HEART FAILURE PREDICTION IN PATIENTS:

A MACHINE LEARNING APPROACH

NINAD DIXIT

SPRINGBOARD DATA SCIENCE BOOTCAMP

BACKGROUND

- Cardiovascular diseases (CVDs) death toll – 17.9 million 2019 (32% of all global deaths)
- At least 75% deaths occurred in low- and middle-income countries
- Patients may not have access to advanced care

SOURCE: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

PROBLEM

HOW CAN WE RELIABLY PREDICT HEART FAILURE IN PATIENTS?

SOLUTION

USE MACHINE LEARNING ALGORITHMS

- SAVE LIVES
- SAVE HEALTHCARE EXPENSES

THE DATA

- Obtained from Kaggle competition: https://www.kaggle.com/fedesoriano/heart-failure-prediction

- Combination of 5 datasets:
  - Cleveland, Hungary, Switzerland, Long Beach VA, and Stalog (Heart) datasets

- 918 unique observations

DATA WRANGLING

- Target feature: ‘HeartDisease’ (1: heart disease, 0: normal)
- No missing values, but many entries with ‘Cholesterol’ = 0
- Dropped all observations with at least one feature with ‘0’ value

DATA WRANGLING

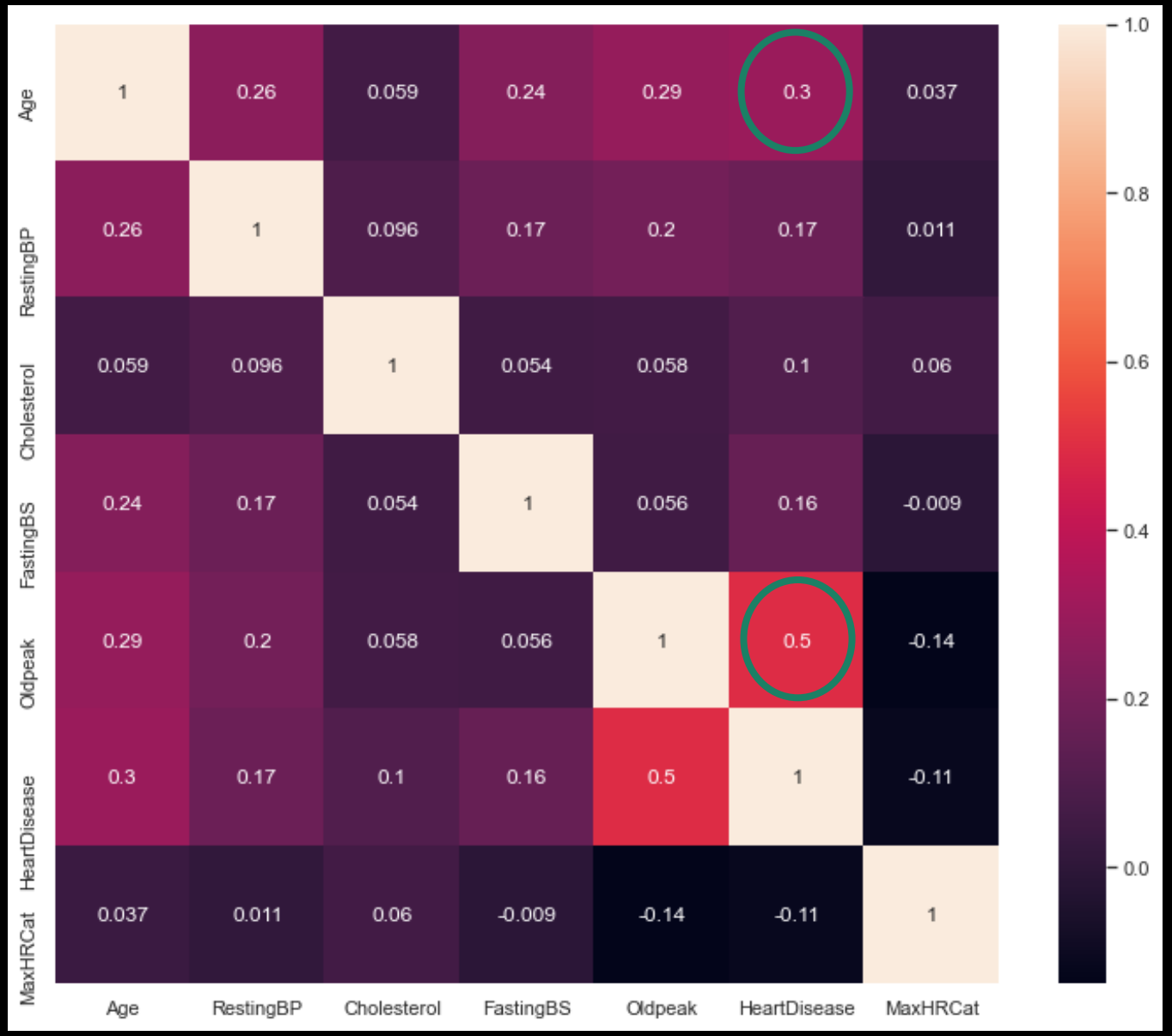
- ‘MaxHR’: maximum heart rate achieved (numerical feature)

