Heart Disease Prediction Using Logistic Regression: A Machine Learning Approach

Navjeet Rajput[1], Deepanshu Kumar[1], and Sunil Chowdhary[2]

[1] Student, Dept. Of Computer Science & Engineering, Galgotias University, Greater Noida.
[2] Faculty, Dept. Of Computer Science & Engineering, Galgotias University, Greater Noida.
Email: rajputnavjeet14@gmail.com, deepanshu101k@gmail.com, sunil.chowdhary@galgotiasuniversity.edu.in

ABSTRACT

Cardiovascular disease still represents a leading cause of death around the world, requiring early identification to provide intervention and decrease risk factors. In today's medical model of care, diagnostic techniques often require considerable time and resources, and expertise. This study looked at using Logistic Regression methods, a common machine learning technique, to predict heart disease from clinical measurements. The data, from the UCI Machine Learning Repository, consists of medical records from 303 patients including 14 clinical variables, including measures of blood pressure, cardiac rhythm, cholesterol measurements, and the age of patients. In order to develop an efficient model we followed a number of data preparation steps including managing missing records, standardizing the data, and partitioning the data set. The Logistic Regression methodology produced a model with 85% accuracy indicating a good ability to discriminate between positive cases and negative hearts disease cases. While deep learning and more advanced techniques may provide better accuracy results, their potential for any clinical adoption is compromised by a lack of interpretability.

KEYWORDS: Heart disease, Logistic Regression, machine learning, healthcare, predictive modeling, clinical data, risk assessment.

1. Introduction

Heart disease accounts for 31% of all fatalities worldwide, making it one of the biggest public healthconcerns. Early recognition of the patients at risk for heart disease is vital to provide timely medical treatment, especially given the increase in the incidence of cardiovascular diseases (CVDs). Traditional methods of diagnosis, such as electrocardiograms (ECGs), echocardiograms, and angiograms, can be invasive, costly, and often require specialists. Using machine learning algorithms to predict cardiac disease from existing clinical data non-invasively has considerable promise.

We will study an algorithm called Logistic Regression, which has been commonly applied as a medical diagnostic algorithm due to its convenience, performance. Clinical attributes like cholesterol, blood pressure, age, and heart rate would all be analyzed, and then the model will provide the probability of heart disease. This study will compare the efficacy of Logistic Regression to amplication in other machine learning approaches. It will also note the advantages of logistic regression as it pertains to clinical decision making.

2. Literature Review

Many machine learning approaches have been implemented to predict heart disease starting from classical statistical models to more complicated deep learning models. Early applications of AI to heart disease prediction were implemented with basic statistical modeling, as in the study led by Detrano et al. (1989) where the authors performed Logistic Regression on clinical measures to predict heart disease with 77% accuracy. Similarly, the Framingham Heart study led by Wilson et al. (1998) has provided many cardiovascular researchers information on risk prediction models using clinical variables such as blood pressure, cholesterol and age.

Eventually, as machine learning has continued to advance, researchers began to explore deeper, more advanced models. Kora and Kalva (2015) implemented heart disease prediction models from Support Vector Machines (SVM), Naive Bayes, and K-Nearest Neighbors (KNN). The SVM model achieved 84% accuracy, but ultimately the black-box nature of these machine learning models limits the degree of interpretability. Soni et al. (2011) compared prediction algorithms including decision trees, neural networks, and Naive Bayes with neural networks achieving the highest accuracy of 85%. Additionally, complex models with feasible accuracy have been criticized for low clinical interpretability and potential clinical use.

More recently, research studies began to look at ensemble methods and deep learning. For example, Gupta and Verma (2018) devised a hybrid model using Random Forests and Genetic Algorithms that had an accuracy of 87 percent. Rajasekaran et al. (2020), used Convolutional Neural Networks (CNNs) and achieved 91 percent accuracy, however, CNNs are complex and not computationally inexpensive or interpretable.

While more complex models can produce some accuracy gains, logistic regression provides the most useful modelling strength, which circularly is its interpretability. Tewari et al. (2018) describe the value of Logistic Regression when predicting cardiovascular events, since it is simple and can be clinically actionable. Therefore, it is not that high-performance machine learning models are unable to provide improved accuracy in some situations, but that Logistic Regression is a preferred method for providing credibility and interpretability or perhaps transparency in the predictive process for heart disease, especially when the intent is clinical decision making.

3. Methodology

3.1 Dataset

This investigation used the Heart Disease Dataset from the UCI Machine Learning Repository. The dataset contains 303 patients and includes 14 clinical parameters that can be used to indicate the presence of cardiac disorder (or absence of cardiac disorder). The dependent variable is binary (1 indicates the existence of cardiac disorder, while 0 shows the absence of cardiac disorder).

