Detecting Cardiovascular Disease

Using Classification Models

Yasir Karim

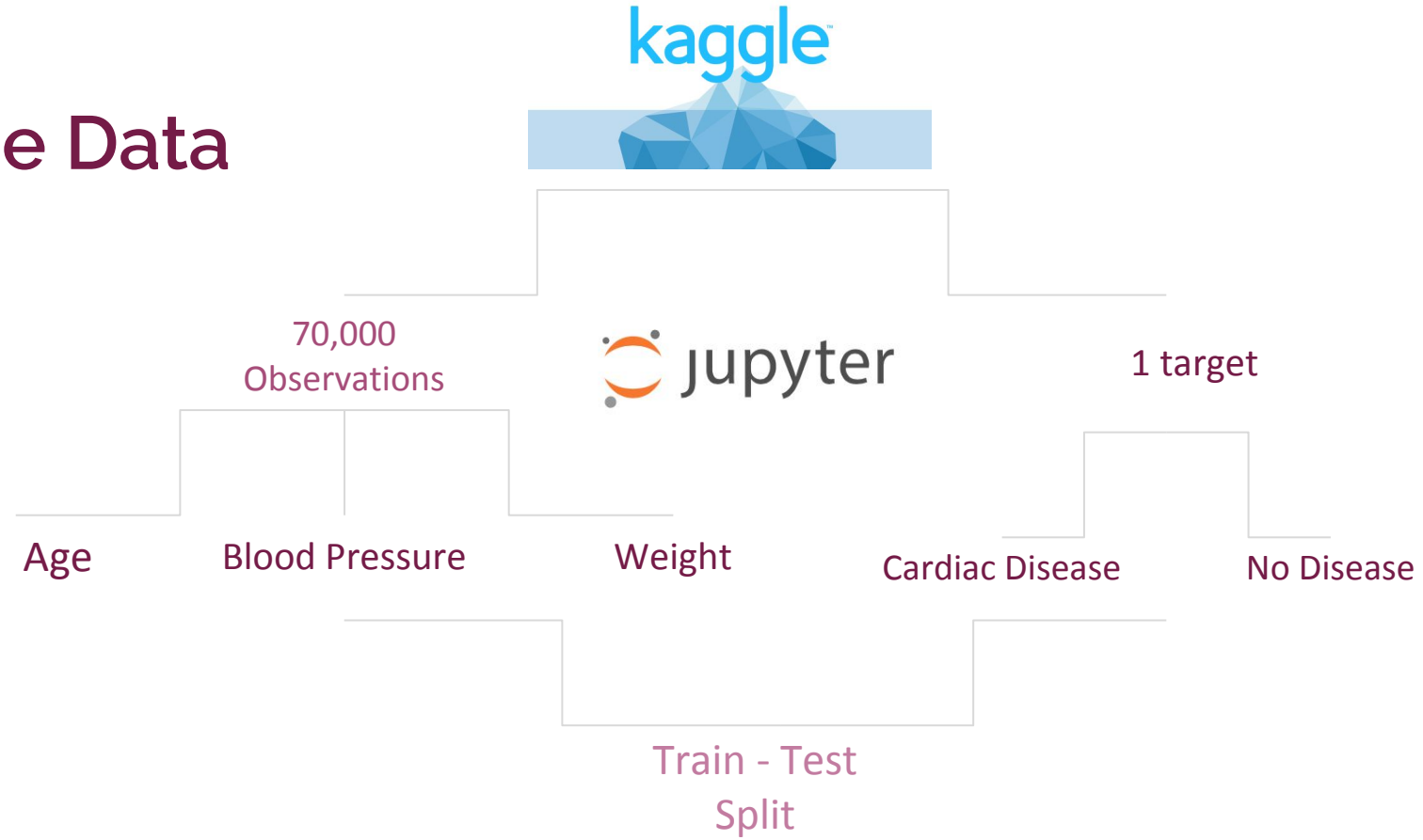
Overview

- *Heart disease* is the **leading cause of death** for men, women, and people of most racial and ethnic groups in the United States.
- About **655,000** Americans die from heart disease each year—that's **1 in every 4 deaths**

Problems to solve

1. How accurately can we predict heart condition in incoming patients in advance?
2. Which metric contribute most in a patient having a heart condition or not?
3. Can we look at a trimmed list of metrics that help us make a prediction?
4. Using the data, can we decrease the number of heart disease related fatalities?

