Random Forest Model to predict if a person is prone to a heart attack or not

Author: Yuhao Sui

Professor: Feng Mai

Stevens Institute of Technology

BIA-652 Multivariate Data Analysis

Contents

Introduction

A heart attack, also called a myocardial infarction, happens when a part of the heart muscle doesn't get enough blood. In the United States, someone has a heart attack every 40 seconds. Every year, about 805,000 Americans have a 'heart attack'. Of these, 605,000 are a first 'heart attack', 200,000 happen to people who have already had a heart attack. About 1 in 5 heart attacks is silent—the damage is done, but the person is not aware of it.

Research question

'Your eating habits affect the health of your heart and blood vessels in a number of ways.' Although food is very important to our body. However, I also want to know what other factors can affect our heart health. The data I used in this study is recorded 303 cases with 13 characteristics of the human body. I hope to find out which of these characteristics can affect our heart health through this research. So, through the characteristics (as 'age', 'sex', 'Resting blood pressure', etc.) of a person, we can judge the possibility of a 'heart attack'.

Hypotheses

Here, I set 'output' to 'response variable' because I would like to find which variables can affect the 'output'. In our traditional concept, the health of the elderly will decline with 'age'. At the beginning, I hope to confirm through experiments that 'age' determines whether heart disease occurs.

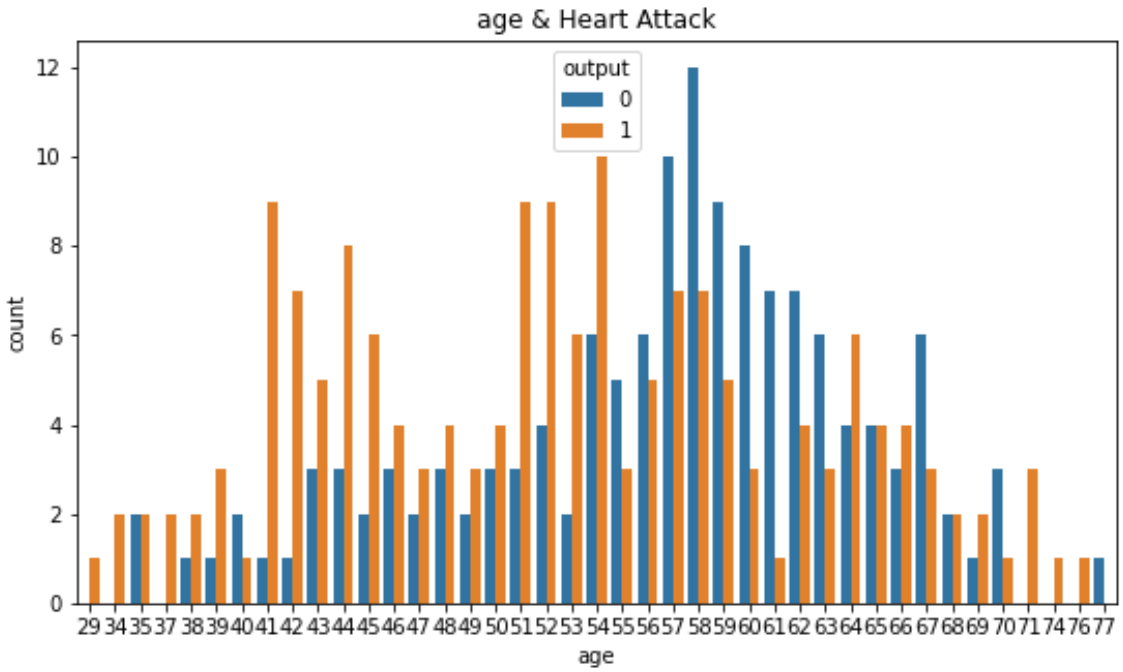


Figure 1. age & Heart Attack

But when I tried to check the distribution (as Figure 1) of 'age' with respect to heart attack, I found that the situation is contrary to my hypotheses. The highest count of '1' (more chance of heart attack) is more concentrated in the front part of the middle (between '40' and '55' years old).

Not only that, the 'heatmap' also can shows 'age' is not significant association with 'output'. So, I assume people who have exercise-induced angina will have more chance of heart attack. And next step I will try to use some other characteristics to compare.

Data collection

The analyzed dataset is provided by Rashik Rahman on Kaggle, and the data source is from 'UCI Machine Learning Repository'. This data has 303 rows and 14 columns.

