10_JURNAL_S_Coronary Artery Disease Prediction Using Decision Trees and Multinomial

By Yulia Resti

WORD COUNT 5419
TIME SUBMITTED 27-APR-2023 10:47AM
PAPER ID 99138689

Inovasi Matematika (Inomatika)
Vol.3 No.2, Juli 2021, hal.172 –187
DOI: 10.35438/inomatika. v3i2.266
https://inomatika.unmuhbabel.ac.id/index.php

Coronary Artery Disease Prediction Using Decision Trees and Multinomial Naïve Bayes with k-Fold Cross Validation

Endang S Kresnawati[1], Yulia Resti [2*], Bambang Suprihatin[3], M. Rendy Kurniawan[4], Widya Ayu Amanda[5]

[1,2,3,4,5]Jurusan Matematika, FMIPA, Universitas Sriwijaya, Indonesia
*yulia_resti@mipa.unsri.ac.id

Abstrak

Penyakit arteri koroner (*coronary artery disease*) menjadi penyebab utama kematian penduduk di dunia setidaknya selama dua dekade (2000-2019) dan mengalami peningkatan kematian terbesar dalam rentang waktu tersebut dibandingkan dengan penyebab kematian lainnya. Keberhasilan memprediksi penyakit arteri koroner secara dini berdasarkan data medis bermanfaat bagi pasien dan juga bagi kestabilan perekonomian negara. Tujuan penelitian ini adalah memprediksi penyakit arteri koroner jantung dengan mengimplementasikan dua metode *statistical learning* yaitu Multinomial Naïve Bayes dan pohon keputusan dengan validasi silang 10-fold, dimana variabel-variabel numerik didiskritisasi untuk memperoleh variabel-variabel kategorik. Hasil penelitian menunjukkan bahwa metode Pohon Keputusan memiliki kinerja yang lebih baik dibandingkan metode Multinomial Naïve Bayes dalam memprediksi penyakit arteri koroner. Ukuran kinerja metode Pohon Keputusan memperoleh tingkat akurasi 99,63 %, sensitivitas 100 %, spesifisitas 99,33%, presisi 99,23 %, dan nilai prediksi negatif (NPV) 100 %. Ukuran-ukuran ini mengindikasikan bahwa metode Pohon Keputusan layak digunakan untuk memprediksi penyakit arteri coroner, termasuk data independent berupa data penyakit arteri coroner lainnya dengan variable predictor yang sama. Hasil penelitian ini juga menunjukkan bahwa perbedaan rujukan dengan penelitian-penelitian sebelumnya dalam mendiskritisasi variabel numerik mampu meningkatkan kinerja metode dalam memprediksi penyakit arteri coroner.

Kata kunci: Penyakit Arteri Koroner, Kinerja, Prediksi.

Abstract

Coronary artery disease has been the leading cause of death in the world population for at least two decades (2000-2019) and has experienced the largest increase in mortality in that time span compared to other causes of death. The success of predicting coronary artery disease early based on medical data is not only beneficial for patients, but also beneficial for the stability of the country's economy. This paper discusses the prediction of coronary artery disease risk by implementing two statistical learning methods, namely Multinomial Naïve Bayes and Decision Tree with 10-fold cross validation, where numerical variables are discretized to obtain categorical variables. The results showed that the Decision Tree method has better performance than the Multinomial Naïve Bayes method in predicting coronary artery disease. The performance measure of the Decision Tree method obtained an accuracy rate of 99.63%, 100% sensitivity, 99.33% specificity, 99.23% precision, and 100% Negative Prediction Value. These measures indicate that the Decision Tree method is appropriate for predicting coronary artery disease, including independent data (other coronary artery disease data with the same predictor variables). The results of this study

also show that the different references to previous studies in discretizing numerical variables can improve the performance of the method in predicting coronary artery disease.
Keywords: Coronary Artery Disease, Performance, Prediction.

Received: Mei 3, 2021/ Accepted: Juli 28, 2021/ Published Online: Juli 29, 2021

PENDAHULUAN

Coronary artery disease (CAD) or also called heart disease (heart disease) occurs due to decreased blood flow to the heart muscle due to plaque buildup (atherosclerosis) in the heart arteries (Mendis et al., 2015). Another name for this disease is coronary heart disease or ischemic heart disease (Bhatia, 2010). In some literature, this disease is also called cardiovascular (Purushottam et al., 2016; Chowdary et al., 2020). This disease has been the leading cause of death in the world's population for at least two decades (2000-2019) and has experienced the largest increase in deaths in that time span compared to other causes of death. In high-income countries, coronary artery disease has long been a major contributor to the overall disease burden, in addition to stroke and cancer. The burden of this disease is also increasing in middle-income countries, and also in low-income countries. The success of early detection of coronary artery disease based on medical data is not only beneficial for patients but also beneficial for economic stability (WHO, 2019).

Purushottam et al., (2016) predicted coronary artery disease using the same dataset, but they filled in the missing data using the AllPossible-MV algorithm (Alcalá-Fdez, et al., 2009; Alcalá-Fdez, et al., 2011). They proposed several machine learning methods, namely Support Vector Machine (SVM), Decision Tree C4.5 Algorithm, Neural Network (NN), PART, Multiple Layer Perceptron (MLP), Radial Basis Function (RBF), TSEAFS, and Efficient System. The highest level of accuracy achieved using the 10-fold cross-validation model was 86.3% using the Efficient System method, followed by the RBF method (78.53%), TSEAFS (77.45%), NN (76.47%), Algorithm C4.5 Decision Tree (73.53%), PART (73.53%), and SVM (70.59%).

