Jurnal Ekonomi Teknologi & Bisnis (JETBIS)
Volume 2, Number 4 April 2023
p-ISSN 2964-903X; e-ISSN 2962-9330

DATA PRE-PROCESSING AND FEATURE SELECTION TECHNIQUES BACKWARD ELIMINATION FOR NAÏVE BAYES CLASSIFICATION ON HEART DISEASE DETECTION

Julius Warih Angkasa[1], Edi Noersasongko[2], Purwanto[3]
Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia[1,2,3]
warih70@gmail.com

KEYWORDS:
Heart Failure; Naïve Bayes; Backward Elimination; Classification; Prediction

ABSTRACT

According to research published in the International Journal of Cardiology with the title "Heart failure across Asia: Same healthcare burden but differences in organization of care", the mortality rate due to heart failure in Indonesia is relatively high. The results indicate that about 5% of the total population in Indonesia suffers from heart failure. Heart disease is a condition that occurs when the heart is impaired, either due to infection or congenital abnormalities. It is important to pay attention to heart disease to reduce mortality rates. However, there are several factors that are less accurate in identifying heart disease, and it is necessary to calculate using one of the prediction approaches using data mining techniques. One of the data mining methods used is the Naïve Bayes algorithm, which acts as a classification technique. In addition, before classifying, problems are often found in the content of the data, namely there are missing values. This problem can interfere with the classifier, therefore special techniques are needed, namely pre-processing techniques to remove missing values, so that it supports obtaining good prediction results. Also to support the classification, this research applies feature selection using the Backward Elimination method to improve accuracy. In this study, through the application of data pre-processing techniques and feature selection, it succeeded in increasing the accuracy rate to 98.31%. So the purpose of this study is to determine the prediction results of heart disease using the Naive Bayes algorithm based on Backward Elimination and to determine the effect of missing values on the performance of the Naive Bayes algorithm based on Backward Elimination on the heart disease classification model.

INTRODUCTION

The heart is a hollow muscular organ located in the center of the chest. The heart is required to pump blood throughout the body. Heart disease is the second highest cause of death in Indonesia. Heart disease occurs due to partial blockages that gradually accumulate, resulting in a disturbance in the balance of blood supply and demand (Maulana & Yahya, 2019).

In Indonesia, people with heart failure account for 5% of the total population (Reyes et al., 2016). And the mortality rate due to heart failure is also high, 17.2 percent of heart failure patients in Indonesia die during hospitalization, 11.3 percent die within 1 year of treatment, and 17 percent of patients experience the need for repeated hospitalizations as a result of worsening symptoms and signs of heart failure. World Health Organization (WHO) data in 2012 showed 31% of 56.5 million

deaths worldwide. More than ¾ of deaths from heart disease occur in developing countries (Handayani et al., 2022).

In addition, if the treatment of heart disease is late, this can threaten the lives of sufferers. This problem arises due to the difficulty of detecting heart disease at an early stage, because patients often ignore the initial symptoms that appear. In addition, the costs required for examination of heart disease are quite high, because it involves visits to specialists and laboratory tests (Handayani et al., 2022). Prediction systems can be one of the options used to detect heart disease early at a more affordable cost. This is due to the elimination of the cost of visits to specialists and laboratory tests that can be replaced by a prediction system. The purpose of this research is to develop a prediction system for heart disease using the Naive Bayes algorithm based on Backward Elimination. This research is based on historical data of patients who will undergo examination. Research Data (Arie & Suryandari, 2023).

The Naïve Bayes algorithm belongs to the category of the most straightforward and simple probability classifiers based on the general assumption that all features are independent of each other. based on the general assumption that all features are independent of each other. each other. And Naïve Bayes is a kind of probabilistic classification mechanism, which stems from the Bayesian theorem proposed by Thomas Bayes (Nguyen et al., 2014). The modeling is simple and uncomplicated, so it is considered suitable for use with large databases. Also, the Naïve Bayes method is very good at classifying text based on machine learning over other probabilities, and is widely used in big data analysis due to its simple and simple algorithm structure. because of its simple and fast algorithm structure (Nguyen et al., 2014). So this research is suitable to be applied with the Naïve Bayes method as a classifier method. Naïve method Bayes method has advantages and disadvantages in terms of classifiers.

