Heart Attack Prediction

Anastasia Popova

Outline

Outline

Introduction

motivation

Cardiovascular diseases are a leading cause of mortality worldwide. The best practice is preventing these diseases, which requires a personalized monitoring approach.

By integrating ML algorithms with medical data, healthcare professionals can detect anomalies and assess risk factors long before symptoms are detected.

goal

In the current project, I aim to *explore the dataset* and *build a model to predict a heart-related condition* based on the provided clinical and demographic factors. I also try to use fewer factors to reach the same accuracy, which can be valuable for the price reduction of the early diagnosis.

dataset

The Heart Attack Dataset (License CC0: Public Domain) offers a comprehensive collection of medical attributes designed to aid in the prediction of heart-related conditions.

It consists of various clinical and demographic factors that play a significant role in assessing a patient's cardiovascular health.

For example, it includes **age, sex, resting blood pressure, electrocardiographic results, cholesterol, and fasting blood sugar levels.**

Outline

Exploratory Data Analysis

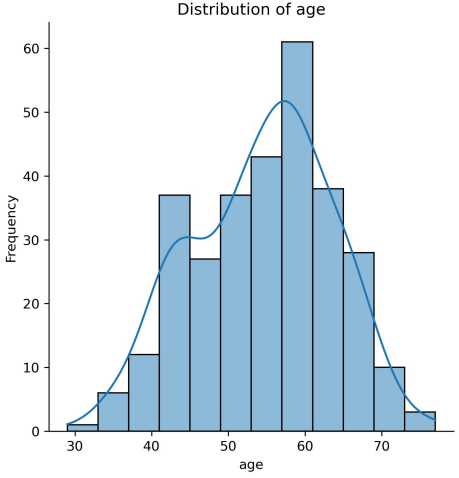
The dataset has 303 observations and 13 features; 5 features are numerical, and 8 features are categorical.

No missing values are present in the dataset, but outliers exist.

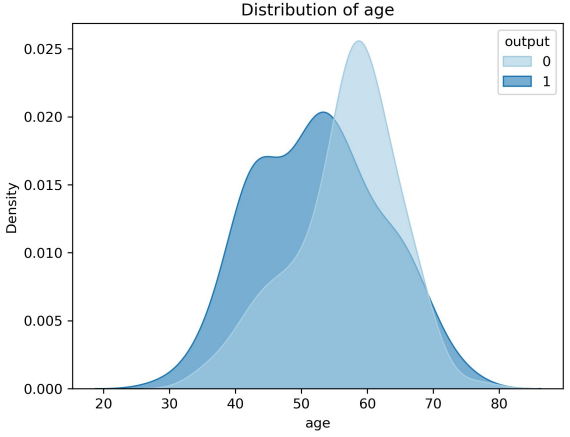
The questions I was interested in are:

- *If the age distribution is normal?*

Yes, the age distribution in the dataset closely resembles a normal distribution, ranging from 29 to 77 years.



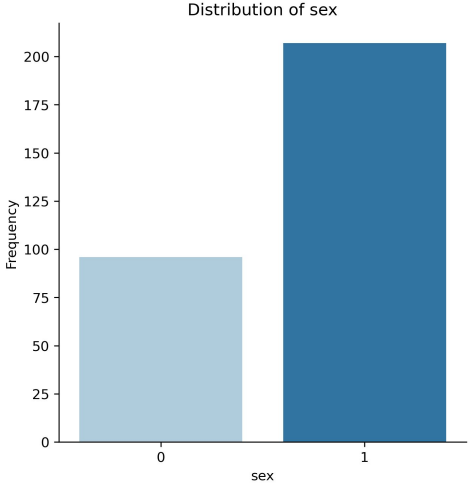
- *According to this dataset, do older patients have a higher chance of heart attack?*