These are the detail of column:

Variable Name	Data Type	Description
age	Int	Age of the patient.
sex	Int	Sex of the patient. 1 = male; 0 = female.
cp	Int	Chest pain type. 0 = Typical Angina; 1 = Atypical Angina; 2 = Non-anginal Pain; 3 = Asymptomatic.
trtbps	Int	Resting blood pressure (in mm Hg on admission to the hospital).
chol	Int	Cholesterol in mg/dl fetched via BMI sensor.
fbs	Int	Fasting blood sugar > 120 mg/dl. 1 = True; 0 = False.
restecg	Int	Resting electrocardiographic results. 0 = Normal; 1 = having ST-T wave normality; 2 = Left ventricular hypertrophy.
thalachh	Int	Maximum heart rate achieved.
exng	Int	Exercise induced angina. 1 = Yes; 0 = No.
oldpeak	Float	Previous peak, ST depression induced by exercise relative to rest.
slope	Int	Slope, the slope of the peak exercise ST segment. 1 = up sloping; 2 = flat; 3 = down sloping.
caa	Int	Number of major vessels. Colored by fluoroscopy.
thall	Int	Maximum heart rate achieved 2 = normal; 1 = fixed defect; 3 = reversable defect.
output	Int	Target variable. The predicted attribute – diagnosis of heart disease (angiographic disease status) 0= less chance of heart attack; 1= more chance of heart attack.

Table 1: I from heart attack dataset

According to the table 1, I didn't find any null data, and all type of data are Numerical variables. Therefore, this dataset is perfect and does not require me to do extra processing.

Data preparation

About Y

Y means the target variable 'output' of the project. It represents the diagnosis of heart disease (angiographic disease status) of 303 test patients, '0' represents the low probability of attack, and '1' represents the high probability of attack. Next, I will use linear analysis and EDA methods to predict Y to see what kind of people are most chance to suffer from heart attack.

About X

X means predictor variable. Because currently considering that we have 13 predictor variables, if we use all the variables to compare and analyze with Y, is very large work. I decided to use the heat map to find the predictor variables that have a large correlation with Y. I will consider comparing other predictors in the future. But in order not to affect the accuracy of the model, after removing the 'output', I set X to 13 other predictor variables.