The Data

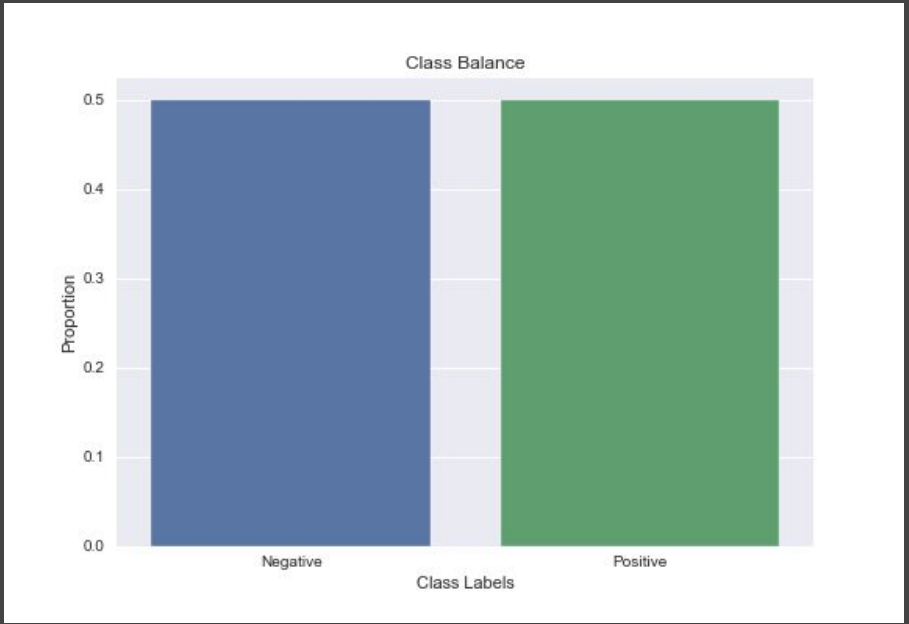

kaggle
70,000
Observations
jupyter
1 target
Age
Blood Pressure
Weight
Cardiac Disease
No Disease
Train - Test
Split

Class Imbalance

"Perfectly balanced,

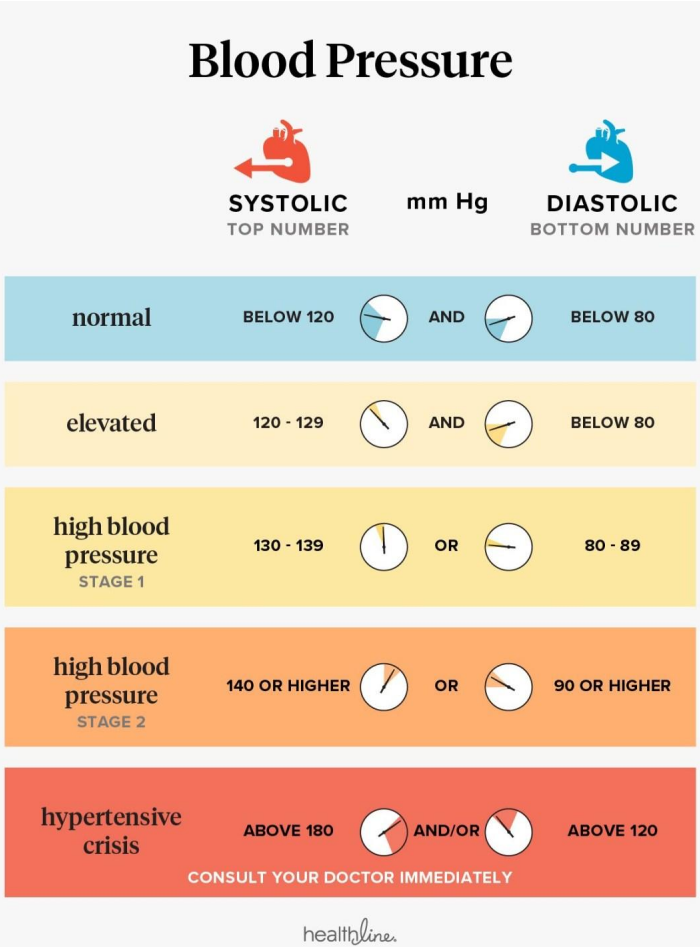
as all things should be"