- ‘MaxHR’ converted to categorical variable:
  - MaxHR < avg max heart rate: ‘0’
  - MaxHR > avg max heart rate: ‘1’

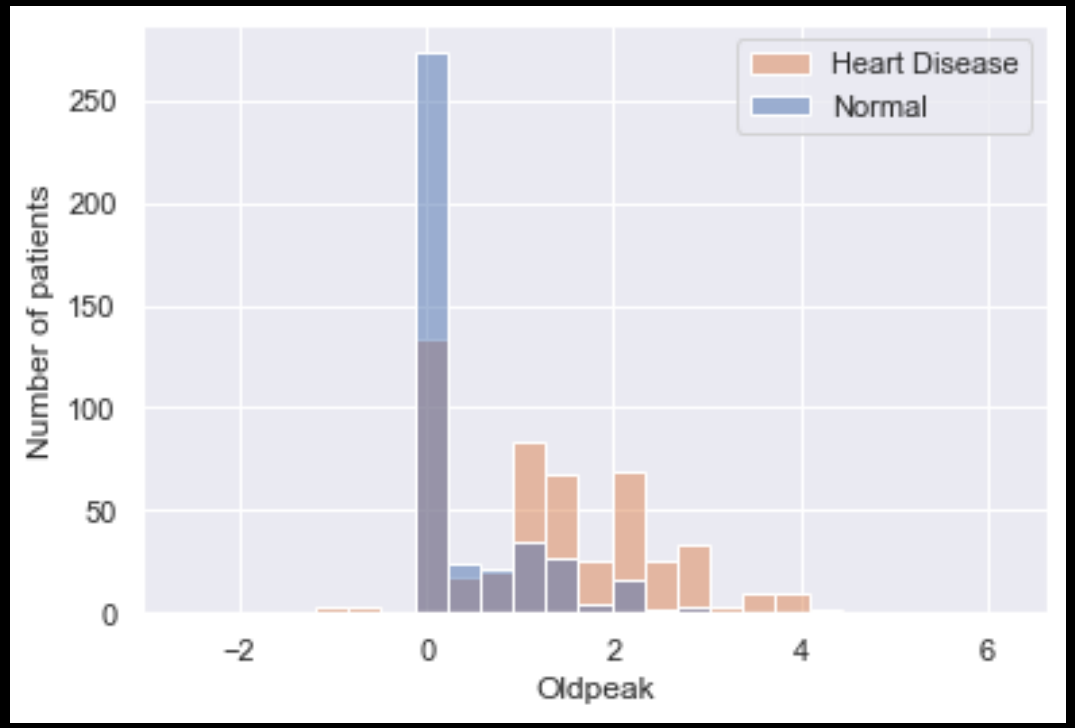
Age	Target HR Zone 50-85%	Average Maximum Heart Rate, 100%
20 years	100-170 beats per minute (bpm)	200 bpm
30 years	95-162 bpm	190 bpm
35 years	93-157 bpm	185 bpm
40 years	90-153 bpm	180 bpm
45 years	88-149 bpm	175 bpm
50 years	85-145 bpm	170 bpm
55 years	83-140 bpm	165 bpm
60 years	80-136 bpm	160 bpm
65 years	78-132 bpm	155 bpm
70 years	75-128 bpm	150 bpm

