JOIN (Jurnal Online Informatika)
Volume 7 No. 2 | December 2022: 219-225
DOI: 10.15575/join.v7i2.919
p-ISSN: 2528-1682
e-ISSN: 2527-9165

Comparative Analysis of Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) Algorithms for Classification of Heart Disease Patients

Aina Damayunita[1], Rifqi Syamsul Fuadi[2], Christina Juliane[3]
[1,2,3]Department of Information Systems, STMIK LIKMI, Bandung, Indonesia

Article Info

Article history:

Received September 07, 2022
Revised December 05, 2022
Accepted December 06, 2022
Published December 26, 2022

Keywords:

Classification algorithms
Heart disease
K-Nearest Neighbors (KNN)
Naive Bayes
Support Vector Machine (SVM)

ABSTRACT

Heart disease is still the leading cause of death. In this study, we tried to test several factors that can identify patients with heart disease using 3 classification algorithms: Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). This study conducts a comparison of algorithms to choose which one invents high accuracy in classifying, analyzing, and obtaining confusion matrix values along with the accuracy of predicting heart disease based on several factors or other comorbidities that the patient has, ranging from BMI to the patient's skin cancer status. From the results of trials conducted by the SVM algorithm, it has the highest accuracy value, which is 92% while the Naive Bayes algorithm is the lowest with an accuracy value of 88%.

Corresponding Author:

Aina Damayunita,
Department of Information Systems
STMIK LIKMI Bandung
Jl. Ir. H. Juanda No.96, Lebakgede, Kota Bandung, Jawa Barat 40132, Indonesia
Email: aina.damayy@gmail.com

1. INTRODUCTION

Heart disease remains the leading cause of death in the United States today [1]. Cardiovascular disease is the leading cause of death worldwide, killing approximately 17.9 million people each year. A cardiovascular disease is a group of heart and blood vessel disorders, including coronary heart disease, cerebrovascular disease, rheumatic heart disease, and others [2]. To predict heart disease, several tests are needed. A lack of expertise in medical staff can result in incorrect predictions [3]. Early diagnosis of the disease becomes very difficult. Challenges to the surgical treatment of heart disease are increasingly complex, especially in developing countries that lack trained medical personnel and other resources necessary for diagnosing and treating patients with heart problems [4]. Evaluation based on an accurate prediction of the risk of heart failure will go a long way in helping patients prevent severe heart attacks and increasing the percentage of patient safety figures [5]. One effective way to identify and predict heart disease is to utilize machine learning algorithms [6]. Early detection of heart disease is necessary once to save many lives. Realizing the usefulness of data mining to help diagnose heart disease is very important.

The application of data mining in the health industry is a new model and is widely used; data mining can be extracted and a hidden pattern of the data can be found that can be used as a decision support [7]. The data mining classification technique is one of the data mining methods whose purpose is to provide predictions of related variables [8]. The high mortality factor due to heart disease can be prevented and suppressed risk factors. Lack of public knowledge about the symptoms of heart disease. The lack of accuracy of the equipment used if it only controls blood sugar and blood pressure, and an unhealthy lifestyle. Laboratory data that has not been effectively functioned can be used to detect heart disease.

Machine learning algorithms and techniques have been applied to various medical data sets to automate the analysis of large and complex data. Machine learning techniques are widely used in research to

p-ISSN: 2528-1682
e-ISSN: 2527-9165

help the medical profession diagnose diseases, including heart disease. Pahwa et al. (2017) used Naive Bayes to predict heart disease [9]. Pouriyeh et al. (2017) used K-Nearest Neighbor to predict the heart disease [10]. Meanwhile, This study conducts a comparison of algorithms to choose which one invents high accuracy in classifying, analyzing, and obtaining confusion matrix values along with the accuracy of predicting heart disease based on several factors or other comorbidities that the patient has, ranging from BMI to the patient's skin cancer status.[11].

The results of this study are expected to help subsequent researchers design and build a coronary heart disease prediction system that can produce diagnostic results that do not affect the diagnosis made by experts. Prior to the gold standard method, medical staff could use the advice in this study to avoid invasive, risky and uneconomical diagnoses of heart disease.

2. METHOD

The problem-solving approach in this research is as follows:

a. Understanding the issue
b. Creating a plan to solve the problem
c. Implementing the project in phase 2
d. Double-checking the results obtained by the problem-solving method:
    1) Data Understanding;
    2) Data Preparation;
    3) Modeling;
    4) Evaluation;
    5) Performance measures: accuracy, precision, recall

Steps for the algorithm implementation in this study are presented in figure 1 below.

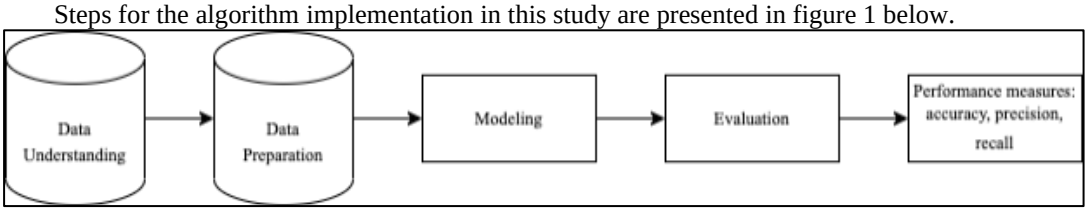

Figure 1. *Stages of Research*

2.1. Classification

Classification is the process of finding a model or function that describes or distinguishes concepts or classes in data in order to infer the type of an object whose label is unknown [12]. The model can be an "if-then" rule in the form of a decision tree, a mathematical formula, or a neural network. Classification methods include C4.5, Random Forest, Naïve Bayesian, neural networks, genetic algorithms, fuzzy, case-based reasoning, and K-Nearest Neighbor [13]. Classification is a process aimed at placing objects into a particular class or category. Data or document classification can also start by creating specific classification rules using training data. This is often called the learning phase, and the tests are used as data tests. The classification process in this study uses the “Google Colab” application; the main purpose of this study is to find the highest accuracy value in the classification algorithm used to predict heart disease accompanied by other disease factors.

2.2. Naïve Bayes

The Naive Bayes algorithm is known as Bayes' theorem because it predicts future odds based on past experience. The main feature of this naive Bayes classifier is the strong assumption (simple) of independence from each condition/event [14].

The advantage of using this method is that only a small amount of training data is required to determine the estimated parameters required for the classification process. Since this is assumed to be the independent variable, only the variance of one variable within the class is required to determine the classification, not the entire covariance matrix [14].

2.3. K-Nearest Neighbors (KNN)

The KNN algorithm is a method that uses a supervised algorithm. The difference between supervised learning and unsupervised learning is that supervised learning aims to find new patterns in data by connecting

existing data patterns with new data [12]. In unsupervised learning, data does not yet have any patterns, and the purpose of unsupervised learning is to find patterns in data. The KNN algorithm uses neighborly classification as the predicted value of the new test sample.

2.4. Support Vector Machine (SVM)

SVM is one of the classification methods in data mining. SVM can also make predictions on both classification and regression [15]. SVM has a linear principle, but now SVM has developed to work on non-linear problems [16]. The way SVM works on non-linear problems is to include the concept of the kernel in a high-dimensional space [17]. In this dimensioned space, a separator or what is often called a hyperplane will be sought. Hyperplanes can maximize distances or margins between data classes. The best hyperplane between the two classes can be found by measuring the margin and then looking for its maximum point. The effort to find the best hyperplane as a class separator is at the core of the process in the SVM method [18].

3. RESULTS AND DISCUSSION

3.1 Dataset

Table 1. shows a brief description of the dataset attributes used in this study. This dataset is derived from Kaggle, CDC (Center for Disease Control) data [19]. This dataset has 18 dimensions and 320000 data from medical records. The existing dimensions are as follows:

Table 1. Attribute Description

No	Nama Atribut	Deskripsi
1	HeartDisease	represents whether the patient has heart disease (yes/no)
2	BMI	represents the magnitude of the BMI number (float)
3	Smoking	represents whether the patient is a smoker (yes/no)
4	AlcoholDrinking	represents whether the patient consumes alcohol (yes/no)
5	Stroke	represents whether the patient has suffered a stroke (yes/no)
6	PhysicalHealth	represents the value of physical health (float)
7	MentalHealth	represents the value of mental health (float)
8	DiffWalking	represents whether the patient has difficulty walking (yes/no)
9	Sex	represents the patient's gender (female/male)
10	AgeCategory	represents the age category of the patient (per 10 years, 80 and above)
11	Race	represents the patient's race (white, black, Latin, etc.)
12	Diabetic	represents whether the patient has diabetes (yes/no)
13	PhysicalActivity	represents whether the patient is doing physical activity (yes/no)
14	GenHealth	represents the patient's health in general (very bad - very good)
15	SleepTime	represents how long the patient sleeps each day (float)
16	Asthma	represents whether the patient has asthma (yes/no)
17	KidneyDisease	represents whether the patient has kidney disease (yes/no)
18	SkinCancer	represents whether the patient has skin cancer (yes/no)

3.2 Data Preparation

Things to do in data preparation are to load data, sort the data, and data integrate. In this study, the data taken came from a data column of the float data type. The number of rows in this dataset is 320000. In this study only the first 200000 data were taken.

```
df.head(21)
```

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes	Yes	Very good	5.0	Yes	No	Yes
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No	Yes	Very good	7.0	No	No	No
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes	Yes	Fair	8.0	Yes	No	No
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No	No	Good	6.0	No	No	Yes
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No	Yes	Very good	8.0	No	No	No
5	Yes	28.87	Yes	No	No	6.0	0.0	Yes	Female	75-79	Black	No	No	Fair	12.0	No	No	No
6	No	21.63	No	No	No	15.0	0.0	No	Female	70-74	White	No	Yes	Fair	4.0	Yes	No	Yes
7	No	31.64	Yes	No	No	5.0	0.0	Yes	Female	80 or older	White	Yes	No	Good	9.0	Yes	No	No
8	No	26.45	No	No	No	0.0	0.0	No	Female	80 or older	White	No, borderline diabetes	No	Fair	5.0	No	Yes	No
9	No	40.69	No	No	No	0.0	0.0	Yes	Male	65-69	White	No	Yes	Good	10.0	No	No	No
10	Yes	34.30	Yes	No	No	30.0	0.0	Yes	Male	60-64	White	Yes	No	Poor	15.0	Yes	No	No
11	No	28.71	Yes	No	No	0.0	0.0	No	Female	55-59	White	No	Yes	Very good	5.0	No	No	No
12	No	28.37	Yes	No	No	0.0	0.0	Yes	Male	75-79	White	Yes	Yes	Very good	8.0	No	No	No
13	No	28.15	No	No	No	7.0	0.0	Yes	Female	80 or older	White	No	No	Good	7.0	No	No	No
14	No	29.29	Yes	No	No	0.0	30.0	Yes	Female	60-64	White	No	No	Good	5.0	No	No	No
15	No	29.18	No	No	No	1.0	0.0	No	Female	50-54	White	No	Yes	Very good	6.0	No	No	No
16	No	26.26	No	No	No	5.0	2.0	No	Female	70-74	White	No	No	Very good	10.0	No	No	No
17	No	22.59	Yes	No	No	0.0	30.0	Yes	Male	70-74	White	No, borderline diabetes	Yes	Good	8.0	No	No	No
18	No	29.86	Yes	No	No	0.0	0.0	Yes	Female	75-79	Black	Yes	No	Fair	5.0	No	Yes	No
19	No	18.13	No	No	No	0.0	0.0	No	Male	80 or older	White	No	Yes	Excellent	8.0	No	No	Yes
20	No	21.16	No	No	No	0.0	0.0	No	Female	80 or older	Black	No, borderline diabetes	No	Good	8.0	No	No	No

Figure 2. Dataset fill that is being used

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Check the contents of the fields, making sure that nothing is *null*.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   HeartDisease      200000 non-null  object
 1   BMI               200000 non-null  float64
 2   Smoking           200000 non-null  object
 3   AlcoholDrinking   200000 non-null  object
 4   Stroke            200000 non-null  object
 5   PhysicalHealth    200000 non-null  float64
 6   MentalHealth      200000 non-null  float64
 7   DiffWalking       200000 non-null  object
 8   Sex               200000 non-null  object
 9   AgeCategory       200000 non-null  object
 10  Race              200000 non-null  object
 11  Diabetic          200000 non-null  object
 12  PhysicalActivity  200000 non-null  object
 13  GenHealth         200000 non-null  object
 14  SleepTime         200000 non-null  float64
 15  Asthma            200000 non-null  object
 16  KidneyDisease     200000 non-null  object
 17  SkinCancer        200000 non-null  object
dtypes: float64(4), object(14)
memory usage: 27.5+ MB
```

Figure 3. table info and data type of dataset

Set independent variable yaitu BMI, PhysicalHealth, MentalHealth, dan SleepTime `X = df.iloc[:, [1,5,6,14]]`

Set dependent variable `y = df['HeartDisease']`

X

	BMI	PhysicalHealth	MentalHealth	SleepTime
0	16.60	3.0	30.0	5.0
1	20.34	0.0	0.0	7.0
2	26.58	20.0	30.0	8.0
3	24.21	0.0	0.0	6.0
4	23.71	28.0	0.0	8.0
...	...	...	...	...
149995	28.32	0.0	10.0	9.0
149996	26.13	0.0	0.0	6.0
149997	36.02	0.0	0.0	7.0
149998	33.64	29.0	3.0	8.0
149999	21.59	0.0	3.0	6.0

150000 rows × 4 columns

Figure 4. Set independent variable

Some contents are not on the same scale as BMI and SleepTime so it needs to be created on the same scale as StandardScaler().

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaler.fit(X)
X = scaler.transform(X)

X

array([[-1.83371293, -0.04400324,  3.28853757, -1.45126186],
       [-1.24527113, -0.4236761 , -0.49337388, -0.07330719],
       [-0.26348587,  2.10747629,  3.28853757,  0.61567015],
       ...,
       [ 1.221779  , -0.4236761 , -0.49337388, -0.07330719],
       [ 0.84731604,  3.24649487, -0.11518274,  0.61567015],
       [-1.0485994 , -0.4236761 , -0.11518274, -0.76228452]])
```

Figure 5. Standardize the value of the dataset

3.3. Confusion Matrix

The calculation of classification performance evaluation is carried out with the Confusion Matrix. Confusion Matrix is a performance measurement for machine learning classification problems where the output can be two or more classes [20]. Confusion Matrix is a table with 4 different combinations of predicted values and actual values. Four terms are representations of the results of the classification process in the confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

```
array([[42930,  2889],
       [ 3336,   845]])
```

Figure 6. results of Confusion Matrix in Naive Bayes algorithm

Figure 6 shows the results of the Confusion matrix using the Naïve Bayes algorithm, TP = 42930; FP = 2889; FN = 3336; TN = 845.

TP is positive data, namely yes (having heart disease) which is classified as yes, FP is data yes but is classified as no by the system, then TN is the number of people who do not have heart disease and is classified as no, and FN is the data number that is classified yes by system.

```
[[45376   288]
 [ 4286    50]]
```

Figure 7. Results of Confusion Matrix in KNN algorithm

Figure 7 shows the results of the Confusion matrix using the KNN algorithm, TP = 45376; FP = 288; FN = 4286; TN = 50.

```
array([[45817,     2],
       [ 4181,     0]])
```

Figure 8. results of Confusion Matrix in SVM algorithm

Figure 8 shows the results of the Confusion matrix using the SVM algorithm, TP = 45817; FP = 2; FN = 4181; TN = 0.

Calculation of classification performance evaluation is done by confusion matrix. The confusion matrix value of each algorithm is shown in table 2.

Table 2. Confusion Matrix

Algoritma	TP	TN	FP	FN
Naïve Bayes	42930	845	2889	3336
KNN	45376	50	288	4286
SVM	45817	0	2	4181

3.4. Accuracy Measurement

The algorithms Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) were used in this study, and the values of precision, recall, accuracy, and F Measure were used as a result of the classification of this study. The accuracy value will be used as a result to predict heart disease caused by other disease factors suffered by the patient. Figure 9 illustrates the results of Naive Bayes' accuracy measure.

	precision	recall	f1-score	support
No	0.93	0.94	0.93	45819
Yes	0.23	0.20	0.21	4181
accuracy			0.88	50000
macro avg	0.58	0.57	0.57	50000
weighted avg	0.87	0.88	0.87	50000

Figure 9. Accuracy measurement with Naive Bayes algorithm

Figure 10 illustrates the results of the KNN accuracy measure.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

	precision	recall	f1-score	support
No	0.91	0.99	0.95	45664
Yes	0.15	0.01	0.02	4336
accuracy			0.91	50000
macro avg	0.53	0.50	0.49	50000
weighted avg	0.85	0.91	0.87	50000

Figure 10. Accuracy measurement with KNN algorithm

Figure 11 illustrates the results of the SVM accuracy measure.

	precision	recall	f1-score	support
No	0.92	1.00	0.96	45819
Yes	0.00	0.00	0.00	4181
accuracy			0.92	50000
macro avg	0.46	0.50	0.48	50000
weighted avg	0.84	0.92	0.88	50000

Figure 10. Accuracy measurement with SVM algorithm

It can be seen in Table 3 the details of measuring the accuracy of each algorithm.

