Modeling naïve bayes imputation classification for missing data

B K Khotimah\textsuperscript{1,3}, Miswanto\textsuperscript{1,2}, H Suprajitno\textsuperscript{1,2}
\textsuperscript{1}Faculty of Science and Technology, University of Airlangga, Surabaya, Indonesia
\textsuperscript{2}Department of Mathematics, University of Airlangga, Surabaya, Indonesia
\textsuperscript{3}Department of Informatic Engineering, University of Trunojoyo Madura, Bangkalan, Indonesia

E-mail: bain@trunojoyo.ac.id, miswanto@fst.unair.ac.id, herry-s@fst.unair.ac.id

*Corresponding author E-mail: herry-s@fst.unair.ac.id

Abstract. Naïve Bayes Imputation (NBI) is used to fill in missing values by replacing the attribute information according to the probability estimate. The NBI process divides the whole data into two sub-sets is the complete data and data containing missing data. Complete data is used for the imputation process at the lost value. The process is repeated for each missing attribute to generate complete data for classification. This research applies NBI for imputation and preprocessing as preparation of classification process. The trial of this study used NBI for imputation compared to using the mean and mode to predict the missing data. The data used for imputation is full train of complete data as a whole to predict the missing value so as to represent the entire data. The results of this study prove that imputation with NBI produces the right imputation with higher accuracy than other imputations. NBI with single imputation and multiple imputation results in better performance because of the right features. This study aims to calculate the effect of missing values on Naïve Bayes Imputation Algorithm is based on a probabilistic model using mixed data. Empirically shows that the interaction between several methods of imputation and supervised classification results in differences in the performance of classification for the same imputation method.

1. Introduction

Missing data based on dependent characteristics causes have been divided into three categories [1-6][13]. Missing Completely at Random (MCAR) data group occurs when missing data have no relation with other variables. Missing at Random (MAR) is when the missing data has a relationship with other features. So, Missing Not at Random (MNAR) is missing data has relation with other missing data, also they could not make estimation by using existing variable [2]. Preprocessing data by imputation is substituting the lost value by using certain values. The imputation method is grouped by imputation of global constant imputation, conventional imputation and prediction imputation model using machine learning algorithm as development of modern imputation method [10].

The general technique of imputation in missing data in large datasets has used an unsupervised algorithm such as average, median, modus and standard deviation [16]. Research to review lost data is very important a technique in pattern classification that used to measure the relationship between data...
is k-Nearest Network Imputation (k-NNI). The value of k is used to measure the level of similarity of data if the higher the value of k then make the classification becomes more blurred [9]. The weakness of the k-NNI method is that learning outcomes are strongly influenced by features and feature weights that are suitable for data resulting in high computation [18]. The research hybrid method for outlier detection is KNN. This application has two stages in the first stage of the classic DBSCAN algorithm identifying several points as noise points and the second stage all noise points are reconsidered with the KNN classifier. KNN depends on k value and parameters which must be chosen are (i) minimum point (ii) Epsilon (iii) parameter k. Future work will consist of finding methods for optimal automatic selection parameter [19][23][24]. So, Support Vector Machine Imputation (SVMI) can handle linear and nonlinear kernel classification and regression problems on large dimension data issues [12]. SVMI has a weakness that is difficulty determining the optimal parameter values in large dimensional data [12]. Since, Neural Network Imputation can be used on both continuous and category data, but requires advance preprocessing of data with the resulting computations to be longer [7]. Classification to handle missing data is one of them with Naïve Bayes Imputation. Modeling naïve Bayes imputation involves specification of structure and model parameters, result an error in imputation processes that affect classification accuracy [4]. Missing data in datasets comes from data whose attributes have no value. The existence of missing data can also cause duplicate data because there is more than one data with the same name with different data completeness [8][19].

