Jurnal Matematika MANTIK
Vol. 8, No. 1, June 2022, pp. 67-77

ISSN: 2527-3159 (*print*) 2527-3167 (*online*)

Application of Machine Learning for Heart Disease Classification Using Naive Bayes

Siti Hadijah Hasanah

Department of Statistics, Faculty of Science and Technology, Universitas Terbuka
,

Article history:
Received Nov 21, 2021
Revised May 27, 2022
Accepted May 31, 2022

Kata Kunci:
Penyakit jantung,
Klasifikasi,
Machine learning,
Naive Bayes

Abstrak. Naive Bayes classifier menggunakan pendekatan dari suatu teorema Bayes dengan penggabungan pengetahuan sebelumnya dengan yang baru. Tujuan penelitian ini adalah untuk mengembangkan machine learning dengan menggunakan teknik klasifikasi Naive Bayes dan sebagai sistem keputusan dalam menghasilkan ketepatan klasifikasi yang cepat dan akurat dalam mendiagnosis penyakit kardiovaskuler seperti penyakit jantung. Penyakit kardiovaskular merupakan penyebab kematian utama, 32% dari seluruh kematian global dimana 85% di antaranya disebabkan oleh stroke dan serangan jantung. Berdasarkan hasil analisis didapatkan bahwa akurasi ketepatan klasifikasi di data training pada data pasien diklasifikasikan tepat memiliki dan tidak memiliki penyakit jantung masing-masing sebesar 83,21% dan 83,81%. Pada data testing persentase data pasien diklasifikasi tepat memiliki dan tidak memiliki penyakit jantung masing-masing sebesar 83,78% dan 87,50%. Berdasarkan nilai AUC di data training dan data testing masing-masing sebesar 83,15% dan 85,24%. Maka dari hasil tersebut dapat disimpulkan bahwa metode Naive Bayes baik digunakan pada klafisikasi data pasien penyakit jantung.

Keywords:
Heart disease,
classification,
Machine learning,
Naive Bayes

Abstract. The Naive Bayes classifier uses an approximation of a bayes theorem by combining previous knowledge with new ones. The purpose of this research is to develop machine learning using Naive Bayes classification techniques and as a decision system in producing fast and accurate classification accuracy in diagnosing cardiovascular diseases such as heart disease. Cardiovascular disease is the leading cause of death, 32% of all global deaths, of which 85% are caused by stroke and heart disease. Based on the results of the analysis, it was found that the accuracy of classification accuracy in the training data on patient data was classified as having and not having heart disease, respectively 83,21% and 83,1%. In data testing, the percentage of patient data classified as having and not having heart disease was 83,78% and 87,50%, respectively. Based on the AUC values in the training data and testing data, they are 83,15% and 85,24%, respectively. So, from these results, it can be concluded that the Naive Bayes method is good for classifying heart disease patient data.

How to cite:
S. H. Hasanah, “Application of Machine Learning for Heart Disease Classification Using Naive Bayes”. *J. Mat Mantik*, vol. 8, no. 1, pp. 68-77, Jun. 2022.

CONTACT:
Siti Hadijah Hasanah, sitihadijah@ecampus.ut.ac.id Department of Statistics, Universitas Terbuka, Jl. Pd. Cabe Raya, Pd Cabe Udik Pamulang, Tangerang Selatan, 15418, Indonesia

The article can be accessed here. *https://doi.org/10.15642/mantik.2022.8.1.68-77*

1. Introduction

According to WHO globally cardiovascular disease (CVDs) is the leading cause of death, in 2019 an estimated 17.9 million died from CDV, and this number represents 32% of all global deaths of which 85% are due to stroke and heart attack [1]. Currently, globally, developing low- and middle-income countries have the highest cardiovascular risk (CVD) and are common in adults aged 40 years and over [2]. Several factors that cause cardiovascular disease include heredity, uncontrolled high blood pressure, increased prevalence of diabetes, and obesity [3].

