JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 7, No. 4, October 2023, pp. 1167-1177

MICE Implementation to Handle Missing Values in Rain Potential Prediction Using Support Vector Machine Algorithm

Aina Latifa Riyana Putri[1], Bayu Surarso[2], Titi Udjiani SRRM[3]
[1,2,3]Departement of Mathematics, University of Diponegoro, Indonesia
ainalatif47@gmail.com[1], bayus@lecturer.undip.ac.id[2], udjianititi@lecturer.undip.ac.id[3]

Article History:
Received : 20-07-2023
Revised : 21-09-2023
Accepted : 22-09-2023
Online : 12-10-2023

Keywords:
Adversity quotient;
Reflective thinking;
Problem-solving;
Pythagorean.

ABSTRACT

Support Vector Machine (SVM) is a machine learning algorithm used for classification. SVM has several advantages such as the ability to handle high-dimensional data, effective in handling nonlinear data through kernel functions, and resistance to overfitting through soft margins. However, SVM has weaknesses, especially when handling missing values in data. The use of SVM must consider the missing values strategy chosen. Missing values in data mining is a serious problem for researchers because it causes many problems such as loss of efficiency, complications in data handling and analysis, and the occurrence of bias due to differences between missing data and complete data. To overcome the above problems, this research focuses on understanding the characteristics of missing values and handling them using the Multiple Imputation by Chained Equations (MICE) technique. In this study, we utilized secondary data experiments that contain missing values from the Meteorological, Climatological, and Geophysical Agency (called BMKG) related to predictions of potential rain, especially in DKI Jakarta. Identification of types or patterns of missing values, exploration of the relationship between missing values and other variables, incorporation of the MICE method to handle missing values, and the Support Vector Machine Algorithm for classification will be carried out to produce a more reliable and accurate prediction model for rain potential. It shows that the imputation method with the MICE gives better results than other techniques (such as Complete Case Analysis, Imputation Method Mean, Median, Mode, and K-Nearest neighbor), namely an accuracy of 89% testing data when applying the Support Vector Machine algorithm for classification.

https://doi.org/10.31764/jtam.v7i4.16699

A. INTRODUCTION

Support Vector Machine (SVM) is a machine learning algorithm used for classification. In this algorithm, an optimal hyperplane is found that can separate different classes in a dataset. The main goal of the Support Vector Machine (SVM) algorithm is to find the optimal hyperplane that maximizes the margin, namely the perpendicular distance between the hyperplane and the closest data point of each class. These data points are known as support vectors. Support Vector Machine (SVM) has been widely used in various applications including the classification of English text and documents in research Luo (2021), using 1033 datasets and obtaining accuracy exceeding 90% when using more than 4000 features. In the research of Vijayarajeswari et al. (2019), Support Vector Machine (SVM) in the field of bioinformatics is used to classify mammogram images for the detection of breast cancer using 95 datasets images and it was found that Support Vector Machine (SVM) is an effective algorithm for classifying abnormal classes in cancer cases breast. Support Vector Machine (SVM) has several advantages such as

the ability to handle high-dimensional data Gaye et al. (2021), is effective in handling nonlinear data through kernel functions Chen et al. (2016), and resistance to overfitting through soft margins (Xu et al., 2020). However, SVM has weaknesses, especially when handling missing values in data (Stewart et al., 2018). So according to Stewart et al. (2018), the use of SVM must consider the missing values strategy chosen.

Missing values are a common for data sets in real problems (Alamoodi et al., 2021). Missing values in a data set refer to the absence of data points in one or more variables from a data set. The existence of missing values can be caused by many things such as data collection errors, data corruption, or incomplete data entry. There are various types of missing values such as Missing Completely at Random (MCAR) as missing values that occur randomly and are not related to all other variables, Missing at Random (MAR) as missing values that occur randomly but mathematically there is a relationship with the response variable/observations, and Missing Not at Random (MNAR) as missing values that occur non-randomly, have a pattern, and are related to other variables. The existence of missing values in data usually requires a preprocessing stage where the data is prepared and cleaned so that it can produce knowledge. To handle missing values, several general approaches can be used, such as removing data (or Complete Case Analysis) by removing rows or columns with missing values from the data set Bartlett et al. (2014) and imputation by filling in missing values using the mean or median Jadhav et al. (2019), k-Nearest Neighbor (KNN) imputation Jadhav et al. (2019) and Multiple Imputation by Chained Equation (MICE) (Z. Zhang, 2016).

