Predicting Heart Disease using Logistic Regression

by Mochammad Anshori

Submission date: 14-Aug-2023 03:10PM (UTC+0700)
Submission ID: 2145626104
File name: 29576-144582-4-PB_2.pdf (531.98K)
Word count: 4971
Character count: 26714

Knowledge Engineering and Data Science (KEDS)
Vol 5, No 2, December 2022, pp. 188–196

pISSN 2597-4602
eISSN 2597-4637

Predicting Heart Disease using Logistic Regression

Mochammad Anshori [1,*], M. Syauqi Haris [2]

Department Informatics, Institute of Health and Science Technology Rs. dr. Soepraoen Malang, Jl. S. Supriyadi No. 22, Malang, 65147, Indonesia
[1] moanshori@itsk-soepraoen.ac.id; [2] haris@itsk-soepraoen.ac.id;*
** corresponding author*

ARTICLE INFO

Article history:
Received 11 September 2022
Revised 25 October 2022
Accepted 23 December 2022
Published online 30 December 2022

Keywords:
Heart disease
Cardiovascular disease
Classification
Machine learning
Logistic regression

ABSTRACT

A common risk of death is caused by heart disease. It is critical in the field of medicine to be able to diagnose cardiac disease in order to adequately prevent and treat patients. The most accurate method of prediction has the potential to both extend the patient's life and reduce the severity of their cardiac disease. The use of machine learning is one approach that may be taken to generate predictions. In this study, patient medical record information was used in conjunction with an algorithm for logistic regression in order to make heart disease diagnoses. The outcomes of the logistic regression have been utilized to achieve a high level of accuracy in the prediction of heart disease. To get the model coefficients needed for the equation, the experiment uses an iterative form of the logistic regression test. Iteration 14 produced the best results, with an accuracy of 81.3495% and an average calculation time of 0.020 seconds. The best iteration was reached at that point. The percentage of space that lies beneath the ROC curve is 89.36%. The findings of this study have significant implications for the field of heart disease prediction and can contribute to improved patient care and outcomes. Accurate predictions obtained through logistic regression can guide healthcare professionals in identifying individuals at risk and implementing preventive measures or tailored treatment plans. The computational efficiency of the model further enhances its applicability in real-time decision support systems.

I. Introduction

Heart disease (HD), or cardiovascular disease, is a major cause of death worldwide. Based on World Health Organization (WHO) report, there are 17.9 million deaths yearly, and almost 32% of all are passed away [1][2]. According to the WHO page, the cause of heart disease is a heart attack, stroke, and rheumatic. Everyone has the potential for heart disease, especially men compared to the woman. Unhealthy lifestyles, such as smoking, cholesterol, high blood pressure, obesity, alcohol, and hereditary history, become the most critical risk of heart disease [3]. Not all sufferers of heart disease end in death. A controlled lifestyle, such as eating habits and physical activity, can prevent the risk.

Symptoms indicate heart disease, such as shortness of breath [4], physical fatigue [5], and pain in the chest, arms, shoulders, or back [6]. Heart disease can attack the sufferer and is not easy to cure because it needs special treatment. As a vital organ, heart health care must be highly guarded. The most effortless action to take as a preventive measure is to reduce smoking habits, have a healthy diet, be active in physical activities and stop consuming alcohol [7]. The various causes of heart disease may increase the prediction complexity.

With the development of medical data sourced from the patient's health record, there is a great opportunity as a basic material in developing patient health. Currently, the use of computers has been applied in various fields. In health, it can be used to improve the decision-support system in medicine [8]. Especially, implementing machine learning as an analytical tool can find hidden patterns in the data [9]. This development follows up a high degree of prediction in terms of proper prevention.

Prior studies on predicting and classifying heart disease using machine learning techniques are offered. These studies explore various features, methods, and their corresponding accuracies. Some of the notable findings include research that used K-nearest neighbors (KNN) with an accuracy of

https://doi.org/10.17977/um018v5i22022p188-196

74% [10], information gain combined with KNN achieving 99.65% accuracy [11], decision tree (DT) method with 99.62% accuracy [12], and GCSA-DCNN model with 95.34% accuracy [5]. Additionally, feature selection and classification methods such as Chi-squared combined with BayesNet achieved 85% accuracy [13], the FCMIM-support vector machine (SVM) method attained an accuracy of 92.37% [14], PCA combined with random forest (RF) achieved 98.7% accuracy [8]. Other methods like logistic regression (LR) achieved accuracies of 92.58% [15] and 92.76% [9], while a machine learning framework utilizing PSO and support vector machine (SVM) classifier achieved 84.36% accuracy [16]. Ensemble classification techniques, including naive Bayes (NB), Bayesian network (BN), random forest (RF), and multilayer perceptron (MLP), achieved an accuracy of 85.48% [11]. These prior studies contribute to the understanding and development of machine learning approaches for heart disease prediction and classification.

