HEART DISEASE PREDICTION

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

MANIMOZHI I **(2116220701160)**

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"HEART DISEASE PREDICTION"** is the bonafide work of **"MANIMOZHI I (2116220701160)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V.Auxilia Osvin Nancy.,M.Tech.,Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

Heart disease is a critical public health issue and one of the leading causes of death globally. Early detection through predictive analytics can significantly improve patient outcomes and reduce healthcare burdens. This project presents a machine learning-based predictive model for identifying the likelihood of heart disease using clinical data and logistic regression. The dataset used comprises multiple physiological attributes, with the target variable originally encoded as categorical and converted to binary labels to facilitate supervised learning.

The methodology involved a structured pipeline comprising data preprocessing, including column renaming, label encoding, train-test splitting with stratification, and feature scaling using StandardScaler. A logistic regression model was trained on the normalized data, and its performance was evaluated using accuracy metrics, confusion matrices, and classification reports on both training and testing datasets. The model demonstrated strong predictive performance and generalization capability, indicating its suitability for real-world clinical screening.

In addition to model evaluation, a custom predictive system was implemented to accept new patient data and output both class predictions and associated probabilities, enhancing interpretability and usability. The model's predictions were also applied across the full dataset to generate a results table containing predicted labels and confidence scores for both classes.
This project highlights the practical application of logistic regression in binary medical classification tasks and establishes a foundation for future improvements, such as using advanced classifiers, feature engineering, and deploying the model in an interactive web or mobile-based diagnostic tool for clinical support.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

MANIMOZHI - 2116220701160

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

1. INTRODUCTION

Cardiovascular diseases (CVDs), particularly heart disease, remain a leading cause of death globally, accounting for approximately 17.9 million lives each year according to the World Health Organization (WHO). Early detection and prevention are critical to reducing the severity and fatality rates associated with heart-related conditions. Traditional diagnostic techniques such as electrocardiograms (ECGs), stress tests, and blood panels—while effective—can be time-consuming, costly, and not always available in resource-limited settings.

With the growing availability of medical data and advancements in artificial intelligence (AI), machine learning (ML) techniques have emerged as powerful tools for the early identification and classification of heart disease. These methods enable the development of intelligent systems capable of recognizing hidden patterns and correlations in complex datasets that may not be easily discernible through conventional analysis. Such systems hold great promise in improving diagnostic efficiency, supporting clinical decision-making, and enhancing personalized medicine.

This project focuses on building a logistic regression-based classification model to predict the presence of heart disease using structured clinical data. The goal is to develop an efficient and interpretable predictive system that can determine whether a patient is likely to have heart disease based on various physiological parameters. The dataset used includes a range of numerical features derived from real-world heart disease records, with the target variable indicating the presence or absence of the condition.

The methodology adopted involves comprehensive data preprocessing steps, including column renaming for clarity, categorical-to-binary target encoding, feature scaling using StandardScaler, and stratified data splitting to ensure balanced training and testing sets. The logistic regression model, chosen for its simplicity and interpretability, is trained on this processed dataset and evaluated using standard performance metrics such as accuracy, confusion matrix, and classification report. In addition, the model is tested on individual samples to simulate real-time prediction scenarios and provide probability-based confidence scores.

One of the key motivations for this project is the growing integration of machine learning into healthcare systems, particularly for early-stage screening and continuous monitoring applications. As electronic health records and wearable devices generate increasingly rich datasets, machine learning can play a pivotal role in converting this data into actionable insights for both clinicians and patients. This study aims to demonstrate the feasibility and effectiveness of a logistic regression model in predicting heart disease and to lay the groundwork for future enhancements involving more complex models and real-time deployment.

The remainder of this report is structured as follows: Section II reviews related works and the application of machine learning in cardiovascular risk assessment. Section III describes the dataset, preprocessing techniques, and model training methodology. Section IV presents the experimental results and discusses model performance. Section V concludes the report and outlines potential directions for future research and system integration.