Missing values in data mining is a serious problem for researchers because it causes many problems such as loss of efficiency, complications in data handling and analysis, and the occurrence of bias due to differences between missing data and complete data (Luengo et al., 2012). In the case of classification, continuing the Support Vector Machine (SVM), the presence of missing values can interfere with the performance of the algorithm. There are reasons for how this could happen, such as missing values can affect the hyperplane construction by creating gaps or changing the feature space distribution. This can cause the classification limit to be not optimal. The effectiveness of the general missing values approach chosen can also affect the performance of the Support Vector Machine (SVM) (Stewart et al., 2018). Therefore, it is important to understand the analysis of missing values and apply the right strategy to handle them before applying the data mining algorithm (in this study the Support Vector Machine algorithm) so that the validity and reliability of the analysis results are maintained.

To overcome the above problems, this research will focus on understanding the characteristics of missing values and handling them using the Multiple Imputation by Chained Equations (MICE) technique. The MICE method handles imputed missing values based on their relationship with other variables in the dataset. This method involves the repetition of several steps, where each variable step that has a missing value is imputed based on linear regression with available predictor variables. Then the process will be repeated until convergence is reached. This study uses online data experiments that contain missing values from Meteorology, Climatology and Geophysics Agency (BMKG) related to predictions of potential rain, especially in DKI Jakarta. In the context of predicting rainfall with a Support Vector Machine (SVM), missing value analysis becomes very important to ensure the integrity of the input data. Identification of types or patterns, exploration of the relationship between missing

values and other variables, incorporation of the MICE method to handle missing values, and the Support Vector Machine Algorithm for classification will be carried out to produce a more reliable and accurate prediction model for rain potential. A comparison was also made between MICE and other missing values handling techniques such as Complete Case Analysis, Mean, Mode, Median, and kNN imputation by creating several Support Vector Machine (SVM) models. The value of the confusion matrix will be seen with the highest model accuracy value. Confusion Matrix is a table used to evaluate the performance of a classification model (Navin J R & R, 2016). The findings from this research are expected to contribute to the development of better forecasting models and to increase understanding of the relationship between missing values and the performance of the Support Vector Machine (SVM) model in predicting rainfall potential.

B. METHODS

Experimental trials for the analysis of missing values in this study used online data (https://dataonline.bmkg.go.id/home) from the Meteorology, Climatology and Geophysics Agency (BMKG), especially at the Tanjung Priok Maritime Meteorological Station, North Jakarta from 01 January 2019 to 10 November 2022 with parameters that can be seen in Table 1. The daily climate data used in this study has 10 variable parameters.

Table 1. Variable Description

	Variable	Description
1	ddd_x	The wind direction at maximum speed (°)
2	ff_x	Maximum wind speed (m/s)
3	ff_avg	Average wind speed (m/s)
4	RH_avg	Average Humidity (%)
5	ss	Length of sunshine (hours)
6	Tx	Maximum temperature (°C)
7	Tn	Minimum temperature (°C)
8	Tavg	Average temperature (°C)
9	ddd_car	Most wind direction (°)
10	RR	Rainfall (mm)

The original dataset in Table 1 will be downloaded and saved as a file with a .xlsx extension. The data will also describe information related to attribute variables, target variables, and existing missing values. Table 1 totals 1410 observations and 10 variables, the first 8 variables being numeric data types and the last 2 variables being categorical data types. Furthermore, the original dataset in Table 1 has to modify the variable, where the variables used in this study consist of the target variable (Y) and the predictor variable (X). Predictor variables are determined by 9 parameter variables and the target variable ("RR") will show on one day whether it rains or not.

The original dataset that contains missing values will be handled using MICE techniques. That dataset will be identified regarding patterns, types of missing values (e.g. MCAR, MAR, or MNAR), and their implications by applying relevant statistics or visualization techniques. MICE techniques to imputing missing values with iterations until convergence should also be achieved. After getting the result in the form of complete data, apply the classification with the

Support Vector Machine, which will be checked by the confusion matrix to see the model's performance. Also apply other missing values handling techniques such as Complete case Analysis and Imputation Method (Mean, Median, Mode, and K-Nearest neighbor). Classify predictions using SVM models with imputed datasets from each of these techniques. Compare the accuracy of the confusion matrix of models using MICE imputation with those of models using other imputation techniques.