The Naïve Bayes classification specifies values that are missing in an imputation element based on the order of variables one by one. Naïve Bayes depends on the use of the number of features and the selection of elements in the database. Naïve Bayes Imputation (NBI) has a strong feature independence assumption if there is no occurrence of label class, then the estimated probability based on frequency has zero. According to the assumption of conditional independence, if the probability value is multiplied by zero it affects the estimated posterior probability[16][20][21].

NBI is based on a probabilistic model by considering all its values when testing, but can reduce some values to some extent to improve efficiency [17]. Problems with some classification methods can be overcome by using Naive Bayes with weighting because it does not require long computational time and the accuracy may increase.

The study used Naïve Bayes Imputation (NBI) modeling to address missing data on the Naïve Bayes classification for mixed data. The performance of NBI is compare by imputation process with imputation mean, mode, median. NBI is done by using complete data as a whole to predict missing values on incomplete data. This method is expected to find the missing data model according to both category and discrete data. NBI is very suitable for varied data with data on definitive features.

2. Naïve Bayes Imputation (NBI)

The dataset in vector form with the m sequence attribute is expressed \( x^d = [x_{11}, x_{12}, \ldots, x_{im}] \) and the class is shown \( c_j \) consist \( C = \{c_1, c_2, \ldots, c_j\} \). Data containing the missing attributes declared with probability \( P(X_1 = x_1, X_2 = x_2 ...X_j =? ...X_d = x_d|y) \) [3]. Naive Bayes Imputation for handling missing data by using complete data to predict the value of a variable that is missing on incomplete data, thus affecting the calculation of the probability [5]. The probability equation for describing the missing features is as follows:

\[
P(x_1, x_2, ...X_j ...x_d|y) = \Pi_{i\neq j}^d P(x_i|y) \tag{1}
\]

Naïve Bayes imputation for charging empty data depending on the number of data features. If the data types of categories in the small data set that is using a way of charging the value that allows in the same attribute series.

For example, there is a missing value \( \{X_1 = H, X_2 =?, X_3 = T\} \) then in predicting the value on the missing data required two options \( X_1 = H, X_2 = T \) and \( X_1 = H, X_2 = T, X_3 = T \).
Since \( P(x_1, x_2, \ldots, x_j \ldots x_d | y) = P(H, H, T) + P(H, T, T) \) \hspace{1cm} (2)

The compute \( X_j \) which contained missing data according to :

\[
P(x_1, \ldots, x_j \ldots x_d | y) = P(x_1 | y) \ldots P(x_j | y) \ldots P(x_d | y)
\]
\[
\sum_{x_j} P(x_1, \ldots, x_j \ldots x_d | y) = \sum_{x_j} [P(x_j | y)] \ldots P(x_d | y)
\] \hspace{1cm} (3) \hspace{1cm} (4)

They requires identification of probability in each class \( P(x_1 | c_i)P(x_2 | c_i), \ldots, P(x_n | c_i) \), \( x_k \) refers to attribute value \( A_k \) for observation data \( x | [3][4] \). Class determination \( C_j \) on NBI show :

\[
c^* = \operatorname*{arg\,max}_c p(c)p(x_1 | c)p(x_2 | c)p(x_4 | c)
\] \hspace{1cm} (5)

Naïve Bayes Imputation steps for missing value and large data. If the \( X_D \) dataset is represented by a set of vectors \( x^i \) (i = 1, ..., N) with each corresponding class label \( c_i \in C \). The \( X_D \) dataset is the whole data that contains the complete attribute data \( X_C \) and the data with the missing value attribute \( X_M \). So the overall data is stated \( X_D = X_C \cup X_M \). Notation adopted for hypothetical dataset with symbol '?' represents the missing value and \( c \) is the class on x the observed data [1][19].

