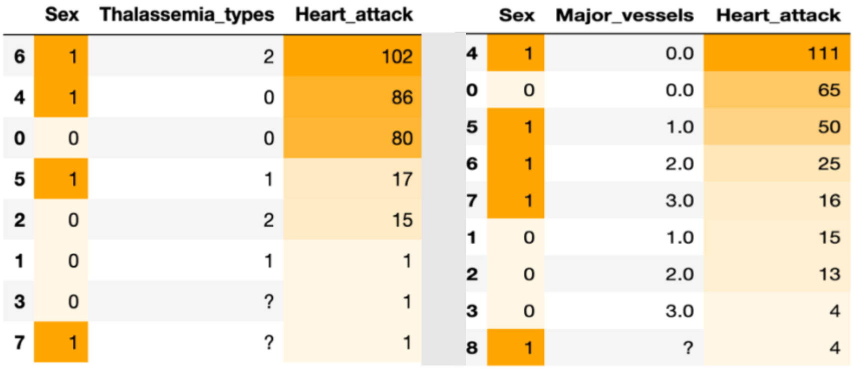
1. Introduction

Heart disease describes a range of conditions that affect our heart. Heart disease is one of the biggest causes of mortality among the population of the world. For this reason, prediction of cardiovascular disease is considered as one of the most important case studies in the field of clinical data analysis.

2. Case Study and EDA

The objective of this study is to build a model to predict if a patient is likely to have a heart attack. The Cleveland dataset containing 303 patients and 14 features (both categorical and numerical features) is used for this scope. We are dealing with a balanced dataset (54 % of healthy patients and 46% of unhealthy patients). The data looks quite clean, exept for some missing values for Major_vessels (number of major vessels 0–3 colored by flourosopy) and Thalassemia_types (displaying the thalassemia) features; since the number of missing values is quite small, the mode value is used to fill the missing data for both the features.



	Sex	Thalassemia_types	Heart_attack
6	1	2	102
4	1	0	86
0	0	0	80
5	1	1	17
2	0	2	15
1	0	1	1
3	0	?	1
7	1	?	1

	Sex	Major_vessels	Heart_attack
4	1	0.0	111
0	0	0.0	65
5	1	1.0	50
6	1	2.0	25
7	1	3.0	16
1	0	1.0	15
2	0	2.0	13
3	0	3.0	4
8	1	?	4

Table1

Table1 shows that as the number of major blood vessels increases, the number of heart disease cases decreases. Also, most of the heart attack cases present reversable defect thalassemia type.

3. Algorithm Definition

For this particular task, I have decided to choose a decision tree (DT) algorithm. DT algorithms work quite well for classification tasks, easy to interpret and explain. The idea behind those algorithms is to represent the data with a tree structure which has per nodes the features, per branch a rule, and per leaf an outcome.