1. Missing Values

Missing values can be interpreted as missing or unavailable data or information regarding research subjects on certain variables. (Little & Rubin, 2019)divides three types of missing values, namely: Suppose $y_i = (y_{i1}, \dots, y_{ip})^T$ that $i = 1,2,\dots,n$ where n is the sample size, is a vector sequence with p and $Y = (y_1, \dots, y_n)^T$ is the data matrix n x p. Observed entries and entries containing missing values are expressed as Y_o and Y_m respectively Y. Let $r_i = (r_{i1}, \dots, r_{ip})^T$ that show as an indicator whether each component in the vector y_i is observed, i.e. $r_{ik} = 1$ if y_{ik} it is observed and $r_{ik} = 0$ if y_{ik} are the missing values for $i = 1,2,\dots,n$ and $k = 1,2,\dots,p$. $Y = (y_1, \dots, y_n)^T$is a stacked matrix r (Li et al., 2013).

a. Missing Completely at Random (MCAR)
   In this type, missing values occur randomly for all observation units and are independent, meaning they are not related to the values of all other variables (missing values or the observed variables) in the observation. Then the MCAR assumption can be defined as:

$$\Pr(R|Y_m, Y_0) = \Pr(R) \tag{1}$$

   This equation implies that the indicator of the missing value does not depend on the data containing the missing values and the observed data.

b. Missing at Random (MAR)
   In this type, the missing values have a systematic relationship with the observed variables. This type of missing value can be predicted based on other data available in observation. While MAR can be defined as:

$$\Pr(R|Y_m, Y_0) = \Pr(R|Y_0) \tag{2}$$

   In other words, the distribution of missing value indicators depends on the observed data.

c. Missing Not at Random (MNAR)
   In this type, missing values are not randomly distributed. MNAR is the most complicated missing value because the missing value in a variable is related to the missing values of the variable itself so it cannot be predicted from the variables in a dataset. MNAR can be defined as:

$$\Pr(R|Y_m, Y_0) = \Pr(R|Y_m) \tag{3}$$

In other words, the distribution of missing values indicators depends on the missing values variable itself. The Rstudio software has packages that are available and ready to be used for handling missing values such as MICE (Multivariate imputation by chained equations). The MICE procedure follows a series of regression models that are run, where each variable from the missing values is conditionally modeled on other variables in the data.

2. Support Vector Machine

In this study, we will focus on the case of binary classification. Consider the problem of binary classification with two classes, labeled -1 and +1. Given a training dataset with data points 'n' and corresponding labels, denoted as $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ where 'x_i' is a feature vector and 'y_i' is a class label (-1 or +1). The resulting model on Support Vector Machine can be used to predict or classify a data x with a decision function that can be formulated as follows.

$$f(x) = d(y(x)) = d(\sum_n \alpha_n y_n k(x_n, x) + b) \quad (4)$$

where function $d(y(x))$ is:

$$d(y(x)) = \begin{cases} 1, \text{jika } y(x) > 0 \\ -1, \text{jika } y(x) < 0 \end{cases}$$

3. Confusion Matrix

The Confusion Matrix is a table used to evaluate the performance of a classification model (Navin J R & R, 2016). The confusion matrix visualizes model performance by comparing the predicted classes with the actual classes of the data set, as shown in Table 2.

Table 2. Confusion Matrix

	True Class	
Predicted Class	**Positive (c_1)**	**Negative (c_2)**
Positive (c_1)	TP (True Positive)	FP (False Positive)
Negative (c_2)	FN (False Negative)	TN (True Negative)

These metrics provide insights about the performance of the model, such as accuracy. The mathematical equation of accuracy can be seen in Equation (5).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (5)$$

C. RESULT AND DISCUSSION

1. Data Collection

In this study, an experiment was designed and implemented from a case using the primary data described in the previous chapter, which is related to the prediction of the potential for rain to describe the analysis of missing values. In Table 3 the average and percentage of data subjects containing missing values for each variable (numerical and categorical variables) are

shown. The percentage of data subjects containing missing values ranged from as low as 0.9% "ddd_car" to 31% "RR". It can also be seen based on the plot in the figure below to make it easier for the reader to understand the pattern of missing values for each variable, as shown in Table 3 and Figure 1.

