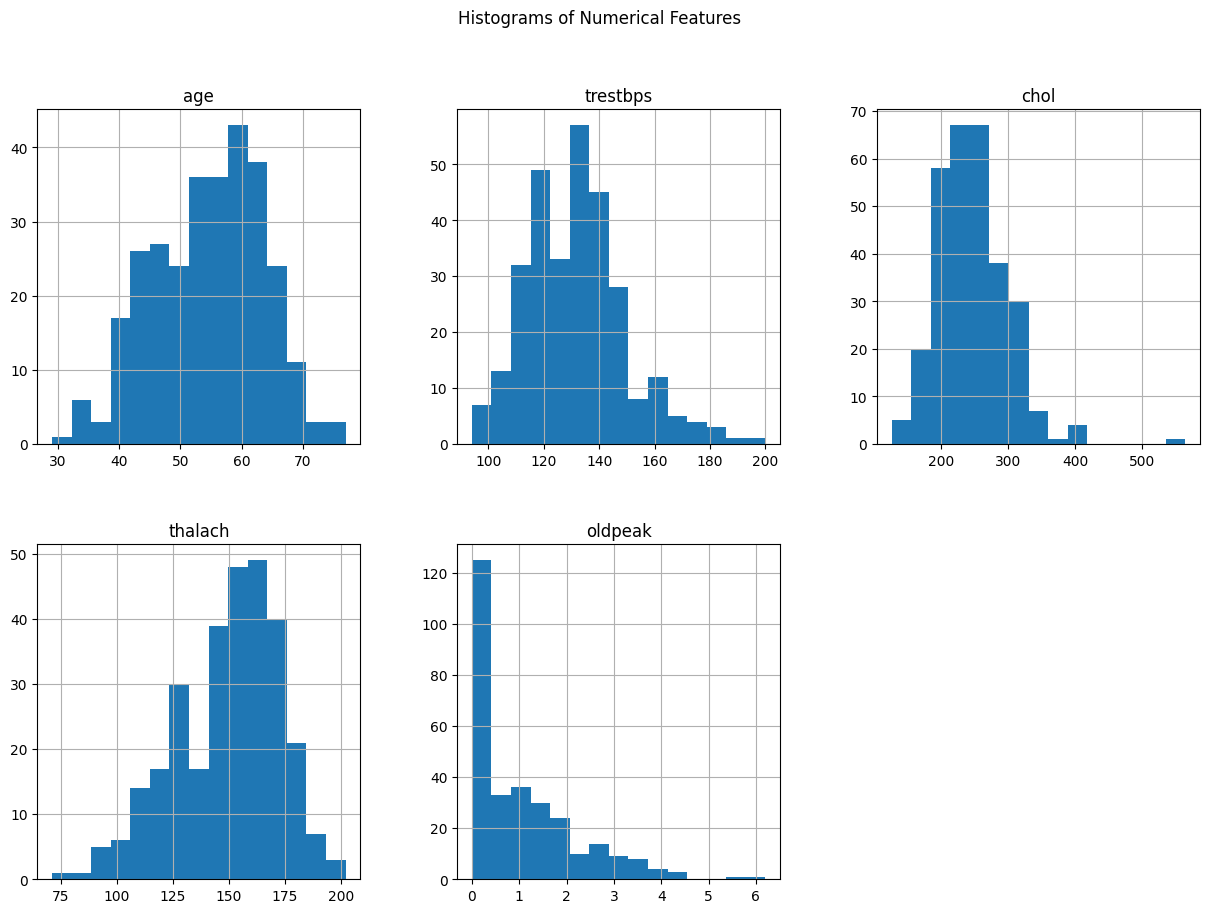
**Week 1**

**Heart Disease Prediction using Patient Data**

**Data Analysis**

****

**Age:**

The distribution of age is somewhat right-skewed, with a higher concentration of patients in the middle age range (50s and 60s). This suggests a prevalence of heart disease screening or incidents among this age group.

