JUSIKOM PRIMA (Jurnal Sistem Informasi dan Ilmu Komputer Prima)
Vol. 7 No. 2, February 2024 E-ISSN:2580-2879

APPLICATION OF DATA MINING USING THE RANDOM FOREST METHOD TO PREDICT HEART DISEASE

Felix[1], Delima Sitanggang[2], Yonata Laia[3], Amalia[4], Muhammad Radhi[5]
Prima Indonesia University
felixsen17@gmail.com , delima@unprimdn.ac.id , yonata@unprimdn.ac.id , amalia@unprimdn.ac.id , muhammadradhi@unprimdn.ac.id

ABSTRACT

A heart attack is when fatty deposits block the arteries. This causes symptoms such as shortness of breath and chest pain. In addition, obstructed blood flow to the heart can cause damage to the heart muscle. Heart attacks are still the highest cause of death in Indonesia to date. The problem today is that it is tough to predict and identify heart disease. The appropriate method needed to predict heart disease is the Random Forest method. This research aims to calculate the level of accuracy in predicting heart attacks. Based on research and data processing carried out by previous study by comparing two K-Neighbor algorithms, which produced an accuracy value of 83% and the Logistic Regression algorithm produced an accuracy value of 88% and it was found that the Random Forest algorithm had an accuracy of 86.88%. Thus, other algorithms are better at predicting heart attacks than the Random Forest algorithm.

Keywords: Heart Attack, Random Forest, Prediction.

INTRODUCTION

Heart attack is a condition where the arteries are blocked due to fat deposits, according to data from the World Health Organization (WHO). This disease causes symptoms such as shortness of breath and chest pain[1]. This is also caused by obstructed blood flow to the heart and can destroy the heart muscle. Until now, heart attacks are still the highest cause of death in Indonesia [2]. This is caused by narrowing the blood vessels, viruses or bacteria, heart valve abnormalities, unhealthy lifestyles, and side effects of certain medications.

The heart is a muscle that is divided into four chambers. At the top are two atria, the right atrium, and the left atrium. Below are two other ventricles: the right and left ventricles of the heart [3]. Between the left and right chambers is a muscular wall (septum) that prevents oxygenated blood from mixing with deoxygenated blood. The heart's primary function is to circulate oxygen-rich blood to all body parts [4]. When all the organs in the body run out of oxygen, the deoxygenated blood returns to the heart and then to the lungs filled with oxygen [5].

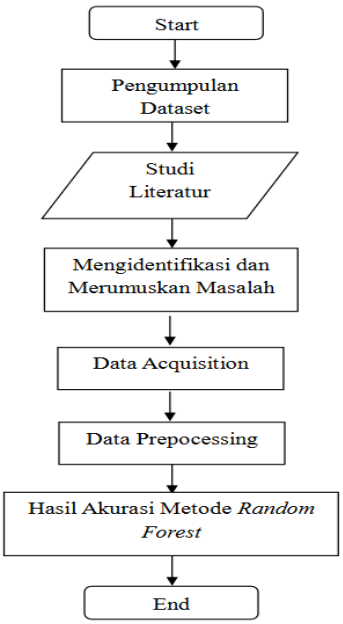
The problem currently faced is that it is tough to predict heart disease and determine whether someone suffers from heart disease. Appropriate methods are needed to predict heart disease [6]. According to Harvard Health Publishing, heart disease is often considered a disease that attacks suddenly [7]. This disease appears when plaque builds up over the years and blocks the heart arteries. Blockage of blood vessels in the heart in the form of a buildup of fat, cholesterol, and plaque in the arteries, where the plaque breaks and forms blood clots, disrupting blood circulation and damaging the heart muscle [8].

Based on research carried out by Ade Riani (2019), Implementation of Data Mining to Predict Heart Disease Using the Naive Bayes Method, This research carried out calculations using the Naive Bayes algorithm, which produced an accuracy value of 86% so that more accurate values are still needed in predicting heart disease using other methods [9]. Research conducted by Nicholas (2022), Implementation of Data Mining to Predict Heart Disease Using the K-Nearest Neighbor Method and Logistic Regression. This research carried out calculations using the K-Nearest Neighbor algorithm, which produced an accuracy value of 83%, and the Logistic Regression algorithm, which produced an accuracy value of 88%

In this study, it was found that the Random Forest algorithm had an accuracy of 86.88%.

