Heart Failure Prediction

**Legend of features:**

Age: age of the patient [years]Sex: sex of the patient [M: Male, F: Female]ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]RestingBP: resting blood pressure [mm Hg]Cholesterol: serum cholesterol [mm/dl]FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]ExerciseAngina: exercise-induced angina [Y: Yes, N: No]Oldpeak: oldpeak = ST [Numeric value measured in depression]

ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

HeartDisease: output class [1: heart disease, 0: Normal]

**Dataset:**

The dataset was created by merging different datasets that had similar attributes. 5[[1]](#footnote-1) different heart datasets are combined over 11 features. The five datasets that were used:

* Cleveland: 303 observations
* Hungarian: 294 observations
* Switzerland: 123 observations
* Long Beach VA: 200 observations
* Stalog (Heart) Data Set: 270 observations

Total: 1190 observations  
Duplicated: 272 observations

Final dataset: 918 observations

This dataset has a few interesting aspects such as:

1. Some features are interacting differently with other features. e.g. a feature with value k is normal for a teenager but might be harmful for an elderly person.
2. While some features are made of only two values whereas other features have a wider range. (Sex VS. Cholesterol [0,603])
3. As a result of merging different datasets there is the question how will it affect the accuracy of the results.

**Feature Analysis:**

In a few articles we read about feature selection it’s been said that the method is recommended excluding medical research where it’s important to understand the features which are dropped and their importance for the prediction.

Another aspect we wanted to study was which were the most important features, and to see if it changes per algorithm.

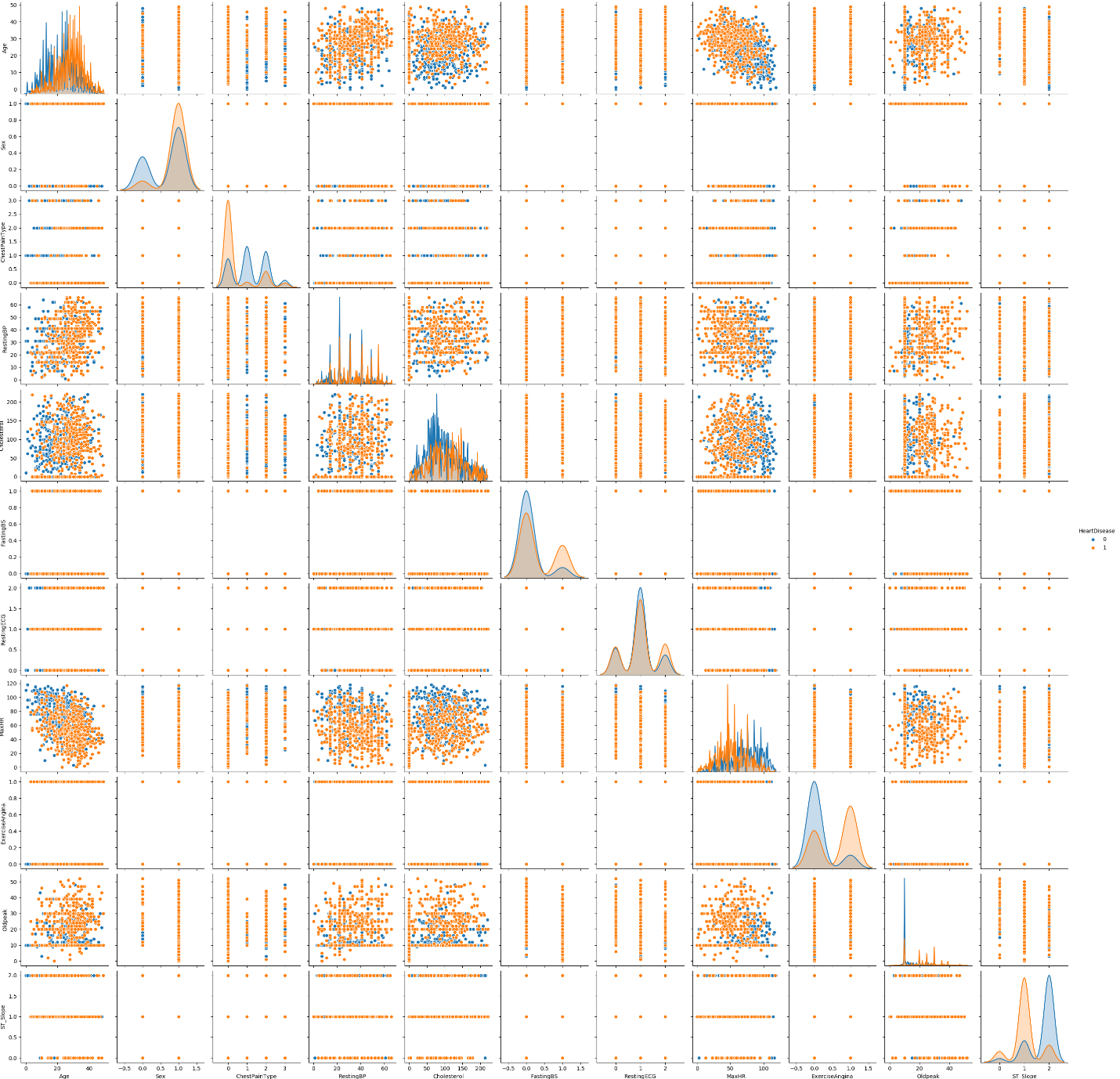
Using the graphs below its easy to point out that some ranges for some of the features are already potential for heart failure such as: Old age, High cholesterol, low MaxHR.

