



# Diagnosing Heart Attacks using Machine Learning

Caroline Foshee, Evan Woodard, Jamee Jones, Sivangi Raychoudhury

# What is a Heart Attack?

A **heart attack**, also called a myocardial infarction, happens when a part of the heart muscle doesn't get enough blood.

There are multiple risk factors for heart attack including age, family history, and lifestyle. Also, half of all Americans have at least one of the three risk factors for heart disease: high blood cholesterol, high blood pressure, and smoking.

This dataset contains some medical information of patients which tells whether the chance of that person getting a heart attack is less or more.

This project will make predictions on the probability of individuals experiencing a heart attack, using different Machine Learning models and comparing their outputs to get the best model.



# Variable Definitions

## Numeric Variables:

age - Age of the Patient

trtbps - Resting Blood Pressure

chol - Cholesterol

thalachh - Maximum Heart Rate

oldpeak - ST Depression

## Categorical Variables:

sex - Gender

cp - Chest Pain Type

fbs - Fasting Blood sugar

restecg - Resting Electrocardiographic Results

exng - Exercise Induced Angina

slp - The Slope of ST Segment

caa - Number of Major Vessels

thall - Thalassemia

output - Target



# The Data

- CSV from kaggle.com
- 303 samples age range 29-77
- 14 explanatory variables

|    | age | sex | cp | trtbps | chol | fbs | restecg | thalachh | exng | oldpeak | slp | caa | thall | output |
|----|-----|-----|----|--------|------|-----|---------|----------|------|---------|-----|-----|-------|--------|
| 0  | 63  | 1   | 3  | 145    | 233  | 1   | 0       | 150      | 0    | 2.3     | 0   | 0   | 1     | 1      |
| 1  | 37  | 1   | 2  | 130    | 250  | 0   | 1       | 187      | 0    | 3.5     | 0   | 0   | 2     | 1      |
| 2  | 41  | 0   | 1  | 130    | 204  | 0   | 0       | 172      | 0    | 1.4     | 2   | 0   | 2     | 1      |
| 3  | 56  | 1   | 1  | 120    | 236  | 0   | 1       | 178      | 0    | 0.8     | 2   | 0   | 2     | 1      |
| 4  | 57  | 0   | 0  | 120    | 354  | 0   | 1       | 163      | 1    | 0.6     | 2   | 0   | 2     | 1      |
| 5  | 57  | 1   | 0  | 140    | 192  | 0   | 1       | 148      | 0    | 0.4     | 1   | 0   | 1     | 1      |
| 6  | 56  | 0   | 1  | 140    | 294  | 0   | 0       | 153      | 0    | 1.3     | 1   | 0   | 2     | 1      |
| 7  | 44  | 1   | 1  | 120    | 263  | 0   | 1       | 173      | 0    | 0.0     | 2   | 0   | 3     | 1      |
| 8  | 52  | 1   | 2  | 172    | 199  | 1   | 1       | 162      | 0    | 0.5     | 2   | 0   | 3     | 1      |
| 9  | 57  | 1   | 2  | 150    | 168  | 0   | 1       | 174      | 0    | 1.6     | 2   | 0   | 2     | 1      |
| 10 | 54  | 1   | 0  | 140    | 239  | 0   | 1       | 160      | 0    | 1.2     | 2   | 0   | 2     | 1      |



