Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2023, 2023:81
https://doi.org/10.28919/cmbn/7873
ISSN: 2052-2541

TUBERCULOSIS CLASSIFICATION USING RANDOM FOREST WITH K-PROTOTYPE AS A METHOD TO OVERCOME MISSING VALUE

EKA MALA SARI ROCHMAN[1,2], MISWANTO[1,*], HERRY SUPRAJITNO[1], ISTIFADATUL KAMILAH[2], AERI RACHMAD[2], IWAN SANTOSA[3]

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

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

[3]Information Management Department, School of Management, National Taiwan University of Science and Technology, Taiwan

Abstract: Tuberculosis is a disease that attacks the core of the respiratory organs, which affects many people. This disease is one of the contributors to high mortality cases, especially in Indonesia. Based on its anatomical location, tuberculosis is divided into two classes, namely pulmonary for tuberculosis detected in lung parenchymal tissue and extrapulmonary for tuberculosis detected in organs other than the lungs. Detecting the location of the infection in the lungs requires some analysis of laboratory results for the triggering parameters where the analysis process is still done manually, so it takes longer, and because the input process is still done manually, patient data which causes the possibility of human error to be very high. Therefore, the solution offered and the aim of this study is the ease of patient diagnosis in determining the classification of TB disease. The method used in this study is k-prototype imputation to repair missing values that have different data types, then for tuberculosis data classification methods

*Corresponding author
E-mail address: miswanto@fst.unair.ac.id
Received January 04, 2023

and medical record data processing using the Random Forest, Support Vector Machine, and Backpropagation methods. Of the three classification methods proposed in this study, all three have an excellent level of accuracy. However, the Random Forest method performs more than other methods, reaching 98.8%.

Keywords: tuberculosis; imputation; missing values; K-prototype; classification; random forest.

2020 AMS Subject Classification: 92C60.

1. INTRODUCTION

Every year the development of an increasingly sophisticated era is marked by increasing industrial technology in various fields, resulting in factories' high growth. However, it is undeniable that the rapid growth of factories in Indonesia has a negative impact, namely, air pollution produced after production makes air quality decrease so that it is not suitable for the health of the respiratory organs, namely the lungs. Bacteria or viruses will more easily infect unhealthy lungs; one example of a disease caused by a bacterial infection is tuberculosis. Tuberculosis (TB) is an infectious disease caused by Mycobacterium tuberculosis in the lungs and is one of the diseases with the highest death rate [1], [2], and [3] [4]. Based on Indonesian TB data in 2022, cases of death from tuberculosis are 93,000 people per year or the equivalent of 11 deaths in one hour [5]. Even though the number of deaths caused by TB has decreased from the previous year, you have to watch out for it precisely because this disease is contagious.

To data from the World Health Organization (2019), Indonesia is listed as one of the countries with the third highest level of TB cases, with a total of 842,000 or 46% of all cases, especially in East Java province, which ranks second in Indonesia in terms of highest TB cases with a total of 57,014 [6]. It can be seen that TB cases have increased from year to year in various regions due to a lack of socialization about the dangers of TB and how to prevent and treat it. This can be seen from how people underestimate health and do not have a complete treatment for TB disease, which increases this case in Indonesia. Therefore, it is necessary to utilize data mining techniques to build a classification system that can facilitate diagnosing TB disease so that treatment can be carried out immediately.

In data mining, structured and complete data is needed for accurate results. However, not all raw data can be processed immediately because there is noise or an empty value called the missing value. In general, datasets in the medical field have incomplete data [7]. A missing value can reduce the accuracy of data [8] [9]. Therefore, it is necessary to take an approach to overcome the problem of missing values. Overcoming the missing value can be done by ignoring the missing value during the analysis, or you can also do the imputation. Imputation is a technique of replacing missing values with values obtained from a method [10].

One of the technologies that are currently popular is machine learning. In machine learning, classification methods play a role in various fields, one of which is the health sector. In this field, machine learning can predict disease and present medical diagnosis data. Many machine learning methods are used to classify and analyze diseases, one of which is the Random forest. In previous research on Coronary Artery Disease classification, which discussed a comparison of several machine learning methods, the best accuracy results were obtained when applying the Support Vector Machines, K-Nearest Neighbors, Neural Network, and Random Forest algorithms [11]. Therefore, this research will classify using three different methods, namely, Naïve Bayes, SVM, LSTM, and Backpropagation, to find out which method best performs categorizing data involving the imputation process using the K-Prototype method.

2. PRELIMINARIES

System errors, such as no response to sensors or input receiving devices, as well as human errors when entering data, are common things that often occur so that there is incompleteness in data which causes missing data. In data mining, some methods can only be processed when the data has complete features or data, and therefore special handling is needed for missing data. The following are the methods used to deal with missing data problems: Case Deletion, Parameter Estimation, and Imputation Techniques.

Case deletion is the simplest method, namely by deleting data that contains missing data so that it is not used in further processing. However, this method has a weakness because some important

information is deleted. The imputation technique is a technique that is widely used because, in this technique, no data is deleted for nothing. The way this technique works is to estimate missing data by obtaining patterns from data that have complete features. Mean, Median, and clustering are some of the most frequently used imputation methods [12].

Data mining can be known as the process of mining data information or disclosing information from printed data sets [13]. Data mining is characterized as a process of extracting data where clients communicate with various reports involving analytical tools as part of information mining. The purpose of data mining is to obtain valuable data from a set of reports so that the information sources used in data mining are different data whose arrangement is unstructured or perhaps semi-coordinated. Based on the tasks that can be performed, data mining is grouped into several sections, namely description, prediction, estimation, classification, clustering, and association [14]. Below are the stages in data mining shown in Figure 1.

