International Journal for Modern Trends in Science and Technology
Volume 10, Issue 01, pages 82-87.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol10issue01.html
DOI: https://doi.org/10.46501/IJMTST1001011

Cardiovascular Disease Prediction using Logistic Regression

L. Jyothi[1] | D. Sattibabu[2]

[1]Department of Computer Science and Engineering, Godavari Institute of Engineering and Technology, Rajahmundry, Andhra Pradesh, India. jyothilekkala24@gmail.com

[2]Department of Computer Science and Engineering, Godavari Institute of Engineering and Technology, Rajahmundry, Andhra Pradesh, India. sattibabu538@gmail.com

To Cite this Article
L. Jyothi and D. Sattibabu, Cardiovascular Disease Prediction using Logistic Regression, International Journal for Modern Trends in Science and Technology, 2024, 10(01), pages. 82-87. https://doi.org/10.46501/IJMTST1001011

Article Info
Received: 02 January 2024; Accepted: 22 January 2024; Published: 22 January 2024.

ABSTRACT

Cardiovascular disease (CVD) poses a significant global health threat, necessitating effective preventive strategies. This study employs logistic regression to predict CVD and diabetes based on demographic, lifestyle, and medical history data. Utilizing a diverse dataset, logistic regression models analyze relationships between risk factors and disease likelihood. Feature selection enhances model accuracy, identifying key risk factors for valuable insights. Results demonstrate logistic regression's efficacy in assessing disease probability. Identified key features enable healthcare professionals to personalize risk assessments, providing practical tools for early intervention and prevention. This research emphasizes personalized medicine in healthcare predictive analytics, establishing logistic regression as a reliable method for disease prediction. The study lays the foundation for refining predictive models to improve patient outcomes.

Keywords: Cardiovascular disease (CVD), Logistic regression, Risk prediction, Preventive measures

1. INTRODUCTION

The primary challenge in the medical industry is to enhance health infrastructure for early disease diagnosis and timely treatment, particularly considering that 31% of global mortality is attributed to cardiac diseases [1,2]. Developing and underdeveloped nations face obstacles in predicting diseases early due to limited infrastructure and medical resources. The growth of information and telecommunication technology has bridged this gap, providing real-time, cost-effective information to patients and significantly increasing detailed health records. However, managing vast medical data poses a challenge.

Machine learning becomes crucial in swiftly transforming this extensive medical data into valuable insights. Researchers employ machine learning models to uncover hidden patterns and correlations within datasets [3,4]. Given the inconsistency and redundancy in medical datasets, proper preprocessing is pivotal [5]. Feature selection, especially considering the 14 prevalent features associated with cardiac diseases, significantly impacts the accuracy of predictive models [6].Machine

learning, when applied accurately, can greatly improve the accuracy of predicting cardiac diseases, enabling early-stage diagnosis and treatment. This advancement has the potential to save numerous lives by facilitating swift disease diagnosis for a large number of patients.
In this study, the application of a Logistic Regression classifier model is employed, and the results are compared with findings from prior research.

Several studies have delved into the application of machine learning models for the classification of heart disease. FirdaAninditaLatifah et al. conducted a comparative analysis, employing logistic regression and random forest on the Framingham dataset, achieving a notable 85.04% accuracy [7]. Zameer Khan et al. expanded the exploration, utilizing multiple machine learning algorithms, including logistic regression, with their model achieving an accuracy of 85.71% on the UCI Cleveland dataset [8]. ThanujaNishadi A S proposed a logistic regression model on the Framingham dataset, attaining an accuracy of 86.66% [9]. Montu Saw et al. focused on logistic regression for cardiac disease classification, reporting an accuracy of 87.02% using Framingham datasets [10]. Saba Bashir et al. explored various ML algorithms on the UCI dataset, with logistic regression reaching 82.56% accuracy and logistic regression support vector machine achieving 84.85% accuracy [12]. Kannan, R et al. compared multiple ML algorithms, highlighting logistic regression's superior accuracy of 86.51% using the UCI Cleveland dataset [13]. Ganesan M et al. proposed four ML algorithms, with logistic regression achieving 83.70% accuracy on the UCI dataset [14]. Dinesh Kumar G et al. introduced five machine learning algorithms, with logistic regression leading with an overall accuracy of 86.51% for cardiac disease classification [15]. These studies collectively showcase the versatility and effectiveness of logistic regression in the context of cardiac disease prediction.