The rapid development of science in all fields automatically produces a very large amount of data, data mining has a role in handling large amounts of data. The basic functions of data mining include classification, clustering, and association [4]. Classification is an important data mining technique with a wide range of applications in classifying various types of data [5]. The application of classification can be predicted quickly and accurately by using several machine learning algorithms such as support vector learning (SVM) [6], [7], random forest, and CART [8], K-Nearest Neighbor, Genetic Algorithm [4], and Artificial Neural Network (ANN) [9].

The purpose of this research is to develop machine learning using Naive Bayes classification techniques and as a decision system in producing fast and accurate classification accuracy in diagnosing cardiovascular diseases such as heart disease. Naive Bayes is a data mining algorithm that is quite popular in classification [10]. A naive Bayes Classifier uses an approximation of a Bayes theorem by combining previous knowledge with new ones [11]. The advantage of this method is the use of a simple algorithm [12] and has high accuracy [11].

Several studies that apply the Naive Bayes classification method include in the field of education such as regarding the interest of students in determining majors in high school [13], in the economic field, namely the classification of data for underprivileged citizens so that the aid funds provided by the government to them are right on target [13], [14], classifying customers who are right on target in receiving credit and as a way to avoid the risk of default in the future [15], in the social sector regarding the analysis of public sentiment towards the development of e-sports education [16].

This article will present the application of classification using Naive Bayes whose implementation is assisted by the Python program, in the hope that it can continue our research. In the future, the author will apply several machine learning methods and at the same time use optimization in machine learning.

2. Methods

The data for this article was obtained from UCI Machine Learning in the form of a Heart Disease Data Set [17], with a total of 330 data records. The data is divided into 2, namely training data (80%) and testing data (20%) [18], and consists of 14 variables including 13 predictor variables, 1 response variable which is seen in table 1. The response variable describes the target in the classification with 2 categories, namely if < 50 % diameter narrowing means that they do not have heart disease (category 0) and if > 50% diameter is narrowing then it has heart disease (category 1) and flowchart of Naive Bayes Classification can be seen in Figure 1.