Chowdary et al. (2020) also predict coronary artery disease using the same dataset, but they change some of the categorical type data to numeric type. The machine learning methods they implement are quite a lot, namely Logistic Regression, Random Forest, Decision Tree, Gaussian Naïve Bayes, Binomial Nave Bayes, Multinomial Naïve Bayes, K-Nearest Neighbor, Artificial Neural Network, and Voting of Logistic Regression and K-Nearest Neighbor (VLRAKN). Their funding shows that the VLRAKN method has the highest level of accuracy at 89%. The accuracy that has been achieved using the split system validation model is 67% as

training data and 33% as test data. The other methods have an accuracy rate of between 80%-88%. This work also calculates the performance of prediction methods based on sensitivity, specificity, precision, and F-Measures, where the VLRAKN method is the method that has the highest performance measure on all of these measures.

Multinomial Nave Bayes and Decision Trees are two of the most popular and easy to understand classification methods. The Multinomial Naïve Bayes method uses Bayes' theorem in determining its decisions, where each predictor variable must be categorical following a multinomial distribution if there are more than two categories, and a binomial distribution if there are only two categories (Chen & Fu, 2018). The Decision Tree method uses a tree structure representation where each node describes the variable, the branch describes the value of the variable, and the leaf describes the class. Decision Trees have a fairly high level of accuracy in various cases (Santoso, 2012).

This study discusses risk prediction for coronary artery disease, which can also be called early detection of heart disease, by implementing two statistical learning methods, namely Multinomial Naïve Bayes and Decision Trees with 10-fold cross-validation as a model validation technique. The novelty in this study is a technique for categorizing five numerical variables in research data, namely age (years), cholesterol levels (mg/dl), fasting blood sugar levels (mg/dl), maximum heart rate (bpm), and old peak (mV) conducted with different criteria from Purushottam et al., (2016), as well as David and Belcy, (2018) and Riani et al., (2019). The categorization of the five numerical variables is based on valid references that specifically discuss these numerical variables. In addition, in this study, the missing data was not included in the data processing because the majority of the data was incomplete. The performance of the two statistical learning methods is then measured based on the level of accuracy, sensitivity, specificity, precision, and negative predictive value (NPV). This performance measure is very important in practice, because it guides the choice of learning method or model, and provides a measure of the quality of the method or model that is finally selected, including for independent data (Hastie et al., 2009).

METODE

The steps in this study are presented in Figure 1. The research data is Heart Disease data from the Cleveland Clinic Foundation, which was donated as public data to the Center for Machine Learning and Intelligent Systems (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). The data consists of a target variable (dependent) and a predictor variable (independent). The target variable is the health status of patients related to heart

disease, which consists of two categories, namely patients who have heart disease and patients who do not have heart disease. The predictor variables consisted of two personal data variables in the form of age and gender and 11 data variables from medical examination results.

