Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2022, 2022:66

https://doi.org/10.28919/cmbn/7538

ISSN: 2052-2541

OVERCOMING MISSING VALUES USING IMPUTATION METHODS IN THE CLASSIFICATION OF TUBERCULOSIS

EKA MALA SARI ROCHMAN[1,2,*], MISWANTO[1], HERRY SUPRAJITNO[1]

[1]Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia.

[2]Department of Informatics, Faculty of Engineering, University of Trunojoyo Madura, Bangkalan, Indonesia

Abstract: Indonesia is one of the countries with the highest population density in the world with a very high number of Tuberculosis (TB). This TB disease is very serious because it is very easily transmitted through the air, namely, droplets that come from a TB patient who coughs or sneezes. In diagnosing a disease, missing data often occurs, resulting in researcher errors in the data collection process, so this study proposes the mean Imputation method to overcome missing data. For the classification of TB disease data in Bangkalan Regency, Indonesia, which consists of 886 data, the method used is Naive Bayes compared to Logistics Regression. For the distribution of training and testing data, this research uses multiple trains and tests K-Fold cross-validation with a total of k=10. Based on research trials using the mean imputation method is better than the one imputation method in filling in the missing data for this case with an average accuracy is 97.36% and the F1 score is 95.01% better than one imputation with an average accuracy is 97.35% and F1 score is 94.35 % on the Naive Bayes method. For TB classification, the Naive Bayes method produces an average accuracy is 97.36% and the F1 score is 95.01% better than the logistic regression method in classifying tuberculosis with an accuracy rate is 97.36% with an F1 score is 89.58%.

[*]Corresponding author

E-mail address: miswanto@fst.unair.ac.id

Received June 7, 2022

Keywords: tuberculosis; imputation; missing value; classification; Naive Bayes; logistics regression.

2010 AMS Subject Classification: 92C60.

1. INTRODUCTION

According to the World Health Organization (WHO), TB is a dangerous disease because it is one of the top ten causes of death worldwide. Indonesia is one of the highest countries most infected with tuberculosis [1] and [2]. This disease comes from various strains of mycobacteria, namely Mycobacterium tuberculosis. Tuberculosis generally attacks the lungs, but can also have an impact on other body parts and most TB sufferers are people with an age range of 15-65 years so around 80% of TB bacteria attack the lungs [3].

Tuberculosis in Indonesia has become very serious because it is very easy to transmit. The disease, which used to be known as TB, has now become pulmonary TB, which can be transmitted through the air. Usually, when the patient sneezes or coughs, the bacteria will come out so that other healthy people can inhale it.

Diagnosing requires complete data because it affects the results of an accurate diagnosis. Problems in a data process contained in data mining are called missing data. Missing data or missing data or a missing value is a situation where there are empty values or incomplete values in the data. The missing data phenomenon happens very often in many kinds of research. Missing data can occur due to several things including the error of researchers who collect data, limited data collection tools, program errors when collecting data, and so on [4] and [5].

The technique for dealing with missing values is the Average Imputation. Where the attribute containing the missing value is replaced by calculating the mean of the missing value. Several imputation methods have been developed to minimize the negative impact of missing values, including by replacing them with a maximum or minimum value of zero, or one.

Classification of disease data in the medical is an important task in predicting disease, it can even help doctors in making decisions about diagnosing the disease, thus it is very important to make an early diagnosis to reduce TB transmission to the wider community. Many researchers have carried out activities in predicting TB disease with data mining classification methods, but it is not yet known what method is the most accurate in classifying TB disease.

Classification is an important technique in deep data mining which covers all areas of life [6]. Data mining is the process of extracting data from a database using certain techniques to gain new knowledge or information. Techniques used to extract new information in data mining include estimation, clustering, association, prediction, and classification [6] and [7]. Classification techniques can be carried out using various methods including the Naive Bayes classifier (NBC) classification method. Naïve Bayes is a simple and effective method for prediction [8]. This classification method estimates the probability value of each class and classifies it linearly in an effective way [9]. This classification technique can be used to assist in identifying diseases, so this technique is suitable to be applied to diagnosis. The process to implement the Naive Bayes Classifier method on the identification system is done by collecting some training data containing the experience of the body conditions of a large number of people to classify.

On this occasion, the research that will be carried out is to impute tuberculosis data using average imputation and then classify it using the Naive Bayes algorithm compared to Logistic Regression.