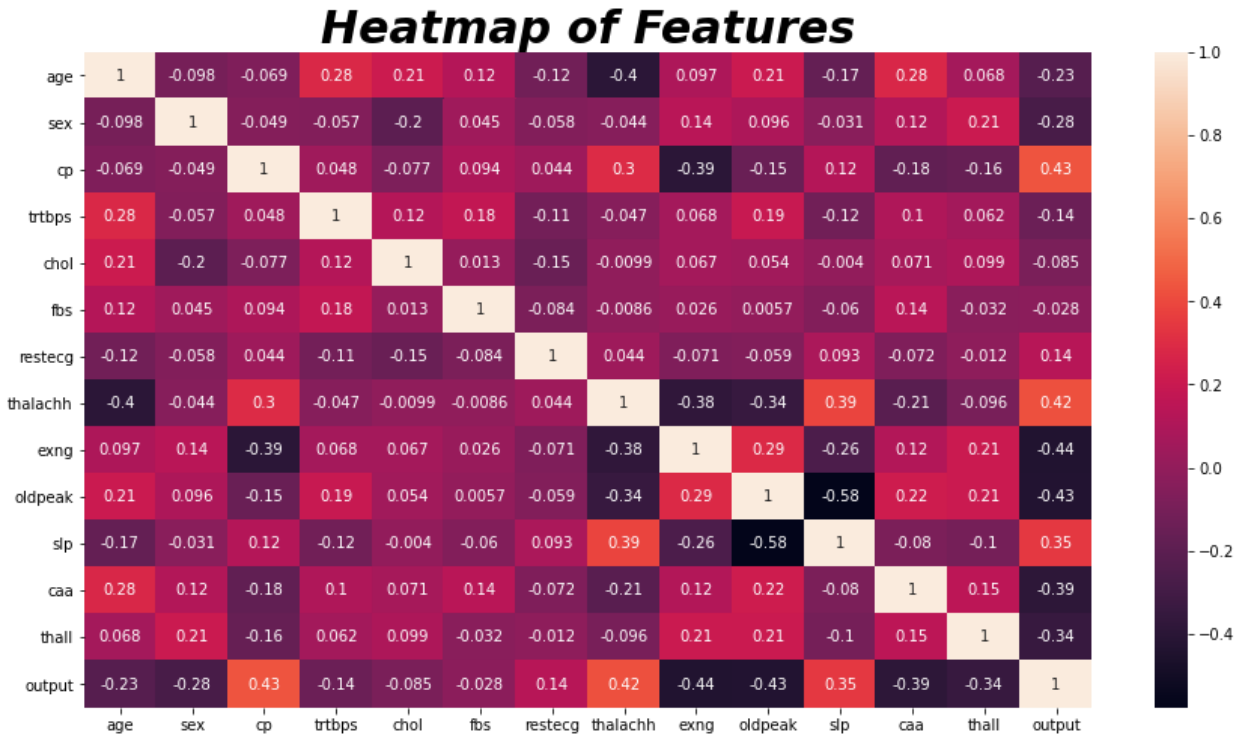


Figure 2. Heatmap of Features

As shown in picture 2, it tells you some highest (close +1 and -0.4) correlated columns. like 'cp' (Chest pain type) and 'output', 'thalachh' (Maximum heart rate achieved) and 'output' and 'exng' (Exercise induced angina) and 'output', their correlations are: 0.43, 0.42 and -0.44.

However, since the numerical values of ‘cp’ and ‘exng’ are ‘ordinal feature’, they are converted from non-numeric character strings, which are inaccurate when compared with other numerical values. Therefore, I will model and analyze several other X variables and Y variables.

EDA

Before modeling, first compare some X variables and Y variables in the data, and try to find some rules from them.

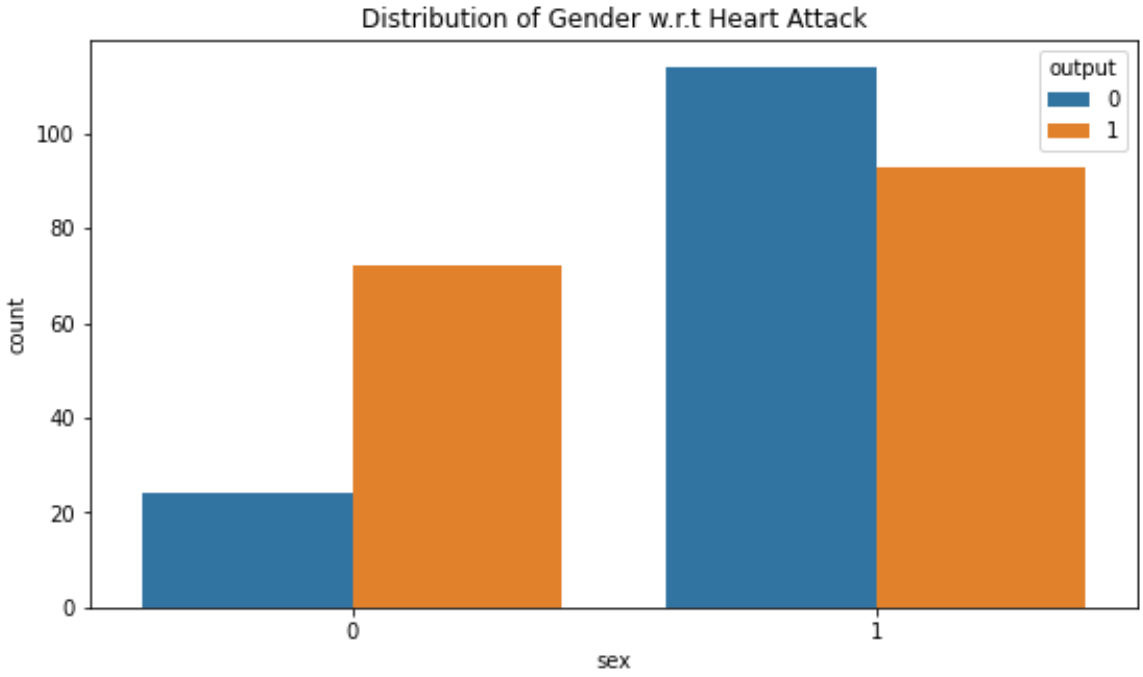


Figure 4

The Figure 4 shows the effect of gender on heart attack. Blue is low probability of heart attack, orange is high probability of heart attack; 0 means female, 1 means male; X-axis is sex, and Y-axis is the number of patients.