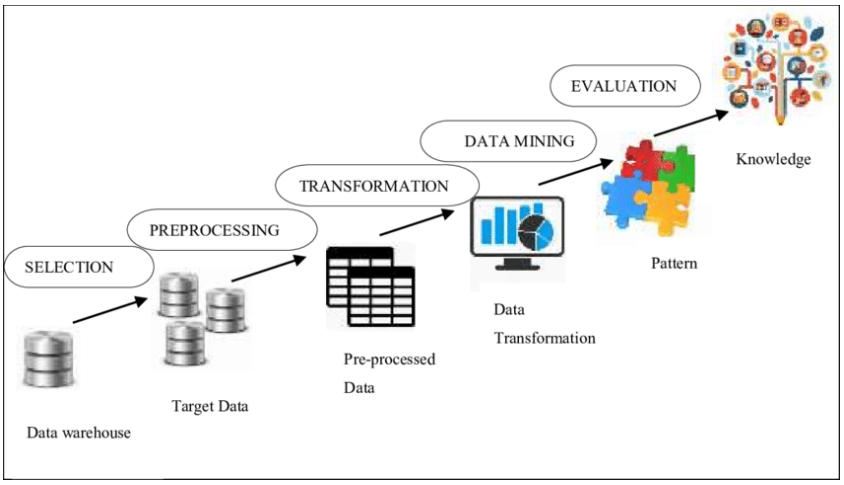

Figure 1 Data Mining Stages

In Figure 1 above, there are several stages of the process of data mining, namely:

a. Data selection
b. Preprocessing/Cleaning
c. transformation
d. Data Mining

e. Evaluation

2.1. Data Preprocessing

The data preprocessing process in this study begins with data transformation to change the categorical data type to numeric so that computation can be carried out. Then, next is the data imputation process to overcome the missing value condition. The final data preprocessing stage is normalization for scaling the range of values between data.

A. Data Transformation

The data transformation process is the stage for changing the data type into another data type [15]. Data transformation aims to simplify the classification process. In this study, the transformation is carried out by changing the values in the dataset in the form of categorical data converted into numeric data.

B. Missing Value Imputation

1. A missing value is a condition where there are several attribute values in data that have no value or are missing. There are several causes for missing values, one of which is due to an error during the data entry process [16]. Therefore, it is necessary to apply an algorithmic method to overcome missing value conditions, one of which is the imputation process.
2. Imputation is a process that can handle the phenomenon of missing values in datasets by filling in missing data using new values based on data that has complete attributes or other information available from the dataset [17]. This imputation process can handle missing values better than other methods. In carrying out the imputation process, several algorithm methods can be applied, one of which is the K-Prototype.
3. K-Prototype is an algorithm that is commonly used in grouping data with mixed data types, namely numeric and categorical [18]. This method works by combining distance calculations in the k-means algorithm, namely the euclidean distance and dissimilarity measures found in k-modes [19]. The steps for imputing using the K-Prototype algorithm are as follows:
   a. Determine the number of group parts to be used.

b. Determine the center value of the group (centroid).

c. Calculate the distance of each data with the selected center value; calculate the distance using the following formula:

$$D(x,y) = \sum_{j=1}^{d_1} (xj - yj)^2 + \lambda \sum_{j=d1+1}^{d} \delta(xj, yj) \quad (1)$$

$$\delta(x_j, y_j) = \begin{cases} 1, if\ x_j \neq y_j \\ 0, if\ x_j = y_j \end{cases} \quad (2)$$

$$\lambda = 0{,}5\ \times std(Xnumerical) \quad (3)$$

Description:

d_1 : distance size limit for numeric attributes

d : distance size limit for category attributes

xj : data weight at the center of the cluster (centroid)

yj : data weight in the cluster you want to find the distance

4. Grouping data based on the smallest distance.

5. Determine the new centroid value using the average value of numeric data and using the mode of categorical data. Here is the formula:

$$Ck = \frac{1}{nk} \sum_{i=1}^{n} di \quad (5)$$

Description:

nk : total data in the cluster,

di : total in each cluster, Repeat steps three to five until no member of the cluster has moved.

Normalization

Data normalization is a scaling technique that aims to change values between data that have the same range of values [20]. By using the normalization technique, the dataset will have a new field or range of values so that no data is too large or too small to simplify the statistical analysis process. Min-Max Normalization is one method that is commonly applied in carrying out data normalization processes [21]. The equation below is the formula for the min-max normalization

method.

$$x' = \frac{x - \min(x)}{[\max(x) - \min(x)]} \quad (6)$$

Description:

x’ : normalization result,

x : actual value of variable x

min(x) : minimum value

max(x) : maximum value

2.2. Data Mining Process

The data mining process involves the process of dividing training data and test data before entering the classification process. Training data is needed in the learning process of the classification model, while test data is used to evaluate the results of the model obtained when the learning process is carried out. The data-sharing process is carried out using several methods, one of which is by using the k-Fold Cross Validation method. This method uses the value of k to determine the number of partitions. Below is an illustration of the use of the value of k = 5 [21].