- Thanos



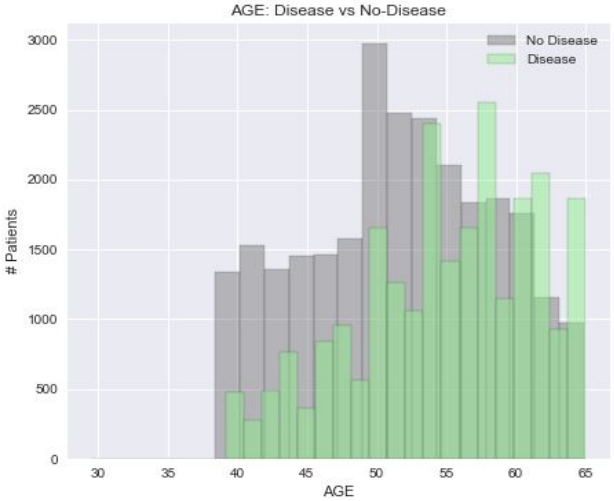
Data Cleaning

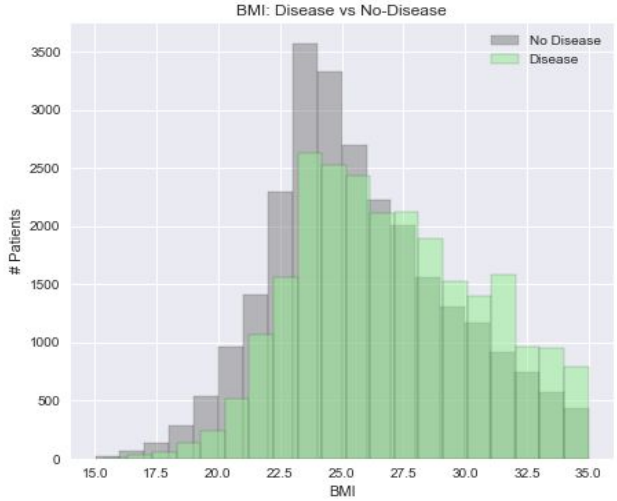
- Many values in the **Blood Pressure** columns were **out of range**. Error during input may have resulted in this. We had to divide these values appropriately to get the correct numbers.
- The units for some of columns had to be **rescaled**.
- There many **extreme outliers** in height and weight columns that had to be **removed**.

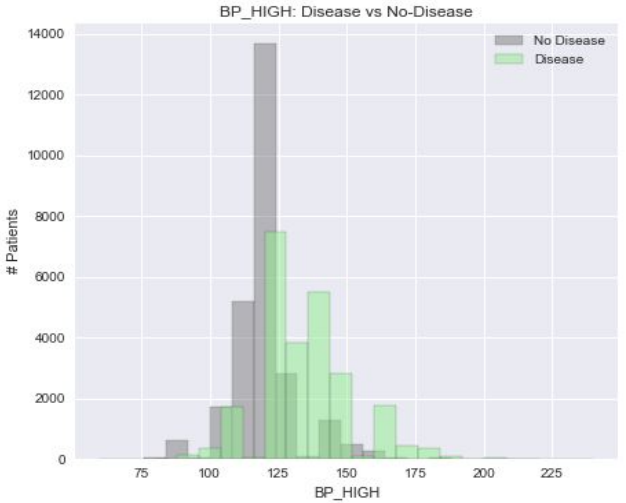


EDA

- From our initial analysis, it seems like **Age**, **BMI**, and **BP** are positively correlated with having a heart condition.