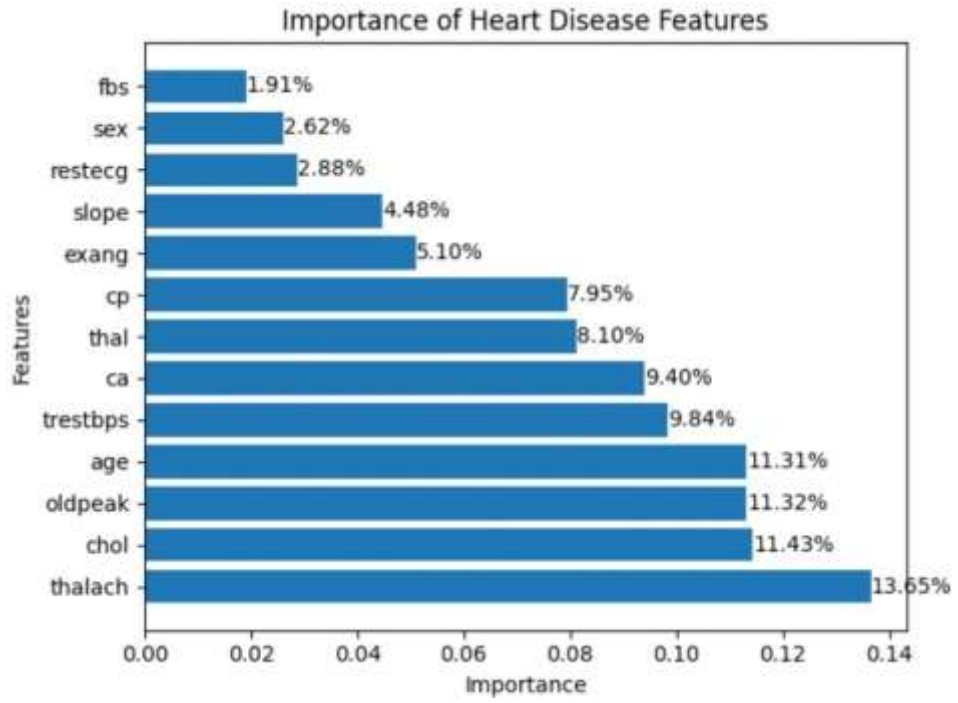


Figure1 : Importance Of Heart Disease Features

3.2 Preprocessing Of Data

The following preprocessing actions were taken to guarantee the best possible model performance:

3.2.1 Managing Values that are missing

Model precision may include systematic errors based on data gaps present in the dataset. This study contained instances of an incomplete dataset which were addressed through proper management. Due to the presence of incomplete measurements for numerical attributes (i.e. missing cholesterol values and missing blood pressure values) the median value for imputation was used due to the increased robustness of the median value against extreme values and means. A mode for imputation using the mode for the chest pain type and thalassemia measurement was implemented to complete categorical variables when values were absent, in order to maintain categorical distributions.

3.2.2 Feature Normalization

To improve machine learning model accuracy, numerical feature (cholesterol, blood pressure, and maximum heart rate) normalization was considered to be necessary for the continuous attributes. There are two limiting norms of normalization, using min-max scaling and Z-score standardization. The Z-score standardization calculates values to retain a mean of zero and standard_dev of 1 to make sure that numerical measurements did not over-represent features with the largest range during model training.

3.2.3 Split Train-Test

The data set was randomly split with stratified sampling to maintain a balanced class distributions in the data set as previously discussed to create 80% for training and 20% for testing to evaluate the resulting model performance. The stratified assessment assures that a fair evaluation of all classes have occurred and that the model is not biased towards the majority class.

