Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease

Adi Purnomo¹, Mula Agung Barata¹, Moch Arief Soeleman¹, Farrikh Alzami¹
¹Department of Computer Science, Dian Nuswantoro University, Semarang, ID
E-mail: adipurnm745@gmail.com

Abstract. Feature selection in the classification model has a role to choose relevant and interconnected features in the data mining task. In the medical world, feature selection can help the classification model in predicting heart attack. Naive Bayes is one of the most popular classification learning methods that can help to predict patients in helping paramedics to make decisions. The addition of feature selection in the form of backward elimination can increase the accuracy of Naive Bayes by 89.45% from 84.29% previously. The results of this study indicate the accuracy of a backward selection method in predicting heart attack is quite high in adding accuracy.

1. Introduction
One of the important human organs is the heart. The modern lifestyle of people in big cities in developed countries is a factor causing high blood pressure, diabetes, and complication of heart disease. This happens because they do not care about dietary habits and are also less moving [1]. A heart attack is currently experiencing a very significant increase from year to year. According to WHO research estimates that 12 million deaths occur annually worldwide [2][3].

The WHO statistical report in 2012 clarifies the circumstances that one in three adults experience high blood tension which can cause a heart attack or stroke [4]. An early sign of heart disease can be identified by the following characteristics [5]: headache, dizziness, frequent fainting, discomfort when eating, shortness of breath, fatigue without cause, chest pain, and palpitations. While the factors that cause heart disease are as follows: high blood pressure, a history of diabetes, smoking, high cholesterol, and obesity [5].

The process of diagnosing heart disease is a very tedious job for medical employees or doctors [5]. The process of data mining or also called knowledge discovery in databases can help to find hidden information in the heart disease dataset that can replace the role of the doctors. Data mining is the procedure of recognizing information or knowledge for decision making and extracting data into useful information [6]. Various methods in machine learning as well as k-NN, Naïve Bayes, SVM, Linear Regression, etc can provide benefits in processing large data to assist humans in making decisions in various matters for example in the field of medical, education, manufacturing and industry.

2. Related Work
Subbalakshmi et al. concluded that the technique of processing patient data using the Naïve Bayes method is very suitable when the dataset has many attributes [6]. Although this method is simple or
easy to do, it often can outperform other classification methods. Several previous studies conducted by Rajkumar et al. on Naïve Bayes classification concluded that an accuracy of the Naïve Bayes is better than the k-NN algorithm [5]. Soni Jyoti et al. conducted an experiment by comparing the accomplishment of several learning algorithms as though Naïve Bayes method, Neural Network method, Decision Tree method, and k-NN to detect heart attack illness results in a conclusion that Decision Tree accuracy is better than the other three algorithms. But after being combined with genetic algorithms on Naïve Bayes it has increased accuracy significantly.

Bathla, Nidhi., Jyoti, Kiran predicts heart disease employ Decision Tree method and Neural Networks algorithm. The experiments produce that the Neural Network method delivers the upper accuracy up to 100% compared to the Decision Tree which only produces an accuracy in the amount of 99.2% [4]. Researchers [7] propose to use the Naïve Bayes approach in classification to predicted heart illness. With this method, the result shows that the Naïve Bayes algorithm has an accuracy value of 86.41% with minimum time.

3. Research Methodology

This research describes various data mining techniques, especially on the Naïve Bayes algorithm which can help paramedic analysis to predict a heart attack. The methodology in this research is by reading journals, publications, reviews in the discipline of computer science especially data mining and heart disease. This study applies the Naïve Bayes combination algorithm by adding feature selection in the classification of heart disease. So that better accuracy can be generated.

![Proposed Method](image)

Figure 1. Proposed Method.

