Understanding Coronary Artery Disease causes with machine learning techniques

Martins Hugo

University of Tsukuba's internship

June 10, 2024

Table of Contents

1 Problem

2 Project's work presentation

3 Conclusion

Problem

General Point Of View:

- Coronary Artery Disease (虚血性心疾患（狭心症・心筋梗塞)) is very common (18M adults in the US)
- Choices made in everyday life add up to a big portion of the risk of CAD, there are many risk factors.

Medical Point Of View

- How to better diagnose, predict, and prevent the disease.
- What are the factors for becoming healthy again.

Goal

The aim of this project will be to:

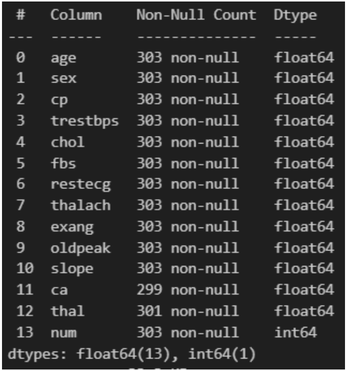
- try to improve the understanding of the disease, in particular the causes of it
- compare the results obtain through this project to the already existing medical knowledge

To achieve this, we will use machine learning techniques:

- the model's own understanding and classification of the features is knowledge
- we will use an appropriate dataset to train and test our model with

Basis of the work: the dataset

The dataset is issued from the UCI ML Repository

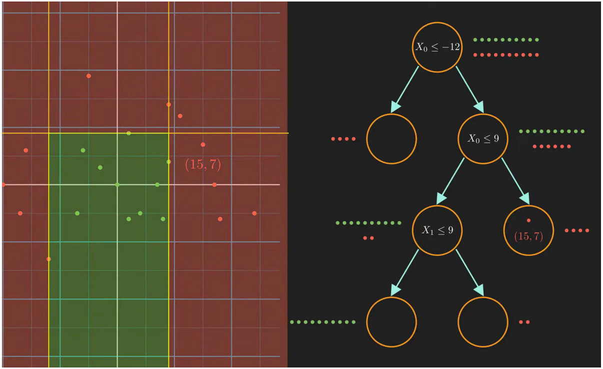


```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    float64
 1   sex       303 non-null    float64
 2   cp        303 non-null    float64
 3   trestbps  303 non-null    float64
 4   chol      303 non-null    float64
 5   fbs       303 non-null    float64
 6   restecg   303 non-null    float64
 7   thalach   303 non-null    float64
 8   exang     303 non-null    float64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    float64
 11  ca        299 non-null    float64
 12  thal      301 non-null    float64
 13  num       303 non-null    int64
dtypes: float64(13), int64(1)
```

- CP: Chest Pain type
- FBS: Fasting Blood Sugar
- Restecg: Resting Electrocardiographic data
- Thalach: maximum heart rate achieved
- CA: number of major vessels
- Thal: blood disorder called Thalassemia
- num: heart disease

Basis of the work: the random forest

Predicting if a patient is healthy allows to see how the model uses the features passed as parameters. A random forest is a machine learning model composed of a large number of decision trees.



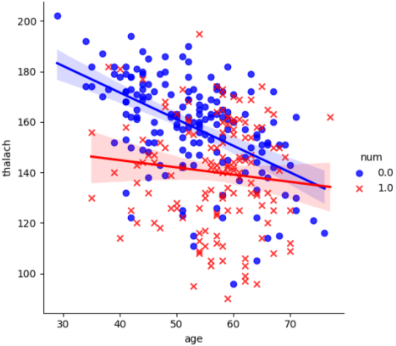
Core methods: Features importance and SHAP Values

Features importance consists in classifying the features of the model by importance

Name	**Description**	**Benefit**	**Problem**
Random Forest's own tool: feature importance	It's possible to display the corresponding graph directly with the random forest's own tools.	The function comes with the same library as the model.	We can't explain the ranking.
SHAP values: summary plot	SHAP values also allow the creation of such graphs, but those are much more detailed and precise. It's based on cooperative game theory	More detailed, the ranking is intelligible and can be explained with SHAP elements.	We need to import the library, and to start working with it