# EXPLORATORY DATA ANALYSIS

Numerical

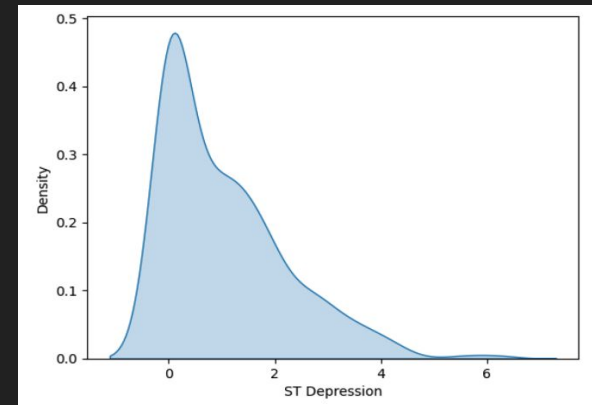
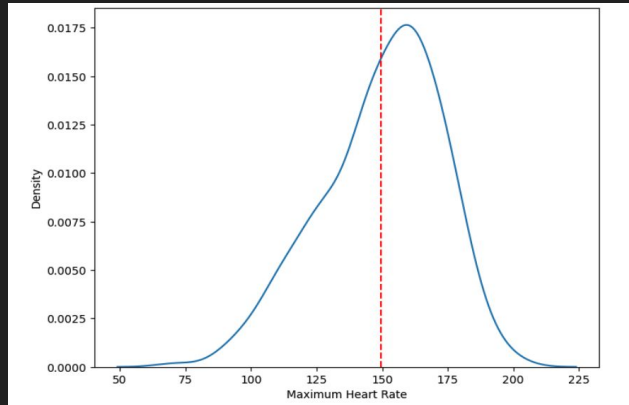
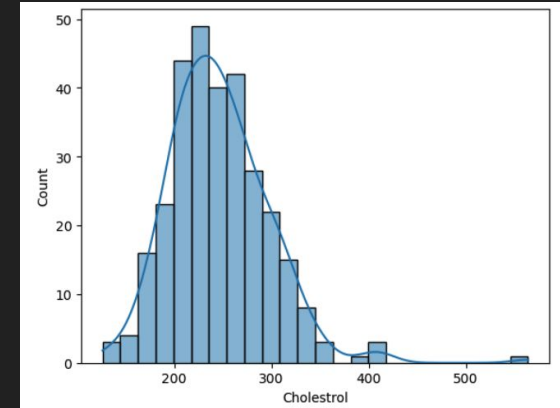
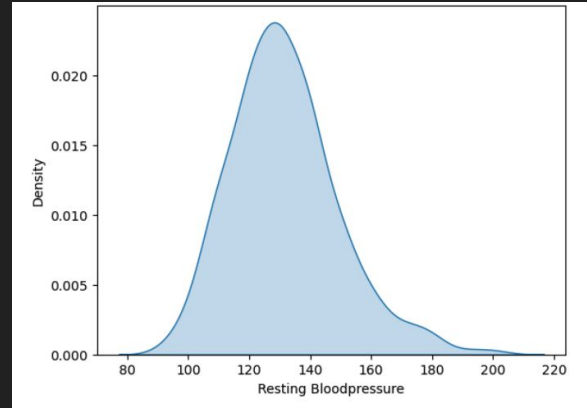
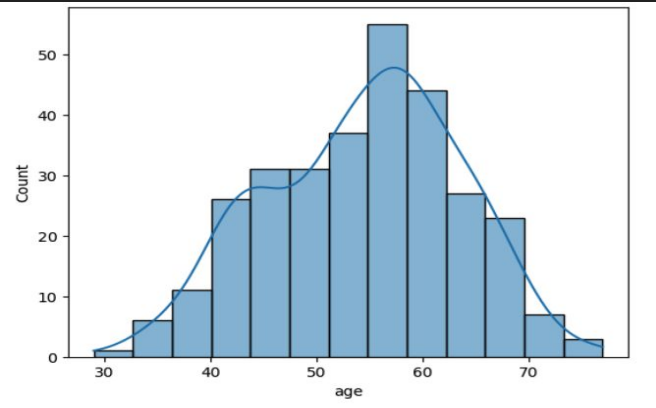
Age  
Trtbps  
Chol  
Thalachh  
Oldpeak