However, machine learning techniques are useful for predicting heart disease. Implementing the machine learning technique may be more advantageous and effective in terms of cost [17]. Various methods are used to predict heart disease accurately and with maximum accuracy. The methods used range from simple to hybrid methods with other methods aimed at increasing the accuracy of the classifier model. Several methods have been used, including NB [18], BN [19], RF [20], MLP [21], SVM [22], KNN [23], LR [24], DT [25], and deep convolutional neural network (DCNN) [26]. The method for preprocessing uses principal component analysis (PCA), chi-squared, and information gain. Optimization methods include particle swarm optimization (PSO), and ant colony optimization (ACO).

This research applied a machine learning algorithm called logistic regression to predict heart disease risk based on risk factors from the patient health records. The logistic regression used is simple logistic regression without any optimization. With this reliability, this study offers the use of logistic regression in classifying heart disease. Previous studies use the same dataset with 14 features, which has resulted in an accuracy of 92.76% [9] and a total of 13 with an accuracy of 92.58% [13]. Based on the result above, logistic regression can provide high accuracy. The difference between the research conducted with previous research is based on the dataset used. This study uses a dataset with a number of features = 9. For the comparison to get the best model, a comparison method is implemented. The model comparisons are based on function classifiers, such as SVM (support vector model) and LDA (linear discriminant analysis). The aim of this study is to know the model of log regression while implemented in this dataset. the fundamental difference between this study and previous research lies in the dataset used. In this research, we used a new dataset that covers symptoms of heart disease that have a total feature less than previous research.

The motivation behind this research stems from the pressing need to improve the accuracy of heart disease prediction models, given the significant impact of heart disease on global health. Accurate and reliable prediction models can aid healthcare professionals in identifying high-risk individuals and implementing timely preventive measures. By leveraging machine learning algorithms and exploring various features and methods, we aim to contribute to the development of more effective and efficient heart disease prediction models. The findings of this research can potentially enhance medical decision-making processes, improve patient outcomes, and ultimately reduce the burden of heart disease on individuals and healthcare systems.

This research contributes to the existing body of knowledge on heart disease prediction by focusing on a specific dataset with a reduced number of features. While previous studies have achieved high accuracies using more comprehensive datasets, this research explores the potential of logistic regression with a limited feature set. By evaluating the performance of logistic regression and comparing it with other classifiers, such as SVM and LDA, we aim to provide insights into the effectiveness of logistic regression in predicting heart disease using a more compact dataset. The findings of this study can shed light on the trade-offs between feature selection and predictive accuracy, offering valuable guidance for future research and the development of practical heart disease prediction models.

The remaining sections of this paper are organized as follows. Section II provides a detailed explanation of the methodology used, including data collection, data preparation, and the

implementation of logistic regression, SVM, and LDA classifiers. Section III presents the experimental results and performance evaluation metrics, comparing the accuracies of different classifiers. Additionally, a discussion of the findings and their implications will be provided in this section. Finally, Section IV concludes the paper, by summarizing the key findings and their significance in the field of heart disease prediction, the limitations of the study, and potential areas for future research.

II. Method

In this research, a systematic methodology consisting of four stages represent in Figure 1. Figure 1 provides an overview of these stages, which include dataset loading, dataset preparation, model creation using the selected method, and result evaluation.

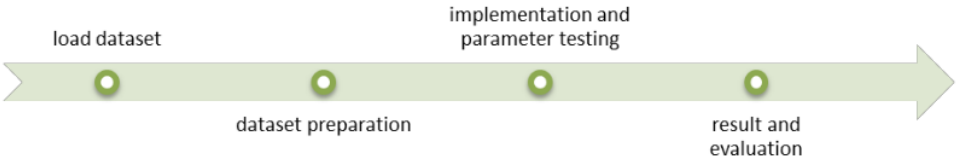

Fig. 1. Research methodology

The initial stage involves preparing the dataset for analysis. The dataset used in this research was obtained from the Mendeley dataset [1]. This dataset contains information on observable characteristics and risk factors associated with heart attacks. The data instances were collected from electronic health records of patients. In total, the dataset comprises 1319 data instances, each representing a patient's information. The data comparison with positive and negative labels can be seen in Figure 2.

