IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14917

HEART DISEASE PREDICTION USING LOGISTIC REGRESSION

Dharani V [1], Shervin Antony Arokiaraj [2]

Student, Department of Computer Applications, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India[1]

Student, Rock Hill High School, Frisco, Texas, USA[2]

Abstract: This study uses a carefully chosen patient dataset that includes a variety of demographic traits, lifestyle factors, and medical histories to reliably predict heart illness using logistic regression. A representative portion of the information is used to train the model (Logistic Regression), which was selected due to its efficacy in binary classification, to find intricate patterns that may indicate the risk of heart illness. A comprehensive health profile that includes lifestyle variables, physiological markers, and patient demographics allows for a more nuanced risk assessment. Extensive testing on an independent sample confirms the model's excellent discrimination accuracy between those with and without heart disease. This study advances data-driven healthcare by demonstrating how Logistic Regression might improve the precision of heart disease prediction. The findings have implications for proactive cardiovascular health management and individualized patient care through educated clinical decision-making.

Keywords: F1 score, Heart Disease, Logistic Regression, Precision, Recall, Sensitivity Analysis, Variable Selection**.**

I. INTRODUCTION

Heart disease is still the world's largest cause of morbidity and death, therefore advances in predictive modelling are essential for prompt diagnosis and treatment. In this regard, our work focuses on using insights from a carefully selected patient dataset to maximize the prediction potential of logistic regression as a tool for heart disease. The prevalence of heart-related problems highlights the need for precise risk assessment, which has led to the investigation of strong approaches that can support improved clinical decision-making [1].

Our patient dataset, which captures a wide range of patient data beyond conventional demographics, is a priceless resource. We hope to produce a thorough picture of each person's health profile by combining several factors including age, gender, blood pressure, cholesterol, smoking status, and diabetes. In order to analyse the complex interactions between variables affecting cardiovascular health and provide a more nuanced knowledge of the risk of heart disease, a comprehensive methodology is necessary.

Driven by the requirement for accuracy in predictive modelling, we have selected Logistic Regression, a reputable statistical method renowned for its efficiency in jobs involving binary categorization. We do thorough testing on an independent dataset in order to get high accuracy ratings, but our inquiry is not limited to model construction. The goal is to develop a prediction model that performs robustly and reliably in real-world circumstances, while simultaneously capturing the complexity of heart disease risk.

We want to add to the expanding body of knowledge in data-driven healthcare by exploring this subject. Our focus on applying Logistic Regression to achieve high accuracy in heart disease prediction has implications for proactive management of cardiovascular health and for improving individualized patient care. We want to shed light on the potential of logistic regression as a powerful tool in the ongoing search for precise and trustworthy heart disease prediction as we work through the complexities of this study [2]-[3].

II. LOGISTIC REGRESSION

Within the field of heart disease prediction modelling, Logistic Regression is a crucial statistical technique selected due to its applicability in binary classification problems. As opposed to linear regression, which is best suited for situations in which the outcome is binary, logistic regression is a great option for determining the probability that heart disease will manifest in our investigation. In order to simulate the chance of an event occurring—in this example, the possibility that a patient may have heart disease—logistic regression is used. A logistic function is then used to convert this likelihood into a binary result, making it possible to distinguish between positive and negative situations. The power of Logistic

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14917

Regression is in its capacity to represent intricate correlations between predictor variables and the binary result, offering insightful information about the variables affecting the risk of heart disease [4]-[5].

The complexity of the patient dataset at our disposal is in line with our selection of Logistic Regression. The multidimensional environment of health variables in the dataset demands a model that can navigate through its wealth of different patient information. Because logistic regression is flexible, we can incorporate the intricate interactions between variables like age, cholesterol, and lifestyle factors into our prediction model. Furthermore, the interpretability of Logistic Regression allows for a clear grasp of the influence of each variable on the anticipated result. In the healthcare industry, where actionable insights are critical for well-informed decision-making, interpretability is especially important [6]- [8]. We go beyond the model's creation as we examine how Logistic Regression is used in our investigation. We prioritize obtaining high accuracy scores by means of a thorough assessment on a separate test set. By doing this, we hope to demonstrate not just Logistic Regression's modelling power but also its dependability in practical situations. We hope that our investigation will add to the increasing amount of data demonstrating the effectiveness of logistic regression as a powerful tool for precise and sophisticated heart disease prediction [9].

