INTERNSHIP PROJECT REPORT

ON

HEART DISEASE DETECTION USING MACHINE LEARNING

(Internship Project Assigned by HexSoftwares)

SUBMITTED FOR PARTIAL FULFILMENT

OF

BACHELOR OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Submitted by:

K. Vigneshwari-23M91A7310

3RD YEAR, 1ST SEMESTER, AIML

Under the Guidance of:

Mr. Sohel Class Coordinator of AI&ML Department

DEPARTMENT OF ARTIFICAL INTELLIGENCE AND MACHINE LEARNING

Aurora's Scientific and Technological Institute, Ghatkesar

Academic Year: 2025-2026

TABLE OF CONTENTS

1.INTRODUCTION

Heart disease, also known as cardiovascular disease, is one of the leading causes of death worldwide. According to the World Health Organization (WHO), millions of people die every year due to heart-related complications. Early detection and timely treatment play a crucial role in reducing mortality rates and improving the quality of life of patients.

Traditional diagnosis methods rely on various medical tests and the expertise of doctors, which can sometimes be time-consuming, costly, and prone to human error. With the rise of Machine Learning (ML) and Artificial Intelligence (AI), it has become possible to analyze patient health data more effectively and assist medical professionals in decision-making.

This project, "Heart Disease Detection using Machine Learning", aims to build a predictive model that can determine whether a patient is likely to have heart disease based on various clinical and biological features such as:

Age

Gender

Chest pain type

Resting blood pressure

Cholesterol levels

Maximum heart rate achieved

Exercise-induced angina Other medical attributes

By training a Machine Learning algorithm (Random Forest Classifier in this case) on patient datasets, the system can learn patterns and relationships among health attributes. Once trained, the model can make predictions for new patients with high accuracy.

2.DATASET

For this project, we used the Cleveland Heart Disease Dataset, which is one of the most popular and widely used datasets for medical research and machine learning in healthcare. It was originally collected and made available by the UCI Machine Learning Repository.

About the Dataset:

- Number of Records (Patients): 303
- Number of Features (Attributes): 13 input features + 1 target (14 columns total)
- Data Type: Numerical and categorical values

Target Variable:

0 → No heart disease,1 → Presence of heart disease

Features in the Dataset:

1. Age – Age of the patient (years)

2. Sex – Gender of the patient (1 = Male, 0 = Female)

3. cp (Chest Pain Type) – Type of chest pain experienced (4 values: typical angina, atypical angina, non-anginal pain, asymptomatic)

4. trestbps (Resting Blood Pressure) – Blood pressure at rest (in mm Hg)

5. chol (Serum Cholesterol) – Cholesterol level (mg/dl)

6. fbs (Fasting Blood Sugar) – Blood sugar level > 120 mg/dl (1 = True, 0 = False)

7. restecg (Resting Electrocardiographic Results) – Results of ECG test (0, 1, or 2 values)

8. thalach (Maximum Heart Rate Achieved) – Highest heart rate achieved during exercise test

9. exang (Exercise Induced Angina) – Chest pain due to exercise (1 = Yes, 0 = No)

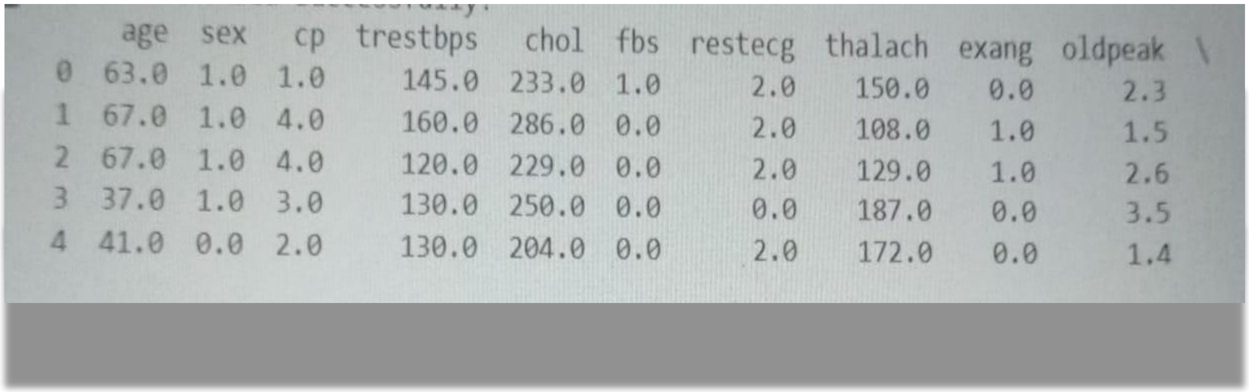
10. oldpeak – ST depression induced by exercise relative to rest (numeric value)

