Heart Disease Prediction

Project report – Valentin Formont – Autumn 2021

The dataset used for the project is called 'Heart Failure Prediction Dataset' and is available on Kaggle. It gathers 5 different datasets, with 12 variables and a total of 918 unique observations.

There are **11 predictors**: Age(*n)*, Sex(*b)*, Chest Pain Type(*c)*, Resting Blood Pressure(*n)*, Cholesterol(*n)*, Fasting Blood Sugar(*b)*, Resting ECG(*c)*, Maximum Heart Rate(*n)*, Exercice Angina(*b)*, OldPeak(*n)* and ST slope(*c)*.

The **response** is Heart Disease*(b)* – (*b* for binary, *c* for categorical, *n* for numerical)

1. Feature Exploration

First, a correlation matrix of the entire dataset shows no strong correlation between one predictor and the response. The most correlated predictors are OldPeak(+0.40), Maximum HearRate (-0.40) and ST_slope(+0.39) – (strength of association coefficient) .

No strong correlation is found between the predictors, which lessen the risk of multicollinearity.

Every numerical predictor, apart from Old Peak seems to follow a gaussian distribution, with their own means and standard deviation. As a result, no correlation is directly found if the raw features are compared to the response. However, when the response is transformed to be a percentage of the sample population having a HeartDisease per variable value, it leads to more interpretable graphs.

Using data smoothing, trends are exhibited and confirm the correlation values.

Most noticeable elements from Feature Exploration:

- Age (+0.28), Resting Blood Pressure (+0.11) and OldPeak (+0.40) have a positive correlation with HeartDisease;
- Maximum Heart Rate (-0.40) has a negative correlation;
- Most of the time, a heart disease is not accompanied by Chest Pain symptoms;
- Manifesting exercise related angina more than double the odds of heart disease;
- If the ST Slope is not going upward the risk of heart disease goes from 20% to 80% (+400%);
- Resting Blood Pressure contains one outlier (0), and Cholesterol (-0.23) has a negative correlation with heart disease.

2. Preprocessing

After removing the outlier from RestingBP, it is important to deal with the Cholesterol negative correlation with the heart disease. Medical knowledge so far is very confident in the positive correlation between high cholesterol level and cardiac complications. The cholesterol distribution shows 172 values with 0 (171 when the RestingBP outlier is dropped). These values are not possible to achieve for a living human. Over these 171 observations, 88% of them have a 1 value for HeartDisease. To be certain that these are missing values and not categorical values, 0 standing for no high level of cholesterol, the cholesterol values are split into 2 groups. Doing so yields to the following results: more than 60% of the sample population without cholesterol had a heart disease, compared to about 50% of the sample population with cholesterol. The fact that 88% of the 0 cholesterol values have a value of 1 for HeartDisease combined to a high level of cholesterol being healthier leads to treat the 172 values as missing values.

Because these observations contain a lot of useful data, dropping them would be the worst-case scenario. In order to avoid it, different imputing methods are used. A first round of accuracy assessment is performed, using a Random Forrest Classifier prediction accuracy as a score metric. Simple imputers and Iterative imputers from sklearn.impute are compared. In almost 100% of the runs, the 'clean' data 'imputer' has test best accuracy, but the highest standard deviation. This is when the observations with missing values are removed from the dataset.

A second round of accuracy comparison is performed, using custom imputers. Slightly tuned regressors are used as imputers. Because of the stochastic nature of the cross validation, the results vary from one run to another (random_state is not set). The best accuracy score generally does not drop under 0.87.

Once the cholesterol missing values have been imputed or dropped, depending on the best imputer, the data set undergoes different components analysis methods (PCA,FAMD,MCA). The initial number of predictors is 11. Once the categorical data are one hot encoded, this number increases to 20. To reach this number, one binary predictor is added per categorical variables values. For instance, RestingECG has 3 possible values: LVH,Normal and ST. Once encoded, the binary predictors RestingECG_LVH, RestingECG_Normal and RestingECG_ST are added to the dataset.

Linear Principal Components Analysis (PCA) and PCA using *cosine* kernel reach a 100% decomposition explained variance using 15 components, and about 80% with 9 components. Comparing the explained variance with 4 components, Factor Analysis of Mixed Data (47.17%) and Linear PCA (48,78%) get the highest explained variance.

3. Model Exploration

Several classifiers are chosen to perform an accuracy comparison. The estimators are slightly tuned to get a good accuracy in a "small" amount of time. For each classifier a comparison of the scaling impact is performed, as well as an accuracy comparison using different data (PCA, FAMD, KernelPCA and original). Two metrics are used to compare models' performance – Accuracy and ROC-AUC score

Given the stochastic nature of the exercise, the results differ from one run to another. As an example, values from one run will be given. The analysis might differ.

The different classifiers are Logistic Regression (accuracy=0.880 - AUC=0.885), Support Vector Machine (0.880-0.883), K-Nearest Neighbors (0.873-0.873), Naïve Bayes (0.848-0.847), Random Forrest (0.880,0.885) and two different classifiers using Ensemble Methods, AdaBoost (0.880 – 0.888) and Voting Classifier (0.884 – 0.890)

The classifiers using the features generated by PCA underperforms compared to the scaled data for every single classifier, regardless of the number of components used.

However, the ones using FAMD features manages to outperform the original data with about 8 components every time reaching more than 0.90 accuracy for ensemble methods.

The impact of scaling method depends on the method used. No method outperforms the others for every classifier. However, the RobustScaler method (use of median/interquartile values instead of the mean and standard deviation) has the worst accuracy test rate for every classifier.

Overall, ensemble methods combined with FAMD transformation yields the best result, reaching 0.906 accuracy and 0.909 ROC-AUC score on the test set.

Improvement possibilities

This project could yield better results with the following additions:

- Hyperparameter tuning using multivariate optimization. It would lead to a better accuracy, but significantly increase the runtime
- Analyze the impact of predictors scaling methods on components analysis
- Use custom code to include a validation set for imputer optimization (scikitlearn.impute uses train and test set for cross validation, but no validation which can result in information leakage)