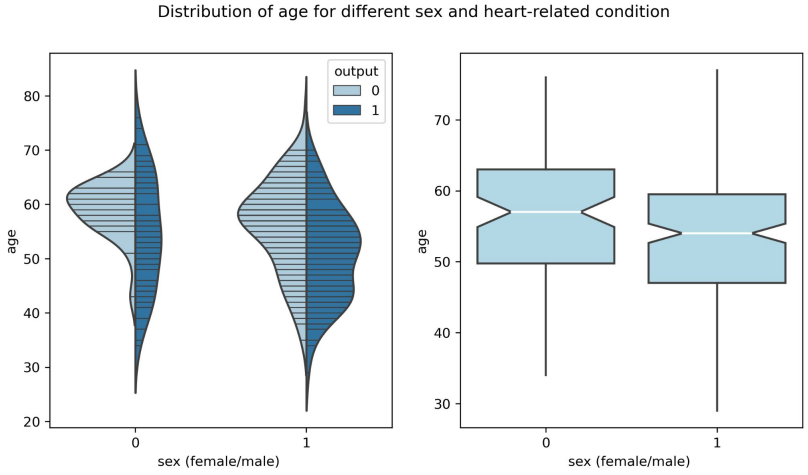


The answer is **no**: Younger patients have higher chance of heart attack

Exploratory Data Analysis

- *Is there a significant gender imbalance in the dataset, and how does the age distribution differ between male and female patients?*
- *What can we infer about heart attack risk based on gender?*





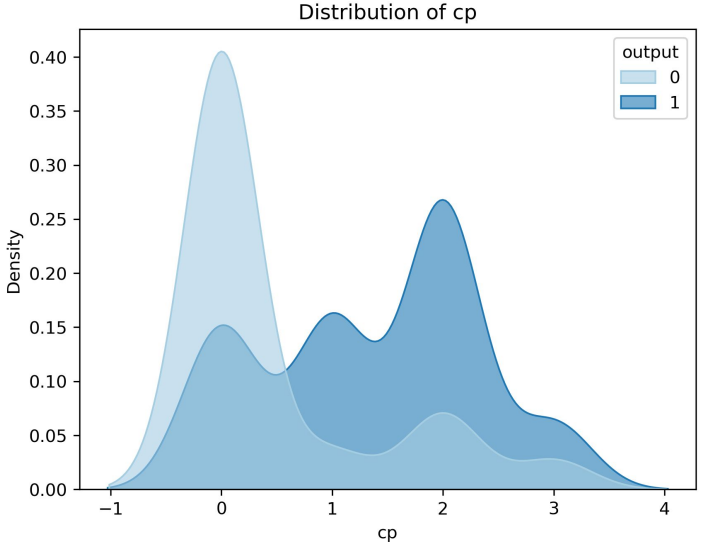
There are **twice** as many male patients as female patients in the dataset.

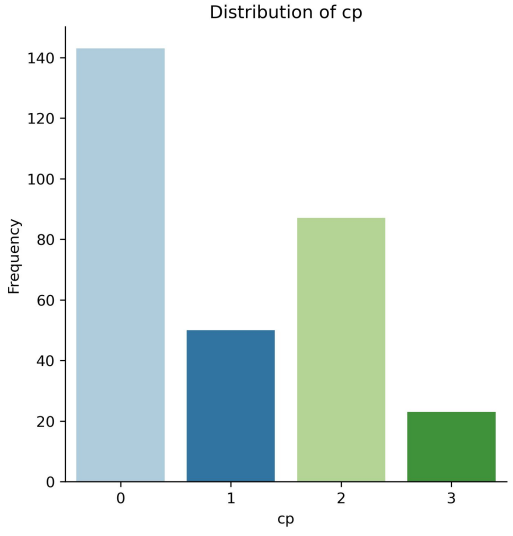
Female patients tend to be older on average than male patients in the dataset.

There is a notable difference in heart attack risk between genders.