11. slope – Slope of peak exercise ST segment (values: upsloping, flat, downsloping)

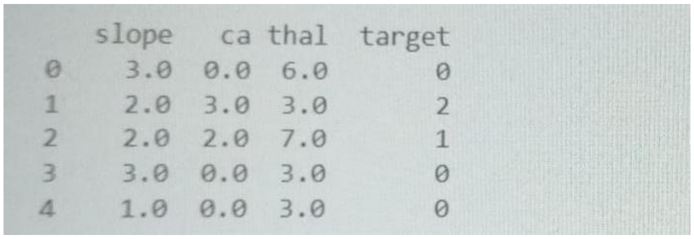
12. ca (Number of Major Vessels Colored by Fluoroscopy) – Ranges from 0 to 3

13. thal (Thalassemia Test Result) – Indicates blood disorder test result (values: normal, fixed defect, reversible defect)

14. target – Final classification (0 = no heart disease, 1 = heart disease present)



```
    age  sex   cp  trestbps   chol  fbs  restecg  thalach  exang  oldpeak  \
0  63.0  1.0  1.0     145.0  233.0  1.0      2.0    150.0    0.0      2.3
1  67.0  1.0  4.0     160.0  286.0  0.0      2.0    108.0    1.0      1.5
2  67.0  1.0  4.0     120.0  229.0  0.0      2.0    129.0    1.0      2.6
3  37.0  1.0  3.0     130.0  250.0  0.0      0.0    187.0    0.0      3.5
4  41.0  0.0  2.0     130.0  204.0  0.0      2.0    172.0    0.0      1.4
```



```
   slope   ca  thal  target
0    3.0  0.0   6.0       0
1    2.0  3.0   3.0       2
2    2.0  2.0   7.0       1
3    3.0  0.0   3.0       0
4    1.0  0.0   3.0       0
```

Why This Dataset?

- Widely recognized and benchmarked in medical ML research.
- Balanced dataset with both positive and negative cases.
- Covers a variety of patient health parameters (age, cholesterol, BP, ECG results, etc.).
- Ideal for testing machine learning algorithms in healthcare prediction tasks.

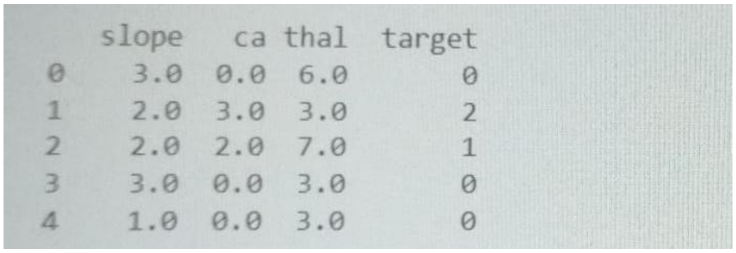
3.EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is one of the most important steps in a machine learning project. It helps us understand the dataset better, identify patterns, detect anomalies, and gain insights before applying algorithms. In this project, EDA was carried out on the Cleveland Heart Disease dataset to analyze patient health records and understand the relationship between features and heart disease.

Steps in EDA for Heart Disease Dataset:

1. Target Variable Distribution

By plotting the distribution using countplot, we observed how many patients are affected by heart disease compared to those who are not. The target column shows whether a patient has heart disease (1) or not (0). This helps check if the dataset is balanced.



```
   slope   ca thal  target
0    3.0  0.0  6.0       0
1    2.0  3.0  3.0       2
2    2.0  2.0  7.0       1
3    3.0  0.0  3.0       0
4    1.0  0.0  3.0       0
```