The advantage of the Naive Bayes classifier is that it requires little training data to estimate the parameters needed for classification. So only the variance of the variables for each class needs to be determined and not all covariance matrices (Pattekari & Parveen, 2012).

The weakness of the Naïve Bayes classifier in prediction is that the predicted value is zero when any parameter is zero. However, zero probability can be overcome by smoothing techniques in the numerator and denominator to avoid the appearance of zero values (Adnan & Husain, 2012). And the probability estimation results are not maximized and features or attributes are still often wrong. The solution to overcome these problems is to add an effective feature selection method to improve accuracy (Nugroho et al., 2022).

The data in this study has relatively many features, so it needs to be selected to handle only the important features. handle only important features. So an initial process stage is needed which is called the pre-processing stage. The pre-processing stage is a stage for processing raw data to correct irregular and incomplete data. Normalization is a pre-processing stage. Normalization technique is a technique changing all attributes on the same scale or mapping technique at the preprocessing stage. And this technique helps in generating predictions (Patro & Sahu, 2015). Due to the data data used in this study are many attributes and need to be normalized, because some attributes have different values. attributes have different values.

Missing values are a common problem in data processing. The management of missing values becomes important when only a small number of samples are available (Ghannad-Rezaie et al., 2010). Many researchers work with such datasets, previously filling in the missing values with the

average of each column. with the average value of each column. However, in this study, researchers However, in this study, the researcher eliminated 6 data due to missing values. And it really impact on the classification results.

This research was taken from the kaggle website which is Cleveland heart disease data. The data amounted to 303 data, but only 297 data were used. Because 6 of them are incomplete or have missing values in them. So this research uses pre-processing techniques by removing missing values. The results will be combined with the Backward Eliminaton method as attribute feature selection. And continued classification using Naïve Bayes. So that this research is very influential for medical detecting heart disease and based on the research objectives described earlier, the benefits of this research are:

1. Practical Benefits :For medical personnel, the concept of methods that can be used to perform early diagnosis is through the classification process. Someone has heart disease or not using the Naïve Bayes algorithm based on Backward Elimination.
2. Theoretical Benefits :Able to know the implementation process in using pre-processing and classification techniques using the Naïve Bayes algorithm based on Backward Elimiantion.
3. Policy Benefits: Directing technology development policies for researchers involved in studies related to the materials and methods used in overcoming the problem of heart disease in Indonesia.
4. For future research: Hopefully, this proposed method can be applied to other datasets so as to expand the theory related to pre-processing techniques for the classification of heart disease predictions. This method can also be a consideration for other researchers in overcoming missing values.

RESEARCH METHODS

In this study, various methods and techniques were used, and the followed in Figure 1 below is the flow model in this study.

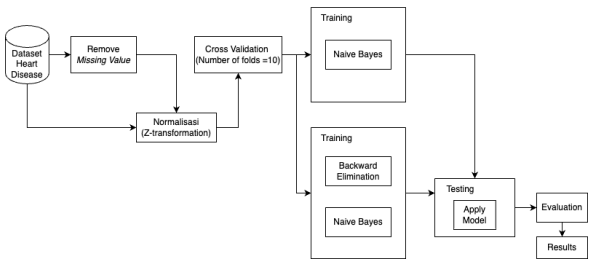

Figure 1
Research Workflow

Figure 1 explained that started from the removal of missing values in the dataset, then the data that has been pre-processed will be normalized using z-transformation. The used of normalization to converted data into a standard form which aims so that when classification is carried out no errors occur. Next to the dataset division stage in this study used cross validation. This research uses the value of k = 3, k = 5, and k = 10. Each sampled will be trained using the proposed method. After the data is trained, it will be brought and tested using testing data. From these results, accuracy is

obtained which will be used as an evaluation in the study.

Data Collection

This research uses public data taken from kaggle, the data contains 297 records that have been removed missing value as much as 6 data. The dataset consists of 14 variables included Age, Sex, Cp (chest pain type), trestbps (resting blood pressure), chol (Cholesterol), fbs (FastingBS), restecg (resting electrocardiographic results), thalach (maximum heart rate achieved), exang (exercise induced angina), Oldpeak (ST depression induced by exercise relative to rest), Slope (the slope of the peak exercise ST segment), ca (number of major vessels (0-3) colored by flourosopy), thal (0 = normal; 1 = fixed defect; 2 = reversable defectand the label), condition (0 = no disease, 1 = disease). This dataset will be used for experimental data for predicted those affected by heart disease.

