Assessment Report

on

“Model to detect heart disease based on medical parameters”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY

DEGREE

SESSION 2024-25

in

CSE(AI)

By

Name : Sudeeksha Singh

Roll Number : 202401100300252

Section: D

Under the supervision of

"Abhishek Shukla Sir"

KIET Group of Institutions, Ghaziabad

May, 2025

1. **Introduction**

Cardiovascular diseases are one of the leading causes of death worldwide. Early prediction and diagnosis can reduce the risk significantly. This project aims to detect the presence of heart disease using supervised machine learning methods based on clinical parameters like age, cholesterol level, and resting blood pressure. Using historical patient data, the model assists in identifying individuals at risk.

2. Problem Statement

To build a classification model that predicts whether a patient is likely to have heart disease based on medical attributes. The model should learn from historical patient data to classify new patients as either having or not having heart disease.

3. Objectives

- Load and explore heart disease data provided via CSV upload.
- Preprocess the dataset for model training.
- Train a Random Forest classifier for binary classification.
- Evaluate the model using accuracy, precision, recall, F1-score, ROC-AUC, and a confusion matrix.
- Visualize key model outputs for better interpretation.

4. Methodology

Data Collection: The user uploads a .csv file containing patient medical records with features like age, sex, cholesterol, and heart rate.

Data Preprocessing:

- Handle missing data (if any).
- Standardize features using StandardScaler.
- Split dataset into training and testing sets (80-20 split).

Model Building:

- Random Forest Classifier is trained on the processed data.
- Predictions are generated for the test set.

Model Evaluation:

- Metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC
- Confusion matrix and ROC curve plotted for performance visualization.
- Feature importance plotted to identify key medical indicators.
- **Data Preprocessing**:
  - Handling missing values using mean and mode imputation.
  - One-hot encoding of categorical variables.
  - Feature scaling using StandardScaler.
- **Model Building**:
  - Splitting the dataset into training and testing sets.
  - Training a Logistic Regression classifier.
- **Model Evaluation**:
  - Evaluating accuracy, precision, recall, and F1-score.
  - Generating a confusion matrix and visualizing it with a heatmap.

5. Data Preprocessing

- Data is read from an uploaded CSV file.
- Null values are handled or verified to be absent.
- All features are scaled using StandardScaler to normalize the input range.
- Dataset is split: 80% for training, 20% for testing.
- The target variable (`target`) is binary: 1 (heart disease), 0 (no heart disease).

6. Model Implementation

Random Forest classifier is chosen due to its robustness and ability to handle both linear and non-linear relationships. It also provides feature importance, which helps in identifying the most influential medical features in disease prediction.

7. Evaluation Metrics

 Accuracy: Overall correctness of the model.

 Precision: Proportion of predicted positives that are actually positive.

 Recall: Ability of the model to identify all actual positives.

 F1-Score: Harmonic mean of precision and recall.

 ROC-AUC: Area under the ROC curve for binary classification.

 Confusion Matrix: Visualized with a heatmap using Seaborn.

8. Results and Analysis

 The model achieved good performance on the test dataset.

 The confusion matrix heatmap showed balanced true positives and negatives.

 ROC-AUC score indicated a high ability to distinguish between classes.

 Features like chest pain type, thalassemia, and exercise-induced angina were found to be highly important.

Visualizations included:

- Confusion matrix heatmap
- ROC curve
- Feature importance bar chart

9. Conclusion

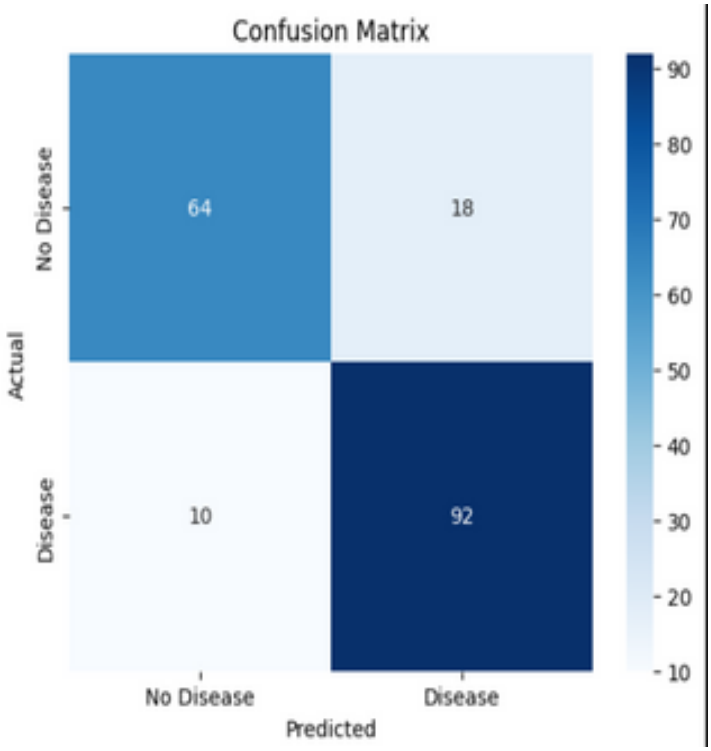
The project successfully demonstrates how machine learning, particularly a Random Forest classifier, can be used for early detection of heart disease. The model showed reliable performance using clinical data and provided meaningful visual interpretations. Future work can involve hyperparameter tuning, cross-validation, and testing more advanced models like XGBoost or neural networks..