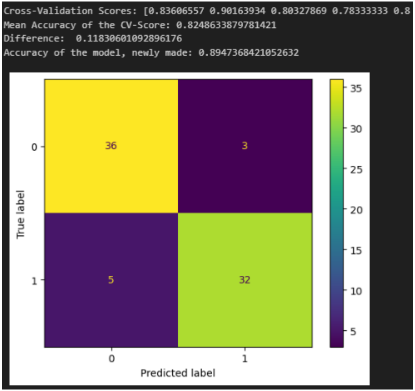
Core methods: Data visualization

Consists in displaying graphs, diagrams, basically using graphical means to represent data; this method was used to gather informations concerning some peculiarities or intriguing aspects of the data.
Below, a graph showing the relationship between thalach and age:

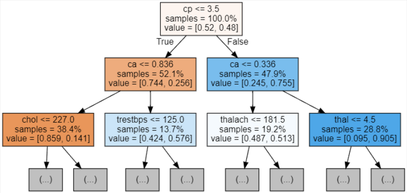


Conducted experiments: Random Forest and features importance

The random forest is created with a 750 trees, each with a maximum depth of 7; the data from the dataframe is separated into two sets, one for training and one for testing. Some graphical results:



```
Cross-Validation Scores: [0.83606557 0.90163934 0.80327869 0.78333333 0.8
Mean Accuracy of the CV-Score: 0.8248633879781421
Difference:  0.11830601092896176
Accuracy of the model, newly made: 0.8947368421052632
```



cp <= 3.5
samples = 100.0%
value = [0.52, 0.48]

True

False

ca <= 0.836
samples = 52.1%
value = [0.744, 0.256]

ca <= 0.336
samples = 47.9%
value = [0.245, 0.755]

chol <= 227.0
samples = 38.4%
value = [0.859, 0.141]

trestbps <= 125.0
samples = 13.7%
value = [0.424, 0.576]

thalach <= 181.5
samples = 19.2%
value = [0.487, 0.513]

thal <= 4.5
samples = 28.8%
value = [0.095, 0.905]

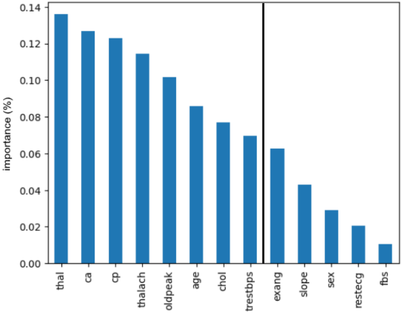
(...) (...) (...) (...) (...) (...) (...) (...)

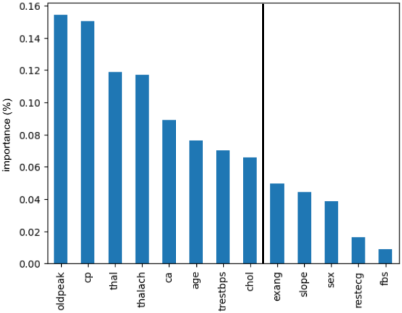
Decision tree visualization

$< -$ CV-score, model accuracy and CM

Conducted experiments: Random Forest and features importance

Here are 2 graphs, caused by different training set:

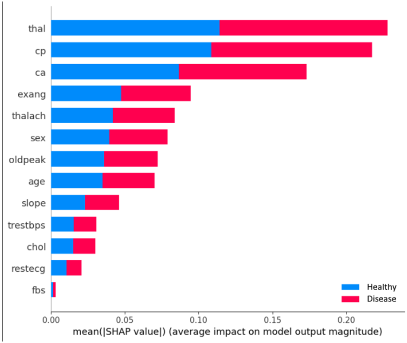


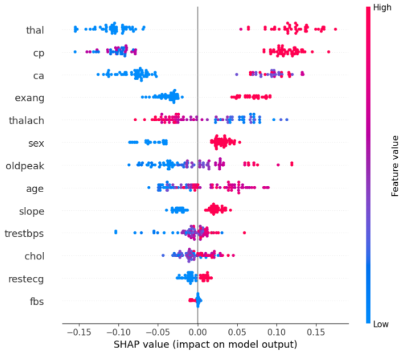


Some similarities (right part) but nothing stands out: it's inefficient.

