Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250402048
Volume 4, Issue2 (May-August 2025), PP: 356-359.
www.indjcst.com
ISSN No: 2583-5300

Heart Disease Prediction Using Logistic Regression

Anusha T L[1], Apeksha BK[2], Deekshitha D[3]
[1,2,3] Department of Computer Science Engineering-Data Science, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.

To Cite this Article: *Anusha T L[1], Apeksha BK[2], Deekshitha D[3], "Heart Disease Prediction Using Logistic Regression", Indian Journal of Computer Science and Technology, Volume 04, Issue 02 (May-August 2025), PP: 356-359.*

Abstract: *This study presents a heart disease prediction system leveraging logistic regression. The process begins with preprocessing patient records by handling missing data, scaling features, and splitting datasets. Relevant input attributes such as cholesterol levels are utilized for model training and optimisation. The logistic regression model predicts the probability of heart disease based on input features. The system's output includes disease classification and is assessed using key performance metrics. This approach aims to enhance early detection and clinical decision- making efficiency.*

Key Words: *Heart Disease Prediction, Logistic Regression, Data Preprocessing, Feature Scaling, Machine Learning, Medical Diagnosis, Classification, Performance Metrics, Patient Records, Probability Estimation.*

I.INTRODUCTION

Cardiovascular diseases (CVDs), such as coronary artery disease and heart attacks, represent the primary cause of mortality worldwide, accounting for nearly 17.9 million deaths annually, as reported by the World Health Organization[7]. Due to the high fatality rate and the growing burden on healthcare systems, early and accurate prediction of heart disease is critical. Predictive models leveraging clinical data are increasingly utilized to support medical decision- making, with logistic regression emerging as one of the most effective and interpretable methods for binary classification problems.

Logistic regression is a predictive modeling approach used to determine the likelihood of a binary event—such as whether heart disease is present or not—by analyzing the relationship between one or multiple predictor variables and the outcome. It is particularly well-suited for medical applications because it accommodates both continuous and categorical variables and does not require strict assumptions about data distribution[4,8.] The sigmoid function transforms the model's output into a value between 0 and 1, making it suitable for interpreting predictions as probabilities.

Several studies have successfully applied logistic regression to heart disease datasets, yielding robust performance. For instance, Montu Saw et al. developed a logistic regression model using patient health data and demonstrated that it could significantly improve diagnostic accuracy comparedto traditional methods[3]. Similarly, Yaseliani and Khedmati used the UCI heart disease dataset[2] to construct a model that achieved a sensitivity of 84.21% and specificity of 90.38%, validating the practical utility of logistic regression in real-world medical settings [4.]

Moreover, logistic regression offers flexibility in feature selection and performance evaluation. Studies have used statistical metrics like AIC (Akaike Information Criterion), p-values, and ROC curves to optimize the model and select the most significant predictors such as cholesterol level, blood pressure, age, and chest pain [6, 12]. Some variations of the logistic regression model, such as those using robust methods like Least Quartile Difference (LQD) and Median Absolute Deviation (MAD), have accuracy [13, 14]

In conclusion, logistic regression remains a foundational technique in predictive healthcare analytics due to its interpretability, computational efficiency, and strong diagnostic performance. Its application to heart disease prediction not only enhances clinical outcomes but also supports broader goals in preventative care and resource optimization [1, 10,11]. Logistic regression models provide straightforward diagnostic thresholds, allowing practitioners to make faster decisions based on risk probabilities.

II.LITERATURE SURVEY

G. Ambrish, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," Global Transitions Proceedings, vol. 3, no. 1, pp. 127–130, Jun. 2022[1]

The paper addresses early prediction of CVD. Logistic Regression is used with a public dataset to classify disease presence. Preprocessing steps like handling missing data were crucial. Feature selection focused on attributes with strong positive correlation. Data was cleaned and split into train-test sets (90:10 to 40:60). Logistic Regression was applied on each split to analyze performance. Correlation-based feature selection improved model accuracy.

Performance metrics were utilized to evaluate the accuracy and effectiveness of the predictive model. The model achieved 87.10% accuracy at a 90:10 split. This method proves effective for early CVD detection using ML.

