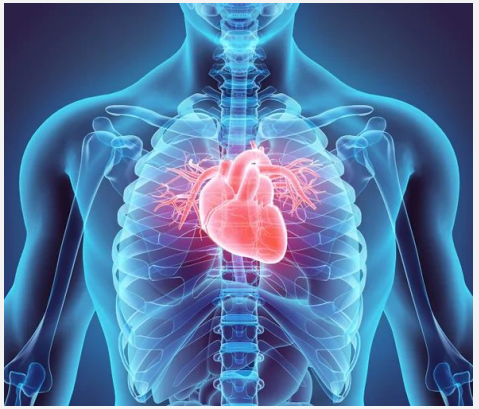
Predicting Heart Disease

Overall Problem

- According to the CDC, heart disease is one of the leading causes of death in the US
- Some risk factors we cannot change
  - Diabetes
- But we can lower risk but changing habits
  - Eating less saturated fats, trans fats, and cholesterol
  - Physical Activity
  - Moderate alcohol consumption
  - Cutting down on cigarettes



Problem Statement

Can we predict whether patients suffered from heart disease using other variables?

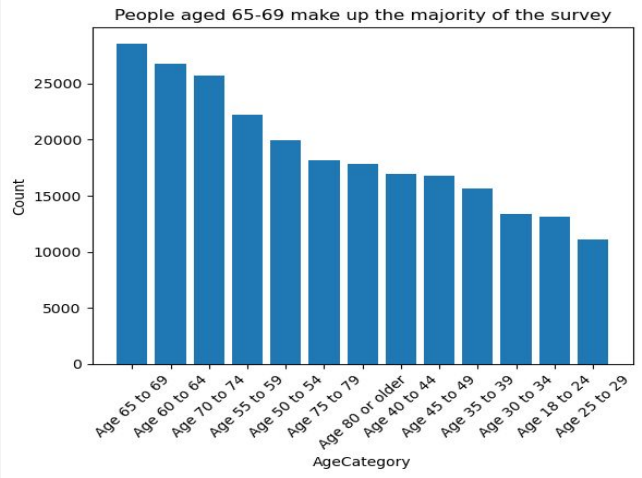
What conditions increase a patient's chance of suffering from heart disease?