**Resting Blood Pressure (trestbps):**

The distribution is approximately normally distributed with a peak around 120 to 140 mm Hg. This indicates that most patients have blood pressure within this range, which is typically considered normal to slightly elevated.

**Cholesterol (chol):**

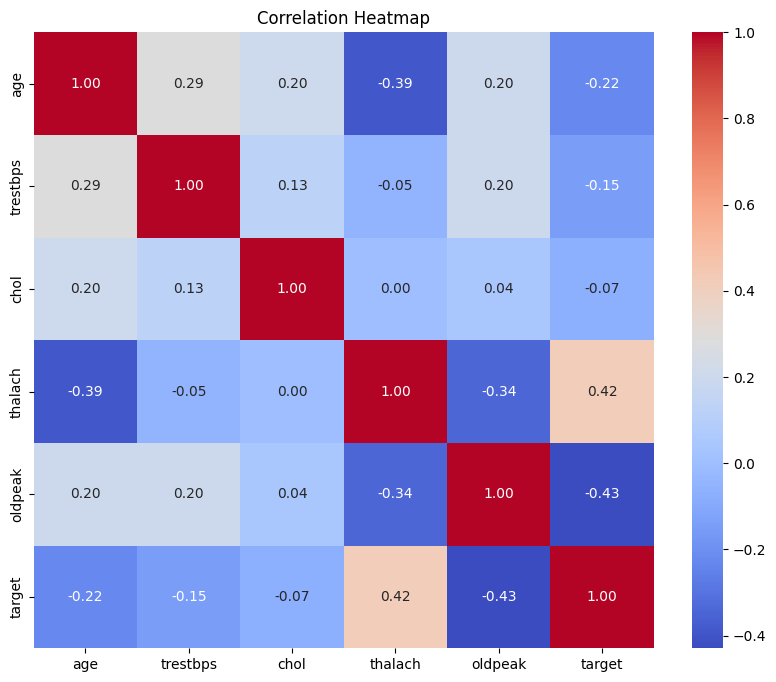
The cholesterol levels are also somewhat right-skewed, with most values concentrated between 200 and 300 mg/dl. High cholesterol levels (above 240 mg/dl) are prevalent, indicating a risk factor for heart disease among the patients.

**Maximum Heart Rate Achieved (thalach):**

The distribution is left-skewed with most individuals achieving a peak heart rate between 150 to 170 beats per minute during stress. The decline in maximum heart rate as it approaches higher values suggests that fewer individuals achieve very high heart rates, which could be related to age or heart condition.

**ST Depression Induced by Exercise Relative to Rest (oldpeak):**

The distribution is heavily skewed towards lower values, with a large number of observations close to 0. Higher values of ST depression are less common but could indicate more severe exercise-induced cardiac stress.

****

**Age:**

Positively correlated with trestbps (resting blood pressure) and chol (cholesterol), suggesting that these parameters tend to increase with age.

Negatively correlated with thalach (maximum heart rate achieved), meaning older individuals tend to have lower peak heart rates.

**Trestbps and Chol:**

Mild positive correlation with each other, suggesting that individuals with higher blood pressure may also have higher cholesterol levels.

**Thalach (Maximum Heart Rate):**

Negatively correlated with age and oldpeak (ST depression), suggesting that younger individuals and those with less exercise-induced ST depression achieve higher maximum heart rates.

Positively correlated with the target (presence of heart disease), indicating that higher maximum heart rates are associated with a greater likelihood of having heart disease in this dataset.

**Oldpeak (ST Depression):**

Positively correlated with age and negatively correlated with thalach, suggesting that older age and lower heart rates are associated with greater exercise-induced ST depression.

Strong negative correlation with the target, indicating that higher values of ST depression are linked with a lower likelihood of having heart disease, which could be indicative of more severe symptoms or advanced disease progression.

**Target (Heart Disease):**

Correlations with age, thalach, and oldpeak are notably significant in understanding how the likelihood of heart disease is associated with these clinical measurements.