2. PRELIMINARIES

Data mining is a process of extracting data or filtering data by utilizing a collection of data through a series of processes to obtain information on the data using pattern recognition technology such as statistics and mathematical techniques [10]. Data mining is also a method used in the large-scale data processing. Data mining is often referred to as knowledge discovery in database (KDD) which includes data collection, historical use to determine regularity patterns and relationships between large data sets. Based on the task, data mining is grouped into 6 namely descriptions, estimation, prediction, classification, clustering, and association [11]. The stages carried out in the data mining process, are shown in Figure 1.

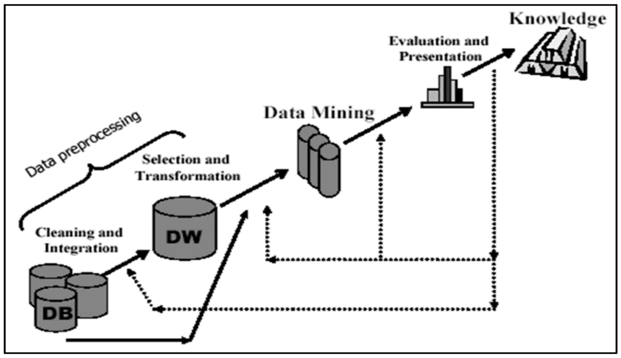

Figure 1 Stages of Data Mining

In Figure 1 there are the first and second stages called stages preprocessing, the third is the mining stage, and the fourth is the evaluation stage.

2.1. Preprocessing

Preprocessing stage is the stage to prepare data so that it can be used for the mining process as needed. At this stage, there is the handling of Missing Value data and data transformation.

A. Cleaning and Imputation

The missing value is a condition where there are empty values or incomplete values in the data [4]. In the TB data, there are some missing data on the features of the chest X-ray, TCM (Molecular Rapid Test) results, HIV status, and history of diabetes. To handle the missing value, an imputation technique was used. The imputation method with the value of Mean and Mode is an imputation method from a simple statistical method, the missing value is replaced with a reasonable estimated value (one estimate per *missing value*) before being entered into the whole [12].

The next missing data method is to replace it with a maximum, minimum, zero, or one value: As the name suggests, replace the missing attribute value with the maximum or minimum training observation value or replace the missing value with zero or one.

B. Transformation Data

Data transformation is a way to change data so that the data is following the needs so that the mining process can be carried out. One technique for data transformation is Discretization and normalization. Discretization helps to get accurate and easier results [13].

Discretization is the process of converting continuous attribute values into a finite number of intervals and associating each interval with a discrete numeric value. Min-Max data normalization is the stage before starting the clustering. Min-Max is a method by doing the linear transformation to raw data [14]. Intending to describe the value of each variable to the same range, namely [0,1], it enters the normalization process. Min Max normalization is shown by equation (1)

$$x_n = \frac{x_0 - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Where:

x_n = normalized data

x_0 = data to be normalized

x_{min} = minimum value of all data

x_{max} = maximum value of all data

2.2. Mining Process

The mining process involves the sharing of data. The dataset is divided into several k partitions. Then the learning process is carried out k times to get the model. In each process, the kth partition is used as test data, and the rest of the other partitions as learning data. In this data separation, it is called a validation technique, namely K-Fold Cross Validation [15] and [16].

Data sharing is done using K-Fold Cross Validation with a value of k= 10 [15] The following illustration of 10 -fold cross-validation can be seen in Figure 2.

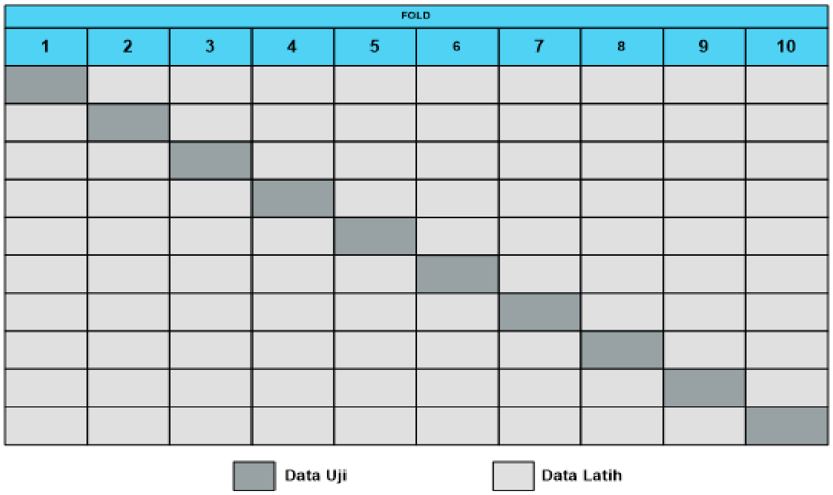

Figure 2 illustration of 10 *-fold cross-validation*

The way *k-fold cross-validation works* is as follows:

1. Total data is divided into *k* parts
2. *the fold* is when the 1st part becomes test data (*testing data*) and the rest becomes training data (*training data*).
3. Repeat step 2 until you reach the *k* - *fold*

Data *mining* itself can be interpreted as a process of looking for patterns in data using certain techniques or methods. Methods in data *mining* vary widely. Among them are classification with Naive Bayes and logistic regression,

A. Naive Bayes

Naïve Bayes Classifier is a classification method using *probability* and statistical methods, this method is a popular machine learning method and has been widely used in predicting various types of diseases [9], [17], and [18]. The main characteristics of this method are very strong (naive) assumptions that are independent of each event. This method certainly has several advantages and disadvantages, which are as follows [19]:

a. Advantages *of Naive Bayes:*

1. Can be used on both quantitative and qualitative data
2. Does not require a large amount of data
3. Does not require a lot of *training data*
4. If there is a missing value, it can be ignored in the calculation
5. Fast and efficient calculation
6. Can be used for binary or *multiclass problem classification.*
7. Document classification can be personalized, and tailored to the needs of each person.

b. Disadvantages *of Naive Bayes:*

1. If the conditional probability is zero, then the prediction probability will be zero.
2. The assumption of each independent variable causes a decrease in accuracy. This is because

there is a correlation between one variable and another.

3. The level of accuracy cannot be measured using only one probability.

The equation of the *Naïve Bayes method* is as follows:

$$P\ (H|X)\ =\ \frac{P\ (X|H)\ .PH}{P(X)} \tag{2}$$

Where:

X : Data with unknown class

H : Hypothesis data

P(H|X) : Probability of hypothesis H based on condition X

P(H) : Hypothesis probability H

P(X|H) : Probability of X based on the conditions on the hypothesis H

P(X) : Probability X

The classification process requires a number of instructions to determine the appropriate class for the sample to be analyzed. Therefore, the *Naïve Bayes method* above can be adjusted as in equation 3 below:

$$P(C|F1\ldots.Fn) = \frac{P(C)\ P(F1\ldots Fn|C)}{P(F1\ldots Fn)} \tag{3}$$

Where the variable C is a class, while the variable F1 ... Fn is a characteristic of the instructions needed for classification. The equation explains that the probability of entering a characteristic sample in class C (posterior) is the probability of having a sample characteristic in class C (likelihood) which is then divided by the probability of the appearance of a global sample characteristic (evidence). Therefore, equation 2 can be written simply as in equation 4:

$$Posterior\ =\ \frac{prior\ x\ likelihood}{evidence} \tag{4}$$

The value of evidence is always fixed in each class of each sample. Posterior values will be compared with other class posterior values. This is for the classification of sample classes. Then the assumption of independence is used. With these assumptions, equation 5 is obtained, which is as follows:

$$P\ (F_i|F_j)\ =\ \frac{P\ (F_i \cap F_j)}{P(F_j)}\ =\ \frac{P\ (F_i)P(F_j)}{P(F_j)}\ = P(F_i) \tag{5}$$

Equation 5 is a model of the Naïve Bayes theorem which will then be used in the classification process. For continuous data classification, you can use the Gaussian Density equation as in equation 6

$$P\ (X_i\ =\ x_i|\ Y\ =\ y_j)\ =\ \frac{1}{\sqrt{2\pi\sigma_{ij}}}\ e^{-\frac{(x_i--\mu_{ij})^2}{2\,\sigma^2 ij}} \qquad (6)$$

Where:

P : Opportunity

Xi : Attribute i

xi : Value of attribute i

Y : The class you are looking for

y i : Subclass Y you are looking for

µ : Average off all attribute (mean)

σ : A variance of all attributes (Standard Deviation)

The mean can be generated using equation 7, which is as follows:

$$\mu\ =\ \frac{x_1+x_2+x_3+\ldots+x_n}{n} \qquad (7)$$

Where:

xi : the value of the i-th sample

n : number of samples

The standard deviation value (standard deviation) can be generated using the following 8 equations:

$$\sigma\ =\ \sqrt{\frac{\sum_{i=1}^{n}(x_i-\ \mu)^2}{n-1}} \qquad (8)$$

After calculating the average value and standard deviation, the next step is to find the probabilistic value by calculating the amount of data that matches the amount of data in that category.