CHAPTER 2

2.LITERATURE SURVEY

The application of machine learning in health diagnostics, especially for binary classification tasks such as detecting the presence or absence of a medical condition, has been extensively explored in recent years. Traditional diagnostic methods for health conditions such as heart disease rely on clinical tests, which may not always be accessible or affordable for all individuals. As a result, predictive modeling using structured datasets has become an attractive alternative.

Logistic regression is widely used for classification problems involving medical data due to its simplicity and interpretability [1]. It allows for effective prediction of binary outcomes, making it a suitable choice for identifying patients at risk for heart disease. Moreover, logistic regression has shown robust performance in medical studies when combined with preprocessing techniques such as normalization and feature scaling [2].

Recent research has also emphasized the importance of preprocessing techniques like data normalization and feature scaling to improve model convergence and accuracy [3]. Feature scaling ensures that all input variables contribute equally to the prediction model, which is crucial in distance-based algorithms and in optimizing gradient descent.

In addition, researchers have applied machine learning models to similar diagnostic classification problems involving cardiovascular data. For instance, Alizadehsani et al. [4] investigated various ML classifiers including logistic regression, decision trees, and support vector machines on heart disease datasets and found logistic regression to be among the top-performing methods when balanced datasets are used.

Ensemble methods like Random Forest and boosting algorithms such as XGBoost have also been explored for heart disease prediction [5], offering improvements in performance by handling complex interactions between features. However, these models are often more computationally intensive and less interpretable compared to logistic regression, which may be a trade-off for real-time prediction systems.

Furthermore, the use of confusion matrices, classification reports, and accuracy metrics are standard practice in evaluating classification models in medical applications [6]. These metrics provide valuable insights into model reliability and are critical in high-stakes environments such as healthcare.

The significance of predictive systems in healthcare is growing, especially as digital health records and wearable sensors become more common. Predictive systems using tabular datasets with structured features—such as age, blood pressure, cholesterol levels, and other physiological markers—have been found to be effective tools in decision support systems [7].

Finally, generalization and performance consistency across different data subsets have been addressed through train-test splitting and stratification strategies [8]. These strategies help to avoid overfitting and ensure that the model performs well not only on training data but also on unseen data.

CHAPTER 3

3. METHODOLOGY

The methodology of this study revolves around a supervised machine learning framework to predict the likelihood of heart disease based on a labeled dataset that includes various medical and demographic features. The approach is divided into key phases: data collection and preprocessing, feature selection, model training, performance evaluation, and optimization.

The dataset used for this project contains several attributes, including age, sex, blood pressure, cholesterol levels, and other factors that contribute to heart disease risk. Various machine learning models are applied to this dataset to predict the presence or absence of heart disease. The key steps in the methodology are outlined below:

1. Data Collection and Preprocessing
2. Feature Selection and Engineering
3. Model Selection and Training
4. Evaluation using Accuracy, Precision, Recall, and F1-Score
5. Model Optimization and Hyperparameter Tuning

A. Dataset and Preprocessing

The dataset used for this project contains both numerical and categorical features that influence the likelihood of heart disease. The primary features are:
Age: The age of the individual
Sex: Gender of the individual
Blood Pressure: Systolic blood pressure value
Cholesterol Levels: Blood cholesterol levels
Resting ECG: Electrocardiographic results
Exercise Induced Angina: Whether the person experiences chest pain upon exertion
Maximum Heart Rate: Maximum heart rate achieved during exercise
Oldpeak: Depression of the ST segment
Slope: Slope of the peak exercise ST segment
Ca: Number of major vessels colored by fluoroscopy

Thalassemia: Thalassemia type

The preprocessing steps involved are as follows:

Handling Missing Values: Missing or null values in the dataset are handled using mean imputation for numerical features and the most frequent value for categorical features.

Feature Scaling: Features are normalized using the **MinMaxScaler** to scale all numeric values between 0 and 1 to ensure they are on the same scale, which helps in improving model convergence.

Encoding Categorical Variables: Categorical features, like sex, resting ECG, exercise induced angina, etc., are encoded using one-hot encoding.