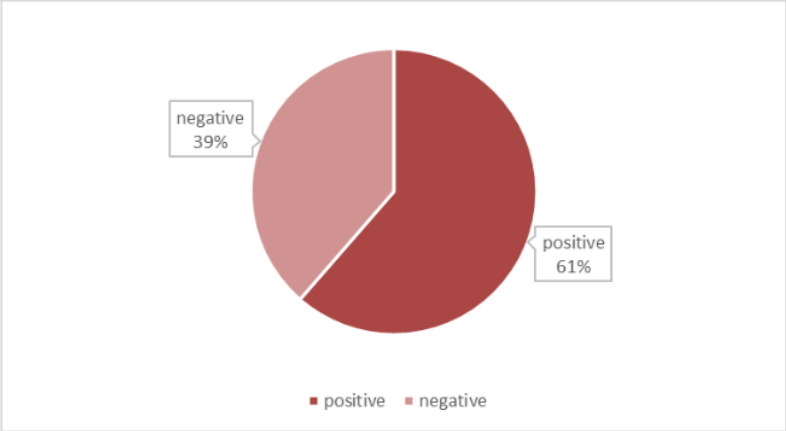

Fig. 2. Target class demographics

Figure 2 provides a visualization of the distribution of positive and negative labels in the dataset. Based on Figure 2, 61% of the data was labeled positive, and the remaining 39% was labeled negative. From the figure, the instance data with a positive class has more quantity than those with a negative label. The dataset has features unlocked 9. The details of the features in the dataset are shown in Table 1. If observed, all data types of each feature are numeric. It indicates that the nominal data has been converted to numeric, making it easier for the model to perform calculations. On the other hand, it makes it easier for researchers to process data because there is no need to convert nominal data types.

Table 1. Dataset details

No	Feature	Data type	Range	Description
1	age	numeric	14, 103	Age of patient
2	gender	numeric	0, 1	1 = male, 0 = female
3	impulse	numeric	20, 1111	Heart rate
4	Pressure high	numeric	42, 223	Systolic blood pressure
5	Pressure low	numeric	38, 154	Diastolic blood pressure
6	glucose	numeric	35, 541	Blood sugar
7	kcm	numeric	0.321, 300	CK-MB
8	troponin	numeric	0.001, 10.3	Test troponin
9	class	nominal	0, 1	Positive/negative

The second stage is separating the data between training data and test data. The data shared is used to build a classifier model. The scheme used in data sharing is the k-fold cross-validation method. This method is applied because the resulting model is more general and can avoid overfitting [27]. Cross-validation works based on the value of the parameter k. The value of k here determines how many data segments are shared between test data and training data. The illustration of cross-validation can be seen in Figure 3.

Fig. 3. Illustration of cross-validation with k-fold = 10

Figure 3 shows cross-validation for this research with a value of k = 10. The gray cells will be the test data for each section and run iteratively for the value of k. The parameter k used in this study is 10-fold cross-validation, meaning the data is divided into 10 subsets. Each subset is used as the test set once, while the remaining nine subsets are combined to form the training set. This iterative process ensures that the model is evaluated on different combinations of training and test data, providing a more robust assessment of its predictive capabilities. By utilizing the k-fold cross-validation technique, this research aims to build a classifier model that can generalize well to unseen data. This approach helps to assess the model's performance and determine its ability to accurately predict heart disease in new and unseen cases

The third stage is creating a LR classification model. LR is a mathematical model that uses probability estimation for each class [28]. LR is one of the supervised learning methods. In this case, LR uses to overcome the binary classification. However, generally, LR is also reliable in the case of multi-label classification. The advantages of LR are that it does not require a lot of parameter optimization and is easy to implement [29].

The LR model operates similarly to linear regression, as seen in (1). However, the primary distinction lies in the function used. In LR, the sigmoid function, shown in (2), is employed within the equation. By substituting the sigmoid function into (1), (3) is derived. Equation (4) represents the formulation of logistic regression as a logit, known as the log probability function. The term inside the brackets is referred to as the odds, representing the ratio of the probability of success to the

probability of failure. The LR coefficients are estimated using the iteratively reweighted least squares (IRLS) method [30]. In each iteration, the dependent variable is adjusted to obtain the optimal LR coefficient.

$$\hat{y} = E(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \in \quad (1)$$

$$\sigma(Z) = \frac{1}{1+e^{-z}} \quad (2)$$

$$E(y|x) = sigma(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n) \quad (3)$$

$$E(y|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \quad (4)$$

where $\hat{y}$ represents the predicted value of the dependent variable y given the independent variables x_1, x_2, ..., x_n. The coefficients β_0, β_1, ..., β_n are estimated parameters that determine the relationship between the independent variables and the dependent variable. The term $\in$ represents the error term or residual. Z represents the linear combination of the coefficients and independent variables.