METHODS

The stages in completing the research can be seen in the following picture:



Picture1. Flow Stages Study

1. Dataset Collection

In this study, disease data was obtained from a heart attack (Heart Attack) using data via Kaggle. This data will be processed and become an essential source of information in this research.

2. Study of literature

Collection: Previous research journals and books related to the research objectives and the research title, namely predicting heart attacks [10].

3. Identifying and Formulating Problems

At this stage, it is essential to increase the search for information sources about heart disease related to the causes of the problems experienced by sufferers.

4. Data Acquisition

The heart attack dataset was taken via Kaggle data. This dataset will be processed through a data processing process to identify the leading causes of heart attacks [11].

5. Data Preprocessing

The process of converting raw data into an easy-to-understand format. This process is carried out because raw data is often irregular. The aim is to be used as a source of information through data collection that can be forwarded for data processing [12].

6. Random Forest Method Accuracy Results

This research will obtain a comparison of accuracy values between the Random Forest method with previous research in predicting heart attacks.

RESULTS AND DISCUSSION

Problem analysis

Heart disease is a condition where the heart stops working. There are various disorders, such as disorders of the blood vessels, heart valves, and heart muscle. This disease can also be caused by infection or congenital disabilities [13]. The common cause of this disease is blockage, inflammation, or damage to the heart and surrounding blood vessels. This condition blocks blood flow to nature, reducing the circulation of oxygen and nutrition to the muscles and

tissues around the nucleus [14]. The high death rate due to this disease is caused by a lack of public awareness of the need for regular heart health checks and people's poor lifestyle patterns.

Data analysis

The dataset used in this research was obtained from Kaggle as the data source. Heart attack data is used as sample data for data processing.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

Figure 2. Heart Disease Dataset

The total data that will be processed consists of 303 samples with 14 attributes that will be processed data, and there are also 2 labels, which include 0 = less likely to have a heart attack, 1 = more likely to experience heart disease. The following is an explanation of the dataset attributes:

Table 1. Explanation of Research AttributesHeart attack

No.	Name Field	Information
1	Age	Patient Age.
2	Sex	Patient Gender (1 = Male, 0 = Female).

3	Cp	Types of Chest Pain (typical angina, atypical, non-anginal, and asymptomatic). (0 = General symptoms of chest pain with the possibility of blockage of the coronary arteries are Typical Angina. 1 = Symptoms are not detailed, and the case of blockage is lower in Atypical Angina. 2 = Sharp feeling and long or short-term pain is Non-anginal pain. 3 = Not showing symptoms of the disease is Asymptomatic)
4	Trestbps	In units of mm/Hg, the patient's blood pressure when at rest.
5	Chol	Serum cholesterol in mg/dl.
6	Facebook	(Fasting Blood Sugar) The amount of blood sugar in mg/dl, greater or less than 120mg/dl. (0 = less than 120mg/dl, 1 = more than 120mg/dl).
7	Restecg	Resting electrocardiography results. (0 = Normal, 1 = ST waves increase / decrease more than 0.5 mV, 2 = Left ventricular hypertrophy).
8	Thalach	Maximum Heart Rate.
9	Exang	Chest pain due to exercise (0 = no pain, 1 = pain).
10	Oldpeak	ST segment size from exercise relative to resting conditions.
11	Slopes	The magnitude of the ST segment slope at peak or maximum exercise conditions. (0 = *downsloping*, 1 = flat, 2 = upsloping).
12	Ca	Number of major blood vessels (0-3).
13	Thal	StatusThe heart is divided into 4, 0 = unknown, 1 = permanent defect, 2 = normal, 3 = reversible defect.
14	Target	Indicated heart attack or not. (0 = less likely to have a heart attack, 1 = more likely to have a heart attack).