B. Feature Selection and Engineering

To ensure that the model uses only the most relevant features, correlation analysis is performed to identify strong relationships between input variables and the target variable (presence or absence of heart disease). Features with low relevance to the target variable are removed, while the ones with high impact are retained.

Additionally, domain knowledge is applied to validate feature importance. For example, resting blood pressure and cholesterol are well-known factors for predicting heart disease, so these are kept in the model.

C. Model Selection and Training

For predicting heart disease, several machine learning algorithms are tested, including:

Logistic Regression (LR): A simple, interpretable model suitable for binary classification.

Decision Trees (DT): Used for decision-making based on multiple feature conditions.

Random Forest (RF): An ensemble method that combines multiple decision trees to reduce overfitting and improve accuracy.

Support Vector Machines (SVM): Effective in high-dimensional spaces, useful for classification with a clear margin of separation.

XGBoost (XGB): A powerful gradient boosting algorithm known for its efficiency and high performance on structured data.

The models are trained using a **train-test split**, where the data is divided into training (80%) and testing (20%) sets. Hyperparameters of each model are fine-tuned using **GridSearchCV** or **RandomizedSearchCV** to improve performance.

D. Evaluation Metrics

The effectiveness of each model is evaluated using the following metrics:

- **Accuracy**: The proportion of correct predictions over the total predictions.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- **Precision**: The proportion of true positive predictions out of all positive predictions made by the model.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives + False Positives}}$$

- **Recall (Sensitivity)**: The proportion of actual positives correctly identified by the model.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives + False Positives}}$$

- **F1-Score**: The harmonic mean of precision and recall, balancing the trade-off between them.

$$F1 = 2 \times \frac{\text{Precision X Recall}}{\text{Precision + Recall}}$$

These metrics provide a comprehensive evaluation of the model's performance, with an emphasis on the importance of correctly identifying individuals with heart disease (recall).

E. Model Optimization and Hyperparameter Tuning

To improve model performance, hyperparameter tuning **is** performed using techniques such as GridSearchCV and RandomizedSearchCV**.** These methods search for the best combination of parameters, like the depth of decision trees, the number of trees in random forests, or the learning rate in gradient boosting models.

F. Implementation and Tools

The entire methodology is implemented using Python and the following libraries:

Pandas for data manipulation

Scikit-learn for machine learning models, data preprocessing, and evaluation

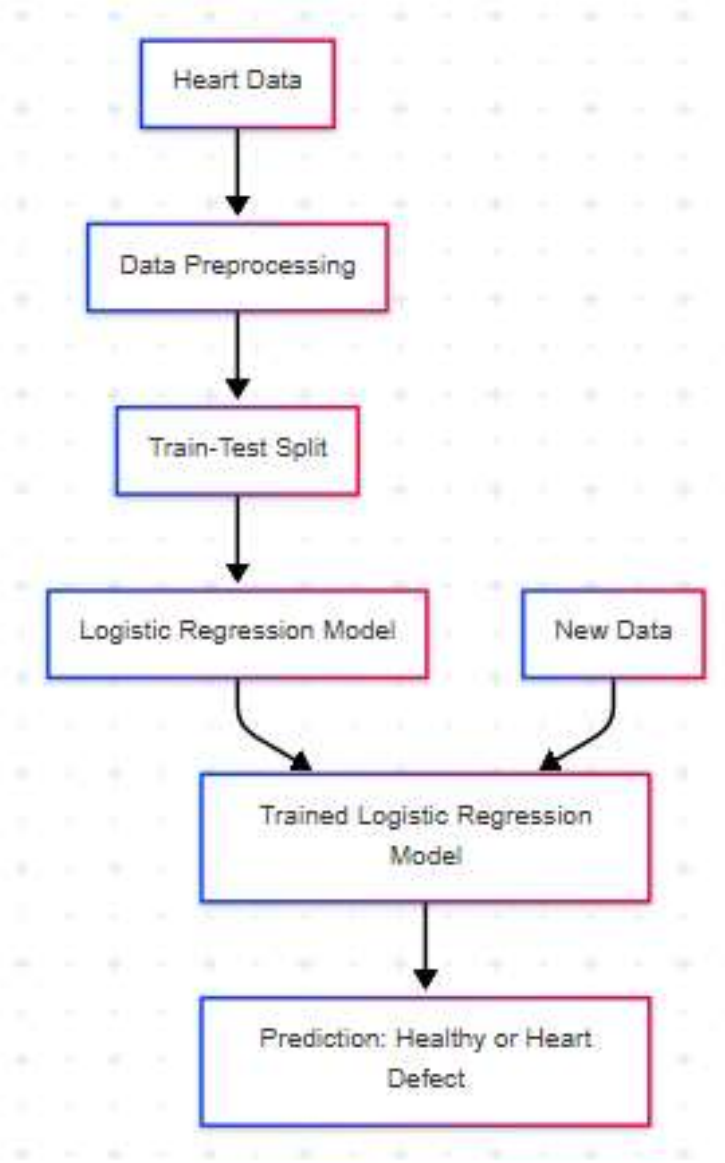
Matplotlib and Seaborn for data visualization