Table 3. Description of Data Statistics

Variable	Mean	Size of Subject Data with Values	Size of Subject Data Containing Missing Values	Percentage of Subject Data Containing Missing Values
Variable Numerical				
Tn	26.25	1309	101	7.2%
Tx	32.14	1357	53	3.8%
Tavg	28.71	1393	17	1.2%
RH_avg	77.8	1393	17	1.2%
ss	5.802	1352	58	4.1%
ff_x	5.387	1396	14	1%
ddd_x	166	1396	14	1%
ff_avg	2.246	1396	14	1%
Variable Categorical				
ddd_car		1397	13	0.9%
RR		972	438	31,3%

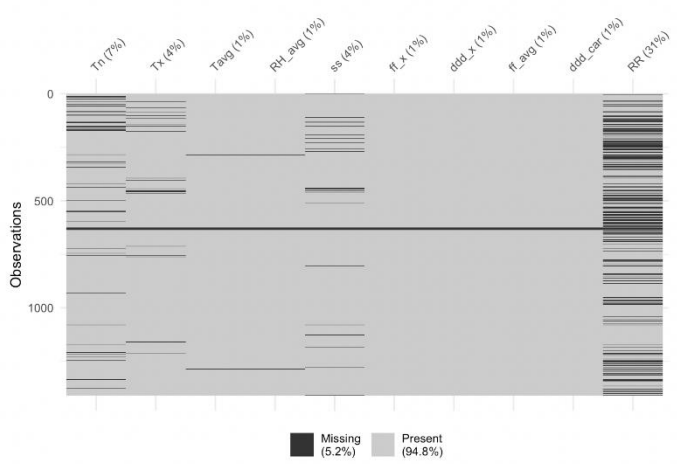

Figure 1. The plot of The Missing Values Pattern

2. Handling Missing Values with MICE

In this study, the technique for handling missing values imputation uses MICE. In the imputation method, including missing values prevents potentially useful information from being lost. The specification of the imputation model obtained is the most challenging step in imputation, especially MICE, including the following:

a. The missing values assumption must be decided beforehand whether they are plausible MAR (Missing at Random) types or not because MICE can only handle MAR (Missing at Random) (Ahn et al., 2021). Finch (2021) wrote that MICE resulted in biased results on the MNAR data. The MAR assumption itself is a suitable type of missing value and can be handled in many practical cases but needs to be suspected due to its imputation which must be plausible. In determining the type of missing values from the data obtained, an

analysis of the relationship of each variable is carried out from the data subjects that have values and the data subjects that contain missing values. To determine whether the data are MAR, MAR testing can be done by statistical tests using the logistic regression approach on each variable that has missing values, one of which is as shown below for the "RR" variable.