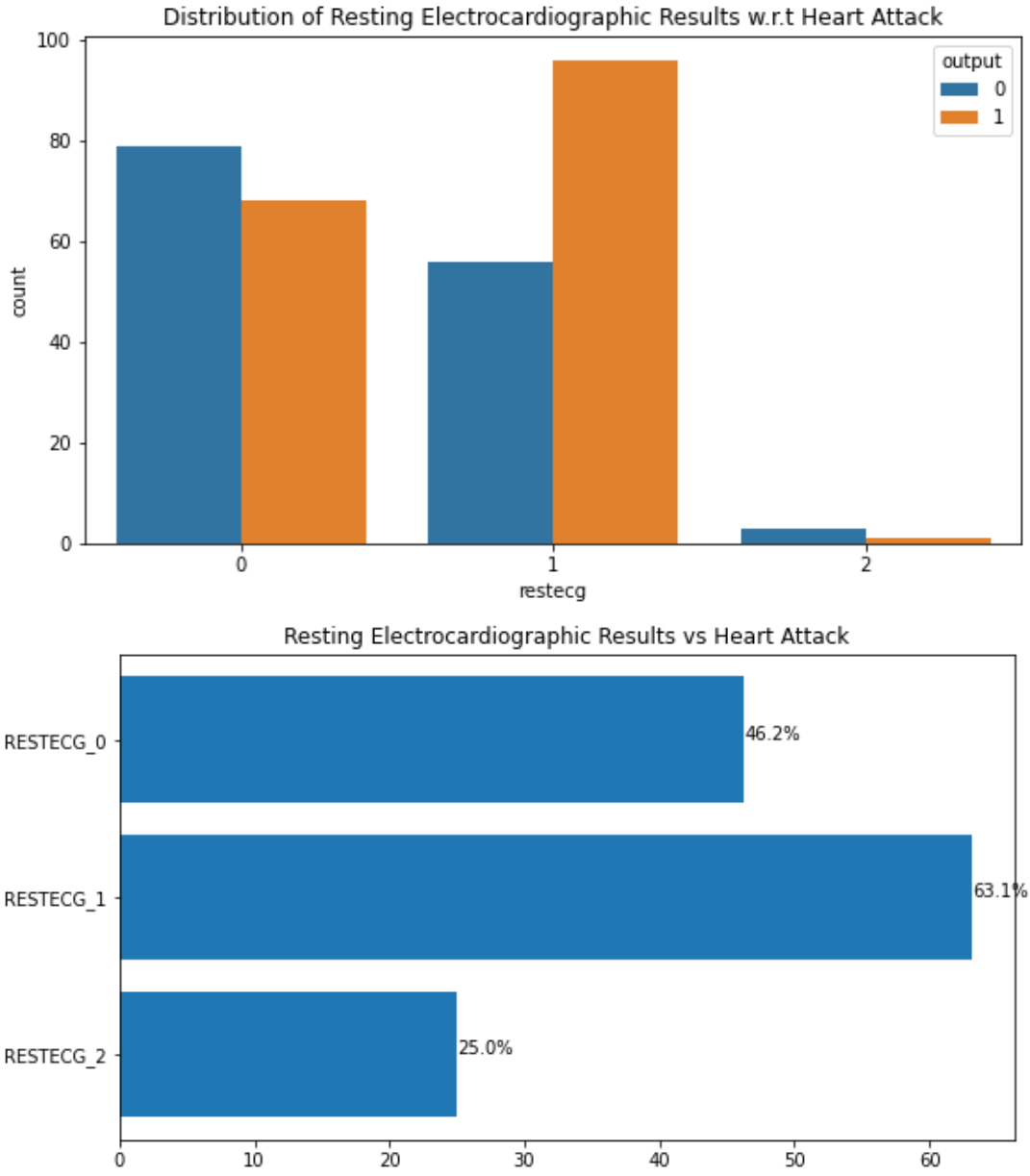


Figure 5

Figure 5 shows the effect of resting electrocardiographic results on heart attack. The above is the description of resting electrocardiographic results. Blue is low probability, orange is high probability; the three values on the X axis represent: 0 = Normal; 1 = having ST-T wave normality; 2 = Left ventricular hypertrophy, and the Y axis is the number of patients. The figure below shows the proportion of resting electrocardiographic results with a high probability of heart attack. The X-axis is the proportion of high probability of heart attack, and the Y-axis is the three results (for the convenience of observation). It can be found that patients having ST-T wave normality are more prone to heart attacks.