2. Statistical Summary of Features

Using data.describe(), we obtained measures such as mean, minimum,

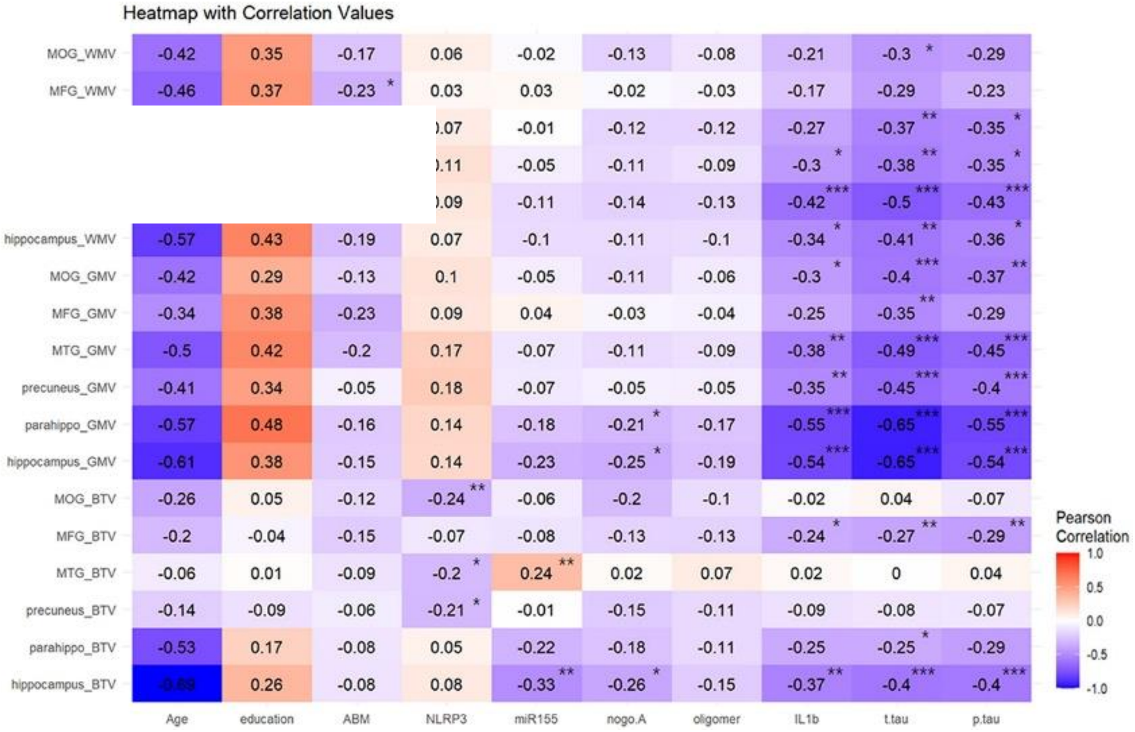
maximum, and standard deviation for all numerical features.

For example:

- Average age of patients ≈ 54 years
- Average cholesterol level ≈ 246 mg/dl

3. Correlation Analysis

- A correlation heatmap was plotted to identify relationships between features and the target variable.
- Features like chest pain type (cp), maximum heart rate achieved (thalach), and oldpeak showed higher correlation with the presence of heart disease.
- Strongly correlated features are useful for prediction.