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.8448468  3.6218126  -1.062  0.28843
## Tn           0.2962315  0.0960238   3.085  0.00204 **
## ss           0.0314969  0.0242256   1.300  0.19355
## Tx          -0.0518625  0.1021966  -0.507  0.61182
## Tavg         0.0873758  0.1627176   0.537  0.59128
## RH_avg      -0.0727669  0.0160446  -4.535 5.75e-06 ***
## ff_x         0.0305462  0.0469473   0.651  0.51527
## ddd_x       -0.0003783  0.0007157  -0.529  0.59705
## ff_avg      -0.1555583  0.1291423  -1.205  0.22838
## ddd_carE    -0.0344908  0.2447290  -0.141  0.88792
## ddd_carN     0.3873459  0.3602607   1.075  0.28229
## ddd_carNE    0.1823306  0.2193635   0.831  0.40587
## ddd_carNW   -0.9929150  0.4440597  -2.236  0.02535 *
## ddd_carS     0.3849493  0.3539669   1.088  0.27680
## ddd_carSE    0.3087495  0.2666069   1.158  0.24684
## ddd_carSW   -1.3970840  1.0641950  -1.313  0.18925
## ddd_carW    -0.3072618  0.2672054  -1.150  0.25018
```

Figure 2. Logistic Regression MAR Testing

In Figure 2, it can be seen that the relationship between the missing values for the "RR" variable has mostly p-values with other variables greater than 0.05. This test provides evidence that the missing values in the variable "RR" are related to several other variables, implying that the missing values are MAR with a significance level of 0.05. By Rouzinov & Berchtold (2022), for the type of missing values MAR, some of the information obtained by the variables can at least partially explain the presence (has a relationship) with the missing values. Furthermore, it is also very important to check diagnostically whether value imputation is plausible or not (Bondarenko & Raghunathan, 2016). Unreasonable data imputed values (such as negative values at variable “hour” etc.) must not appear in the data. A diagnostic check on the data for each variable will provide a way to check the reasonable value of the imputation.

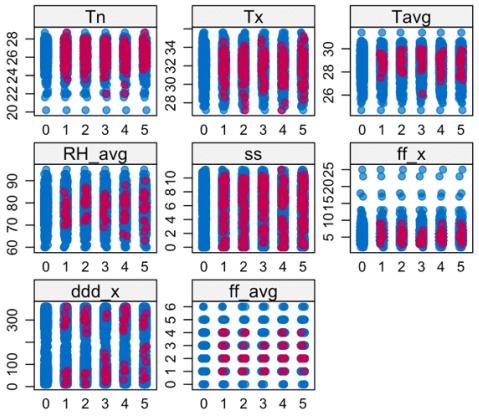

Figure 3. Plot to Check The Plausible Value of Imputation

In Figure 3, the red dot represents the value that was just imputed while the blue dot represents the value from the original data (data that has values and missing values). It can be seen that the red dot follows the blue dot nicely. A suitable and precise pattern like the image shows that the imputed value is indeed a reasonable (plausible) value.

b. Referring to the imputation of the missing values obtained, it should be checked the distribution of the original data (data that has values and missing values) and data that has just been imputed (Nguyen et al., 2017). There is a scatterplot for each imputed variable as shown in Figure 4.

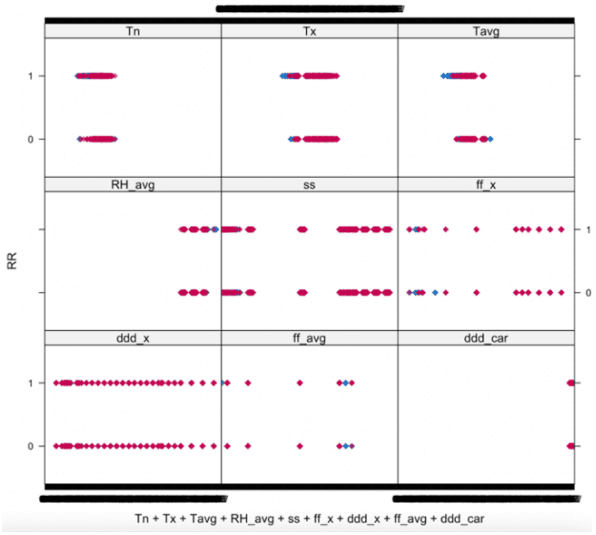

Figure 4. Distribution Data

In Figure 4 the imputed data values are colored red and the values from the original data (data that have values and missing values) are colored blue. In MAR-type data it is assumed that the red dots have almost the same pattern as the blue dots (the distribution is assumed to be identical). It can be seen that the blue dots and red dots are constant across the data set which means they represent certainty regarding the true value of the missing values.

c. Based on the specified number of iterations, MICE convergence should also be monitored. To determine whether the MICE algorithm has converged to plot one ("RR") or more parameters against the specified number of iterations (in this study iterations= 5) as shown in Figure 5.

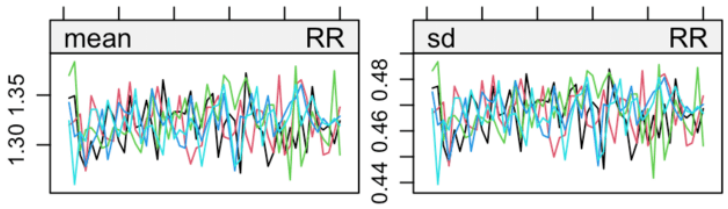

Figure 5. MICE Convergence

At the desired convergence, the streams must be distinct and free to mingle with each other without showing a definite trend. Convergence is diagnosed when the variance between the different sequences is not greater than the variance with each sequence. Figure 5 shows the mean (left side) and standard deviation (right side) of the imputation for the "RR" variable. It can be obtained that the variable "RR" has a healthy convergence because there is a trend that the streams mix very well from the start.

3. Implementation of Handling Missing Values Each technique with the Support Vector Machine Algorithm

