Journal FORTEI-JEERI, Vol. 1, No. 1, May 2020: 41-48
https://jeeri.fortei.org/index.php/fortei-jeeri

Improving Feature Selection on Heart Disease Dataset with Boruta Approach

Muhammad Arzanul Manhar[1*], Indah Soesanti[2], Noor Akhmad Setiawan[3]
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada Yogyakarta, Indonesia
*Corresponding Author Email: muh.arzanul.m@mail.ugm.ac.id

Abstract -- Coronary artery disease (CAD) is one of the deadliest diseases globally, including in Indonesia. CAD occurs due to the narrowing or blockage of coronary arteries, which is usually caused by atherosclerosis. Various studies have been conducted to predict the nature and characteristics of this disease. Some researchers use the Z-Alizadeh Sani dataset, which consists of 54 attributes with two classification results, CAD and Normal, to classify its data. Feature selection is one way to reduce the number of attributes that exist by leaving the attributes that have a high effect on the dataset. In this study, the Boruta method used as a feature selection to minimize the attributes and leave the attributes with a high relative with the dataset. By reducing the attributes in the dataset through the feature selection process, sets of 17 and 18 attributes selected as attributes with a high relative with the dataset. These attributes then used to calculate the accurate value of the dataset using the several classification methods, and 90,3% accuracy is obtained from this study.

Keywords:
Boruta
Classification
Coronary Artery Disease
Feature Selection

Article History:
Received: May 13, 2020
Revised: May 16, 2020
Accepted: May 30, 2020
Published: June 3, 2020

DOI: 10.46962/forteijeeri.v1i1.05

I. INTRODUCTION

The heart is responsible for pumping around 3000 gallons of blood throughout our body every day. The heart needs a continuous blood supply to continue pumping blood through the coronary arteries [1]. Without normal blood flow, the heart will lack oxygen and nutrients needed to function normally. When the blood supply needed by the heart becomes limited, or the energy required by the heart becomes higher than the blood supply, the fatal thing that may occur is a heart attack [2][3].

Many things can cause CAD, such as smoking, high blood cholesterol, high blood pressure, excessive diabetes, rarely active or frequent uncontrolled diets [4]. In America, heart disease is one of the leading causes of death for men and women. The American Heart Association (AHA) estimates that every 40 seconds, there is one heart attack in America [5].

CAD is the most common occurrence in the Cardiovascular Disease (CVD) class [6]. The WHO estimates that deaths from CVD remain the leading cause in countries throughout the world. CVD can cause potentially life-threatening complications, especially when the subject has a stroke and heart attack that requires immediate medical treatment [7][8]. With so many cases of death that occur due to heart disease, the right method needed to predict this disease. This prediction is based on several factors that can trigger heart diseases such as diabetes, hypertension, and the influence of smoking [9]. In this case, the data used in research must be data that have a substantial impact in supporting the research conducted.

There are many studies with the purpose of increasing the accuracy of CAD prediction. Robert Detrano used the Cleveland Dataset with the Logistic Regression method to obtain 77% results, and Newton Cheung who focused on the Naive Bayes method obtained an accuracy value of around 81% [10]. Ratnakar, et al. conducted research by removing several attributes that contained missing values and a 79% accuracy value obtained from this experiment. Pedreira et al. used the Neural Network method in the Cleveland dataset and an accuracy of 89%.

These studies mostly use UCI Dataset, which is the popular dataset for CAD research. This dataset itself made from 1989 and is very old now. Besides the UCI dataset, there is the new dataset for CAD research which is Z-Alizadeh Sani Dataset that contains 303 data. František Babič et al. uses Predictive and Descriptive Analysis with this dataset and obtained 86,67% accuracy value for the SVM Method and 86,32% with the Neural Network method [11]. Because this dataset contains 54 attributes, lots of research has a purpose reducing the attributes of this dataset. Many attributes didn't mean that the quality of the dataset is also high. Sometimes the attribute has a little effect of the dataset as a whole. In this cause, feature selection will be used to decrease the attribute until only the high effect attributes for the dataset left. Ümit KILIÇ et al. uses the Artificial Bee Colony method as the feature selection method and obtained 89,4% accuracy with the SMO algorithm and 88,4% with the Random Forest algorithm [12]. This research aims to find the best attributes of the Z-Alizadeh Dataset with the feature selection process using the Boruta algorithm.

II. METHODS AND MATERIAL

A. Dataset

In this study, the Z-Alizadeh Sani dataset is used as one of the datasets that are often used in CAD research [13][14][15]. This dataset was collected from random patients at Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Center [16]. IT consisted of 303 data where 216 of it has CAD label, and the rest are labeled Normal. These data are divided into four groups which are:

1. Demographic.
2. Symptom and examination.
3. ECG.
4. Laboratory and echo features.

Each data has two possible classification categories, which are 'CAD' and 'Normal' where the narrowing of the arteries by 50% or more means that the patient has CAD and if the narrowing of the arteries is less than 50% it means the patient is in the normal category.

TABLE I.
Dataset's Attributes For Laboratory And Echocardiography Category

Attributes	**Value**
FBS (*Fasting Blood Sugar*)	62 – 400
Cr (*Creatine*)	0.5 – 2.2
TG (*Triglyceride*)	37 – 1050
LDL (*Low-Density Lipoprotein*)	18 – 232
HDL (*High-Density Lipoprotein*)	15 – 111
BUN (*Blood Urea Nitrogen*)	6 – 52
ESR (*Erythrocyte Sedimentation Rate*)	1 – 90
Hb (*Hemoglobin*)	8.9 – 17.6
K (*Potassium*)	3.0 – 6.6
Na (*Sodium*)	128 – 156
WBC (*White Blood Cell*)	3700 - 18000
Lymph (Lymphocyte)	7 – 60
Neut (Neutrophil)	32 – 89
PLT (Platelet)	25 – 742
EF (*Ejection Fraction*)	15 – 60
Region With RWMA (Region Wall Motion Abnormality)	0, 1, 2, 3, 4
VHD (*Valvular Heart Disease*)	*Normal, Mild, Moderate, Severe*

TABLE II.
Dataset’s Attributes For Symptoms and Examination Category

Attributes	Value
BP (*Blood Pressure*)	90 – 190
PR (*Pulse Rate*)	50 – 110
Edema	*Yes, No*
Weak Peripheral Pulses	*Yes, No*
Lung Rales	*Yes, No*
Systolic Murmur	*Yes, No*
Diastolic Murmur	*Yes, No*
TCP (*Typical Chest Pain*)	*Yes, No*
Dyspnea	*Yes, No*
Function Class	1, 2, 3, 4
Atypical CP	*Yes, No*
Non-anginal CP	*Yes, No*
ECP (*Exertional Chest Pain*)	*Yes, No*
Low Threshold Angina	*Yes, No*

TABLE III.
Dataset’s Attributes For Demographic Category

Attributes	Value
Age	30 – 86
Weight	48 – 120
Length	140-188
Sex	*Male, Female*
BMI (*Body Mass Index*)	18 – 41
DM (*Diabetes Mellitus*)	Yes, No
HTN (Hypertension)	Yes, No
Current Smoker	Yes, No
Ex-Smoker	Yes, No
FH (*Family History*)	Yes, No
Obesity	*Yes* (if MBI>25), *No* (if MBI<=25)
CRF (*Chronic Renal Failure*)	*Yes, No*
CVA (*Cerebrovascular Accident*)	*Yes, No*
Airway Disease	*Yes, No*
Thyroid Disease	*Yes, No*
CHF (*Congestive Heart Failure*)	*Yes, No*
DLP (Dyslipidemia)	*Yes, No*

Table IV.
Dataset’s Attributes For ECG Category

Attributes	Value
Rhythm Sin	Sin, Af
Q Wave	*Yes, No*
ST-Elevation	*Yes, No*
ST Depression	*Yes, No*
Tinversion	*Yes, No*
LVH (*Left Ventricular Hypertrophy*)	*Yes, No*
Poor R Regression	*Yes, No*

B. Pre-Processing

The pre-processing process is done by replacing non-numeric values in the data with numerical values. Then these data will be normalized using the following equation:

$$f(x) = \sum_{i=1}^{N} \frac{x_i - min_i}{max_i - min_i} \tag{1}$$

$f(x) = \sum_{i=1}^{N} \frac{x_i - min_i}{max_i - min_i}$ The value of N is the number of attributes in the dataset, while x is the data to be normalized. One of the advantages of normalizing the data is facilitating the feature selection process to minimize the attributes that will be used later. This pre-processing process is carried out before the classification process, where the classification processes require numerical values of each data provided.

C. Feature Selection

Feature Selection is one way to eliminate attributes that have little effect on the dataset as a whole. These days many datasets usually have many variables. One problem with the significant amount of data in a dataset is that not all of the data has the same effect when compared to other data. Sometimes there are some irrelevant data and have little impact on the classification given. By reducing the dimensions of the dataset, it is expected that the data used are data that have a high relative on the dataset so that the research results obtained are the optimal results.

In this study, the feature selection process we use is the Boruta method. This method used by utilizing a random forest equation using the Z-Score value, which is calculated by dividing the average loss value against its standard deviation [17] and the Z-Score itself determined by the MDI equation. This calculation used to see the importance of the existing attributes. The Z score itself does not directly relate to the importance of an attribute because the value obtained from the random forest is not in the form of exact values such as 0 and 1.

$$z = (x - \mu)/\,\sigma \tag{2}$$

$z = (x - \mu)/\,\sigma$Because Z-Score cannot be used to determine the importance of an attribute directly, an additional system is made, which consists of new attributes consisting of random values from the old attribute. This attribute is then called the shadow attribute. The processes in Boruta's approach are [18] :

1. Make the shadow attribute.
2. Randomize the value of new attributes so that the correlations with old data are lost.
3. Calculate the Z-Score using a random forest equation.
4. Determine the maximum value of Z-Score in Shadow Attribute and mark the attribute with a better value than that value.
5. For each attribute whose importance has not been determined, a Z-Score test will be performed.
6. Attributes that have a lower value than the maximum Z-Score will be marked as 'unimportant' and will be deleted from the dataset.
7. Mark the attribute with a value higher than the maximum value of the Z-Score as ‘important’.
8. Elimination of all shadow attributes made at the beginning.
9. Repeatition of the procedure above until all attributes have their importance.

D. Impurity

Z-Score calculation itself can be done by utilizing the importance score of each attribute. This calculation is related to the impurity of an attribute, which is a value that explains the state of an attribute that does not have clear certainty about the result decision. The importance score can be identified in two ways, Mean Decrease of Impurity (MDI) and Mean Decrease of Accuracy (MDA). MDI is obtained by summing the total reduction in the impurity of each node where the attribute appears. At the same time, the MDA is generated by measuring the value of the accuracy reduction from out-of-bag samples, where the attribute is generated randomly [19][20].

In this study, the importance score was obtained using MDI calculations because the calculations were carried out faster and did not require bootstrap sampling. The calculation will take place in each iteration where the attribute that has gotten the 'important' label is not included in the next iteration.

$$imp\,(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t\, \in T: v(t) = X_m} p(t) \Delta i(t) \tag{3}$$

Where p (t) = N_t / N and Δi (t) is the reduction of impurity at the node t produced with the following equation:

$$\Delta i(t) = 1 - \sum_{n=1}^{C} p(n)^2 \tag{4}$$

The C value is the number of classes that the variable has.

E. Experimental Design

This research aims to minimize the high dimensions of the dataset by removing the attributes that have low importance to the dataset.

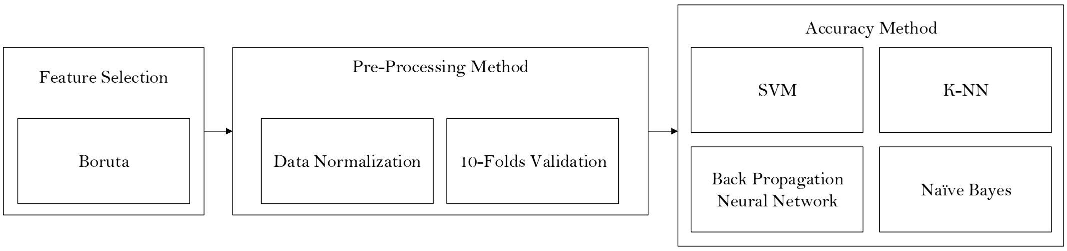

FIGURE. 1.
Proposed System

In the process of calculating the accuracy, the data used comes from the dataset that has been through the feature selection process so that the dataset used is a dataset with the reduced attributes. Boruta's approach is used to process the feature selection, and several classification process functions used to calculate the accurate value of the dataset used. The system used in this study is as follows:

1. Perform a feature selection process to reduce the dimensions of the dataset.
2. It is pre-processing the new dataset by converting all data to numeric values.
3. Normalize data on each attribute so that data distribution is better.
4. Calculate the accuracy value in the new dataset with several methods and choose the method with the highest accuracy as the best method.

The feature selection process used in this study is the Boruta method. There are other feature selection processes like Pearson Correlation [21]. They determine the absolute value of the target and numeric value on the dataset and Chi-Square [22], which use the chi-square metric between the target and numeric variables and choose the variable with the highest chi-square value. In this research, the Boruta method was chosen because the process determines the importance of each attribute with the dataset as a whole.

To calculate the value of accuracy, we will use the 10-folds validation approach where the data used will be divided into ten sections where one part will be a test-set, and the rest will be a train-set [23]. These sections later will be used in every iteration to obtained the accuracy value. The classifications used in this research are K-NN, SVM, Backpropagation Neural Network and Gaussian Naïve Bayes. That processes were chosen compared with the old research about Z-Alizadeh Sani Dataset’s feature selection using Artificial Bee Colony as the feature selection method.

III. RESULT AND DISCUSSION

The table below shows the result with all the attributes in the dataset.

TABLE V.
Accuracy Result with All Attributes

Method	**Accuracy**
K-NN	83,6 %
Naïve Bayes	82 %
SVM	86 %
Back Propagation Neural Network	85 %

As a comparison, we used other research with different feature selection methods, and the result is shown below.

TABLE VI.
Accuracy Result with Artificial Bee Colony as Feature Selection Method

Method	**Num. of Attribute**	**Accuracy**
K-NN	6	84,4%
Naïve Bayes	16	86,7%
SMO	16	89,4%
Random Forest	23	88,4%

From the table above, we get the best accuracy result with 89,4% from the SMO method, and other classification methods also give a good accuracy value. But the number of attributes on the classification method is mostly very different for each method, which means that even though the accuracy value is good, there is no good way to determine the best attributes.

For the Boruta method, we get three sets of attributes that contain 17 and 18 attributes.

TABLE VII.
Selected Attributes Result

Number of Attributes	**Attribute's Index in Dataset**
17	0, 5, 6, 17, 23, 24, 27, 28, 34, 38, 39, 40, 48, 49, 52, 53, 54
18	0, 5, 6, 17, 18, 24, 27, 28, 34, 38, 39, 40, 44, 49, 50, 52, 53, 54

Most of the attributes on those two sets of new attributes contain the same attributes. The shadow attributes created on the feature selection process before it become the main point of the difference of the final result of attributes obtained.

TABLE VIII.
Accuracy Result With Boruta Method as Feature Selection

Method	Num. of Attribute	Accuracy
K-NN	17	85,3 %
Naïve Bayes	18	86,3 %
SVM	18	90,3 %
Back Propagation Neural Network	17	89 %

In the table above, we get the best result of 90,3% by using the Support Vector Machine as a classification method. The best number of attributes for the process is 17 attributes that labeled as 'important' attributes. These attributes come from all the four categories from the Z-Alizadeh Sani Dataset, and with the Laboratory and Echocardiography category provides most attributes.

This study's results indicate that by reducing the dimensions of the existing dataset, the accuracy value obtained from the dataset can increase from 89,4% in the previous research to 90,3% with a sensitivity value of around 93% and 82% specificity. The reason with the low specificity value is because this dataset only contains a few data with Normal classification. Along with the 303 data in this dataset, only 87 of them were obtained from normal patients, which is just about 28% of the dataset.

With these selected attributes, the new dataset can be generated with several classification methods. As a result, shows in Table 5. This classification process has better accuracy results than the previous study, where several classification methods have been carried out for research using this dataset. The very little increase in accuracy value is caused by a very small amount of data on the machine learning process, which usually requires a lot of data to increase the accuracy value properly.

IV. CONCLUSION

The problem often faced in using a dataset is that sometimes these datasets have dimensions that are too large, where several data that are actually not so important to the dataset but are still often used in research. One way to fix this problem is to use the feature selection approach, to reduce the number of attributes or data in the dataset so that the research carried out only uses data that is important to the dataset used.

In this study, the Boruta approach is used to reduce attributes that are not highly important to the whole dataset. By using this method, 17 and 18 important attributes were obtained to test its accuracy value. The Backpropagation Neural Network algorithm, K-NN, Naïve Bayes, and SVM are used to calculate the accuracy value and obtained a value above 85%. The best result is 90,3% using the SVM method, and this means that the Boruta method has good results in the feature selection process. The distribution of the number of attributes obtained in this study is better because it can determine several attributes that have the highest importance value. While in previous study, the numbers of obtained attributes that have good accuracy value are 6, 16 to 23 attributes. In this study, the number of attributes that produce a good accuracy value only ranges from 17 and 18 attributes.

