International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Comparative Study of Heart Failure Prediction Algorithm: Logistic Regression and SVM

Anang Prasetyo[1], Erick[2], Diva Angelika Mulia[3], Keiko Kimberly Octavina[4]

Submitted: 11/11/2022 **Accepted**: 12/02/2023

Abstract: Heart and blood vessel function is the biggest causes of death globally. Some researchers used machine learning approaches to predict it earlier and effective. There is a various machine learning to predict heart disease including Logistic Regression and Support Vector Machine. Many researchers using some of dataset, the dataset used is a dataset of heart disease patients obtained from the UCI Machine Learning Repository. This study comparing the performance of Logistic Regression and Support Victor Machine to predict it. Finally, the results Logistic Regression and SVM has average same performance, the AUC Support Vector Machine value is around 93.58% and Logistic Regression is around 92.95%, the SVM value is 0.63% higher than Logistic Regression.

Keywords: *Machine Learning, Heart Disease, Support Vector Machine, Logistic Regression*

1. Introduction

Heart and blood vessel function is important cause of death globally, according to the Indonesian Ministry of Health. It caused by various factors such as age, gender, family history, hypertension, diabetes, and lifestyle [1]. Some of researcher gather more information that cause heart disease, they can use machine learning approaches to predict it earlier and more effective. This allows some notice to interventions and help patients prevent heart disease.

Machine learning, as defined by IBM, is a discipline of artificial intelligence (AI) that focuses on using data and algorithms to replicate the way humans learn, give some decisions [2]. There are two types of machine learning technique: unsupervised learning and supervised learning. Unsupervised learning is involving of analyses and classifies unlabelled data sets and supervised learning is learning that works by labelled data. In this study used a labelled heart disease dataset from the UCI Machine Learning Repository, which can be found on the Kaggle website [3].

Classification is a process of categorizing based on properties or characteristic. There are a variety of supervised machine learning algorithms for classifying such as Naive Bayes, Decision Tree-Based Method, Rule-based Methods, Neural Network, K-Nearest Neighbour (KNN), Logistic Regression and Support Vector Machine (SVM). This study used Logistic Regression and Support Vector Machine. Logistic regression an algorithm with simple but effective model for large datasets, but it can be difficult to implement in complex data. Support vector machine is ideal model to handle a complex issue and high-dimensional datasets, and more difficult to set up a need more time to train the datasets [4].

Therefore, this study aims to compare the performance of the Support Vector Machine and Logistic Regression for predicting heart disease based on dataset from UCI Machine Learning Repository.

1234Computer Science Deapartment, School of Computer Science, Bina Nusantara University, Jakarta Indonesia 11480
** Corresponding Author Email: anang.prasetyo@binus.edu*

2. Literature Review

A total of 17.9 million cases of death are caused by constriction of blood vessels in the heart, this makes heart disease ranks first among the top ten causes of death in the world [5] [6]. Heart disease has symptoms such as an irregular heartbeat, blood vessel problems, or chest pain. There are several factors that can cause heart disease, including: age, gender, diabetes, cholesterol, hypertension, and diet [7]. Over time, the number of deaths from heart disease is also increasing, which makes technology important in predicting heart disease as a preventative measure [8]. By predicting heart disease early, it can help treat heart disease better and more accurately, so that it can have an impact on improving the patient's quality of life [9]. There are a variety of methods that can be used to classify heart disease against data from the health industry about the factors that cause heart disease. Various AI methods have been applied to help predict heart disease, namely machine learning that can learn from datasets from various hospitals to produce accurate predictions [10].

Machine learning is a technique or method for solving a problem using a dataset. Machine learning aims to consistently interpret patterns from data, analyse the relationship between data, and systematically predict a discovery in the dataset [11]. Support Vector Machine, K-Nearest Neighbour (KNN), Artificial Neural Network (ANN), Deep Residual Neural Network, Logistic Regression, and other algorithms could be used to identify heart failure [12]. The supervised learning model of Logistic Regression and Support Vector Machine will be classified in the paper. Logistic Regression is one of a classification algorithm when the prediction results have a binary value (yes or no). This method is used to explain the relationship between the dependent and independent variables so that they can perform classification [13]. While the Support Vector Machine is an algorithm used to find the best hyperplane in a space with two or more dimensions, the hyperplane is used as a separator between classes [14] [15].