Regression Analysis

Modeling

Before modeling, we must first determine what kind of model fits the data. First, I will compare three models of logistic regression, random forest classifier, extreme Gradient Boosting classifier (XGBoosst classifier), K neighbors classifier, Support vector machines, Decision tree classifier to test their accuracy for heart attack data.

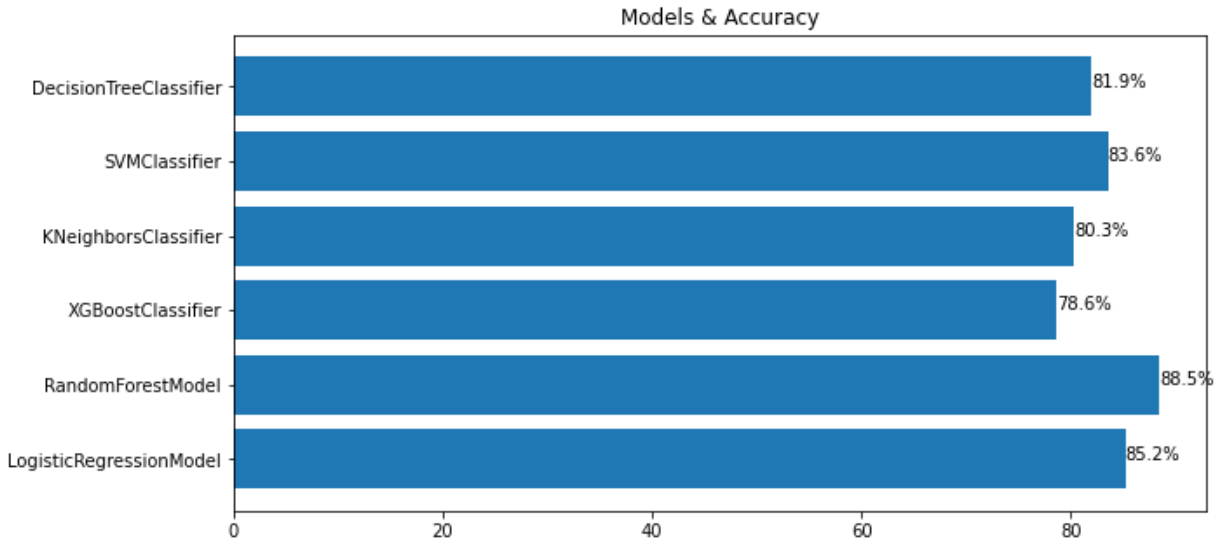
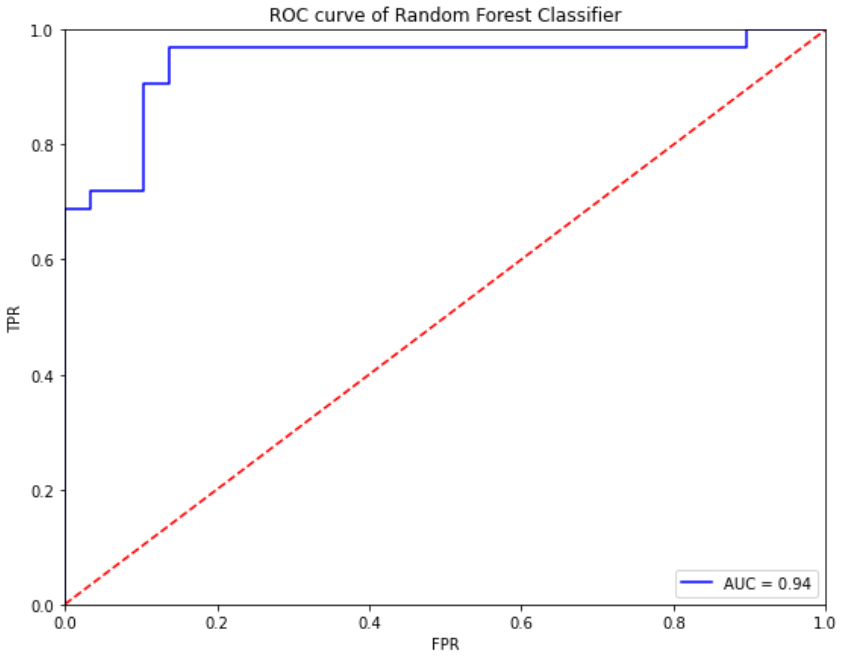


Figure 6. Models & Accuracy

As shown in Figure 6, the random forest model is the model with the highest accuracy. Therefore, I will build a random forest model and optimize it to make it more accurate.