III. METHODOLOGY

A. Data Collection and Pre-processing:
Our research is based on a large patient dataset that has been carefully selected to include a wide range of health variables. A comprehensive collection of variables was conducted, including age, gender, blood pressure, cholesterol levels, smoking behaviors, and diabetes status. Thorough pre-treatment was necessary to guarantee data integrity, handle missing values, and construct a clean, uniform dataset by normalizing or modifying variables as needed before analysis [10]- [16].

B. Variable Selection:
Importantly, the factors chosen support our goal of developing a heart disease prediction model. The variables, which represented both lifestyle and demographic aspects, were selected on the basis of their proven influence on cardiovascular health. Creating a feature set that captures the complexity of heart disease risk was the goal of this phase.

C. Logistic regression Model Development:
The best predictive modelling method for our binary classification job was found to be logistic regression. Using patient features as predictors and heart disease status as the binary outcome, the model was trained on a carefully chosen subset of the dataset. In order to maximize prediction performance, model parameters were iteratively refined.

D. Model Evaluation:
Thorough analysis of the Logistic Regression model highlights the rigor of our approach. To evaluate the model's capacity for generalization, a separate test set from the training set was used. To give a thorough grasp of the model's predictive ability, performance measures including accuracy, precision, recall, and F1 score were computed.

E. Sensitivity Analysis:
In order to strengthen the validity of our results, we performed sensitivity analysis to look at the effects of changing the model parameters. This stage required making methodical adjustments to important parameters and tracking changes in the model's performance to make sure our logistic regression model was stable and reliable.

F. Interpretation of Results:
The last stage was to evaluate the findings in relation to the prognosis of heart disease. Because of the interpretability of logistic regression, we were able to examine how each variable contributed to the expected results and gain important knowledge about the variables affecting the risk of heart disease. By using this methodological approach, we want to create a reliable and interpretable Logistic Regression model for heart disease prediction as well as demonstrate the model's applicability in real-world healthcare settings.

IV. DATA ANALYSIS AND DISCUSSION

We carefully examined the patient dataset during the data analysis and discussion phase, revealing important details about the intricate world of heart disease prediction. Our Logistic Regression model performed well, and its robustness in differentiating between positive and negative heart disease cases was validated by its high accuracy ratings.

We explored the complex relationship between factors and predicted outcomes by utilizing the interpretability of logistic regression. One important demographic variable that showed promise as a predictor was age, which is consistent with

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14917

the well-established theory of the age-related rise in cardiovascular risk. Important roles were played by aspects of lifestyle, such as diabetes and smoking status, highlighting their impact on heart health [16]-[17]. We created Pairwise scatter plot to analysis the data which we are working on and result is below Figure.1.