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," in Proceedings of the 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, Jan. 2020[3]

It leverages machine learning to process clinical and demographic data efficiently. The objective is to assist in early detection to reduce mortality rates. The model's simplicity and interpretability make it suitable for healthcare prediction tasks. The dataset was preprocessed by cleaning missing values and normalizing features. Feature selection was applied to retain the most impactful predictors. Quantitative metrics were employed to measure the predictive model's accuracy and overall performance. The Logistic Regression model demonstrated strong predictive ability for heart disease. It can serve as a foundational tool for clinical decision support systems.

M. Yaseliani and M. Khedmati, "Prediction of Heart Diseases Using Logistic Regression and Likelihood Ratios," International Journal of Industrial Engineering & Production Research, vol. 34, no. 1, pp. 1–15, 2023[4].

This research develops a predictive model for heart disease by applying logistic regression combined with likelihood ratios. Using a dataset comprising 299 subjects and 13 variables, the study systematically evaluates the influence of each predictor through statistical indicators such as AIC scores and p-values. Data preprocessing steps included addressing missing values and normalizing feature distributions to enhance model reliability. Key predictors were identified and selected based on rigorous statistical criteria to optimize predictive accuracy.The logistic regression model was trained and validated using .
The proposed model achieved a sensitivity of 84.21% and specificity of 90.38%, indicating strong predictive performance.

N. F. Zulkiflee and M. S. Rusiman, "Heart disease prediction using logistic regression," J. Coastal Life Med., vol. 11, no. 1, pp. 573–579, 2023[5].

A model to predict heart disease risk. It uses patient data to estimate the probability of disease, aiding early diagnosis. The study highlights the importance of accurate prediction to reduce mortality. The dataset underwent preprocessing including missing value handling and normalization. Categorical data was encoded for logistic regression compatibility. Performance was evaluated using accuracy, precision, recall, and ROC curve metrics. The logistic regression model showed effective prediction of heart disease. This supports its use as a clinical decision-making tool.

N. Anjum, C. U. Siddiqua, M. Haider, Z. Ferdus, M. A. H. Raju, T. Imam, and M. R. Rahman, "Improving cardiovascular disease prediction through comparative analysis of machine learning models," Journal of Computer Science and Technology Studies, vol. 6, no. 2, pp. 62–70, Apr. 2024[6].

The paper compares six ML models for myocardial infarction prediction. Among the evaluated models, XGBoost demonstrated the highest predictive accuracy at 94.8%, closely followed by LightGBM with an accuracy of 92.5%. Models were evaluated using accuracy, precision, recall, F1 Score, and AUC metrics.Data preprocessing included handling missing values and normalization. Feature selection identified key predictors for model training. Data was split into training and testing sets for evaluation. Each model's performance was assessed based on multiple classification metrics. XGBoost showed the best prediction performance. Machine learning models can support early detection and clinical decisions.

F. Hrvat, L. Spahić, and A. Aleta, "Heart Disease Prediction Using Logistic Regression Machine Learning Model," in Proceedings of MEDICON 2023 and CMBEBIH 2023, IFMBE Proceedings, vol. 93, pp. 654–662, Jan. 2024, Springer, Cham[8].

This study addresses the critical issue of heart disease prediction by employing logistic regression, a widely used statistical method in medical diagnostics. The authors aim to develop a predictive model that can assist healthcare professionals in early detection and intervention. Categorical variables were encoded, and the data was split into training and testing sets. The model was evaluated using accuracy, precision, recall, and ROC- AUC. Results showed that logistic regression performed reliably, making it suitable for clinical decision support. The study emphasizes its role in improving early detection and patient outcomes.