Pre-processing techniques

Data Pre-processing technique is a process of converting raw data into a form of data that is easily understood by the system. There are several stages in data pre-processing, among others:

1. **Data Cleaned**

   The first stage is cleaned the data, which means that the raw data that has been obtained will be re-selected. Removed or eliminated data - incomplete data, will avoid analyzing a data. Some problems that often occur in datasets are Missing Values. Missing Values are values that are not filled or empty in the dataset. Missing Values can arise because there are respondents who do not answer all the questions in the questionnaire or during the manual or trial data entry process is wrong so that there are empty values in the datasets (Jiri, 2014). And there are several ways to overcome the problem of missing values, namely:

   a. **Reduced the Dataset**

      This dataset reduction is a simple solution in imputed missing values. This solution involves removed samples (rows) that have missing values (Kantardzic, 2011) or removed attributes (columns) that have missing values (Lakshminarayan et al., 1999). Both approaches can be combined, such removal of all samples is known as complete case analysis (Acuna & Rodriguez, 2004).

   b. **Replaced Missing values with the average value (mean)**

      This method replaced each value that has missing values with an average value (Kantardzic, 2011). The mean value is calculated based on all known values, this method can only be used for numeric value attributes and is combined with replacing missing values.

   c. **Replaced Missing values with the median value.**

      This method is almost the same as the previous one but this method uses the median value calculation. The median value calculation is based on all known values (Acuna & Rodriguez, 2004).

2. **Data Cleaned**

   The first stage is cleaned the data, which means that the raw data that has been obtained will be re-selected.

3. **Data Integration**

   The stage where combined data from various sources is collected into one larger data.

4. **Data Transformation**

Then there is the Transformation stage, this stage changes the data structure, data format so as to produced a dataset that is in accordance with the designed algorithm.

5. Data Reduction

The data reduction stage aims to reduce the amount of data taken. This stage needs to be considered and adjusted to the needs of whether the data being processed is large, medium, or even needs to be compressed and will risk being detrimental.

Backward Elimination Feature Selection

Backward Elimination is a method to eliminated attributes that are typical of the global best herd to improved search capabilities, the method is designed to be fast and effective by using a filter size based on mutual information (Nguyen et al., 2014).

This method involved all independent variables and then eliminates variables that are considered insignificant to the model. Steps to perform backward elimination:

1. First enter all independent variables into the model
2. If the independent variable produces a probability of F that is greater than the probability to remove then the independent variable is eliminated from the model.
3. The process can be stopped if all the probability of F is smaller than the probability to remove.

Naïve Bayes Classification

Naïve Bayes is a data classification method used the probability method proposed by an English scientist named Thomas Bayes. In a Naïve Bayes network, all features are conditionally independent. The Naïve Bayes algorithm is suitable for classifying high-dimensional datasets. This classification algorithm uses conditional independence. Conditional independence assumes that attribute values are independent of the values of other attributes in a class (Latha & Jeeva, 2019).

Naïve Bayes is a common approach used to predicted classes for various types of datasets such as educational data mining and medical data mining. This model is also useful for classifying various types of datasets such as sentiment analysis and virus detection (Nguyen et al., 2014).

To implemented the Naïve Bayes method, the followed equation can be used (Damanik et al., 2019) :

1. Calculated the probability value (total chance)

To calculated the probability value for each category, it is done by calculated the amount of data included in the category, then dividing it by the total amount of data in the category. If there are numerical values, the next step is to found the mean and standard deviation of each parameter that has numerical data.

$$\mu = \frac{\Sigma_{i=1}^{n} X_i}{n} \qquad (1)$$

Description:

μ : average count (mean)

Xi : x value to -i

n : number of samples

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{n} (X_i - \mu)^2}{n}} \qquad (2)$$

Description:
σ: standard deviation
xi : values x
μ: average count
n: number of samples

2. **Found the probability value for each feature in each class**

After calculated the mean and standard deviation, the next step is to found the probability value of the features in each class. To calculated the value, it is necessary to count the number of corresponding data in the same category, then divided it by the total number of data in that category.