B. Logistic Regression

The logistic regression method is one of the classification methods. In this case, TB has 6 attributes so it can be applied using logistic regression methods [21] and [22]. Logistic regression is part of the regression analysis used when the dependent variable (response) is a dichotomous variable [21] and [23]. Dichotomous variables usually consist of only two values that represent the occurrence or absence of an event which is usually assigned a number 0 or 1. Unlike ordinary linear regression, logistic regression does not assume a linear relationship between the independent and dependent variables. Logistic regression is a non-linear regression where the specified model will follow a linear curve pattern.

There are several advantages of this method, here are some of the advantages of logistic regression which has several advantages over other analytical techniques, namely:

1. Logistic regression does not have normality and heteroscedasticity assumptions on the independent variables used in the model so the classical assumption test is not needed even though the independent variables are more than one.
2. The independent variables in logistic regression can be a mixture of continuous, district, and dichotomous variables.
3. Logistic regression does not require the limitations of the independent variables.
4. Logistic regression does not require that the independent variables be in the form of intervals.

The logistic model is expressed in the form of a probability model where the response variable in this model is the logit of the probability of an attribute that will apply with the condition or condition of the presence of certain independent variables. The *logistic function* has a range of $0 \leq y \leq 1$. The following *logistic regression* model is found in equation (9).

$$y = f(x) = \frac{1}{1+e^{-x}} \tag{9}$$

Where :

$$x = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_dx_d \tag{10}$$

w 0 = intercept

w1 … .w d = weight of variable 1 to d

From the information above, it can be seen in equation (11).

$$f(x;\omega) = \frac{1}{1+e^{-(\omega_0+\omega_1x_1+\omega_2x_2+\cdots+\omega_dx_d)}} \tag{11}$$

The formula for updating the weights using Gradient Descent can be seen in equation (12).

$$\omega_j = \omega_j + \eta(y^i - f(x^i))f(x^i)(1 - f(x^i))x_j^i \tag{12}$$

Where :

$w_j = weight\ to\ j$

η = learning rate

y^i = target

$f(x^i)$ = prediksi

x_j^i = fitur to j with data i

2.3. Evaluation

Evaluation is the process of testing the performance of the method or algorithm used. In general, this performance evaluation uses a confusion matrix [5]. Evaluation with confusion matrix is used to calculate accuracy in data mining. The confusion matrix is depicted by a table that states the number of test data that is correctly classified and the number of test data that is incorrectly classified. The higher the accuracy value, the better the resulting model. Recall calculations, precision and accuracy are contained in equations (13) - (16) [24].

The recall is used to measure the positive pattern that has been classified correctly.

$$recall = \frac{TP}{TP+FN} \tag{13}$$

Precision is used to measure the positive pattern that was predicted correctly from all the total positive predicted patterns.

$$Precision = \frac{TP}{TP+fp} \tag{14}$$

Accuracy is used to measure the ratio of the correct prediction results to the total amount of data being evaluated

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

The F1-Score value or also known as the F-Measure is one of the evaluation calculations that combines recall and precision. Recall and precision values in a situation can have different weights. The best value for F1-Score is 1.0 and the worst value is 0. In representation, if the F1-Score has a good score, it indicates that our classification model has good precision and recall.

$$F1-score\ =\ 2x\ \frac{Precision *\ Recall}{Precision + recall} \qquad (16)$$

3. MAIN RESULTS

A. Data Collection

The data used in this classification process is tuberculosis data from Syarifah Ambami Hospital Rato Ebhu Bangkalan Regency, Indonesia consists of 886 records and 6 attributes, namely age, gender, chest X-ray, HIV status, history of diabetes mellitus (DM), TCM (Rapid Molecular Test) results, and anatomical location [3]. Of the seven attributes, the anatomical location is the target, namely pulmonary TB or extra pulmonary TB.

The discretization process is carried out by converting TB data into a numeric form so that it can be processed in the programming process. The dataset can be seen in Table 1, which at the beginning was in the form of text data, which was converted into categorical data. For age, it is a continuous numeric data so it doesn't need to be changed, but for age, the data normalization process is carried out so that the data range is the same.

Table 1. Data Set Description

Age	**Numerical continuous**		
Gender	L (Male) : 0	P (Female) : 1	
Thoracic Photos	Negative : 0	Positive: 1	
HIV Status	NR/Negative : 0	Positive : 1	
Diabetes history	No : 0	Yes: 1	
TCM Results	Sensitive Rif:0	Negative:1	Rif Resistance:2
Anatomical Location	Lung: 0	Extra Lung: 1	

B. Analysis

Figure 3. describes the IPO diagram as follows:

1. Input Stage

   The input stage begins by entering data as input for the classification process. The data entered is based on the features of TB disease, which are 6 features.