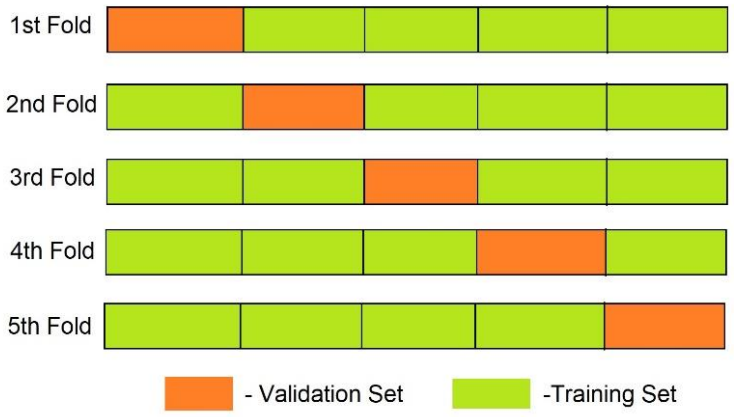

Figure 2. Illustration of 5-Fold Validation

Figure 2 above shows the distribution of data divided by as many as k values, namely 5. The five partitions in each fold have almost the same amount of data. As shown in the figure, the first fold iteration has four training data partitions and 1 test data partition where the 1st partition is the test

data and the 2nd to 4th partitions are the training data. In research on the data mining process, several algorithms are used to carry out classification modeling of Tuberculosis datasets. These algorithms include Support Vector Machine, Backpropagation, and Random Forest.

A. Random Forest

Random Forest is a method that makes modeling using a collection of several decision trees. The method is included in the supervised learning algorithm, which can classify data based on samples and attributes of the training data. Random forest is also one of the algorithms that use ensemble techniques by applying bagging and random feature selection methods. The ensemble learning applied to this method is useful for reducing the problem of less stable classifications by combining some basic learning to reduce prediction errors [22]. The stages of the Random Forest algorithm are as follows.

1. The first stage is inputting the dataset and then bootstrapping the data to create a subset by taking random samples with a return size of n from the training data.
2. The second stage is the process of random feature selection to build a tree until it reaches the maximum size.
3. The third stage is to calculate the value of all features to build a tree using the entropy formula below.

$$entropy(S) = \sum_{i=1}^{c} p_i log_2 p_i \tag{7}$$

Description:

S : the set of datasets

C : the number of classes

Pi : probability class i-th frequency in the dataset

$$entropy(T, X) = \sum_{i=k}^{k} P(c)E(c) \tag{8}$$

Description:

(T, X) : features of T and X

P(c) : the probability of feature class

E(c) : entropy result of a feature class

4. The fourth stage is calculating the information gain obtained from the entropy value of each attribute, where the attribute with the highest information gain value will be used as the root node in the tree. Below is the formula for calculating information gain.

$$gain(A) = entropy(S) = \sum_{i=1}^{k} \frac{|Si|}{|S|} x\ entropy\ (Si) \quad (9)$$

Description:

S : the set of datasets

A : features

|Si| : the number of sample values to-i

|S| : sum of all data

Entropy(Si) : entropy of the i-th value sample

5. The fifth stage builds "k" trees, if the tree has not reached the target, then repeat the first step until the data is split. Meanwhile, if the tree has reached the target (k), then the next step is a majority vote to get the final prediction result.

B. Support Vector Machine

Support Vector Machine was invented by Vapnik in 1992 as an AI strategy that works with standard SRM or Structural Risk Minimization. Support Vector Machine hopes to determine the best hyperplane that will isolate two classes in the information space. A hyperplane is a separator between the two classes that aims to maximize the distance (margin) between data classes [23] and [24]. To find the optimal separator function (classifier) and can separate two different classes, the best hyperplane must be found among the unlimited number of other hyperplanes. A good hyperplane if it is located right between two sets of objects from two classes. Figure 3 shows how SVM maximizes the distance between two different sets of classes (margins) by determining the best hyperplane.

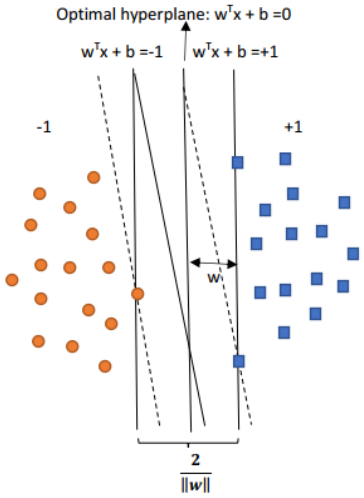

Figure 3 SVM Visualization

Figure 3 shows how the hyperplane acts as a separator between two different classes in the classification process to achieve good accuracy by measuring the hyperplane margin and determining the maximum point. The distance between the hyperplane and the closest pattern of each class is called the margin. In the image above, the dots on the dotted line represent the pattern closest to the hyperplane, which is called the support vector.

In the SVM method, several kernel functions can be used to separate data based on its class which has high dimensions or cannot be solved using a linear dividing line. The following are kernel functions that are commonly used to classify non-linear data [25]:

1. Linear Kernels

   $K(x_i, x) = x_i^T x$

2. Polynomial Kernels

   $\mathrm{K}(x_i, x) = (\gamma x_i^T + r)^p, \gamma > 0$

3. Kernel RBF or Radial Basis Function

   $\mathrm{K}(x_i, x) = \exp(-\gamma ||x - x_i||^2)$

4. Sigmoid Kernels

   $\mathrm{K}(x_i, x) = \tanh(\gamma x_i^T + \mathrm{r})$

The steps in classifying using the SVM algorithm are as follows:

1. Enter data to be classified.
2. Calculate the value of the dot product by the kernel function used.
3. Compute the hessian matrix. The Hessian matrix is the product of the kernel function and the value. The value in question is a vector value that has a value of 1 and -1. The Hessian matrix can be calculated using the following formula.

$$D_{ij} = y_{i,}y_j\left(K\left(x_{i,}x_{j,}\right) + \lambda^2\right) \quad (10)$$

Description:

x_i : i-th data

x_j : jth data

y_i : i-th data class

y_j : jth data class

4. Calculate the error value, delta alpha, and new alpha using the following formula.
   a. Calculates the error value

$$E_i = \sum_{i=1}^{l} a_i D_{ij} \quad (11)$$

   b. Calculating delta alpha

$$\delta a_i = min\{\max[\gamma(1 - E_I), -a_i), \mathrm{C} - a_i\} \quad (12)$$

   c. Got a new alpha

$$a_i = a_i \delta a_i \quad (13)$$

Description:

E_i : average error

γ : gamma

C : alpha limit

a_i : alpha i-th

δa_i : delta alpha ith

5. Calculate the bias value using the formula below.

$$b = -\frac{1}{2}(w.x^{+} + w.x^{-}) \quad (14)$$

Description:

b : bias value

$w.x^{+}$: support vector weight in the positive class

$w.x^{-}$: support vector weight in the negative class

6. Calculate the dot product value between the training data and test data.
7. Determine the test data class using the equation below.

$$f(x) = \sum_{i=1}^{l} a_i y_i K(x_i, x) + b \quad (15)$$

Description:

x_i : i-th data

x_j : j-th data

y_i : i-th data class

b : biased

D. Backpropagation

Backpropagation is a supervised learning algorithm that uses several layers to update or change the weight values connected to several neurons in the hidden layer [26]. This algorithm works by minimizing the error value in the network's output results by updating the weight value in the backward phase using the output error value. The output error value is obtained after performing calculations in the forward phase. In this algorithm, the learning process is carried out in two phases, namely the forward propagation and backward propagation stages. The following is the flow of the backpropagation algorithm in each phase.

Phase 1: Forward

1. Initialize the weights using small random values, max Epoch, error, and learning rate.
2. If the epoch value is smaller than the maximum value, do steps 3 to 4. This condition will repeat until it meets the requirements.

3. Any data contained in the input layer will receive a signal and be forwarded to the hidden layer.
4. Calculate all outputs in the hidden layer z_k (j = 1,2, … ,p), using the following formula:

$$z\_net_j = v_{jo} + \sum_{i=1}^{n} x_i + v_i \quad (16)$$

$$z_j = f(z\_net_j) = \frac{1}{1+e^{-z\_net}} \quad (17)$$

Description:

v_{0j} : bias weight in hidden units z_j

z_j : jth hidden unit

i : 1, …, n

n : number of input units

5. Calculate all results on the output layery_k (k = 1,2, … , m), sing the following formula:

$$y\_net_k = w_{ko} + \sum_{j=1}^{p} z_i + w_{kj} \quad (18)$$

$$y_k = f(y\_net_j) = \frac{1}{1+e^{-y\_net}} \quad (19)$$

Description:

w_{ko} : bias weight at the output unit yk

z_i : hidden unit

y_k : k-th output unit

i : 1, …, n

p : number of hidden units

Phase 2: Backward

6. Calculate the unit output factor based on the error value for each output unit y_k (k = 1,2,…,m)

$$\delta_k = (t_k - y_k)\, f'Ynet_j = (t_k - y_k) y_k (1 - y_k) \quad (20)$$

Calculate the change in weight using the formula below.

$$\Delta W_{jk} = \alpha\, \delta_k\, z_y\, ; k = 1,2, \dots, m\, ; j = 0,1, \dots . p \quad (21)$$

Description:

w_{jk} : line weight from z_j to output unit yk

y_k : k-th output unit

t_k : target output

j : 1, … , p

p : number of hidden units

δ_k : predictive output

7. Calculate the hidden unit factor based on the error value at each output unit y_k (k = 1,2,…, m)

$$\delta_{\_netj} = \sum_{k=1}^{m} \delta_k \; W_{kj} \tag{22}$$

$$\delta_j = \delta\_net_j \; f'Znet_j = \; \delta_{\_netj} \; z_j(1 - \; z_j) \tag{23}$$

Calculate the change in weight using the formula below.

$$\Delta V_{ji} = \; \alpha \, \delta_j \, x_j \; ; j = 1,2, \dots , m \; ; i = 0,1, \dots . p \tag{24}$$

8. The sum of all changes in weight on the line leading to the unit output.

$$W_{jk}(baru) = \; W_{jk}(lama) \; + \; \Delta W_{jk} \; ; k = 1,2, \dots , m \; ; j = 0,1, \dots . p \tag{25}$$

9. The sum of all weight changes on the line leading to the hidden unit.

$$\Delta V_{ji}(baru) = V_{ji}(lama) \; + \; \Delta V_{ji} \; ; j = 1,2, \dots , m \; ; i = 0,1, \dots . p \tag{26}$$

Description:

W_{jk} : changes in weight values in the output layer

ΔV_{ji} : change in weight values in the input layer

2.3. Evaluation

The system performance evaluation process is needed to measure how well the method used is. The technique commonly used in evaluating the system is the confusion matrix. The confusion matrix can produce calculations of the level of accuracy, precision results, recall, and f1 score [27]. The following shows a representation of the confusion matrix in Table 1.