Exploratory Data Analysis

- *What is the most common type of chest pain in the dataset, and how does it relate to heart attack risk?*
- *Are there any notable patterns associated with different chest pain types?*



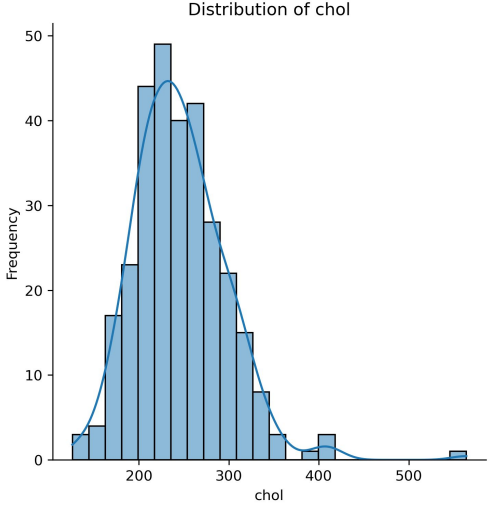


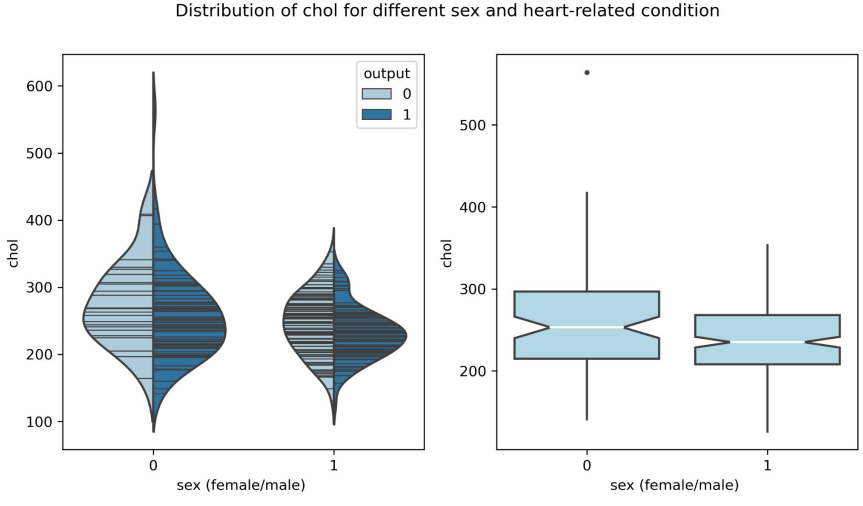
The most common chest pain type is typical angina ("0").

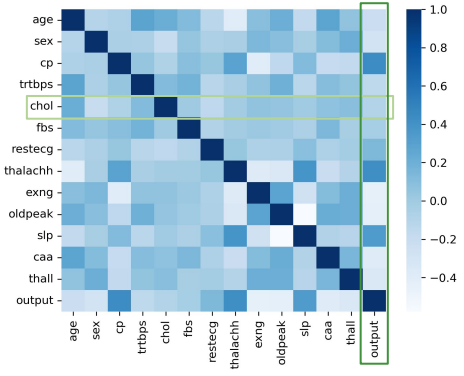
Typical angina ("0") is common even in healthy patients, while non-cardiac chest pain ("2") is most frequently associated with heart attacks.

Exploratory Data Analysis

- *Are there any outliers in the cholesterol measurements of patients, and how does cholesterol level vary between male and female patients?*
- *Is there any correlation between cholesterol levels and heart attack risk?*







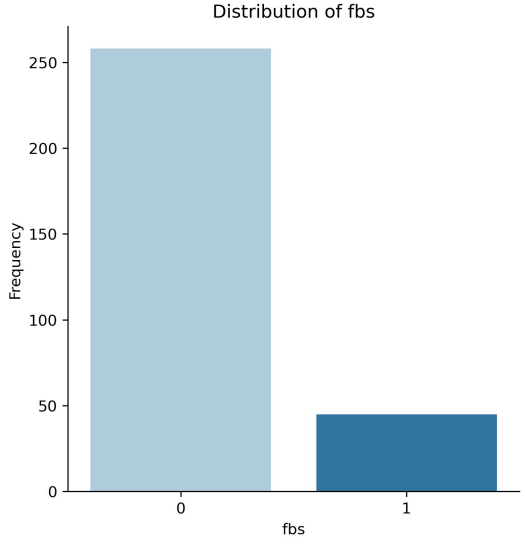
Cholesterol measurements have outliers, with a maximum value of 564.