Modelling Process

- We **normalized** our train set before splitting it into two parts for model fitting purposes.
- Then, we used **grid search** to optimize the hyperparameters for each classification model.
- Random Forest and **Decision Tree** had the best recall scores.
- Our best Voting Classifier had those two models with “Soft” voting.

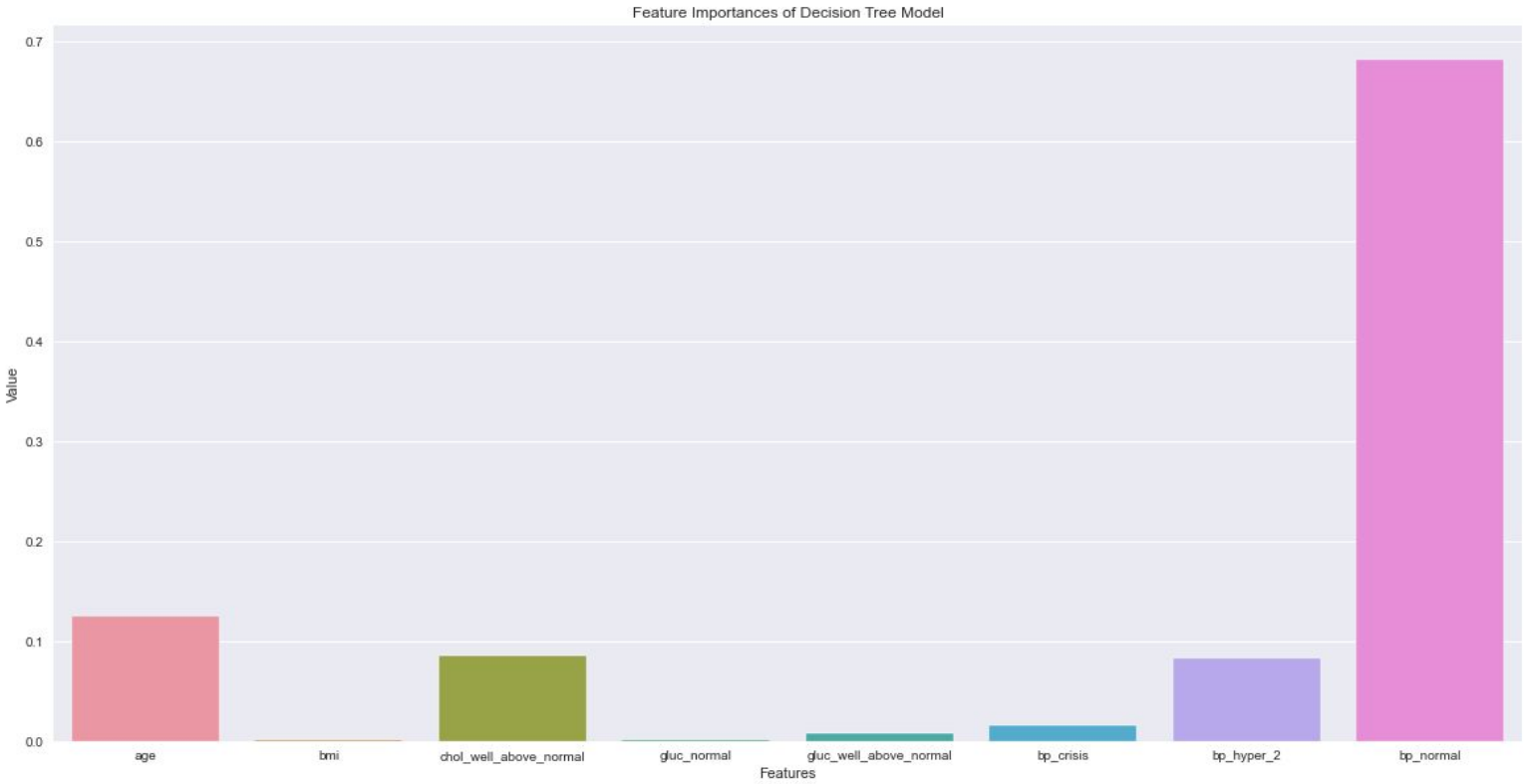
	Recall Score
LogReg	0.621
Decision_Tree	0.715
knn	0.677
random_forest	0.716
voting_clf	0.710