Table 1 Confusion Matrix

	Actual Positive	**Actual Negative**
Predicted Positive	*True Positive* (TP)	*False Positive* (FP)
Predicted Negative	*False Negative* (FN)	*True Negative* (TN)

Description:

True Positive (TP): It is said to be true positive if the classification result is predicted to be in a

positive class and the actual class is also positively labeled.

True Negative (TN): It is said to be true negative if the classification result is predicted to be in a negative class and the actual class is also negatively labeled.

False Positive (FP): It is said to be false if the classification result is predicted to be in a positive class while the actual class is labeled negative.

False Negative (FN): It is said to be false if the classification result is predicted to be in a negative class while the actual class is labeled positive.

The confusion matrix results will calculate the AUC value, accuracy, recall, precision, and f1 score using the following formula. The confusion matrix results will calculate the AUC value, accuracy, recall, precision, and f1 score using the following formula.

1. AUC (area under the curve), or it can be called probability, is a method for calculating under the ROC curve. The higher the AUC value, the classification method used can be applied properly in a study. The following is the calculation formula [27]:

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FP} + \frac{TN}{TN + FP}\right) \tag{27}$$

The following is a table of classification accuracy with AUC values:

Table 2 AUC Accuracy Range

AUC Value	Target
0.90-1.00	Very Good
0.80-0.90	Good
0.07-0.80	Good Enough
0.60-0.70	Less Good
0.50-0.60	Bad

2. Accuracy is the percentage of the predicted classification result data that is correct compared to all the input data. The accuracy formula can be written as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (28)$$

3. A recall is the number of reviews that are correctly classified as positive divided by the number of positive reviews in the dataset. Recall can be written and calculated using the following formula:

$$Recall \quad = \frac{TN}{TN + FN} x\ 100\% \quad (29)$$

4. Precision has the same meaning as recall, it's just that precision is used to calculate the negative class. Here is the formula:

$$Presisi = \frac{TP}{TP + FP} x\ 100\% \quad (30)$$

5. F-measure is the result of mean recall and precision, where the range of the f1 score itself is 0-1.

$$f1\ score \quad = 2\ x\ \frac{Recall\ x\ Precision}{Recall + Precision} x\ 100\% \quad (31)$$

3. MAIN RESULTS

A. Data collection

In the study, the dataset used was information on patients diagnosed with tuberculosis, as many as 985 records with six attributes which included age, sex, chest X-ray, HIV status, history of diabetes, and TCM results. In the raw data, there are still several empty values or so-called missing values, which will later be resolved by applying the imputation technique using the K-Prototype method. Missing values in the dataset can be seen in Table 3.

Table 2 Missing Value on Each Attribute

Attribute	Number of Blank Data
Age	0
Sex	0
Thorax X-Ray	28
HIV	261
Diabetes	7
TCM	439