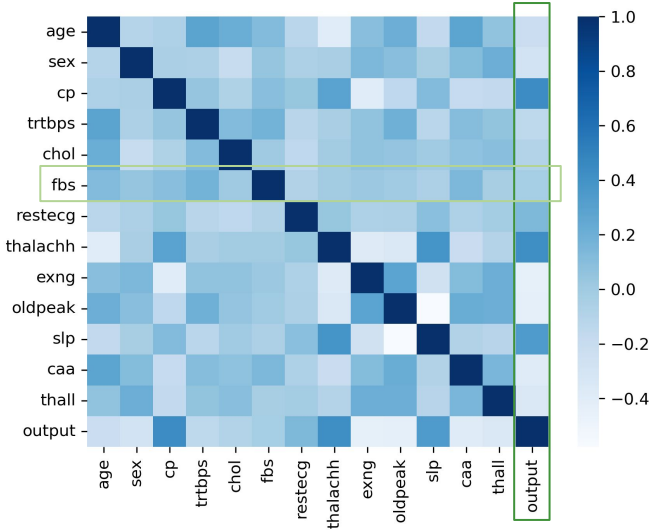
Female patients exhibit higher variance and median cholesterol levels.

No correlation between *cholesterol levels and heart attack risk*.

Exploratory Data Analysis

- *What percentage of patients have normal fasting blood sugar levels, and how does this relate to heart attack risk?*





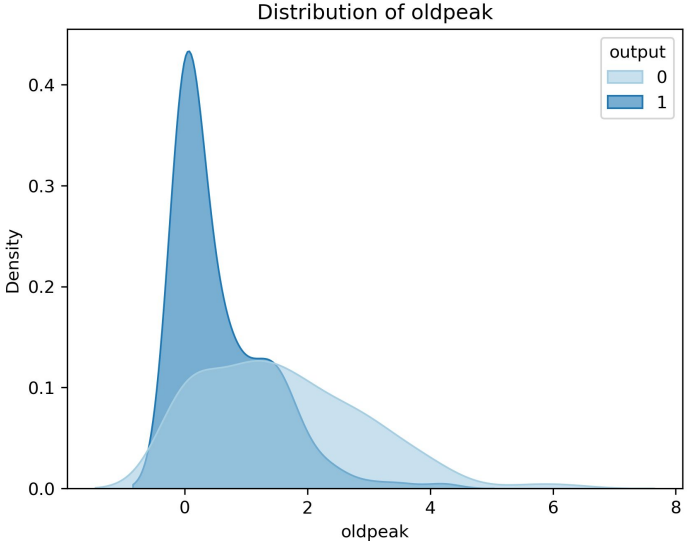
Over 80% of patients have normal fasting blood sugar levels.

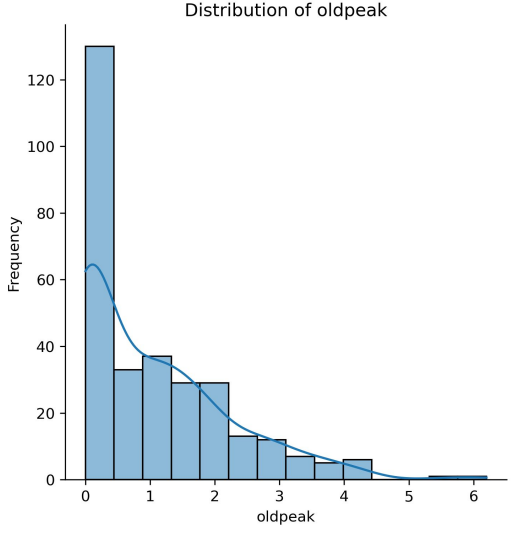
The correlation between fasting blood sugar levels and heart attack risk *is low.*

Exploratory Data Analysis

The previous peak reflects the ST depression induced by exercise relative to rest. It can signify the presence of coronary artery disease.