Random Forest Flowchart

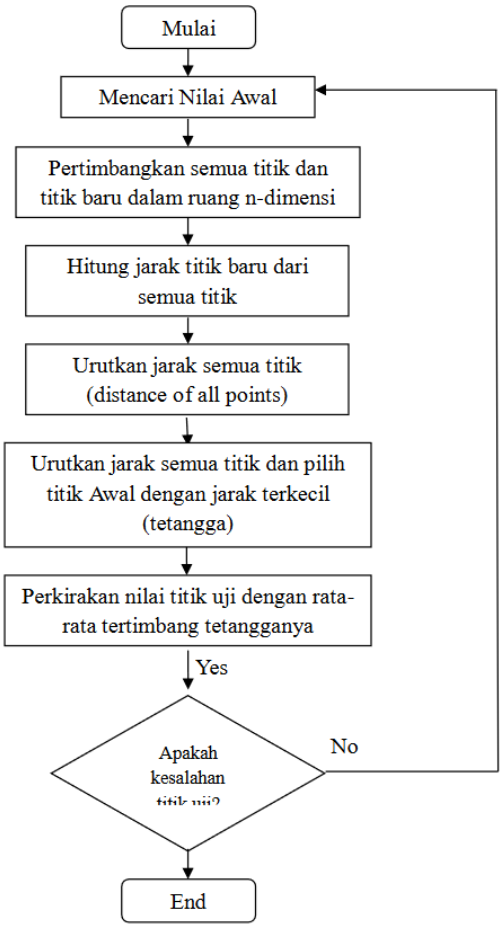


Figure 3. Random Forest Flowchart

Data processing

Heart disease is the highest cause of death that can occur at any age. The data will be processed using the Random Forest method as a research method to produce accurate values. The following data processing processes are carried out as follows:

Import Libraries

The following are the initial steps in the import library data processing process:

Source Code Import Library:

```
import seaborn as sns
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, accuracy_score
import warnings
warnings.filterwarnings("ignore", category=DeprecationWarning)
import os
for dirname, _, filenames in os.walk('/content'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

After the library import process is carried out, then carry out the process of checking the number of datasets:

```
print('Number of rows are',heart.shape[0], 'and number of columns are ',heart.shape[1])
```

```
Number of rows are 302 and number of columns are  14
```

```
heart.describe()
```

	count	mean	std	min	25%	50%	75%	max
age	303.0	54.366337	9.082101	29.0	47.5	55.0	61.0	77.0
sex	303.0	0.683168	0.466011	0.0	0.0	1.0	1.0	1.0
cp	303.0	0.966997	1.032052	0.0	0.0	1.0	2.0	3.0
trtbps	303.0	131.623762	17.538143	94.0	120.0	130.0	140.0	200.0
chol	303.0	246.264026	51.830751	126.0	211.0	240.0	274.5	564.0
fbs	303.0	0.148515	0.356198	0.0	0.0	0.0	0.0	1.0
restecg	303.0	0.528053	0.525860	0.0	0.0	1.0	1.0	2.0
thalachh	303.0	149.646865	22.905161	71.0	133.5	153.0	166.0	202.0
exng	303.0	0.326733	0.469794	0.0	0.0	0.0	1.0	1.0
oldpeak	303.0	1.039604	1.161075	0.0	0.0	0.8	1.6	6.2
slp	303.0	1.399340	0.616226	0.0	1.0	1.0	2.0	2.0
caa	303.0	0.729373	1.022606	0.0	0.0	0.0	1.0	4.0
thall	303.0	2.313531	0.612277	0.0	2.0	2.0	3.0	3.0
output	303.0	0.544554	0.498835	0.0	0.0	1.0	1.0	1.0

```
heart.isnull().sum()/len(heart)*100
```

```
age         0.0
sex         0.0
cp          0.0
trtbps      0.0
chol        0.0
fbs         0.0
restecg     0.0
thalachh    0.0
exng        0.0
oldpeak     0.0
slp         0.0
caa         0.0
thall       0.0
output      0.0
dtype: float64
```

Figure 4. Initialize Dataset Header

Checking For Data Types of The Attributes

Separating the columns into categorical and continuous

The following process separates the columns into categorical and continuous forms to make the data processing process more accessible.

