Online Version Available at: http://bsti.ubd.ac.id/e-jurnal

JOURNAL OF TECH-E

|2581-1916 (Online) | 2598-7585 (Printed) |

Article

Comparison of Data Mining Methods Using C4.5 Algorithm and Naive Bayes in Predicting Heart Disease

Rino

Buddhi Dharma University, Faculty of Sains & Technology, Banten, Indonesia

SUBMISSION TRACK

Recieved: Februari 19, 2021
Final Revision: Februari 26, 2021
Available Online: March 15, 2021

KEYWORD

Heart Disease, Data Mining, C4.5, Naïve Bayes

CORRESPONDENCE

E-mail: rino@ubd.ac.id

A B S T R A C T

Heart disease is a condition of the presence of fatty deposits in the coronary arteries in the heart which changes the role and shape of the arteries so that blood flow to the heart is obstructed. Data mining methods can predict this disease, some of the methods are C4.5 Algorithm and Naive Bayes which are often used in research.

The data set in this research was obtained from the uci machine learning repository site, where the dataset has 3546 records and 13 attributes.

The accuracy value of the Naïve Bayes algorithm has a high value of 81.40% compared to the C4.5 algorithm which only has an accuracy value of 79.07%. Based on the calculation results, it can be concluded that the Naïve Bayes Algorithm is a very good clarification because it has a value between 0.709 - 1.00.

From conclusion above, the Naïve Bayes algorithm has a higher accuracy value than the C4.5 algorithm so the researchers decided to use the Naïve Bayes algorithm in predicting heart disease.

INTRODUCTION

Heart disease is a condition of the presence of fatty deposits in the coronary arteries in the heart which changes the role and shape of the arteries so that blood flow to the heart is delayed [1]. The World Health Organization in 2013 states that the death rate caused by heart disease is 45%, and it is estimated that in 2030 it will increase by 23.3 million people each year. In 2013, the prevalence of heart disease diagnosed by doctors in Indonesia was 883,447 people. Central Java Province is in third place with a total of 120,447 heart disease sufferers (basic health research).

Heart disease occurs indirectly, usually a person will experience a process of narrowing of the coronary vessels in a long enough time, therefore everyone has a risk of heart disease. In addition, there are other factors that cause a person to experience heart disease, namely lifestyle and genetic factors.

Currently, technological developments have entered many fields, one of which is in the medical world. The use of information technology (IT) in the medical world is familiar to the wider community and one of these uses is used to predict heart disease. There are many ways or methods to predict

heart disease, one of which is using data mining.

Data mining (DM) is a combination of a number of computer science disciplines [2], which defines it as the process of discovering new patterns from very large data sets, including methods that are a slice of artificial intelligence, machine learning, statistics, and database systems [2]. Data mining or word addition is a relatively fast and easy technique to find patterns and / or relationships between data, automatically.

By combining four computer science disciplines as defined above, knowledge can be found in five sequential processes: selection, processing, transformation, data mining, and interpretation / evaluation [3]. Data mining is a data processing method where does it work to find hidden patterns from some dataset and the results its can be used to make decisions in the future. This data mining is also known as pattern recognition [4].

Data mining is a large-scale data processing method, therefore data mining has an important role in industry, finance, weather, science and technology. In general, data mining studies discuss methods such as clustering, classification, regression, variable selection and market basket analysis. From the above definitions, it can be concluded that in general data mining is a data analysis activity to look for a certain pattern, with a large amount of data and aims to produce information that can be used and developed further.

The C4.5 algorithm is a program that contributes to a data set labeled and produces a decision tree as output [5]. This follow-up decision tree is then verified against invisible labeled test data to calculate generalizations. C4.5 is a program used to generate taxonomic rules using a decision tree from a given data set. The C4.5 algorithm is an extension of the basic ID3 algorithm and was designed by Quinlan. C4.5 is one of the most widely used learning algorithms. The C4.5 algorithm constructs a decision tree from a series of training data similar to the ID3 algorithm, using the information entropy concept. C4.5 is also known as statistical classification.

One of the classification algorithms that is often used and has received a lot of attention from researchers in predicting heart disease is Naive Bayes and Decision Tree (C45). Simplicity of the Naive Bayes algorithm and the Decision Tree (C4.5). What makes these two algorithms attractive are because their have high accuracy for prediction and can be implemented in various applications, such as expert system, data mining for prediction and classification [3]. Because of this, we compare the two algorithms from their level of accuracy and time performance to improve prediction performance.

Naive Bayes is one of the algorithms contained in the clarification technique. Naive Bayes is a clarification with the probability and statistical method proposed by the British scientist Thomas Bayes, namely predicting future opportunities based on previous experiences so it is known as the Bayes Theorem.

The theorem is combined with Naive where it is assumed that the conditions between the attributes are mutually independent. Clarification Naive Bayes assumed that the presence or absence of certain characteristics of a class has nothing to do with characteristics.

I. LITERATURES REVIEW

The definition of the C4.5 algorithm was introduced by Quinlan (1996) as an improved version of ID3 [3]. In ID3, decision tree induction can only be done for categorical features (nominal or ordinal), while numeric types (interval or ratio) cannot be used. Improvements made are being able to handle features with numeric types, prunning decision trees, and deriving rule sets. Bayes is a simple probability-based prediction technique that differs from the application of the Bayes theorem (or Bayes rule) with the assumption of strong (naive) independence [3]. In other words, in Naïve Bayes, the model used is an independent feature model.

In Bayes (especially naïve Bayes), the meaning of strong independence on features is that a feature in a data set is not related to the presence or absence of other features in the same data. For example in the case of animal clarification with features of skin covering, childbirth, weight, and breastfeeding.

II. FRAMEWORK

Model-View-Controller (MVC) is a programming model that implements an application architecture into three parts, namely separating the process, views and parts that connect to the database. MVC aims to separate business processes from user interface considerations so that developers can more easily develop one part of the application so that it does not affect other parts [6].

In the MVC model describes information (data) and business processes. The view contains interface elements such as text, images, or 16 input forms, while the controller manages the communication between the view and the model [7]. If mapped an MVC workflow will look like the following picture.

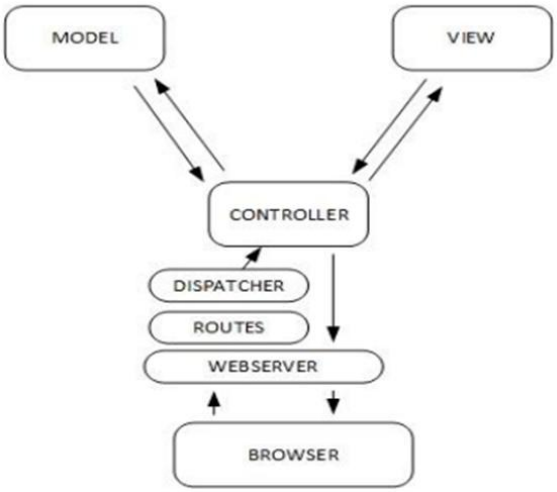

Figure 1: MVC Workflow

The MVC concept describes the Model-View-Controller as follows [8]:

a. Model

A model is a class that underlies the process logic in a software application and the classes associated with it. Model is an object that does not contain information about the user interface. Model is also a class that contains methods / functions and is used to store data and relevant business rules.

b. View

The view view is a collection of classes that represent the elements in the interface, in the view there are names that are used to identify 17 view script files when called via the render function. The view name is the same as the view script file name.

c. Controller

Controller is a class that connects the model and view, used to communicate between classes in the model and view. Controllers have standard actions. When the user request does not specify which action to run, the program executes the standard action.

d. Expert system

Understanding Expert Systems According [6], some definitions of expert systems according to some experts are as follows.

1. According to Durkin: An expert system is a computer program designed to model the problem-solving abilities of an expert.
2. According to Ignizo: Expert system is a model and related procedures, in a particular domain, where the level of expertise can be compared with the expertise of an expert.
3. According to Giarratano and Riley: An expert system is a computer system that can match or imitate the abilities of an expert

III. METHODS

The dataset to be studied is obtained from the uci machine learning repository data provider site: https://archive.ics.uci.edu/ml/datasets/Heart+Disease where the dataset has 3546 records, 13 attributes of these attributes are the result attributes.

Table 1. Dataset From UCI Website

Sex	Age	CSk	CPD	PS	PHyp	Dbts	TCh	SBP	DBP	BMI	HR	TYCHD
1	39	0	0	0	0	0	195	106	70	26,97	80	0
0	46	0	0	0	0	0	250	121	81	28,73	95	0
1	48	1	20	0	0	0	245	127,5	80	25,34	75	0
0	61	1	30	0	1	0	225	150	95	28,58	65	1
0	46	1	23	0	0	0	285	130	84	23,1	85	0
0	43	0	0	0	1	0	228	180	110	30,3	77	0
0	63	0	0	0	0	0	205	138	71	33,11	60	1
0	45	1	20	0	0	0	313	100	71	21,68	79	0
1	52	0	0	0	1	0	260	141,5	89	26,36	76	0
1	43	1	30	0	1	0	225	162	107	23,61	93	0
0	50	0	0	0	0	0	254	133	76	22,91	75	0
0	43	0	0	0	0	0	247	131	88	27,64	72	0

1. Algorithm C4.5

The following is the model used in processing the dataset using the C4.5 algorithm:

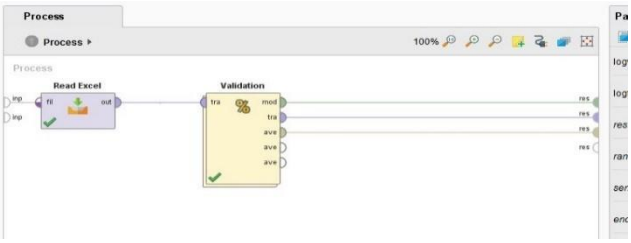

Figure 2: Dataset Processing Model

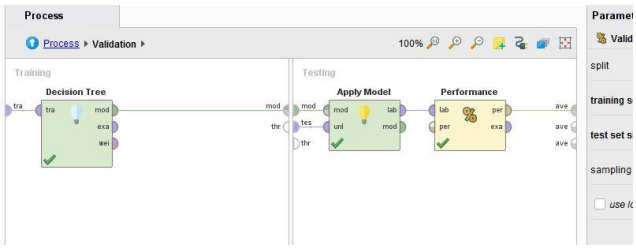

Figure 3: C4.5 Algorithm Processing Process in Rapid Miner Studio

2. Naive Bayes

The following is a model used in dataset processing using the Naive Bayes algorithm:

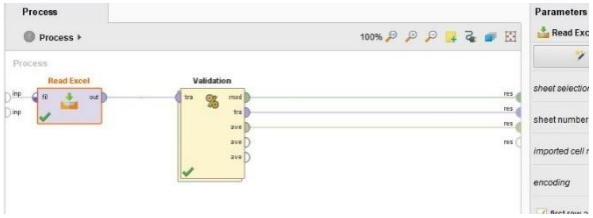

Figure 4: Naive Bayes Algorithm processing model in Rapid Miner Studio

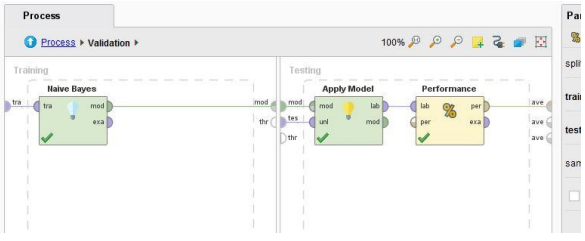

Figure 5: Naive Bayes Algorithm Model Processing in Rapid Miner Studio

The use case of an expert information system using the naïve Bayes algorithm is as follows.

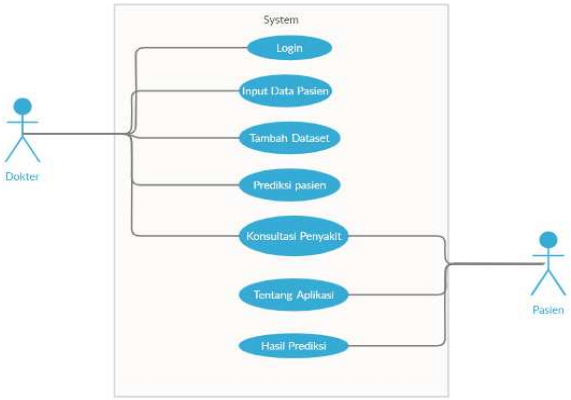

Figure 6: Use case diagram of a heart disease expert system

IV. RESULT

For the diagnosis of heart disease, there are 13 criteria used, which are as follows.

1. Age of the patient
The patient age ranges in the table below are a reference for heart disease.

Table 2. Age Range Table in Heart Disease

No	*Range*	Information
1	0-34	Young
2	35-50	Middle-Aged
3	51-68	Old

2. Gender
The sexes generally recognized throughout the world are male and female.

Table 3. Table Description of Gender

No	*Value*	Information
1	0	Woman
2	1	Man

3. Active Smoker
People who regularly consume the smallest cigarette even though it's only one cigarette a day.

Table 4. Table Description of Active Smokers

No	*Value*	Information
1	0	Passive Smoker
2	1	Active Smoker

4. A day's cigarette stick
People who smoke a certain amount in one day

Table 5. Table of Number of Cigarettes in a Day

No	*Range*	Information
1	0-9	Value 4: without symptoms
2	10-20	Value 3: non-angina pain
3	21-30	Value 2: angina atipikal
4	30-43	Value 1: angina tipikal

5. The usual stroke
People who have a stroke

Table 6. Stroke Information Table

No	*Value*	Information
1	0	Not Have stroke
2	1	Have stroke

6. Sugar levels (> 120 mg / dl)
Sugar level is the level of glucose in the blood. The unit used for sugar content is mg / dl (milligrams per deciliter). Examination of blood sugar levels in venous blood when the patient is fasting 12 hours before the examination (blood sugar satisfied nuchter) or 2 hours after eating (post prandial blood sugar).

Table 7. Information Table of Blood Sugar

No	*Value*	Information
1	0	No
2	1	Yes

7. Diabetes
People who have diabetes

Table 8. Information Table Diabetes

No	*Value*	Information
1	0	No
2	1	Yes

8. Cholesterol
People who have high cholesterol have a major role in triggering the occurrence of blockage of blood vessels that leads to heart attacks, strokes and others.

Table 9. Cholesterol Information Table

No	*Value*	Information
1	0-129	Low
2	130-237	Middle
3	238-464	High

9. Pressure when the Heart Pumps the Whole Body
The pressure in the blood vessels when the heart contracts to pump clean blood around the body.

Table 10. Table Description of Pressure when the Heart Pumps Throughout the Body

No	*Range*	Information
1	0-94	Low
2	95-132	Middle
3	133-212	High

10. Blood Pressure when the Heart Relaxes
Blood pressure when the heart is relaxing or resting. On a heart rate curve, diastolic pressure is the blood pressure represented in the range between the heart rate charts.

Table 11. Table Description of Blood Pressure when the Heart Relaxes

No	*Range*	Information
1	0-57	Low
2	58-81	Middle
3	82-124	High

11. Body Weight BMI
This is calculated by dividing the body weight in kilograms by the height in meters squared. Suppose you have a weight of 75 kg and a height of 1.65m (165cm). Your Body Mass Index or BMI is: BMI = 75kg / (1.65 x 1.65) = 27.55

Table 12. Table Information on Body Weight BMI

No	*Range*	Information
1	0-16	Thin
2	17-25	Ideal
3	26-45	Fat

12. Heartbeat
Heartbeat is the pulse released by the heart and as a result of blood flow through the heart. Doctors usually use a stethoscope when examining a patient to listen for a heartbeat

Table 13. Table Description of Heart Rate

No	*Range*	Information
1	0-48	Slow
2	49-75	Normal
3	76-110	Fast

13. Diagnosis of heart disease
Narrowing and hardening of the arteries due to plaque buildup on the walls of blood vessels. This condition is a common cause of heart disease.

Table 14. Table Description of Diagnosis of Heart Narrowing

No	*Range*	Information
1	0	<50% Narrowing of the diameter
2	1	Narrowing of the diameter 50%

V. DISCUSSION

The model generated from the C4.5 and Naïve Bayes algorithms is tested using cross validation, it can be seen that the Naïve Bayes algorithm has the highest accuracy while the lowest is the C4.5 algorithm.

Table 15. Comparison of Accuracy and AUC Values

	C4.5 Algorithm	*Naïve Bayes Algorithm*
Accuracy	79.07	81.40%
AUC	0.472	0.709

It can be seen that the accuracy value of the Naïve Bayes algorithm has a high value of 81.40% compared to the C4.5 algorithm which only has an accuracy value of 79.07%. Based on the calculation results, it can be concluded that the Naïve Bayes Algorithm is a very good clarification because it has a value between 0.709 - 1.00. The graph of the accuracy value and AUC value of each algorithm is shown in Figure 4.8.

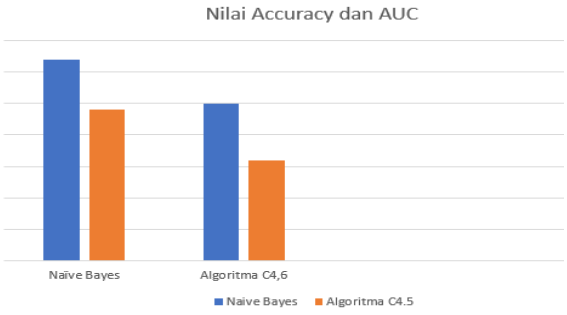

Figure 7: Accuracy and AUC values for each algorithm

Screen Display Design

In this screen display design the author will provide an image of the application and its functions.

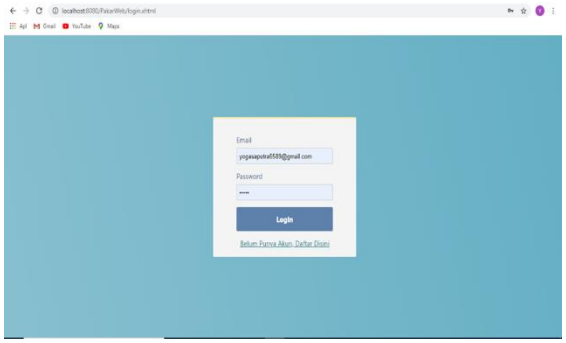

Figure 8: Login Menu Display Design

Login is the initial display when the application will be run / used

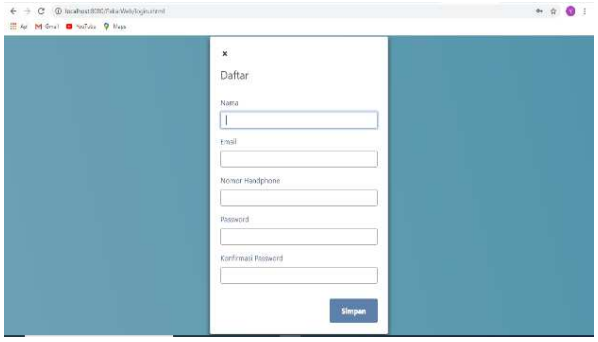

Figure 9. Display Login List

Register is when the doctor wants to create or register a new ID and password to log into the application.

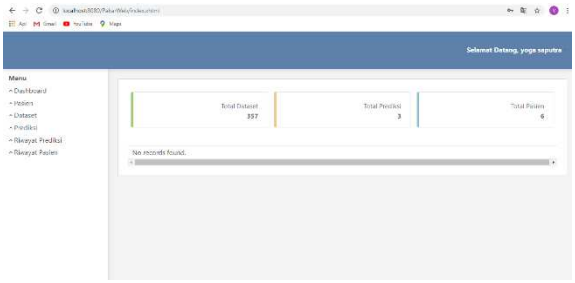

Figure 10. Dashboard view

Dashboard is the initial display when doctors enter and use the application.

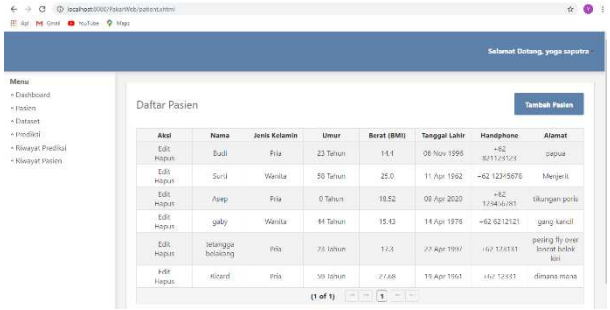

Figure 11. Patient Views

VI. CONCLUSION

Based on the results of research conducted by researchers with the title Comparison of Data Mining Methods Using the C.45 Algorithm and Naive Bayes in Predicting Heart Disease, the following conclusions can be drawn:

1. Expert system for Comparison of Data Mining Methods Using the C.45 Algorithm and Naive Bayes in Predicting Heart Disease can be built using the Naïve Bayes algorithm as an infrastructure machine in calculating the possibility of a patient suffering from heart disease.
2. The expert system using the Naïve Bayes method has an accuracy of 83.17% while the C4.5 algorithm only has an accuracy of 81.77%, so the Naïve Bayes algorithm has a superior accuracy value.
3. The Naïve Bayes algorithm has a higher accuracy value than the C4.5 algorithm so that researchers use the Naïve Bayes algorithm in predicting heart disease with 3546 records of training data, and 13 attribute data obtained from the UCI Repository.

The expert system uses the Naïve Bayes method to diagnose heart disease which can be used to diagnose new patients with results in the form of a percentage of positive and negative probability values and is proven by the calculation results of a program that has been built in the Java programming language.

REFERENCES

[1] Brunner & Suddarth, 2013. Word Health Organization 2013.
[2] Dr. Suyanto, S. M. (2017). Data Mining Untuk Klasifikasi Dan Klasterisasi Data.bandung: Informatika
[3] Retno Tri Vulandari, 2017, Data Mining Teori dan Aplikasi Rapid Miner, Yogyakarta
[4] Romney, Marshall B., dan Paul John Steinbart. 2015. Accounting informasi systems,13Thed. England: Pearson Educational Limited.
[5] Suryana, Taryana dan Koesheryatin. 2014. Aplikasi Internet Menggunakan HTML, CSS, & JavaScript. Jakarta: PT Elex Media Komputindo.
[6] Gellinas, U. J., Dull, R.B. (2012). Accounting information systems, 9th ed. USA: South-Western Cengage Learning
[7] Abdul Kadir. 2014. Pengenalan Sistem Informasi Edisi Revisi. Andi.Yogyakarta
[8] Alexander F. K. Sibero, 2011, Kitab Suci Web Programing, MediaKom,Yogyakarta

BIOGRAPHY

Rino, received his Bachelor of Informatics Engineering (S.Kom) from STMIK Buddhi, Indonesia in 2008 and his Master of Computer Science (M.Kom) concentration in Software Engineering from STMIK Eresha, Indonesia in 2012. He is a lecturer in the Engineering Study Program Informatics, Faculty of Science & Technology, Buddhi Dharma University.