- *Is there a significant difference in heart attack risk based on previous peak values?*

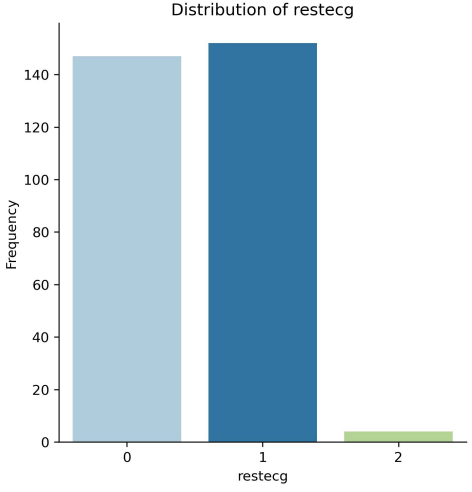


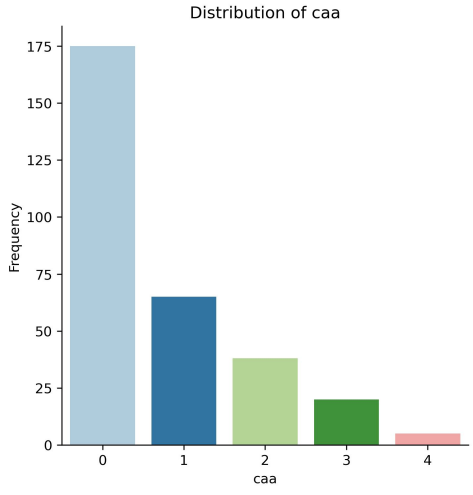


Patients with lower previous peak have significantly more chances of heart attacks.

Exploratory Data Analysis

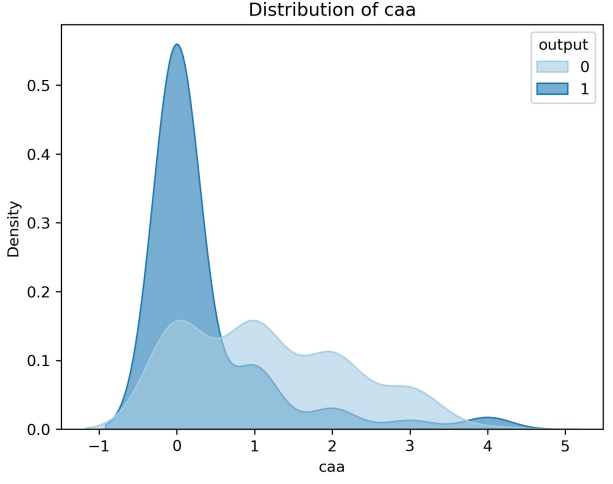
- *What can be observed in the resting electrocardiographic results of the patients?*





Patients rarely have left ventricular hypertrophy in resting electrocardiographic results, with most patients having normal ST-T waves.

- *Does having a lower number of major vessels increase the risk of a heart attack?*

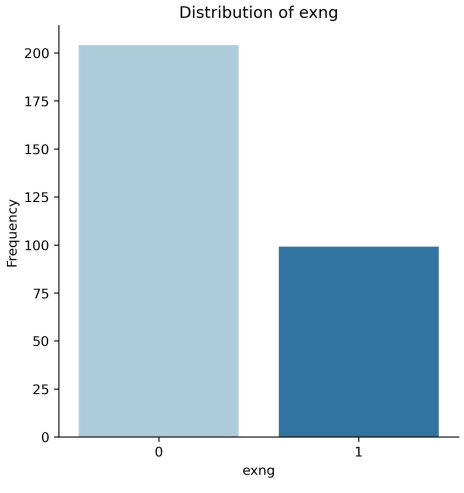


Heart attacks are strongly associated with a lower number of major vessels.