3. **Gaussian Distribution Value**

After this step, the next step is to calculated the probability value for test data features that consist of numeric data.

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sigma\sqrt{2\pi}} . \; e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2} \quad (3)$$

Description :
P : Probability
Xi : value x
C : The class you are looking for
Ci : Y subclass to be searched
μ : Average value
σ : Standard Deviation

4. **Final Probability of Each Class**

The process of calculated the final probability for each class involved combined all the Gaussian distribution values into one similar class.

$$P(X|Class) = P(V1|Class * P(V2|Class) * P(V3|Class) * P(V4|Class) * P(V5|Class) * P(V6|Class) * P(V7|Class) * P(V8|Class) \quad (4)$$

5. **Final Probability**

The final probability is obtained by calculated the final probability value of each class used the Naïve Bayes Classifier method. The process of calculated the final probability is done as followed. Equation of Bayes' theorem (Latha & Jeeva, 2019) :

$$P(C_i|X) = \frac{P(X|C_i) . P(C_i)}{P(X)} \quad (5)$$

Description:
X : Data that has an unknown class
Ci : Hypothesis data is a specific class

$P(Ci| X)$: A posteriori probability of hypothesis
$P(Ci)$: Prior probability of hypothesis
$P(X|Ci)$: Probability associated with the condition in the hypothesis
$P(X)$: Probability of X

RESULTS AND DISCUSSION

This research is a classification study in the use of naive bayes method based on backward elimination using heart disease patient data. The heart disease dataset is processed through the removal of missing values, and z-score normalization then applying a set of feature selection methods using Backward Elimination. And the selected subset of features is fed into the Naïve Bayes method. Performance evaluation using cross validation. Cross validation is used to separate two subsets, namely learning process data and validation data. This research uses 3 sampling k namely 3, 5 and 10 fold. As seen in table 1.

Tabel 1
Experimental results of three sampling types using Naïve Bayes and Naïve Bayes based on Backward Elimination

K Fold	Sampling Type	Pre-processing technique + Naïve Bayes	Pre-processing technique + Naïve Bayes + Backward Elimination	Improved accuracy of deleted data
K = 3	Stratified Sampling	86.44%	91.53%	5.09%
K = 5	Stratified Sampling	79.66%	93.22%	13.56%
K = 10	Stratified Sampling	84.75%	98.31%	13.56%

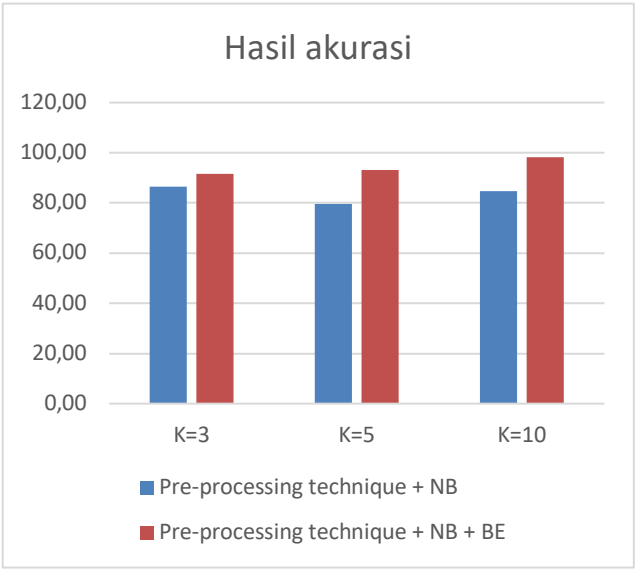

Figure 2
Final evaluation results

Figure 2 shows that the accuracy results using additional selection features, namely Backward elimination, affect the classification as shown by the value of K = 3, K = 5 and K = 10 which continues to increase at first from K = 3, namely 91.53%, then K = 5, namely 93.22% and K = 10, namely 98.31%. Compared to methods that do not use selection features, it decreases at the value of k = 3 as much as 86.44% then decreases at the value of k = 5 to 79.66%. Compared to the method that does not use the selection feature it has decreased at the value of k = 3 as much as 86.44% then dropped at the value of k = 5 to 79.66%. And back up at the value of k = 10 as much as 84.75%. So the result in this study is that the Backward elimination method is very influential in classification.