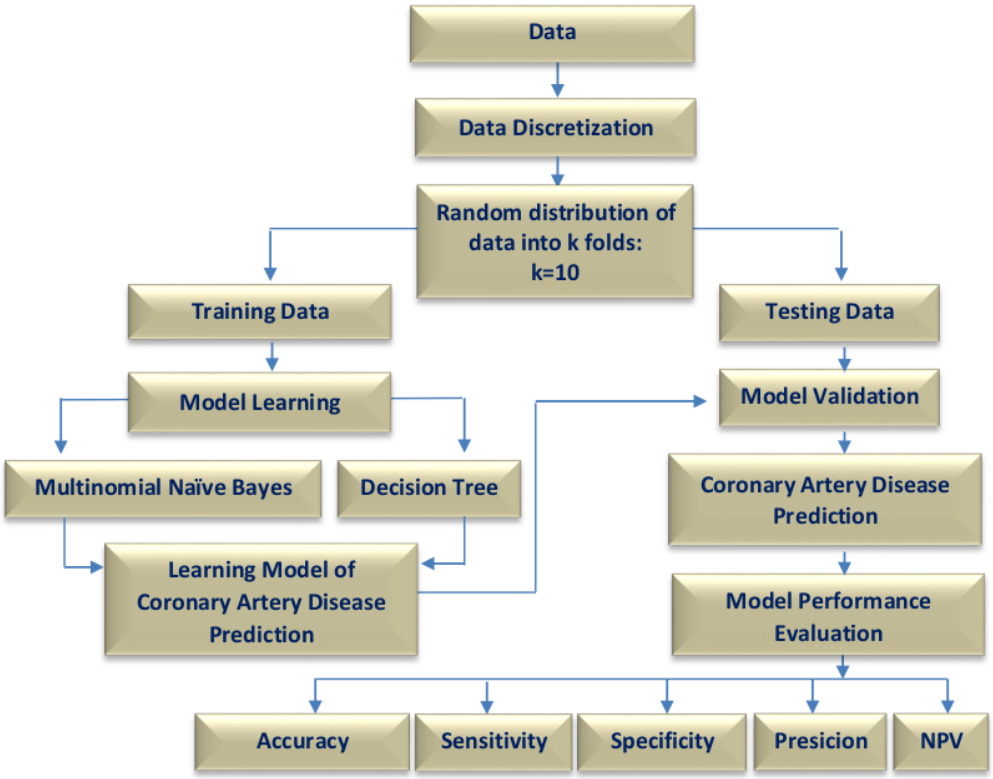

Figure 1. Research Methodology

The dependent variable is denoted as $Y_j, j =$ no, yes, where "no" represents patients who do not have heart disease and "yes" represents patients who have heart disease. The thirteen independent variables are each denoted as $X_i, i = 1,2,3,\cdots,13$ namely age (X_1), sex (X_2), types of chest pain (X_3), blood pressure at rest (X_4), cholesterol level (X_5), fasting blood sugar level > 120 mg/dl (X_6), ECG results at rest (X_7), maximum heart rate (X_8), exercise causes angina-type chest pain (X_9), oldpeak or ST segment obtained from exercise relative to rest (X_{10}), ST segment slope (X_{11}), number of major pulses stained by fluoroscopy (X_{12}) and thalassemia (X_{13}). Of the thirteen explanatory variables, there are 5 numerical variables and 7 categorical variables. Both the Multinomial Naïve Bayes and Decision Tree methods require that all variables be categorical types, so that the five numeric variables are categorized first. The

complete details of the response variables and the thirteen explanatory variables are presented in Table 1.

Tabel 1. Data Penelitian

Variable	Data Type	Information
X_1	Numeric	29-77 years
X_2	Categorical	male; female
X_3	Categorical	typical angina; atypical angina; non-anginal pain; asymptomatic
X_4	Numeric	94-200 mm/Hg
X_5	Numeric	126-564 mg/dl
X_6	Categorical	false; true
X_7	Categorical	normal; have ST-T wave abnormalities; demonstrate left ventricular hypertrophy according to estes criteria
X_8	Numeric	71- 202 bpm
X_9	Categorical	No; Yes
X_{10}	Numeric	0 – 6,2 mV
X_{11}	Categorical	leaning up; flat; slightly sloping
X_{12}	Categorical	0; 1; 2; 3
X_{13}	Categorical	three (normal); six (permanent disability); seven (temporary disability)
Y	Categorical	no; yes

In the process of predicting, the data is divided into two parts, namely training data and test data. The training data is used to build a prediction/classification learning model, while the test data is used to validate the previously built model. The model validation method using a cross-validation technique with many folds was chosen in this study because it has a small bias (Rodrı´guez et al, 2010). The division refers to (Burger, 2018) as presented in Figure 2.

test	train	train	train	train	train	train	train	train	train
train	test	train	train	train	train	train	train	train	train
train	train	test	train	train	train	train	train	train	train
train	train	train	test	train	train	train	train	train	train
train	train	train	train	test	train	train	train	train	train
train	train	train	train	train	test	train	train	train	train
train	train	train	train	train	train	test	train	train	train
train	train	train	train	train	train	train	test	train	train
train	train	train	train	train	train	train	train	test	train
train	train	train	train	train	train	train	train	train	test

Figure 2. Splitting Training and Test Data for 10-Fold Cross Validation

The statistical learning methods used to build a predictive model of heart disease status in this work are the Multinomial Naïve Bayes method (Pan et al, 2018) and the Decision Tree (Han et al., 2012). These two methods are often successful in carrying out prediction/classification tasks with a high degree of accuracy (Retnasari and Rahmawati, 2017).

The Multinomial Naïve Bayes method works based on the Bayes theorem, which determines the maximum posterior probability of each observation obtained as the product of the prior probability and the likelihood probability. Let $P(X_k|Y_{yes})$ and $P(X_k|Y_{no})$ Let A and B are the likelihood probabilities of the occurrence/diagnosis of cardiac arrest and no, respectively, which are written as,

$$P(X_k|Y_{yes}) = \frac{\sum_c^m n_c(X_k|Y_{yes}) + 1}{n(X_k|Y_{yes}) + m} \quad (1)$$

$$P(X_k|Y_{no}) = \frac{\sum_c^m n_c(X_k|Y_{no}) + 1}{n(X_k|Y_{no}) + m} \quad (2)$$

Posterior probability for Y_j, j is,

$$P(Y_j|X_1, \cdots, X_d) = \arg\max P(Y_j) \prod_{k=1}^{d} P(X_k|Y_j) \quad (3)$$

where the prior probability of each group is defined as,

$$P(Y_{yes}) = \frac{\sum_{k=1}^{d} n(X_k|Y_{yes}) + 1}{n + g} \quad (4)$$

$$P(Y_{no}) = \frac{\sum_{k=1}^{d} n(X_k|Y_{no}) + 1}{n + g} \quad (5)$$

For the group of patients diagnosed with heart disease, $n_c(X_k|Y_{yes})$ is the number of patients diagnosed with heart disease in variable X_k with category c, $n(X_k|Y_{yes})$ is the number of patients diagnosed with heart disease in variable X_k, $n(Y_{yes})$ is the number of patients diagnosed with heart disease, m is the number of categories in the variable X_k, and g is the total number of groups in the study.