Comparison is needed to obtain the best method. The model comparison that will be used is SVM and LDA. SVM generally works by splitting data class based on the hyperplane. The SVM function is shown in (5).

$$L_D = \sum_{i=1}^{n} \propto_i - {}^1/_2 \sum_{i=1}^{n} \sum_{j=1}^{n} \propto_i \propto_j y_i y_j x_i^T x_j \quad (5)$$

L_D represents the SVM function, $\propto_i$ and $\propto_j$ are the weights assigned to the data points, y_i and y_j are the class labels, and x_i and x_j are the feature vectors. The objective of SVM is to find the optimal weights that maximize the margin between the classes.

On the other hand, LDA works by projecting all data vectors linearly. LDA optimize the distance between class and minimize the distance between inner class. The LDA formula is shown in (8). The equation is formed from covariance at (6) and pooled covariance at (7).

$$c_i = \frac{(x_i^0)^T x_i^0}{n_i} \quad (6)$$

$$c_{(r,s)} = \frac{1}{n} \sum_{i=1}^{g} n_i c_i(r,s) \quad (7)$$

$$f_i = \mu_i C^{-1} x_k^T - {}^1/_2 \mu_i C^{-1} x_i^T + \ln(p_i) \quad (8)$$

where c_i represents the covariance for each class, n_i is the number of instances in class i, x_i^0 denotes the centered data for class i, g is the total number of classes, μ_i is the mean vector for class i, C^{-1} is the inverse of the covariance matrix, x_k^T is the transpose of the centered data, and p_i is the prior probability of class i.

The researcher uses a performance reference as an accuracy value as a benchmark in comparing the results in the fourth stage. The formula for calculating accuracy is shown in (9) below. It also uses TPR (true positive rate) and FPR (false positive rate) to get the ROC curve value [31]. ROC here is valid for modeling errors/errors from the built classification model. FPR and TPR can be seen in (10) and (11) below for the accuracy formula. TP means that it is correct and predicted correctly, TN is correct, but the prediction is wrong, FP is wrong but predicted right, and FN is wrong and predicted wrong.

$$accuracy = \frac{TP+TN}{total\ data} \quad (9)$$

$$TPR = \frac{TP}{TP+FN} 100\% \quad (10)$$

$$FPR = \frac{TP}{TP+FN} 100\% \quad (11)$$

III. Results and Discussion

The results of this study are by observing the results of logistic regression performance. The application of logistic regression uses the Weka application [32]. There is no data preprocessing here because the data obtained is considered clean. The IRLS iteration test carried out to obtain the logistic regression coefficient. The parameter values tested are 2 to 30 with multiples of 2. The iteration test results can be seen in Figure 4.