| No. instance | \( a_1 \) | \( a_2 \) | \( a_3 \) | \( a_4 \) | Class |
|--------------|--------|--------|--------|--------|-------|
| \( x_{1i} \)  | \( x_{11} \) | \( x_{12} \) | \( ? \)  | \( x_{14} \) | \( c_1 \) |
| \( x_{2i} \)  | \( x_{21} \) | \( x_{22} \) | \( x_{23} \) | \( x_{24} \) | \( c_2 \) |
| \( x_{3i} \)  | \( x_{31} \) | \( x_{32} \) | \( ? \)  | \( x_{34} \) | \( c_1 \) |
| \( x_{4i} \)  | \( x_{41} \) | \( x_{42} \) | \( x_{43} \) | \( x_{44} \) | \( c_2 \) |
| \( x_{5i} \)  | \( x_{51} \) | \( x_{52} \) | \( ? \)  | \( x_{54} \) | \( c_1 \) |
| \( x_{6i} \)  | \( x_{61} \) | \( x_{62} \) | \( x_{63} \) | \( x_{64} \) | \( c_2 \) |
| \( x_{7i} \)  | \( x_{71} \) | \( x_{72} \) | \( x_{73} \) | \( x_{74} \) | \( c_1 \) |
| \( x_{8i} \)  | \( x_{81} \) | \( x_{82} \) | \( ? \)  | \( x_{84} \) | \( c_2 \) |
| \( x_{9i} \)  | \( x_{91} \) | \( x_{92} \) | \( x_{93} \) | \( x_{94} \) | \( c_1 \) |
| \( x_{10i} \)| \( x_{101} \)| \( x_{102} \)| \( x_{103} \)| \( x_{104} \)| \( c_1 \)|

Divide data into two ie complete data and incomplete data containing missing data. Preparing training data is complete data for the NB classification process whose value is used for imputation.

| No. instance | \( a_1 \) | \( a_2 \) | \( a_3 \) | \( a_4 \) | \( y(\text{Class}) \) |
|--------------|--------|--------|--------|--------|----------------|
| \( x_{2i} \)  | \( x_{21} \) | \( x_{22} \) | \( x_{23} \) | \( x_{24} \) | \( c_2 \) |
| \( x_{4i} \)  | \( x_{41} \) | \( x_{42} \) | \( x_{43} \) | \( x_{44} \) | \( c_2 \) |
| \( x_{6i} \)  | \( x_{61} \) | \( x_{62} \) | \( x_{63} \) | \( x_{64} \) | \( c_2 \) |
| \( x_{7i} \)  | \( x_{71} \) | \( x_{72} \) | \( x_{73} \) | \( x_{74} \) | \( c_1 \) |
| \( x_{9i} \)  | \( x_{91} \) | \( x_{92} \) | \( x_{93} \) | \( x_{94} \) | \( c_1 \) |
| \( x_{10i} \)| \( x_{101} \)| \( x_{102} \)| \( x_{103} \)| \( x_{104} \)| \( c_1 \)|
b. $X_M$ (incomplete data)

| No. | Attribute | $a_1$ | $a_2$ | $a_3(x_{mis})$ | $a_4$ | $y$ (Class) |
|-----|-----------|-------|-------|----------------|-------|-------------|
| $x_{1i}$ | $x_{11}$ | $x_{12}$ | ?     | $x_{14}$       | $c_1$ |
| $x_{3i}$ | $x_{31}$ | $x_{32}$ | ?     | $x_{34}$       | $c_1$ |
| $x_{5i}$ | $x_{51}$ | $x_{52}$ | ?     | $x_{54}$       | $c_1$ |
| $x_{8i}$ | $x_{81}$ | $x_{82}$ | ?     | $x_{84}$       | $c_2$ |

If data is incomplete $x_{mis}^m$, there is Missing value. the NBI classification will replace the missing value with the $x_{ci}$ value and $a_c$ is the predicted class value.