This paper specifically applies logistic regression without optimization to predict heart disease risk based on patient health records. Despite its simplicity, logistic regression exhibits reliability in classifying heart disease, with comparable accuracy to previous studies that used larger datasets. The study focuses on a dataset with nine features and employs model comparisons with classifiers like Support Vector Model (SVM) and Linear Discriminant Analysis (LDA). The objective is to identify the most effective logistic regression model for this dataset, emphasizing the difference in features from previous research.

This investigation is motivated by the urgent necessity to enhance the precision of models predicting heart disease, given the substantial global impact of this health concern. The development of accurate models is crucial for aiding healthcare professionals in identifying individuals at high risk and implementing timely preventive measures. By employing machine learning algorithms, we aim to explore diverse features and methodologies to contribute to the refinement of more effective heart disease prediction models. The potential outcomes of this study could significantly improve medical decision-making, elevate patient outcomes, and alleviate the burden of heart disease on individuals and healthcare systems.

This research builds upon existing knowledge in heart disease prediction by concentrating on a specific dataset with a reduced number of features. While prior studies have achieved notable accuracies with larger datasets, our investigation delves into the effectiveness of logistic regression using a more concise feature set. Through comparisons with alternative classifiers such as SVM and LDA, our objective is to offer insights into the efficiency of logistic regression in predicting heart disease within the constraints of a more streamlined dataset. The findings from this study may provide valuable information on striking a balance between feature selection and predictive accuracy, thereby guiding future research toward the development of practical heart disease prediction models.

2. LOGISTIC REGRESSION

Logistic Regression stands out as one of the most straightforward and effective machine learning classification algorithms. Widely employed in various applications, it operates as a supervised binary classification algorithm, particularly suited for categorical dependent variables with outcomes in the form of discrete or binary categories. The logistic function is a crucial element in logistic regression, a machine learning approach used for classification tasks. This function generates an S-shaped curve that reflects probabilities, making it suitable for applications such as predicting cardiac disease. It is versatile, capable of

handling both continuous and discrete datasets, and excels at categorizing observations.The logistic function, visualized in the Figure 1below, plays a central role in this classification process.

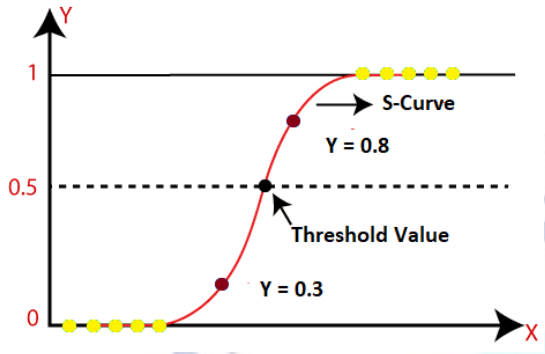

Figure 1: Model of logistic regression.

The algorithm utilizes the sigmoid function as its cost function, which transforms predicted real values into probabilistic values within the range of '0' and '1.Logistic Sigmoid function is shown in Eq. (1).

$$P(x) = \frac{1}{(1+e\Lambda(-x))} \quad (1)$$

Here, $P(x)$ represents the probability estimation function, producing a value between 0 and 1. $x$denotes the input to the probability function, which corresponds to the algorithm's prediction value. The mathematical constant$e$, Euler's number, holds an approximate value of 2.71828.

3. PROPOSED METHOD

This paper follows a systematic methodology outlined in Figure 2. The process involves loading the dataset, preparing the dataset, creating the model using the selected method, and assessing the results. Figure 2 provides a visual summary of these crucial stages in the proposed workflow.

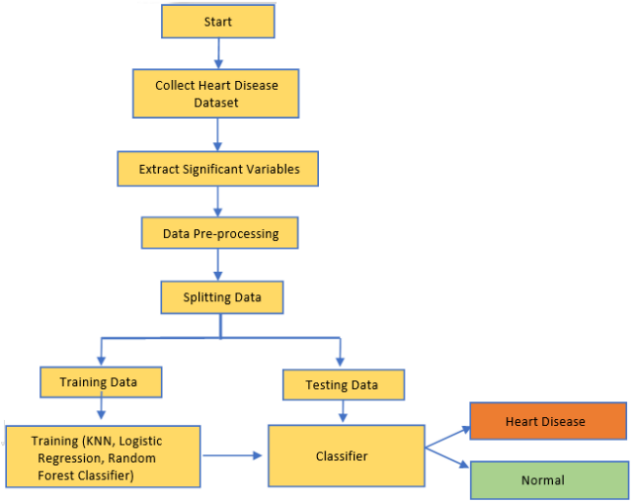

Figure 2: Proposed logistic regression model for CVD.