This study will also explain the differences between other missing values handling techniques (such as Complete Case Analysis, Imputation Method Mean, Median, Mode, and K-Nearest neighbor) based on accuracy from experimental data. In terms of flexibility with data types, MICE can handle various types of data including continuous, categorical, or binary variables (Mera-Gaona et al., 2021). In contrast to the mean Wissler et al. (2022), median Hunt (2017), and K-Nearest neighbor imputation methods which can only handle continuous variables. The imputation mode method can only handle categorical or binary variables (Triguero et al., 2019). Then in terms of missing handling mechanisms values as described in the previous section, MICE can handle MAR (Missing at Random) and MCAR (Missing Completely at Random) missing values. MICE is also relatively easy to implement Zhai & Gutman (2022) in statistical software that automates the iterative imputation process and analyzes imputation datasets.

Good or bad estimates with various techniques will be analyzed about the accuracy obtained from the Support Vector Machine classification algorithm. Accuracy is simply measured by how likely an algorithm can correctly predict negative and positive events. The performance of each missing values handling technique can be seen in the following table, after being tried to be implemented with the Support Vector Machine algorithm, as shown in Table 4.

Table 4. Accuracy of Various Missing Value Handling

No.	Handling Missing Value Technique	Accuracy Training Data	Accuracy Testing Data
1	Mode-Mean	0.8434	0.5252
2	Mode-Median	0.7344	0.695
3	Complete Case Analysis	0.8247	0.8736
4	K-Nearest Neighbor	0.8723	0.8511
5	MICE	0.8322	0.8936

The results in Table 4, illustrate that the accuracy value is closely related to the selection of missing values handling techniques. A comparison of the accuracy of the training data and data testing needs to be done to ensure that the model obtained is not underfitting or overfitting. Underfitting is a condition that occurs when the model is unable to capture the variability of the data (H. Zhang et al., 2019). As a result, it creates very bad patterns when using training data. Meanwhile, overfitting is a condition that occurs when the model has a low error during training but functions very poorly when predicting new data (Santos et al., 2021). So when the

accuracy of the data training significantly outperforms data accuracy testing, then there will be a high probability of overfitting. In Table 3, it can be seen that the Mode-Mean, Mode-Median, and K-Nearest Neighbor techniques show a lower level of accuracy for testing data compared to training data. Therefore, it was concluded that these three techniques were not suitable for application, because there was overfitting in the model. On the other hand, the Complete Case Analysis and MICE techniques show that the accuracy of the testing data is higher than the accuracy of the training data, indicating that there is no overfitting or underfitting in the model. However, it should be noted that the MICE technique shows higher testing data accuracy, namely 89% compared to the Complete Case Analysis technique. Therefore, of the five techniques for handling missing values in applying the SVM algorithm for classification, the choice fell on MICE.

D. CONCLUSION AND SUGGESTIONS

Based on the results obtained, it shows that the rain potential prediction data is a type of missing value MAR using a logistic regression approach. To overcome the missing values, the MICE method is used. It shows that the imputation method with the MICE gives better results than other techniques (such as Complete Case Analysis, Imputation Method Mean, Median, Mode, and K-Nearest neighbor), namely an accuracy of 89% testing data when applying the Support Vector Machine algorithm for classification. The accuracy value is closely related to the selection of missing values handling techniques.

REFERENCES

Ahn, H., Sun, K., & Kim, K. P. (2021). Comparison of Missing Data Imputation Methods in Time Series Forecasting. *Computers, Materials & Continua*, *70*(1), 767–779. https://doi.org/10.32604/CMC.2022.019369

Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., Almahdi, E. M., Chyad, M. A., Tareq, Z., Albahri, A. S., Hameed, H., & Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Systems with Applications*, *167*. https://doi.org/10.1016/J.ESWA.2020.114155

Bartlett, J. W., Carpenter, J. R., Tilling, K., & Vansteelandt, S. (2014). Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics*, *15*(4), 719–730. https://doi.org/10.1093/BIOSTATISTICS/KXU023