2. Preprocessing Stage

   The process stage is divided into 2, namely preprocessing and processing. In the preprocessing section, the missing value is handled for each blank data using the Mean Imputation method which is compared by giving a value of 1 to the missing value. Then proceed with changing the data into continuous categorical and numerical data. Furthermore, the data normalization process is carried out.

3. The stage of sharing training and testing data

   In the process section, the training and testing data is divided using K-Fold Cross-Validation.

4. Process Stage

   Learning by Naïve Bayes Algorithm to get a classification model. The model obtained is used as an alternative to predict the test data. In this case, the Naive Bayes method will be compared with the logistic regression algorithm to get the results of the classification of tuberculosis whether it is pulmonary TB or extra pulmonary TB.

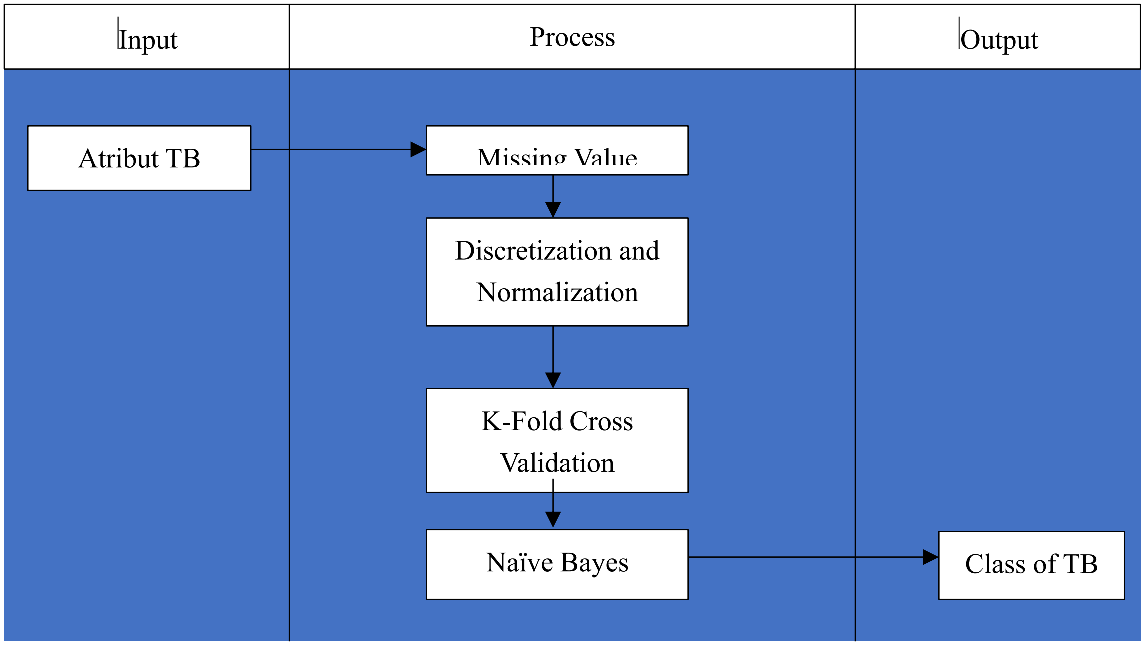

Figure 3 IPO Diagram

4. RESULT AND DISCUSSION

A. Naive Bayes

The flow of the Naïve Bayes method is to first read the training data, then calculate the number and probability, if numerical data is used, then look for the mean and standard deviation of each parameter which is numerical data. The results of the Naïve Bayes trial using the Mean. The imputation method can be seen in Table 2 while the results of trials using imputation one can be seen in Table 3.

Table 2. Results of the Naïve Bayes trial with k=10 and imputation mean

Fold	Accuracy	Precision	Recall	F1 Score
1	94.94%	82.75%	100%	90.56%
2	97.97%	93.33%	100%	96.55%
3	98.98%	96%	100%	97.95%
4	97.97%	91.66%	100%	95.56%%
5	93.93%	79.31%	100%	88.46%
6	98.97%	96%	100%	97.95%
7	97.95%	91.30%	100%	95.45%
8	98.97%	96.42%	100%	98.18%
9	95.91%	88.23%	100%	93.75%
10	97.95%	91.66%	100%	95.65%
Average	97.36%	90.67%	100%	95.01%

Table 3 shows the best accuracy value is in the 3rd fold with an accuracy value is 98.98%. for the average in the trial using 10 fold produces an accuracy is 97.36% and an F1 score is 95.01%.

