COMPARISON OF DATA MINING CLASSIFICATION METHODS TO DETECT HEART DISEASE

Ira Ekanda Putri[1]; Dwi Rahmawati[2]; Yufis Azhar[3]

Informatics Engineering Study Program
University of Muhammadiyah Malang
www.umm.ac.id
[1]iraekandaputri@webmail.umm.ac.id, [2]dwirahma@webmail.umm.ac.id, [3]yufis@umm.ac.id

Abstract*— Heart disease is a disease that is deadly and must be treated as soon as possible because if it is too late, it has a big risk to one's life. Factors causing the disease of the heart is the use of tobacco, the physical who are less active, and an unhealthy diet. With existing data, the study is to compare the three algorithms, namely: Naive Bayes, Logistic Regression, and Support Vector Machine (SVM) which aims to determine the level of accuracy of the best of the dataset that is used to predict disease heart. This research produces the best accuracy of 87%, which is generated by the Naive Bayes method.*

Keywords: *Heart Disease, Classification, Data Mining*

Abstrak— *Penyakit jantung merupakan salah satu penyakit yang mematikan dan harus segera diobati karena jika terlambat beresiko besar bagi nyawa seseorang. Faktor penyebab penyakit jantung adalah penggunaan tembakau, fisik yang kurang aktif, dan pola makan yang tidak sehat. Dengan data yang ada maka penelitian ini membandingkan tiga algoritma yaitu: Naive Bayes, Logistic Regression, dan Support Vector Machine (SVM) yang bertujuan untuk mengetahui tingkat akurasi terbaik dari dataset yang digunakan untuk memprediksi penyakit jantung. Penelitian ini menghasilkan akurasi terbaik sebesar 87% yang dihasilkan dengan metode Naive Bayes.*

Kata Kunci*: Penyakit Jantung, Klasifikasi, Data Mining*

INTRODUCTION

The heart is one of the organs of the body of a man which is very important because the heart has the function of pumping blood to the whole body. The heart has a risk of diseases that are very large and can result in death. One of the heart diseases is coronary heart disease. Cardiac Coronary an interruption that occurred in the system vessel blood great. Thus causing the heart and circulation of blood not working as it should (Aeni et al., 2014). Cause primary disease of the heart is the use of tobacco, physically not inactive, diets that are not healthy and the use of alcohol, the risk of disease heart increases with increasing age, pressure blood high, having cholesterol high, and excess weight body (Lestari, 2015). Knowledge society about the symptoms of disease heart is still very low (Adrian, 2020) and less accurate as of the equipment that is used to detect diseases of the heart (Aeni et al., 2014) if only to control the sugar and pressure of blood (Aeni et al., 2014) and Data laboratory which has not functioned as effectively be used to detect heart disease. To deal with the problem, the effectiveness and accuracy in detecting diseases of the heart are then made system detecting disease heart using implementations of the algorithm or method of classification data mining (Lestari, 2015), (Rohman, 2016).

Many studies on the prediction of heart disease by using the method of classification data mining, research previously were using datasets from statlog heart disease, gain accuracy highs of 84.7% by using methods Naïve Bayes (Putra & Rini, 2019). Then in previous research which also used a dataset of heart disease catalogue, with the highest accuracy using the Logistic Regression method with an accuracy of 85% (Dwivedi, 2018). Furthermore, the research that uses the dataset from Cleveland's heart dataset, where the results of the accuracy of the highest obtained by a method Support Vector Machine (SVM) obtained an accuracy of 86.87% (Amin et al., 2019).

Several studies that discuss heart disease using data mining classification methods have resulted in different accuracy with different datasets and the number of records. The purpose of this study was to determine the most optimal method of the three best methods produced by previous research, namely the Naive Bayes method, Logistic Regression, and Support Vector Machine (SVM) with the same dataset and amount of data, so in this study a comparison of the three algorithms. This study also conducted k-fold cross-validation testing to test the model in the training stage (data validation) in order to limit problems such as overfitting. In addition to solving the problem of overfitting, testing using k-fold cross

DOI: 10.33480/pilar.v16i2.1481

validation also serves to produce stable results even though the data used is random.

