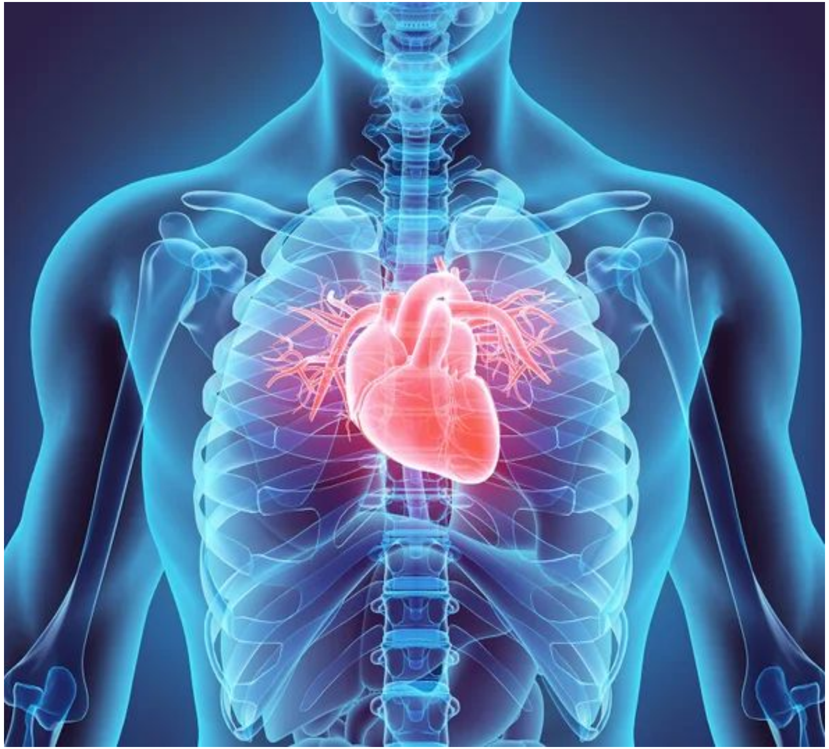
Predicting Heart Disease

Tasnima Jahan

Overall Problem

- According to the CDC, heart disease is one of the leading causes of deaths in the US
- Some risk factors we cannot change
  - Diabetes
  - Age
- 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 are at risk from heart disease using other variables?

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

Data

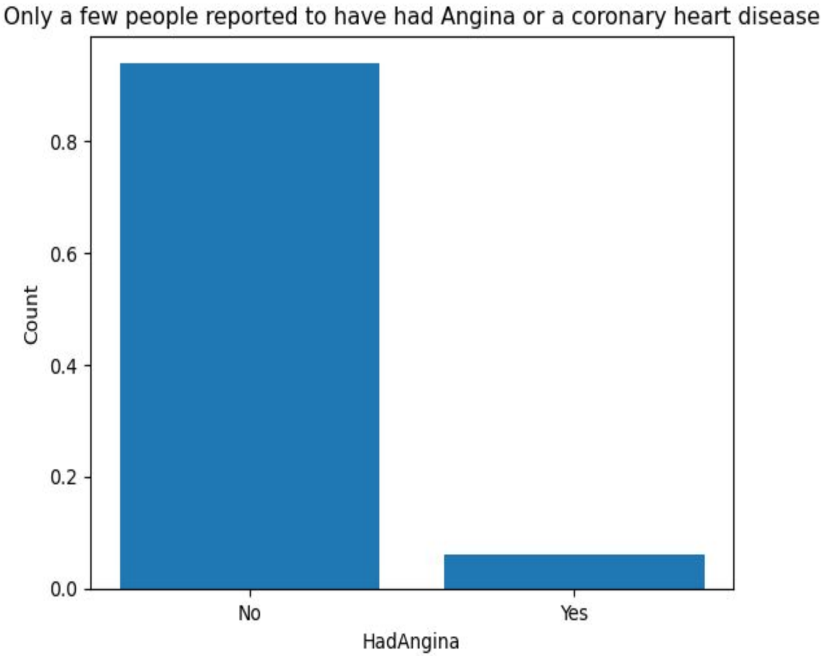
Telephone survey by the Behavioral Risk Factor Surveillance System