3.1. Heart attack dataset

The dataset used in this paper is from UC Repository. The UC Repository is the center of a public dataset that can be taken for free for research. This heart disease dataset contains 100 patient records consisting of 7 attributes and 1 label. The seven attributes are the following: age, marriage, weight category, cholesterol, stress training, and stress level. Below is the input attribute of the heart attack illness dataset.
Table 1. Heart attack illness dataset.

| Number | Attribute          | Description          |
|--------|-------------------|----------------------|
| 1      | Marriage          | 0: Single            |
|        |                   | 1: Married           |
|        |                   | 2: Widow             |
| 2      | Gender            | 0: Woman             |
|        |                   | 1: Man               |
| 3      | Weight            | 0: Normal            |
|        |                   | 1: Middle            |
|        |                   | 2: Obesity           |
| 4      | Stress Management | 0: No                |
|        |                   | 1: Yes               |
| 5      | Stress Level      | 0-100                |
| 6      | Coronary          | 0: No                |
|        |                   | 1: Yes               |

4. Research Results

In this section, we execute experiments carried out using the RapidMiner application as many as 100 datasets of heart disease patients were entered with 7 attributes and 1 label. These experiments were committed against Core i3 with 1.9 GHz processing pace and 6 GB main memory. This dataset comes from UCI [8]. Other than that, this dataset also used in the literature [9][10]. To boost the accuracy of prediction in the classification, many researchers have been addressed the dimensionality problem of using the feature selection approach [14]. Feature selection needed to decrease the dimension of an attribute for large data [11]. Before the data is entered in the model, the data separated into training data and data testing employ the 10-fold method of cross-validation [13]. After that, three methods of feature selection are backward elimination, forward selection, and weight by chi-square are added. Naïve Bayes inserted into the model because of the simplicity of the method [15]. The Naïve Bayes + Forward Selection method is the best method among other methods. Naïve Bayes + Backward Elimination produces an accuracy of 90.27%, then followed by Naïve Bayes + Optimize Selection produces an accuracy of 94.89% and the last and highest accuracy is the Naïve Bayes + Forward Selection method with an accuracy of 95.44%.

The result of the experiment about classification accuracy of heart attack dataset using three feature selection methods as a forward selection method, backward elimination method, and optimize selection method in Table 2. Table 2 shows the differences in accuracy of four methods, namely Original Naïve Bayes, Backward Elimination in Naïve Bayes, Optimize Selection in Naïve Bayes, and Naïve Bayes with Forwarding Selection.

Table 2. Accuracy of various approaches with Naïve Bayes.

| Our Approach                        | Accuracy |
|-------------------------------------|----------|
| Naïve Bayes                         | 84.29%   |
| Naïve Bayes + Backward Elimination  | 90.27%   |
| Naïve Bayes + Optimize Selection    | 94.89%   |
| Naïve Bayes + Forward Selection     | 95.44%   |
Furthermore, the performance evaluation is conducted in terms of the classifier model based on accuracy [16]. The results of a test with RapidMiner using the Naïve Bayes method without a combination of feature selection obtained an accuracy of 84.29%, a precision value of 78.63% and AUC of 0.961. Figure 2 presents the accuracy results of the Naïve Bayes method without feature selection. Figure 3 presents the accuracy results of Naïve Bayes combined with the backward elimination method. Figure 4 presents the accuracy results of Naïve Bayes combined with the optimized selection method. Figure 5 presents the accuracy results of Naïve Bayes combined with the forwarding selection method. Figure 6 presents overall from other feature selection methods combined with Naïve Bayes.

![Figure 2. Naive Bayes without feature selection accuracy results.](image1)

For the test results using a Naïve Bayes classification algorithm and added with a backward elimination feature, we get an accuracy of 90.27%, a precision of 87.36%, and an AUC value of 0.973. Meanwhile, the results using the Naïve Bayes and the addition of optimized selection, the accuracy is 94.89%, the precision is 98.57%, and the AUC value is 0.979. The chart result of Naïve Bayes classification algorithm added with backward elimination can be seen in Figure 3 and chart result of Naïve Bayes classification algorithm added with optimized selection can be seen in Figure 4 below.