In Muhammad Zeeshan Younas' research, Logistic Regression also gives a good accuracy value of around 86.8% [16]. Another study by Tania Ciu and the team from Indonesia predicts heart disease using the Logistic Regression algorithm, producing an accuracy of

85%, with an error rate of only 0.145 [17]. Then the research conducted by Reddy Prasad and a team from India, which compared Logistic Regression, KNN, and Naïve Bayes, concluded that the Logistic Regression accuracy value was higher, around 86.89% [18]. Furthermore, research conducted by Yamala and the team from India, using the SVM algorithm and compared with other algorithms such as ANN, Naïve Bayes, and KNN, concluded that SVM is more accurate, with an accuracy value of 85.6% [19]. Research conducted by Faria Rahman and the team from Bangladesh compared several techniques such as KNN, SVM, and Decision Tree, which showed that SVM had more accurate results, with an accuracy value of 93.33% [20]. Then research conducted by Megha and the team from India, comparing the SVM, KNN, Naïve Bayes, and Decision Tree algorithms, concluded that SVM is superior, with an accuracy value of 85% [21].

In this paper we choose Support Vector Machine and Logistic Regression as the algorithm because in SVMs always try to maximize the margins between support vectors to distinguish classes, SVM also have kernel methods that can classify features by mapping higher dimensional data by using kernel function, therefore SVM is suitable for handling complex data. Logistic regression uses the sigmoid function to finds the predicted value either 0 or 1 output which that is what we are going to do, to predict whether a person is potentially has a heart failure or not.

With models of machine learning that we created using Support Vector Machine and Logistic Regression method, we are trying to compare which algorithm can give the better result to produce heart failure prediction.

3. Methods

As shown in Fig. 1, there are steps that must be taken so that the model can produce good predictions.

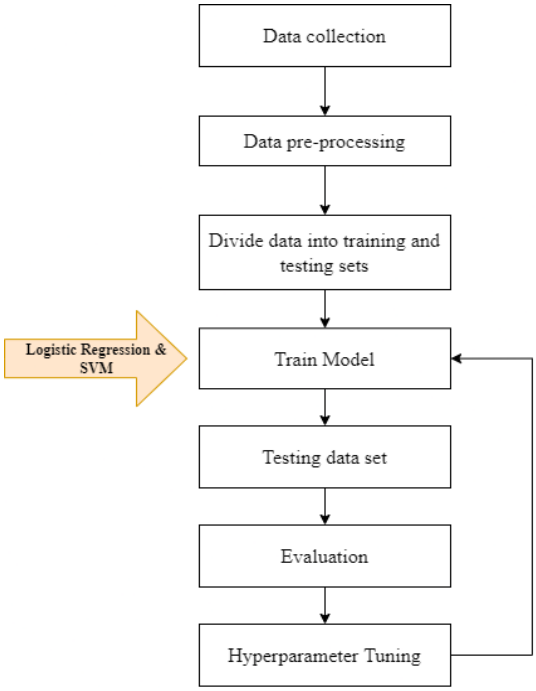

Fig. 1. Steps to conduct research

3.1. Data Collection

The initial stage in conducting this research is to collect datasets that will be used to classify heart failure. The datasets used in this study were obtained from the UCI Machine Learning Repository and are available on the Kaggle website. The dataset is a collection of observations on patients from Switzerland, Cleveland, Long Beach VA, Hungary, and the Stalog Dataset. This dataset is combined into eleven common features, for more details can be seen in Table 1.