In the first stage, prepare the dataset for analysis. The dataset obtained from the UCI [1], contains information about observable characteristics and risk factors related to heart attacks. These data points were extracted from electronic health records, amounting to 1319 instances, each representing an individual's details. Figure 3 offers a visual depiction of the data distribution between +ve and -ve labels. The figure demonstrates how +ve and -ve labels are distributed in the dataset. According to the visual representation, 61% of the data is categorized as positive, whereas the remaining 39% is classified as negative. Positive instances dominate over negative instances in the dataset. The dataset encompasses nine unlocked features, all possessing numeric data types, suggesting the conversion of nominal data to numeric for simplifying modeling and analysis.

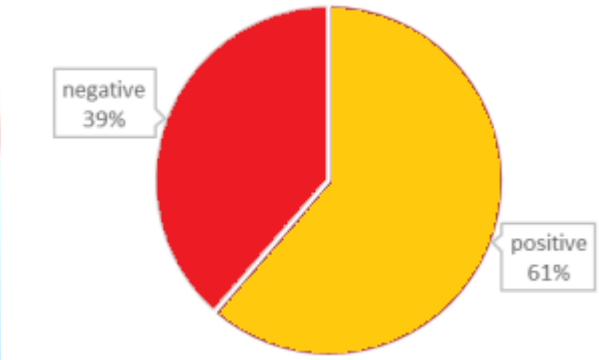

Figure 3: Distribution of Target Class Demographics.

During the second phase, the data undergoes segmentation into training and test sets through the utilization of the k-fold cross-validation technique. The parameter k dictates the number of data segments shared between the test and training datasets. Figure 4 presents a visual representation of the cross-validation procedure. The figure outlines the application of the 10-fold cross-validation method in this research. Cells highlighted in the figure signify the test data for each section, iteratively executed for the specified value of k. In this investigation, a 10-fold cross-validation with k = 10 is implemented, dividing the data into ten subsets. Each subset functions as the test set once, while the remaining nine subsets amalgamate to create the training set. This iterative approach ensures a robust evaluation, enhancing the model's capacity to generalize and make accurate predictions concerning heart disease in novel and unseen cases.

Figure 4: 10-Fold Cross-Validation Process.

In the third stage, a Logistic Regression (LR) classification model is created. LR is a mathematical model that utilizes probability estimation for each class, making it suitable for binary and multi-label classification. LR is advantageous for its simplicity, requiring minimal parameter optimization, and ease of implementation [29].Table 1 below illustrates the division of the dataset into training and testing sets, expressed as percentages. Training involves four splitting conditions, and subsequent testing with specific data aims for optimal accuracy and understanding of the model's behavior, yielding categorizations of 1 and 0 for disease presence and absence.

Table 1: Percentage of Training and Test Data.

S.No	Training Data	Test Data
Category-1	80%	20%
Category-2	70%	30%
Category-3	60%	40%
Category-4	50%	50%

/* Algorithm for Linear Regression Model for Cardiovascular Disease Prediction */

```
// Input: Feature-selected dataset with predictors (X) and target variable (y)
// Output: Trained linear regression model

// Definehyperparameters
learning_rate = 0.01
num_iterations = 1000

// Initialize weights and bias
initialize_weights()

// Gradient Descent for Model Training
for iteration in range(num_iterations):
    // Forward Pass
predictions = predict(X)

    // Calculate loss
loss = calculate_loss(predictions, y)

    // Backward Pass (Gradient Descent)
gradient = compute_gradient(X, predictions, y)

    // Update weights and bias
update_weights(gradient, learning_rate)

// Output: Trained linear regression model

// Function to initialize weights
functioninitialize_weights():
    // Initialize weights (coefficients) and bias
    // You can set them to small random values or zeros

// Function to make predictions
function predict(X):
    // Calculate linear combination of predictors and weights
return X * weights + bias

// Function to calculate mean squared loss
functioncalculate_loss(predictions, y):
    // Mean Squared Error (MSE) loss
return mean((predictions - y)^2)

// Function to compute gradient for gradient descent
functioncompute_gradient(X, predictions, y):
    // Gradient of the loss with respect to weights and bias
return 2 * mean(X * (predictions - y), axis=0), 2 * mean(predictions - y)

// Function to update weights and bias
functionupdate_weights(gradient, learning_rate):
    // Update weights and bias using gradient descent
weights = weights - learning_rate * gradient[0]
bias = bias - learning_rate * gradient[1]
```