Categorical

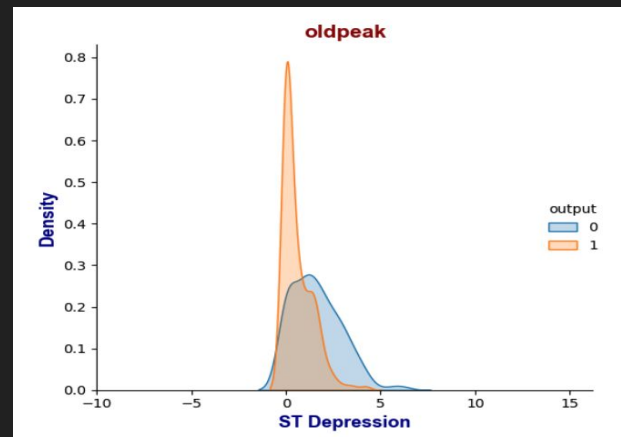
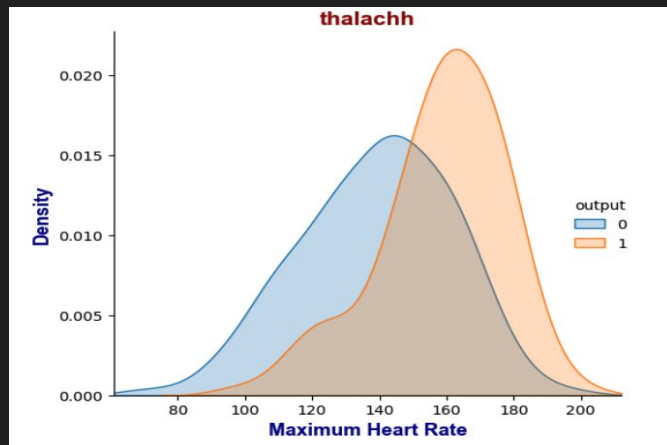
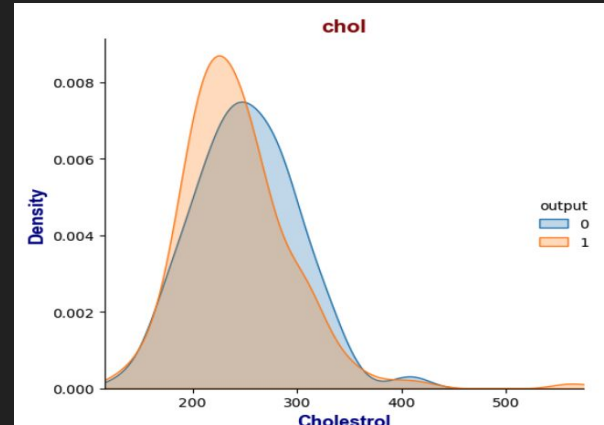
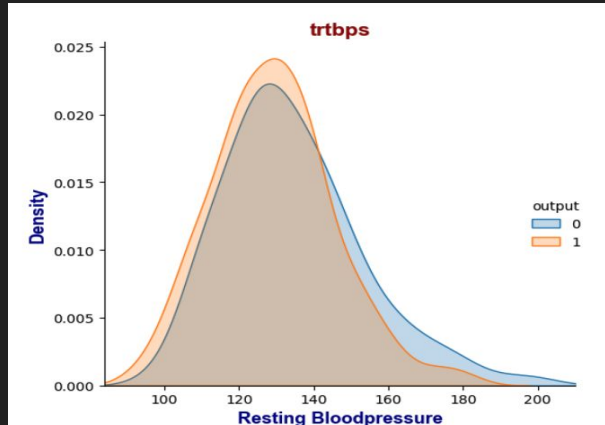
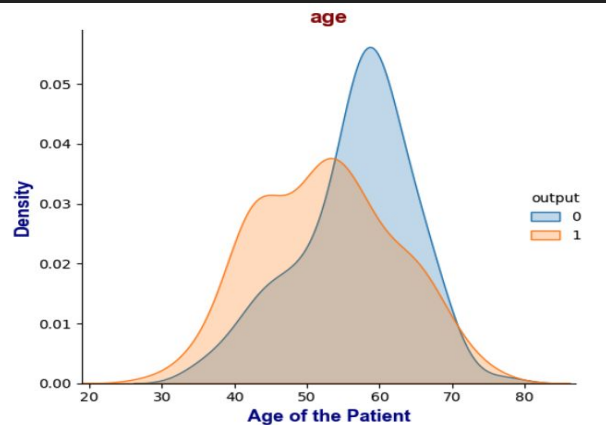
Gender  
Cp  
Fbs  
Restecg  
Exng  
Slp  
Caa  
Thall  
Output