Table 1 Attribute Information

Feature Name	Description
Age	Patient age (in years)
Sex	Patient gender (male and female)
ChestPainType	Types of chest pain (Typical Angina, Atypical Angina, Non-Anginal Pain, and Asymptomatic)
RestingBP	Blood pressure levels at rest (in mm Hg)
Cholesterol	Serum cholesterol (in mm/dl)
FastingBS	Blood sugar levels after fasting
RestingECG	Electrocardiography results at rest (ST, LVH and Normal)
MaxHR	Reached maximum heart rate (60 – 202)
ExerciseAngina	Exercise includes angina
Oldpeak	Relative exercise-induced ST depression.
ST_Slope	The ST segment's slope during the peak workout (upsloping, flat, and down sloping)
Heart Disease	Output • 1: heart disease • 0: normal

3.2. Data Preprocessing

"Garbage in, trash out" is an often-used cliche in machine learning. If a machine learning model learns data that is poorly labelled or has lots of null values, then the resulting model is also bad. Data pre-processing is a technique to convert raw data into data that can be processed, so that the resulting model is also good.

3.2.1. Understanding datasets

Understanding datasets such as knowing the amount of data, data types, and variables or dataset sizes can help determine the steps that must be taken at the next stage of data pre-processing.

There are 918 samples in the dataset used in this study. Age, RestingBP, Cholesterol, FastingBS, MaxHR, Oldpeak, and HeartDisease are numerical variables, while Sex, ChestPainType, RestingECG, ExerciseAngina, and ST_Slope are categorical variables.

3.2.2. Data cleaning

After understanding the dataset, it is necessary to do data cleaning such as handling null values and data duplication, normalizing data, removing irrelevant variables, and so on.

The dataset used in this paper does not have null values and duplicate data, but categorical variables need to be converted into numeric variables because machine learning models are based on mathematical calculations.

Since there are not many unique values of categorical variables in

this dataset, the ideal encoding method is one hot encoding. One hot encoding is the process of adding a new feature to each categorical variable whose values are 0 and 1. In Table II are the new variables after performing one hot coding.

Table 2. Categorical Variables After One Hot Encoding

Feature Name	Feature Name After One Hot Encoding
Sex	Sex_Female
	Sex_Male
Certainties	ChestPainType_AtypicalAngina
	ChestPainType_TypicalAngina
	ChestPainType_Asymptomatic
	ChestPainType_NonAnginalPain
RestingECG	RestingECG_LVH
	RestingECG_ST
	RestingECG_Normal
ExerciseAngina	ExerciseAngina_Yes
	ExerciseAngina_No
ST_Slope	ST_Slope_Upsloping
	ST_Slope_Downsloping
	ST_Slope_Flat

3.2.3. Splitting Dataset

The datasets must be separated into training and testing sets before the model can be trained. The model is trained using the training set, while the testing set is used to evaluate the performance of the model that has been trained [22]. In this paper we used the 80-20 ratio to split the datasets.

3.2.4. Feature Scaling

Because machine learning models are based on mathematical calculations, feature scaling is needed for some sensitive algorithms to calculate the distance between data because the values in each feature have different distances, some are of small value, and some are of large value. Algorithms such as KNN, K-means or Support Vector Machine (SVM) perform classifications based on the distance between data, which makes these algorithms sensitive to the distance between data.

There are several ways to perform feature scaling, including Min-Max Scaling, Robust Scaling and Standard Scaling. In this research, we use the Standard Scaling method, which changes the distribution of the data centered on the average value with a standard deviation of one.

3.2.5. Hyperparameter Tuning

Machine learning models have various parameters in carrying out data learning, but there are some parameters that cannot be studied directly which are the architecture of the model called hyperparameters. Hyperparameters in the Logistic Regression model such as determining the solver algorithm to perform optimization, or in the Support Vector Machine specifying the type of kernel to be used. The process of finding the most optimal set of hyperparameters for the model architecture is called hyperparameter tuning.

3.3. Machine Learning Model

In this study, we used Logistic Regression algorithm and Support Vector Machine to predict the possibility of heart failure in patients..