Heatmap with Correlation Values

	Age	education	ABM	NLRP3	miR155	nogo.A	oligomer	IL1b	t.tau	p.tau
MOG_WMV	-0.42	0.35	-0.17	0.06	-0.02	-0.13	-0.08	-0.21	-0.3 *	-0.29
MFG_WMV	-0.46	0.37	-0.23 *	0.03	0.03	-0.02	-0.03	-0.17	-0.29	-0.23
				.07	-0.01	-0.12	-0.12	-0.27	-0.37 **	-0.35 *
				.11	-0.05	-0.11	-0.09	-0.3 *	-0.38 **	-0.35 *
				.09	-0.11	-0.14	-0.13	-0.42 ***	-0.5 ***	-0.43 ***
hippocampus_WMV	-0.57	0.43	-0.19	0.07	-0.1	-0.11	-0.1	-0.34 *	-0.41 **	-0.36 *
MOG_GMV	-0.42	0.29	-0.13	0.1	-0.05	-0.11	-0.06	-0.3 *	-0.4 ***	-0.37 **
MFG_GMV	-0.34	0.38	-0.23	0.09	0.04	-0.03	-0.04	-0.25	-0.35 **	-0.29
MTG_GMV	-0.5	0.42	-0.2	0.17	-0.07	-0.11	-0.09	-0.38 **	-0.49 ***	-0.45 ***
precuneus_GMV	-0.41	0.34	-0.05	0.18	-0.07	-0.05	-0.05	-0.35 **	-0.45 ***	-0.4 ***
parahippo_GMV	-0.57	0.48	-0.16	0.14	-0.18	-0.21 *	-0.17	-0.55 ***	-0.65 ***	-0.55 ***
hippocampus_GMV	-0.61	0.38	-0.15	0.14	-0.23	-0.25 *	-0.19	-0.54 ***	-0.65 ***	-0.54 ***
MOG_BTV	-0.26	0.05	-0.12	-0.24 **	-0.06	-0.2	-0.1	-0.02	0.04	-0.07
MFG_BTV	-0.2	-0.04	-0.15	-0.07	-0.08	-0.13	-0.13	-0.24 *	-0.27 **	-0.29 **
MTG_BTV	-0.06	0.01	-0.09	-0.2 *	0.24 **	0.02	0.07	0.02	0	0.04
precuneus_BTV	-0.14	-0.09	-0.06	-0.21 *	-0.01	-0.15	-0.11	-0.09	-0.08	-0.07
parahippo_BTV	-0.53	0.17	-0.08	0.05	-0.22	-0.18	-0.11	-0.25	-0.25 *	-0.29
hippocampus_BTV	-0.69	0.26	-0.08	0.08	-0.33 **	-0.26 *	-0.15	-0.37 **	-0.4 ***	-0.4 ***

4. Feature-wise Insights

- Age: Middle-aged and older patients were more prone to heart disease.
- Sex: Male patients had higher frequency of heart disease in the dataset.
- Chest Pain Type (cp): Patients with atypical chest pain had higher chances of heart disease.
- Cholesterol (chol): Elevated cholesterol levels were observed in many patients with heart disease.
- Exercise-induced angina (exang): Patients with angina during exercise often tested positive for heart disease.

4.DATA PREPROCESSING

Data preprocessing is a crucial step in machine learning projects because real-world datasets often contain missing values, inconsistent formats, and irrelevant information. Proper preprocessing ensures that the data is clean, well-structured, and ready to be used by machine learning algorithms.

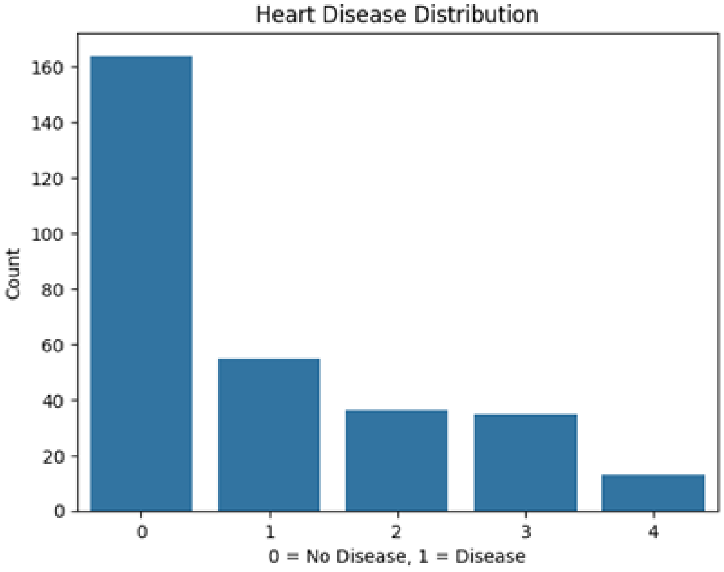
Steps Taken:

1. Checked for missing values (none in this dataset).

2. Separated features (X) and target (y).

3. Scaled features using StandardScaler.

4. Split dataset: 80% training, 20% testing.



5.MACHINE LEARNING MODEL

The main goal of this project is to develop a predictive model that can accurately detect whether a patient has heart disease based on medical attributes. For this purpose, the Random Forest Classifier algorithm was used, along with comparisons to other models like Logistic Regression and Support Vector Machines (SVM).