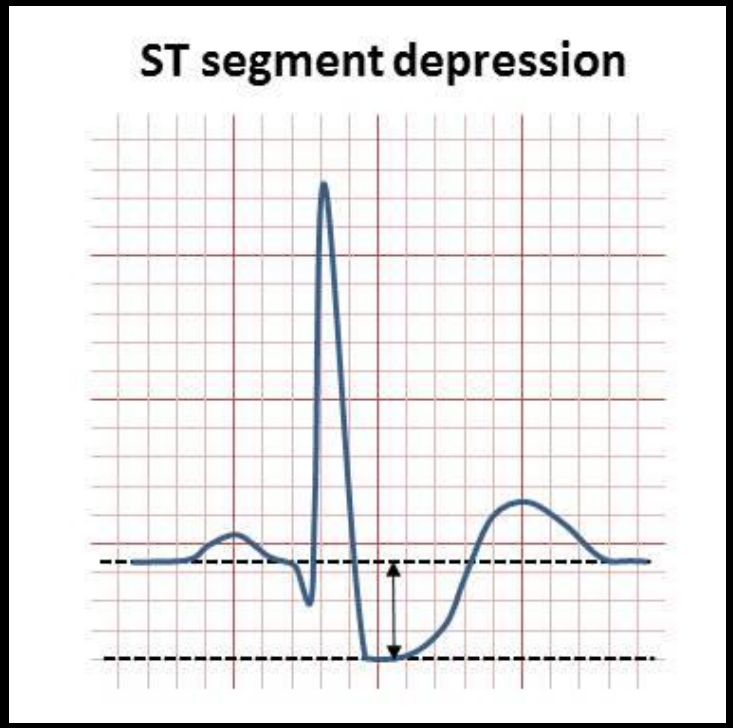
SOURCE: https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates

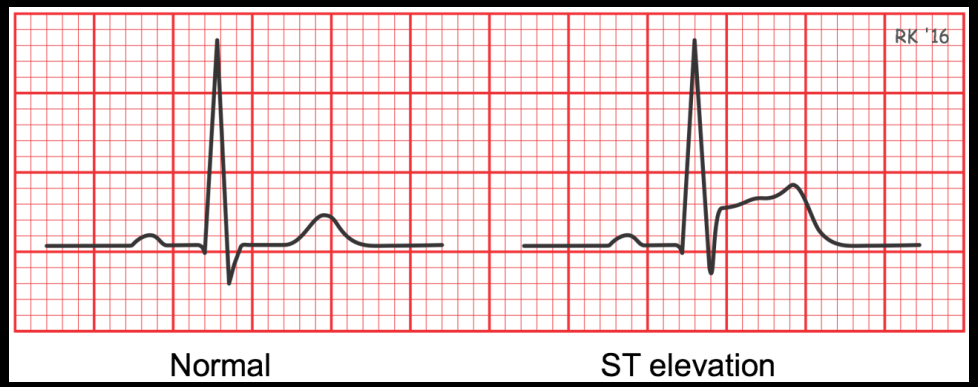
EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS

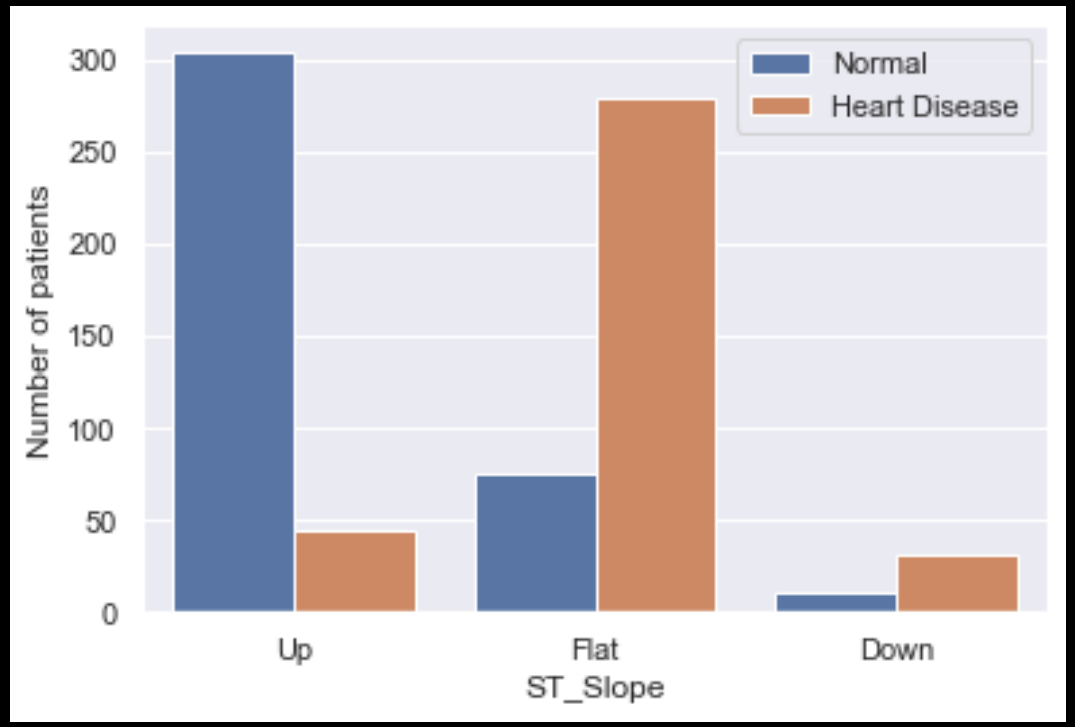


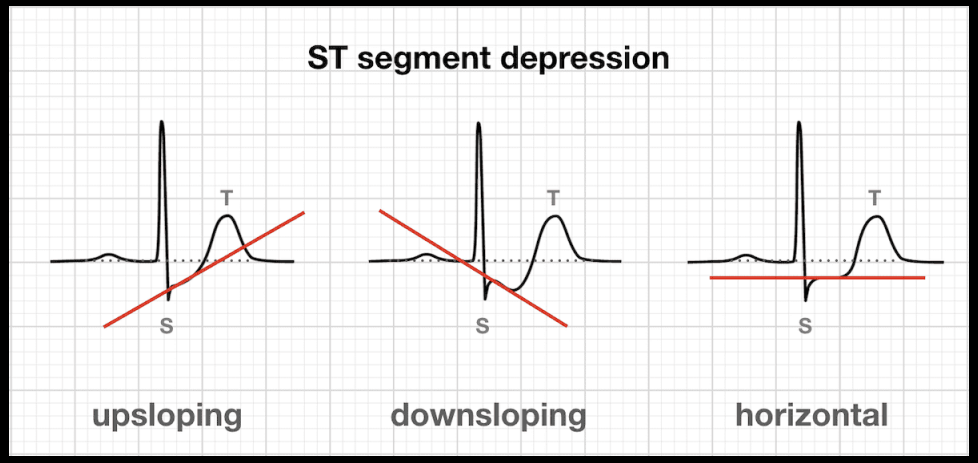




SOURCES:
https://en.wikipedia.org/wiki/ST_depression
https://www.cvphysiology.com/CAD/CAD012