3.3.1. Logistic Regression Model

Logistic regression is an algorithm that is used to predict something that is true or false (0 or 1). In this dataset, if the output has a value of 0 it means that it has a normal heart, while if it has a value of 1 it means heart failure. Logistic regression works by measuring the relationship between independent variable and dependent variable (heart disease). Equation (1) is the sigmoid function that used in this model for the predicted value to be 0 or 1.

$$y = \frac{1}{1+e^{-x}} \tag{1}$$

As seen in Fig.2, the sigmoid function has an "S" shaped curve. The predicted value will become one if the value of x goes to positive infinity and zero if it goes to negative infinity.

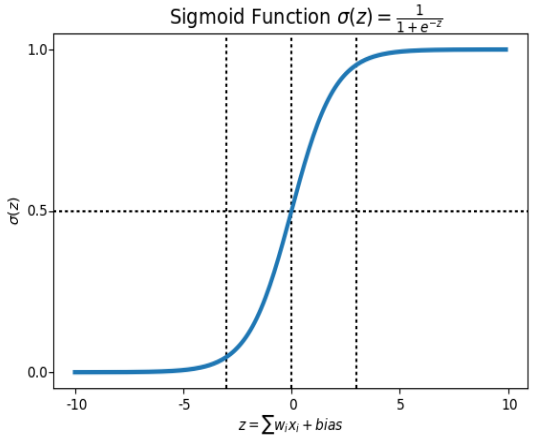

Fig. 2 Sigmoid function
Source: Adapted from [23]

3.3.2. Support Vector Machine (SVM) Model

SVM is a supervised machine learning model that is used to solve classification, regression and outliers problems. As seen in the Fig. 3 this model works by mapping data points into N-dimensional space so that the data can be categorized, and then looking for the most optimal hyperplane as a separator between one class and another class as shown in Fig. 4. [24].

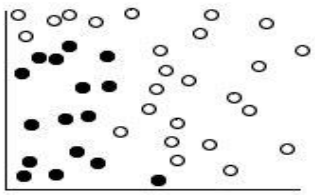

Fig. 3 Mapped dataset
Source: Adapted from [24]

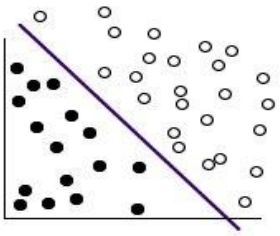

Fig. 4 Mapped dataset with hyperplane
Source: Adapted from [24]

3.4. Metric Evaluation

Metric evaluation is a method used to determine how well a model to performs classification. Evaluation methods include recall, accuracy, precision, and f-score [25]. Accuracy is defined as the average number of right predictions. A higher accuracy value can be present as the better performance. A recall value is the ratio of the accurate prediction to the total amount of correct data, the higher the recall value indicate the better performance. Precision is the ratio of correct prediction to the overall positive prediction result, the higher precision value is better performance. F-score is the average value between precision and recall, the f-score has value from 0 to 1, closer to one is the better performance. Numerous studies have been conducted to classify heart disease. Research conducted by Faria Rahman uses several classification methods such as KNN, SVM, and Decision Tree to predict heart disease, SVM has the highest accuracy value of 93.33% [20].
After the SVM and Logistics Regression models are trained using the dataset, an evaluation will be conducted to determine the performance of the two models. The evaluation metrics that will be used are accuracy, precision, recall and F1-Score.

4. Result

As seen in the Table III and IV. The outcomes of the evaluation metrics f1-score, precision, recall, and accuracy in both models mean that the Logistic Regression model performs better in terms of predicting the occurrence of heart failure in patients than the Support Vector Machine model.

Table 3. Evaluation Results

Model	Logistic Regression	Support Vector Machine(SVM)
F1-Score	0.900	0.743
Precision	0.872	0.750
Recall	0.931	0.735
Accuracy	0.886	0.717

Table 4. Average Performance Score

Model	Average Performance Score
Logistic Regression	0.897
Support Vector Machine (SVM)	0.736

However, in Tables V and VI, after performing the feature scaling and hyperparameter tuning process, the two models have the same performance, this is because SVM performs a classification based on the distance between the data, where the average Logistic Regression performance is 89.3% and the average Logistic Regression performance is 89.3%. SVM performance of 89.7% which makes SVM performance better about 0.4% than Logistic Regression.