B. Analysis

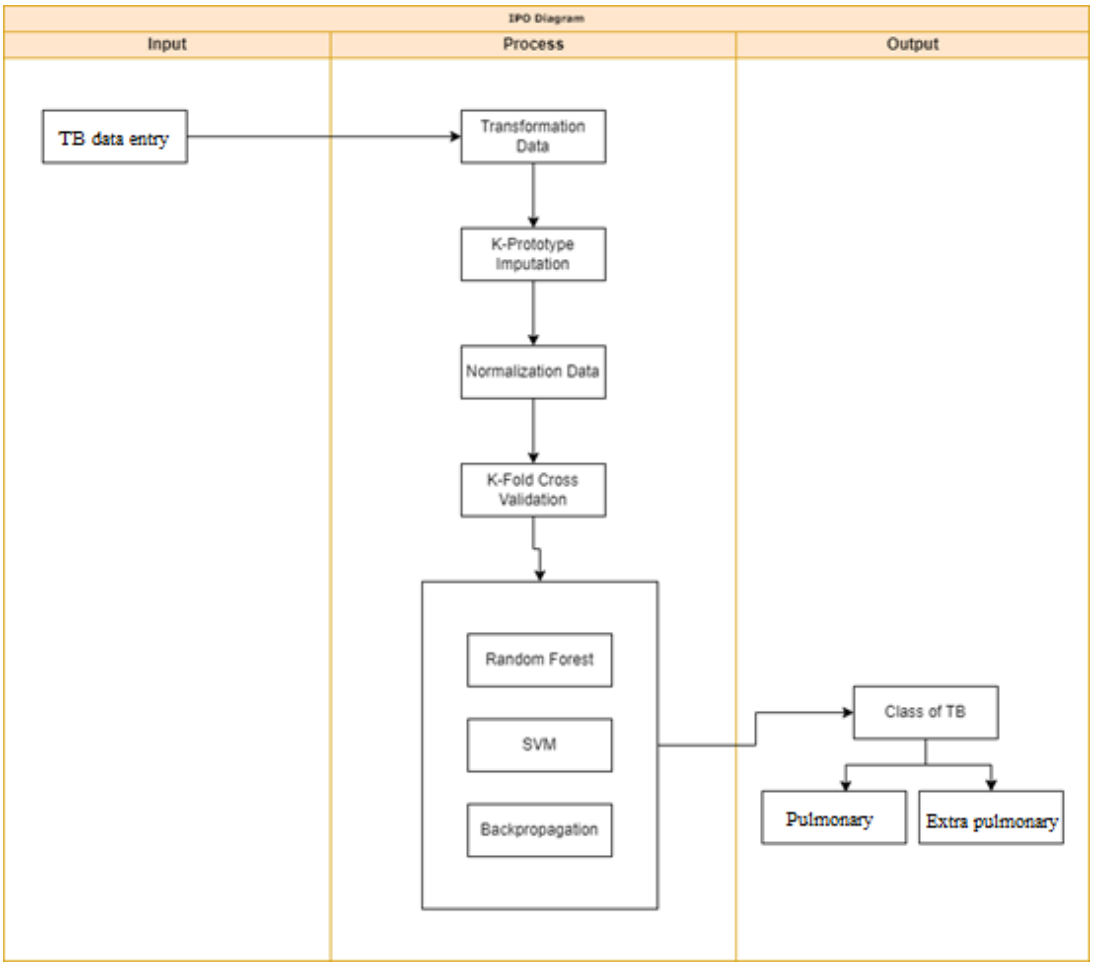

Figure 3 Process Flow Chart

Figure 3. describes the IPO diagram as follows:

1. Input

   The input section is the initial part. Namely, there is a data input process. The data is the result of records from the medical records of patients with tuberculosis, which consists of 7 attributes.

2. Preprocessing

   This stage aims to change the raw data to be more structured to facilitate and increase the accuracy in classifying data. In this study, there were several stages of data preprocessing, including the following:

   a. Transformation

      Data transformation is the initial stage in the data preprocessing carried out in this study. This stage is necessary to convert categorical data into numeric data so that it can be computed.

   b. Imputation

      After transforming the data, the next step is to perform the imputation technique to handle missing values using the K-Prototype algorithm.

   c. Normalization

      The new dataset obtained from the imputation process will be normalized before entering the classification process, this is necessary so that the range or range of values in the data is not too far away and a 0 to 1 value range scaling is required.

3. Data Sharing Process

   The process of dividing the data in this study applies the K-fold Cross Validation method with k = 5 and k = 10.

4. Classification

   In the classification process, a learning process is carried out to obtain a classification model using several different methods, namely Support Vector Machine, Backpropagation, and

Random Forest. The other classification models will be compared with the accuracy results to obtain optimal classification results.

5. Outputs

   The resulting output is a class prediction of TB attributes based on the modeling method proposed in this study.

4. RESULT AND DISCUSSION

A. Imputation

Data preprocessing is crucial before data is processed or processed in Machine Learning (ML). Data preprocessing is used so the system can properly process the data. This preprocessing stage includes several things, namely, filling in the missing values because the presence of missing values can affect the results of the dataset classification itself. The preprocessing stages carried out in this study were imputation or filling in the missing values using the K-Prototype Imputation method. Before that, do a way to change categorical data and normalize numeric data in the columns 'Age,' 'Gender', and 'TCM Results'. There are several ways to do categorical encoding using label encoding and one hot encoding. In its implementation, the use of encoding labels for categorical data that has more than two types of values can be misinterpreted by the algorithm as having some hierarchy or sequence. One-hot encoding is a technique that changes each value in a column as a new column and fills it with a binary value, namely 0 or 1. An example of data results after a one-hot encoding process:

Table 4. Data after one-hot encoding

Age	Gender_P	Gender_L
12	0	1
54	0	1
55	1	0
57	0	1
46	1	0

Determining the number of clusters in this study is K = 3, while the distance between data to each centroid using the K-Prototype formula because there are 2 types of data, namely categorical and numerical data in calculating distances. Grouping data according to the minimum distance from the centroid. Perform iterations to determine the new centroid until the centroid and the number of members does not change.

B. Random Forest

The classification process using the Random Forest method involves dividing the dataset using the K-Fold Cross Validation method using the values k=5 and k=10. The parameter of the number of trees used is 10. Following are the results of the evaluation using the confusion matrix method that has been carried out on the tuberculosis dataset of 985 records.

Table 4 Random Forest Evaluation Results

Fold	AUC	Classification Accuracy	F1 Score	Precision	Recall
5	98.6 %	97.7 %	97.6 %	97.7 %	97.7 %
10	98.8 %	97.4 %	97.3 %	97.4 %	97.4 %

Table 4 shows the test results using two different k-fold values where the accuracy of the two results are not much different where the use of k-fold = 5 is slightly better than k-fold = 10, this is indicated by the Classification Accuracy value when using k- fold = 5 reaches 98.6 % and 97.4 % on k-fold = 10. However, the AUC value using k-fold = 10 is better, 98.8 %.

C. Support Vector Machine

The results of testing SVM modeling on the RBF kernel using the k-fold method to divide training data and test data with k-fold values = 5 and 10. As for the C parameter value used, 1, and an error tolerance value of 0.0010 with a maximum of 100 iterations, The following results of the evaluation of SVM modeling on the RBF kernel are shown in the table below.

Table 5. Support Vector Machine Evaluation Results

Fold	AUC	Classification Accuracy	F1 Score	Precision	Recall
5	95.3 %	97.4 %	97.3 %	97.5 %	97.4 %
10	94.0 %	97.4 %	97.3 %	97.5 %	97.4 %

Table 5 shows the test results using two different k-fold values where the two accuracy results have the same value. This is indicated by the Classification Accuracy value when k-fold = 5 and 10, which reaches 97.4%. However, the AUC value using k-fold = 5 is superior to k-fold = 10, which is 95.3%.

D. Backpropagation

Testing on Backpropagation modeling begins by dividing the dataset into test data and training data using the k-fold method with values of k = 5 and 10. As for the number of neurons, as many as 100 layers and the activation, the function uses ReLu with SGD and Adam optimizer. Then, for a learning rate value of 0.0001 with some iterations (epochs) of 200. The following are the evaluation results of modeling using the Backpropagation algorithm shown in Table 6.

