme (RIGS): RIGS16-346-0510.

References

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

[2] Emran & Nurul A, *Data completeness measures*, *Pattern Analysis, Intelligent Security and the Internet of Things*, 117–130, 2015.

[3] Tahani Aljuaid, & Sreela Sasi, Proper imputation techniques for missing values in data sets *International Conference on Data Science and Engineering (ICDSE)*, 2016.

[4] R. S. Somasundaram, & R. Nedunchezhian, Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values, *International Journal of Computer Applications*, 21(10):14–19, 2011.

[5] Hyun Kang, The prevention and handling of the missing data, *Korean Journal of Anesthesiology*, 402–406, 2013.

[6] Peter Schmitt, Jonas Mandel & Mickael Guedj, A Comparison of Six Methods for Missing Data Imputation *Journal of Biometrics & Biostatistics*, 6(1), 2015.

[7] Marsh, & H. W., "Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes" in *Pairwise deletion for missing data in structural equation models. Structural Equation Modeling: A Multidisciplinary Journal* 5, 22–36, 1998.

[8] Soley-Bori & Marina, Dealing with missing data: Key assumptions and methods for applied analysis, *Boston University*, 2013.

[9] Susianto, Y and Notodiputro, KA and Kurnia, A and Wijayanto, H, "A Comparative Study of Imputation Methods for Estimation of Missing Values of Per Capita Expenditure in Central Java" in *IOP Conference Series: Earth and Environmental Science*, 58(1), 012017, 2017.

[10] Allassonniere, Stéphanie and Kuhn, Estelle, Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation, *Computational Statistics & Data Analysis*, 91, 4–19, 2015.

[11] Rahman, Md Geaur, Islam and Md Zahidul, Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques, *Knowledge-Based Systems*, 53, 51–65, 2013.

[12] Patidar, Preeti and Tiwari, Anshu, Handling missing value in decision tree algorithm, *International Journal of Computer Applications*, 70(13), 2013.

[13] Gavankar, Sachin and Sawarkar, Sudhirkumar, Decision Tree: Review of Techniques for Missing Values at Training, Testing and Compatibility, *Artificial Intelligence, Modelling and Simulation (AIMS), 2015 3rd International Conference on*, 122–126, 2015.

[14] Gerardnico, Data Mining - K-Nearest Neigbors, *CC Attribution-Noncommercial-Share Alike 4.0 International*, 2017.

[15] Beretta, Lorenzo and Santaniello, Alessandro, Nearest neighbor imputation algorithms: a critical evaluation, *BMC medical informatics and decision making*, 16(3):74, 2016.

[16] Kenward, M. G., The handling of missing data in clinical trials, *Clinical Investigation*, 3(3):241–250, 2013.

[17] Liu, Yuzhe and Gopalakrishnan, Vanathi, An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data, *Data*, 2(1):8, 2017.

[18] Glickman, Mark E and Van Dyk, David A, Basic bayesian methods, *Topics in Biostatistics*, 319–338, 2017.

[19] Barga, Roger and Fontama, Valentine and Tok, Wee Hyong and Cabrera-Cordon, Luis, Predictive analytics with Microsoft Azure machine learning, 2015.