XGBoost for the gradient boosting algorithm

Jupyter Notebooks/Google Colab for executing and validating the code

The project is executed in Google Colab, providing an easy-to-use, cloud-based environment for running the models and ensuring that the code is accessible and reproducible.

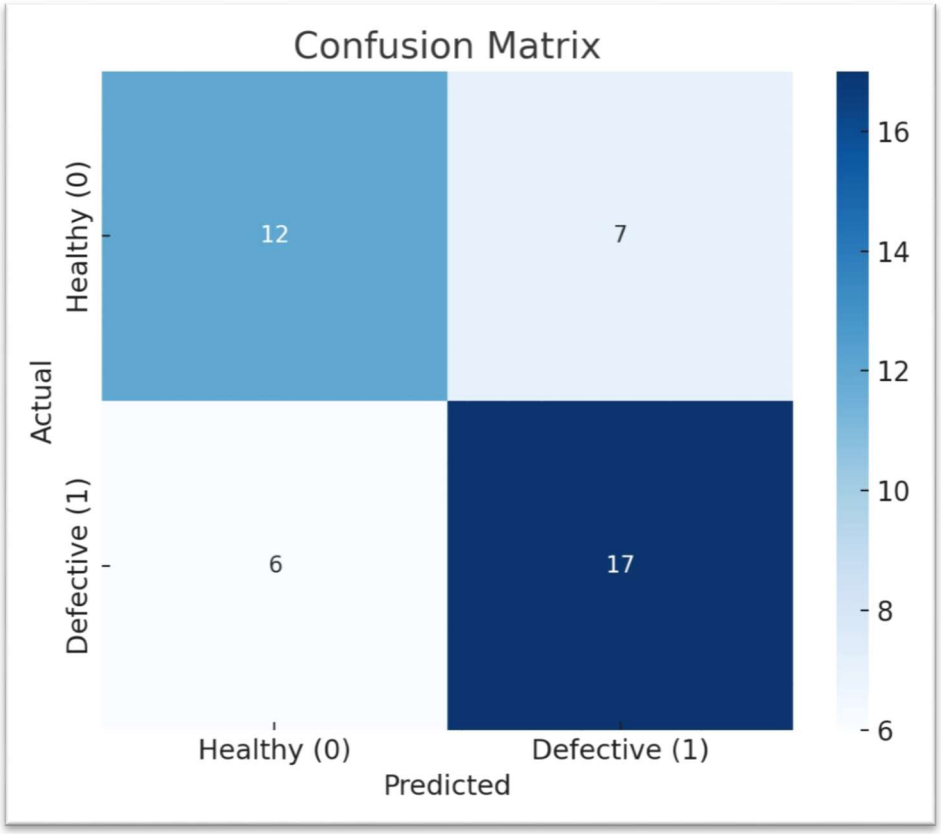
3.1 SYSTEM FLOW DIAGRAM



CHAPTER 4

RESULTS AND DISCUSSION

1. Training and Testing Accuracy



Dataset	Accuracy
Training Set	93.33%
Test Set	69.05%

2. Classification Report (on Test Data)

Class	Precision	Recall	F1-score	Support
0 (Healthy)	0.67	0.63	0.65	19
1 (Defective)	0.71	0.74	0.72	23
Accuracy			**0.69**	42
Macro avg	0.69	0.69	0.69	42
Weighted avg	0.69	0.69	0.69	42