V. REFERENCES

[1] Z. Zhao and C. Ma, "An intelligent system for -oninvasive diagnosis of coronary artery disease with EMD-TEO and BP neural network," *2008 Int. Work. Educ. Technol. Train. 2008 Int. Work. Geosci. Remote Sensing, ETT GRS 2008*, vol. 2, no. 1, pp. 631–635, 2009.

[2] M. G. Tsipouras *et al.*, "Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 4, pp. 447–458, 2008.

[3] A. Cüvitoğlu and Z. Işik, "Classification of CAD dataset by using principal component analysis and machine learning approaches," *2018 5th Int. Conf. Electr. Electron. Eng. ICEEE 2018*, pp. 340–343, 2018.

[4] *et al.*, "A review on prevalence, causes, preventions, and treatments of coronary artery disease," *Asian Pacific J. Heal. Sci.*, vol. 4, no. 4, pp. 104–107, 2017.

[5] E. J. Benjamin *et al.*, *Heart disease and stroke statistics - 2018 update: A report from the American Heart Association*, vol. 137, no. 12. 2018.

[6] W. M. Baihaqi, T. Hariguna, and T. Astuti, "Review on fuzzy expert system and data mining techniques for the diagnosis of coronary artery disease," *Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017*, vol. 2018-Janua, pp. 353–358, 2018.

[7] U. Desai, C. G. Nayak, G. Seshikala, and R. J. Martis, "Automated diagnosis of Coronary Artery Disease using pattern recognition approach," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 434–437, 2017.

[8] S. M. J. Jalali, M. Karimi, A. Khosravi, and S. Nahavandi, "An efficient neuroevolution approach for heart disease detection," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 2019-Octob, pp. 3771–3776, 2019.

[9] N. P. Waghulde and N. P. Patil, "Genetic Neural Approach for Heart Disease Prediction," *Int. J. Adv. Comput. Res.*, vol. 4, no. 3, pp. 778–784, 2014.

[10] R. Devi Priya and S. Kuppuswami, "A genetic algorithm based approach for imputing missing discrete attribute values in databases," *WSEAS Trans. Inf. Sci. Appl.*, vol. 9, no. 6, pp. 169–178, 2012.

[11] 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*, vol. 11, pp. 155–163, 2017.

[12] Ü. Kiliç and M. Kayakeleş, "Feature Selection with Artificial Bee Colony Algorithm on Z-Alizadeh Sani Dataset," *Proc. - 2018 Innov. Intell. Syst. Appl. Conf. ASYU 2018*, pp. 8–10, 2018.

[13] R. Alizadehsani *et al.*, "A data mining approach for diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 111, no. 1, pp. 52–61, 2013.

[14] S. N. R. Alizadehsani, M.H. Zangooei, M.J. Hosseini, J. Habibi, A. Khosravi, M. Roshanzamir, F. Khozeimeh, N. Sarrafzadegan, "Coronary artery disease detection using computational intelligence methods," *Knowledge-Based Syst.*, no. 109, pp. 187–197, 2016.

[15] A. A. Y. Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," *Comput. Methods Programs Biomed.*, vol. 141, pp. 19–26, 2017.

[16] R. Alizadehsani, M. J. Hosseini, Z. A. Sani, A. Ghandeharioun, and R. Boghrati, "Diagnosis of coronary artery disease using cost-sensitive algorithms," *Proc. - 12th IEEE Int. Conf. Data Min. Work. ICDMW 2012*, pp. 9–16, 2012.

[17] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta - A system for feature selection," *Fundam. Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.

[18] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010.

[19] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in Forests of randomized trees," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2013.

[20] J. H. Hur, S. Y. Ihm, and Y. H. Park, "A variable impacts measurement in random forest for mobile cloud computing," *Wirel. Commun. Mob. Comput.*, vol. 2017, 2017.

[21] N. J. Gogtay and U. M. Thatte, “Principles of correlation analysis,” *J. Assoc. Physicians India*, vol. 65, no. MARCH, pp. 78–81, 2017.

[22] A. Ugoni and B. F. Walker, “The Chi square test: an introduction.,” *COMSIG Rev.*, vol. 4, no. 3, pp. 61–4, 1995.

[23] A. Cüvitoğlu and Z. Işik, “Classification of CAD dataset by using principal component analysis and machine learning approaches,” *2018 5th Int. Conf. Electr. Electron. Eng. ICEEE 2018*, pp. 340–343, 2018.