Collects health status of 400,000 US citizens each year

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

Reduced to 250,000 after cleaning

EDA Findings

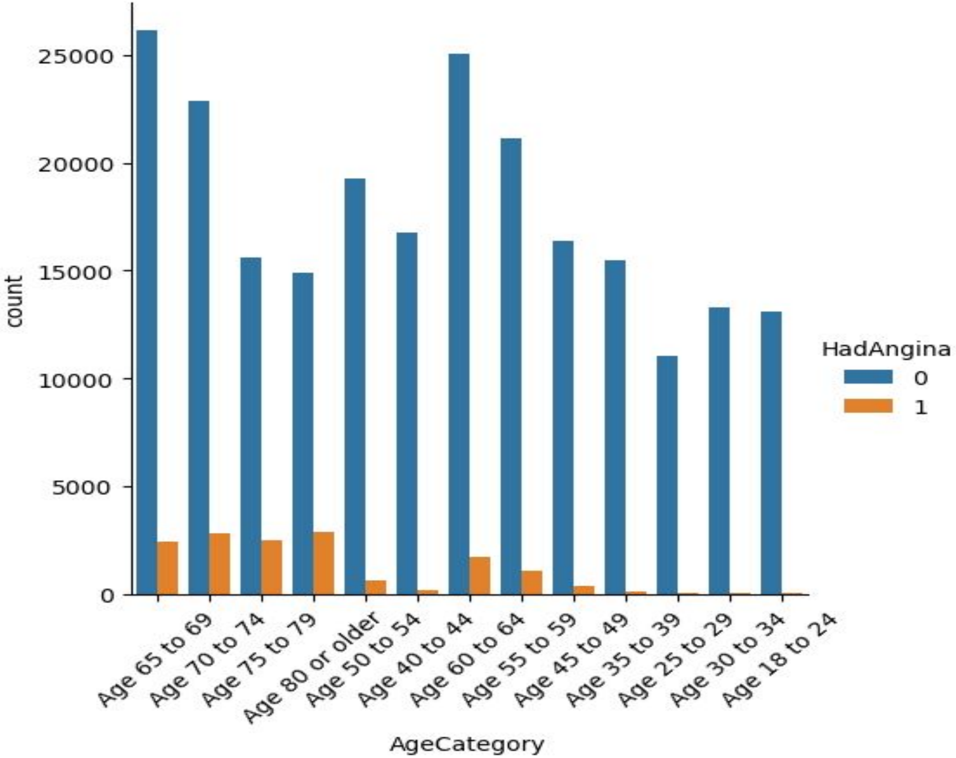


Target Column, what are we trying to predict?

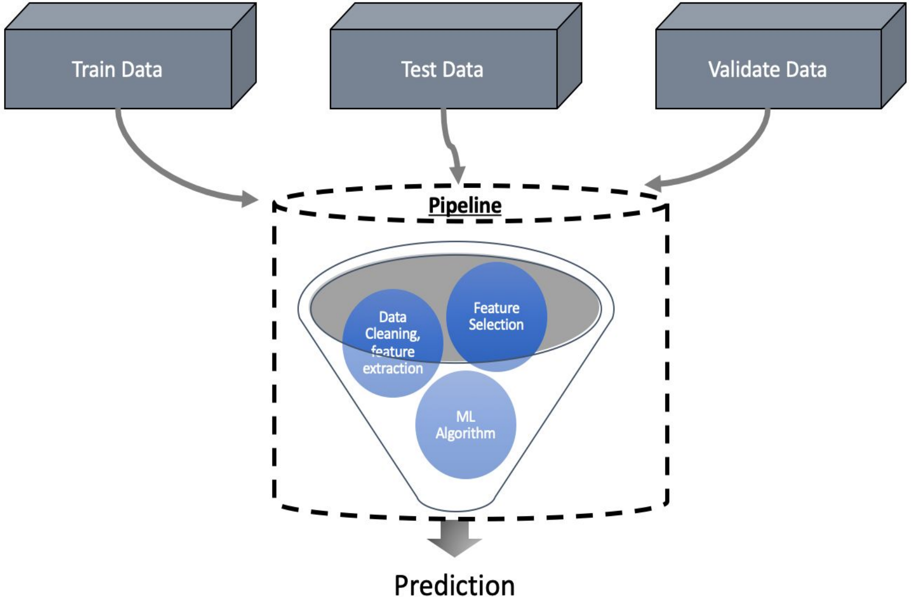
6% have reported that they have suffered from angina or any coronary heart disease.