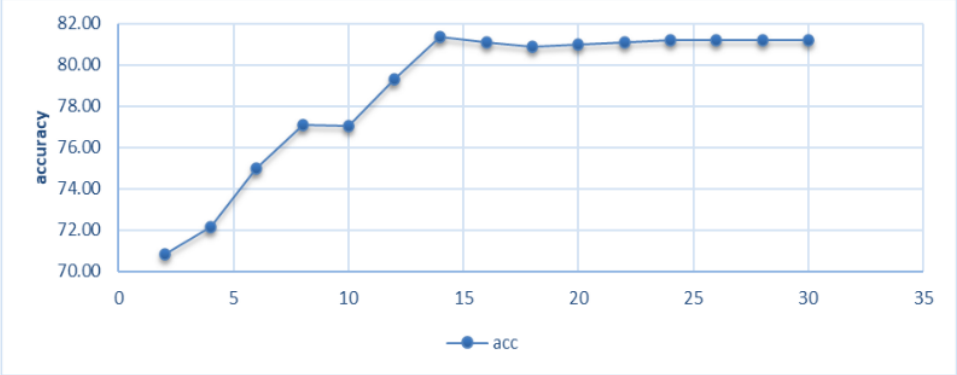

Fig. 4. Iteration parameter testing

Figure 4 shows a graph of the change in accuracy of each iteration test. When the iteration is low, the accuracy obtained is also low. The greater the iteration value, the higher the accuracy value. At iteration = 10, there is a decrease in the accuracy value compared to the accuracy at iteration 8. It shows that iteration = 10 is the optimal locale because the accuracy increases and decreases again. Furthermore, at iteration = 14, it produces an accuracy that tends to be high, namely 81.35%. During this iteration, the logistic regression model can produce the best accuracy because when the accuracy is increased again, it decreases accuracy, and there tends to be no change in the increase or decrease in accuracy. Based on these findings, it can be concluded that the logistic regression model achieved the best accuracy at iteration = 14. This information is crucial for selecting the optimal logistic regression coefficients and maximizing the predictive power of the model.

The accuracy of logistic regression was obtained, then the model was compared. The comparison is shown in Table 2. The table shows the evaluation measure such as accuracy, TPR, FPR, and computational time. The time value is second and obtained from ten times rials. The table shows the accuracy of log regression = 81.35%, SVM with linear kernel = 78.17%, and LDA gives accuracy = 69.75%. These results give the highest accuracy from the log regression model. Linearly, the TPR value is also rising to the increase in the accuracy value. Unlike the FPR value, which is inversely proportional, the value will be smaller if the TPR value increases. For the computational time, LDA gives the worst time equal to 0.17 seconds. SVM reach about 0.06 second, better than LDA. The best computational is gained from log regression, which only needs 0.02 seconds to do classification.

Table 2. Classification results based on the LR model and its comparison.

Evaluation	Log Regression	SVM (linear)	LDA
Accuracy	**81.3495**	78.1653	69.7498
TPR	**81.3**	78.2	69.7
FPR	**18.7**	23.4	39.5
Time (s)	**0.02**	0.06	0.17

In Table 2 several evaluations of the performance of the logistic regression obtained from the confusion matrix. Based on these results, it can be said that logistic regression can be used to predict heart disease with high accuracy. The TPR (sensitivity) was correctly calculated, and the calculated FPR was incorrectly identified [33]. Computational time is also included in the calculation. The computational time obtained resulted in 10 times of testing to get the average. The average value of

computing time is 0.02 seconds. Based on the computational time generated, the prediction model with logistic regression has a relatively fast computation time. Next, consider Figure 5.

Table 3. Confusion matrix

Actual	Predicted	
	Positive	Negative
Positive	660	150
Negative	96	413

Since we know if log regression is the best model in this case, let us see the confusion matrix. Using iteration = 14, the results of the evaluation of the implementation of logistic regression are shown in the confusion matrix table in Table 3. The confusion matrix/error matrix is used to visualize the performance of the logistic regression algorithm. The confusion matrix represents the result between the actual and predicted values. The table shows the value of TP = 660, TN 413, FP = 96, FN = 150. Table 3 shows that the classifier cannot predict all the data accurately. From the confusion matrix table above, there are still misclassifications. Next, consider Figure 5. The picture represents ROC of the model performance. The ROC is generated based on the log regression model.

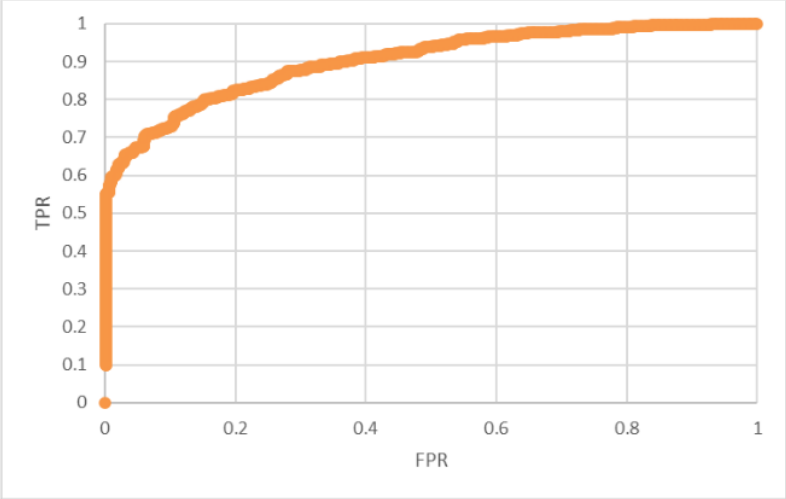

Fig. 5.ROC curve

Figure 5 shows the ROC curve, which is a combination of the x and y axes, TPR occupies the x-axis and FPR on the y-axis. By being able to visualize the performance of the classifier in making predictions [33]. ROC The value of the ROC curve in Figure 5 is 89.36. This value is good because it is close to 1, which is the best value of the ROC curve. A good curve has a value between 0.5 up to 1 it means that the curve produced by logistic regression is close to its best value. It is proven that the classifier's performance is suitable for predicting heart disease.