Table 3. Results of the Naïve Bayes trial with 10 folds and one imputation

Fold	**Accuracy**	**Precision**	**Recall**	**F1 Score**
1	97.97%	100%	100%	96.29%
2	98.98 %	96%	100%	97.95%
3	97.97%	93.33%	100%	96.55%
4	94.94%	82.75%	100%	90.56%%
5	97.97%	92.85%	100%	96.29%
6	96.93%	86.36%	100%	92.68%
7	96.93%	90.62%	100%	95.08%
8	95.91%	85.71%	100%	92.30%
9	97.95%	91.30%	100%	95.45%
10	97.95%	92.30%	100%	96%
Average	97.35%	90.41%	100%	94.91%

Table 3 shows the best accuracy value is in the 2nd fold with an accuracy value is 98.98%. for the average in the trial using 10 fold produces an accuracy is 97.35% and an F1 score is 94.91%

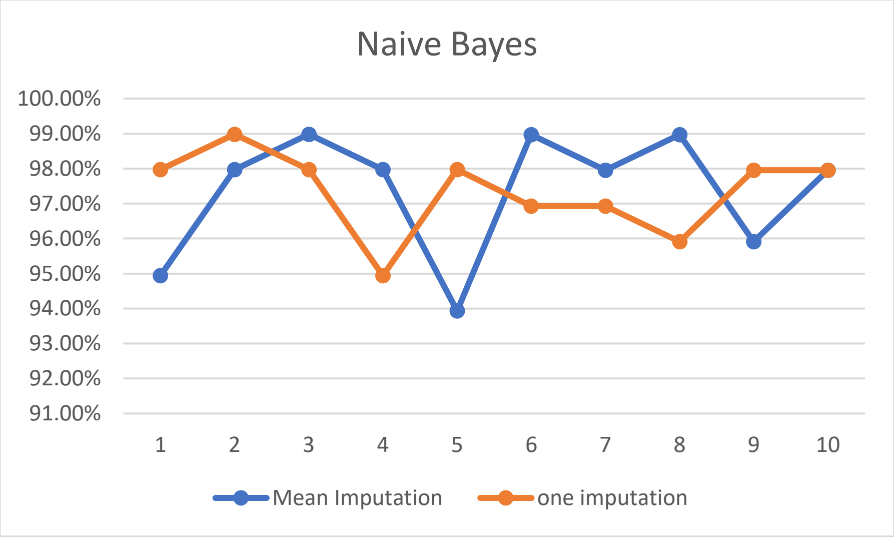

Figure 4. Comparison of Accuracy with Imputation Mean and One in Naïve Bayes

Figure 4 shows the difference using the Naive Bayes classification algorithm with imputed mean getting accuracy, recall precision and F1 score in the 3rd fold. While on imputation one, the best fold is at k=2. However, from the graph, it can be seen that the accuracy value tends to be high at the imputation mean.

B. Logistics regression

For the test results of the logistic regression method using the imputation method, the highest mean is in the 2nd fold with an accuracy value of 100%. Changes in the value of accuracy in this logistic regression method are in the range of 80%-100%. As for the imputation value of one, this method produces the highest accuracy value in the 2nd fold. However, this one imputation method also produces an accuracy that is far from the other folds, so that it reaches 40.47% in the 10th fold position. For the difference in accuracy in the logistic regression method using the imputation mean method compared to the imputation one, it can be seen in Figure 5.