4. RESULTS AND DISCUSSION

The proposed logistic regression model is evaluated using the UCI dataset across four different training and testing split ratios, with the corresponding

accuracies presented in the Table 1.In data processing for the UCI Heart Disease Dataset, the target variable initially includes values (0, 1, 2, 3, 4), where 0 signifies good health (no cardiovascular illness), and (1, 2, 3, 4) indicate the presence of cardiovascular disease. To focus on detecting the absence or presence of heart disease, the classes have been condensed to (0, 1) by reducing levels (1, 2, 3, 4) to 1.In this study, MinMax normalization was employed, a technique also known as feature scaling. This approach rescales the numeric values within a data feature to a range between 0 and 1.Through linear transformation, the original data undergoes normalization. The formula used for replacement is:

$$X_{ij}^{'} = \frac{X_{ij} - X_{min}}{X_{max} - X_{min}} \quad (2)$$

Here, j represents the index of the feature being normalized. The outcome of this normalization process is that all the features are brought to a common scale, facilitating more effective comparisons and analyses in subsequent stages of the study.

The dataset exhibits missing values in key patient records like fasting blood sugar and major vessels, posing a challenge for analysis. The initial suggestion of removing or filling missing values with defaults proved inadequate, as it led to a reduction in training data. To address this, a more comprehensive approach was adopted, utilizing techniques such as Mean Value and K Nearest Neighbor (KNN) for filling missing cells. Mean Value replaces missing values with the mean of each column, while KNN approximates missing values based on the characteristics of the nearest neighbors. The selection of the optimal k value for KNN involved training classifiers (SVM and BN), with the best accuracy achieved at k = 6 for both. This strategy aims to enhance dataset robustness and mitigate the impact of missing data for subsequent analyses.

Fig. 5 visually depicts the increasing accuracy trend as the training ratio is augmented, and the detailed accuracy results are summarized. Based on results of Figure 5, the proposed LR model demonstrates an increase in accuracy from 50% to 90% training, with the highest accuracy of 96.15% achieved at 80% training and 20% testing ratio.

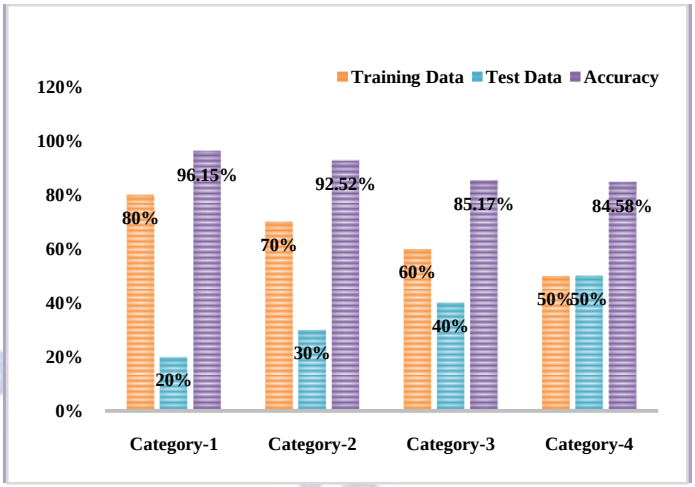

Figure 5: Performance analysis of logistic regression.

Table 2 summarizes the LR classifier's classification report metrics on the UCI dataset, revealing consistent and effective performance with precision (PR), recall (RC), F1-score (F1), and accuracy (AC) all at 0.956, utilizing a 80% training and 20% testing split.Table 4 highlights the superior classification performance of proposed LR methodology compared to previous approaches. Our LR model employs KNN for handling missing values, contrasting with earlier methods advocating for record removal, yielding suboptimal results.