The Decision Tree is a classification method that has a tree structure such as a flow chart (Figure 3), where each internal node shows a test on a variable, each branch shows the results of the test, and a leaf node shows the results of the test node (classes), while the topmost node is called the root node. The concept of a decision tree is to partition data based on the highest

gain value so that a decision tree is formed which is then used to form decision rules using IF-THEN logic (Han et al., 2012).

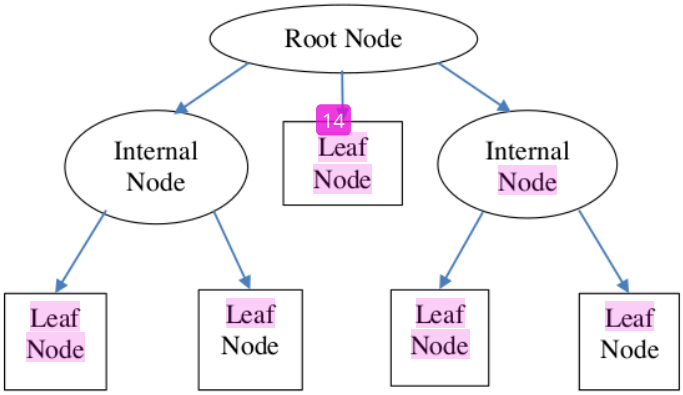

Figure 3. Decision Tree Model

The main steps in constructing a decision tree are: first, selecting a variable as the root; second, loading the branch for each value; third, dividing each branch into classes; and fourth, repeating the process for each branch until all cases in each branch have the same class. The basis for choosing a variable as the root is the highest information gain value of all variables. Before getting the highest gain value, first calculate the entropy value of all values in the variable. Entropy acts as a parameter to measure the variance of the sample data. After the entropy value in the sample data is known, the most influential variable will be a measure of classifying the data. This measure is referred to as information gain.

Entropy and Information Gain are obtained using (6) – (8), respectively.

$$Entropy\ (S) = \sum_{i=1}^{k_s} -P_i\ log_2 P_i \tag{6}$$

$$Entropy\ (S_c) = \sum_{c=1}^{k_X} -P_c\ log_2 P_c \tag{7}$$

$$Information\ Gain(S,X) = Entropy\ (S) - \sum_{c=1}^{k_X} \frac{|S_c|}{|S|} Entropy\ (S_c) \tag{8}$$

where S, k_s, S_c k_X, P_i, and P_c respectively as the total number of patients, the number of patient groups, the total number of patients in the c-th category of the predictor variable X, the number of categories in the variable X, the prior probability in the i-th group of the predictor variable X, and the prior probability in the c-th category of the predictor variable X.

Furthermore, after the prediction results of heart disease status using the Multinomial Naïve Bayes and Decision Tree methods were obtained, the performance of the two methods was evaluated. Regarding medical data, the performance of prediction results can be evaluated using the level of accuracy, sensitivity, specificity, precision, and negative predictive value (NPV) (Maniruzzaman et al., 2017) as shown in equation (9) – (13) based on the confusion matrix as in Table 2 (Gathak, 2017); (Burgers, 2018).

Table 2. Confusion Matrix

Predict	Actual	
	Yes	**No**
Yes	True Positive (TP)	False Positive (FP)
No	False Negative (FN)	True Negative (TP)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (9)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (10)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (11)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (12)$$

$$\text{Negative Prediction Value (NPV)} = \frac{\text{TN}}{\text{FN} + \text{TN}} \quad (13)$$

RESULT

This research begins by discretizing predictor variable data of numeric type based on references as presented in Table 3. Next, divide the data into 10-fold randomly and then separate them as training data and test data as illustrated in Figure 1. Table 4 presents the composition of each training (learning) data and test data for model development and validation.

Table 3. Numerical Variable Discretization

Variabel	Diskritisasi	Sumber
X_1	< 40 years; 40-64 years; ≥ 65 years	WHO, 2019 Woodward et al., 2012
X_4	90-119 mm hg (normal); 120-139 mm hg (pra-hipertentioni); ≥140 mm hg (hipertention)	Borghi et al. (2003)
X_5	< 200 (normal); 200-239 (high limit); ≥ 240 (high)	Third Report of the National Cholesterol Education Program (NCEP), 2001
X_8	≤100 (normal); > 100 (takikarbi)	Palatini, 1999
X_{10}	< 3,2 (no); ≥ 3,2 (yes)	Riani et al., 2019

Table 4. Composition of Training and Test Data

Test Data (One-fold)										
Grup	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Yes	15	13	16	12	13	8	10	12	12	9
No	12	14	11	15	14	19	17	15	15	18
Total	27	27	27	27	27	27	27	27	27	27
Learning Data (Nine-fold)										
Grup	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Yes	105	107	104	108	107	112	110	108	108	111
No	138	136	139	135	136	131	133	135	135	132
Total	243	243	243	243	243	243	243	243	243	243

Table 5 presents a learning model using the Multinomial Naive Bayes method for the first learning data with observations of 105 patients who have heart disease and 108 who do not have heart disease (healthy). The learning model using the Decision Tree method for the first learning data with the same observations as using the Multinomial Naive Bayes method is presented in Figure 4.