Accurately predicting heart disease risk is crucial for developing effective decision-support systems in healthcare. The findings of this research contribute to the development of such systems by providing insights into the performance and feasibility of logistic regression as a predictive model. Integrating logistic regression-based algorithms into decision support systems can assist healthcare professionals in identifying individuals at high risk of heart disease and making informed decisions regarding prevention and treatment strategies. These findings highlight the effectiveness of logistic regression as a predictive model for heart disease. Despite misclassifications, the model exhibited high accuracy, relatively fast computational time, and a good ROC curve. These results donate to understanding logistic regression's potential in heart disease prediction and can inform the development of more accurate and efficient prediction models.

IV. Conclusion

Referring to the results and discussion, the machine learning method, namely logistic regression, can predict heart disease based on the patient's electronic medical record. A dataset used in this study has a total feature = 9 and 1319 instances of data. Based on the iteration parameter test results, the increase in the iteration value affects the accuracy value of the classifier model. It was found that the best iteration that can produce the highest accuracy at iteration = 14. The given accuracy is 81.3495%. The difference in iteration values affects the performance of logistic regression, as evidenced by the increasing iteration value providing an increase in accuracy until finding the optimal point. Log regression is proven more reliable in making predictions with relatively high accuracy and relatively fast computation time. Further research for this study by comparing some machine learning models, namely SVM and LDA. Feature selection can be made in further research from this study to get a better model.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

References

[1] S. S. Maghdid and T. A. Rashid, "An Extensive Dataset for the Heart Disease Classification System," Mendeley Data, 2022.
[2] WHO, "Cardiovascular diseases," World Health Organization, 2020. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed Aug. 08, 2022).
[3] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics Med. Unlocked, vol. 16, p. 100203, 2019.
[4] A. Alshukry et al., "Clinical characteristics of coronavirus disease 2019 (COVID-19) patients in Kuwait," PLoS One, vol. 15, no. 11, p. e0242768, Nov. 2020.
[5] S. M. Nagarajan, V. Muthukumaran, R. Murugesan, R. B. Joseph, M. Meram, and A. Prathik, "Innovative feature selection and classification model for heart disease prediction," J. Reliab. Intell. Environ., vol. 8, no. 4, pp. 333–343, Dec. 2022.
[6] S.-J. Kim, "Global Awareness of Myocardial Infarction Symptoms in General Population," Korean Circ. J., vol. 51, no. 12, p. 997, 2021.
[7] R. Ndejjo, G. Musinguzi, F. Nuwaha, H. Bastiaens, and R. K. Wanyenze, "Understanding factors influencing uptake of healthy lifestyle practices among adults following a community cardiovascular disease prevention programme in Mukono and Buikwe districts in Uganda: A qualitative study," PLoS One, vol. 17, no. 2, p. e0263867, Feb. 2022.
[8] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Informatics Med. Unlocked, vol. 19, p. 100330, 2020.
[9] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, and M. A. Hossain, "Comparative Analysis of Classification Approaches for Heart Disease Prediction," in 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Feb. 2018, pp. 1–4.
[10] M. Anshori, F. Mar'i, and F. A. Bachtiar, "Comparison of Machine Learning Methods for Android Malicious Software Classification based on System Call," in 2019 International Conference on Sustainable Information Engineering and Technology (SIET), Sep. 2019, pp. 343–348.
[11] P. Thombare, M. Ghalme, S. Raut, N. Dhakne, and P. R. Dholi, "Prediction of Heart Disease using Machine Learning Techniques," Int. Res. J. Mod. Eng. Technol. Sci., vol. 04, no. 06, pp. 1099–1102.2022.