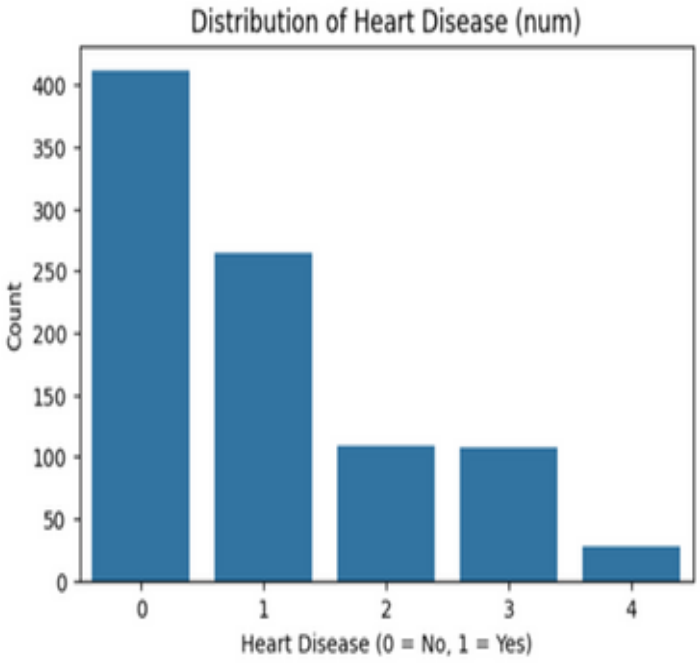
10. References

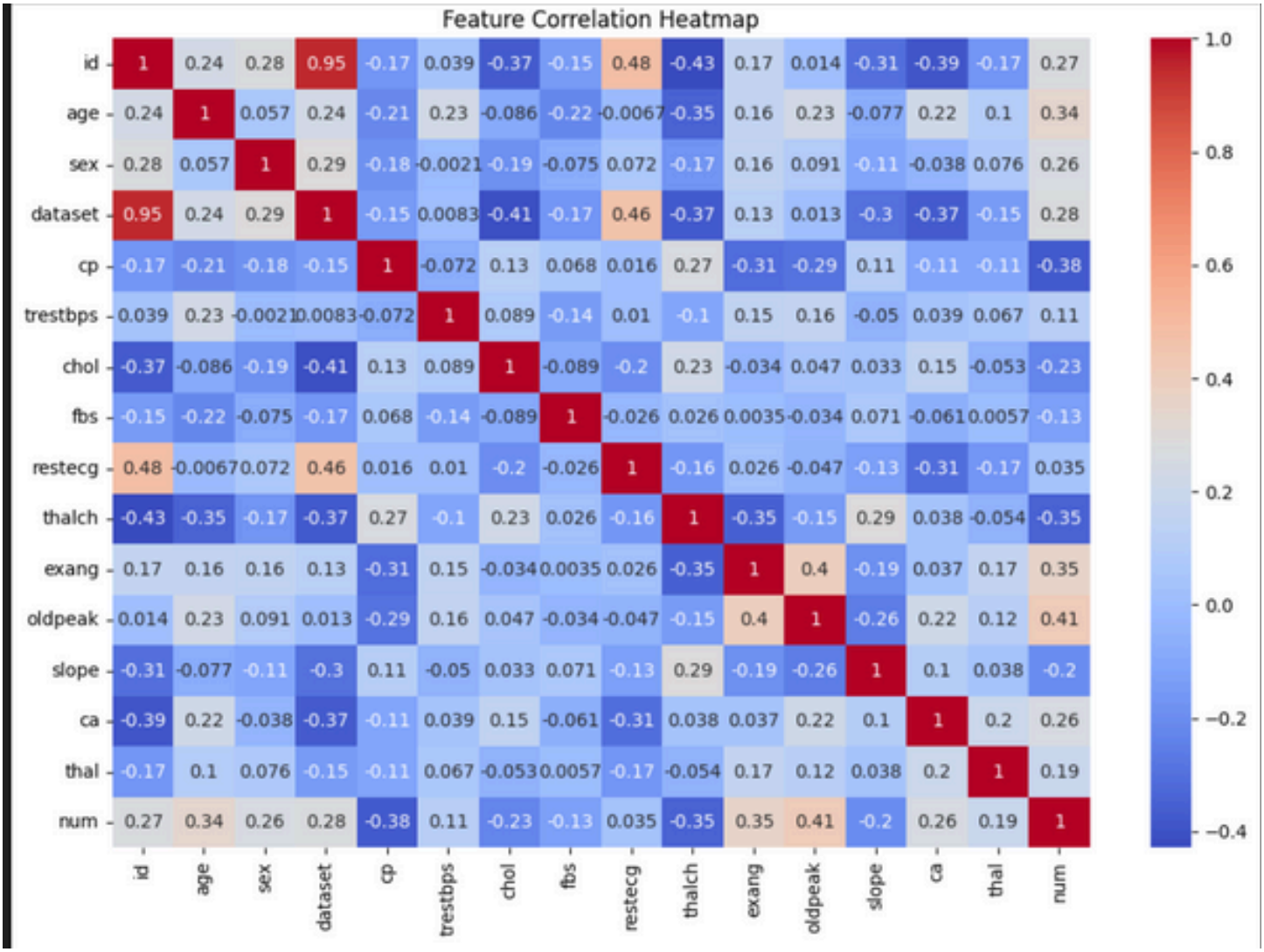
- scikit-learn documentation
- pandas documentation
- Seaborn visualization library