Table 6 Backpropagation Evaluation Results

Optimizer	AUC	Classification Accuracy	F1 Score	Precision	Recall
SGD	96.9 %	97.4 %	97.3 %	97.5 %	97.4 %
Adam	97.3 %	97.3 %	97.2 %	97.3 %	97.3 %

As can be seen in Table 4, the use of adam and sigmoid optimizer gives an almost as good performance. Overall the evaluation results, the sigmoid optimizer gives slightly superior results compared to the adam optimizer as shown in the table above where the CA, F1 Score, Precision, and Recall values have higher values. However, the AUC value of Adam Optimizer has a higher value, reaching 97.3%.

Based on the evaluation results on the AUC value of each of the algorithms used, the ROC graph will be formed as shown in Figures 4 and 5.

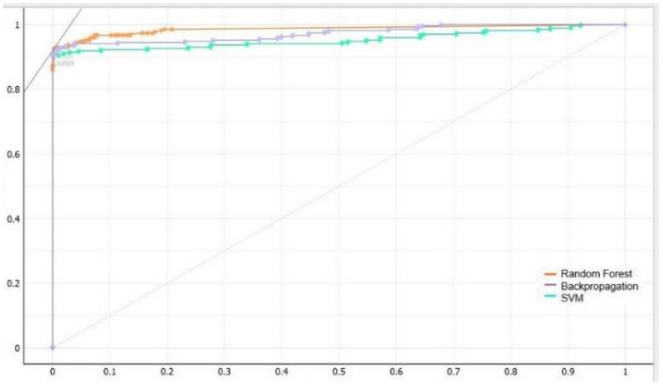

Figure 4 Graph ROC 5 Fold

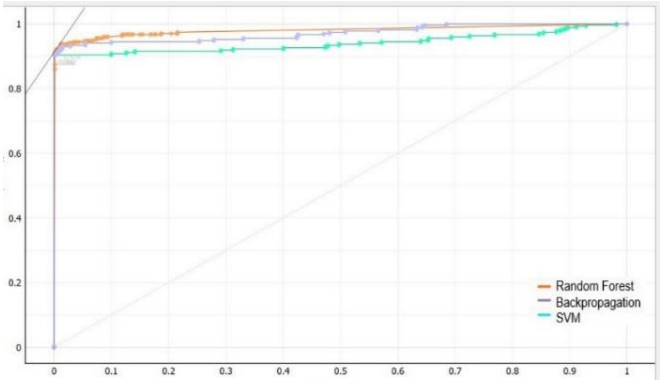

Figure 5 Graph ROC 10 Fold

Figure 4 and Figure 5 show the ROC graphs comparing the AUC value to the Random Forest, SVM, and Backpropagation methods. Based on the ROC graph both on the use of k-folds 5 and 10, it is found that modeling using the Random Forest method has the best level of accuracy, which is indicated by the point line that represents the AUC Random Forest value that is closest to the number 1 because the AUC value is closer to number 1 then this method has excellent accuracy based on Table 4. So, in this case, the Random Forest method has the best accuracy in classifying the TB dataset.

5. CONCLUSION

Based on the analysis and discussion that has been carried out on the results of the classification evaluation by dividing training and testing data using 5-fold and 10-fold on 985 records with seven

attributes, namely age, sex, district, chest X-ray, HIV status, history of diabetes, outcome TCM (Rapid Molecular Test), it is concluded that:

1. Using K-Prototype imputation with K=3 can overcome gaps in data with conditions of different data types.
2. Using k-fold values 5 and 10 does not provide a significant difference as shown in Tables 2 and 3. However, from the evaluation results, the use of k-fold five values is slightly superior to k-fold 10 when implemented in research.
3. The application of different classification methods gives different evaluation results. Where based on the discussion of the performance evaluation results for each of these methods, it is known that in this case, the Random Forest method with a k-fold value = 5 has a better performance value compared to the Support method Vector Machines and Backpropagation.

ACKNOWLEDGMENT

The author would like to thank Universitas Airlangga Indonesia, which has facilitated doctoral education. As well as the University of Trunojoyo Madura, Indonesia, where the author resides as a teacher.

CONFLICT OF INTERESTS

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

REFERENCES

[1] S.B. Rakhmetulayeva, K.S. Duisebekova, A.M. Mamyrbekov, et al. Application of classification algorithm based on SVM for determining the effectiveness of treatment of tuberculosis, Procedia Computer Sci. 130 (2018), 231–238. https://doi.org/10.1016/j.procs.2018.04.034.

[2] A. Soni, A. Rai, S.K. Ahirwar, Mycobacterium tuberculosis detection using support vector machine classification approach, in: 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), IEEE, Bhopal, India, 2021: pp. 408–413. https://doi.org/10.1109/CSNT51715.2021.9509635.

[3] K. Kanesamoorthy, M.B. Dissanayake, Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm, Int. J. Mycobacteriol. 10 (2021), 279-284.

[4] E.M.S. Rochman, Miswanto, H. Suprajitno, Comparison of clustering in tuberculosis using fuzzy c-means and k-means methods, Commun. Math. Biol. Neurosci. 2022 (2022), 41. https://doi.org/10.28919/cmbn/7335.

[5] Admin PLK UNAIR, Waspadai TB dikala Pandemi, Pusat Layanan Kesehatan UNAIR, 24 Maret 2021. [Online]. Available: http://plk.unair.ac.id/waspadai-tbc-di-kala-pandemi/. [Accessed 10 August 2022].