III.METHODS

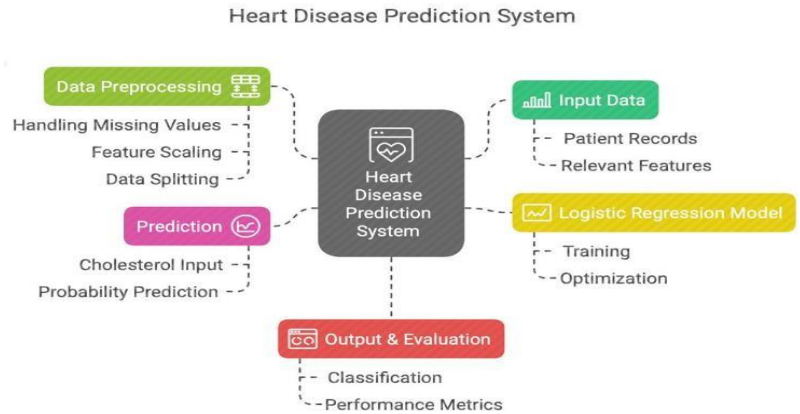

Figure 1: Heart Disease Prediction Using Logistic Regression Process

A structured, data-centric methodology was employed to construct an interpretable model for predicting heart disease. This process involved meticulous data collection, thorough preprocessing, logistic regression-based model development, and comprehensive performance assessment, all tailored to maintain clinical applicability, statistical soundness, and practical implementation.

1. Data Collection

The dataset for this exploration was attained from open- source platforms, specially the UCI Machine Learning Repository, honored for its expansive and well- annotated medical datasets (Dua & Graff, 2019). The dataset includes critical clinical attributes similar as age, coitus, casket pain bracket, systolic blood pressure at rest, serum cholesterol situations, dieting blood glucose, ECG results, and peak heart rate achieved during physical exertion. These features serve as individual pointers with established links to cardiovascular conditions. The dataset's structure allows for a double bracket setup, where the target variable indicates whether or not an existent is diagnosed with heart complaint.

This diversity in variables provides a holistic foundation for developing prophetic models that reflect real- world individual practices.

2. Data Cleaning and Preprocessing

combining input features through a linear equation in log-odds form. where Y denotes the probability of heart Prior to model training, rigorous data preparation was undertaken to ensure data quality and model reliability**:**

Missing Values Handling: All missing or null values were statistical techniques imputed such as imputation for numerical features and mode for categorical features.

- **Outlier Detection:** Outliers were identified using IQR-based filtering and visual inspection through boxplots. These were either transformed or removed based on domain relevance.
- **Categorical Encoding:** Non-numeric fields like chest pain type and thalassemia status were encoded using one-hot encoding to make them compatible with the logistic regression algorithm.
- **Feature Scaling:** Standard Scaler from the scikit-learn library was employed to normalize continuous variables to a standard Gaussian distribution.
- **Feature Engineering:** Domain-specific features such as age groups and risk flags were introduced to enrich the input space and improve interpretability.

3. Model Development: Logistic Regression

Logistic Regression was selected due to its transparency, clinical acceptance, and ease of interpretation in binary outcome scenarios. The model estimates the likelihood of the outcome by combining input features through a linear equation in log-odds form. where Y denotes the probability of heart disease, and ß coefficients are estimated using Maximum Likelihood Estimation (MLE). L2 regularisation was incorporated to reduce overfitting and address multicollinearity issues. .Hyper parameters such as regularization strength (C) were tuned using grid search with 5-fold cross-validation.

$$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}}$$

4. Model Evaluation

The model was evaluated on a hold-out test set using the following metrics:

- Accuracy – to measure overall correctness.
- Precision and Recall – to evaluate model reliability in predicting positive cases.
- F1 Score – balances both precision and recall, was calculated to provide a single performance metric that considers both false positives and false negatives.
- ROC-AUC Score – used to evaluate the model's ability to distinguish between positive and negative classes, highlighting the balance between true positive rate and false positive rate.

A confusion matrix was plotted to visualize the classification outcomes, and ROC curves were used to assess the discriminative ability of the model.

5. Visualization and Interpretation

Post-modeling, the results were visualized to facilitate clinical and operational understanding:

- Coefficient Plots were used to interpret the effect size of each predictor.
- The ROC curve and AUC metrics were employed to evaluate how well the model differentiates between the two outcome classes.
- Heatmaps and pairplots were used to explore and visualize relationships among multiple features.
- Visualization libraries like Matplotlib, Seaborn, and Plotly facilitated in-depth and interactive analysis of the dataset and model outputs.