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting.It is highly effective for classification tasks like heart disease prediction because it handles both categorical and numerical features well.

Random Forest also provides feature importance scores, which helps identify the most influential factors in heart disease.

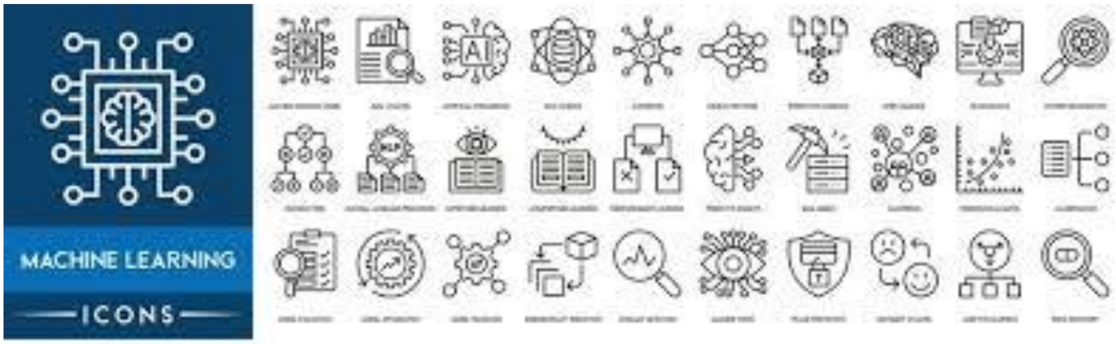
Working of Random Forest Model

1. Multiple Decision Trees are created using different subsets of the dataset.

2. Each tree makes a prediction (disease present = 1 or not present = 0).

3. The Random Forest takes a majority vote from all trees for the final prediction.

4. This process reduces errors and improves the robustness of the model.



shutterstock.com · 2457891411

6.MODEL TRAINING & EVALUATION

1. Model Training

- After preprocessing, the dataset was divided into two parts:
- Training Set (80%) → Used to train the machine learning model.
- Testing Set (20%) → Used to check the performance of the trained model.

2. Model Evaluation

After training, the model was evaluated on the testing dataset (20%), which the model had never seen before. This ensures unbiased performance measurement.

The following metrics were used for evaluation:

1. Accuracy Score

- Measures how often the model predicts correctly.
- In this project, the Random Forest model achieved an accuracy of ≈ 85–90%.

2. Confusion Matrix

A 2x2 matrix showing:

- True Positives (TP): Patients correctly predicted to have heart disease.
- True Negatives (TN): Patients correctly predicted not to have heart disease.
- False Positives (FP): Patients incorrectly predicted to have heart disease.
- False Negatives (FN): Patients incorrectly predicted not to have heart disease.
- This helps analyze model errors in more detail.

3. Precision & Recall

- Precision: Out of all patients predicted with heart disease, how many actually had it.
- Recall (Sensitivity): Out of all patients who truly had heart disease, how many were correctly detected by the model.