MATERIALS AND METHODS

Naive Bayes classifier

One of the best operative classifications is the simplest Bayesian network. This network consists of a network- like structure with the accompanying probabilities. But can be combined with estimates of the density of the kernel and achieve the level of accuracy that is much higher. Definition others say Naive Bayes is a classification by the methods of probability and statistics were presented by scientists British Thomas Bayes, and predictor opportunities in the future will be based on the experience in the past before (Informatikalogi, 2017)

The naïve bayes equation

$$P(a_i \mid v_j) = \frac{n_c+mp}{n+m} \text{ .... (1)}$$

p = Prior estimate
n_c = The amount of training data where $v = v_j$dan
$a = a_i$
n = Where is the amount of training data $v = v_j$
m = Equivalent sample size
Source : (Informatikalogi, 2017)

Logistic Regression

Logistic regression is an analysis of regression that is appropriate to do when a variable dependent is dichotomous (binary) (Huang, 2019). Like all regression analyzes, logistic regression is a predictive analysis. Logistic regression is used to describe data and describe the relationship between one binary dependent variable and one or more independent variables nominal, ordinal, interval, or -level ratio (Huang, 2019)

Logistic Regression Equations

$$ln\left[\frac{p^\wedge}{1-p^\wedge}\right] = \beta_0 + \beta_1 x \text{ .... (2)}$$

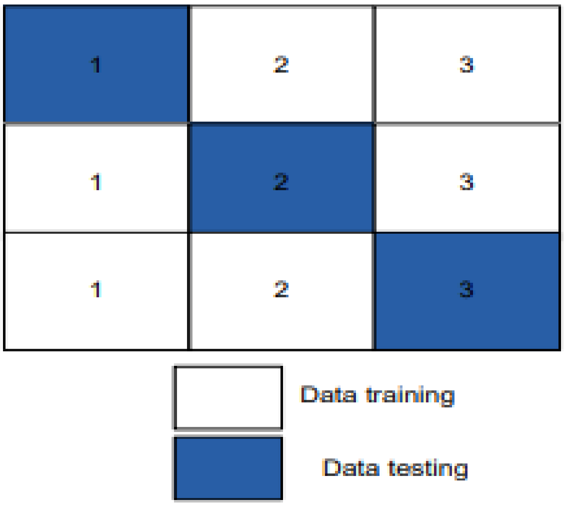
ln = Natural Logarithm
$p^\wedge$ = Logistic probability where $\frac{e^{\beta_0+\beta_1 x}}{1+ e^{\beta_0+\beta_1 x}}$
$\beta_0 + \beta_1 x$ = equation are the usual known in the OLS.
Source : (Huang, 2019)

Support Vector Machine

Support Vector Machine (SVM) is a linear model for classification and regression problems. SVM can solve the problem of linear or nonlinear and work with both for many problems practically. The idea of a simple SVM is an algorithm that creates a line or hyperplane that separates data into classes (Samsudiney, 2019). SVM has several kernels that are often used, namely the Linear, Sigmoid, and Radial Basis Function (RBF) kernels which will then be tested for the accuracy of each kernel to get the best kernel.

K-Fold Cross-Validation

Cross-validation is a model validation technique to assess how the results of statistical analysis will generalize to independent dataset (Salsabila, 2019). This technique is mainly used to make model predictions and estimate how accurate a predictive model will be when it is run in practice. On the approach to method K-Fold Cross Validation, the dataset is divided into several pieces of partition random. Furthermore, do several k-this time experiments with each experiment using the data partition took as a data testing and using the rest of the partition other as training data. Experiments will be carried out following the number of partitions that do (Supartini et al., 2017).