```
cat_cols = ['sex','exng','caa','cp','fbs','restecg','slp','thall']
con_cols = ["age","trtbps","chol","thalachh","oldpeak"]
target_col = ["output"]
print("The categorial cols are : ", cat_cols)
print("The continuous cols are : ", con_cols)
print("The target variable is :  ", target_col)
```

```
The categorial cols are :  ['sex', 'exng', 'caa', 'cp', 'fbs', 'restecg', 'slp', 'thall']
The continuous cols are :  ['age', 'trtbps', 'chol', 'thalachh', 'oldpeak']
The target variable is :   ['output']
```

Data Visualization

This data visualization aims to transform a collection of datasets into a more complex or straightforward form that is easy to understand and analyze. In this data visualization, visualizations related to each other are carried out. Here is the data visualization source code:

```
heart.head().style.background_gradient(cmap='PuRd').hide_index().set_properties(**{'font-family': 'Segoe UI'})
```

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.300000	0	0	1	1
37	1	2	130	250	0	1	187	0	3.500000	0	0	2	1
41	0	1	130	204	0	0	172	0	1.400000	2	0	2	1
56	1	1	120	236	0	1	178	0	0.800000	2	0	2	1
57	0	0	120	354	0	1	163	1	0.600000	2	0	2	1

Figure 4. Coloring table for heart disease

Judging from the visualized table, the dark red dramatically influences heart disease. The following colors, namely pink, purple, and light purple, can potentially cause heart disease.

Data Preprocessing

The process of converting raw data becomes more straightforward to understand. This process is carried out because the raw data often needs to be more appropriate. The goal is to become a source of information with a collection of data that can be processed.

```
x = heart.iloc[:, 1:-1].values
y = heart.iloc[:, -1].values
x,y
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size= 0.2, random_state= 0)
print('Shape for training data', x_train.shape, y_train.shape)
print('Shape for testing data', x_test.shape, y_test.shape)
scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
x_train,x_test
```

Random Forest

Apart from using Random Forest, This research also uses Random Forest as an algorithm for predicting heart attacks. Here is the Random Forest source code:

```
rf = RandomForestClassifier()
rf.fit(x_train, y_train)
predicted = rf.predict(x_test)
print("The accuracy of RF is : ", accuracy_score(y_test, predicted)*100, "%")
```

CONCLUSION

The method used in this research is the Random Forest algorithm, which aims to predict heart attack symptoms. The data processing process is then compared with the algorithm accuracy results, namely the algorithm from the

previous research

with Random Forest. The following are the results of the comparison of accuracy between algorithms:

Random Forest

Source Code Random Forest:

```
rf = RandomForestClassifier()
rf.fit(x_train, y_train)
predicted = rf.predict(x_test)
print("The accuracy of RF is : ", accuracy_score(y_test, predicted)*100, "%")
```