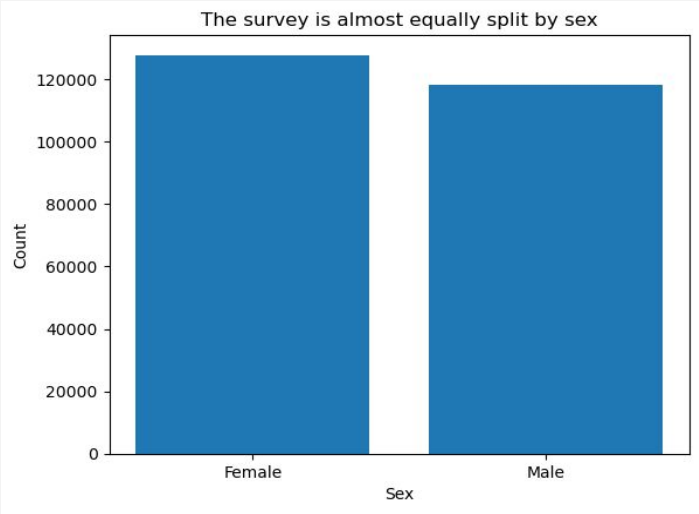
Data

Telephone survey by the Behavioral Risk Factor Surveillance System

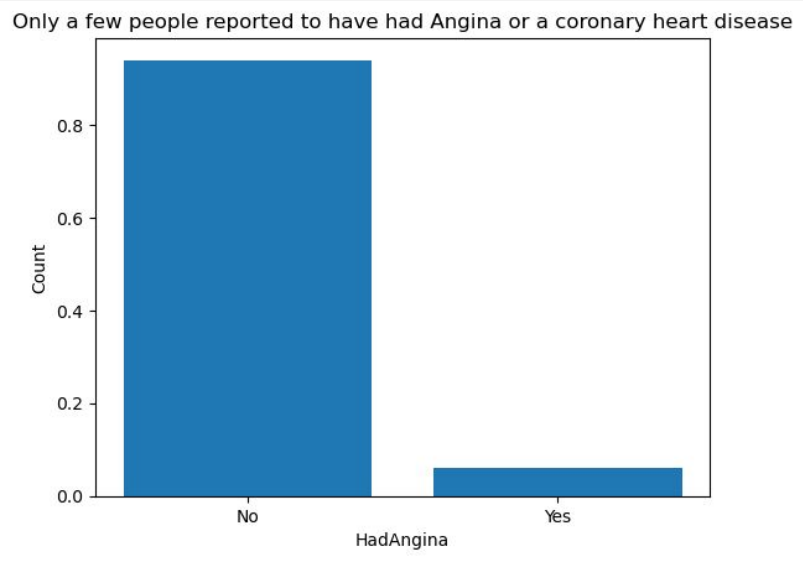
Collects health status of US citizens

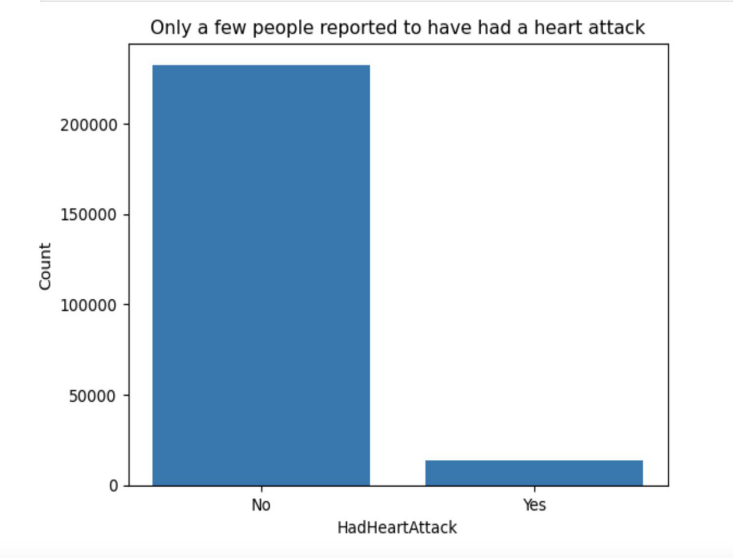
Kamil Pytlak has narrowed down the original surveys from almost 300 variables to 40 key variables for this topic





EDA Findings





6% have reported that they have suffered from angina or any coronary heart disease.

5% have had a heart attack.

The odds of someone developing heart disease, who has previously had a heart attack is 14.9.

Pre-processing + Modeling

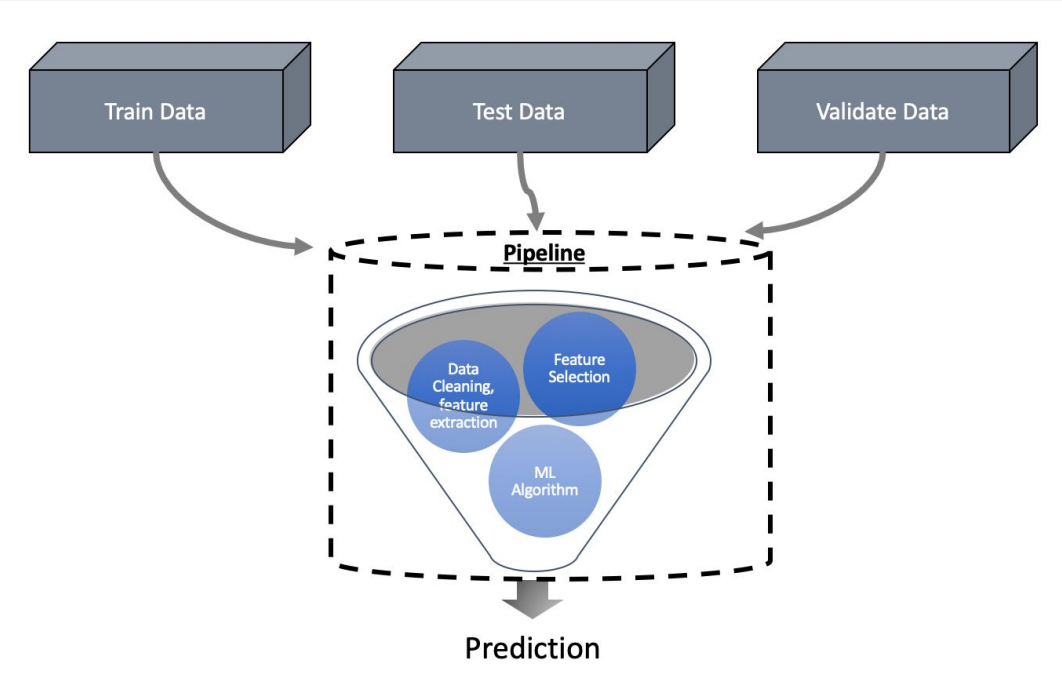


Image source: https://towardsdatascience.com/building-machine-learning-pipelines-7bbd613054cc

Model Types:

- Logistic Regression
- Decision Trees

Pipelines and gridsearch to fine tune and find best models

SMOTE

- Imbalanced Data so that can impact our model.
- Repeat model

Model Evaluation

- Accuracy high with imbalanced data
- Low recall
- Mid precision

After SMOTE

- Lower accuracy
- Higher recall
- Low precision

	F1 score	Recall score	Precision score	Accuracy
Baseline LogReg	38.06	27.98	59.49	94.46
Best LogReg	28.24	18.46	60.07	94.30
Best DT	15.03	8.79	51.77	93.96
Best SMOTE LogReg	25.69	48.41	17.49	82.98
Best SMOTE DT	21.24	41.49	14.27	81.29

	predicted 0	predicted 1
true 0	45845	367
true 1	2439	552

Best LogReg model and its confusion model

Predicted vs Actual

Whats next?