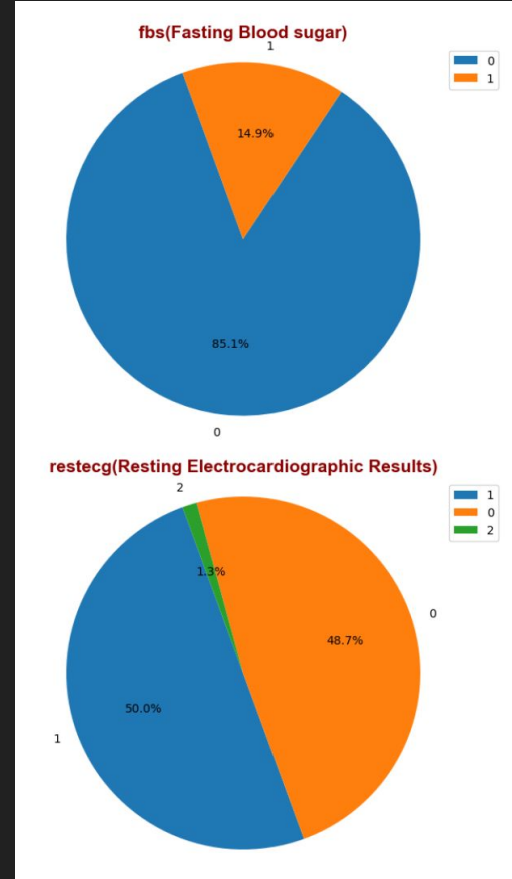
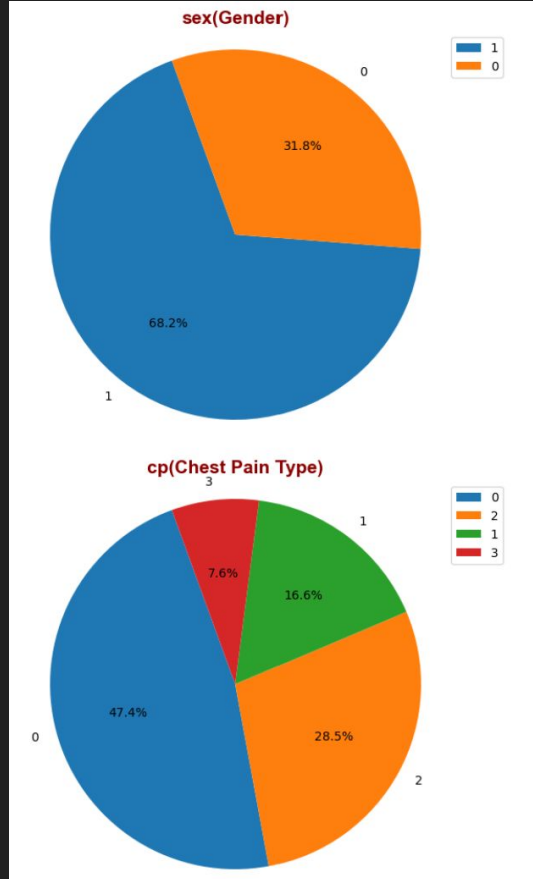
# UNIVARIATE ANALYSIS OF NUMERIC VARIABLES