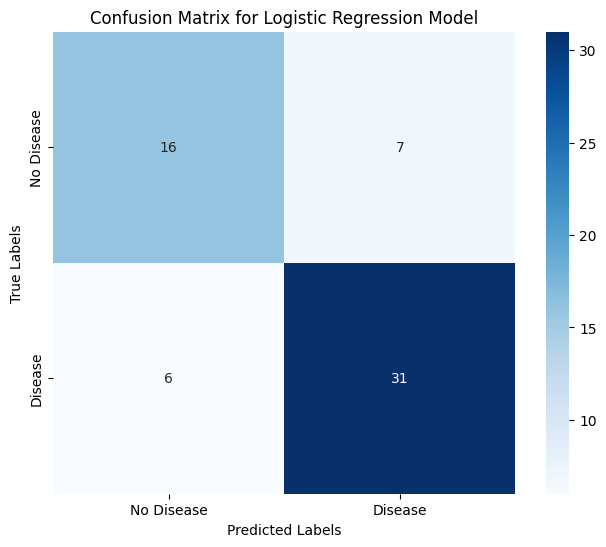
**Model Performance**

1. **Using Logistic Regression:**

**Accuracy:** 0.78

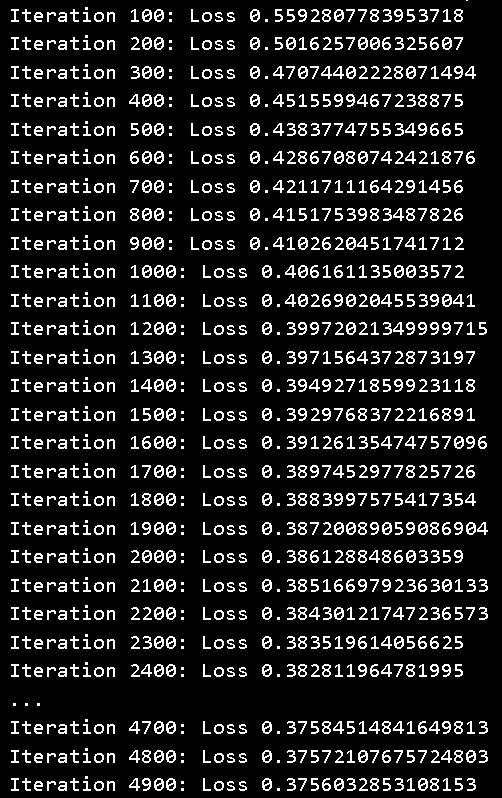
**'C':** 0.1**, 'penalty':** 'l2'**, 'solver':** 'saga'

Using more optimized approach the model can reach a maximum accuracy level of 0.81 using the above mentioned hyper-paramters.

****

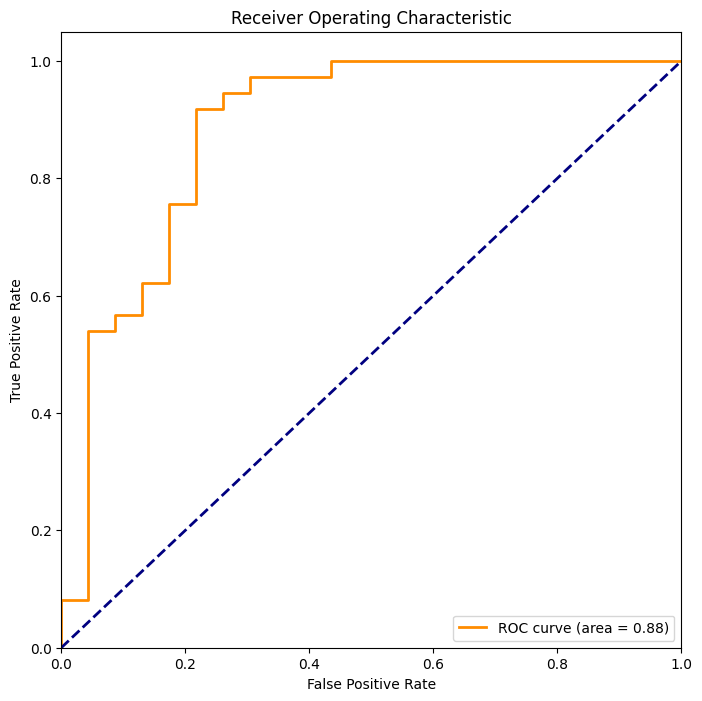
**Gradient Descent:**

Used gradient descent to lower the cost function and came to conclusion that more datasets will help in getting more and more closer to that minimum cost. So in short, model is performing really well.