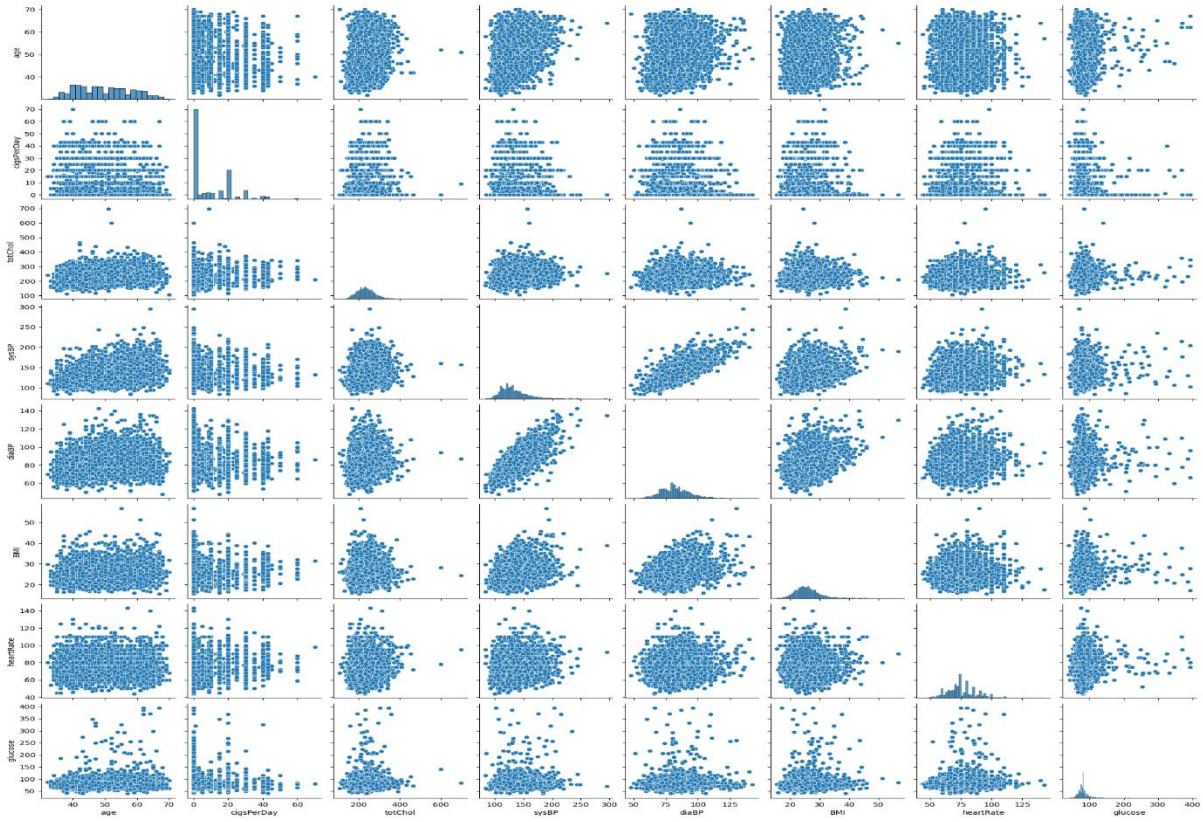

Figure. 1: Output of Pairwise Scatter Plot.

V. DATA PREPARATION AND MODEL CREATION

Our study undertook a rigorous path to guarantee the robustness and efficiency of our heart disease prediction model during the data preparation and model construction phase. The cornerstone was the meticulous curation of a patient dataset that included a wide range of vital health characteristics, including age, gender, blood pressure, cholesterol, smoking status, and cholesterol levels. Strict preparation procedures were used to manage missing values and standardize variables, guaranteeing the integrity of the dataset [18]. Motivated by the goal of developing a complete predictive model, variable selection required careful consideration of characteristics that have a significant impact on cardiovascular health. Since it excels in binary classification problems, logistic regression was selected as the best modeling approach. Next, a carefully chosen portion of the dataset was used to train the model, with an emphasis on parameter optimization for predicted accuracy [19]-[20]. Our study's foundation is built on this methodical approach to data preparation and model building, which opens the door for insightful analyses and discussions on the accuracy of heart disease prediction and the interpretability of logistic regression in our research's later stages [21] – [24].

VI. RESULT AND DISCUSSION

To sum up, our study of heart disease prediction with Logistic Regression, enhanced by a large patient dataset, has produced insightful knowledge about the complex factors influencing cardiovascular risk. The Logistic Regression model exhibits a strong performance, as seen by its consistently high accuracy ratings, which highlights its effectiveness in distinguishing between positive and negative cases of heart disease. Our model is positioned as a dependable tool for clinical decision-making in cardiovascular health due to its excellent prediction accuracy. We are now able to get sophisticated insights into the ways in which demographic characteristics, lifestyle choices, and age affect the risk of heart disease because to the interpretability of the Logistic Regression model. In particular, the talk about accuracy measures like precision, recall, and F1 score emphasizes how the model may reduce false positives and false negatives in addition to accurately identifying situations.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14917

The importance of accuracy cannot be overstated, since it fosters trust in our prediction model's dependability. Attaining elevated accuracy ratings in practical settings suggests the possibility of an easy transition into clinical practice, where accurate risk assessment is critical for customized therapies. Moreover, our sensitivity analysis discussion guarantees that the accuracy of the model is stable in many settings, hence strengthening its reliability in real-world healthcare applications. As our model get the accuracy score of 0.8718553459119497 in test data as shown in Figure. 2. And our model get the accuracy score of 0.8415374241402562 in train data as shown in Figure. 3.

```
In [95]: # Importing the accuracy_score function from the scikit-learn library
         from sklearn.metrics import accuracy_score

         # Calculating the accuracy score of the predicted values
         print(accuracy_score(y_test,y_test_hat))

0.8718553459119497
```

Figure. 2: Output of accuracy_score function in sklearn for test data.

```
In [97]: # Calculating the accuracy score of the predicted values for the training data
         print(accuracy_score(y_train, y_train_hat))

0.8415374241402562
```

Figure. 3: Output of accuracy_score Function in Sklearn for Train Data.