Table 5. The First Learning Model using Multinomial Naive Bayes

Y	$P(Y)$	$P(X_1\|Y)$			…	$P(X_{13}\|Y)$		
		< 40 years	40-64 years	≥ 65 years		normal	permanent disability	temporary disability
Jantung	0.43	0.06	0.80	0.14	…	0.78	0.05	0.17
Tidak	0.57	0.05	0.80	0.15	…	0.31	0.07	0.62

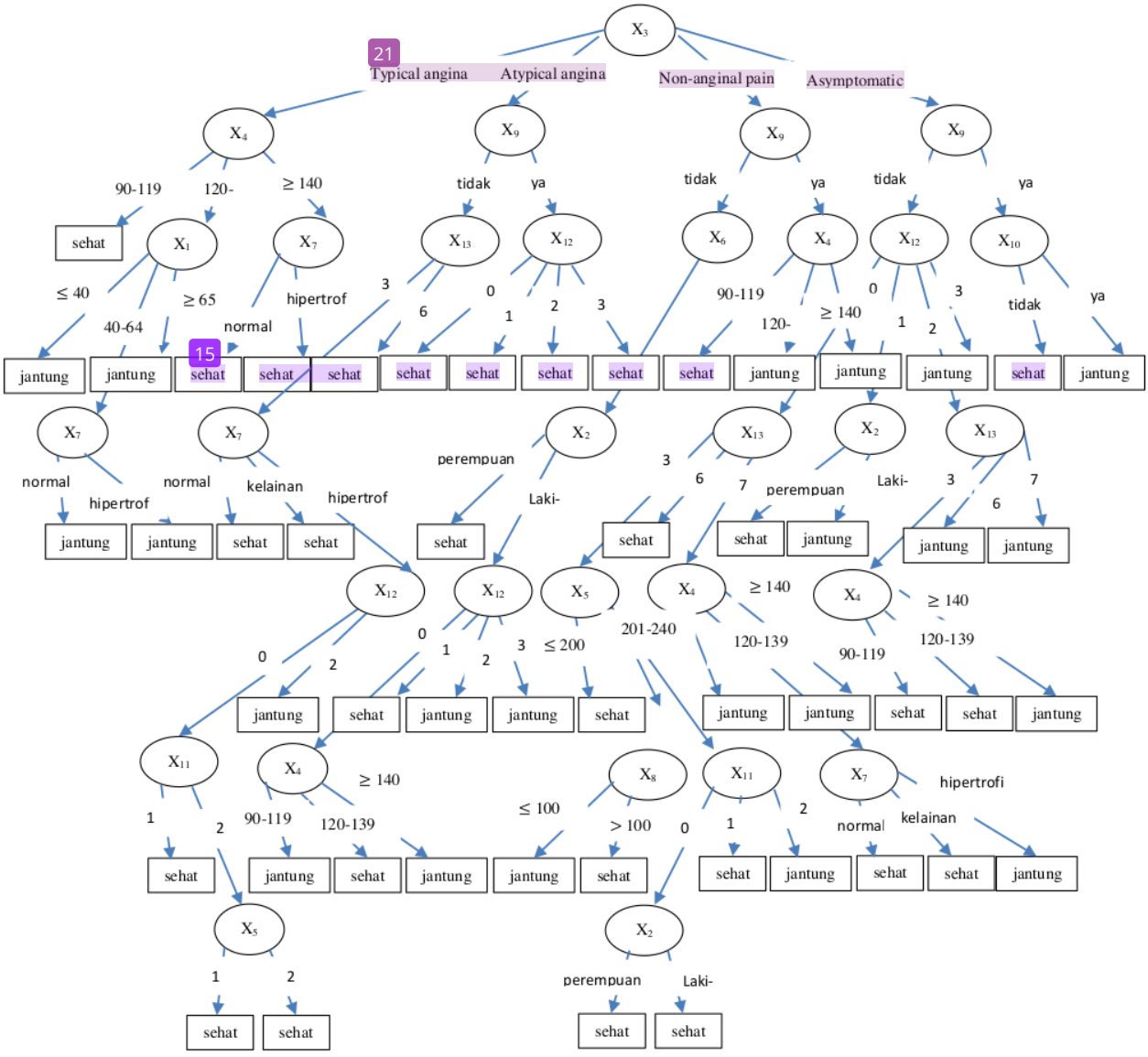

Figure 4. Decision Tree method learning model for the first learning data

Furthermore, Table 6 presents the prediction results for fold 1 to fold 10 using the Multinomial Naive Bayes method based on each learning model. Prediction results using the Decision Tree method for fold 1 to fold 10 based on each learning model are presented in Table 7.

Table 6. Prediction of Heart Disease Status using Multinomial Naïve Bayes

Testing	Accuracy	Sensitivity	Specificity	Precision	NPV
Fold 1	88.89	93.33	83.33	87.50	90.91
Fold 2	92.59	92.31	92.86	92.31	92.86
Fold 3	92.59	87.50	100.00	100.00	84.62
Fold 4	85.19	75.00	93.33	90.00	82.35
Fold 5	81.48	69.23	92.86	90.00	76.47
Fold 6	85.19	50.00	100.00	100.00	82.61
Fold 7	100.00	100.00	100.00	100.00	100.00
Fold 8	92.59	83.33	100.00	100.00	88.24
Fold 9	100.00	100.00	100.00	100.00	100.00
Fold 10	100.00	100.00	100.00	100.00	100.00
Mean	91.85	85.07	96.24	95.98	89.80
Standard Deviation	6.72	16.26	5.60	5.31	8.41

Prediction results using the Decision Tree method for fold 1 to fold 10 based on each learning model are presented in Table 7.