Source : (Tempola et al., 2018)
Figure 2. Model *3-Fold Cross-Validation*

Figure 2 shows the use of 3-fold cross-validation. Wherein each of the data will be in the execution as much as 3 times and each subset of data will have the opportunity as a data testing or training data. The model test as follows with the assumed name of each division of the data, which is D1, D2, and D3 (Tempola et al., 2018):

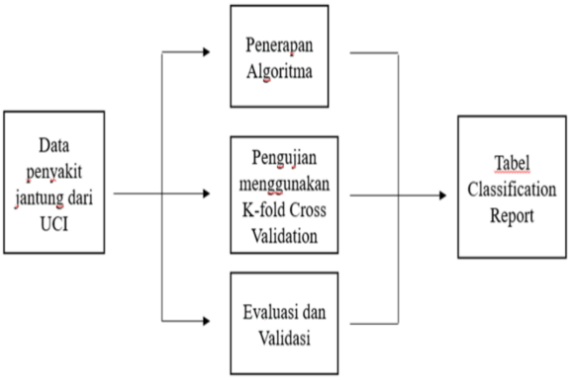
1. The first experiment was D1 data as testing data, while D2 and D3 were used as training data.
2. The second experiment was D2 data as testing data, while D1 and D3 data were used as training data.
3. In the last experiment or the third experiment, D3 data was used as testing data, while D1 and D2 were used as training data.

This study used data taken from the University of California, Irvine (UCI) Machine Learning Repository (Janosi et al., 1988). The dataset used is derived from data from patients at the Cleveland Clinic Foundation, which conducted heart disease screening. Results from the data that there were 303 patients were examined, and as many as 165 patients detected pain, and 138 patients detected healthy. Datasets used have 14 attributes are used to diagnose diseases of heart, namely,

1. Age: age (written in the form of numbers with units of the year)
2. Sex: the type of sex (1 for men and 0 for women)
3. Cp: type of chest pain
4. Trestbps: pressure blood breaks (in mm Hg when the entrance to the house sick)
5. Col: serum cholesterol in mg / dl
6. Fbs: fasting blood sugar > 120 mg / dl → (1 = true; 0 = false)
7. Restecg: rest electrocardiography results
8. Thalach: the beating heart of the maximum
9. Exang: exercise- induced angina (1 = yes; 0 = no)
10. Oldpeak: ST depression induced by exercise relative to rest
11. Slope: the slope of the peak exercise ST segment
12. Ca: number of main blood vessels (0–3) stained with fluoroscopy
13. Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

The third thirteen attributes refer to one attribute that is a target. Target is the result of the calculation of 13 attributes that were used and resulted in the conclusion that the patient is detected sick or healthy were written 1 of the patient's pains and 0 if the patient is healthy. The data used in this study were divided into two, namely training data of 80% or 242 data and testing data of 20% or 61 data. Then the dataset is processed using the Naive Bayes method, Logistic Regression, and Support Vector Machine (SVM) to see which algorithm has the highest accuracy.

Stage The first who performed in the research of this is to collect the data in which the researchers used data taken from the University of California, Irvine (UCI) Machine Learning Repository (Janosi et al., 1988) . Based on the source of data, the source of the data is divided into two, namely primary data and secondary. Primary data is data that is obtained is directly from the researchers, while the data secondary is data that is obtained from investigators from sources that already exist either published or not published (C.R Khotari, 2004). In this study, researchers used a comparative research type. Research comparatively a study that is comparing (Hughes, 2008). This research was conducted to compare the similarities and differences between two or more facts and the properties of the object under study based on the research problem. The model proposed in this study is to use three data mining classification methods. The third method that is proposed is Naive Bayes, Logistic Regression, and Support Vector Machine (SVM). In this study, it will do some of the steps or stages of research, such as Figure 1.