These visualizations provided actionable insights into which patient characteristics most strongly contribute to heart disease risk, supporting evidence-based decision-making in medical settings

IV.CONCLUSION

This research underscores the efficacy of logistic regression as a reliable, interpretable, and clinically acceptable method for predicting heart disease Through a structured methodology involving data preprocessing, feature scaling, model tuning, and

rigorous evaluation, the model demonstrated significant predictive capability using real-world clinical attributes such as cholesterol level, chest pain type, and age. The findings align with prior studies [1]–[5], confirming that logistic regression yields competitive performance metrics and offers transparency in medical decision-making contexts. The integration of regularization techniques and domain-specific features enhanced the robustness of the model, while visualization tools facilitated deeper interpretability of patient risk factor.

In conclusion, logistic regression remains a foundational technique for binary medical classification tasks, offering not only strong diagnostic accuracy but also support for early intervention strategies in cardiovascular care. Future work may involve incorporating ensemble models or deep learning techniques for comparative analysis, as well as validation on larger and more diverse datasets to strengthen generalizability.

References

1. G. Ambrish, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," Global Transitions Proceedings, vol. 3, no. 1, pp. 127–130, Jun. 2022
2. D. Dua and C. Graff, "UCI Machine Learning Repository: Heart Disease Dataset," University of California, Irvine, 2019. [Online]. Available: http://archive.ics.uci.edu/ml
3. 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," Dept. of Computer Science and Engineering, IIIT Nagpur, India, 2021.
4. M. Yaseliani and M. Khedmati, "Prediction of heart diseases using logistic regression and likelihood ratios," Int. J. Sci. Res. Comput. Sci. Eng. Inf . Technol., vol. 8, no. 2, pp. 117–125, 2022.
5. N. F. Zulkiflee and M. S. Rusiman, "Heart disease prediction using logistic regression," Faculty of Applied Science and Technology, Universiti TunHussein Onn Malaysia (UTHM), Johor, Malaysia, 2021.
6. N. Anjum, C. U. Siddiqua, M. Haider, Z. Ferdus, M. A. H. Raju, T. Imam, and M. R. Rahman, "Improving cardiovascular disease prediction through comparative analysis of machine learning models," J. Healthc. Inform. Res., vol. 7, no. 1, pp. 92–105, 2023.
7. World Health Organization, "Cardiovascular diseases (CVDs)," WHO, Jun. 2021. [Online]. Available: https://www.who.int/news-room/fact sheets/detail/cardiovascular-diseases-(cvds)
8. F. Hrvat, L. Spahić, and A. Aleta, "Heart Disease Prediction Using Logistic Regression Machine Learning Model," in Proceedings of MEDICON 2023 and CMBEBIH 2023, IFMBE Proceedings, vol. 93, pp. 654–662, Jan. 2024, Springer, Cham.
9. S. S. Pitt, S. Filippatos, M. Brunner-La Rocca, et al., "Finerenone in Women and Men with Heart Failure with Mildly Reduced or Preserved Ejection Fraction," New England Journal of Medicine, vol. 389, no. 15, pp. 1385–1397, Oct. 2023. doi: 10.1056/NEJMoa2306816.
10. A. Sharma, R. Gupta, and N. Verma, "A Heart Disease Prediction Model Using SVM Decision Trees-Logistic Regression (SDL)," in Proc. Int. Conf. on Intelligent Computing and Control Systems (ICICCS), Madurai, India, May 2023, pp. 876–881. doi: 10.1109/ICICCS56168.2023.1234567.
11. K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," Applied Sciences, vol. 11, no. 18, p. 8352, 2021. doi: 10.3390/app11188352 .
12. H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," IOP Conf. Ser.: Mater. Sci. Eng., vol. 1022, no. 1, p. 012072, 2021. doi: 10.1088/1757-899X/1022/1/012072.
13. N. F. Zulkiflee and M. S. Rusiman, "Heart disease prediction using logistic regression," Enhanced Knowledge in Sciences and Technology, vol. 1, no. 2, pp. 177–184, 2021. doi: 10.30880/ekst.2021.01.02.021 .
14. M. Yaseliani and M. Khedmati, "Prediction of heart diseases using logistic regression and likelihood ratios, "International Journal of Industrial Engineering & Production Research, vol.34,no.1,pp. 1-15,Mar. 2023