EXPLORATORY DATA ANALYSIS

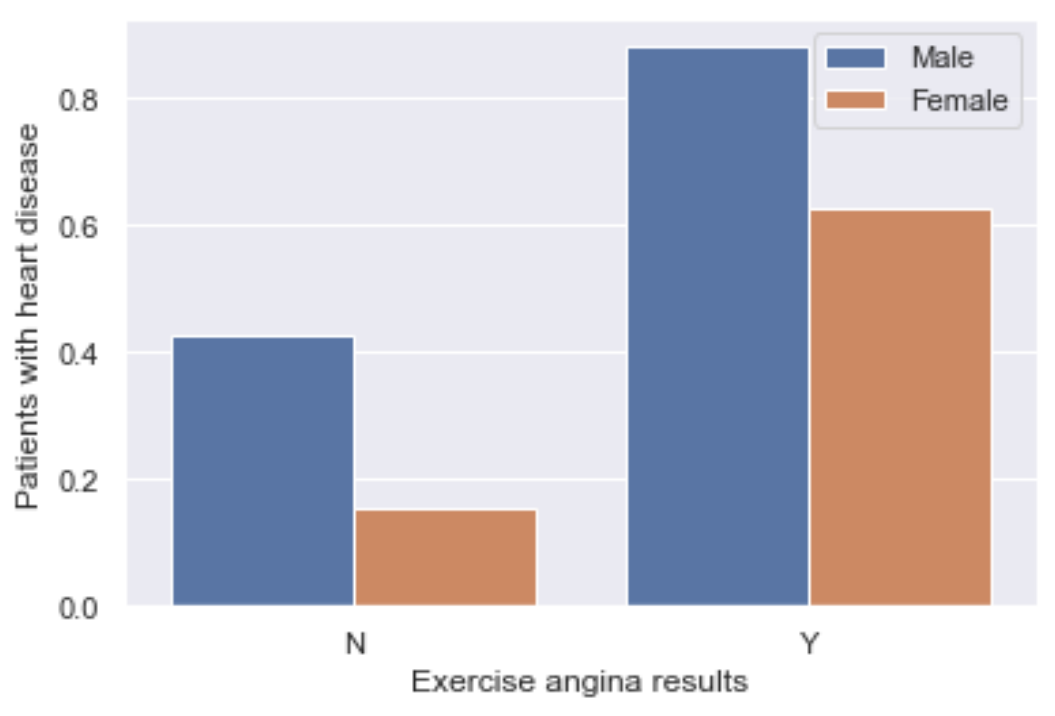


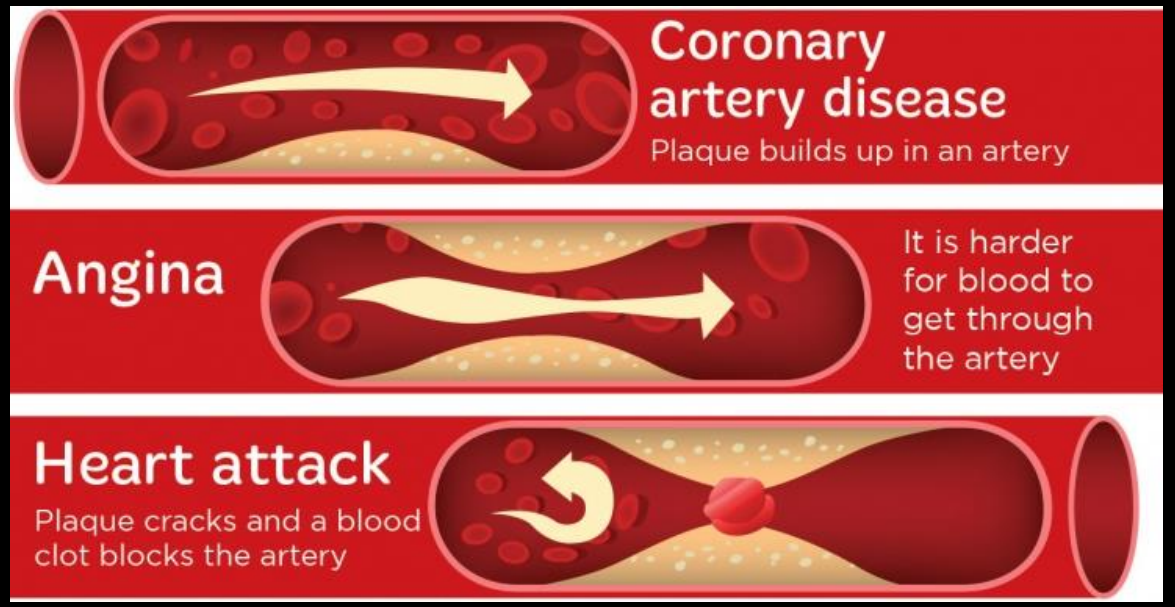


Up: Normal
Flat: heart disease
Down: heart disease

SOURCE: https://litfl.com/st-segment-ecg-library/

EXPLORATORY DATA ANALYSIS





SOURCE: https://www.heartfoundation.org.nz/your-heart/heart-conditions/angina