Table 2. Classification report of LR classifier.

Confusion Matrix			Classification Report			Empty Cell
	+ve	-ve	PR	RC	F1	AC
+ve	15	2	0.971	0.98	0.94	95.6
-ve	2	12	NA	NA	NA	NA

Table4: Model performancemetrics.

Metrics	XGBoost	AdaBoost	Gradient Boost	Extra Trees	Proposed LR
AC	92	92	91	95.23	96.15
SP	92.1	92.5	91	94.22	94.61
PR	92	90.1	91	93.5	98
RC	93	92.8	91	93.56	96.2
F1	92	91.9	92	93.9	96.3

6. CONCLUSIONS

The rising incidence of heart failure necessitates the development of a robust diagnostic system. Our LR approach addresses the challenge of missing data in the preprocessing stage, leveraging the KNN model as the optimal algorithm for handling non-existent values. Employing XGboost, Adaboost, gradient boosting, extra trees and the stacking algorithm in the classification step, we achieved an impressive accuracy score of 96% with the proposed LR model.This work has the potential to significantly advance automatic diagnosis, aiding physicians in timely and accurate patient assessments. Future studies will build upon these findings to create a predictive system that enhances medical treatment and reduces costs.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] World Health Organization and J. Dostupno, cardiovascular diseases: key facts, vol. 13, no. 2016, p. 6, 2016.

[2] Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks Proced. Comput. Sci., 120 (2017), pp. 588-593

[3] N. Kausar, S. Palaniappan, B.B. Samir, A. Abdullah, N. Dey Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients in Applications of Intelligent Optimization in Biology and Medicine, Cham, Switzerland: Springer (2016), pp. 217-231.

[4] M. Shouman, T. Turner, R. Stocker Integrating clustering with different data mining techniques in the diagnosis of heart disease J. Comput. Sci. Eng., 20 (1) (2013), pp. 1-10.

[5] M.S. Amin, Y.K. Chiam, K.D. Varathan Identification of significant features and data mining techniques in predicting heart disease Telemat. Inf., 36 (2019), pp. 82-93.

[6] D. Singh, J.S. Samagh A comprehensive review of heart disease prediction using machine learning J. Crit. Rev., 7 (12) (2020), p. 2020.

[7] F.A. Latifah, I. Slamet, Comparison of heart disease classification with logistic regression algorithm and random forest algorithm.

[8] Z. Khan, D.K. Mishra, V. Sharma, A. Sharma Empirical study of various classification techniques for heart disease prediction Proceedings of the IEEE 5th International Conference on Computing Communication and Automation (ICCCA) (2020), pp. 57-62.

[9] Nishadi, A.S.T. (n.d.). International journal of advanced research and publications predicting heart diseases in logistic regression of machine learning algorithms by python jupyterlab.

[10] M. Saw, T. Saxena, S. Kaithwas, R. Yadav and N. Lal Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning saw (Ed.), 2020 International Conference on Computer Communication and Informatics (ICCCI), IEEE, Coimbatore, India (2020), pp. 1-6, January 22-24.

[11] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K.H., Lee, S., &Froelicher, V. (n.d.). International application of a new probability algorithm for the diagnosis of coronary artery disease.

[12] S. Bashir, Z.S. Khan, F. Hassan Khan, A. Anjum, K. Bashir Improving heart disease prediction using feature selection approaches Proceedings of the 16th International Bhurban3 Conference on Applied Sciences and Technology (IBCAST) (2019), pp. 619-623, 10.1109/IBCAST.2019.8667106

[13] R. Kannan, V. Vasanthi, Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease, Springer Briefs in Applied Sciences and Technology, Springer Verlag (2019), pp. 63-72, 10.1007/978-981-13-0059-2_8

[14] M. Ganesan, N. Sivakumar, IoT based heart disease prediction and diagnosis model for healthcare using machine learning models Proceedings of the IEEE (ICSCAN) (2019), pp. 1-5, 10.1109/ICSCAN.2019.8878850

[15] K.G. Dinesh, K. Arumugaraj, K.D. Santhosh, V. Mareeswari Prediction of cardiovascular disease using machine learning algorithms Proceedings of the International Conference on Current Trends towards Converging Technologies (ICCTCT) (2018), pp. 1-7, 10.1109/ICCTCT.2018.8550857