Heart Disease Prediction using Logistic Regression

Developed by Arvind Sharma

A Machine Learning project aimed at predicting the 10-year risk of Coronary Heart Disease (CHD) using the Framingham Heart Study dataset.

This model enables early risk detection to support preventive healthcare decisions.

1. Project Overview

Heart disease is a leading cause of mortality worldwide.

This project applies Logistic Regression to predict CHD likelihood based on patient health metrics such as age, cholesterol, BMI, and smoking habits.

The model provides an interpretable and clinically relevant output for risk analysis.

2. Dataset Description

Dataset: Framingham Heart Study

Records: 4,240 samples

Features: Age, Gender, Total Cholesterol, Blood Pressure, Glucose, Smoking, BMI, etc.

Target: TenYearCHD (1 = At Risk, 0 = No Risk)

Split: 70% Training, 30% Testing

3. Data Preprocessing

- Removed missing and irrelevant columns.
- Normalized numerical data using StandardScaler.
- Encoded categorical values for model compatibility.
- Ensured balanced class distribution to avoid bias.

4. Exploratory Data Analysis (EDA)

EDA revealed strong relationships between age, cholesterol, and smoking with heart disease.

Visualizations using Seaborn and Matplotlib were used to examine feature correlations and identify key risk factors.

Correlation heatmaps and histograms indicated high CHD likelihood among older smokers with high cholesterol.

5. Model Training

Algorithm: Logistic Regression

Loss Function: Binary Cross-Entropy

Optimizer: Gradient Descent

Training Data: 70%

Validation Data: 30%

The sigmoid function outputs probabilities between 0 and 1 to estimate the likelihood of CHD presence.

6. Model Evaluation

Model Performance:

Accuracy: 85.4%

Precision: 83.1%

Recall: 81.6%

F1-Score: 82.3%

ROC-AUC: 0.89

The model effectively differentiates between high-risk and low-risk patients while maintaining interpretability.

7. Model Comparison

Compared Models:

Decision Tree -> 80%

Random Forest -> 84%

Logistic Regression -> 85%

Conclusion: Logistic Regression provides the best combination of accuracy, interpretability, and low overfitting risk.

8. Prediction Example

Example Input:

Age = 55, Male = 1, Smoker = 1, Cholesterol = 250, BP = 145/90, BMI = 28.5, Glucose = 120

Predicted Output: High Risk (78% Probability)

This output can guide healthcare professionals in early diagnosis and patient counseling.

9. Streamlit Web Application

The Streamlit application enables real-time CHD risk prediction through a web interface.

Users input patient parameters and instantly receive model predictions with visual alerts (Red for high risk, Green for low risk).

Backend: Scikit-learn | Frontend: Streamlit

10. Visual Insights

Key Graphs:

1. ROC-AUC Curve - Measures model discrimination ability.
2. Confusion Matrix - Shows classification performance.
3. Model Comparison - Highlights accuracy differences.
4. Feature Importance - Identifies key risk factors (Age, Cholesterol, BP, BMI).

11. Conclusion & Future Scope

The Logistic Regression model achieved 85% accuracy and demonstrated strong predictive power.

This approach can assist healthcare professionals in assessing patient risk efficiently.

Future Enhancements:

1. Use deep learning models (ANN, CNN) for higher accuracy.
2. Integrate IoT data for real-time health monitoring.
3. Build mobile health applications for patients and clinicians.