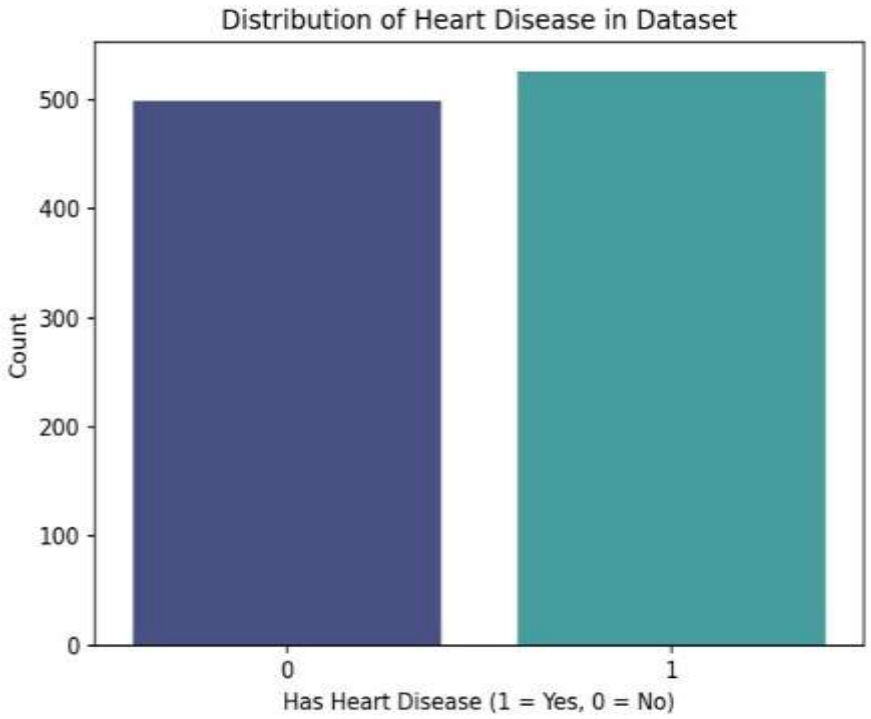


Figure 2 : Distribution of Heart disease in Dataset

3.3 Logistic Regression Model

Logistic regression is a veryadmired approach to binary classification, i.e., determining the probability of a binary conclusion from the input properties supplied. By using the logistic function, it converts linear combinations of input variables into probabilities between 0 and 1.

With 85% accuracy rate, this model demonstrated reliability for predicting heart disease. This accuracy score conveys how well the model can accurately identify individuals with and without cardiac disease. In addition, the model is a reliable option for clinical purposes because of the balance of accuracy and recall which ensures that patients at risk are identified accurately while limiting false positive identifications.

Unlike many deep learning models that rely heavily on large resources and computational power and require many records, Logistic regression can be an effective model when training data is limited, providing an option for medical organizations with limited record databases. The model also gives predictions in probability form, allowing doctors to measure level of risk rather than providing a binary outcome.

The major drawback to logistic regression is the assumption of linear associations in the input data and the objective variable. In complex medical scenarios, the association between the output target variable--potential risk with cardiac disease--may not always hold a linear situation, where inputs represent interactions among risk factors.