**Table 3. Sample Data Complete with a grade considered as an attribute**

| No. instance | attribute | $a_1$ | $a_2$ | $a_c$ | $a_4$ | class |
|--------------|-----------|-------|-------|-------|-------|-------|
| $x_{2i}$     | $x_{21}$  | $x_{22}$ | $x_{c2}$ | $x_{24}$ | $x_{23}$ |
| $x_{4i}$     | $x_{41}$  | $x_{42}$ | $x_{c2}$ | $x_{44}$ | $x_{43}$ |
| $x_{6i}$     | $x_{61}$  | $x_{62}$ | $x_{c2}$ | $x_{64}$ | $x_{63}$ |
| $x_{7i}$     | $x_{71}$  | $x_{72}$ | $x_{c1}$ | $x_{74}$ | $x_{73}$ |
| $x_{9i}$     | $x_{91}$  | $x_{92}$ | $x_{c1}$ | $x_{94}$ | $x_{93}$ |
| $x_{10j}$    | $x_{101}$ | $x_{102}$ | $x_{c1}$ | $x_{104}$ | $x_{103}$ |

**Table 4. Sample Data incomplete with missing variables is considered predicted value**

| No. instance | attribute | $a_1$ | $a_2$ | $a_3$ | $a_c$ | $C_{misj}$ |
|--------------|-----------|-------|-------|-------|-------|-------------|
| $x_{1j}$     | $x_{11}$  | $x_{12}$ | $x_{14}$ | $x_{c1}$ | ? |
| $x_{3j}$     | $x_{31}$  | $x_{32}$ | $x_{34}$ | $x_{c1}$ | ? |
| $x_{5j}$     | $x_{51}$  | $x_{52}$ | $x_{54}$ | $x_{c1}$ | ? |
| $x_{8j}$     | $x_{81}$  | $x_{82}$ | $x_{84}$ | $x_{c2}$ | ? |

Naïve Bayes imputation process is used to fill in missing values. Then the process is repeated to calculate every missing attribute, using the equation:

$$P(C_{misj}|X_1 \ldots X_c \ldots X_i) = \frac{P(C_{misj})P(X_1 \ldots X_c \ldots X_i|C_{misj})}{P(X_1 \ldots X_c \ldots X_i)}$$  \hspace{1cm} (5)

$$P(x_1|C_{misj}) = \prod_{k=1}^{n} P(x_1|C_{misj})$$
$$P(x_i|C_{misj}) = P(x_1|C_{misj}) \cdots xP(x_c|C_{misj}) \cdots xP(x_1|C_{misj})$$  \hspace{1cm} (6)

NBI will predict the class label $C$, $P(X|C_j)P(C_j)$ evaluated for each class $C_j$. Then the classifier predicts the X pair in the $C_{misj}$ class if and only if:

$$P(X|C_{misj})P(C_{misj}) > P(X|C_j)P(C_j)$$
\hspace{1cm} for $1 \leq misj \leq m, \ misj \neq j$  \hspace{1cm} (7)

$$v_{mis} = \arg \max_{c \in C} P(C_{misj}) \prod_i P(x_1|C_{misj})$$  \hspace{1cm} (8)
\[ v_{mis} = \arg \max_c P(C_{misj})p(x_1|C_{misj})p(x_c|C_{misj}) ... p(x_1|C_{misj}) \]  

\[ (9) \]

\[ v \text{ is the output value of Naïve Bayes classification as the class label. The process of filling missing data by entering the class prediction value in } x_{mis} \text{ column table:} \]

\[ x_{mis} = v_{mis} \]  

\[ (10) \]

3. Research Methods

The process of eliminating data is done randomly from 5%, 10%, 20%, and 30%. If there is at least one empty value in a record, then the data will be grouped into missing data. After that in preliminary data done preprocessing the selection of data with and conducted the selection of attributes used in the study. The next process is to supplement missing values with Naive Bayes Imputation (NBI) and learning that used NBC. NBI has its own advantages that can handle incomplete or missing data.

Algorithm NBI imputation of mixed data values:

**Begin of algorithm**

**Step-1:** if the data uses mixed data it is necessary to do a z-score transformation.

**Step-2:** Share data to build NBI using imputation attributes \( x_i \) as class \( C_{misj} \).

**Step-3:** Exchange where classes are considered as missing attributes and attributes are considered as the sought value.