Bondarenko, I., & Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, *35*(17), 3007–3020. https://doi.org/10.1002/SIM.6926

Chen, J., Zhang, X., & Gao, Y. (2016). Fault detection for turbine engine disk based on an adaptive kernel principal component analysis algorithm. *Http://Dx.Doi.Org/10.1177/0959651816643670*, *230*(7), 651–660. https://doi.org/10.1177/0959651816643670

Finch, H. (2021). Cite this article: Holmes FW. A Comparison of the Heckman Selection Model, Ibrahim, and Lipsitz Methods for Deal-ing with Nonignorable Missing Data. *J Psychiatry Behav Sci*, *4*(1), 1045. http://meddocsonline.org/

Gaye, B., Zhang, D., & Wulamu, A. (2021). Improvement of Support Vector Machine Algorithm in Big Data Background. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/5594899

Hunt, L. A. (2017). Missing Data Imputation and Its Effect on the Accuracy of Classification. *International Federation of Classification Societies*, *0*, 3–14. https://doi.org/10.1007/978-3-319-55723-6_1

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, *33*(10), 913–933. https://doi.org/10.1080/08839514.2019.1637138

Li, C., Li, & Cheng. (2013). Little's test of missing completely at random. *Stata Journal*, *13*(4), 795–809. https://EconPapers.repec.org/RePEc:tsj:stataj:v:13:y:2013:i:4:p:795-809

Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data. *Statistical Analysis with Missing Data*, 1–449. https://doi.org/10.1002/9781119482260

Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, *32*(1), 77–108. https://doi.org/10.1007/S10115-011-0424-2/METRICS

Luo, X. (2021). Efficient English text classification using selected Machine Learning Techniques. *Alexandria Engineering Journal*, *60*(3), 3401–3409. https://doi.org/10.1016/J.AEJ.2021.02.009

Mera-Gaona, M., Neumann, U., Vargas-Canas, R., & López, D. M. (2021). Evaluating the impact of multivariate imputation by MICE in feature selection. *PLOS ONE*, *16*(7), e0254720. https://doi.org/10.1371/JOURNAL.PONE.0254720

Navin J R, M., & R, P. (2016). Performance Analysis of Text Classification Algorithms using Confusion Matrix. *International Journal of Engineering and Technical Research (IJETR), 6(4), 75-8.* www.erpublication.org

Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model checking in multiple imputation: An overview and case study. *Emerging Themes in Epidemiology*, *14*(1), 1–12. https://doi.org/10.1186/S12982-017-0062-6/TABLES/5

Rouzinov, S., & Berchtold, A. (2022). Regression-Based Approach to Test Missing Data Mechanisms. *Data 2022, Vol. 7, Page 16*, *7*(2), 16. https://doi.org/10.3390/DATA7020016

Santos, A. E. M., Lana, M. S., & Pereira, T. M. (2021). Evaluation of machine learning methods for rock mass classification. *Neural Computing and Applications*, *34*(6), 4633–4642. https://doi.org/10.1007/S00521-021-06618-Y

Stewart, T. G., Zeng, D., & Wu, M. C. (2018). Constructing support vector machines with missing data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*(4), e1430. https://doi.org/10.1002/WICS.1430

Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(2), e1289. https://doi.org/10.1002/WIDM.1289

Vijayarajeswari, R., Parthasarathy, P., Vivekanandan, S., & Basha, A. A. (2019). Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. *Measurement*, *146*, 800–805. https://doi.org/10.1016/J.MEASUREMENT.2019.05.083

Wissler, A., Blevins, K. E., & Buikstra, J. E. (2022). Missing data in bioarchaeology II: A test of ordinal and continuous data imputation. *American Journal of Biological Anthropology*, *179*(3), 349–364. https://doi.org/10.1002/AJPA.24614

Xu, C., Tannant, D. D., Zheng, W., & Liu, K. (2020). Discrete element method and support vector machine applied to the analysis of steel mesh pinned by rockbolts. *IJRMM*, *125*, 104163. https://doi.org/10.1016/J.IJRMMS.2019.104163

Zhai, R., & Gutman, R. (2022). *A Bayesian Singular Value Decomposition Procedure for Missing Data Imputation*. https://doi.org/10.6084/M9.FIGSHARE.20405770.V1

Zhang, H., Zhang, L., & Jiang, Y. (2019). Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems. *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*. https://doi.org/10.1109/WCSP.2019.8927876

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of Translational Medicine*, *4*(2). https://doi.org/10.3978/J.ISSN.2305-5839.2015.12.63