Table 7. Prediction of Heart Disease Status using Decision Tree Method

Testing	Accuracy	Sensitivity	Specificity	Precision	NPN
Fold 1	100.00	100.00	100.00	100.00	100.00
Fold 2	100.00	100.00	100.00	100.00	100.00
Fold 3	100.00	100.00	100.00	100.00	100.00
Fold 4	100.00	100.00	100.00	100.00	100.00
Fold 5	100.00	100.00	100.00	100.00	100.00
Fold 6	100.00	100.00	100.00	100.00	100.00
Fold 7	100.00	100.00	100.00	100.00	100.00
Fold 8	100.00	100.00	100.00	100.00	100.00
Fold 9	96.30	100.00	93.33	92.31	100.00
Fold 10	100.00	100.00	100.00	100.00	100.00
Mean	99.63	100.00	99.33	99.23	100.00
Standard Deviation	1.17	0.00	2.11	2.43	0.00

DISCUSSION

Prediction of coronary artery disease has been carried out using many methods. This study proposes two statistical learning methods to predict coronary artery disease, namely Multinomial Naïve Bayes and Decision Trees. Both of these methods require that all variables be categorical type so that numerical variables in the research data are discretized first to obtain categorical type variables. The results of the discretization of the five numerical variables presented in Table 3 show that each age variable (X_1), blood pressure at rest (X_4), and cholesterol levels (X_5) has three categories, while each variable maximum heart rate (X_8) and oldpeak or ST segment obtained from exercise relative to rest (X_{10}) has two categories. The results of this discretization are different from those in Purushottam et al., (2016), David and Belcy, (2018), Riani et al., (2019) (only the variableX_{10} is the same), as well as Chowdary et al. (2020). This study also did not involve missing data like Purushottam et al., (2016).

Randomly dividing the data into 10-folds and then separating them as training data for model learning and test data to validate the model as presented in Table 4 shows that each fold has the same size, both training data, which contains nine folds, and test data. which contains one-fold. However, the size of the data on being diagnosed with heart disease and not having heart disease is not exactly the same. In the 1st and 3rd fold data, the size of the data diagnosed as having heart disease is larger than the data size diagnosed as not having heart disease, while in other folds it is the opposite.

Model learning using the Multinomial Naïve Bayes method shows that each variable has a different probability (likelihood) in each category as presented in Table 5 for the 1st learning data. The same thing happened to the other nine learning data. In the learning model using the Decision Tree method as shown in Figure 4, the variable that becomes the root node is the type of chest pain variable (X_3) which has four categories where each of the four categories has a size different internal nodes and leaves. The typical angina category has the smallest internal node and leaf size compared to the atypical angina, non-anginal pain, and asymptomatic categories, while the asymptomatic category has the largest size.

17 The performance measures of the two methods as presented in Table 6 and Table 7 show that the average of the five performance measures and the standard deviation of the 10-fold test data as validated using the Decision Tree method is higher than the Multinomial Naïve Bayes method. In the Decision Tree method, the five performance measures at nine-fold are all 100%. In the 9th fold, not all of the performance measures are 100%, so the average of the five performance measures is 99.63% accuracy, 100% sensitivity, 99.33% specificity, 99.23%

precision, and 100% NPV. In the Multinomial Naïve Bayes method, only fold 7, fold 9, and fold 10 have five performance measures that are all 100%. The average performance measures of the five measures for fold 1 to fold 10 in a row, namely accuracy, sensitivity, specificity, precision, and NPN are 91.85%, 85, 07%, 96.24%, 95.98%, and 89.90%, while the standard deviations are 6.72%, 16.26%, 5.6%, 5.31%, and 8.41%, respectively. Overall, both the mean and standard deviation of all folds in this study indicate that the Multinomial Naïve Bayes method is not better than the Decision Tree method in predicting coronary artery disease. The results obtained in this study are also better than those of Purushottam et al. (2016), who obtained the highest level of accuracy of 86.3% with the Efficient System method, one of the several methods it uses. The results of this study are also better than David and Belcy (2018), which obtained a precision of 81% with the Random Forest method. Likewise, when compared with Chowdary et al. (2020) who obtained accuracy, sensitivity, specificity, and precision of 89%, 86%, 91%, and 91.6%, respectively, in predicting coronary artery disease using the VLRNAK ensemble method. Several other studies using the same dataset as Normawati and Winiarti, (2017), Retnasari and Rahmawati, (2017), Indrajani et al., (2018), Aini et al., (2018), Aulia, (2018), Riani et al., (2019), and Pangaribuan et al., (2019) have accuracy that is not better than our work.

CONCLUSION

This study succeeded in predicting the risk of coronary artery disease using two statistical learning methods, namely Multinomial Naïve Bayes and Decision Trees. Numerical variables in the research data were discretized to obtain categorical variables by referring to valid sources. The learning model validation technique used is 10-fold cross validation. The results showed that the performance measures of the Decision Tree method were more consistent, higher, and had a relatively smaller standard deviation than the Multinomial Naïve Bayes method. These results indicate that the performance of the Decision Tree method is better than the Multinomial Naïve Bayes method in predicting coronary artery disease. The results of this study also indicate that differences in reference in discretizing numerical variables can affect the performance of the method in predicting the risk of coronary artery disease.