[12] H. Gulfam Ahmad and M. Jasim Shah, "Prediction of Cardiovascular Diseases (CVDs) Using Machine Learning Techniques in Health," Azerbaijan J. High Perform. Comput., vol. 4, no. 2, pp. 267–279, Dec. 2021.
[13] S. D. Desai, S. Giraddi, P. Narayankar, N. R. Pudakalakatti, and S. Sulegaon, "Back-Propagation Neural Network Versus Logistic Regression in Heart Disease Classification," in Advances in Intelligent Systems and Computing, 2019, pp. 133–144.
[14] W. Książek, M. Gandor, and P. Pławiak, "Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma," Comput. Biol. Med., vol. 134, p. 104431, Jul. 2021.
[15] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, pp. 107562–107582, 2020.
[16] J. Vijayashree and H. P. Sultana, "A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier," Program. Comput. Softw., vol. 44, no. 6, pp. 388–397, Nov. 2018.
[17] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, vol. 16, no. 2, p. 88, Feb. 2023.
[18] L. Ali et al., "A Feature-Driven Decision Support System for Heart Failure Prediction Based on X2 Statistical Model and Gaussian Naive Bayes," Comput. Math. Methods Med., vol. 2019, pp. 1–8, Nov. 2019.
[19] A. Elsayad and M. Fakhr, "Diagnosis of cardiovascular diseases with bayesian classifiers," J. Comput. Sci., vol. 11, no. 2, pp. 274–282, 2015.
[20] S. Asadi, S. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," J. Biomed. Inform., vol. 115, p. 103690, Mar. 2021.
[21] K. Subhadra and B. Vikas, "Neural network based intelligent system for predicting heart disease," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 5, pp. 484–487, 2019.
[22] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," IEEE Access, vol. 7, pp. 54007–54014, 2019.
[23] R. TR, U. K. Lilhore, P. M, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive analysis of heart diseases with machine learning approaches," Malaysian J. Comput. Sci., pp. 132–148, Mar. 2022.
[24] S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques," IETE J. Res., vol. 68, no. 4, pp. 2488–2507, Jul. 2022.
[25] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," Comput. Methods Programs Biomed., vol. 192, p. 105400, Aug. 2020.
[26] T. K. Sajja and H. K. Kalluri, "A Deep Learning Method for Prediction of Cardiovascular Disease Using Convolutional Neural Network," Rev. d'Intelligence Artif., vol. 34, no. 5, pp. 601–606, Nov. 2020.
[27] S. Nusinovici et al., "Logistic regression was as good as machine learning for predicting major chronic diseases," J. Clin. Epidemiol., vol. 122, pp. 56–69, Jun. 2020.
[28] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," J. Appl. Sci. Technol. Trends, vol. 1, no. 4, pp. 140–147, 2020.
[29] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm," IEEE Access, vol. 10, pp. 3284–3293, 2022.
[30] A. Swift, R. Heale, and A. Twycross, "What are sensitivity and specificity?," Evid. Based Nurs., vol. 23, no. 1, pp. 2–4, Jan. 2020.
[31] E. Frank, M. A. Hall, and I. H. Witten, The WEKA workbench. Morgan Kaufmann, 2016.
[32] K. Kirasich, T. Smith, and B. Sadler, "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," SMU Data Sci. Rev., vol. 1, no. 3, p. 9, 2018.
[33] L. de S. Rodrigues, E. T. Matsubara, and B. M. Nogueira, "Learning a Fast Bipartite Ranker for Text Documents Using Lexicographical Rankers and ROC Curves," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Nov. 2017, pp. 1307–1312.

Predicting Heart Disease using Logistic Regression

ORIGINALITY REPORT

20%	17%	14%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repository.uin-malang.ac.id Internet Source	3%
2	arxiv.org Internet Source	2%
3	epdf.pub Internet Source	1%
4	discovery.researcher.life Internet Source	1%
5	hdl.handle.net Internet Source	1%
6	ejournal.raharja.ac.id Internet Source	1%
7	"Inventive Communication and Computational Technologies", Springer Science and Business Media LLC, 2021 Publication	<1%
8	Farhanna Mar'i, Gusti Pangestu. "Classification of Fake GPS in GOJEK Application using Logistic Regression", 6th	<1%

	International Conference on Sustainable Information Engineering and Technology 2021, 2021 Publication	
9	dokumen.pub Internet Source	<1 %
10	thesai.org Internet Source	<1 %
11	Submitted to Liverpool John Moores University Student Paper	<1 %
12	www.cs.usask.ca Internet Source	<1 %
13	"Soft Computing and Signal Processing", Springer Science and Business Media LLC, 2022 Publication	<1 %
14	easy.dans.knaw.nl Internet Source	<1 %
15	Alireza Pourkeyvan, Ramin Safa, Ali Sorourkhah. "Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks", Research Square Platform LLC, 2023 Publication	<1 %
16	Submitted to Clark University Student Paper	<1 %