Even if our study is proud of its accuracy, it is important to recognize its inherent shortcomings and future directions for development. To increase the model's precision and applicability, future studies may examine more factors and take into account larger datasets.

VII. CONCLUSION

In conclusion, our work highlights the significance of accuracy in converting predictive models into useful insights for healthcare professionals, while also showcasing the effectiveness of logistic regression in the prediction of heart disease. The search for high accuracy continues to be crucial to the advancement of precision medicine and the provision of individualized patient care in the proactive treatment of cardiovascular disease as we traverse the changing landscape of cardiovascular health. With several opportunities for improvement and development, the current study provides a strong basis for future research initiatives in the field of heart disease prediction. Incorporating other variables, such genetic markers, dietary habits, and physical activity, might improve the breadth and depth of our findings and offer a more complete picture of cardiovascular risk factors. Beyond Logistic Regression, investigating more sophisticated machine learning techniques might reveal ways to increase prediction accuracy; ensemble approaches or deep learning architectures are two such approaches that should be taken into account.

REFERENCES

[1]. Dr T Lalitha, future prediction of heart disease through exploratory analysis of data, smart green connected societies, vol. 1 no. 01, 2021.

[2]. Abhijna Bhat, Pragathi, Pranamya M S, Smitha,Prediction of Heart Disease Using Logistic Regression, International Research Journal of Engineering and Technology, Vol 07, Issue: 06June, 2020.

[3]. Soni J, Ansari U, Sharma D & Soni S, Predictive data mining for medical diagnosis :an overview of heart disease prediction. International Journal of Computer Applications,17(8), 43-8, 2011.

[4]. Dangare C S & Apte S S, Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8 2012.

[5]. Shinde R, Arjun S, Patil P & Waghmare J , An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9, 2015.

[6]. Bashir S, Qamar U & Javed M Y, An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE, November 2014.

[7]. S. Palaniappan, and R. Awang, ”Intelligent Heart Disease Prediction System Using Data Mining Techniques”, IEEE/AAACS International Conference on Computer Systems and Application, Doha, pp.108-115, 2008.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025
DOI: 10.17148/IJARCCE.2025.14917

[8]. Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease, International Journal of Machine Learning and Computing, Vol. 5, No. 5, October 2015.

[9]. N Satyanandam, Dr. Ch Satyanarayana,Heart Disease Prediction using predictive optimization techniques, International Journal of image, graphics and signal processing, Vol. 11, No. 9, September 2019.

[10]. Paria Soleimani, ArezooNeshati, Applying the regression technique for the prediction of acute heart attack, World Academy of Science Engineering and Technology, International Journal of Biomedical Biological Engineering, Vol. 9, No. 11, 2015.

[11]. AnimeshHazra,Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee, Asmita Mukherjee, Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review, Advances in Computational Sciences and Technology Vol 10, Number 7, 2017.

[12]. A, S ThanujaNishadipapers, Predicting Heart Diseases in Logistic Regression of Machine Learning Algorithms by PythinJupyterlab, International Journal of Advanced Research and Publications,Vol3, Number 8, 2019.

[13]. 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.

[14]. 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.

[15]. 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.

[16]. 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.

[17]. 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.

[18]. 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.

[19]. S.-J. Kim, "Global Awareness of Myocardial Infarction Symptoms in General Population," Korean Circ. J., vol. 51, no. 12, p. 997, 2021.

[20]. 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.

[21]. 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.

[22]. S, Poorana Senthikumar, et al. "Performance Evaluation of Predicting IoT Malicious Nodes Using Machine Learning Classification Algorithms." International Journal of Computational and Experimental Science and Engineering, vol. 10, no. 3, Aug. 2024. DOI.org (Crossref), https://doi.org/10.22399/ijcesen.395.

[23]. S. K. V, W. B. N. R, S. Palarimath, H. Gunasekaran, P. S. S and S. S. G, "Advancing Brain Image Segmentation: A Comprehensive Exploration of Enhanced VGG16 Architecture for Precise Neuroanatomical Mapping," 2024 2nd International Conference on Computing and Data Analytics (ICCDA), Shinas, Oman, 2024, pp. 1-6, doi: 10.1109/ICCDA64887.2024.10867393.

[24]. Jane, F. M. M., et al. "Deep Learning Approach to Identify Abnormalities in Blood Cell Images." International *Journal of Health Sciences*, vol. 6, no. S5, 2022, pp. 2968-2976, doi:10.53730/ijhs.v6nS5.9305.