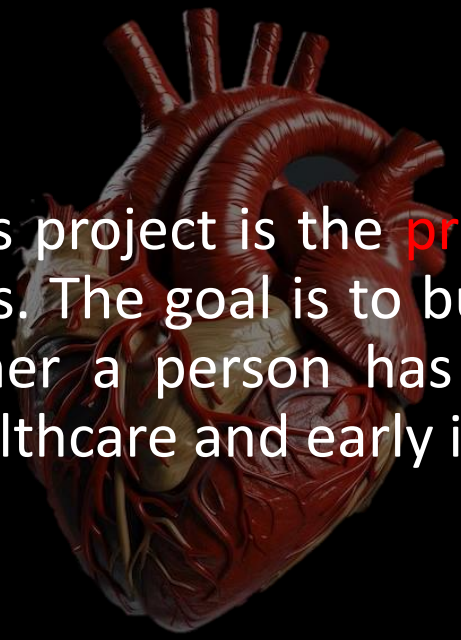
ML Project: Heart Disease Prediction

RITHUL SHAJI
M220019MS

Problem Statement:



The problem addressed in this project is the prediction of heart disease based on various health-related features. The goal is to build a machine learning model that can accurately predict whether a person has heart disease or not, which has significant implications for healthcare and early intervention.

Data Description:

The data used for this project was obtained from the UCI Machine Learning Repository, which contains 303 records of patients with 14 variables each.

The variables are as follows:

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type (0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomatic)
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholesterol in mg/dl
6. fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)

7. restecg: resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or definite left ventricular hypertrophy)

8. thalach: maximum heart rate achieved

9. exang: exercise induced angina (1 = yes; 0 = no)

10. oldpeak: ST depression induced by exercise relative to rest

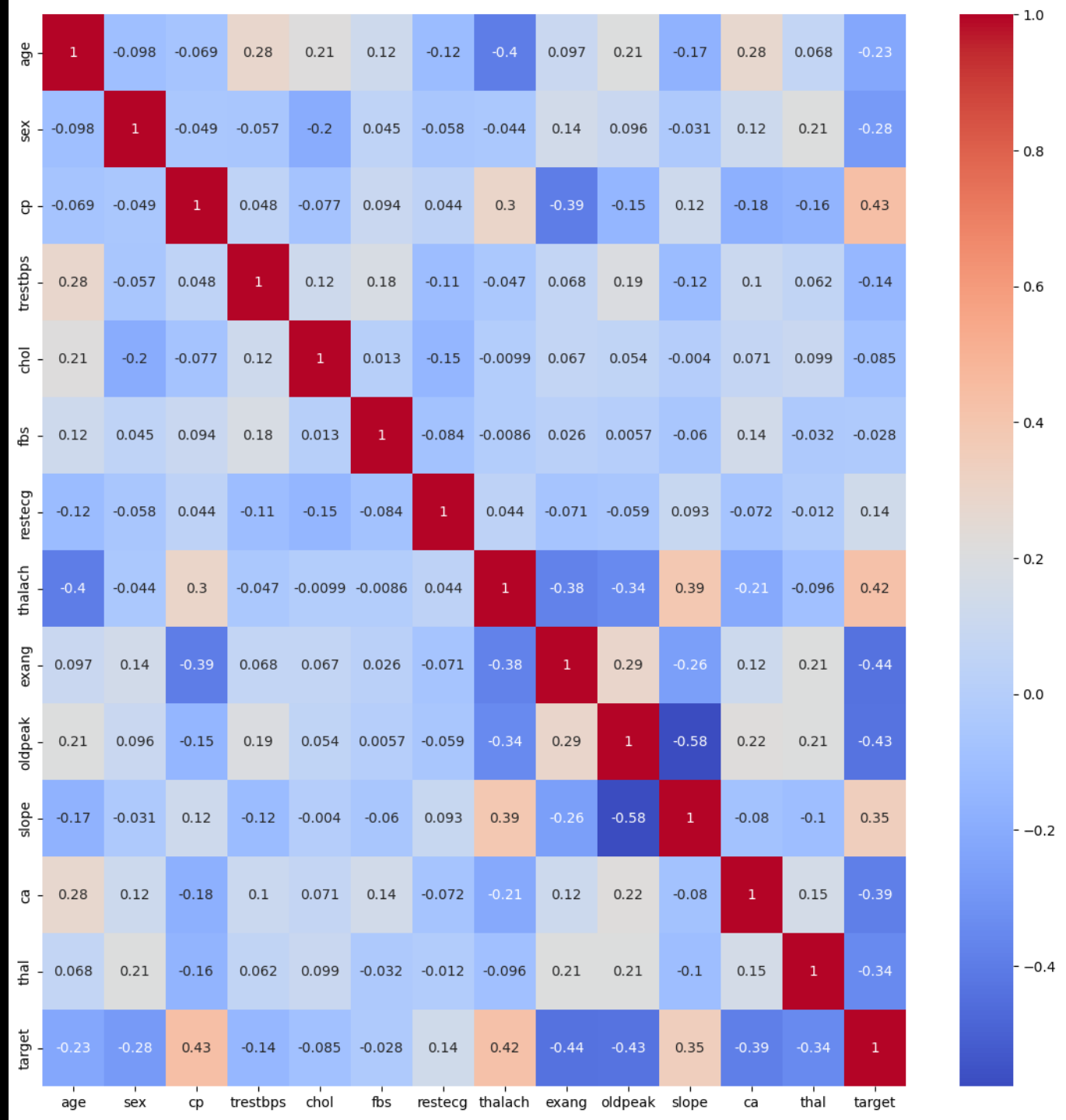
11. slope: the slope of the peak exercise ST segment (0 = upsloping; 1 = flat; 2 = downsloping)