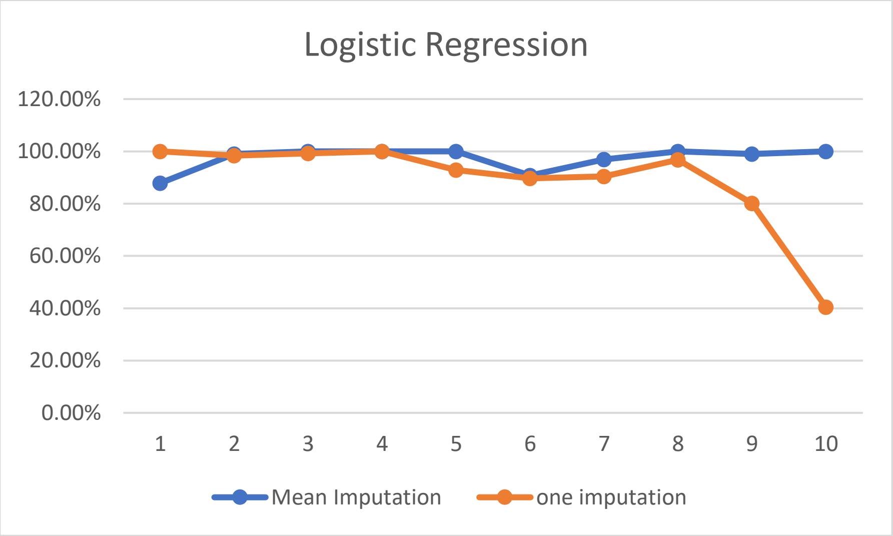

Figure 5. Comparison of Accuracy with Imputation Mean and One on Logistic Regression

Figure 2 shows the difference using the logistic regression classification algorithm with imputation mean to get accuracy, recall precision, and F1 score in the 3rd fold. While on imputation one, the best fold is at k=2. However, from the graph, it can be seen that the accuracy value is more stable and gets a high value on the imputation mean.

To compare the accuracy level between Naive Bayes and Logistic Regression in the classification process, the average accuracy level can be seen in Tables 4 and 5.

Table 4. Imputation using the mean value

No	Method	Precision(%)	Recall(%)	F1 score (%)	Accuracy(%)
1	Naive bayes	90.67	100	95.01	97.36
2	Logistics Regression	86.23	100	89.58	97.36

Table 5. Imputation using the one value

No	Method	Precision(%)	Recall(%)	F1 score(%)	Accuracy(%)
1	Naive bayes	90.41	100	94.91	97.35
2	Logistics Regression	52.71	56.07	53.94	88,88

The two tables above show that the measurement using the Naïve Bayes classification is better than the logistic regression using either the imputed mean or using the imputation one. Table 1 shows that even with the same accuracy value, the F1 score shows that Naive Bayes is better. While table 2 shows the accuracy value which is only slightly different, the value of the F1 score of Naive Bayes is also greater than the logistic regression.

5. Conclusion

Based on the analysis and discussion that has been carried out on training and testing data using 10 folds where there are 886 TB data with 6 attributes, namely age, gender, chest X-ray, HIV status, history of diabetes mellitus (DM), TCM results (Rapid Molecular Tests). then conclude:

1. Based on Tables 4 and 5, the imputation means method better than the imputation one method in filling in the missing data for this case with an average accuracy is 97.36%, and

the F1 score is 95.01% better than imputation one method with an average accuracy is 97.35% and F1 score is 94.35% on the Naive Bayes method. This is because by giving the same value to the data, the mining algorithm will find it interesting. After all, it forms the same value so that it results in unfavorable results.

2. For the measurement of accuracy, the logistic regression method can produce an accuracy value is 100% compared to Naive Bayes which produces an accuracy rate is 98.98% in classifying TB.
3. For the average level of accuracy obtained based on table 4, the Naive Bayes method produces an average accuracy of 97.36% and the F1 score is 95.01% better than the logistic regression method in classifying tuberculosis with an accuracy rate is 97.36% with an F1 score is 89.58%.

ACKNOWLEDGMENT

The researcher is grateful to the promoter and co-promoter who always guide the author can complete this research. The author does not forget to thank Airlangga University for being allowed to be part of the Doctoral program at the Faculty of Saint and Technology (FST).

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] A. Rachmad, N. Chamidah, R. Rulaningtyas, Mycobacterium tuberculosis images classification based on combining of convolutional neural network and support vector machine, Commun. Maths. Biol. neurosci. 2020 (2020), Article ID 85. https://doi.org/10.28919/cmbn/5035.

[2] A. Rachmad, N. Chamidah, R. Rulaningtyas, Mycobacterium tuberculosis identification based on color feature extraction using expert system, Ann. Biol. 36 (2020), 196-202.

[3] E.M.S. Rochman, Miswanto, H. Suprajitno, Comparison of clustering in tuberculosis using fuzzy c-means and

k-means methods, Commun. Maths. Biol. neurosci. 2022 (2022), Article ID 41. https://doi.org/10.28919/cmbn/7335.

[4] I.B.G.N. Giriputra, Missing value imputation using KNN method optimized with memetic algorithm, e-Proc. Eng. 3 (2016), 1098–1105.

[5] A. Kowarik, M. Templ, Imputation with the R package VIM, J. Stat. Soft. 74 (2016), 1-16. https://doi.org/10.18637/jss.v074.i07.