Very small distribution in our dataset

Age categories of survey takers and whether they have Angina



Pre-processing + Modeling



Model Types:

- Logistic Regression
- Decision Trees
- RandomForest

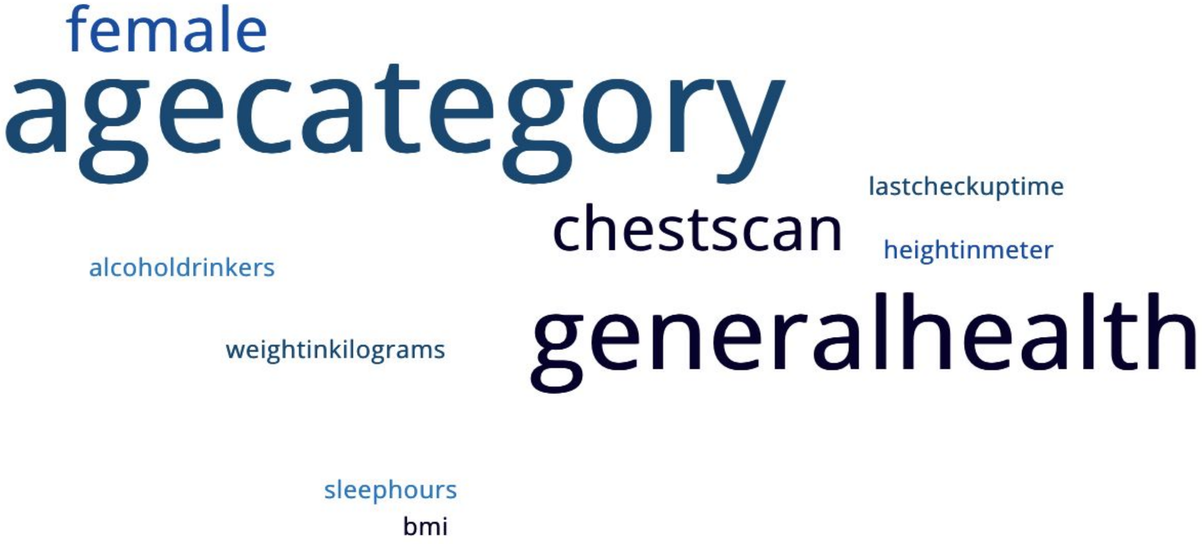
Pipelines and gridsearch to fine tune and find best models

SMOTE

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

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

Top 5 Features



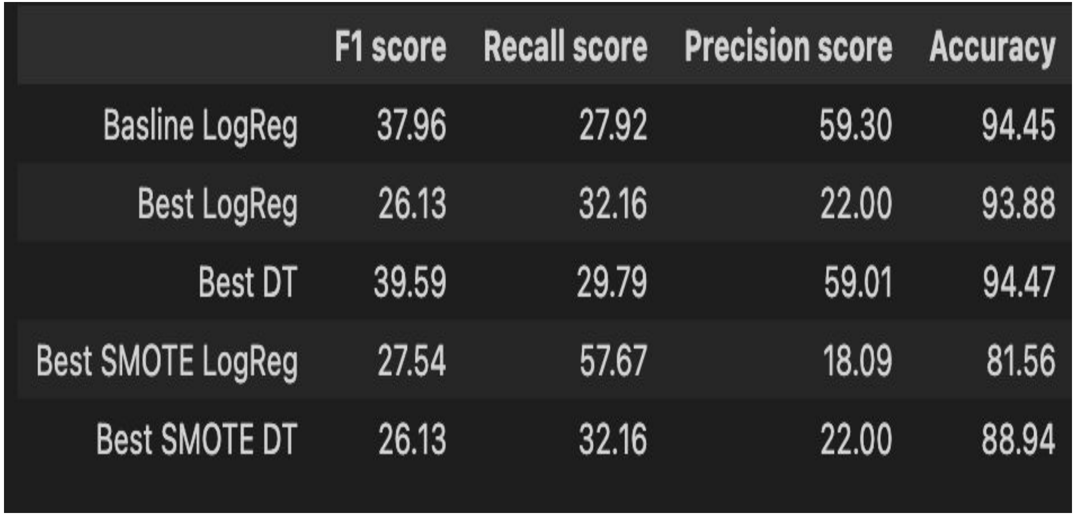
Model Evaluation

Imbalanced Data

- Accuracy high
- Low recall
- Mid precision

After SMOTE

- Lower accuracy
- Higher recall
- Low precision



	F1 score	Recall score	Precision score	Accuracy
Basline LogReg	37.96	27.92	59.30	94.45
Best LogReg	26.13	32.16	22.00	93.88
Best DT	39.59	29.79	59.01	94.47
Best SMOTE LogReg	27.54	57.67	18.09	81.56
Best SMOTE DT	26.13	32.16	22.00	88.94

Whats next?