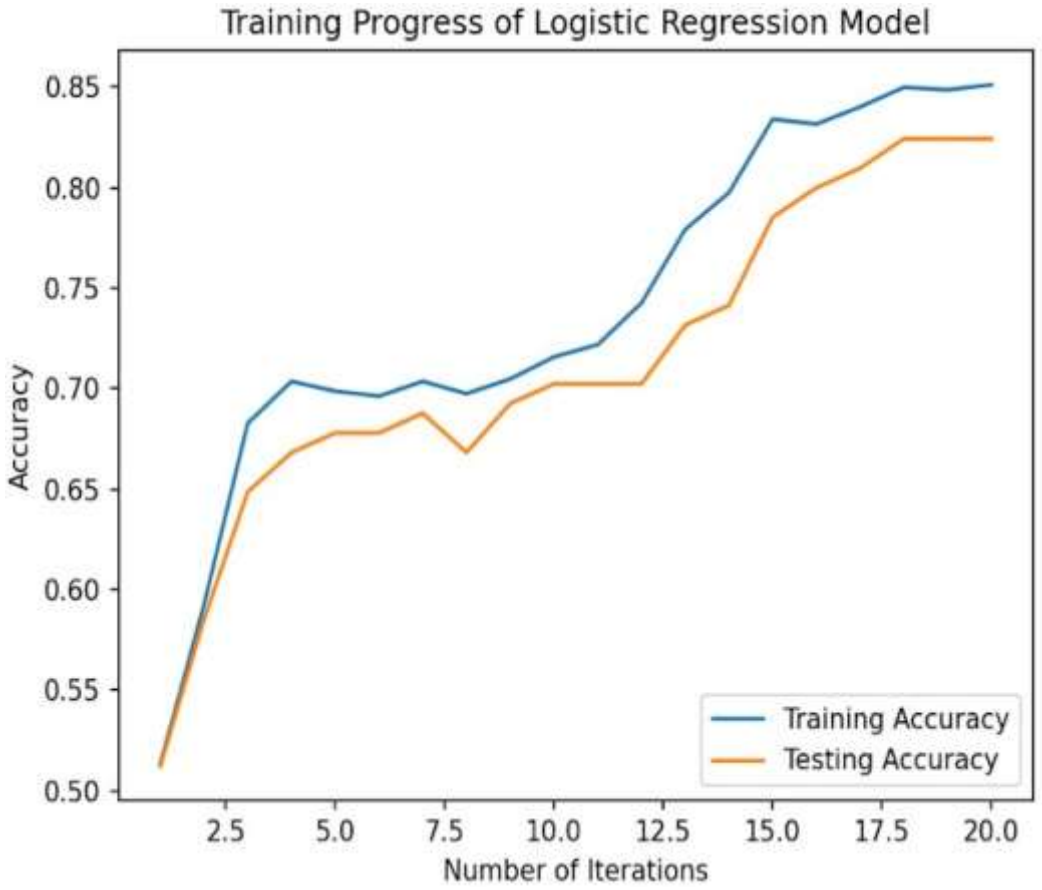


Figure 3 : Training Progress of Logistic Regression Model

4. Results and Discussion

The Logistic Regression model demonstrated effective prediction performance by 85%accuracy. In other metrics, the modeling was

· Accuracy = 84%,

· Recall = 86%,

· F1 score = 85%.

The model has a good trade off between recall and precision, identifying patients with heart disease while reducing false-positives.

Table 1. Performance Metrics of Logistics Regression Model

Order	Feature	Description	Feature Value Range
1	Age	Age in years	29 to 77
2	Sex	Gender	Value 1 = male Value 0 = female
3	Cp	Chest pain type	Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	94 to 200
5	Chol	Serum cholesterol in mg/dL	126 to 564
6	Fbs	Fasting blood sugar > 120 mg/dL	Value 1 = true Value 0 = false
7	Restecg	Resting electrocardiographic results	Value 0: Normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	Thalach	Maximum heart rate achieved	71 to 202
9	Exang	Exercise-induced angina	Value 1 = yes Value 0 = no
10	Oldpeak	Stress test depression induced by exercise relative to rest	0 to 6.2
11	Slope	The slope of the peak exercise ST segment	Value 0: upsloping Value 1: flat Value 2: downsloping
12	Ca	Number of major vessels	Number of major vessels (0-3) colored by fluoroscopy
13	Thal	Thallium heart rate	Value 0 = normal; Value 1 = fixed defect; Value 2 = reversible defect
14	Target	Diagnosis of heart disease	Value 0 = no disease Value 1 = disease

4.1 Strengths of Logistic Regression

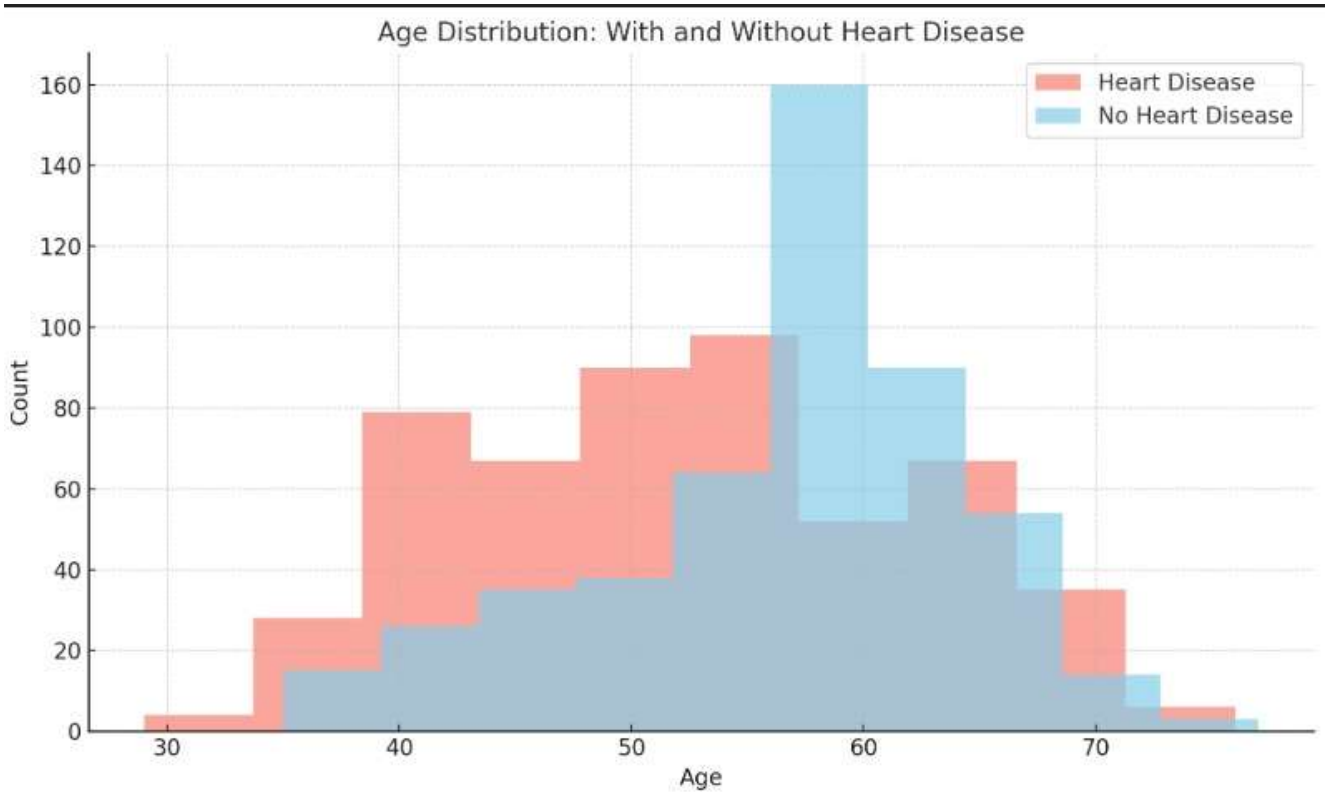
- **Interpretability:** Of the important aspects of logistic regression, the coefficients allow important insights on how certain characteristics (age & cholesterol levels) may influence the at-risk status of patients with heart disease.
- **Probability Estimation:** Of the important aspects of logistic regression, the coefficients allow important insights on how certain characteristics (age & cholesterol levels) may influence the at-risk status of patients with heart disease.