# BIVARIATE ANALYSIS OF NUMERIC - OUTPUT

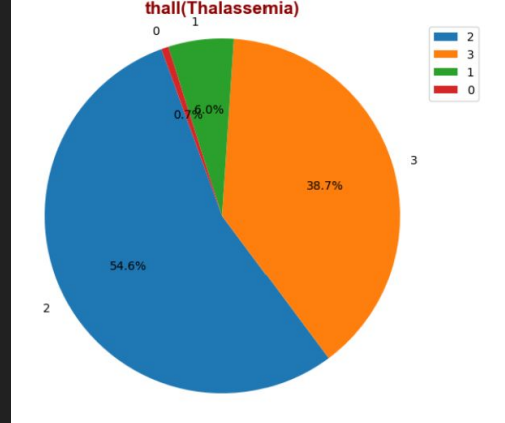
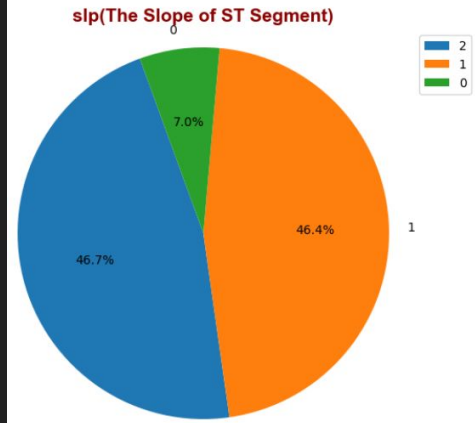
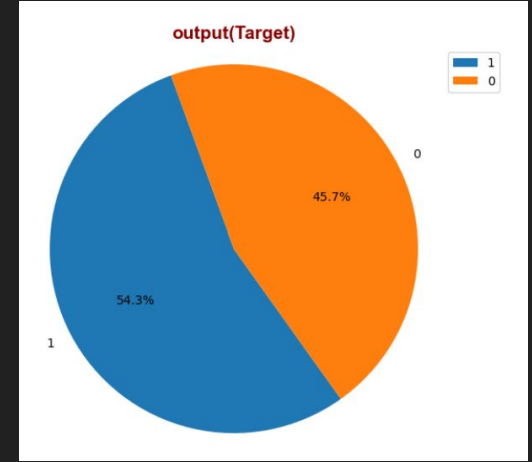
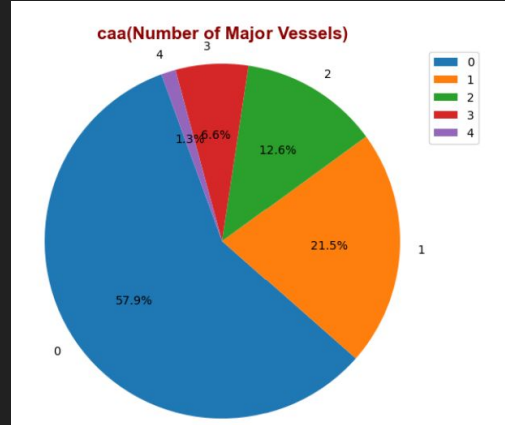
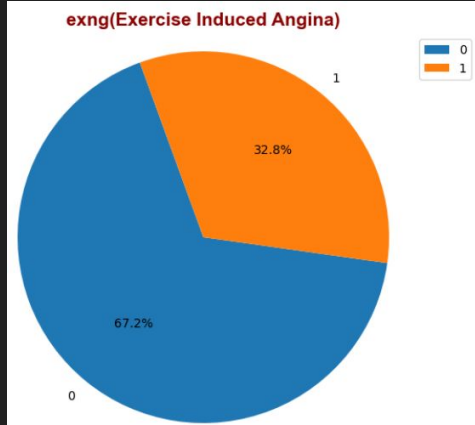


# UNIVARIATE ANALYSIS OF CATEGORIC VARIABLE

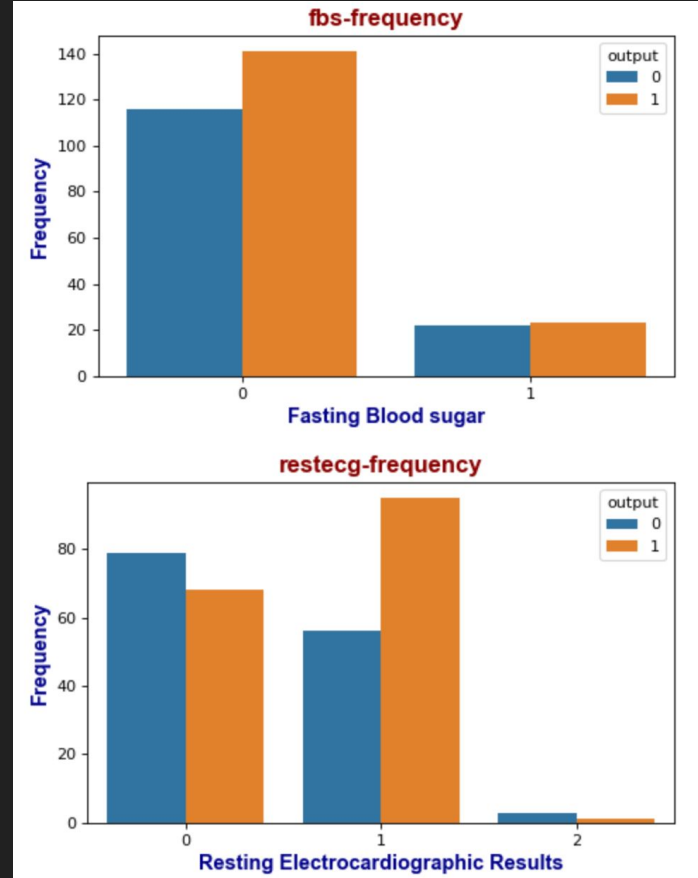
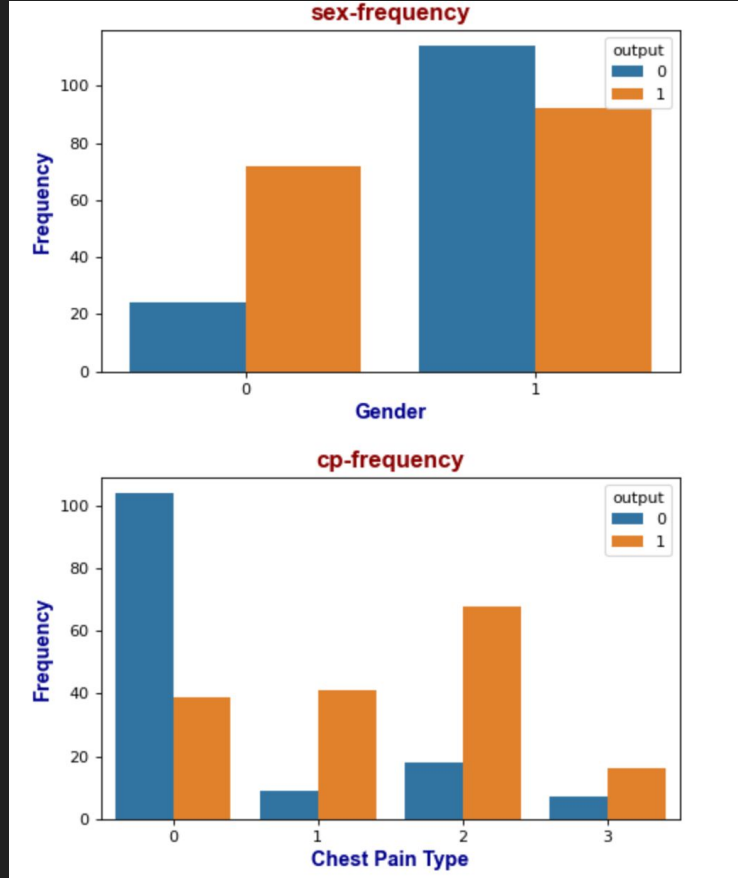