[6] V.V. Nurdiansyah, I. Cholissodin, P.P. Adikara, Klasifikasi Penyakit Tuberkulosis (TB) menggunakan Metode Extreme Learning Machine (ELM), Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 4 (2020), 1387-1393.

[7] A.R. Ismail, N.Z. Abidin, M.K. Maen, Systematic review on missing data imputation techniques with machine learning algorithms for healthcare, J. Robot. Control. 3 (2022), 143–152. https://doi.org/10.18196/jrc.v3i2.13133.

[8] R.H.H. Groenwold, O.M. Dekkers, Missing data: the impact of what is not there, Eur. J. Endocrinol. 183 (2020), E7–E9. https://doi.org/10.1530/eje-20-0732.

[9] E.M.S. Rochman, Miswanto, H. Suprajitno, Overcoming missing values using imputation methods in the classification of tuberculosis, Commun. Math. Biol. Neurosci. 2022 (2022), 66. https://doi.org/10.28919/cmbn/7538.

[10] S.K. Kwak, J.H. Kim, Statistical data preparation: management of missing values and outliers, Korean J. Anesthesiol. 70 (2017), 407-411. https://doi.org/10.4097/kjae.2017.70.4.407.

[11] S. Mathew, J.T. Abraham, S.J. Kalayathankal, Data mining techniques and methodologies, Int. J. Civil Eng. Technol. 9 (2018), 246-252.

[12] M. Yuvalı, B. Yaman, Ö. Tosun, Classification comparison of machine learning algorithms using two independent CAD datasets, Mathematics. 10 (2022), 311. https://doi.org/10.3390/math10030311.

[13] S. Mathew, J.T. Abraham, S.J. Kalayathankal, Data mining techniques and methodologies, Int. J. Civil Eng. Technol. 9 (2018), 246-252.

[14] P.A. Widya, M. Sudarma, Implementation of EM algorithm in data mining for clustering female cooperative, Int. J. Eng. Emerging Technol. 3 (2018), 75-79.

[15] E. Zdravevski, P. Lameski, A. Kulakov, Advanced transformations for nominal and categorical data into numeric data in supervised learning problems, in: Proceedings of the 10th International Conference for Informatics and Information Technology (CIIT 2013), Bitola, Macedonia, 2013.

[16] H. Kang, The prevention and handling of the missing data, Korean J Anesthesiol. 64 (2013) 402-406. https://doi.org/10.4097/kjae.2013.64.5.402.

[17] C.H. Liu, C.F. Tsai, K.L. Sue, et al. The feature selection effect on missing value imputation of medical datasets, Appl. Sci. 10 (2020), 2344. https://doi.org/10.3390/app10072344.

[18] Z. Jia, L. Song, Weighted k-prototypes clustering algorithm based on the hybrid dissimilarity coefficient, Math. Probl. Eng. 2020 (2020), 5143797. https://doi.org/10.1155/2020/5143797.

[19] B. Kim, A fast K-prototypes algorithm using partial distance computation, Symmetry. 9 (2017), 58. https://doi.org/10.3390/sym9040058.

[20] H. Henderi, Comparison of min-max normalization and Z-score normalization in the K-nearest neighbor (kNN) algorithm to test the accuracy of types of breast cancer, Int. J. Inform. Inform. Syst. 4 (2021), 13–20. https://doi.org/10.47738/ijiis.v4i1.73.

[21] M. Gheorghe, R. Petre, The importance of normalization method for mining medical data, J. Int. J. Computers Technol. 14 (2015), 6014-6020.

[22] D. Normawati, D.P. Ismi, K-fold cross validation for selection of cardiovascular disease diagnosis features by applying rule-based datamining, Signal Image Process. Lett. 1 (2019), 23–35. https://doi.org/10.31763/simple.v1i2.3.

[23] G.S. Saragih, Z. Rustam, D. Aldila, et al. Ischemic stroke classification using random forests based on feature extraction of convolutional neural networks, Int. J. Adv. Sci. Eng. Inform. Technol. 10 (2020), 2177-2182.

[24] H. Sain, S.W. Purnami, Combine sampling support vector machine for imbalanced data classification, Procedia Computer Sci. 72 (2015), 59–66. https://doi.org/10.1016/j.procs.2015.12.105.

[25] P.A. Octaviani, Y. Wilandari, D. Ispriyanti, Penerapan Metode Klasifikasi Support Vector Machine Pada Data Akreditasi Sekolah Dasar di Kabupaten Magelang, Jurnal Gaussian, 3 (2014), 811-820.

[26] Z. Liu, H. Xu, Kernel parameter selection for support vector machine classification, J. Algorithms Comput. Technol. 8 (2013), 163-177.

[27] Y.A. Lesnussa, C.G. Mustamu, F.K. Lembang, et al. Application of backpropagation neural networks in predicting rainfall data in Ambon city, Int. J. Artif. Intell. Res. 2 (2018), 1-9. https://doi.org/10.29099/ijair.v2i2.59.

[28] M.U. Devi, Categorizing the age group and measuring accuracy of fuzzy model, Int. J. Electron. Commun. Eng. Technol. 10 (2019), 36-46.

[29] E.M.S. Rochman, A. Rachmad, D.A. Fatah, et al. Classification of salt quality based on salt-forming composition using random forest, J. Phys. Conf. Ser. 2406 (2022), 012021. https://doi.org/10.1088/1742-6596/2406/1/012021.

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