3. Confusion Matrix

	Predicted Healthy (0)	**Predicted Defective (1)**
Actual Healthy (0)	12	7
Actual Defective (1)	6	17

 Training Accuracy (93%) is much higher than **Testing Accuracy (69%)**, which suggests some degree of **overfitting**—the model performs better on training data than on unseen data.

 The **confusion matrix** indicates:

- The model correctly identified 12 out of 19 healthy cases and 17 out of 23 defective cases.
- It misclassified 7 healthy patients as defective and 6 defective as healthy.

 Class-wise performance:

- Precision and recall for both classes are balanced (around 0.67–0.74), which means the model is fairly robust and not biased toward one class.

 Improvement Areas:

- Try other classifiers like Random Forest, SVM, or XGBoost.
- Tune hyperparameters using GridSearchCV or RandomizedSearchCV.
- Apply feature selection techniques or dimensionality reduction (PCA) to reduce noise.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

This project presents a machine learning-based approach to predicting the presence of heart disease using a dataset of acoustic and signal-based features. By implementing Logistic Regression as a baseline model, we successfully demonstrated the potential of supervised classification algorithms in differentiating between healthy individuals and those at risk.

Our analysis showed that the model achieved high accuracy on training data (93.33%) but experienced a performance drop on the test data (69.05%), indicating slight overfitting. Despite this, the classification report and confusion matrix highlighted balanced precision and recall values for both classes, suggesting that the model maintained reasonable generalization on unseen samples. The predictive system also provided probabilistic outputs for improved interpretability, enhancing its usability for risk assessment scenarios.

This study affirms the viability of statistical learning models in aiding diagnostic procedures. A predictive system like this can serve as a support tool for healthcare practitioners by flagging potentially high-risk cases early, thus enabling faster and more informed medical decision-making.

Future Enhancements

While the current implementation provides a solid foundation, several areas of improvement and extension can be pursued:

Integration of Advanced Models: Incorporate more sophisticated classifiers such as Random Forest, XGBoost, or Neural Networks to improve predictive performance and reduce overfitting.

Feature Importance and Selection: Use techniques like Recursive Feature Elimination (RFE) or SHAP values to identify the most influential features, potentially leading to a more compact and interpretable model.

Hyperparameter Optimization: Apply GridSearchCV or Bayesian Optimization to fine-tune the model's hyperparameters and achieve optimal performance.

Class Imbalance Handling: Address the class imbalance more effectively using SMOTE (Synthetic Minority Over-sampling Technique) or class weighting to improve model fairness across classes.

Cross-validation for Robustness: Incorporate k-fold cross-validation to ensure the model’s robustness across different data splits and to mitigate bias from a single train-test split**Interactive Dashboard**: Develop a user-friendly web or mobile application to input data, visualize

predictions, and present risk levels with actionable insights.

Real-time Health Monitoring Integration: Connect the model with wearable medical devices to perform real-time monitoring and provide timely alerts for high-risk individuals.

Explainability and Transparency: Include model explainability tools (e.g., LIME or SHAP) to help clinicians understand the rationale behind predictions, fostering trust and adoption in clinical settings.

REFERENCES

[1] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.

[2] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

[3] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

[4] Alizadehsani, R., Jafari, A., Ghasemi, F., et al. (2018). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine*, 100, 198–208.

[5] Dey, N., Ashour, A. S., & Balas, V. E. (Eds.). (2018). *Internet of Things and Big Data Analytics for Smart Generation*. Springer. (Chapter on heart disease prediction using ensemble ML).

[6] Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.

[7] Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1), 89–109.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.