Conducted experiments: SHAP Values to understand

Better distinction, easier to understand, additional informations

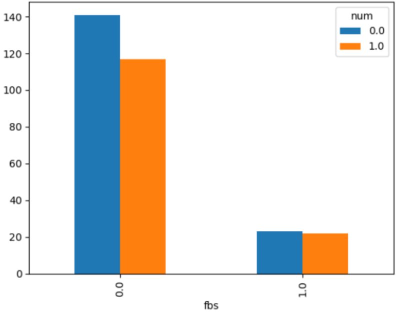


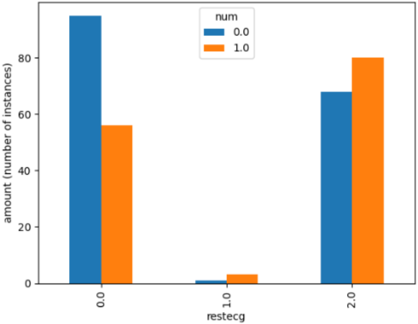


The bigger the shap value (or the gap in the middle for the second graph), the bigger the impact of the feature

Conducted experiments: Data Visualization and comparisons

Most useless features: fbs and restecg

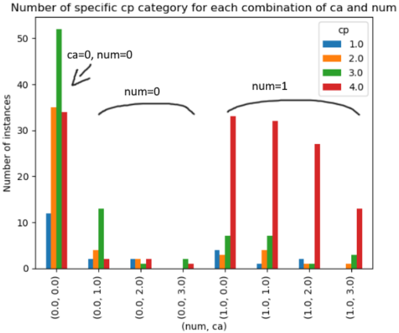


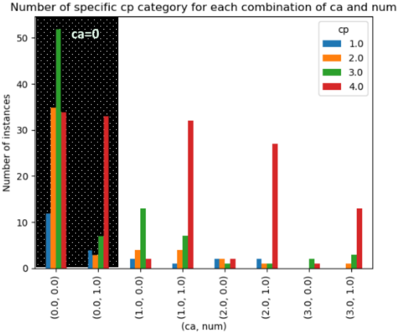


Here, fbs shows basically no direct link and restecg has a sampling size issue

Conducted experiments: Data Visualization and comparisons

Important features graph: cp comparison with ca and num

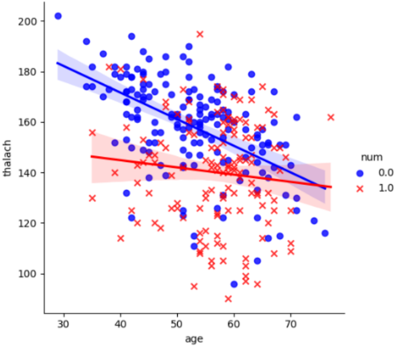




$ca = 0$ and $cp \neq 4$ often means healthy

Discussion/Other details

The correlation with thalach and age, being the possible cause of a lack of importance



thalach goes down as age goes up, meaning that it is possible to predict one with the other and have a fair chance of success

Conclusion

- Coronary Artery Disease: the most common heart condition, and leading first cause of death in the US
- Machine Learning techniques and data visualization

Possible future endeavors and improvements

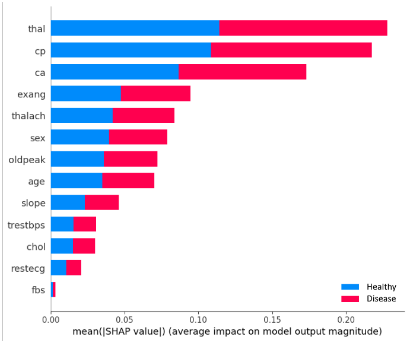
- Collecting more samples for the features with sampling size issues, and more samples overall
- Trying the same experience with different datasets (particularly from different clinics) containing either the same features or different ones.
- trying new ML techniques or trying different ML models

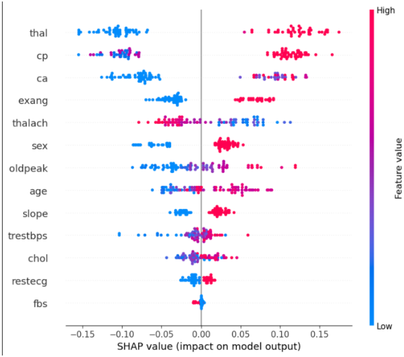
Thank you for listening

If you have any questions, I will be glad to answer them!

Potential question concerning the results

- being an elderly male: those are 2 factors that were mentionned in the book *Be Heart Smart: Understand, Treat and Prevent Coronary Heart Disease (CHD) by Waqar Khan*
- new knowledge (it was less mentionned, if at all) : Thalassemia (thal) and the number of major vessels (ca)


thal
cp
ca
exang
thalach
sex
oldpeak
age
slope
trestbps
chol
restecg
fbs
Healthy
Disease
0.00
0.05
0.10
0.15
0.20
mean(|SHAP value|) (average impact on model output magnitude)


thal
cp
ca
exang
thalach
sex
oldpeak
age
slope
trestbps
chol
restecg
fbs
SHAP value (impact on model output)
High
Feature value
Low