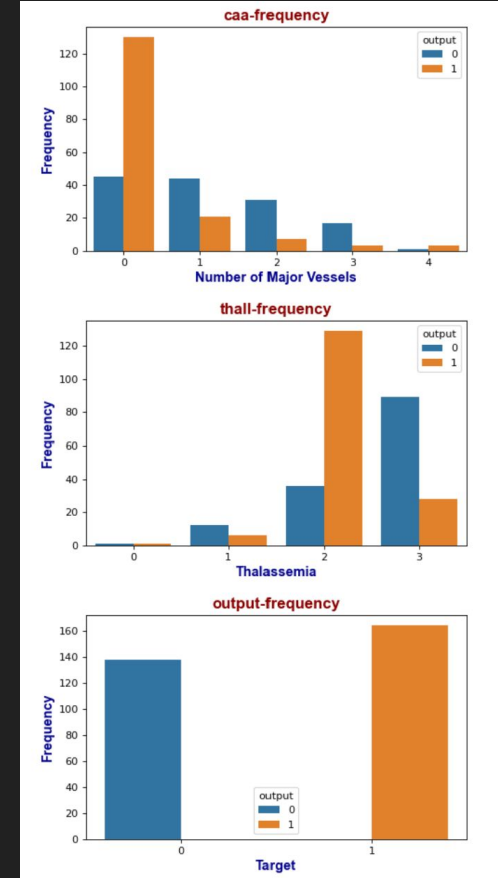
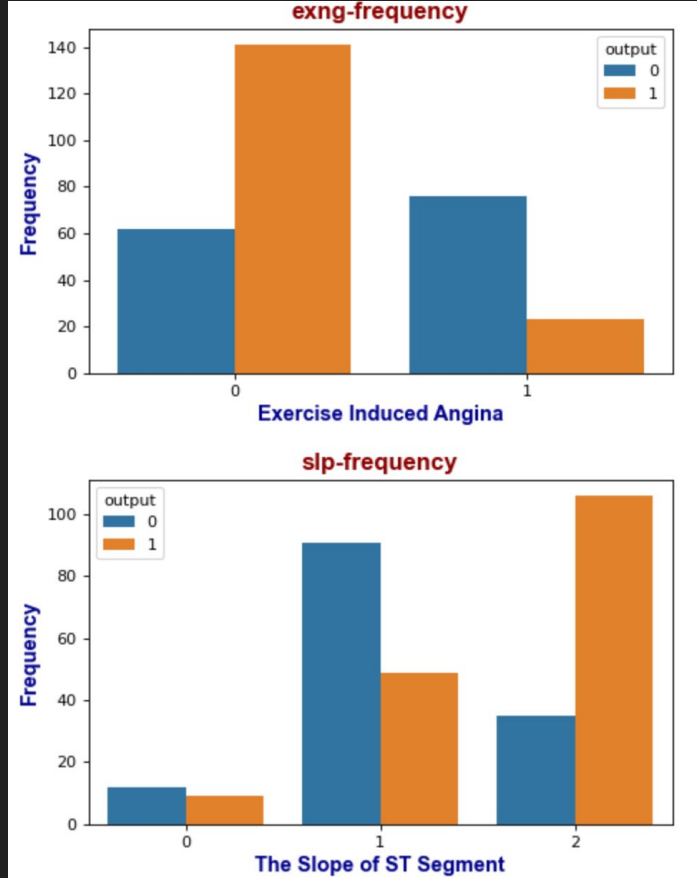
# UNIVARIATE ANALYSIS OF CATEGORIC VARIABLE