Our assumption is probably those 3 features will be in all the algorithms, as it’s easy to see so probably it will be noticeable in the algorithms.

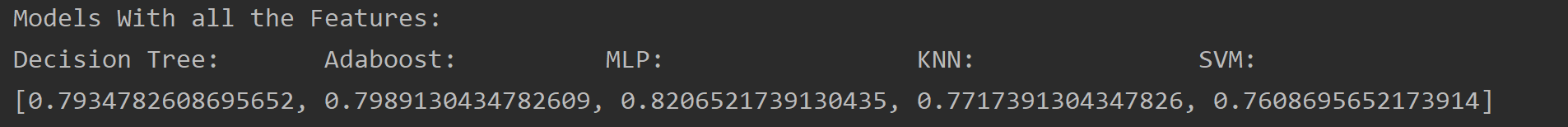
**Before normalizing:**

****

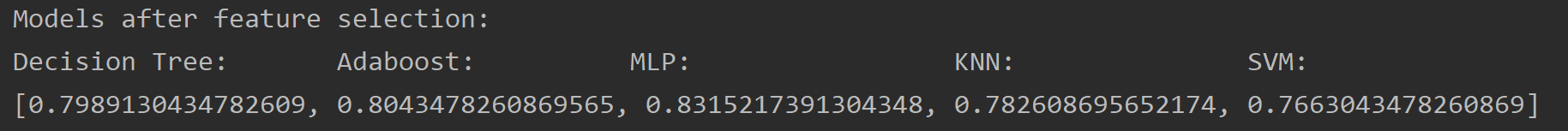
**After normalizing:**



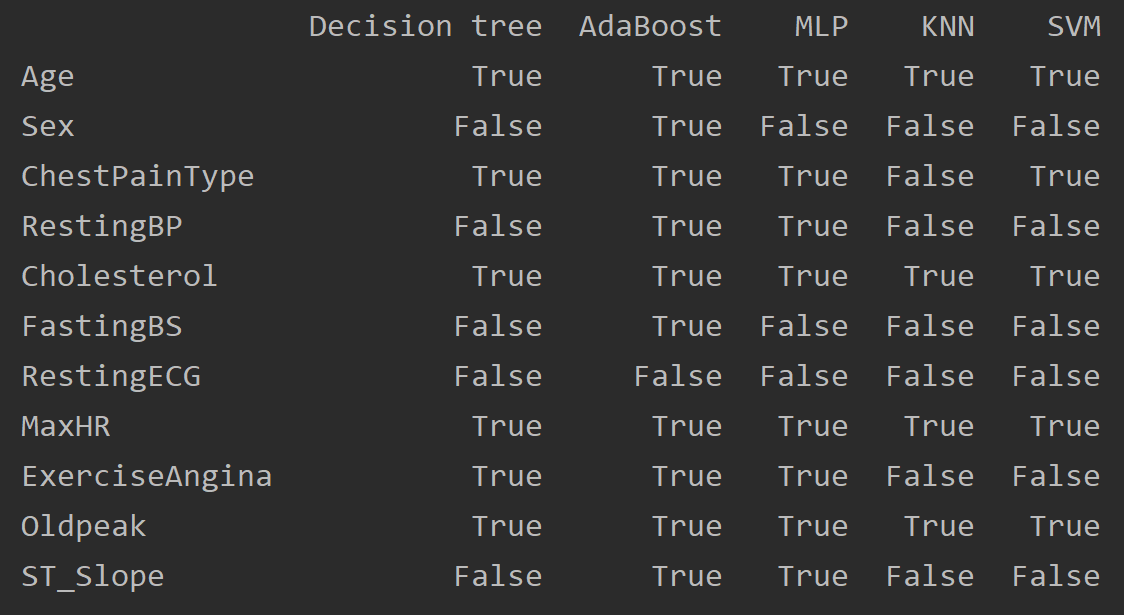
**Models prediction:**

****

**Feature selection:**

****

**Number of features that are true:** 6 10 8 4 5

****

**Impact of test size:**

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

**Challenges:**

**Overfitting –**

We encountered an overfitting in our models: decision tree as well as the Adaboost.

It was interesting to discover that as the Adaboost is constructed from a few decisions tree we expected it to be more resilient than the decision tree itself as its constructed from a few models as classifiers rather than one.

**Encoding the data –**

A few of our features came with string values that aren’t quantified like regular numbers in a regular Machine learning dataset. After some research we found that there are two approaches to handle the problem, so we did the adjustments to apply the method on the dataset.