Exploratory Data Analysis

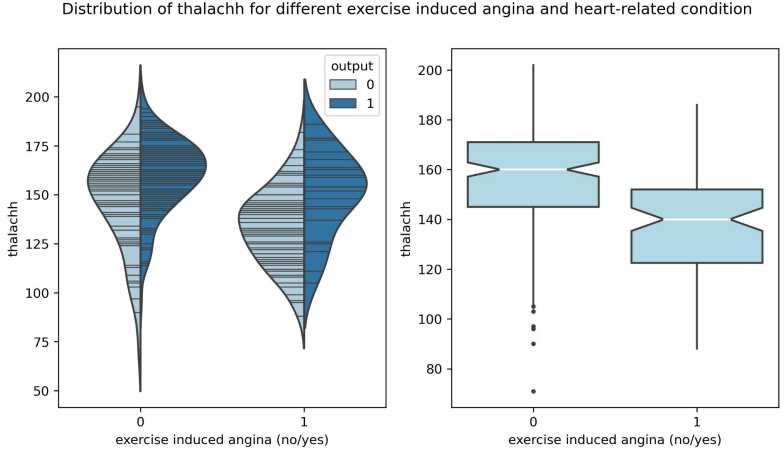
The presence of angina during exercise can indicate underlying heart conditions.

- *What proportion of patients experience exercise-induced angina, and how does this affect their maximum heart rate?*



2/3 of patients have exercise-induced angina in the dataset.

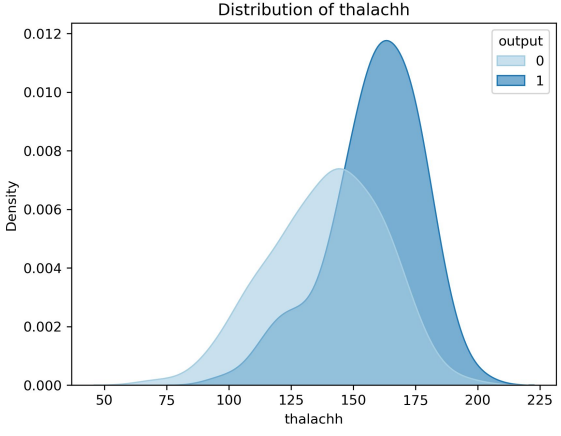
- *Does exercise-induced angina impact the likelihood of a heart attack?*



Patients with exercise-induced angina have a lower maximum heart rate on average.

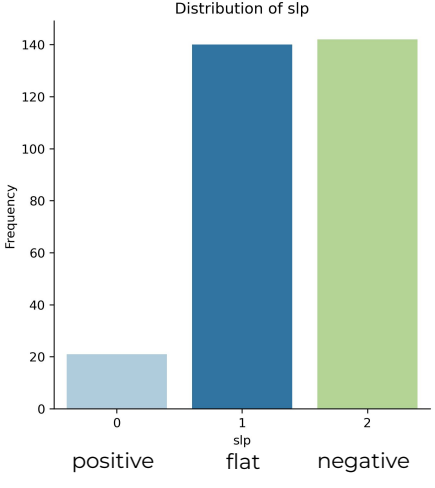
Exploratory Data Analysis

- *Is there a correlation between a higher maximum heart rate and the likelihood of experiencing a heart attack?*