Data penyakit jantung dari UCI
Penerapan Algoritma
Pengujian menggunakan K-fold Cross Validation
Evaluasi dan Validasi
Tabel Classification Report

Source: (Putri et al., 2020)
Figure 1. Stage of research

RESULTS AND DISCUSSION

Table 1. Comparison of SVM kernel accuracy results

Model	Level of Accuracy
Kernel RBF	0,63
Kernel Sigmoid	0,55
Kernel Linear	0,84

Source: (Putri et al., 2020)

Based on the results of the value of accuracy that in the produce of testing 3-fold cross-validation for each Kernel (Table 1), the study is using the function kernel Linear which is used to classify the data because it has the value of the accuracy of the highest compared two models function kernel other, the value of accuracy equal to 0.84 in predicting. Meaning that as much as 84 % of the data right in the prediction.

DOI: 10.33480/pilar.v16i2.1481

Table 2. Confusion matrix from the prediction results of the Naïve Bayes.

	Predicted class	
	1	0
Actual class		
1	28	5
0	3	25

Source: (Putri et al., 2020)

Table 3. Confusion matrix from the prediction results of the Logistic Regression.

	Predicted class	
	1	0
Actual class		
1	26	7
0	2	26

Source: (Putri et al., 2020)

Table 4. Confusion matrix from the prediction results of the SVM.

	Predicted class	
	1	0
Actual class		
1	24	9
0	1	27

Source: (Putri et al., 2020)

Tools to measure the degree of accuracy of classification methods are used among another confusion matrix. Confusion matrices from prediction results are placed in table 2 to table 4 for Naive Bayes, Logistic Regression, and Support Vector Machine (SVM). Based on the three tables that proved that Naive Bayes predict the number of true positives was highest (Table 2). The highest number of true negatives is generated by the SVM method (table 4). Then for the second-highest number of true negative and true positives is generated by the Logistic Regression method (table 3). Further, to see a comparison of the three methods were used on the 14 attributes are used as a parameter for predicting disease heart can be seen in the Table 5 below:

Table 5. Comparison of the Naïve Bayes method, Logistic Regression, and *SVM.*

	Precision	Recall	F1-score	accuracy
Naïve Bayes	86,5%	87%	87%	87%
Logistic Regression	83%	83%	82%	82%
SVM	85%	86%	85%	85%

Source: (Putri et al., 2020)

Heart disease prediction using the 14 attributes that have been mentioned above, can be used as a benchmark to predict a person's heart condition. The Naive Bayes method produces an accuracy value of 87%, then the Logistic Regression method produces an accuracy of 82 %, and the SVM method produces an accuracy of 85 %. From the results, it shows that using 14 attributes is the same for predicting disease heart method Naive Bayes is a method best than Logistic Regression and SVM.

Measurement of the performance of the three methods of classification are used in the study is based on the accuracy of the system by using the equation as follows:

$$accuracy = \frac{\sum klasifikasi\ benar}{\sum data\ uji} \times 100\% \text{ .... (3)}$$

It was then continued with data validation with k-fold cross-validation. Research is using the value K = 3, which means dividing the number of data into three in the same lot that is 101 data. Data validation illustration as in Figure 1 of testing 3- fold cross-validation resulted in an average accuracy of methods Naïve Bayes 0.8 05, then the Logistic Regression of 0.8 35 and method of SVM for 0, 838.

K-fold cross-validation serves as an examiner resilience of a method and can demonstrate the results were stable despite using data randomly. On research is the accuracy of Naive Bayes down after tested using the k-fold cross validation because the method Naïve Bayes have properties of probabilistic where the concept of probabilistic highly dependent on the overall training data that result when the training data change the accuracy of which is produced also change. In contrast well with the method SVM and Logistic Regression are not influential in changing the training data because both methods are only seeing in part data and do not see the entirety of data, as well as data that is on the edge value, is more stable than the value of accuracy that produced more high.

CONCLUSION

Accuracy best is produced without testing k-fold cross-validation is a method Naïve Bayes while accuracy is best after conducted testing of k-fold cross-validation is a method of Support Vector Machine (SVM). Having done testing k-fold cross-validation accuracy of methods Naïve Bayes experienced a decline because highly dependent on the overall training data that resulted in when training data undergo changes in the accuracy of which is produced also undergo changes or not stable, while for the method of SVM when tested using the k-fold cross-validation results are more stable due to the SVM only look at in part data and do not see a whole the data then from the value of accuracy that produced more high.