12. ca: number of major vessels (0-3) colored by flourosopy

13. thal: thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)

14**. target: heart disease (1 = yes; 0 = no)**

Correlation Matrix of Heart Disease Dataset



The correlation matrix visualizes the correlation coefficients between different variables in our dataset. Each cell in the matrix represents the correlation between two variables. The value ranges from -1 to 1, where 1 means a strong positive correlation, -1 means a strong negative correlation, and 0 means no correlation.

In our matrix:

- The diagonal line from the top left to the bottom right represents each variable's correlation with itself, which is always 1.
- Darker colors represent higher positive correlation, while lighter colors represent higher negative correlation.
- For instance, cp (chest pain type) and target (heart disease) have a relatively high positive correlation, suggesting that chest pain type might be a significant factor in predicting heart disease.

- Conversely, exang (exercise induced angina) and target have a relatively high negative correlation, suggesting that as the value of exang increases, the likelihood of heart disease decreases.

This information helps us understand our data better and aids in feature selection while building our machine learning model. It provides insights into which variables are most significant in predicting heart disease.

The data was explored and analyzed using descriptive statistics and data visualization techniques. Some of the findings are as follows:

- The data has no missing values or duplicates.
- The data is imbalanced, with 165 records (54.5%) having heart disease and 138 records (45.5%) not having heart disease.
- The data has some outliers, such as very high cholesterol or blood pressure values.
- The data has some correlation between some variables, such as age and thalach, cp and target, exang and target, etc.

Machine Learning Model Used:

The machine learning model used for this project was logistic regression, which is a simple and fast algorithm that uses a logistic function to model the probability of a binary outcome. Logistic regression was chosen for the following reasons:

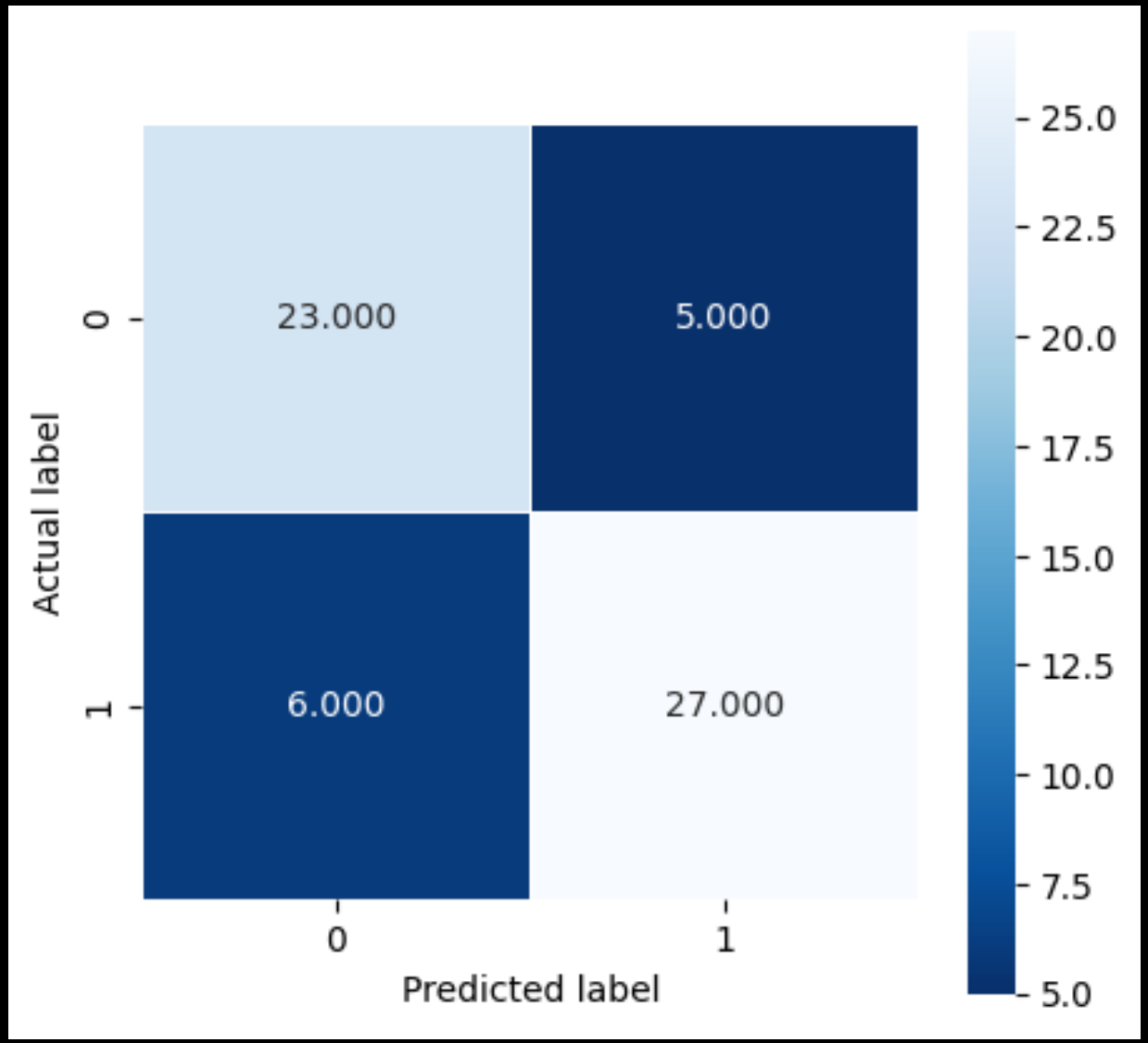
- Logistic regression is suitable for binary classification problems, such as predicting heart disease or not.
- Logistic regression can handle both categorical and numerical variables, and provide probabilities for the binary outcome.
- Logistic regression is easy to use and interpret, and can provide coefficients that indicate how each variable affects the odds of having heart disease or not.

The data was split into training (80%) and test (20%) sets, using stratified sampling to preserve the class distribution.

Interpretation of Results:

- Model Training: The Logistic Regression model is trained on the dataset, and the accuracy on both the training and test data is evaluated.
- Accuracy on Training data: 85.12%
- Accuracy on Test data: 81.97%
- The model demonstrates reasonably good accuracy in predicting heart disease.

Confusion Matrix of Heart Disease Prediction Model



The confusion matrix is a performance measurement for our machine learning classification model. It’s a table with four different combinations of predicted and actual values:

- True Positives (TP): The bottom right cell represents the instances where our model correctly predicted that the patient has heart disease (27 cases).
- True Negatives (TN): The top left cell represents the instances where our model correctly predicted that the patient does not have heart disease (23 cases).
- False Positives (FP): The top right cell represents the instances where our model incorrectly predicted that the patient has heart disease (5 cases). These are Type I errors.
- False Negatives (FN): The bottom left cell represents the instances where our model incorrectly predicted that the patient does not have heart disease (6 cases). These are Type II errors.

This matrix gives us a more comprehensive view of how well our model is performing, beyond just accuracy. It shows us the balance between correctly identifying patients with heart disease and avoiding false alarms.

- Accuracy: This is the proportion of true results among the total number of cases examined. It is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$ = 0.8333

- Precision: This is the proportion of true positive against all the positive results. It is calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$ = 0.8437