![Figure 3. Naive Bayes with backward elimination accuracy results.](image2)

And finally, as seen in Figure 5 shows the Naïve Bayes classification method with the addition of forwarding selection obtained high accuracy of 95.44%, a precision value of 98.75%, and AUC value of 0.960. this approach (forward selection) outperformed other methods (backward elimination, optimized selection, Naïve Bayes original). Therefore, this paper proposes using forward selection in Naïve Bayes compared to three feature selection techniques to get the best results.
5. Conclusion and Future Work

We discover that new approach forward selection in classifier model Naïve Bayes outperforms other feature selection approach. From this research, it is concluded that for predict heart attack, forward selection in Naïve Bayes model obtained high accuracy of 95.44%, meanwhile Naïve Bayes without feature selection obtained an accuracy of 84.29%, Naïve Bayes with backward elimination feature achieve accuracy of 90.27%, Naïve Bayes with optimizing selection achieve accuracy of 94.89%. For further research in predicting heart disease, the Naïve Bayes method can be integrated with feature generation and feature weighting to get maximum accuracy.
References

[1] J. Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction,” vol. 17, no. 8, pp. 43–48, 2011.
[2] H. D. Masethe and M. A. Masethe, “Prediction of heart disease using classification algorithms,” Lect. Notes Eng. Comput. Sci., vol. 2, pp. 809–812, 2014.
[3] S. A. Pattekari and A. Parveen, “Prediction Systems for Heart Disease Using Naive Bayes,” vol. 3, no. 3, pp. 290–294, 2012.
[4] N. Bhatla and K. Jyoti, “An analysis of heart disease prediction using different data mining techniques,” Int. J. Eng. Res. Tecnol., vol. 1, no. 8, pp. 1–4, 2012.
[5] A. Rajkumar and G. S. Reena, “Diagnosis Of Heart Disease Using Datamining Algorithm,” Glob. J. Comput. Sci. Technol., vol. 10, no. 10, pp. 38–43, 2010.
[6] G. S. M. Tech, “Decision Support in Heart Disease Prediction System using Naive Bayes,” vol. 2, no. 2, pp. 170–176, 2011.
[7] M. Kumari and S. Godara, “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction,” Int. J. Comput. Sci. Trends Technol., vol. 2, no. 2, pp. 304–308, 2011.
[8] S. Belciug and F. Gorunescu, “Error-correction learning for artificial neural networks using the Bayesian paradigm. Application to automated medical diagnosis,” J. Biomed. Inform., vol. 52, pp. 329–337, 2014.
[9] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, “A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods,” vol. 14, no. 12, pp. 868–879, 2017.
[10] M. Fatima and M. Pasha, “Survey of Machine Learning Algorithms for Disease Diagnostic,” J. Intell. Learn. Syst. Appl., vol. 09, no. 01, pp. 1–16, 2017.
[11] S. Wang and J. Wei, “Feature selection based on measurement of ability to classify subproblems,” Neurocomputing, vol. 224, no. March, pp. 155–165, 2017.
[12] Panday P and Godara N 2012 Decision Support System for Cardiovascular Heart Disease Diagnosis using Improved Multilayer Perceptron 45 12–20.
[13] Jabbar M A and Samreem S 2017 Heart disease prediction system based on hidden naive bayes classifier 2016 Int. Conf. Circuits, Control. Commun. Comput. I4C 2016.
[14] Ghaemi M and Feizi-Derakhshi M R 2016 Feature selection using Forest Optimization Algorithm Pattern Recognit. 60 121–9.
[15] Aziz R, Verma C K and Srivastava N 2017 A novel approach for dimension reduction of microarray Comput. Biol. Chem. 71 161–9.
[16] Medjahed S A, Saadi T A, Benyettou A and Ouali M 2017 Kernel-based learning and feature selection analysis for cancer diagnosis Appl. Soft Comput. J. 51 39–48.