Data mining techniques with machine learning algorithm to predict patients of heart disease

Mulyawan1*, Agus Bahtiar2, Githera Dwilestari3, Fadhil Muhammad Basysyar4, Nana Suarna5

STMIIK IKMI Cirebon, Cirebon, Indonesia

*wm7488748@gmail.com

Abstract. Data mining is a way of searching for information from large amounts of data for the purposes of various applications. Several techniques in data mining can be used for association, classification, clustering, prediction, and sequential modeling. Machine learning is used in medical science to help medical teams find out the condition of patients with heart disease. A lot of machine learning still has limited predictive capabilities, and is incompatible. This study uses different machine learning techniques, namely PSO-based SVM, Neural Network, Decision Tree, Naïve Bayes and SVM to assist in building, understanding and interpreting different models of heart disease diagnosis. The use of the pso-based svm algorithm in the prediction of heart disease shows a 100% greatest accuracy than the Decision Tree, only 88.68% and Naïve Bayes of 82.15%, Neural Network with an accuracy of 95.71%, SVM with an accuracy of 99.71%. The results of this study are expected to be beneficial for the world of health and for researchers who use machine learning techniques.

1. Introduction

Heart disease is caused by years of build-up of plaque in arteries, and is caused by the patient's elevated plasma levels of low density lipoprotein (LDL). The buildup of plaque occurring on the walls of the coronary arteries reduces blood flow to the heart muscles because the narrowing prevents the aorta from fully expanding. If arteries become clogged, it increases the chance of heart attack. [1]. There is a hidden pattern in the data that can be discovered using data mining; it can be used for the purposes of diagnosing problems. These data need to be comprehensively organized and integrated into a well-suited system for use by its users. Medical diagnosis systems are conducted using computer-based automation systems to handle complex medical work, and can be conducted efficiently and accurately. Using a computerized diagnostic system can lower the costs of clinical trials testing new drugs. [2].

Machine learning, which is part of artificial intelligence, is a way of making artificial intelligence with computers. Machine learning can be divided into 4 parts, namely [3] Supervised Learning, used in classification and regression problems and the dataset has an output label. Unsupervised Learning: used for clustering problems such as k-nearest neighbor, k-mean, hierarchical clustering. Semi Supervised Learning: can be used for predictive cases. Reinforcement Learning is done by learning from mistakes and from the environment to get the best results. Machine learning technology in a computer system can be used in the medical world to help doctors. Besides, machine learning has the capacity for efficient disease management [4].
2. Methodology

Literature review considering heart disease and statistical data mining techniques is widely shared on the Internet. In this study, the focus is on developing the Heart Disease Prediction system using data mining techniques, namely, statistical learning using Support Vector Machines, neural networks, naïve Bayes, and classification trees. to present the best possible results that the algorithm predicts. Previous research related to the writing of this paper consisted of 5 journals with author data, data of the year of writing, and data mining techniques used and the accuracy results of each prediction in the journal as described in table 1 as follows:

| No | Year | Author | DM Technique | Accuracy |
|----|------|--------|--------------|----------|
| 01 | 2016 | Ashmeet Singh1, R Sathyaraj | Random Forest, Naïve Bayes, Decision Tree | 84.34% |
| 02 | 2017 | Rafaqat Alam Khan, Taseer Suleman, Muhammad Sajid Farooq, Muhammad Waist, Hassan Rafiq | Logistic Regression, Linear Regression, SVM | 98.24% |
| 03 | 2018 | R.S Sanjugyasini, R. | Naïve Bayes, Decision Tree | 98.9% |
| 04 | 2019 | Hossam Meshref | Artificial Neural Networks, Support Vector Machines, Naïve Bayes, Decision Trees and Random Forests | 84.25% |
| 05 | 2020 | Raj H. Chauhan1, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel, Mr. A.D.Prajapati | Random Forest, Gaussian Naïve Bayes, Decision Tree | 97.64% |