CONCLUSION

This research uses three different tests. The tests include data that has been removed missing values and does not remove missing values using the same method, namely using the Naïve Bayes method based on Backward Elimination. And the last test uses data that has been removed missing values and only uses the Naïve Bayes method without the help of the Backward Elimination selection feature, and all of these tests use rapidminer tools. Tests using data that has been pre-processed and using the Naïve Bayes method obtained an accuracy of 84.75%. While testing with datasets that have not been pre-processed using the Naïve Bayes method based on Backward Elimination obtained an accuracy of 89.83%. While testing that has been pre-processed using the same model, namely Naïve Bayes based Backward Elimination, attribute selection is carried out, from 13 attributes obtained 12 attributes are considered influential or relevant, namely age, sex, cp, trestbps, chol, fbs, restecg, thalach, oldpeak, slope, ca, thal. And there is 1 attribute that has no effect or is irrelevant, namely exang and the test obtained an accuracy of 98.31%. So it is concluded that the use of selection features has an effect in the classification process which is shown in the increase in accuracy at a value of k = 3 by 5.09%, a value of k = 5 by 13.56% and a value of k = 10 by 13.56%.

BIBLIOGRAPHY

Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*, 639–647. Google Scholar

Adnan, M. H. B. M., & Husain, W. (2012). A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction. *2012 International Conference on Computer & Information Science (ICCIS)*, *1*, 281–285. Google Scholar

Arie, A. A. P. G. B., & Suryandari, N. N. A. (2023). The Effect Of Good Corporate Governance, Company Size, And Leverage On The Integrity Of Financial Statements. *Jurnal Ekonomi, Teknologi Dan Bisnis (JETBIS)*, *2*(3), 310–324. Google Scholar

Damanik, H. J., Irawan, E., Damanik, I. S., & Wanto, A. (2019). Penerapan Algoritma Naive Bayes untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor. *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, *1*, 501–511. Google Scholar

Ghannad-Rezaie, M., Soltanian-Zadeh, H., Ying, H., & Dong, M. (2010). Selection–fusion approach for

classification of datasets with missing values. *Pattern Recognition*, *43*(6), 2340–2350. Google Scholar

Handayani, A., Hutahaean, J., & Nasution, A. (2022). Penerapan Metode Hybrid Case Base Pada Sistem Pakar Diagnosa Penyakit Jantung. *Building of Informatics, Technology and Science (BITS)*, *4*(2), 537–544. Google Scholar

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons. Google Scholar

Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, *11*(3), 259–275. Google Scholar

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203. Google Scholar

Maulana, D., & Yahya, R. (2019). Implementasi Algoritma Naïve Bayes Untuk Klasifikasi Penderita Penyakit Jantung Di Indonesia Menggunakan Rapid Miner. *Jurnal SIGMA*, *10*(2), 191–197. Google Scholar

Nguyen, H. B., Xue, B., Liu, I., & Zhang, M. (2014). Filter based backward elimination in wrapper based PSO for feature selection in classification. *2014 IEEE Congress on Evolutionary Computation (CEC)*, 3111–3118. Google Scholar

Nugroho, B. I., Lestari, N. P., Kurniawan, R. D., & Gunawan, G. (2022). Tinjauan Pustaka Sistematis: Data Mining Dalam Bidang Kesehatan. *Jurnal Ekonomi, Teknologi Dan Bisnis (JETBIS)*, *1*(1), 14–27. Google Scholar

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *ArXiv Preprint ArXiv:1503.06462*. Google Scholar

Pattekari, S. A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. *International Journal of Advanced Computer and Mathematical Sciences*, *3*(3), 290–294. Google Scholar

Reyes, E. B., Ha, J.-W., Firdaus, I., Ghazi, A. M., Phrommintikul, A., Sim, D., Vu, Q. N., Siu, C. W., Yin, W.-H., & Cowie, M. R. (2016). Heart failure across Asia: same healthcare burden but differences in organization of care. *International Journal of Cardiology*, *223*, 163–167. Google Scholar