REFERENSI

Aini, S. H. A., Sari, Y. A., dan Arwan, Achmad. (2018). Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* Vol. 2, No. 9.

Alcalá-Fdez, J., Sánchez, L., García, M.J. del Jesus, S., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernández, J.C., Herrera, F. (2009). KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. *Soft Computing* 307-318

Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F. (2011). KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing* 17:2-3 (2011) 255-287.

Aulia, W. (2018). Sistem Pakar Diagnosa Penyakit Jantung Koroner Dengan Metode Probabilistic Fuzzy Decision Tree. *Jurnal Sains dan Informatika*. 4(12):106-117.

Bhatia, Sujata K. (2010). *Biomaterials for clinical applications* (Online-Ausg. ed.). New York: Springer. p. 23. ISBN 9781441969200. Archived from the original on 10 January 2017.

Borghi, C., Dormi, A., L'Italien, G., Lapuerta, P., Franklin, S.S., Collatina, S., Gaddi, A. (2003). The Relationship Between Systolic Blood Pressure and Cardiovascular Risk-Results of the Brisighella Heart Study. *The Journal of Clinical Hypertension*, Vol. V, No. 1, January/February.

Burger, S. V. (2018). *Introduction to Machine Learning with R: Rigorous Mathematical Analysis*. Oreilly.

Chen, H., Fu, D. (2018). An Improved Naïve Bayes Classifier for Large Scale Text. *Advances in Intelligent Systems Research*, volume 146, pp.33-36.

Chowdary, G., J., Suganya, G., Premalatha, M. (2020). Effective Prediction of Cardiovascular Disease Using Cluster of Machine Learning Algorithms. *Journal of Critical Reviews*, Vol.7 (18), 2192 – 2201.

David, H. B. F., Belcy, S. A. (2018). Heart Disease Prediction using Data Mining Techniques. *ICTACT Journal on Soft Computing* 9 (1), 1817 - 1823, October.

Gathak, A. (2017). *Machine Learning with R*. Springer.

Han, J., Kamber, M., dan Pei, J. (2012). *Data Mining: Concept and Techniques*, Third Edition. Waltham: Morgan Kaufmann.

Hastie, T., Tibshirani, R., Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. California: Springer.

Indrajani, Bahana, R., Kosala., R, Haryadi, Y. (2018). Aplikasi Informasi Kesehatan dan Diagnosa Penyakit Jantung Berbasis Android, *Seminar Nasional Teknologi Informasi, Komunikasi dan Industri* (SNTIKI-10), November.

Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H.S., El-Baz, A. S., Suri, J.S. (2017). Comparative Approaches for Classification of Diabetes Mellitus Data: Machine Learning Paradigm. *Computer Methods and Programs in Biomedicine*, vol. 152, pp. 23–34, 2017, doi: 10.1016/j.cmpb.2017.09.004.

Mendis, S., Puska, P., Norrving, B. (2015). *Global atlas on cardiovascular disease prevention and control*, 1st ed. Geneva: World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. pp. 3–18. ISBN 9789241564373.

Normawati, D., dan Winiarti, S. (2017). Seleksi Fitur Menggunakan Penambangan Data Berbasis Variable Precision Rough Set (VPRS) Untuk Diagnosis Penyakit Jantung Koroner. *Jurnal Ilmu Teknik Elektro Komputer dan Informatika* (JITEKI) Vol. 3, No. 2.

Palatini, P. (1999). Need for a Revision of the Normal Limits of Resting Heart Rate. *Hypertension*, 33:622-625.

Pan, Y., Gao, H., Lin, H., Liu, Z., Tang, L., Li, S. (2018). Identification of Bacteriophage Virion Proteins Using Multinomial Naïve Bayes with g-Gap Feature Tree. *International Journal of Molecular Science*, 19, 1779; doi:10.3390/ijms19061779.

Pangaribuan J. J., Tedja, C., dan Wibowo, S. (2019). Perbandingan Metode Algoritma C4.5 Dan Extreme Learning Machine untuk Mendiagnosis Penyakit Jantung Koroner. *Informatics Engineering Research and Technology* Vol. 1, No.1.

Purushottam, Saxena, K., Sharma, R. (2016). Efficient Heart Disease Prediction System. *Procedia Computer Science,* 85 962 – 969.

Retnasari, T., dan Rahmawati, E. (2017). Diagnosa Prediksi Penyakit Jantung Dengan Model Algoritma *Naïve Bayes* dan Algoritma C4.5, *Konferensi Nasional Ilmu Sosial & Teknologi* (KNiST), pp. 7-12, Maret 2017.

Riani, A., Susianto, Y., Rahman, Nur. (2019). Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Mengunakan Metode Naive Bayes. *Journal of Innovation Information Technology and Application*, Vol.1, No.01, Desember, pp.25-34, DOI: 10.35970/jinita.v1i01.64.

Rodrı´guez, J. D., Rez, A. P., Lozano, J. A. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 569–575, doi: 0162-8828/10/$26.00.