```
The accuracy of RF is :  86.88524590163934 %
```

From the comparison accuracy results between the two Random Forest algorithms, the one that produces the best accuracy value is the Logistic Regression algorithm from previous research. In contrast, in earlier research, the Random Forest produces an accuracy value of 86.88%,

which is a higher accuracy value than the K-Nearest Neighbor method. In this case, in predicting heart attacks, the accuracy level of the Logistic Regression algorithm is better than that of the Random Forest.

LIMITATIONS

In this research, the data processing process on the dataset used was tiny. So, for further investigation, it is necessary to add more datasets to be used and use more profound algorithms or other Deep Learning algorithms, which are more effective than using Machine Learning alone.

REFERENCES

Supriyatna, HA, Away, Y., & Zulhelmi, Z. "Internet of Things (IoT) system design for monitoring and predicting heart attack symptoms." Computer Journal, Technology Information, and Electrical Engineering, vol. 4, no. 1, 2019.

Dhany, HW "Performance of the K-Nearest Neighbor Algorithm in Predicting Heart Disease." In National Seminar on Informatics (SENATIKA), pp. 176-179, 2021, June.

Al Azhima, SAT, Darmawan, D., Hakim, NFA, Kustiawan, I., Al Qibtiya, M., & Syafei, NS "Hybrid Machine Learning Model to predict Heart Disease using Logistic Regression and Random Forest Methods." Journal of Integrated Technology, vol. 8, no. 1, pp. 40-46, 2022.

Annisa, R. "Comparative Analysis of Data Mining Classification Algorithms for Predicting Heart Disease Sufferers." JTIK (Kaputama Information Engineering Journal), vol. 3, no. 1, pp. 22-28, 2019.

Putra, PD, & Rini, DP "Prediction of Heart Disease with Classification Algorithms." In Annual Research Seminar (ARS), vol. 5, no. 1, pp. 95-99, 2020, February.

Pangaribuan, JJ, Tanjaya, H., & Kenichi, K. "Detecting Heart Disease Using Machine Learning with the Logistic Regression Algorithm." Journal of Information System Development (ISD), vol. 6, no. 2, pp. 1-10, 2021.

Nawawi, HM, Purnama, J.J., & Hikmah, AB "Comparison of Neural Network and Naive Bayes Algorithms to Predict Heart Disease." Pilar Nusa Mandiri Journal, vol. 15, no. 2, pp. 189-194, 2019.

Gunawan, MI, Sugiarto, D., & Mardianto, I. "Improving the Accuracy Performance of

Diabetes Mellitus Prediction Using the Grid Search Method in the Logistic Regression Algorithm." JEPIN (Journal of Informatics Education and Research), vol. 6, no. 3, pp. 280- 284, 2020.

Achmad, AI "Binary Probit Regression Method for Modeling Factors that Influence the Diagnosis of Heart Disease." Journal of Statistical Research, pp. 27-33, 2022.

Rahayu, S., Subekhi, A., Astuti, D., Widaningsih, I., Sartika, I., Nurhayani, N., ... & Rafidah, R. "EFFORT TO AWARE OF HEART ATTACKS THROUGH HEALTH EDUCATION." JMM (Independent Community Journal), vol. 4, no. 2, pp. 163-171, 2020.

Majid, AM, & Miharja, MND "APPLICATION OF DISCRETIZATION AND METHODADABOOST TO IMPROVE THE ACCURACY OF THE CLASSIFICATION ALGORITHM

IN PREDICTING HEART DISEASE." Indonesian Journal of Business Intelligence (IJUBI), vol. 5, no. 2, pp. 70-75, 2022

[1] Riani, A., Susianto, Y., & Rahman, N. "Implementation of Data Mining to Predict Heart Disease Using the Naive Bayes Method." Journal of Innovation Information Technology and Application (JINITA), vol. 1, no. 1, pp. 25-34, 2019.

[2] Andiani, L., Sukemi, S., & Rini, DP "Heart Disease Analysis Using KNN and Random Forest Methods." In Annual Research Seminar (ARS), vol. 5, no. 1, pp. 165- 169, 2020, February.

[3] Karyatin, K. "Factors Associated with the Incidence of Coronary Heart Disease." Health Scientific Journal, vol. 11, no. 1, pp. 37-43.

[4] Bianco, MA, Kusrini, K., & Sudarmawan, S."Designing a Heart Disease Classification System Using Naïve Bayes." Creative Information Technology Journal, vol. 6, no. 1, pp. 75-83, 2020.

[5] Qomariyah, N., Hamzah, N., & Mustika, WP "APPLICATION OF ARTIFICIAL METHODSNEURAL NETWORK FOR DETECTING HEART ATTACKS AT AWAL BROS BEKASI HOSPITAL." INTI Nusa Mandiri, vol. 13, no. 1, pp. 27-32, 2019.

[6] Apriyatmoko, R., & Aini, F. "Adolescents Recognize Coronary Heart Attacks." INDONESIAN JOURNAL OF COMMUNITY EMPOWERMENT (IJCE), vol. 2, no. 2, 2020.

[7] Derisma, D. "Comparison of Algorithm Performance for Heart Disease Prediction Using Data Mining Techniques." Journal of Applied Informatics and Computing, vol. 4, no. 1, pp. 84-88, 2020.

[8] Utomo, DP, & Mesran, M. "Comparative Analysis of Data Mining Classification Methods and Attribute Reduction on Heart Disease Data Sets." Budidarma Media Informatics Journal, vol. 4, no. 2, pp. 437-444, 2020.

[9] Aripin, HA "OUTCOME PREDICTION FOR HEART DISEASE WITH ARTIFICIAL NEURAL NETWORK ALGORITHM." INFOKOM (Informatics & Computers), vol. 9, no. 1, pp. 30-45, 2021.

[10] Sitanggang, D., Nicholas, Wilson, V., Sinaga, ARA, and Simanjuntak, AD "IMPLEMENTATION OF DATA MINING TO PREDICTED HEART DISEASE USING THE K-NEAREST NEIGHBOR AND LOGISTIC REGRESSION METHOD" Tekinkom Journal (Information and Computer Engineering), vol 5, no. 2, pp. 493-501,2022