Random forest model

Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.




```
Accuracy: 90.1639344262295
Precision: 0.8611111111111112
AUC for our classifier is: 0.9428879310344828
```

Figure 7. ROC curve of Random Forest Classifier

As you can see from Figure 7, by increasing the maximum number of features and adjusting the max feature to 10, I finally got the highest area under the curve (AUC), which is about 0.94.

TPR in the figure 7 means True Positive Rate, and FPR means False Positive Rate. The Receiver Operating Characteristic (ROC) curve shows the trade-off between sensitivity (TPR) and specificity (FPR). As figure 7 shown, classifiers that give curves closer to the top-left corner indicate better performance.

the AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Its calculation formula is as follows:

$$AUC(f) = \frac{\sum_{t_0 \in D^0} \sum_{t_1 \in D^1} 1[f(t_0) < f(t_1)]}{|D^0| \cdot |D^1|}$$

$1[f(t_0) < f(t_1)]$ denotes an indicator function which returns 1, $f(t_0) < f(t_1)$ otherwise return 0; D0 is the total number of negative examples, and D1 is the total number of positive examples. The denominator is the total number of combinations of positive and negative samples, and the numerator is the number of combinations of positive samples greater than negative samples.

In the process of optimizing the random forest model, I found that this model has some advantages and disadvantages. I will introduce them in detail later.

Limitations

No model is perfect. While enjoying the convenience of random forest, we must also learn to recognize its shortcomings:

a. It is more complicated than the decision tree algorithm and has a higher computational cost. After all, random forest is composed of many decision trees, and it is not difficult to understand that the algorithm is more complicated.
b. Due to their complexity, they require more time to train than other similar algorithms. Fortunately, the heart attack data is very small, with only 303 samples, but even so, the calculation process took nearly 3 minutes. Imagine if the sample size is large, this will be a huge burden, so be accurate Find a balance between performance and calculation speed.

But no algorithm is perfect, we must learn to balance their advantages and disadvantages.

Discussion

When building a model for this data, using random forest can get a very high accuracy. However, this model is very complicated in actual operation and takes a long time. Therefore, I will try to use linear regression and SVM classifier models in the future. The linear regression model is characterized by simple operation. Moreover, in the comparison in Figure 6, we can see that although the accuracy of linear regression is slightly lower, the convenient operation makes it more competitive.

The SVM classifier can handle the interaction of nonlinear features and can solve the problem of large feature spaces.

Conclusion

For heart attacks, I used EDA to compare various variables and found that the probability of a heart attack in women is about 75%, while that of men is only 45%. Therefore, women are more likely to suffer from heart attacks; People with exercise-induced angina had a heart attack rate of 23.2%, while those without exercise-induced angina had a heart attack rate of 69.6%. Therefore, people with exercise-induced angina are less likely to develop heart disease; people with type 1 chest pain (atypical angina) are more likely to have heart attacks; People with

RESPECT 1 are more likely to have heart attacks, followed by REST ECG 0 and RESTECG 2. People with less age has higher chances of Heart Attack; People with age between 50-70 years are more prone to Heart Attack; Out of all the experimented classification models Random Forest Classifier performed better than other models with an AUC of 0.942 and accuracy of 90.16% and precision of 0.86.

Reference

1. CDC, 'Heart Attack Symptoms, Risk, and Recovery', January 11, 2021, from https://www.cdc.gov/heartdisease/heart_attack.htm
2. Healthgrades, 'Types of Diets for Heart Attack Prevention', April 26, 2021, from https://www.healthgrades.com/right-care/heart-attack/types-of-diets-for-heart-attack-prevention
3. CDC, 'Heart Disease Facts', September 8, 2020, from https://www.cdc.gov/heartdisease/facts.htm
4. Kaggle, 'Heart Attack Analysis & Prediction Dataset', March 22, 2021, from https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
5. Wikipedia, 'Random Forest', May 6, 2021, from https://en.wikipedia.org/wiki/Random_forest