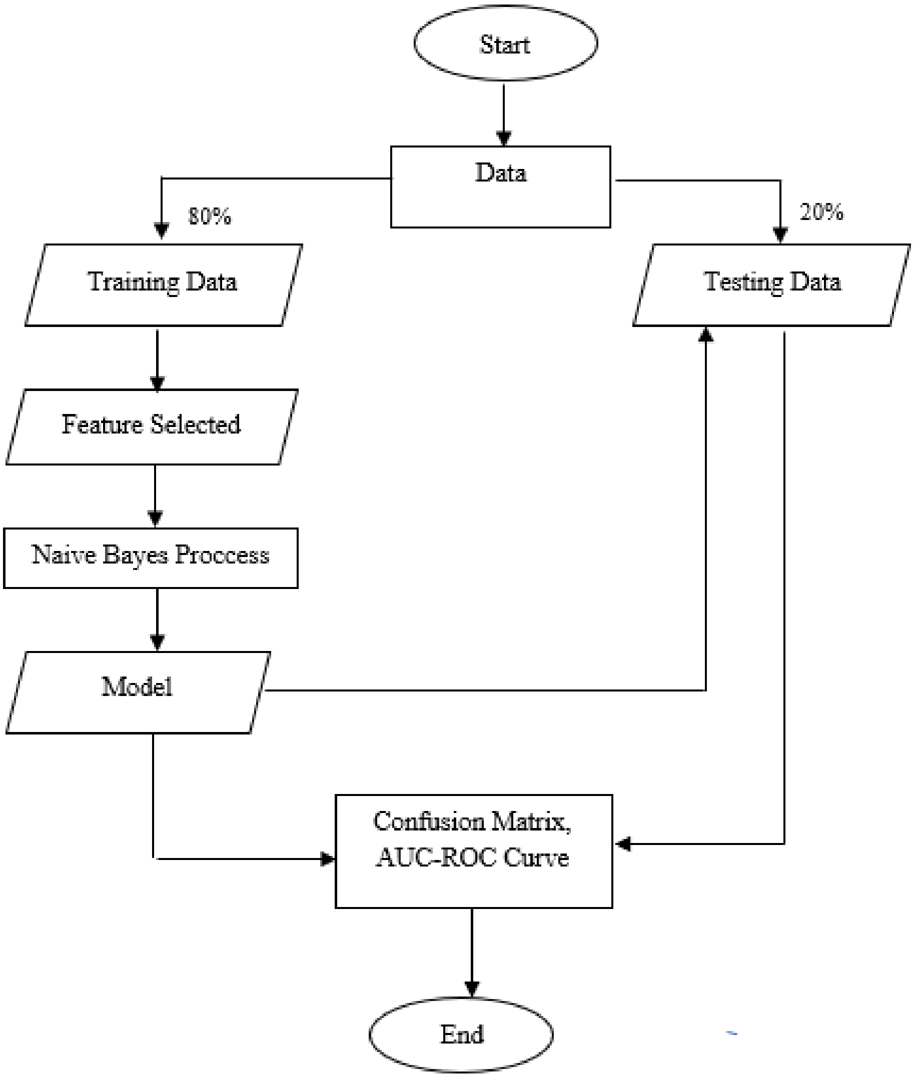


Figure 1. Flowchart of Naive Bayes Classification

The method used in this classification is Naive Bayes which aims to classify patients who have the potential to have or do not have heart disease.

Table 1. Characteristics of the Heart Disease Data Set

Variable	Description	Scale	Category
X1	Age		
X2	Sex	Nominal	0 = male 1 = female
X3	Chest pain type (cp)	Nominal	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic
X4	Resting blood pressure (trestbps)		
X5	Serum cholesterol (chol)		
X6	Fasting blood sugar > 120	Nominal	0 = false

Variable	Description	Scale	Category
	mg/dl (fbs)		1 = true
X7	Resting electrocardiographic results (restecg)	Nominal	0 = normal 1 = having ST-T wave abnormality 2 = showing probable
X8	Maximum heart rate achieved (thalach)		
X9	Exercise induced angina (exang)	Nominal	0 = no 1 = yes
X10	ST depression induced by exercise relative to rest (oldpeak)		
X11	The slope of the peak exercise ST segment (slope)	Nominal	1 = upsloping 2 = flat 3 = downsloping
X12	Number of major vessels (0–3) colored by fluoroscopy (ca)		
X13	Thal		3 = normal 6 = fixed defect 7 = reversable defect
Y	Target/diagnosis of heart disease	Nominal	0 = < 50% diameter narrowing 1 = > 50% diameter narrowing

3. Results and Discussion

3.1 Naive Bayes

Naive Bayes is one of the algorithm methods used in classification techniques in the field of Statistics [19]. This classification can predict the probability of a class which can be calculated based on the following Bayes theorem:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \tag{1}$$

where:
X : Unknown data
H : Hypothesis from class data
$P(H|X)$: Probability of the value hypothesis based on the condition of the value of X
$P(H)$: Hypothesis probability H value
$P(X|H)$: Probability of the value of X based on H value hypothesis
$P(X)$: Probability of value X

3.2. Multicollinearity and Normalization Test

If there is a linear relationship between each variable, it may indicate the existence of multicollinearity, and there are several ways to overcome multicollinearity, one of which is eliminating one of the variables that have a high correlation. The high correlation can be seen from the value of the correlation coefficient which is getting closer to the value 1. The normalization process in data is very important in producing classification outcomes, the use of normalization in classification produces better output results [20].