Research conducted by Rafaqat Alam Khan, Taseer Suleman, Muhammad Sajid Farooq, and Muhammad Hassan Rafiq with Logistic Regression, Linear Regression, SVM, feature selection the best accuracy results are obtained in Logistic Regression with an overall accuracy of 98.24. Meanwhile, in research conducted by Raj H. Chauhan1, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel, Mr. A.D.Prajapati using Random Forest, Gaussian Naïve Bayes, and Decision Tree obtained the best accuracy results in Random Forest of 97.64% after preprocessing. This study uses data mining techniques using machine learning algorithms with the following stages:

2.1 Data Collection

The data sets used were accessed from http://kaggle.com with medical attributes for the diagnosis of heart disease. The dataset is divided into 2 parts, namely the training dataset and the testing dataset. The number of data records was 1025 medical records using several medical attributes. All attributes have a numeric value. The attributes used in the study amounted to 14 attributes.

2.2 Data Preprocessing

Data on patients with heart disease were preprocessed to eliminate noise and missing value data. Data that has been collected in excel format is transferred to the repository tools Rapid Miner 9.5 and preprocessing is carried out.

2.3 Attributes Selection

The total number of attributes used is 14 attributes. Attributes regarding age and gender are information that explains the patient's identity, 12 other attributes are medical data that can be useful for diagnosing and studying heart disease through a machine learning process [4].
2.4 Classification
Classification is a data processing technique used to divide data into different groups based on different criteria. By classifying the data, it can be determined whether the data obtained is accurate or not. In data mining, classification problems can identify prediction accuracy problems using various techniques. [5].

2.4.1 Naïve Bayes
The formula in Naïve Bayes Algorithm is used to determine which class decisions are included in a new instance. The classification rules on Naïve Bayes depend on features and state that features are independent of one another. In calculating class probability Naïve Bayes is used for a specific data set [6]:

$$P(C_k/X) = \frac{P(C_k) \cdot P(X|C_k)}{P(X)}$$

where:
- $X$ is an instance that must be classified
- $C_k$ represents each class
- $P(C_k/X)$ is the probability vector $X$ belonging to the $C_k$ class

2.4.2 Decision Tree
The decision tree algorithm uses entropy and information gain for machine learning. The decision tree utilizes supervised learning to classify or assess every tree node. The root node, the number of leaves, and how long the tree is being important when calculating the overall accuracy and efficiency of the tree. Decision trees can handle a variety of data types. [3].

Entropy ($T$) = $-P \cdot \log_2 P + \cdot P \cdot \log_2 P - E(T, X) = \sum P(c) E(c)$
Gain ($T, X)$ = Entropy ($T$) - Entropy ($T, X$)

2.5 Neural Network
Neural network algorithms can be used for pattern recognition, calculation and decision making. The way the neural network works in the human body, and how the information is disseminated via connections among nerves. Neural Networks are able to be used under both supervised and unsupervised learning problems. [7].

2.6 Support Vector Machine
SVM is machine learning that can be used to create classification or regression models and is included in supervised learning [8][9]. SVM classifier focuses on processing data that is difficult to classify with other algorithm models. In SVM, linear classification aims to maximize the distance between 2 hyperplane. By measuring the distance, a solution will be found.

$wX - b = -1$ for class 1 $wX - b = 1$ for class 2

2.7 Particle Swarm Optimization
The PSO method is a method used to solve complex problems that occur between simple agents and their environment. Every particle that moves in each iteration, one of the particles that is approaching the optimum position describes its position to the others. The PSO algorithm has the principle of moving particles to find the optimum position [8]-[3].
\[ V_i(t+1) = w.v_i(t) + c1r1(pi.best[t] - pi[t]) + c2r2(pg.best[t] - pi[t]) \]

\[ P_i(t+1) = pi[t] + vi[t+1] \]

Where:

- \( i = 1, 2, \ldots, N \), \( N \) is the number of swarm population.
- \( v_i \) = velocity vector
- \( p_i \) = current particle position, \( p_i \), best = is the position of the best previous particle \( p_g \), best = best previous position of all particles.
- \( c1 \) dan \( c2 \) are positive acceleration coefficients
- \( r1 \) dan \( r2 \) are random numbers between 0 and 1

### 2.8 Classification Performance Measurement

To evaluate the accuracy, precision, and recall for prediction and classification, rapid miner tools are used, including software programs that can be used for machine learning processing, data mining, and predictive analysis. Performance measurement from the results of prediction and classification is understandable as is. [10]-[11]:

#### 2.8.1 Accuracy

Accuracy is the accurate predictive value of a class from a total class whose value ranges from 0 to 100. The formula notation is written as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}
\]

#### 2.8.2 Precision

Precision is the actual grade of precision classified and is referred to as positive predictive value, High precision indicates accurate results.

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

#### 2.8.3 Recall

Recall describes the sensitivity of the problem and processes the value or completeness of the product.

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

### 3. Proposed Method

#### 3.1 Work Flow Model

Modeling made in data mining techniques so that it can be processed in Rapid Miner tools, which is described in Figure 1 with optimization and cross validation techniques.

![Figure 1. Work Flow Model](image)

### 4. Result and Discussion

In a study to predict patients who suffer from heart disease using the pso-based svm algorithm, it was found that there are several factors that influence the performance of the accuracy results including the attribute weight: Age, Sex, CP, Trestbps, Chol, Restecg, Fbs, Thalach, Exang, Oldpeak, Slope, Ca, Thal, and the table 3 is described as follows:
Research in predicting heart disease patients using a support vector machine algorithm based on particle swarm optimization with the Rapid Miner tool shows several classification results. The 10 folds Cross Validation technique is used for machine learning algorithms with the target class attribute. The test results from the dataset obtained the classification accuracy level in 5 comparisons using different machine learning algorithms as described in Table 4. From the explanation shown in Table 4, the best accuracy results are 100% by using the pso-based SVM algorithm. These results were obtained by means of optimization techniques with the aim of improving SVM performance.

In the SVM kernel function, the RBF kernel is selected with the parameters $C = 5.0$, $\gamma = 1.0$, convergence epsilon $= 0.05$, kernel cache $= 200$ and max iteration $= 10000$. As for the naïve Bayes, neural network, decision tree, and SVM methods without using optimize. The accuracy of the SVM performance without optimization techniques is 99.71% and the smallest accuracy value is using the naïve Bayes algorithm of 82.15%. The difference in the results of the prediction accuracy of the pso-based SVM method and the svm method is 0.29%.

### Table 5. Performance Comparison Using Different Function Kernel in PSO Based SVM

| Parameter         | RBF | Gama/Kernel | Degree/Kernel | Sigma | C   | Convergence Epsilon | Kernel Cache | Accuracy |
|-------------------|-----|-------------|---------------|-------|-----|---------------------|--------------|----------|
| Degree            | 2   | 5           | 0             |       | 0.05| 0.05                | 200          | 100%     |
| Kernel dot        | -   | -           | -             |       | 0.05| 0.05                | 200          | 83.12%   |

### 4.1 Statistics Based on Class Attributes

From the statistics shown in Figure 8, there are 526 patients who have a risk of heart disease even though it is still at a low level, while patients who do not have a risk of heart disease are 499 people.

![Figure 2. Statistics by Class](image1)

![Figure 3. Statistics by Age](image2)

In the statistics in Figure 9, it can be seen that the highest number of patients who are at risk of heart disease, namely at the age of 58 years to the age of 62 years is 65 patients from a total of 219 patients, around 29.7%. Meanwhile, the smallest number of patients who had a heart attack was 4 people aged 29 to 33 years.
4.2 Comparison

The yellow line in the ROC curve image shows the under curve area of the neural network algorithm and the value is 0.966 which is above the under curve area line of the naïve Bayes algorithm which is light green and the value is 0.907.