REFERENCE

Adrian, K. (2020). *Beberapa Fakta Terkait Penyakit Jantung yang Perlu Diketahui*. Alodokter.Com. https://www.alodokter.com/beberapa-fakta-terkait-penyakit-jantung-yang-perlu-diketahui

Aeni, W. N., Santosa, S., & Supriyanto, C. (2014). Algoritma Klasifikasi data mining naïve bayes berbasis Particle Swarm Optimization untuk deteksi penyakit jantung. *Jurnal Pseudocode*, *1*(1), 11–14. https://ejournal.unib.ac.id/index.php/pseudocode/article/view/57/

Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, 82–93. https://doi.org/10.1016/j.tele.2018.11.007

C.R Khotari. (2004). *Research Methodology* (Second Rev). New Delhi : New Age International (P) Ltd., ©2004 (OCoLC)62197369.

Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications*, *29*(10), 685–693. https://doi.org/10.1007/s00521-016-2604-1

Huang, H. (2019). *Analisis Regresi Logistik Biner*. Globalstat Academic. https://www.globalstatistik.com/analisis-regresi-logistik-biner/

Hughes, R. (2008). KOMPARASI ALGORITMA MULTI LAYER PERCEPTRON DAN RADIAL BASIS FUNCTION UNTUK DIAGNOSA PENYAKIT JANTUNG. *Journal of Chemical Information and Modeling*, *53*(9), 287. https://doi.org/10.1017/CBO9781107415324.004

Informatikalogi. (2017). *Algoritma Naive Bayes*. Informatikalogi.Com. https://informatikalogi.com/algoritma-naive-bayes/

Janosi, A., Steinbrunn, J., Pfisterer, M., & Detrano, R. (1988). *Heart Disease Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+disease

Lestari, M. (2015). Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) untuk Mendeteksi Penyakit Jantung. *Faktor Exacta*, *7*(4), 366–371. http://journal.lppmunindra.ac.id/index.php/Faktor_Exacta/article/view/290

Putra, P. D., & Rini, D. P. (2019). Prediksi Penyakit Jantung dengan Algoritma Klasifikasi. *Prosiding Annual Research Seminar*, *5*(1), 95–99. http://www.seminar.ilkom.unsri.ac.id/index.php/ars/article/view/2118

Putri, I. E., Rahmawati, D., Azhar, Y., Malang, U. M., & Informatika, P. S. (2020). *PERBANDINGAN METODE KLASIFIKASI DATA MINING UNTUK* (Vol. 16, Issue 1).

Rohman, A. (2016). KOMPARASI METODE KLASIFIKASI DATA MINING UNTUK PREDIKSI PENYAKIT JANTUNG. *Neo Teknika: Jurnal Ilmiah Teknologi*, *2*(2), 21–28. http://jurnal.unpand.ac.id/index.php/NT/article/view/766

Salsabila, A. (2019). *Cross Validation of KNN using R*. Medium.Com. https://medium.com/@asalsabila36/cross-validation-of-knn-using-r-84089b21de0f

Samsudiney. (2019). *Penjelasan Sederhana tentang Apa Itu SVM?* Medium.Com. https://medium.com/@samsudiney/penjelasan-sederhana-tentang-apa-itu-svm-149fec72bd02

Supartini, I. A. M., Sukarsa, I. K. G., & Srinadi, I. G. A. M. (2017). Analisis Diskriminan Pada Klasifikasi Desa Di Kabupaten Tabanan Menggunakan Metode K-Fold Cross Validation. *E-Jurnal Matematika*, *6*(2), 106–115. https://doi.org/10.24843/mtk.2017.v06.i02.p154

Tempola, F., Muhammad, M., & Khairan, A. (2018). Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation Comparison of Classification Between Knn and Naive Bayes At the Determination of the Volcanic Status With K-Fold Cross. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *5*(5), 577–584. https://doi.org/10.25126/jtiik20185983