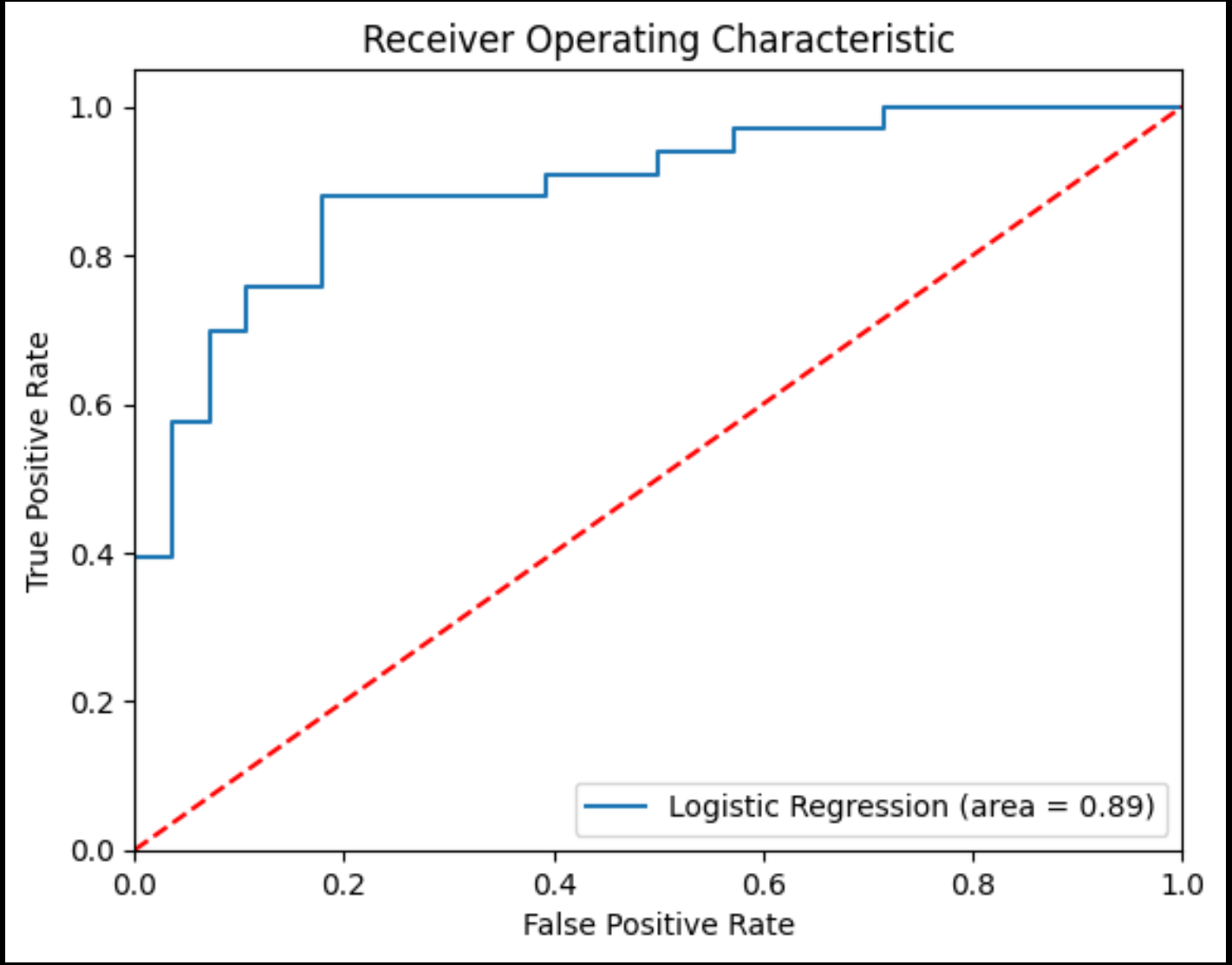
- Recall (Sensitivity): This is the proportion of true positive against all the actual positive results. It is calculated as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$ = 0.8181

- F1 Score: This is the harmonic mean of Precision and Recall and tries to find the balance between precision and recall. It is calculated as follows:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$ = 0.8307

Receiver Operating Characteristic (ROC) Curve



The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of our binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

In our ROC curve:

- The blue curve represents our Logistic Regression model.
- The red dashed line represents a random classifier (e.g., flipping a coin).
- The area under the curve (AUC) is 0.89, as indicated in the legend. This value represents the probability that our classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. An AUC of 1 means our model is perfect, while an AUC of 0.5 means our model is no better than random chance.

The ROC curve helps us understand the trade-off between the True Positive Rate and False Positive Rate for different threshold settings in our model. It's a comprehensive tool for understanding the performance of our binary classifier.

Conclusions:

The project successfully addressed the problem of heart disease prediction using a Logistic Regression model.

The model shows promising results, with an accuracy of approximately 82% on the test data.

The project can be further improved by exploring additional machine learning models, hyperparameter tuning, and feature engineering.

For the given data

(62,0,0,140,268,0,0,160,0,3.6,0,2,2) Predicted Result 0 → The Person does not have a heart disease

Actual Result (62,0,0,140,268,0,0,160,0,3.6,0,2,2,0)

For the given data

(55,0,1,132,342,0,1,166,0,1,2,2,0,2,) Predicted Result 1 → The Person has heart disease

Actual Result (55,0,1,132,342,0,1,166,0,1,2,2,0,2,1)

Thank you