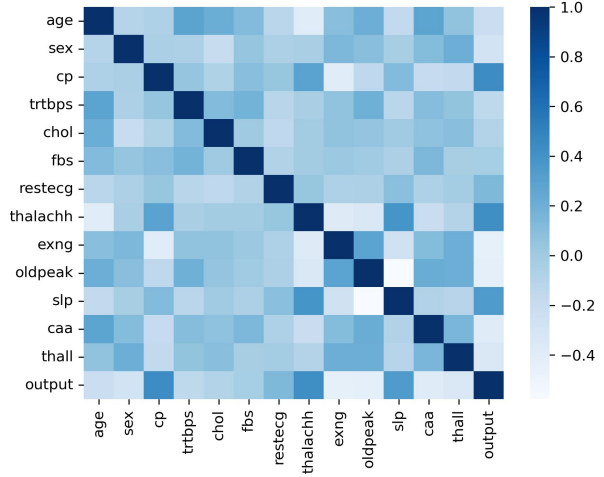


Patients with higher maximum heart rate are more likely to have heart attack

The slope (slp) can provide information about changes in the heart's electrical activity during exercise.



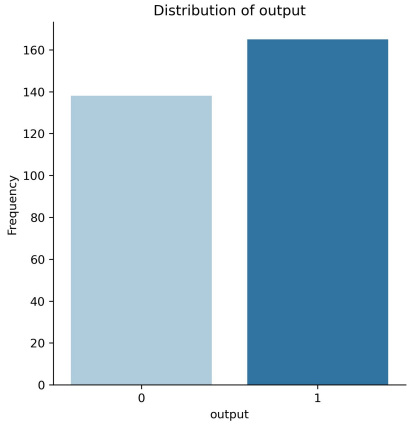
- *Which features in the dataset are most strongly correlated with the "output" variable (heart attacks)?*

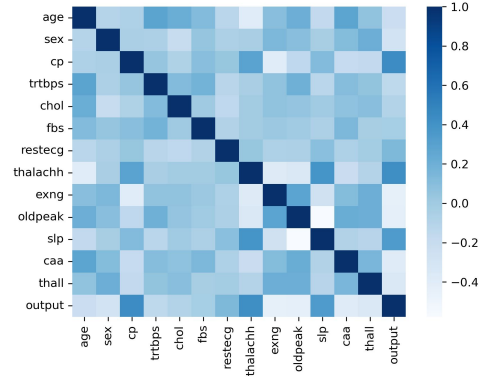


The most significant correlation with the "output" attents for "cp" (chest pain), "slp", and "thalachh" features.

Highlights

- ➔ Younger patients have higher chance of heart attack.
- ➔ There are twice as many male patients as female patients in the dataset. Female patients tend to be **older** on average than male patients in the dataset.
- ➔ **Typical angina** is common even in **healthy patients**, while non-cardiac chest pain is most frequently associated with heart attacks.
- ➔ Over 80% of patients have normal fasting blood sugar levels.
- ➔ **2/3 of patients have exercise-induced angina** in the dataset. Patients with exercise-induced angina have a lower maximum heart rate on average.
- ➔ Patients with lower previous peak have significantly more chances of heart attacks.
- ➔ **Heart attacks are strongly associated with a lower number of major vessels.**
- ➔ The most significant correlation with the heart attack is observed for chest pain, slope of peak exercise segment, and the patient's maximum heart rate during exercise.





Outline

Model Building

Standardization and Feature Engineering

- prepare the categorical features using one-hot encoding
- rescale numerical features to unit variance (StandardScaler from sklearn)
- split data: 20% of the dataset will be used for testing

Building Baseline

Baseline model gives the most naive prediction and the reference, which a more advanced model must beat. Here it is the frequency of the target variable in the training dataset.