Table 5. Evaluation Results After Feature Scaling and Hyperparameter Tuning

Model	Logistic Regression	Support Vector Machine(SVM)
F1-Score	0.896	0.900
Precision	0.863	0.872
Recall	0.931	0.931
Accuracy	0.880	0.886

Table 6. Average Performance Score After Feature Scaling and Hyperparameter Tuning

Model	Average Performance Score
Logistic Regression	0.893
Support Vector Machine (SVM)	0.897

Can be seen in Fig. 5, ROC for the Logistic Regression model is indicated by a blue curve and SVM is indicated by an orange curve, the closer the curve approaches the top left corner, the better the model performs classification. To prove it, we need to calculate the AUC (area under the curve) value, which is the area under the ROC Curve. The closer the AUC value is to one, the better the model. The AUC Support Vector Machine value is around 93.58% and Logistic Regression is around 92.95%, the SVM value is 0.63% higher than Logistic Regression, this shows that the two models have the same performance in classifying heart disease.

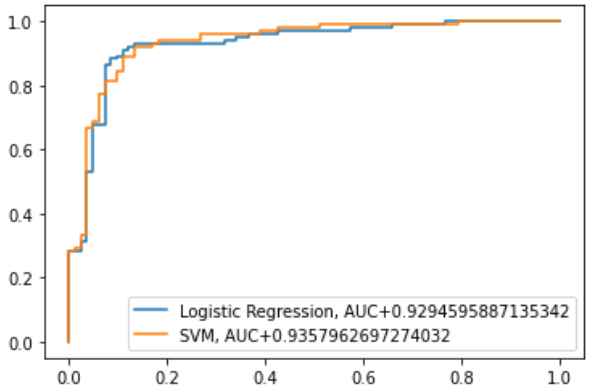

Fig. 2. ROC Curve for Logistic Regression and SVM

5. Conclusions

Both Logistic Regression and Support Vector Machine algorithms have the same performance, which is 89.3% for Logistic regression and 89.7% for Support Vector Machine in classifying heart disease. For the algorithm to have better performance, there are several stages that must be passed, namely processing the dataset used, such as dividing the ratio of training and testing data, encoding, or performing feature scaling, which increases the performance of the SVM model by about 16%. Hyperparameter tuning is also needed to find the appropriate hyperparameter which makes the performance of the two models also increase. From this it can be concluded that the machine learning algorithm has a better performance in classifying if the dataset and parameters used are in accordance with the algorithm.
In further research, it is suggested that implementing PCA on the dataset might improve the performance of the model because the dataset has high dimensions, after that resampling the unbalanced dataset such as the target variable, and finally trying to compare other machine learning models. to perform classification, such as Random Forest, Decision Tree, K-Nearest Neighbours, and others classification algorithm.

References

[1] World Health Organization, "Cardiovascular disaeses," World Health Organizataion, [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases. [Accessed 26 4 2022].

[2] IBM Cloud Education, "Machine Learning," IBM, 15 July 2020. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning. [Accessed 26 4 2022].

[3] Fedesoriano, "Heart Failure Prediction Dataset," Kaggle, September 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction. [Accessed 26 4 2022].

[4] M. Gupta and S. D. Pandya, "A Comparative Study on Supervised Machine Learning Algorithm," Internatonal Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 10, no. 1, pp. 1023-1028, 2022.

[5] G. S. Nayak, S. P. Patro and N. Padhy, "Heart disease prediction by using novel optimization algorithm: A supervised learning prospective," Informatics in Medicine Unlocked, vol. 26, pp. 2-4, 2021.

[6] C. Xiao, Y. Li and Y. Jiang, "Heart Coronary Artery Segmentation and Disease Risk Warning Based on a Deep Learning Algorithm," IEEE Access, pp. 1-1, 2020.

[7] L. Marleni and A. Alhabib, "Faktor Risiko Penyakit Jantung Koroner," Jurnal Kesehatan Politeknik Kesehatan Kementrian Kesehatan Tanjung Karang, vol. 8, no. 3, pp. 478-483, 2017.