[6] M.M. Saritas, A. Yasar, Performance analysis of ANN and naive Bayes classification algorithm for data classification, Int. J. Intell. Syst. Appl. Eng. 7 (2019), 88–91. https://doi.org/10.18201/ijisae.2019252786.

[7] M.S. Chen, J. Han, P.S. Yu, Data mining: an overview from a database perspective, IEEE Trans. Knowl. Data Eng. 8 (1996), 866–883. https://doi.org/10.1109/69.553155.

[8] M. Langarizadeh, F. Moghbeli, Applying naive bayesian networks to disease prediction: a systematic review, Acta Inform. Med. 24 (2016), 364-369. https://doi.org/10.5455/aim.2016.24.364-369.

[9] S. Spino, M.M. Sathik, S.S. Nisha, The prediction of heart disease using naive Bayes classifier, Int. Res. J. Eng. Technol. 6 (2019), 373–377.

[10] S.M. Gorade, A. Deo, P. Purohit, A study of some data mining classification techniques, Int. Res. J. Eng. Technol. 4 (2017), 3112-3115.

[11] J. Han, M. Kamber, Data mining: concepts and techniques, Morgan Kaufmann Publishers, Burlington, MA, 2011.

[12] E. Acuña, C. Rodriguez, The treatment of missing values and its effect on classifier accuracy, in: D. Banks, F.R. McMorris, P. Arabie, W. Gaul (Eds.), Classification, Clustering, and Data Mining Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004: pp. 639–647. https://doi.org/10.1007/978-3-642-17103-1_60.

[13] C.J. Tsai, C.I. Lee, W.P. Yang, A discretization algorithm based on class-attribute contingency coefficient, Inform. Sci. 178 (2008), 714–731. https://doi.org/10.1016/j.ins.2007.09.004.

[14] D. Borkin, A. Némethová, G. Michaľčonok, K. Maiorov, Impact of Data Normalization on Classification Model Accuracy, Research Papers Faculty of Materials Science and Technology Slovak University of Technology. 27 (2019), 79–84. https://doi.org/10.2478/rput-2019-0029.

[15] Z. Nematzadeh, R. Ibrahim, A. Selamat, Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques, in: 2015 10th Asian Control Conference (ASCC), IEEE, Kota

Kinabalu, 2015: pp. 1–6. https://doi.org/10.1109/ASCC.2015.7244654.

[16] J.G. Moreno-Torres, J.A. Saez, F. Herrera, Study on the impact of partition-induced dataset shift on k-fold cross-validation, IEEE Trans. Neural Netw. Learning Syst. 23 (2012) 1304–1312. https://doi.org/10.1109/tnnls.2012.2199516.

[17] X. Liu, R. Lu, J. Ma, L. Chen, B. Qin, Privacy-preserving patient-centric clinical decision support system on naïve Bayesian classification, IEEE J. Biomed. Health Inform. 20 (2016), 655–668. https://doi.org/10.1109/jbhi.2015.2407157.

[18] N. Boyko, K. Boksho, Application of the naive bayesian classifier in work on sentimental analysis of medical data. In: The 3rd International Conference on Informatics & Data-Driven Medicine (IDDM 2020), Växjä, Sweden, November 19–21, pp. 230–239 (2020). http://ceur-ws.org/Vol-2753/paper16.pdf.

[19] S.D. Jadhav, H.P. Channe, Comparative study of K-NN, naive Bayes and decision tree classification techniques, Int. J. Sci. Res. 5 (2016), 1842-1845.

[20] A. Larasati, C. DeYong, L. Slevitch, The application of neural network and logistics regression models on predicting customer satisfaction in a student-operated restaurant, Procedia – Soc. Behav. Sci. 65 (2012), 94–99. https://doi.org/10.1016/j.sbspro.2012.11.097.

[21] H.-A. Park, An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain, J. Korean Acad. Nurs. 43 (2013), 154-164. https://doi.org/10.4040/jkan.2013.43.2.154.

[22] S. Kuhle, B. Maguire, H. Zhang, et al. Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study, BMC Pregnancy Childbirth. 18 (2018), 333. https://doi.org/10.1186/s12884-018-1971-2.

[23] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, J. Biomed. Inform. 35 (2002), 352–359. https://doi.org/10.1016/s1532-0464(03)00034-0.

[24] A. Rachmad, N. Chamidah, R. Rulaningtyas, Classification of mycobacterium tuberculosis based on color feature extraction using adaptive boosting method, AIP Conf. Proc. 2329 (2021), 050005. https://doi.org/10.1063/5.0042283