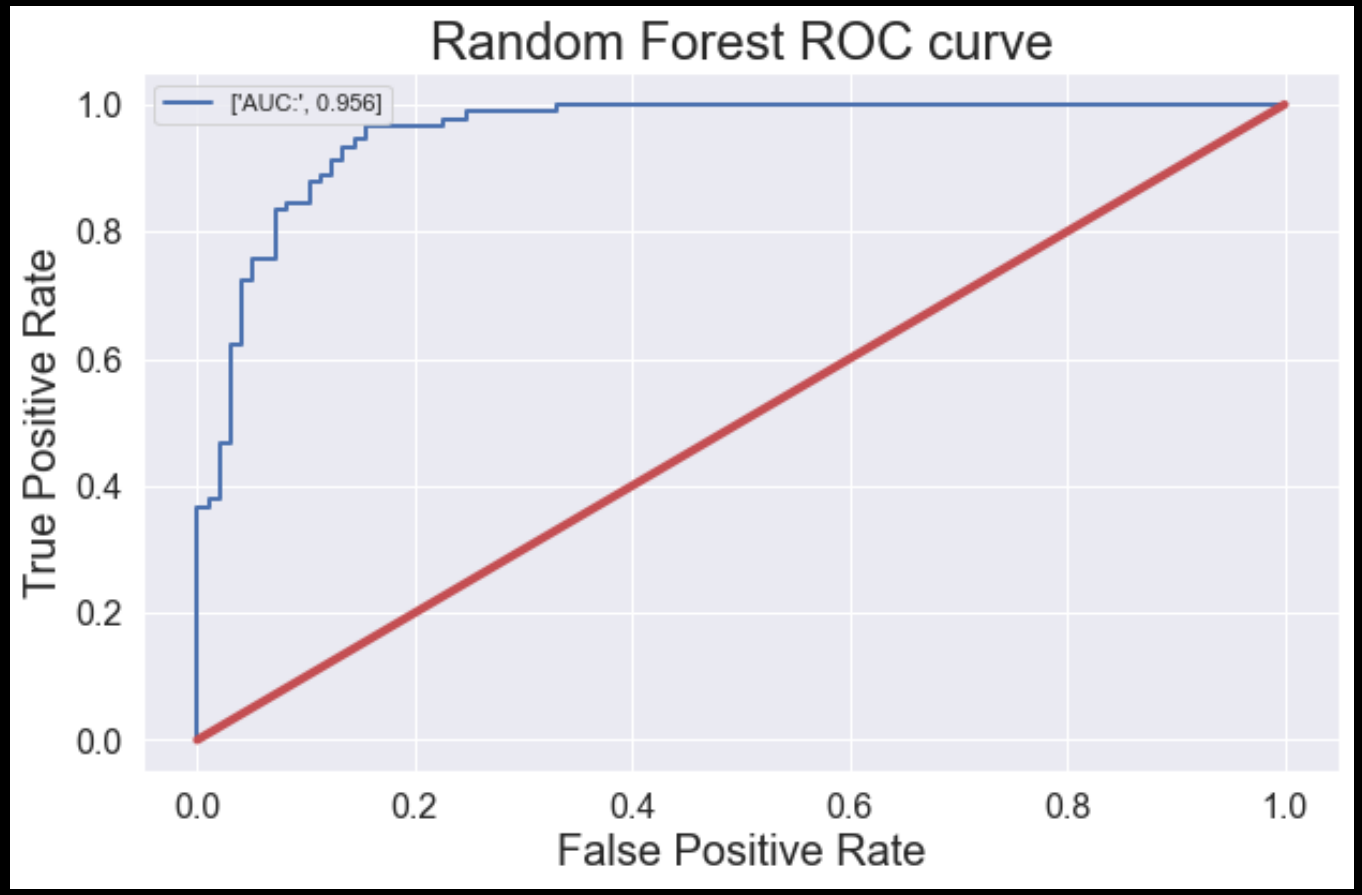
MODEL SELECTION & PERFORMANCE

Model	Accuracy		ROC-AUC
	Cross-validation	Test	
Logistic regression	0.85	0.87	-
K-nearest neighbors	0.84	0.87	-
Random forest	0.85	0.89	0.956
XGBoost	0.85	0.89	0.936

$$accuracy = \frac{(no.\ of\ true\ positives) + (no.\ of\ true\ negatives)}{(total\ no.\ of\ observations)}$$