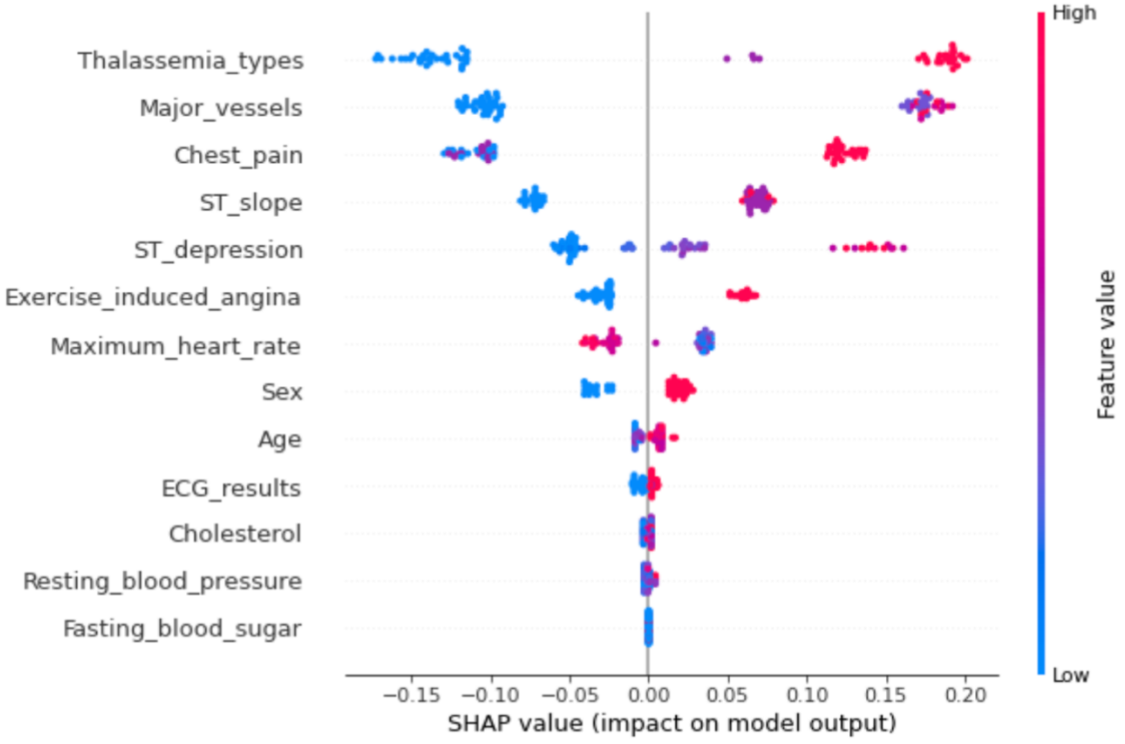
One of the disadvantages of DT algorithms is the risk of overfitting the data (the model learns really well how to represent the training set but it performs poorly on unseen data). Some pruning techniques (e.g. reduced error pruning approach) can be used to limit this phenomenon. Nevertheless, no particularly techniques are used in this analysis for addressing the problem.

Since we are dealing with a balanced dataset (54% vs 46%), I have not proceeded with the Random Forest algorithm which combines many independent decision trees (for this reason it is commonly used for unbalanced dataset), trains each one on a slightly different sent of the observations with limited number of features, and then averages the predictions of each individual tree. I have opted for a gradient boosting algorithm which converts weak learners (decision trees) into robust learners by fitting each tree on a modified version of the original dataset. It is called gradient algorithm because uses the gradient to identify the shortcomings of weak learners. In particular, I have chosen the CatBoostClassifier, an algorithm for gradient boosting on decision trees. High performance, handling categorical features automatically (this is why not much data preprocessing is done on this dataset), robust and easy to use are some of the advantages of this algorithm.

4. Results and Discussion

To evaluate our model, I mainly look at the ROC (Receiver Operating Characteristic) metric which plots the TPR (true positive rate) vs the FPR (false positive rate). The AUC (Area under the curve) reaches the value of 0.93 on the test set indicating a well robust trained model.

What we are mainly interested in this study is how the features influence the predictions. For such purpose, we refer to Shap (Shapley Additive exPlanations), a great tool to explain the output of a ML model.



Plot 1

Plot 1 shows on the Y-axis the most important features in descending order; the X-axis informs us on how the values of a feature affect the outcome, and the intensity of the colour bar on the right side depicts a red colour for high values and a blue colour for low values of a feature.

Features like ECG_results, Resting_blood_pressure, Cholesterol, and Fasting_blood_sugar do not seem to affect the outcome. Features like Thalassemia_tuypes and Major_vessels have a significant impact in determining if a patient has a risk of getting a heart attack or not as we have previously seen (Table1). A strong relationship there is also between high values of St_depression (induced by exercise relative to rest), ST_slope (flat/downsloping), Exercise_induced_angina (exercise induced angina) and the probability to get a heart attack.