Evaluation Metric

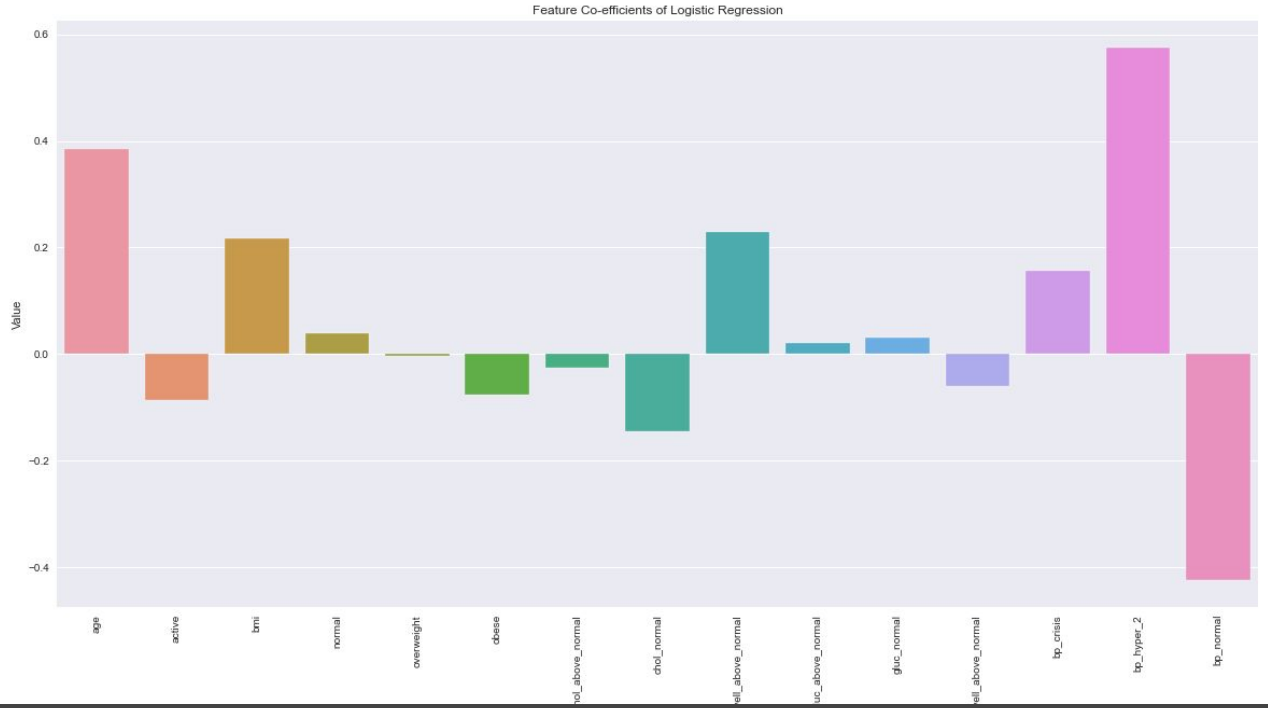
- We used **recall** as our target evaluation metric .
- We wanted to classify the maximum number of patients who end up having a cardiac disease with the **positive class**.
- Even if that results in more **false positives**, it does not outweigh the need to capture more **true positives**.