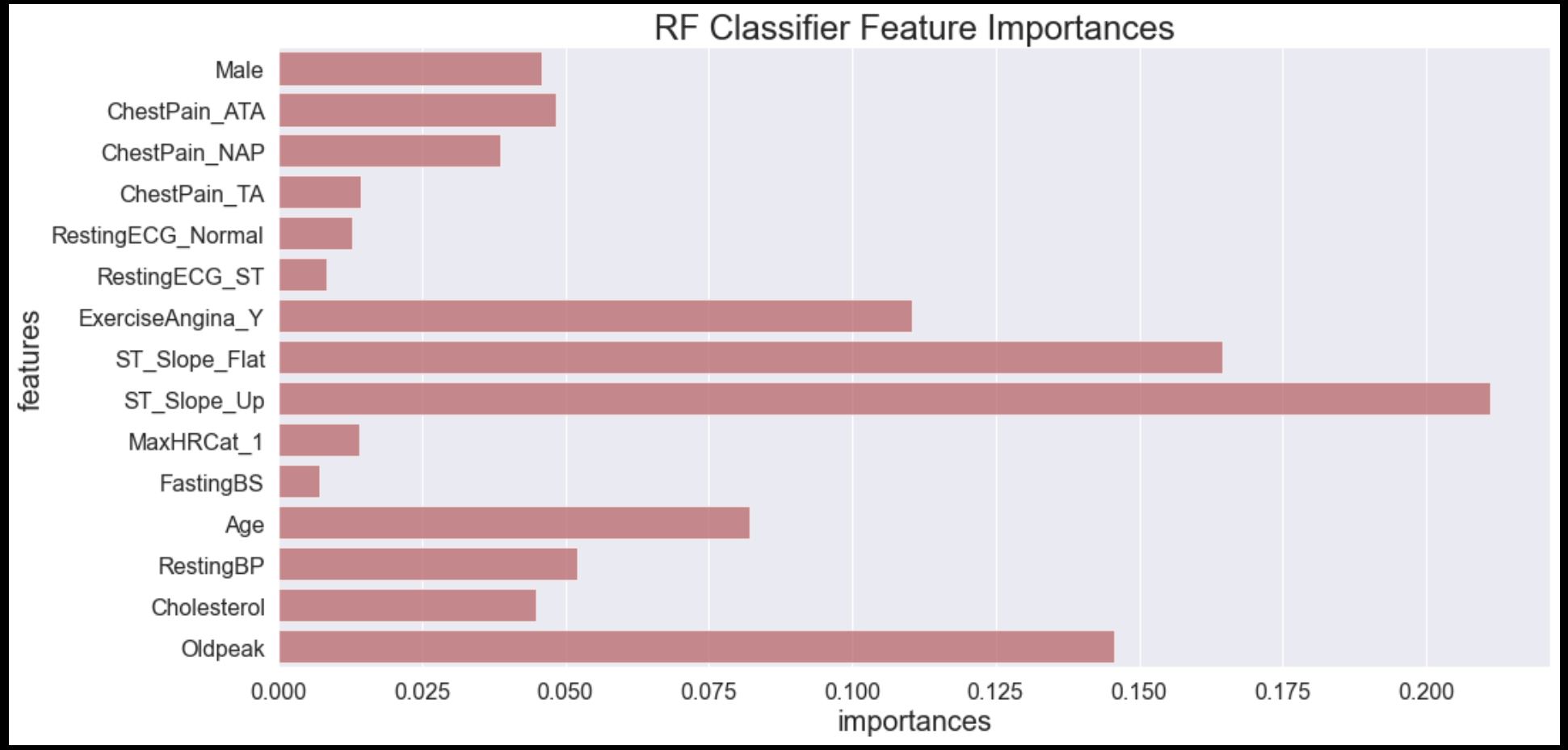
Hyperparameter tuning using Grid Search Cross Validation

MODEL SELECTION & PERFORMANCE



True positive: is ‘1’ and predicted ‘1’
False positive: is ‘0’ and predicted ‘1’

FEATURE IMPORTANCE



ST slope: 0.376

Exercise angina: 0.110

Age: 0.082

Oldpeak: 0.146

Chest pain type: 0.101

CONCLUSIONS

- Random Forest classifier the best performer
- Most important features: ST slope, Oldpeak, Exercise angina

SCOPE FOR IMPROVEMENT

- Choose right metric – precision, recall, f1-score
- Choose right threshold – 0.5 as default
- Use other classifiers – LightGBM

RECOMMENDATIONS FOR CLIENT

- Use the latest medical records for diagnosis
- Seek expert opinion before informing patients

THANK YOU !

- UCI Machine Learning Repository for making the data available publicly
- Ramkumar Hariharan for providing valuable advice

QUESTIONS?