# BIVARIATE ANALYSIS OF CATEGORIC - OUTPUT



# BIVARIATE ANALYSIS OF CATEGORIC - OUTPUT



# Tested Machine Learning Models

- Logistic Regression
  - estimates probability of an event occurring
- Decision Tree
  - all possible outcomes based on present conditions
- Random Forest
  - multiple decision trees to reach a certain result
- Neural Network
  - high-powered, gets very accurate over time with large sample set



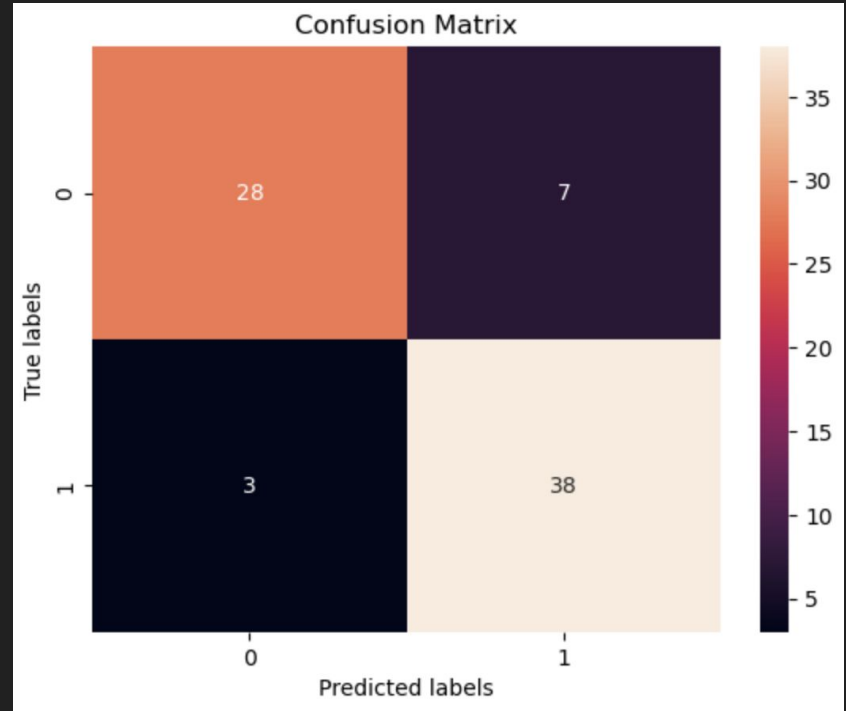
# Accuracy and Model Decision

| <u>Model</u>        | <u>Accuracy</u> |
|---------------------|-----------------|
| Logistic Regression | 87%             |
| Decision Tree       | 83%             |
| Random Forest       | 87%             |
| Neural Network      | 80%             |



# Why Logistic Regression?

- High accuracy
- Best fit for our data and what we set out to accomplish
- We have only 10 misclassified observations out of 76 in total
- Logistic regression models are easier to interpret
- Less prone to overfitting