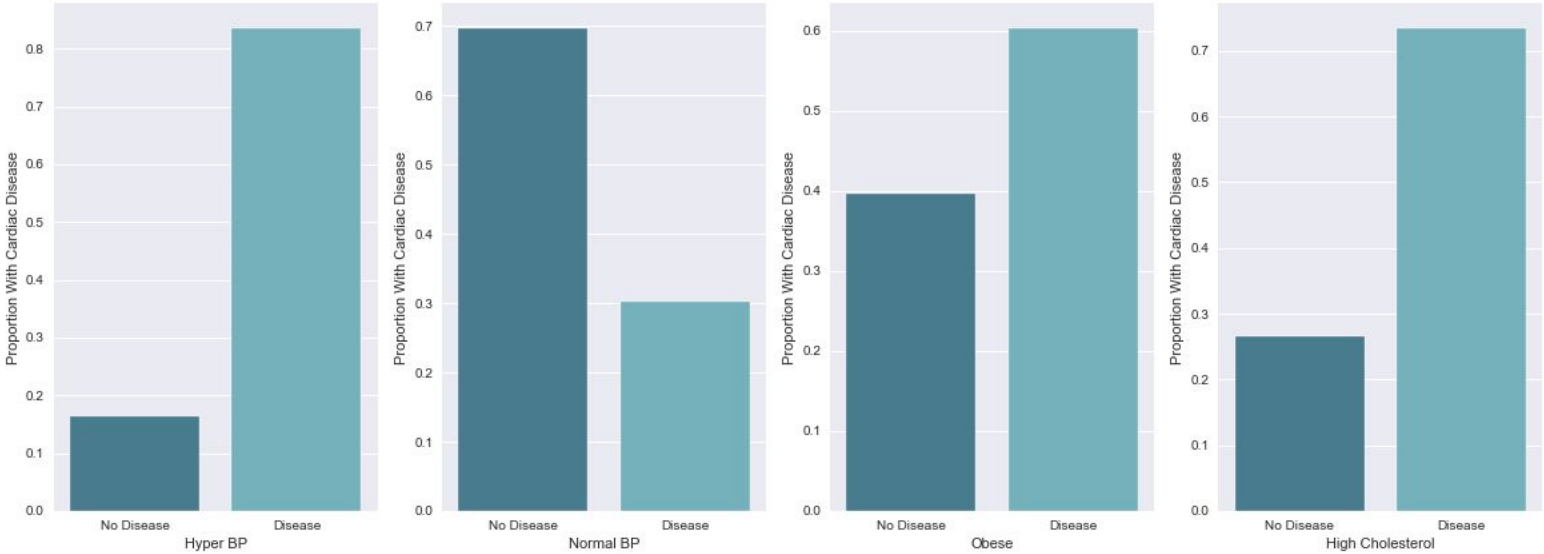
	36%_threshold
recall	0.841
accuracy	0.686
precision	0.631

Feature Importances of Decision Tree



Continued - Feature Coefficients From Logistic Regression




Proportion With Cardiac Disease
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
No Disease
Disease
Hyper BP
Proportion With Cardiac Disease
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
No Disease
Disease
Normal BP
Proportion With Cardiac Disease
0.0
0.1
0.2
0.3
0.4
0.5
0.6
No Disease
Disease
Obese
Proportion With Cardiac Disease
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
No Disease
Disease
High Cholesterol

Applications

Suggestion 1	• Use the predictive model in identifying patients with cardiac disease.
Suggestion 2	• When a patient is checking in, measure the important features to identifying cardiac disease.
Suggestion 3	• Suggest non-emergency patients to take further tests to successfully identify heart condition.
Suggestion 4	• Provide the emergency patients with appropriate healthcare.

Further Research:

- Find the original data source in order to reduce assumptions made and correctly identify background and context.
- Use bagging and boosting methods to try and increase recall score.
- Find similar data sets that have a lot more features to work with.
- Consult with a healthcare expert to in order to gain more knowledge about the ins and outs of medical facility.

Thank you.