Baseline Accuracy: 0.55

Choosing Models and Metrics

I will compare 3 models:

1. Logistic Regression
2. Support Vector Machines
3. XGB Classifier

Across the following metrics:

➔ Accuracy
➔ Precision
➔ Recall
➔ F1-score
➔ ROC curve
➔ Gini Importance

Outline

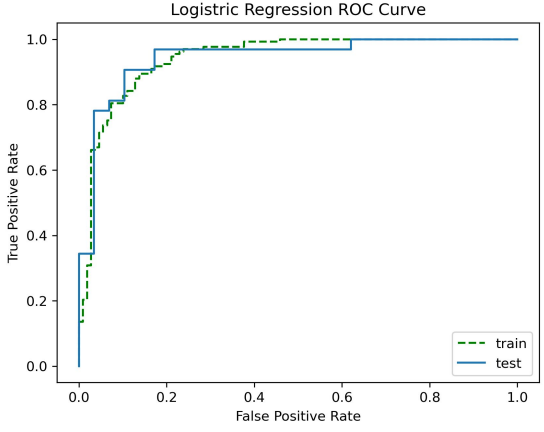
Evaluating performance

Baseline Accuracy: 0.55

Logistic Regression

Training Accuracy: 0.88
Test Accuracy: 0.9

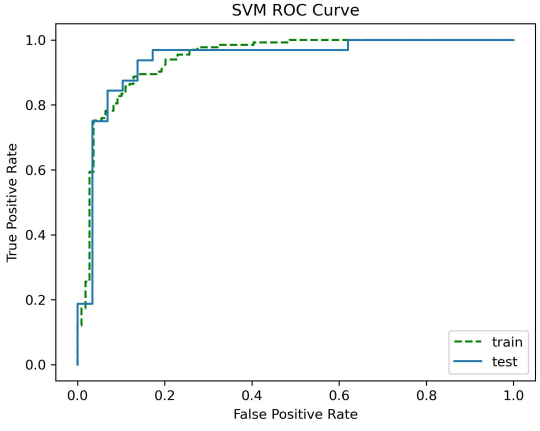
	precision	recall	f1-score	support
0	0.896552	0.896552	0.896552	29.000000
1	0.906250	0.906250	0.906250	32.000000



Support Vector Machines

Training Accuracy: 0.87
Test Accuracy: 0.89

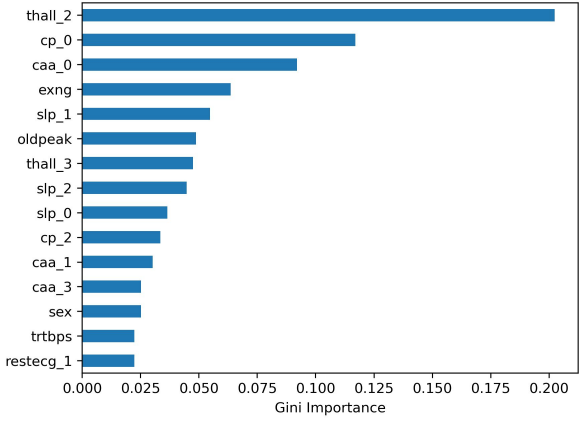
	precision	recall	f1-score	support
0	0.866667	0.896552	0.881356	29.000000
1	0.903226	0.875000	0.888889	32.000000



XGB Classifier

Training Accuracy: 0.92
Test Accuracy: 0.85

	precision	recall	f1-score	support
0	0.833333	0.862069	0.847458	29.000000
1	0.870968	0.843750	0.857143	32.000000



Conclusions

- The best-performing model is Logistic Regression. *The exclusion of fasting blood sugar measurements improved results for all models*.

In conclusion, this project represents a step forward in exploring and the cost-effective prediction of heart-related conditions, using datasets containing clinical and demographic factors.

Key findings from EDA include the strong influence of age, *non-cardiac chest pain, low previous peak values, high maximum heart rate, negative slope of peak exercise segment, the patient's maximum heart rate during exercise, and low number of major vessels on heart attack risk*.

Furthermore, the aspiration to achieve the same level of accuracy while using fewer factors demonstrates the potential for cost-effective early diagnosis in the context of heart-related conditions. By identifying the most influential factors and optimizing the model, this project contributes to the important goal of reducing the economic burden associated with early heart condition detection.