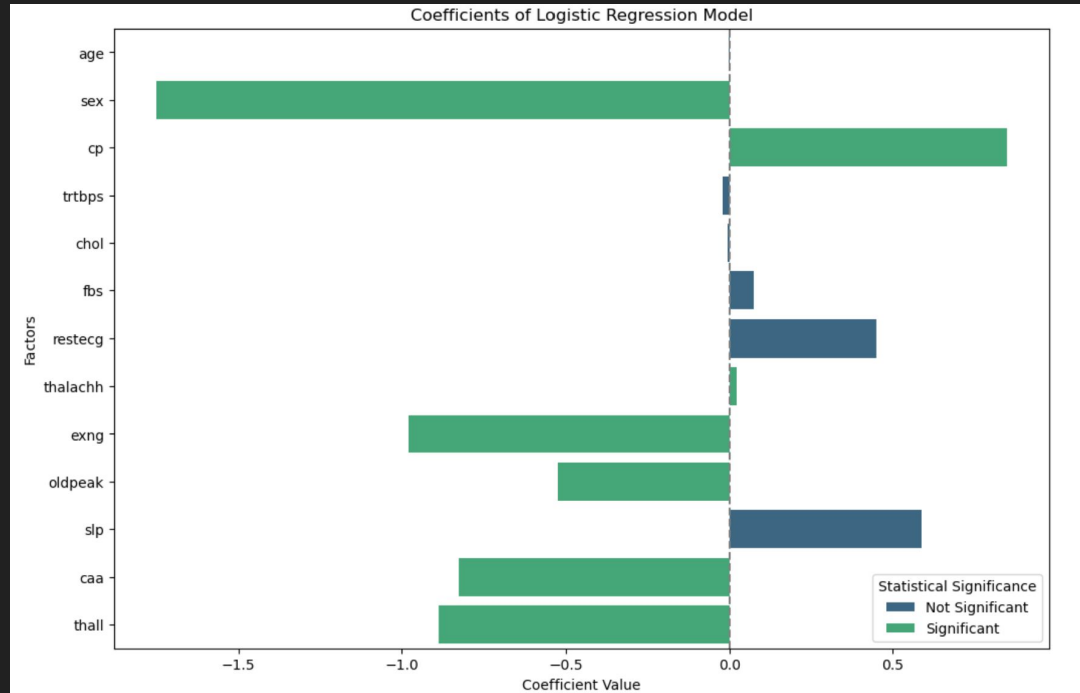


# Finding the Best LR

| Model                               | Accuracy | Confusion matrix errors |
|-------------------------------------|----------|-------------------------|
| #1                                  | 87%      | 10                      |
| #2: Increased number of iterations  | 86%      | 11                      |
| #3: Scaled data                     | 86%      | 11                      |
| #4: Feature selection + scaled data | 86%      | 11                      |



# Variable Significance





## Significant

- Sex
- Type of chest pain
- Incidence of exercise induced angina
- Average ST depression
- Number of major vessels
- Colored by fluoroscopy
- Thalassemia score

## Not Significant

- Age
- Resting blood pressure
- Cholesterol levels
- Fasting blood sugar
- Electrocardiogram results
- Slope of the peak exercise ST segment



|    | Factors  | Coefficients | P-Values | Significance    |
|----|----------|--------------|----------|-----------------|
| 0  | age      | -0.001469    | 0.950062 | Not Significant |
| 1  | sex      | -1.750930    | 0.000184 | Significant     |
| 2  | cp       | 0.847283     | 0.000005 | Significant     |
| 3  | trtbps   | -0.020188    | 0.051916 | Not Significant |
| 4  | chol     | -0.004489    | 0.238252 | Not Significant |
| 5  | fbs      | 0.073463     | 0.890263 | Not Significant |
| 6  | restecg  | 0.450607     | 0.196022 | Not Significant |
| 7  | thalachh | 0.023134     | 0.026835 | Significant     |
| 8  | exng     | -0.981017    | 0.016672 | Significant     |
| 9  | oldpeak  | -0.523604    | 0.014630 | Significant     |
| 10 | slp      | 0.589074     | 0.092236 | Not Significant |
| 11 | caa      | -0.826015    | 0.000043 | Significant     |
| 12 | thall    | -0.887203    | 0.002276 | Significant     |



**DISCLAIMER:**  
I am **NOT** a  
**medical doctor!**