3.3. Receiver Operating Characteristics (ROC) and Area Under Curve (AUC)

ROC is a graph that describes the performance of a binary classification system between sensitivity on the Y-axis and 1-specificity on the X-axis. Sensitivity is a measure of the classification accuracy of an expected event, while specificity is a measure of the classification accuracy of an unexpected event [21]. The overall indication of the diagnostic accuracy of the ROC curve is the Area Under Curve (AUC) value. AUC values can be divided into several groups as follows [22]:

Table 2. Classification of data based on the AUC value

AUC	Classification
AUC>0.9	Outstanding classification
0.8≥AUC>0.7	Excellent classification
0.7≥AUC>0.6	Acceptable classification
0.6≥AUC>0.5	Poor classification
0,50	No classification

3.4. Descriptive Statistics

Based on [23] data on patients who have heart disease can be categorized based on age and gender, so from the exploration results obtained descriptive statistical results are as follows:

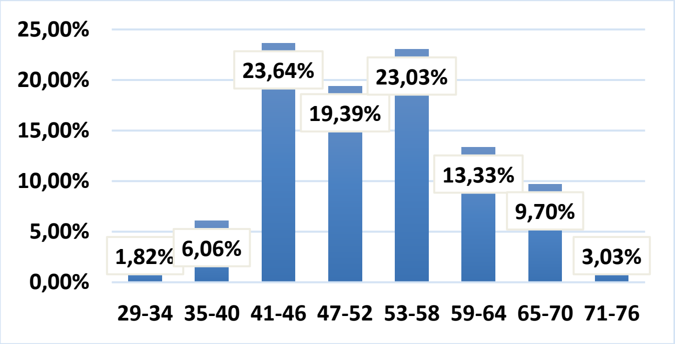


Figure 2. Heart disease data by age

Based on Figure 2, a person's vulnerable age is at risk of having heart disease, namely when they are over 40 years old. However, based on these data, the age under 40 years does not rule out the risk of heart disease even though it has a smaller percentage when compared to those over 40 years old.

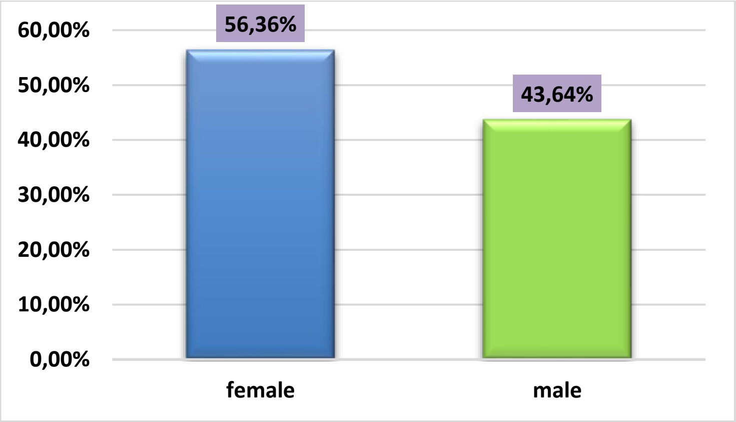
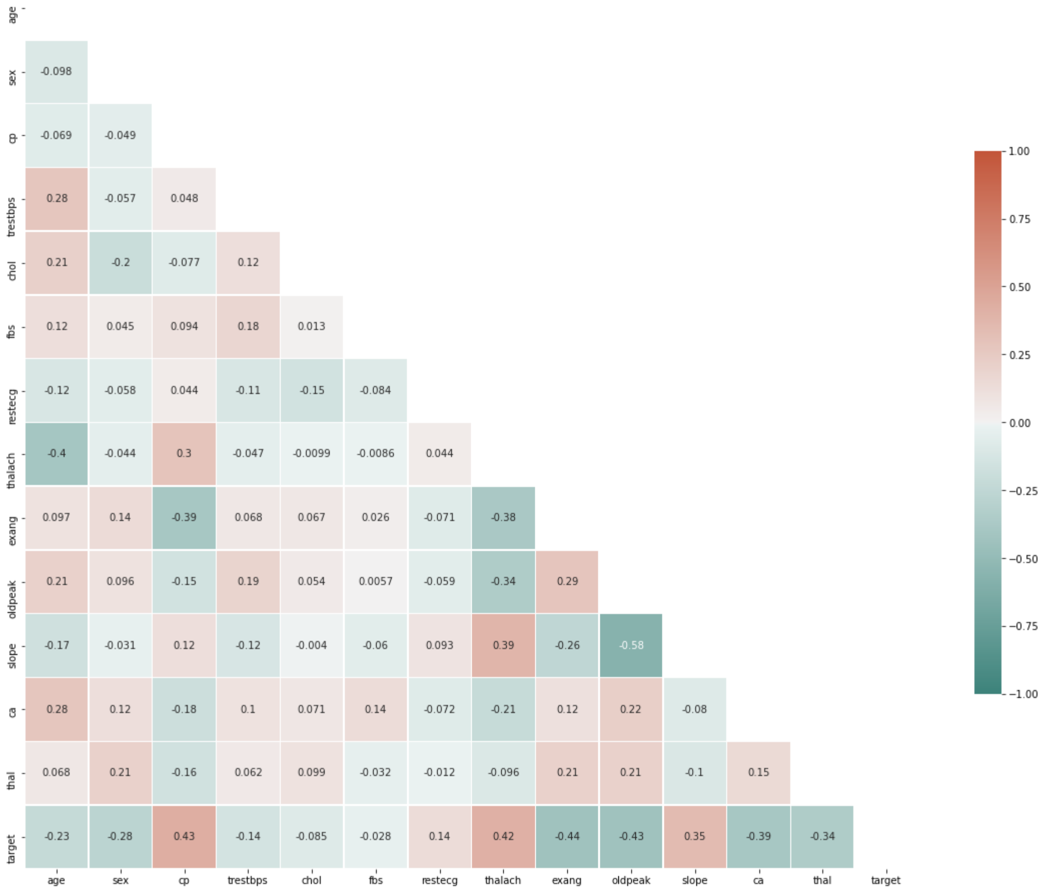