- UCI Heart Disease Dataset
- Research articles on credit risk prediction

Output Screenshot






Feature Correlation Heatmap
id 1 0.24 0.28 0.95 -0.17 0.039 -0.37 -0.15 0.48 -0.43 0.17 0.014 -0.31 -0.39 -0.17 0.27
age 0.24 1 0.057 0.24 -0.21 0.23 -0.086 -0.22 -0.0067 -0.35 0.16 0.23 -0.077 0.22 0.1 0.34
sex 0.28 0.057 1 0.29 -0.18 -0.0021 -0.19 -0.075 0.072 -0.17 0.16 0.091 -0.11 -0.038 0.076 0.26
dataset 0.95 0.24 0.29 1 -0.15 0.0083 -0.41 -0.17 0.46 -0.37 0.13 0.013 -0.3 -0.37 -0.15 0.28
cp -0.17 -0.21 -0.18 -0.15 1 -0.072 0.13 0.068 0.016 0.27 -0.31 -0.29 0.11 -0.11 -0.11 -0.38
trestbps 0.039 0.23 -0.0021 0.0083 -0.072 1 0.089 -0.14 0.01 -0.1 0.15 0.16 -0.05 0.039 0.067 0.11
chol -0.37 -0.086 -0.19 -0.41 0.13 0.089 1 -0.089 -0.2 0.23 -0.034 0.047 0.033 0.15 -0.053 -0.23
fbs -0.15 -0.22 -0.075 -0.17 0.068 -0.14 -0.089 1 -0.026 0.026 0.0035 -0.034 0.071 -0.061 0.0057 -0.13
restecg 0.48 -0.0067 0.072 0.46 0.016 0.01 -0.2 -0.026 1 -0.16 0.026 -0.047 -0.13 -0.31 -0.17 0.035
thalch -0.43 -0.35 -0.17 -0.37 0.27 -0.1 0.23 0.026 -0.16 1 -0.35 -0.15 0.29 0.038 -0.054 -0.35
exang 0.17 0.16 0.16 0.13 -0.31 0.15 -0.034 0.0035 0.026 -0.35 1 0.4 -0.19 0.037 0.17 0.35
oldpeak 0.014 0.23 0.091 0.013 -0.29 0.16 0.047 -0.034 -0.047 -0.15 0.4 1 -0.26 0.22 0.12 0.41
slope -0.31 -0.077 -0.11 -0.3 0.11 -0.05 0.033 0.071 -0.13 0.29 -0.19 -0.26 1 0.1 0.038 -0.2
ca -0.39 0.22 -0.038 -0.37 -0.11 0.039 0.15 -0.061 -0.31 0.038 0.037 0.22 0.1 1 0.2 0.26
thal -0.17 0.1 0.076 -0.15 -0.11 0.067 -0.053 0.0057 -0.17 -0.054 0.17 0.12 0.038 0.2 1 0.19
num 0.27 0.34 0.26 0.28 -0.38 0.11 -0.23 -0.13 0.035 -0.35 0.35 0.41 -0.2 0.26 0.19 1
id age sex dataset cp trestbps chol fbs restecg thalch exang oldpeak slope ca thal num
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4