[8] H. Jindal, S. Agrawal, R. Khera, R. Jain and P. Nagrath, "Heart disease prediction using machine learning algorithms," IOP Conference Series: Materials Science and Engineering, pp. 1-10, 2021.

[9] M. G. C. Leiro, "Advanced heart failure: a position statement of the Heart Failure Association of the European Society of Cardiology," European Journal of Heart Failure, vol. 20, no. 11, pp. 1505-1535, 2018.

[10] F. Z. Abdeldjouad, B. Menaouer and M. Nada, "A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques," ICOST 2020: The Impact of Digital Technologies on Public Health in Developed and Developing Countries, pp. 299-306, 2020.

[11] B. S. F. Astuti, N. A. Firdausanti and S. W. Purnami, "Model Evaluation for Logistic Regression and Support Vector Machines in Diabetes Problem," INFERENSI, vol. 1, no. 2, pp. 77-82, 2018.

[12] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," Computational Intelligence and Neuroscience, vol. 2021, pp. 1-11, 2021.

[13] S. Mehrolia, S. Alagarsamy and V. M. Solaikutty, "Customers response to online food delivery services during COVID-19 outbreak using binary logistic regression," International Journal of Consumer Studies, vol. 45, no. 3, pp. 396-408, 2020.

[14] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," Cancer Genomics & Proteomics, vol. 15, no. 1, pp. 41-51, 2018.

[15] E. F. U. Latifah, "PERBANDINGAN KINERJA MACHINE LEARNING BERBASIS ALGORITMA SUPPORT VECTOR MACHINE DAN NAIVE BAYES," Universitas Indonesia, 2018.

[16] M. Z. Younas, "Effective Heart Disease Prediction using Machine Learning and Data Mining Techniques," INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY (IRJET), vol. 8, no. 4, pp. 3539-3546, 2021.

[17] T. Ciu and R. Oetama, "Logistic Regression Prediction Model for Cardiovascular Disease," International Journal of New Media Technology, vol. 7, no. 1, pp. 33-38, 2020.

[18] R. Prasad, P. Anjali, S. Adil and N. Deepa, "Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning," International Journal of Engineering and Advanced Technology (IJEAT), vol. 8, no. 3, pp. 659-662, 2019.

[19] Y. Sandhya, "Prediction of Heart Diseases using Support Vector Machine," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. VIII, no. 2, 2020.

[20] F. Rahman and A. M. Mahmood, "A Comprehensive Analysis of Most Relevant Features Causes Heart Disease Using Machine Learning Algorithms," 2022.

[21] M. Shahi and E. R. K. Gurn, "Heart Disease Prediction System Using Data Mining Techniques," Oriental Journal of Computer Science and Technology, vol. 6, no. 4, 2013.

[22] Microsoft, "Microsoft technical documentation - training and testing datasets," Microsoft, 2 April 2022. [Online]. Available: https://docs.microsoft.com/en-us/analysis-services/data-mining/training-and-testing-data-sets. [Accessed 26 May 2022].

[23] S. HP, "Logistic regression in Statistics and Machine learning," Medium, 23 July 2019. [Online]. Available: https://medium.com/mlearning-ai/logistic-regression-60694a973bee. [Accessed 26 May 2022].

[24] IBM, "How SVM Works," IBM, 17 August 2021. [Online]. Available: https://www.ibm.com/docs/it/spss-modeler/SaaS?topic=models-how-svm-works. [Accessed 26 May 2022].

[25] Z. Rustam, D. A. Utami, R. Hidayat, J. Pandelaki and W. A. Nugroho, "Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalanced Cerebral Infraction Datasets," International Journal on Advanced Science Engineering Information Technology, vol. 9, no. 2, pp. 685-691, 2019.

[26] Pusat Data dan Informasi Kementrian Kesehatan RI, "Situasi Kesehatan Jantung," Kementrian Kesehatan Republik Indonesia, 2014.