**Step-4:** Variable missing data \( X_M \) is used to predict the missing value attribute.

**Step-5:** NBC is used to estimate and replace lost data in an imputation attribute. Calculate \( v_{mis} \) equal the probability value and determine the highest value probability to predict the missing value.

**Step-6:** After the predicted value is missing, then substitute the missing value.

**Step-7:** Complete data is obtained, then the exchange class as an attribute is returned as before.

**Step-8:** classify the complete data with NBC to find out its performance

**Step-9:** Calculate RMSE for each final result of Naïve Bayes Classification (NBC). If the classification performance is bad, it requires proper data partitioning, because NBC depends on training data.

**End of algorithm**

4. Results and Discussion

The result of data submission use random missing value by using program Rapid Miner. The imputation process in the data is used to generate complete data for the classification process. Data imputation results differ by type of data used. If the data is discrete then the NBI imputation process uses Probability Gaussian formula to calculate the probability prior. While the prediction results in the value of class labels in the form of discrete numbers. The data for the test is correlated with according to the Table 5.

Techniques to determine the relationship between variables in the data then used correlation test. High correlation results so that made missing value that will affect the performance of classification. Iris data was no correlation relationship between variable and flexible. If the missing value is done dynamically the result is shown in Table 2. The result the Pearson's correlation test results \( r \) of 0.788 with Sig. (2-tailed) of 0.00 or less than 0.05. So it is decided sepal length with petal length and petal width correlated significantly and significantly. So the process of data loss can be done on one of the arbitrary variables.
The entire data contained missing data

Separate the complete data and incomplete data

Swapping missing attributes into predictable labels

First training data

Calculate posterior probability with NBI

Calculate the greatest value of the greatest probability

Performing data by using the largest posterior probability

First testing data

Combining data into whole data

Complete data

Partition data for NBC learning

Classification with NBC

Model NBI

End

Figure 1. Flowchart Model Naïve Bayes Imputation

Table 5. The Result Correlations of Iris

|               | sepal length | sepal width | petal length | petal width |
|---------------|--------------|-------------|--------------|-------------|
| sepal length  | Pearson      | -1.109      | .872         | .818**      |
|               | Correlation  |             |              |             |
|               | Sig. (2-tailed) | .183      | .000         | .000        |
|               | N            | 150        | 150          | 150         |
| sepal width   | Pearson      | -.109      | 1            | -.421**     |
|               | Correlation  |             |              |             |
|               | Sig. (2-tailed) | .183      | .000         | .000        |
|               | N            | 150        | 150          | 150         |
| petal length  | Pearson      | .872**     | -.421**      | 1           |
|               | Correlation  |             |              |             |
|               | Sig. (2-tailed) | .000      | .000         | .000        |
|               | N            | 150        | 150          | 150         |
Table 6. Experiment has been used several variables to find out some influential variables. In the data whose value experienced a high difference value resulted in high data distribution. Petal length was the highest standard deviation value is found because the data has a significant difference.

Table 6. The spread of Mean and Standard Deviation data for imputation

| Variables     | Pre Imputation | Post Imputation |
|---------------|----------------|-----------------|
|               | Mean           | Std. Deviation  | Mean | Std. Deviation |
| sepal length  | 5.84333        | 0.82800         | 6.7433| .82800         |
| sepal width   | 3.05400        | 0.43359         | 4.2570| .43359         |
| petal length  | 3.75866        | 1.7644          | 4.9586| 1.76442        |
| petal width   | 1.19866        | 0.76316         | 3.1759| 1.0260         |

Table 7. Result RMSE of research for Iris data

| RMSE Imputation | NBI | Modus | Median | Mean |
|-----------------|-----|-------|--------|------|
| 0.05            | 0.22| 0.35  | 0.58   | 0.76 |
| 0.10            | 0.44| 0.66  | 1.21   | 1.32 |
| 0.20            | 0.76| 1.17  | 1.61   | 1.91 |
| 0.30            | 1.07| 1.61  | 1.97   | 2.37 |