![Figure 4. ROC](image1)

![Figure 5. AUC Neural Network](image2)

![Figure 6. Naïve Bayes](image3)

The threshold value on the neural network algorithm is 0.966 at positive class risk so that this model is good and can be used. The threshold value on the naïve Bayes algorithm is 0.907 on the positive class risk so that this model is good and can be used.

5. Conclusion

This paper uses 5 machine learning algorithms, namely PSO-based SVM, Neural Network, Naïve Bayes, Decision Tree, Support Vector Machine in predicting heart disease patients evaluated according to their respective accuracy results. The prediction of the PSO-based SVM method uses the optimize and cross validation technique, while the Neural Network, Naïve Bayes, SVM, and Decision Tree methods do not use optimization techniques but only use cross validation techniques. The results of the best prediction accuracy are using the PSO-based SVM method with an accuracy value of 100%. The SVM kernel function parameters used are the Radial Basis Function (RBF) kernel, the gamma kernel value is 1.0, kernel cache = 200, the value of Cost (C) = 5, convergence epsilon = 0.05. The lowest accuracy value is 82.15% using the Naïve Bayes algorithm. The difference between the SVM accuracy value based on PSO with optimization techniques and cross validation with the SVM accuracy value using cross validation without optimization is 0.29%

References

[1] F. S. Alotaibi, “Implementation of machine learning model to predict heart failure disease,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 261–268, 2019.

[2] U. A. Son, Jyoti, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction,” Smart Talent Manag. Build. Knowl. Assets Compet. Advant., vol. 17, no. 8, pp. 119–138, 2008.

[3] S. Barik, S. Mohanty, D. Rout, S. Mohanty, A. K. Patra, and A. K. Mishra, “Heart Disease Prediction Using Machine Learning Techniques,” Lect. Notes Electr. Eng., vol. 665, no. 3, pp. 879–888, 2020.

[4] A. Agrawal, H. Agrawal, S. Mittal, and M. Sharma, “Disease Prediction Using Machine Learning,” SSRN Electron. J., no. May, pp. 6937–6938, 2018.

[5] R.S. SANJUVIGASIN, “AN EFFICIENT COMPARISION OF DATA CLASSIFICATION ALGORITHM FOR ANALYSIS OF IRIS DATA SETS,” pp. 1–11, 2018.

[6] F. Khennou, C. Fahim, H. Chaoui, and N. E. H. Chaoui, “A machine learning approach: Using predictive analytics to identify and analyze high risks patients with heart disease,” Int. J. Mach. Learn. Comput., vol. 9, no. 6, pp. 762–767, 2019.

[7] K. Karaduzovic-Hadziabic and R. Köker, “Diagnosis of heart disease using a committee machine neural network,” no. 3, pp. 351–360, 2015.

[8] O. Somantri, D. A. Kurnia, D. Sudrajat, N. Rahaningsih, O. Nurdiawan, and L. Perdana Wanti, “A Hybrid Method Based on Particle Swarm Optimization for Restaurant Culinary Food Reviews,” Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019, pp. 0–4, 2019.

[9] A. Lakshmanarao, Y. Swathi, and P. Sri Sai Sundareswar, “Hybrid Machine Learning Techniques for Heart Disease Prediction S.,” Int. J. Sci. Technol. Res., vol. 8, no. 11, pp. 374–377, 2019.

[10] A. Singh and R. Sathyaraj, “A Comparison Between Classification Algorithms on Different Datasets Methodologies using Rapidminer,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 5, no.
[11] O. Nurdiawan, F. A. Pratama, D. A. Kurnia, Kaslani, and N. Rahaningsih, “Optimization of Traveling Salesman Problem on Scheduling Tour Packages using Genetic Algorithms,” *J. Phys. Conf. Ser.*, vol. 1477, no. 5, 2020.