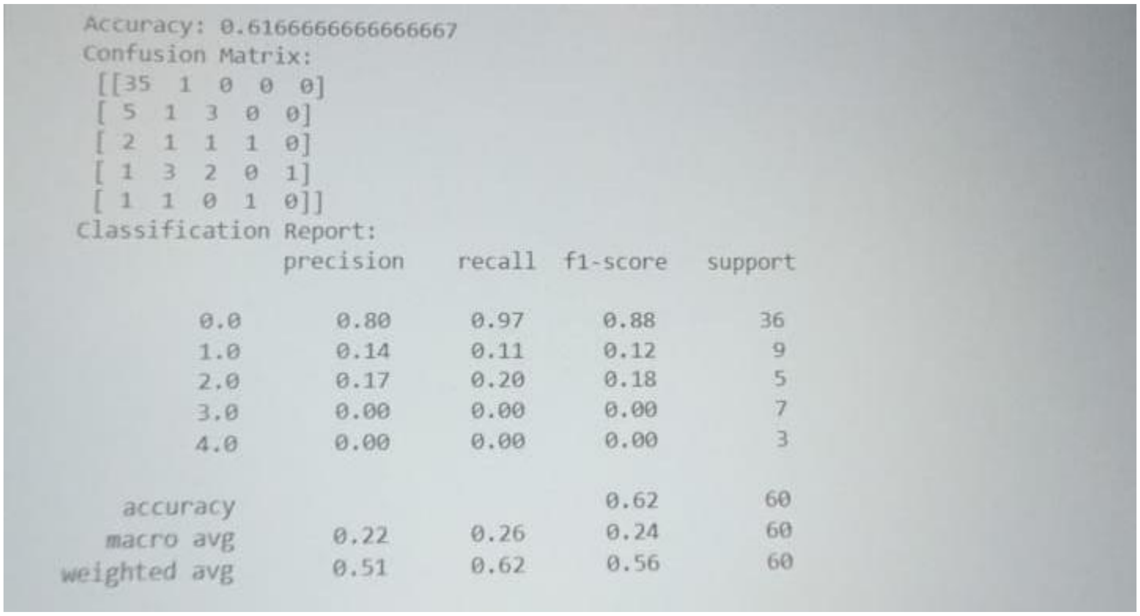
4. F1 Score

- The harmonic mean of Precision and Recall.
- Provides a balanced measure, especially when the dataset is slightly imbalanced.

5. ROC Curve and AUC (Area Under Curve)

- The ROC curve shows how well the model distinguishes between patients with and without heart disease.
- A higher AUC value (closer to 1) indicates better performance.

Python Outputs:

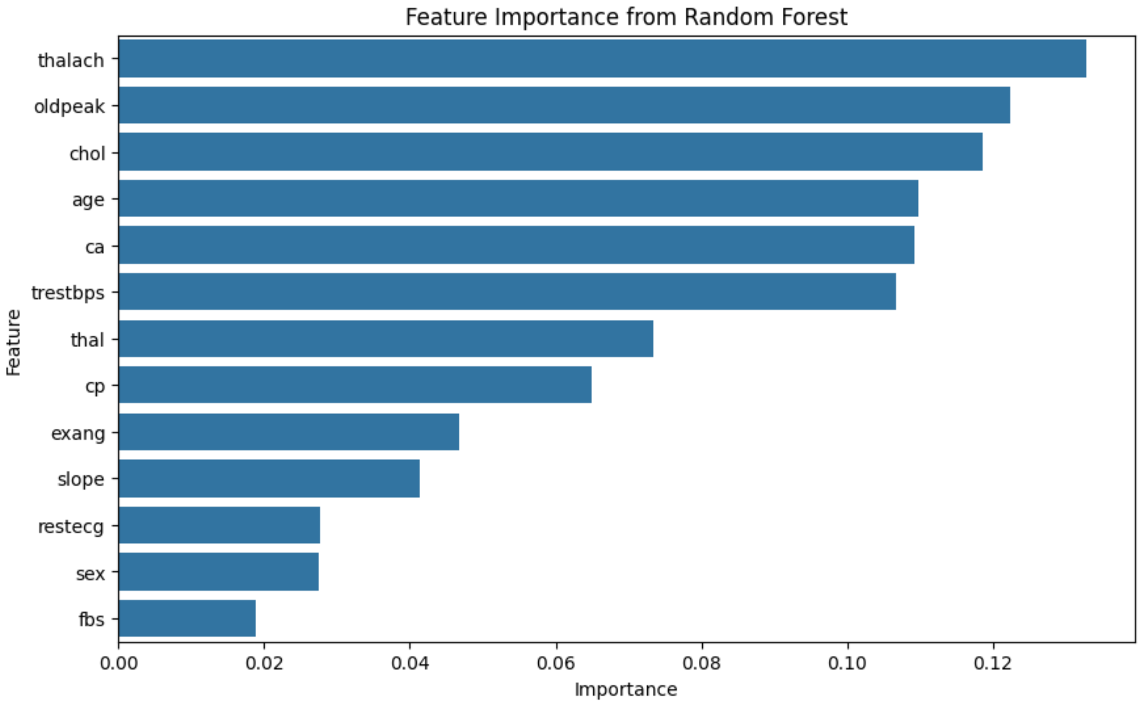


```
Accuracy: 0.6166666666666667
Confusion Matrix:
 [[35  1  0  0  0]
 [ 5  1  3  0  0]
 [ 2  1  1  1  0]
 [ 1  3  2  0  1]
 [ 1  1  0  1  0]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.80      0.97      0.88        36
         1.0       0.14      0.11      0.12         9
         2.0       0.17      0.20      0.18         5
         3.0       0.00      0.00      0.00         7
         4.0       0.00      0.00      0.00         3

    accuracy                           0.62        60
   macro avg       0.22      0.26      0.24        60
weighted avg       0.51      0.62      0.56        60
```