Table 8. Result RMSE of research for Localization Data for Person Activity

| RMSE Imputation | NBI | Mean | Modus | Median |
|-----------------|-----|------|-------|--------|
| 0.05            | 0.68| 1.32 | 1.81  | 1.76   |
| 0.10            | 1.24| 6.66 | 8.21  | 5.32   |
| 0.20            | 2.76| 9.17 | 10.61 | 6.81   |
| 0.30            | 3.09| 11.61| 12.97 | 9.37   |
Simulation results by filling the data with Naïve Bayesian Imputation using the lost value at every 5%, 10%, 20%, 30% showing RMSE results with different percentages. Missing data can reduce prediction model accuracy. The experimental results of the RMSE score on the NBI versus Mean, Median and Standard Deviations show that the RMSE value has increased a clear decrease trend. If the proportion of lost data exceeds 20%, there is a clear error increase in the prediction accuracy. Methods for processing lost data should be chosen carefully to eliminate negative impacts on classification and optimize classification performance. The second test result of the data set The more the amount of data the increased error generated. Fig. 2 and 3 shows the result classification using imputed lost data with NBI and unattended machine learning techniques such as Mean, Median and STD without missing values. The results show that NBI produces the smallest RMSE compared to other methods. NBI replaces the lost data by predicting the value by using the chance value of the
criteria in the complete data. NBI uses the missing value in variable ≤1 because NBI was not be used for some missing values on some variables. The reason charging data on the NBI depends on preprocessing the initial data. The more variables were used on the training data the more valid the missing replacement value.

| Table 9. The results of the research test on the comparison method |
|---------------------|---------------------|---------------------|---------------------|---------------------|
| Percentage          | Type data           | RMSE Imputation     |                     |
|                     |                     | NBI     | Mean    | Modus   | Median  |
| Wine Localization  | Discrete            | 0.76    | 1.17    | 1.61    | 1.91    |
| Data for Person     | Real                | 2.76    | 9.17    | 10.61   | 6.81    |
| Activity            | Discrete and categories | 12.76 | 13.86   | 15.67   | 6.51    |

The results of the trial based on the type of data indicate mixed data needs to be value general. The trick is to change the category data into real values, in addition to normalizing that valued not too far away. The results show results for mixed data showing a higher error. This paper presents a lost and efficient data handling method for the NBC model and is superior to some imputations. In general, NBC requires the selection of relevant attributes to produce NBC that are irrelevant recommended.

5. Conclusión

The Naïve Bayes Imputation model can be used for varying data ie continuous data and categories. Since, method determines the performance of NBI which tested the comparison of two data. The result NBI can be used to overcome the differences in the data. When the NBI is compared with mean, mode and median imputation have produced varying values. The best results are proven by using NBI with some variation of training data obtained by appropriate values for imputation. The application of missing data for the result of the error value depends on the amount of training data. The more missing data then the training data used will be less. So, the more training data will be resulted the more valid the test that showed the smaller error. The next study looked for methods for automatic selection that were optimal for finding parameters. while the use of features requires the right selection techniques to produce efficient performance.

Acknowledgement

We thank the promoters and co-promoters for valuable suggestions and comments to improve quality towards success completion of my studies at Airlangga University. paper. The author also thanked the editor in chief at ICONSME for valuable suggestions in improving the quality of the manuscript.

References

[1] Joaquín A and Javier G C 2017 Improving the Naive Bayes classifier via a quick variable selection method using maximum of entropy Entropy 19( 6) 1-17.
[2] Amanda N, Baraldi and Craig K E 2009 An introduction to modern missing data analyses Journal of School Psychology. Society for the Study of School Psychology 48 (1) 5–37.
[3] Garcia A J T and Hruschka E R 2005 Naïve Bayes as an imputation tool for classification problems Proceedings - HIS 2005: Fifth International Conference on Hybrid Intelligent Systems 2005, pp. 497–499. doi: 10.1109/ICHIS.2005.78.
[4] Grade G F 2017 Naïve Bayesian Classifier 5(2) 237–241.
[5] Cuiping L, Wang S and Wang H 2009 Learning naive Bayes classifiers with incomplete data 2009 International Conference on Artificial Intelligence and Computational Intelligence AICI 4 350–353.