Figure 3. Heart disease data by gender

Based on the results of the percentage of patients who have heart disease in Figure 3, it is found that women patients have a greater percentage of 12.72% than men patients.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age														
sex	-0.098													
cp	-0.069	-0.049												
trestbps	0.28	-0.057	0.048											
chol	0.21	-0.2	-0.077	0.12										
fbs	0.12	0.045	0.094	0.18	0.013									
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084								
thalach	-0.4	-0.044	0.3	-0.047	-0.0099	-0.0086	0.044							
exang	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38						
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29					
slope	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58				
ca	0.28	0.12	-0.18	0.1	0.071	0.14	-0.072	-0.21	0.12	0.22	-0.08			
thal	0.068	0.21	-0.16	0.062	0.099	-0.032	-0.012	-0.096	0.21	0.21	-0.1	0.15		
target	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	

Figure 4. Correlation results between variables

Figure 4 shows that the old peak variable with slope has a negative correlation with the medium category, which is 0,8. So to overcome multicollinearity, the authors delete one of the two variables, namely by eliminating the slope variable in the next analysis.

Table 3. Training Data

Class	**Naive Bayes Result (%)**	
	0	1
0	83,81%	16,79%
1	16,19%	83,21%

Based on table 3, it was found that the percentage of data for patients classified as having no heart disease was 83,81% and patients classified as having heart disease was 83,21%. So it can be concluded that the Naive Bayes method is good for classification in training data.

Table 4. Testing Data

Class	**Naive Bayes Result (%)**	
	0	1
0	87,50%	16,22%
1	12,50%	83,78%

Based on table 4, it was found that the percentage of patients classified as having heart disease was 87,50% and the patients were classified as having heart disease at 83,78%. So, it can be concluded that the Naive Bayes method is good for classification in testing data.