17	Submitted to University of Essex Student Paper	<1%
18	www.ijsr.net Internet Source	<1%
19	Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang. "An improved ensemble learning approach for the prediction of heart disease risk", Informatics in Medicine Unlocked, 2020 Publication	<1%
20	www.journalijar.com Internet Source	<1%
21	www.sciencepubco.com Internet Source	<1%
22	Anzi Ding, Qingyong Zhang, Xinmin Zhou, Bicheng Dai. "Automatic recognition of landslide based on CNN and texture change detection", 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), 2016 Publication	<1%
23	Deepali Yewale, S. P. Vijayragavan. "Comprehensive review on machine learning approach for heart disease prediction: Current status and future prospects", AIP Publishing, 2022 Publication	<1%

24	Sudarshan Nandy, Mainak Adhikari, Venki Balasubramanian, Varun G. Menon, Xingwang Li, Muhammad Zakarya. "An intelligent heart disease prediction system based on swarm-artificial neural network", Neural Computing and Applications, 2021 Publication	<1 %
25	Submitted to University of East Anglia Student Paper	<1 %
26	openaccess.city.ac.uk Internet Source	<1 %
27	vdocuments.mx Internet Source	<1 %
28	www.koreascience.kr Internet Source	<1 %
29	www.mdpi.com Internet Source	<1 %
30	"Advances in Visual Informatics", Springer Science and Business Media LLC, 2021 Publication	<1 %
31	Alina Faskhutdinova, Daria Grigorieva, Bulat Garafutdinov, Vladimir Mokshin. "Investigation of Machine Learning Methods for Stroke Prediction", 2023 IX International Conference on Information Technology and Nanotechnology (ITNT), 2023	<1 %

	Publication	
32	Nor Zakiah Gorment, Ali Selamat, Lim Kok Cheng, Ondrej Krejcar. "Machine Learning Algorithm for Malware Detection: Taxonomy, Current Challenges and Future Directions", IEEE Access, 2023 Publication	<1 %
33	ijcseonline.org Internet Source	<1 %
34	journal.universitasbumigora.ac.id Internet Source	<1 %
35	jurnal.uii.ac.id Internet Source	<1 %
36	koreascience.kr Internet Source	<1 %
37	lppm.unri.ac.id Internet Source	<1 %
38	mdpi-res.com Internet Source	<1 %
39	ngalam.co Internet Source	<1 %
40	pubs.sciepub.com Internet Source	<1 %
41	pubs2.ascee.org Internet Source	<1 %

42	sites.bu.edu Internet Source	<1 %
43	www.coursehero.com Internet Source	<1 %
44	www.seminar.uad.ac.id Internet Source	<1 %
45	Aaron A Izang, Nicolae Goga, Shade O., Olujimi D., Ayokunle A., Adesina K.. "Scalable Data Analytics Market Basket Model for Transactional Data Streams", International Journal of Advanced Computer Science and Applications, 2019 Publication	<1 %
46	Surfraz Mitegar, M. Samba Sivudu, Giridhar Akula. "Chapter 24 Sophisticated Machine Learning Algorithms for Pre-investigation of Heart Disease", Springer Science and Business Media LLC, 2023 Publication	<1 %
47	e-journals.dinamika.ac.id Internet Source	<1 %
48	journal.uinjkt.ac.id Internet Source	<1 %
49	tnsroindia.org.in Internet Source	<1 %
	vdoc.pub	

50	Internet Source	<1%
51	www.hindawi.com Internet Source	<1%
52	www.jatit.org Internet Source	<1%
53	Ahsanullah Yunas Mahmoud, Daniel Neagu, Daniele Scrimieri, Amr Rashad Ahmed Abdullatif. "Early diagnosis and personalised treatment focusing on synthetic data modelling: Novel visual learning approach in healthcare", Computers in Biology and Medicine, 2023 Publication	<1%
54	Shahrokh Asadi, SeyedEhsan Roshan, Michael W. Kattan. "Random forest swarm optimization-based for heart diseases diagnosis", Journal of Biomedical Informatics, 2021 Publication	<1%
55	Shtwai Alsubai, Abdullah Alqahtani, Adel Binbusayyis, Mohemmed Sha, Abdu Gumaei, Shuihua Wang. "Heart Failure Detection Using Instance Quantum Circuit Approach and Traditional Predictive Analysis", Mathematics, 2023 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Predicting Heart Disease using Logistic Regression

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9