Santoso, H. 2012. *Analisis Dan Prediksi Pada Perilaku Mahasiswa Diploma Untuk Melanjutkan Studi Ke Jenjang Sarjana Menggunakan Teknik Decision Tree dan Support Vektor Machine*. Tesis, Universitas Sumatera Utara.

Third Report of the National Cholesterol Education Program (NCEP). (2001). *Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III*). Executive Summary. National Heart, Lung, and Blood Institute. National Institutes of Health, United State, No. 01-3670, May.

Woodward, M., Webster, R., Murakami, Y., Barzi, F., Lam, T-H., Fang, X., Suh, I., Batty, G. D., Huxley, R., Rodgers, A. (2014). The Association Between Resting Heart Rate, Cardiovascular Disease and Mortality: Evidence From 112,680 Men and Women in 12 Cohorts. *European Journal of Preventive Cardiology*, Vol 21 (6), 719-726.

World Health Organization (WHO), (2019). Cardiovascular diseases (CVDs). Diambil dari https://www. who.int/card iovascular_diseases/en/. [Accessed: 24-Des-2020].

10_JURNAL_S_Coronary Artery Disease Prediction Using Decision Trees and Multinomial

ORIGINALITY REPORT

8%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet	50 words — 1%
2	Yulia Resti, Chandra Irsan, Adinda Neardiaty, Choirunnisa Annabila, Irsyadi Yani. "Fuzzy Discretization on the Multinomial Naïve Bayes Method for Modeling Multiclass Classification of Corn Plant Diseases and Pests", Mathematics, 2023 Crossref	38 words — 1%
3	Randi Rian Putra, Hanna Willa Dhany. "Determination of accuracy value in id3 algorithm with gini index and gain ratio with minimum size for split, minimum leaf size, and minimum gain", IOP Conference Series: Materials Science and Engineering, 2020 Crossref	27 words — 1%
4	D R Ente, S Arifin, Andreza, S A Thamrin. "Comparison of C4.5 algorithm with naive Bayesian method in classification of Diabetes Mellitus (A case study at Hasanuddin University hospital Makassar)", Journal of Physics: Conference Series, 2019 Crossref	21 words — < 1%
5	584621ae7566faad44cf-fb5903404bc6e3be139ec7116b0f2cfc.ssl.cf1.rackcdn.com	16 words — < 1%

Internet

6	www.igi-global.com Internet	16 words — < 1%
7	Lecture Notes in Computer Science, 2008. Crossref	14 words — < 1%
8	Liu, Lei. "Leveraging Machine Learning for Pattern Discovery and Decision Optimization on Last-Minute Surgery Cancellation", University of Cincinnati, 2023 ProQuest	13 words — < 1%
9	ebin.pub Internet	13 words — < 1%
10	www.mdpi.com Internet	13 words — < 1%
11	"Proceedings of International Conference on Recent Trends in Computing", Springer Science and Business Media LLC, 2022 Crossref	12 words — < 1%
12	Jan Bohacik, Michal Zabovsky. "Naive Bayes for statlog heart database with consideration of data specifics", 2017 IEEE 14th International Scientific Conference on Informatics, 2017 Crossref	11 words — < 1%
13	De Falco, Ivanoe. "Differential Evolution for automatic rule extraction from medical databases", Applied Soft Computing, 2013. Crossref	10 words — < 1%
14	developpef.blogspot.com Internet	9 words — < 1%

15	digilib.uinsby.ac.id Internet	9 words — < 1%
16	www.researchgate.net Internet	9 words — < 1%
17	"Machine Learning and Knowledge Discovery in Databases", Springer Science and Business Media LLC, 2023 Crossref	8 words — < 1%
18	F. Herrera, M. Lozano. "Chapter 4 Fuzzy Evolutionary Algorithms and Genetic Fuzzy Systems: A Positive Collaboration between Evolutionary Algorithms and Fuzzy Systems", Springer Science and Business Media LLC, 2009 Crossref	8 words — < 1%
19	Oki Dwipurwani, Dian Cahyawati, Eka Susanti. "Analisis Biplot Robust dengan Metode Minimum Covariance Determinant dalam Mendeskripsikan Provinsi Sumatera Selatan Berdasarkan Karakteristik Angkatan Kerja Menganggur Dari Aspek Gender", Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi, 2022 Crossref	8 words — < 1%
20	Paula Mendonça Leite, Maria Auxiliadora Parreiras Martins, Rachel Oliveira Castilho. "Review on mechanisms and interactions in concomitant use of herbs and warfarin therapy", Biomedicine & Pharmacotherapy, 2016 Crossref	8 words — < 1%
21	eprints.uad.ac.id Internet	8 words — < 1%
22	journal.ugm.ac.id Internet	8 words — < 1%

23	kb.psu.ac.th:8080 Internet	8 words — < 1%
24	www.slideshare.net Internet	8 words — < 1%
25	www.tandfonline.com Internet	8 words — < 1%
26	"Innovative Data Communication Technologies and Application", Springer Science and Business Media LLC, 2020 Crossref	7 words — < 1%
27	Devira Dwimarcahyani, Tessy Badriyah, Tita Karlita. "Classification On Category Of Public Responses On Television Program Using Naive Bayes Method", 2019 International Electronics Symposium (IES), 2019 Crossref	7 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF