Prediction Model for Diagnosing Heart Disease Using Classification Algorithm

Risqy Siwi Pradini 1, Mochammad Anshori 2, M. Syauqi Haris 3, Busatto Marilia 4, Tostes Geraldo 5

1 Institut Teknologi Sains dan Kesehatan RS dr. Soepraoen Kesdam V/BRW, Indonesia
2 Institut Teknologi Sains dan Kesehatan RS dr. Soepraoen Kesdam V/BRW, Indonesia
3 Institut Teknologi Sains dan Kesehatan RS dr. Soepraoen Kesdam V/BRW, Indonesia
4 Nanjing Normal University, China
5 University of Nairobi, Kenya

Corresponding Author: Risqy Siwi Pradini  E-mail: risqypradini@itsk-soepraoen.ac.id

ABSTRACT

Heart disease often causes death if not treated quickly and appropriately. Early diagnosis can prevent more serious complications and treat heart disease patients best. The existence of a disease prediction model can help health workers to diagnose diseases more quickly and accurately. The heart disease prediction model using a classification algorithm is a system built using machine learning techniques. The classification algorithm chosen is NN, Naive Bayes, Random Forest, and SVM because it is the best algorithm for predicting heart disease. This study makes a comparison of the four algorithms using a dataset of 918 instances with 11 features. The result is that the Random Forest algorithm produces the highest accuracy, with 86.8%, and has the best ability to distinguish classes based on the ROC curve.

Keywords: Algorithm, Heart Disease, Prediction Model

INTRODUCTION

Heart disease is a disease that often causes death around the world (Thomas & Princy, 2016). Based on the WHO survey, 17% of deaths worldwide are caused by heart attacks. A series of disorders of the heart such as irregular heartbeat, weak heart, and congenital diseases are risk factors for heart disease (Derisma, 2020 (Arnett dkk., 2019; Bechthold dkk., 2019; Powell-Wiley dkk., 2021)). Heart attacks are experienced by many adults and are mostly caused by workload and excessive stress (Dangare & Apte, 2016).
2012 (Bemando dkk., 2021; Mehta dkk., 2018; Vachiéry dkk., 2019)). Thus, it is very important to have an early diagnosis to detect heart disease. Early diagnosis can prevent more serious complications and provide the best care for heart disease patients.

The number of health workers is very limited compared to the patients they treat (Derisma, 2020 (Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India. dkk., 2019)). Doctors may not be able to properly diagnose patients in a short time, especially if the patient has more than one disease (Tarawneh & Embarak, 2019 (Hameetha Begum & Nisha Rani, 2021; Otto dkk., 2021; Vahanian dkk., 2022)). Inaccurate diagnosis can cause disability and even death of the patient. The existence of a disease prediction model can help health workers to diagnose diseases more quickly and accurately.

The prediction model using a classification algorithm is expected to be useful to increase heart disease prediction speed and accuracy. The heart disease prediction model using a classification algorithm is a system built using machine learning techniques by grouping data into classes based on identified patterns (Putra & Rini, 2019 (Hackshaw dkk., 2018; Stout dkk., 2019; Yang & Guan, 2022)). In the context of heart disease prediction, this model will study patient data patterns and classify them into two classes, namely positive and negative heart disease classes. Several previous studies have discussed the topic of heart disease prediction, including research (Ference dkk., 2019; Haq dkk., 2019; Naidu dkk., 2021; Dangare & Apte, 2012; Derisma, 2020; Riani, et al., 2019; Tarawneh & Embarak, 2019; Putra & Riani, 2019; Utomo & Mesran, 2020; Biamto, et al., 2020; Annisa, 2019; Krishnan & Geetha, 2019). The similarities between these studies are using data mining techniques to predict heart disease, while the difference lies in the algorithm they use. Research (Dangare & Apte, 2012 (Mehta dkk., 2018; Tamis-Holland dkk., 2019; Velswamy dkk., 2022)) compared algorithms for the prediction of heart disease and the highest accuracy results when using NN. Research (Derisma, 2020; Riani, et al., 2019; Putra & Rini, 2019; Utomo & Mesran, 2020; Biamto, et al., 2020) resulted in the opinion that the highest accuracy when using Naive Bayes. Research (Tarawneh & Embarak, 2019 (L dkk., 2021; Mohan dkk., 2019; Verma dkk., 2019)) resulted in the opinion that SVM has the highest accuracy for predicting heart disease. Research (Annisa, 2019 (Bechthold dkk., 2019; Benjamin dkk., 2019; Bertrand dkk., 2021; Tuli dkk., 2020)) resulted in the opinion that Random Forest performed the best for the classification of heart disease.

The choice of algorithm depends on the complexity of the problem, dataset, and research objectives, so choosing the best algorithm for predicting heart disease requires comparisons. Based on the four studies above, this study intends to compare the algorithms that are considered to produce the best performance for predicting heart disease (Lakshmi & Devi, 2023; Upadhyay dkk., 2023; Virani dkk., 2020). This study will use classification performance metrics to compare the resulting values. The algorithms to be compared include NN, Naive Bayes, SVM, and Random Forest. The
algorithm that produces the highest CA value will be considered the best algorithm for predicting heart disease.

It is hoped that the results of this study can contribute new knowledge about the best classification algorithm for heart disease prediction models. In addition (Babu & Karthick, 2022; Hackshaw dkk., 2018; Latha & Jeeva, 2019), the prediction model produced in this study can be used to prevent and diagnose heart disease more quickly and accurately. Thus, health workers can take more appropriate prevention and treatment steps.

RESEARCH METHODOLOGY

The process of developing a predictive model for diagnosing heart disease starts with collecting datasets. This study took the dataset for heart disease prediction from Kaggle. The dataset used is 918 instances (number of rows) and 11 columns (number of features). The features used include age, sex, chest pain, resting BP, cholesterol, fasting BP, Resting ECG, Max HR, Exercise Angina, Oldpeak, and ST Slope. The value of each feature will be categorized into 2 classes, namely 0 for negative heart disease and 1 for positive heart disease.

Furthermore, the dataset will be processed using the NN, Naive Bayes, SVM, and Random Forest algorithms. This research will use the help of orange tools to process the dataset when using the algorithm. The results of each algorithm will be compared using classification performance metrics (Bisercic et al., 2023) such as AUC, CA, F1 score, precision, and recall.

AUC (area under the curve) and CA (classification accuracy) are used to measure the performance of the classification model. AUC is used to see the ability of the model to differentiate classes and CA is used to see the accuracy of the resulting predictions (Carrington, et al., 2023; Bhoi et al., 2021). Meanwhile, the three evaluation matrices (F1 score, precision, recall) are used to evaluate the resulting predictive model and it is necessary to pay attention to positive or negative results in data classification (Safder, 2021). Furthermore, based on the results of classification performance metrics, will be used to analyze the ROC curve and the best algorithm for predicting heart disease. The algorithm that will be run for this research can be seen in Algorithm 1 below.

### Algorithm 1. Algorithm for prediction of heart disease

| Inputs: heart disease prediction dataset with 918 instances and 11 columns. |
| Outputs: AUC, CA, F1, precision, recall, confusion matrix, ROC curve. |
| Steps: |
| 1. Open the heart disease diagnosis dataset. |
| 2. Create a prediction model. |
| Each algorithm is used as a learning model. |
| 4. Implement k-fold cross-validation. |
| 5. Determine the average of the series of tests to find the AUC, CA, F1, precision, and recall scores. |
| 6. Determine the CA with the most values as the best algorithm for predicting heart disease. |
heart disease.
7. Scores are compared graphically using the ROC curve.

RESULT AND DISCUSSION
In this section, we will discuss the results that have been analyzed using Algorithm 1. This study uses an open-source platform, orange, to predict heart disease. The dataset is taken from Kaggle and divided into training data and testing data. Data were analyzed by predicting the target class using various classification algorithms namely NN, Naive Bayes, SVM, and Random Forest. Figure 1 shows a machine learning framework designed in orange for heart disease prediction.

This study tries to enter 5 new instances (testing data) and process them using orange to produce the data shown in Table 1. Table 1 shows a comparison of the predicted results using the classification algorithm with actual data. The output (predictive data) generated by the NN and Random Forest algorithms is closer to the actual data when compared to the SVM and Naive Bayes algorithms.

Figure 1. A Machine Learning Framework Designed in Orange for Heart Disease Prediction

<table>
<thead>
<tr>
<th>Prediction Data</th>
<th>Actual Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>NN</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Comparison of Predicted Data and Actual Data
Furthermore, this study conducted k-fold cross-validation with several folds of 10 as suggested in the study (Bhoi, et al., 2021). Thus, the resulting AUC, CA, F1 score, precision, and recall values are shown in Table 3. Based on Table 2, it can be concluded that the highest CA value is the Random Forest algorithm with an accuracy of 86.8%, followed by NN with an accuracy of 86.2%, then Naive Bayes with an accuracy of 85.2%, and the last is SVM with an accuracy of 83.8%.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>CA</th>
<th>F1</th>
<th>Prec</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.916</td>
<td>0.868</td>
<td>0.868</td>
<td>0.869</td>
<td>0.868</td>
</tr>
<tr>
<td>NN</td>
<td>0.923</td>
<td>0.862</td>
<td>0.861</td>
<td>0.862</td>
<td>0.862</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.918</td>
<td>0.852</td>
<td>0.852</td>
<td>0.852</td>
<td>0.852</td>
</tr>
<tr>
<td>SVM</td>
<td>0.912</td>
<td>0.838</td>
<td>0.836</td>
<td>0.840</td>
<td>0.838</td>
</tr>
</tbody>
</table>

Furthermore, to determine the model's ability to differentiate classes, it is necessary to look at the AUC value and the graph in Figure 2. AUC measures the area formed under the ROC curve, so the ROC curve aims to provide an overview of the extent to which the prediction model can distinguish positive and negative classes. The closer the curve follows the left and upper limits of the ROC space, the more accurate the resulting classification will be. Based on Figure 2, the Random Forest (blue color) has the outermost threshold, followed by NN (orange color), then Naive Bayes (purple color), and lastly SVM (green color). Thus, Random Forest has the best ability to distinguish classes and has the most optimal classification accuracy.

Based on the research that has been done and discussed in the previous sub-chapter, it can be concluded that Random Forest is an algorithm that has the highest...
accuracy and can distinguish classes best compared to the NN, Naive Bayes, and SVM algorithms. This is evidenced by the CA value generated by the Random Forest which is 0.868 or the accuracy level is 86.8%. The ranking of the algorithm for predicting heart disease can be seen in Table 3. Of course, the results of this ranking only apply when using the [Kaggle] dataset. If you use a different dataset, the ranking of the algorithm might change. In addition, the results of the ROC curve illustrate that the Random Forest has the best ability to distinguish classes and has the most optimal classification accuracy.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Algorithm</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.916</td>
<td>0.868</td>
</tr>
<tr>
<td>2</td>
<td>0.923</td>
<td>0.862</td>
</tr>
<tr>
<td>3</td>
<td>0.918</td>
<td>0.852</td>
</tr>
<tr>
<td>4</td>
<td>0.912</td>
<td>0.838</td>
</tr>
</tbody>
</table>

Predictive models cannot replace the role of medical personnel in diagnosing diseases. However, the results of the prediction model can help medical personnel to identify early risk factors and early symptoms of disease, one of which is heart disease. Thus, health workers can take more appropriate prevention and treatment steps.

REFERENCES


---

**Copyright Holder:** © Risqy Siwi Pradini et al. (2023)

**First Publication Right:** © Journal of World Future Medicine, Health and Nursing

**This article is under:**

![Creative Commons License](https://creativecommons.org/licenses/by-nc-sa/4.0/)