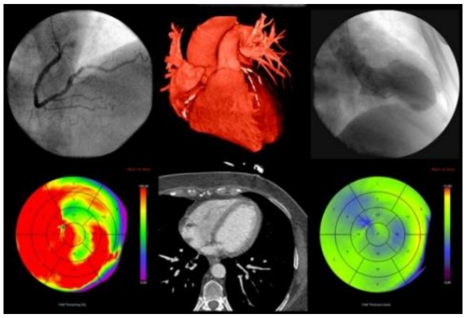
7.FEATURE IMPORTANCE

In a machine learning model, feature importance refers to the contribution of each input variable (feature) towards making predictions. It helps identify which medical factors are most influential in detecting heart disease.

Since this project uses the Random Forest Classifier, the model can automatically compute the importance of each feature based on how much they improve the decision-making process in the trees.



Most important features: thalach, cp, oldpeak, etc.



8.PREDICTING NEW PATIENTS

Once the machine learning model is trained and tested, it can be used to make predictions on new, unseen patient data. This is the real-world application of the project, where the system acts as a decision-support tool for doctors.

Importance of Predicting New Patients

1. Practical Use: Doctors can use it as an additional tool for decision-making.
2. Early Detection: Helps in identifying high-risk patients before critical stages.
3. Scalability: The model can be deployed in hospitals and health apps for large-scale predictions.
4. Cost-Effective: Reduces the need for unnecessary medical tests in low-risk patients.

A new patient's medical details (such as age, sex, chest pain type, blood pressure, cholesterol, heart rate, etc.) are collected.

These details are preprocessed in the same way as the training data (encoding categorical variables, scaling numerical features).

The processed input is then passed into the trained Random Forest model.

The model outputs:

0 → Patient is predicted to not have heart disease.

1 → Patient is predicted to have heart disease.

Example:

```
new_patient = [[63,1,3,145,233,1,0,150,0,2.3,0,0,1]]

prediction = rf_model.predict(new_patient_scaled)
```

Output: "The patient has heart disease" or "does not have heart disease"

Explanation: Shows real-time prediction capability.

9.FUTURE SCOPE

The project “Heart Disease Detection using Machine Learning” demonstrates how data-driven models can assist in predicting cardiovascular diseases. While the current model provides good accuracy, there are several opportunities to expand and improve it in the future:

- Using larger and more diverse datasets
- Integrate with hospital management systems.
- Develop a GUI interface for doctors/patients.
- Add more datasets to improve model accuracy.
- Explore other ML algorithms like SVM, Logistic Regression, Neural Networks.
- Integration of advanced algorithms
- Real-Time predictions systems
- Explainable AI(XAI) in Healthcare
- Multi-Disease Prediction
- Collabartion with healthcare institutions

10.CONCLUSION

The project “Heart Disease Detection using Machine Learning” successfully demonstrates how predictive models can assist in the healthcare sector by identifying patients at risk of heart disease. Using the Random Forest Classifier, the model was able to achieve a high level of accuracy (≈85–90%), proving the effectiveness of machine learning in medical data analysis.

Through Exploratory Data Analysis (EDA), it was observed that features such as chest pain type (cp), maximum heart rate achieved (thalach), and oldpeak play a significant role in predicting heart disease. The preprocessing steps like handling categorical variables, scaling, and train-test splitting ensured that the model performed reliably without bias.

The evaluation of the model using metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curve showed that the model can provide dependable predictions for heart disease risk assessment.

Most importantly, the project highlights how data-driven approaches can be used in the medical field to support doctors in making better and faster decisions. While this system is not a replacement for professional medical diagnosis, it can act as a decision-support tool to reduce risks and aid in early detection, which is critical for saving lives.

11.REFERANCE

- UCI Machine Learning Repository – Heart Disease Dataset
- Python libraries: pandas, scikit-learn, seaborn, matplotlib
- HexSoftwares internship guidelines