Classification of Coronary Heart Disease Using Classification and Regression Trees (CART) Method

Resita Mia Noviana [1, a)] , Mustika Hadijati [1, b)] and Lisa Harsyiah[1, c)]

[1]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Mataram, Mataram, Indonesia.*

[a)] resitamia04@gmail.com
[b)] mustika.hadijati@unram.ac.id
[c)] Corresponding author: lisa_harsyiah@unram.ac.id

Abstract. Coronary Heart Disease (CHD) is a disease caused by narrowing and blockage of the coronary arteries in the heart. According to WHO, CHD is one of the substantial causes of death globally. The exact cause of CHD is not known, although several risk factors cause a person to develop CHD. Risk factors a person can suffer from CHD diverge into two, specifically risk factors that can be changed and cannot be changed. Considering that CHD has complex risk factors with a high risk of death, a solution is necessary to the problem in the form of a mathematical model, namely the classification method.

The classification method used in this study is the Classification and Regression Trees (CART) which uses decision tree techniques. The CART method classifies the dataset on heart disease patients at Siloam Hospital Mataram in 2020 with categorical-scale response variables CHD and NON-CHD. After considering the medical record's patient, seven influential risk factors have been selected, including age, gender, blood pressure, diabetes, dyslipidemia, obesity, and genetics.

The results of the CART analysis produce an optimal classification tree consisting of eight parent nodes with nine terminal nodes so that there are nine classification rules. The variable that most influences patients to have coronary heart disease is obesity. Three other variables that influence patients to suffer from coronary heart disease are age, blood pressure, obesity, and diabetes. Based on the result of the Apparent Error Rate (APER), the classification error value in the CART model is 1.1%, such that the classification accuracy is 98.9% which can be said as a good classification.

INTRODUCTION

Coronary Heart Disease (CHD) is the most common type of heart disorder and is the common cause of death in developed and developing countries, including Indonesia. CHD is one of the non-communicable diseases (NCD) which tends to increase every year and has an impact not only on developed countries but also on developing countries. According to World Health Organization (WHO) [9], CHD is one of the substantial causes of death globally with pneumonia and stroke. By 2020 CHD is estimated will be the most common first killer with 36% of all deaths, which is twice as high as the death rate from cancer. Reported that CHD in Indonesia is the most common and first cause of all deaths at 26.4%, this is four times higher than the death rate caused by cancer (6%). In other words, approximately one in four people who die in Indonesia are due to CHD. Based on WHO calculation data, which estimates that by 2030, cardiovascular disease will account for around 23.6 million deaths in the world. Surely this will be the center of attention for health observers, so the authors are interested in picking this case as a research topic. The exact cause of CHD is not known, although several risk factors cause a person to develop CHD.

The risk factors for a person suffering from CHD can be divided into two, namely modifiable risk factors and non-modifiable risk factors. Modifiable risk factors such as hypertension, diabetes mellitus, smoking, hyperlipidemia, obesity, an inactive lifestyle, and stress are also risk factors [5]. Meanwhile, risk factors that cannot be changed include age, gender, and family or genetic history. Even though you already know the risk factors that cause a person to suffer from coronary heart disease, knowing someone has CHD, can be known by consulting a cardiologist and conducting several laboratory tests to get valid data results. The data obtained from the medical records must be consulted again by a cardiologist to be diagnosed whether the patient has CHD or not. One of the hospitals that are used as a referral for heart disease patients for check-ups is Siloam Hospital Mataram.

Siloam Hospital Mataram is a private hospital located in Mataram City. In 2019 this hospital was accredited by the Hospital Accreditation Commission (KARS) with a plenary score, which means Siloam Hospital Mataram received the highest rating predicate given based on an assessment of quality management and patient safety applied in hospitals. Siloam Hospital Mataram provides facilities for heart disease patients to perform heart check-ups supported by special equipment such as echocardiography examination, examination and installation of Holter monitoring (ambulatory electrocardiogram), and electrocardiography (ECG). In addition, Siloam Hospital Mataram also has an integrated Intensive Coronary Care Unit (ICCU) or intensive care unit for heart disease, especially coronary heart disease, heart attacks, severe heart rhythm disorders, and heart failure.

Considering that CHD has complex risk factors with a high risk of death, a solution to this problem is needed in the form of modeling. In this modeling, it can be predicted whether a person is classified as having CHD or not, so that early prevention can be done. Medical record data from patients that have been stored in the database can form a pattern for determining CHD in the form of a mathematical model, namely the classification method. Classification can classify patients with CHD and NONPJK based on the factors that influence the patient's chances of suffering from CHD. There are various classification methods, both parametric and nonparametric approaches. Nonparametric methods do not rely on certain assumptions so they can be more flexible in analyzing data but still have a high level of accuracy. There are various classification methods with nonparametric approaches, including Classification and Regression Trees (CART), Chi-Squared Automatic Interaction Detection (CHAID), C4.5 algorithm, and other classification methods. Based on references from the results of previous studies that have compared the CART algorithm or method with other algorithms, the average result is that the accuracy of the CART classification is better than other methods, one of which is Nuriyah's [7] research which results in the results of the classification accuracy CART is better than the CHAID method. Therefore, in this study, the method that will be used is the CART method. Based on the description above, this study will classify the factors that influence the chances of patients suffering from CHD using the CART method in heart disease patients at Siloam Hospital Mataram in 2020.

RESULT AND DISCUSSION

Description of Heart Disease Patient Data

The status of coronary heart disease patients was classified into CHD and NONPJK categories. An overview of the characteristics of coronary heart disease patients at Siloam Hospital Mataram will be described statistically based on 7 variables that affect the status of coronary heart disease patients including age, gender, blood pressure, diabetes, and

dyslipidemia, obesity, genetics. To get a comprehensive picture of coronary heart disease patients at Siloam Hospital Mataram, it is necessary to do a statistical description of 7 variables that affect the status of coronary heart disease patients. A description of the status of coronary heart disease patients at Siloam Hospital Mataram is presented in Figure 1 below.

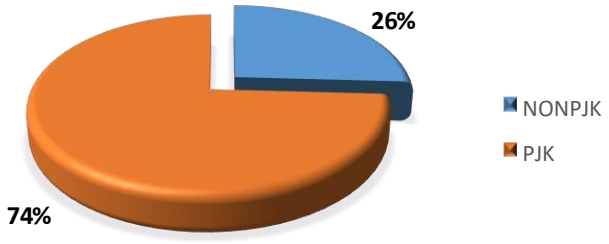

FIGURE 1. Description of Heart Disease Patient Status

Figure 1 explains the status of coronary heart disease patients at Siloam Hospital Mataram in 2020. Of the 89 patients who were research subjects, it can be seen that in 2020, more coronary heart disease patients at Siloam Hospital Mataram were positive for CHD, which was 74.16% percent or as many as 66 patients. While the non-NPJK was 25.84% percent or as many as 23 patients.

Classification and Regression Trees (CART) Analysis

The first step in the CART analysis is the formation of a maximum classification tree. The initial step taken to form a classification tree is to determine the sorting variable and the variable value (threshold). The disaggregating variable and the threshold value were selected from several possible disaggregations of each predictor variable. Next is to calculate the value of the Gini Index variable to get the value of node heterogeneity. The Gini index performs node sorting on each of the rights and left nodes. Then the Gini index value is used as a determinant of the goodness of splits from each sorter. The selected sorter is the sorting variable with the variable value that has the maximum goodness of splits value. The maximum tree construction that is formed based on the factors that influence the diagnosis of coronary heart disease is shown in Figure 2.

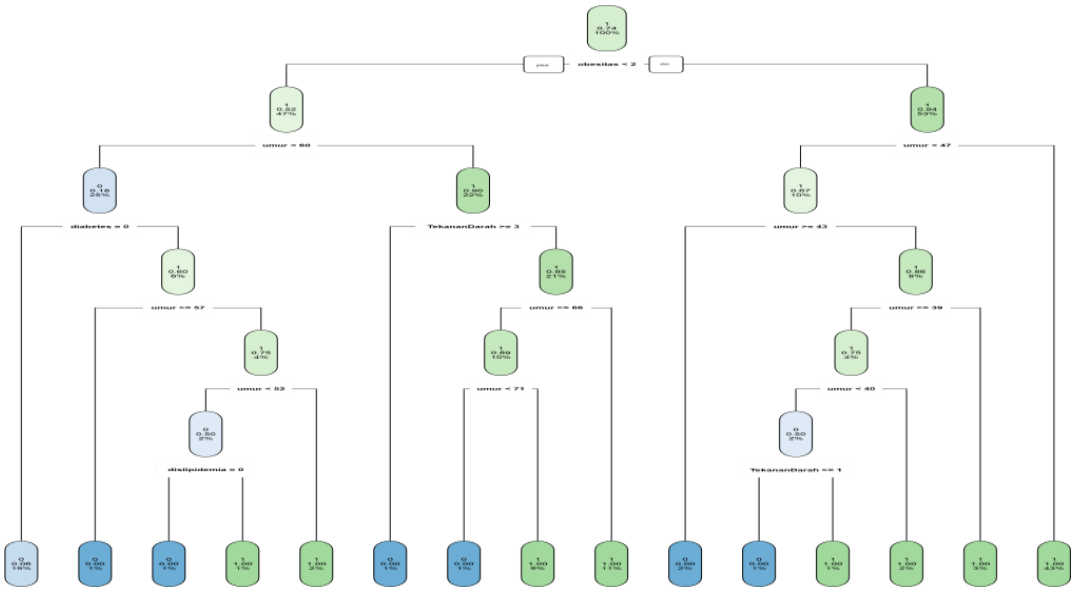

FIGURE 2. Maximum Classification Tree

Figure 2. shows the maximum classification tree (tree) which is large with a total of 27 nodes, consisting of 13 parent nodes and 14 terminal nodes or end nodes with a depth of 7. Because the trees formed are large, pruning or pruning is

done to reduce the complexity of the tree so that it becomes simpler that an optimal classification tree is formed. A good tree can be obtained by pruning based on minimum cost complexity. Judging by the value of the complexity parameter that can minimize the value of the cross-validation error (CV Error), as presented in Table 2:

TABLE 2. Classification Tree Formation Order

	CP	N Split	Rel Error	CV Error	SE
1	0,30434783	0	1,000000	1,000000	0,17956
2*	0,04347826	2	0,391304	0,69565	0,15751
3	0,02173913	8	0,130435	0,82609	0,16807
4	0,01449275	10	0,086957	0,86957	0,17121
5	0,00000000	13	0,043478	0,86957	0,17121

*optimal classification tree

Based on Table 2, the cost complexity parameter value used for pruning is the complexity parameter value that can minimize the value of the cross-validation error, where the CP value of 0.04347826 with a cost cross-validation error value of 0.69565 is used as tree pruning so that the following optimal classification tree is obtained:

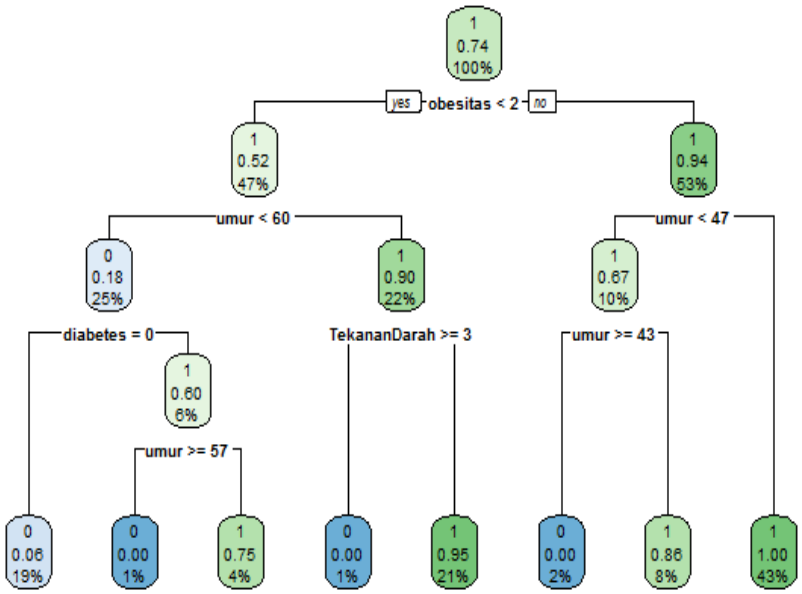

FIGURE 3. Optimal Classification Tree

Based on Figure 3. The obesity variable is the main disaggregating variable and the variable that most determines the classification of coronary heart disease factors at Siloam Hospital Mataram. In addition, three other variables also affect the classification of coronary heart disease patients, namely the patient's age, diabetes, and the patient's blood pressure. Based on Figure 3, the optimal classification tree consists of 8 terminal nodes, meaning that there are 8 classification rules, including:

a. If the patient has normal weight or overweight category with age less than 60 years old, has normal blood sugar levels, or does not suffer from diabetes mellitus, then the patient is classified with NON CHD..

b. If the patient has normal weight or overweight category with age less than 60 years old, has high blood sugar levels or suffers from diabetes mellitus, has normal blood pressure or level 2 hypertension, and has normal blood cholesterol, then the patient is classified with NON CHD..
c. If the patient has normal weight or overweight category with age less than 60 years old, has high blood sugar levels or suffers from diabetes mellitus, has normal blood pressure or level 2 hypertension, and has high blood cholesterol, then the patient is classified with CHD.
d. If the patient has normal weight or overweight category with age less than 60 years old, has high blood sugar levels or suffers from diabetes mellitus, has blood pressure in the pre-hypertension category or hypertension level 1, then the patient is classified with CHD.
e. If the patient has a normal weight category or is overweight with age more than 60 years old, has a blood pressure category of hypertension level 2, then the patient is classified with NON CHD..
f. If the patient has normal weight or overweight category with age more than 60 years old, has normal blood pressure or prehypertension or level 1 hypertension, then the patient is classified with CHD.
g. If the patient is overweight or obese with an age less than 47 years and greater or equal to 43 years, the patient is classified with NON CHD.
h. If the patient is overweight or obese with an age of more than 47 years old, then the patient is classified with CHD.

The next step is to calculate the classification accuracy of the result of the CART tree whether it is good or not. The results of the classification accuracy obtained using the APER measure are shown in the following table:

TABLE 3. CART Classification Accuracy

Observation Results	**Estimated Classification**		**Total**
	NON CHD	CHD	
NON CHD	23	1	24
CHD	0	65	65
Total	23	66	89

The CART accuracy value is calculated using APER as follows:

$$APER\% = \left(\frac{0+1}{89}\right)100\% = 0{,}011 \times 100\% = 1{,}1\%$$
$$Classification\ Accuracy = 100\% - 1{,}1\% = 98{,}9\%$$

Based on the result of the Apparent Error Rate (APER), the classification error value in the CART model is 1.1%, such that the classification accuracy is 98.9% which can be said as a good classification.

CONCLUSION

Characteristics of heart disease patients based on the patient's diagnosis (CHD or NONPJK) it was found that there were more patients diagnosed with CHD than patients diagnosed with NONPJK, with 66 patients being diagnosed with CHD and 23 of them being diagnosed with NONPJK. The results of the CART classification resulted in a classification accuracy of 98.9%. The CART classification produces an optimal classification tree with 8 terminal nodes, of which 4 terminal nodes are classified as CHD patients, namely class labels of 1, and 4 terminal nodes are classified as NONPJK patients, namely class labels of 0. Patient characteristics are based on factors that influence the diagnosis of coronary heart disease patients. Is:

1. Characteristics of patients diagnosed with CHD are patients who have body mass index with normal or overweight category with age less than 60 years old, and suffer from diabetes; patients who have normal weight or overweight category with age more than 60 years old, and have blood pressure category of normal or pre-hypertension or hypertension level I; patients who have body mass index with overweight or obese category and they have an age less than 47 years old; patients who have body mass index with overweight or obese category and their age more than or equal to 47 years.

2. Characteristics of patients diagnosed with NONPJK were patients who have body mass index with normal or overweight category with age less than 60 years old, and did not suffer from diabetes; patients who have body mass index with normal or overweight category with an age range between 57-59 years old; patients who have body mass index with normal or overweight category with age more than or equal to 60 years old and have blood pressure with hypertension II category; patients who have body mass index with overweight or obese category with an age range of 43-46 years old.

ACKNOWLEDGEMENT

First and foremost, the writers would like to express our deepest gratitude for Almighty God, for His marvelous and amazing grace, for the countless blessings and love so the writers have finally completed this paper. The writers are also expressing their extreme gratitude to KPBI Statistics Department of Mathematics, University of Mataram for their support.

REFERENCES

1. Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Tree*, (Chapman And Hall, New York, 1993).
2. Huon H Gray, etc, *Lecture Notes: Kardiologi* 4th ed, (Erlangga Medical Series, Jakarta, 2002).
3. Kemenkes RI, *Riset Kesehatan Dasar: RISKESDAS*, (Balitbang Kemenkes RI, Jakarta, 2013).
4. M.D Lewis dan Roger, J, *An Introduction to Classification and Regression Trees* (CART) *Analysis*. Annual Meeting of Society For Academic Emergency, (UCLA Medical Center, California, 2000).
5. J. Mackay and G.A Mensah, *The Atlas of Heart Disease and Stroke*, (WHO, Geneva 2004).
6. Mozaffarian, etc. (2008). Beyond Established and Novel Risk Factors: Lifestyle Risk Factors For Cardiovascular Disease. *Circulation*. 117(23), 3031.
7. Nuriyah, "Perbandingan Metode *Chi-Square Automatic Interaction Detection* (CHAID) dan *Classification And Regression Trees* (CART) dalam Menentukan Klasifikasi Alumni UIN Sunan Kalijaga Berdasarkan Masa Studi". Skripsi, Program Sarjana Fakultas Sains dan Teknologi Universitas Islam Negeri Sunan Kalijaga, (2013).
8. R. Pratiwi, "*Perbandingan Klasifikasi Algoritma C5.0 Dan Classification And Regression Trees"*. Skripsi, Universitas Mulawarman, 2020.
9. WHO. (2014). *World Helth Statistics* 15 Mei 2014 (https://www.who.int/news/item/15-05-2014-world-health-statistics-2014), downloaded at 08.00 PM, date 12/04/2021.
10. WHO. (2009). Cardiovascular Disease (Cvds) (https://www.who.int/cardiovascular_diseases/en/cvd_atlas_01_types.pdf), downloaded at 08.30 PM, date 12/04/2021.
11. Y Yohannes and J Hoddinot, *Classification and Regression Trees: An Introduction*, (Internal Food Policy Research Institute (IFPRI), USA, 1999).