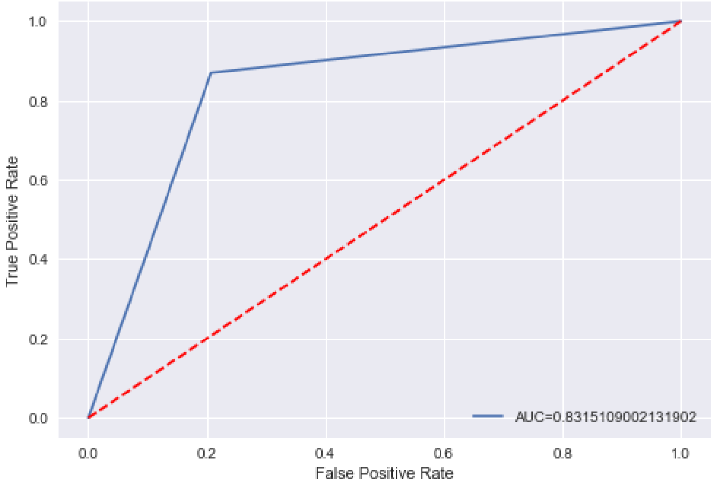


Figure 5. AUC of Training Data

Figure 5 shows that the AUC value is 83,15%, so it can be concluded that the diagnostic accuracy of the classification of patients having or not having heart disease in the training heart disease data can be classified well at 83,15%.

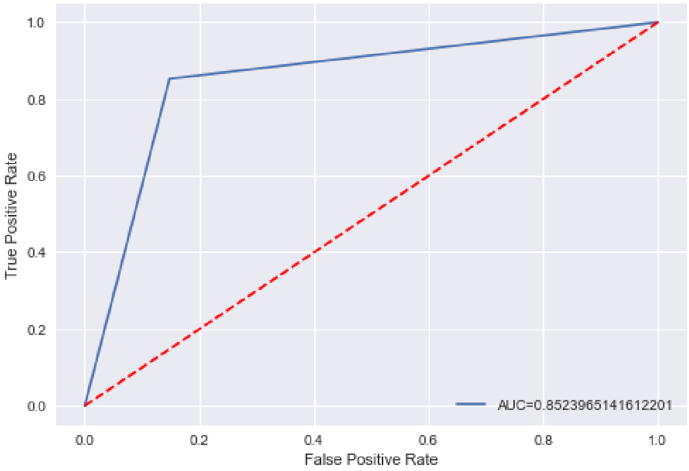


Figure 6. AUC of Testing Data

Figure 6 shows that the AUC value is 85,24%, so it can be concluded that the diagnostic accuracy of the classification of patients having or not having heart disease in the training heart diseases data can be classified as good at 85,24%.

The weakness of this research is that it has not explored the classification of heart diseases using several machine learning methods. Researchers will continue this research by comparing Naive Bayes with several machine learning methods such as Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN) and at the same time applying optimization in machine learning.

4. Conclusions

The results of the accuracy of the classification machine learning with Naive Bayes are good classification, with the results of the accuracy of the classification in the training data on patient data that is classified correctly as not having heart disease by 83,81% and the right patient being classified as having heart disease by 83,21%. In data testing, the percentage of patients classified as having heart disease was 87.50% and the patients were classified as having heart disease at 83,78%. Likewise, the AUC values in the training data and testing data are 83,15% and 85,24%, respectively.

References

[1] WHO, “Cardiovascular diseases (CVDs),” 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Jan. 21, 2022).

[2] A. Maharani, Sujarwoto, D. Praveen, D. Oceandy, G. Tampubolon, and A. Patel, “Cardiovascular disease risk factor prevalence and estimated 10-year cardiovascular risk scores in Indonesia: The SMARThealth Extend study,” *PLoS One*, vol. 14, no. 4, pp. 1–13, 2019, doi: 10.1371/journal.pone.0215219.

[3] M. D. Ritchey, H. K. Wall, M. G. George, and J. S. Wright, “US trends in premature heart disease mortality over the past 50 years: Where do we go from here?,” *Trends Cardiovasc. Med.*, vol. 30, no. 6, pp. 364–374, 2020, doi: 10.1016/j.tcm.2019.09.005.

[4] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” *Procedia Technol.*, vol. 10, pp. 85–94, 2013, doi: 10.1016/j.protcy.2013.12.340.

[5] F. Paquin, J. Rivnay, A. Salleo, N. Stingelin, and C. Silva, “Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic

semiconductors," *J. Mater. Chem. C*, vol. 3, pp. 10715–10722, 2015, doi: 10.1039/b000000x.