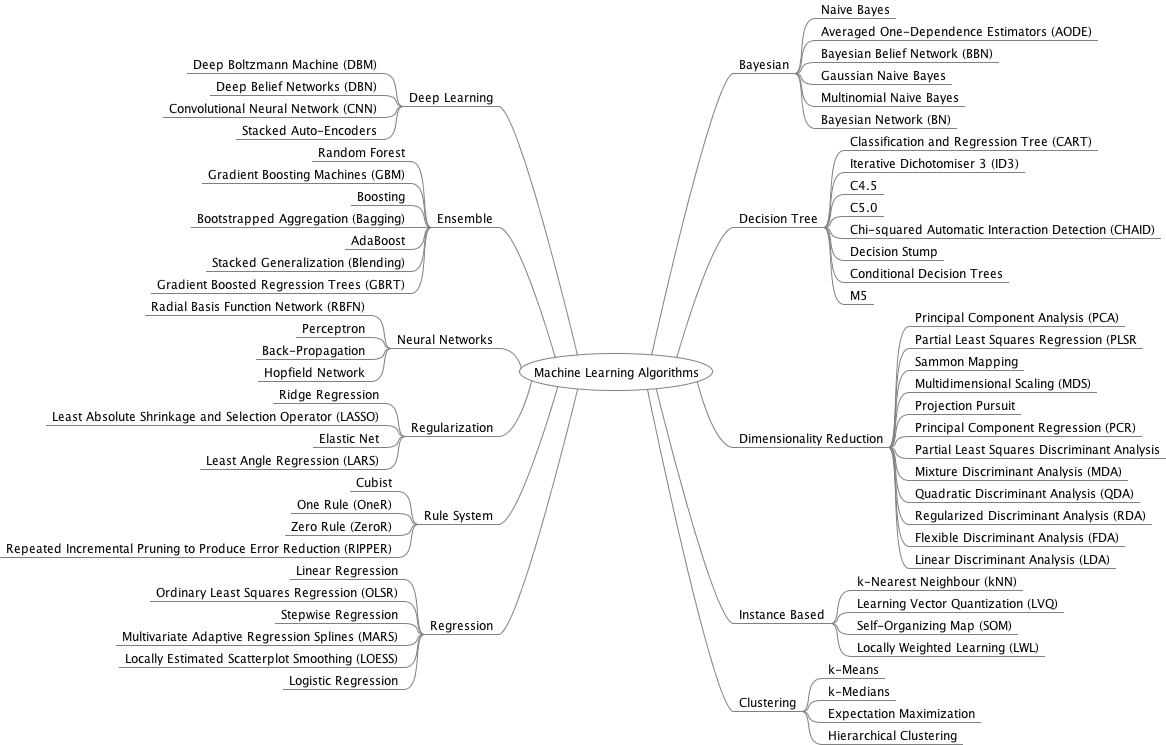
**Inconsistency with test features Vs. feature selection –**

It took us time to stabilize the models as the feature selection showed us different predictions as we expected:

1) All features performed better than the feature selection.

2) Feature selection chose to stop at a low number rather than take all features to get the best results.

3) Results were different between all features results and the feature selection when it took the entire features. (e.g. All features = 90, feature selection that chose all features = 88)

**Machine learning algorithms by groups:**

https://towardsdatascience.com\which-machine-learning-model-to-use-db5fdf37f3dd

1. Creators:

   1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
   2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
   3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
   4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

   [↑](#footnote-ref-1)