Table 3. Accuracy measurement

Algorithm	**Accuracy**	***Recall***	***Precision***	***F-Measure***
Naïve Bayes	88%	88%	87%	87%
KNN	91%	91%	85%	87%
SVM	92%	92%	84%	88%

4. CONCLUSION

Based on the experiment of this study, the Support Vector Machine (SVM) algorithm with the Radial Basis Function (RBF) kernel became the most recommended algorithm than Naïve Bayes and KNN algorithms for making predictions of heart disease classification with the highest accuracy of 92%. The accuracy of the algorithm was affected by the number of trained datasets. Furthermore, research can be carried out using other classification algorithms to predict patients with heart disease and other diseases. Prior to the gold standard method, medical staff could use the advice in this study to avoid invasive, risky and uneconomical diagnoses of heart disease.

5. REFERENCES

[1] C. W. Tsao *et al.*, "Heart Disease and Stroke Statistics-2022 Update: A Report From the American Heart Association," *Circulation*, vol. 145, no. 8, pp. e153–e639, Feb. 2022.

[2] "Global atlas on cardiovascular disease prevention and control." [Online]. Available: https://www.who.int/publications/i/item/9789241564373. [Accessed: 31-Jul-2022].

[3] M. Allahyari *et al.*, "Text Summarization Techniques : A Brief Survey," 2017.

[4] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative Artificial Neural Networks-Based Decision Support System for Heart Diseases Diagnosis," *J. Intell. Learn. Syst. Appl.*, vol. 05, no. 03, pp. 176–183, 2013.

[5] Q. Kadhim Al-Shayea, "Artificial Neural Networks in Medical Diagnosis," *IJCSI Int. J. Comput. Sci. Issues*, vol. 8, no. 2, 2011.

[6] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

[7] A. Rufai, U. S., and M. Umar, "Using Artificial Neural Networks to Diagnose Heart Disease," *Int. J. Comput. Appl.*, vol. 182, no. 19, pp. 1–6, Oct. 2018.

[8] D. A. Firdlous, "Komparasi Algoritma Klasifikasi Data Mining untuk Memprediksi Penyakit Jantung," *Infoman's J. Ilmu-ilmu Manaj. dan Inform.*, vol. 16, no. 1, pp. 79–84, May 2022.

[9] K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features," *2017 4th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Comput. Electron. UPCON 2017*, vol. 2018-January, pp. 500–504, Jun. 2017.

[10] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," *Proc. - IEEE Symp. Comput. Commun.*, pp. 204–207, Sep. 2017.

[11] A. Fadlli and M. I. Rosadi, "Klasifikasi Penyakit Jantung Koroner Menggunakan Seleksi Fitur Dan Support Vector Machine," *Explor. IT J. Keilmuan dan Apl. Tek. Inform.*, vol. 10, no. 2, pp. 32–41, 2018.

[12] "Klasifikasi K-Nearest Neighbors (KNN) Menggunakan Python | by ERZYLIA HERLIN BRILIANT | Medium." [Online]. Available: https://medium.com/@16611077/klasifikasi-k-nearest-neighbors-knn-menggunakan-python-10c64bcb10a1. [Accessed: 21-Jul-2022].

[13] M. Fairuzabdi, "Konsep Dasar Data, Informasi & Pengetahuan - FairuzelsaidFairuzelsaid," 2020. [Online]. Available: http://fairuzelsaid.upy.ac.id/sistem-informasi/konsep-dasar-data-informasi-pengetahuan/. [Accessed: 20-Jul-2022].

[14] "Naive Bayes Classifier Tutorial: with Python Scikit-learn | DataCamp." [Online]. Available: https://www.datacamp.com/tutorial/naive-bayes-scikit-learn. [Accessed: 21-Jul-2022].

[15] M. Awad and R. Khanna, "Support Vector Machines for Classification," *Effic. Learn. Mach.*, pp. 39–66, 2015.

[16] M. T., D. Mukherji, N. Padalia, and A. Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)," *Int. J. Comput. Appl.*, vol. 68, no. 16, pp. 11–15, Apr. 2013.

[17] "Scikit-learn SVM Tutorial with Python (Support Vector Machines) | DataCamp." [Online]. Available: https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python. [Accessed: 29-Jul-2022].

[18] "Support Vector Machine Classification with Python | by Kurnia Sari Pratiwi | Medium." [Online]. Available: https://medium.com/@kurniasp/support-vector-machine-classification-with-python-64521fbd5b57. [Accessed: 29-Jul-2022].

[19] D. for H. D. and S. P. National Center for Chronic Disease Prevention and Health Promotion, "Heart Disease Facts | cdc.gov," 2022. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm. [Accessed: 26-Jul-2022].

[20] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf. Sci. (Ny).*, vol. 507, pp. 772–794, Jan. 2020.