[6] Little R J A and Rublin D 1987 Statistical Analysis with Missing Data (Wiley, New York).

[7] López-Molina T, Pérez-Méndez A and Rivas-Echeverría 2008 Missing values imputation techniques for Neural Networks patterns 12th WSEAS International Conference on SYSTEMS 290–295.

[8] Lüdtke O, Robitzsch A and Grund Psychological Methods Multiple Imputation of Missing Data in Multilevel Designs: A Comparison of Different Strategies Multiple Imputation of Missing Data in Multilevel Designs: A Comparison of Different Strategies American Psychological Association. 1-25.

[9] Malarvizhi and Thanamani A S 2012 K-Nearest Neighbor in Missing Data Imputation International Journal of Engineering Research and Development 5(1) 5–7.

[10] Nakagawa S and Freckleton R P 2008 Missing inaction: the dangers of ignoring missing data Trends in Ecology and Evolution 23(11) 592–596.

[11] Pelckmans K, De Brabanter J, Suykens J A K and De Moor B 2005 Handling missing values in support vector machine classifiers Neural Networks 18 684–692.

[12] Yang B, Janssens D, Ruan D, Cools M, Bellemans T and Wets G 2011 A Data Imputation Method with Support Vector Machines for Activity-Based Transportation Models Foundations of Intelligent Systems, AISC 122 249–257.

[13] Gajawada S and Toshniwal D 2012 Missing Value Imputation Method Based on Clustering and Nearest Neighbours International Journal of Future Computer and Communication 1(2) 206208.

[14] Loukopoulos P, Sampath S and Pilidis P 2016 Dealing with missing data for prognostic purposes Prognostics and System Health Management Conference (PHM- Chengdu).

[15] Garcíaarena U and Santana R 2017 An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers Expert Systems With Applications 89 52–65.

[16] Kanchara S and Thanamani A S 2016 Elevating the Accuracy of Missing Data Imputation Using Bolzano Classifier International Journal of Engineering and Technology (IJET) 8(1) 138-145.

[17] Yelipe U R, Porika S and Golla M 2017 An efficient approach for imputation and classification of medical data values using class-based clustering of medical records Computers and Electrical Engineering 000 1–18.

[18] Gajawada S and Toshniwal D 2012 Missing Value Imputation Method Based on Clustering and Nearest Neighbours International Journal of Future Computer and Communication 1(2) 206208.

[19] Tripathy S and Sahoo L 2017 Efficient Outlier Detection by Integration of Clustering and Classification International Journal of Engineering and Technology (IJET) 9(4) 2766-2770.

[20] Kuppusamy Vand Paramasivam I 2016 An Analysis and Comparison of Various Missing Data Imputation Tools and Techniques International Journal of Engineering and Technology (IJET) 8(5) 1910-1915.

[21] Pedro J, García-Laencina, Luis J, Gomez S, Aníbal R, Vidal F 2009 Pattern classification with missing data: a review Neural (Comput & Applic, Springer-Verlag London Limited).

[22] Vatanen T, Osmala M, Raiko T, Lagus K, Sysi-Aho M, Orešič M, Honkela T and Lähdesmäki H 2015 Self-organization and missing values in SOM and GTM Neurocomputing 147 60–70.

[23] Iskandar N, Nasyrov, Idar I, Nasyrov, Rustam I, Nasyrov, Bulat A and Khairullin 2018 Data Mining for Information Storage Reliability Assessment by Relative Values International Journal of Engineering & Technology 7(4.7) 204-208.

[24] Saar-Tsechansky M, Provost F 2007 Handling Missing Values when Applying Classification Models Journal of Machine Learning Research 8 1625-1657.