[6] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-72685-1.

[7] S. H. Hasanah, "CLASSIFICATION SUPPORT VECTOR MACHINE IN BREAST CANCER PATIENTS," vol. 16, no. 1, pp. 129–136, 2022, [Online]. Available: https://doi.org/10.30598/barekengvol16iss1pp129-136.

[8] S. H. Hasanah and E. Julianti, "Analysis of CART and Random Forest on Statistics Student Status at Universitas Terbuka," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 6, no. 1, pp. 56–65, 2022, doi: 10.29407/intensif.v6i1.16156.

[9] S. H. Hasanah and S. M. Permatasari, "Backpropagation Artificial Neural Network Classification Method in Statistics Students of Open University," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 14, no. 2, pp. 243–252, 2020.

[10] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Syst.*, vol. 192, no. xxxx, 2020, doi: 10.1016/j.knosys.2019.105361.

[11] A. P. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 7, no. 2, p. 91, 2019, doi: 10.3991/ijes.v7i2.10659.

[12] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive Bayes classification algorithm for traffic risk management," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 1, 2021, doi: 10.1186/s13634-021-00742-6.

[13] A. W. Syaputri, E. Irwandi, and M. Mustakim, "Naïve Bayes Algorithm for Classification of Student Major's Specialization," *J. Intell. Comput. Heal. Informatics*, vol. 1, no. 1, p. 17, 2020, doi: 10.26714/jichi.v1i1.5570.

[14] A. Akbar Ritonga, Ibnu Rasyid Munthe, Masrizal, "Jurnal Mantik Jurnal Mantik," *Mobile-Based Natl. Univ. Online Libr. Appl. Des.*, vol. 3, no. 2, pp. 10–19, 2019, [Online]. Available: http://iocscience.org/ejournal/index.php/mantik/article/view/882/595.

[15] A. Krichene, "Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank," *J. Econ. Financ. Adm. Sci.*, vol. 22, no. 42, pp. 3–24, 2017, doi: 10.1108/JEFAS-02-2017-0039.

[16] R. Ardianto, T. Rivanie, Y. Alkhalifi, F. S. Nugraha, and W. Gata, "Sentiment Analysis on E-Sports for Education Curriculum Using Naive Bayes and Support Vector Machine," *(Journal Comput. Sci. Inf.*, vol. 13, no. 2, pp. 109–122, 2020, doi: http://dx:doi:org/10:21609/jiki:v13i2.885.

[17] UCI Machine Learning, "Heart Disease Data Set." archive.ics.uci.edu (accessed Sep. 08, 2021).

[18] N. Boyko and I. Dosiak, "Analysis of machine learning algorithms for classification and prediction of heart disease," *CEUR Workshop Proc.*, vol. 3038, pp. 233–249, 2021.

[19] M. Ismail, N. Hassan, and S. S. Bafjaish, "Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Task," *J. Soft Comput. Data Min.*, vol. 1, no. 2, pp. 1–10, 2020, doi: 10.30880/jscdm.2020.01.02.001.

[20] S. Hasanah and S. Permatasari, "Metode Klasifikasi Jaringan Syaraf Tiruan Backpropagation Pada Mahasiswa Statistika Universitas Terbuka," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 14, no. 2, pp. 243–252, 2020, doi: 10.30598/barekengvol14iss2pp249-258.

[21] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Casp. J. Intern. Med.*, vol. 4, no. 2, pp. 627–635, 2013.

[22] S. Yang and G. Berdine, "The receiver operating characteristic (ROC) curve," *Southwest Respir. Crit. Care Chronicles*, vol. 5, no. 19, p. 34, 2017, doi: 10.12746/swrccc.v5i19.391.

[23] F. Babic, J. Olejar, Z. Vantova, and J. Paralic, "Predictive and descriptive analysis for heart disease diagnosis," *Proc. 2017 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2017*, no. October, pp. 155–163, 2017, doi: 10.15439/2017F219.