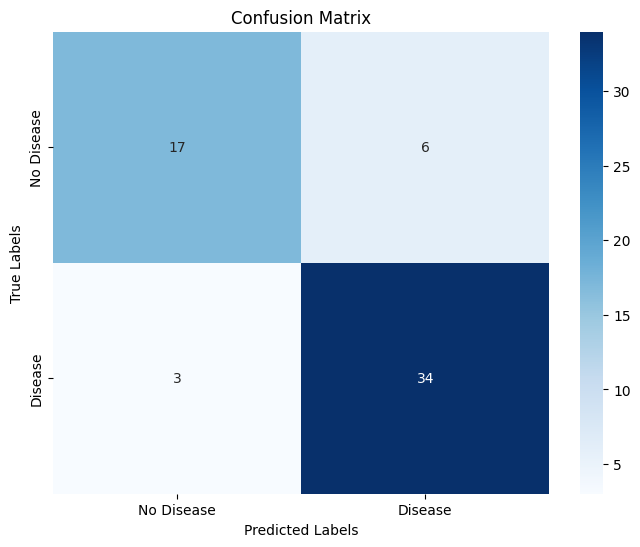
1. **Using Random Forest:**

**Accuracy:** 0.85



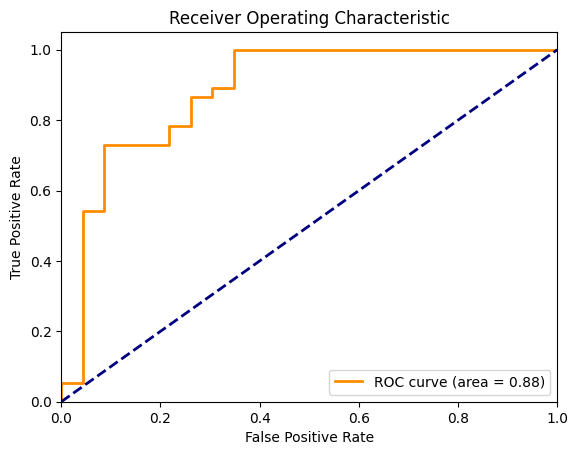
The Receiver Operating Characteristic (ROC) curve shown in the image is a graphical representation used to evaluate the performance of a binary classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. Here’s a breakdown of what you’re seeing in the curve:

1. **True Positive Rate (TPR)**: This is on the y-axis and is also known as sensitivity. It measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).
2. **False Positive Rate (FPR)**: This is on the x-axis and measures the proportion of actual negatives that are incorrectly identified as positives (e.g., the percentage of healthy people who are incorrectly identified as having the condition).
3. **ROC Curve Shape**: The shape of the ROC curve in the image indicates how well the model can distinguish between the two classes. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
4. **Area Under the Curve (AUC)**: The area under the ROC curve is 0.88, as indicated in the legend. This value represents the measure of the model's ability to discriminate between the positive and negative classes. An AUC of 0.88 is considered very good, as it is close to 1, where a perfect classifier would lie.
5. **Diagonal Line**: The dashed diagonal line represents a no-skill classifier that predicts classes randomly. A model with good predictive accuracy will have its ROC curve far above this line.



1. **Using SVM**

**Accuracy:** 0.8



**Best Model Performance:**

Given these metrics, while the AUC values are equal, the **Random Forest model is slightly better** based on its **higher accuracy**. This suggests that for this dataset and the chosen performance metrics, Random Forest might be a better choice as it correctly classifies a higher proportion of instances.