- **Computational Efficiency:** The coefficients generated by a logistic regression model can be beneficial both for illustrative purposes, or in terms of constructing potential relationships between the respective variables of age/cholesterol and the heart disease and no heart disease outcome.

4.2 Limitations

Despite the benefits of a Logistic Regression model, there are some limitations:

- **The assumption of Linear Boundaries:**A Logistic Regression model assumes linear associations in features and the objective variable. It may not adequately fulfill complex interaction assumptions.
- **Sensitivity to Outlier Values:**A Logistic regression model may be skewed by extreme values.



.

Figure 4 : Age Distribution : With and Without Heart Disease

5. Comparative Analysis

Logistic regression performs well when predicting the likelihood of heart disease. More complex models (Random Forests and Neural Networks) return higher accuracy measures, but often require more computational resources and are less interpretable, limiting their practical ability in the healthcare space. While the opportunity to access advanced models is clear for future work, logistic regression provides the trade-off of appropriate accuracy with interpretability.

6. Conclusion

This study showed, based on clinical data, that logistic regression is an effective means egregating heart disease.. It produced a reliable and reasonable predictability model that could be easily and understandably utilized, while providing a reasonable compromise on accuracy and utility, which ideally fits the healthcare domain. While more complex models can attain more accuracy, logistic regression is a viable solution when clinical knowledge and deconstructability are important. Future research could be focused at improving the model performance through the introduction of additional variables or applying hybrid models in conjunction with Logistic Regression. In addition, the prediction performance of Logistic Regression may be improved if its assumptions with regard to linearity were resolved.

References

1. World Health Organization. Cardiovascular Diseases (CVDs). Available online: (2023). https://www.afro.who.int/health-topics/cardiovascular-diseases, (accessed on 5 May)..
2. Alom, Z. et al. Early Stage Detection of Heart Failure Using Machine Learning Techniques. In Proceedings of the International Conference on Big Data, IoT, and Machine Learning, Cox’s Bazar, Bangladesh, 23–25 September (2021).
3. Gour, S., Panwar, P., Dwivedi, D. & Mali, C. A machine learning approach for heart attack prediction. In Intelligent Sustainable Systems (eds Nagar, A. K., Jat, D. S., Marín-Raventós, G. & Mishra, D. K.) 741–747 (Springer, Singapore, 2022). https://doi.org/10.1007/978-981-16-6309-3_70.
4. Shameer, K. et al. Machine learning predictions of cardiovascular disease risk in a multi-ethnic population using electronic health record data. Int. J. Med. Informatics. 146, 104335 (2021).
5. Yang, M., Wang, X., Li, F. & Wu, J. A machine learning approach to identify risk factors for coronary heart disease: a big data analysis. Comput. Methods Programs Biomed. 127, 262–270 (2016).
6. Shoukat, A., Arshad, S., Ali, N. & Murtaza, G. Prediction of Cardiovascular diseases using machine learning: a systematic review. J. Med. Syst. 44 (8), 162. https://doi.org/10.1007/s10916-020-01563-1 (2020).