Sample size and preprocessing:

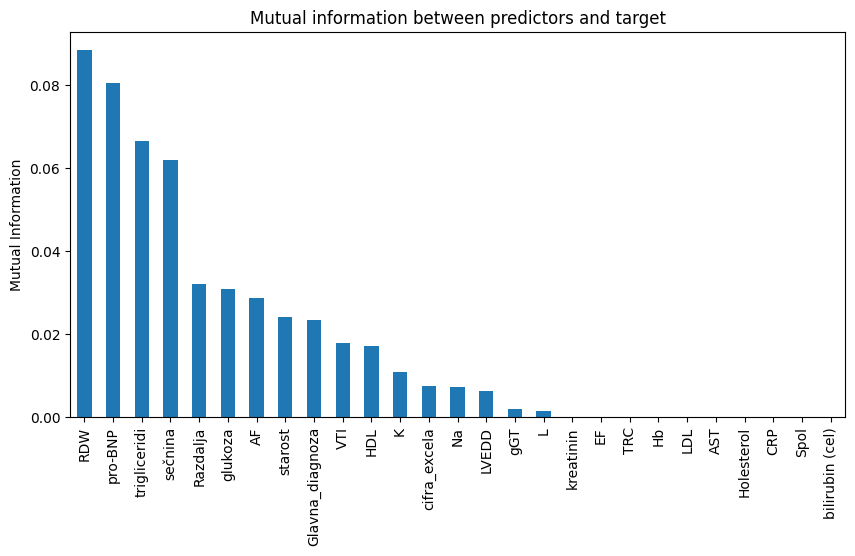
Original dataset included 240 patients with 39 predictor variables. After preprocessing the data (we excluded variables and patients where more than 30% of data was missing), the final cohort included 196 patients and 27 variables. The final class balance of responders to non responders was 93 ([47.4%] non responder) : 103 (responder).

Outcome:

The outcome of interest was a binary variable »responder« indicating whether the patient treatment was successful or not. We implemented 11 machine learning (ML) algorithms and evaluated their performance with a 5-fold cross-validation with stratified folds with respect to the outcome.

Variable importance:

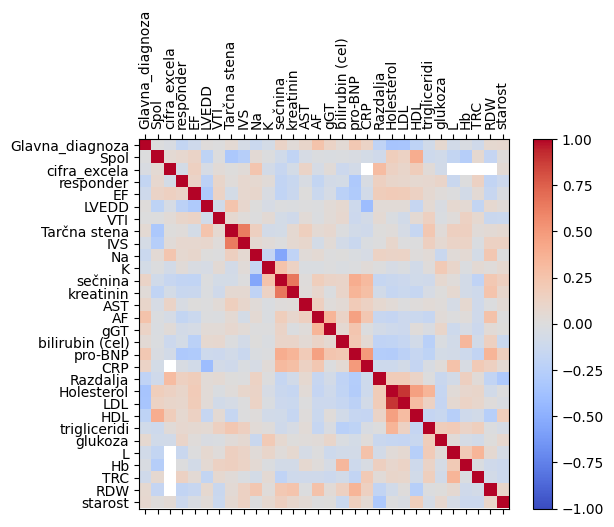
For variable importance we chose mutual information (MI) where we obtain the following plot with the MI of each feature and the outcome. The higher the mutual information, the better the correlation between the variable and the outcome. Mutual information of 0 means that there is no correlation. Note that the mutual information does not measure if the correlation between the variable and the outcome is positive or negative, it just measures the strenght of the correlation. On the graph below, we see that 4 of the variables are more importnat than the others (top 4), and some of the variables have mutual information score of 0.



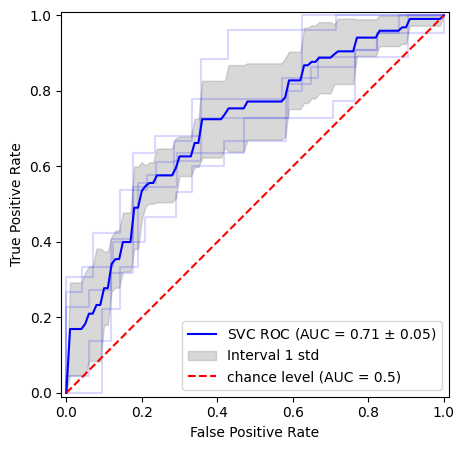
The correlation matrix between the variables can be seen below. In this case, we can see if the correlation is negative or positive. If correlation between A and B is negative, it means that the higher the value of A, the lower the value of B.

Some top correlations:

* »Tarčna stena« and »IVS«
* »kreatin« and »sečnina«
* »holesterol« and »LDL«
* »pro-BNP« and »CRP«

Machine learning algorithms and model performance:

Models were evaluated with a 5-fold cross-validation with stratified folds. We compared the models using accuracy, precision, recall, and area under the receiver operating characteristic curve (ROC AUC). Best performing model according to accuracy was KNeighbors (accuracy mean score 0.65) and the best performing model according to ROC AUC was SVC (roc auc mean score 0.72). ROC curve of the latter model can be found below:



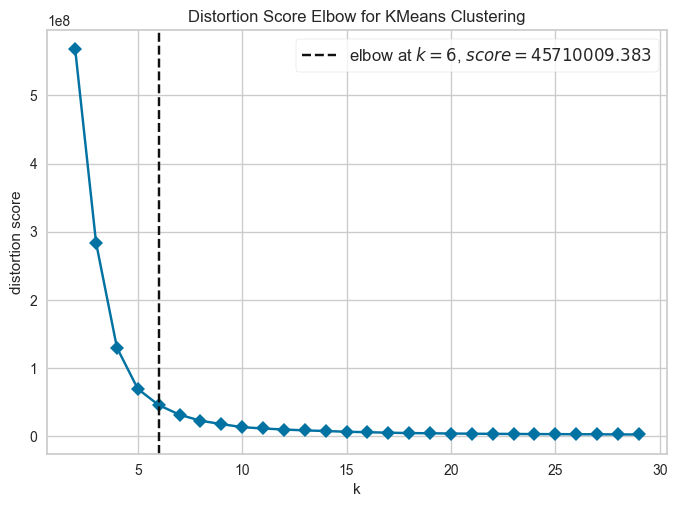
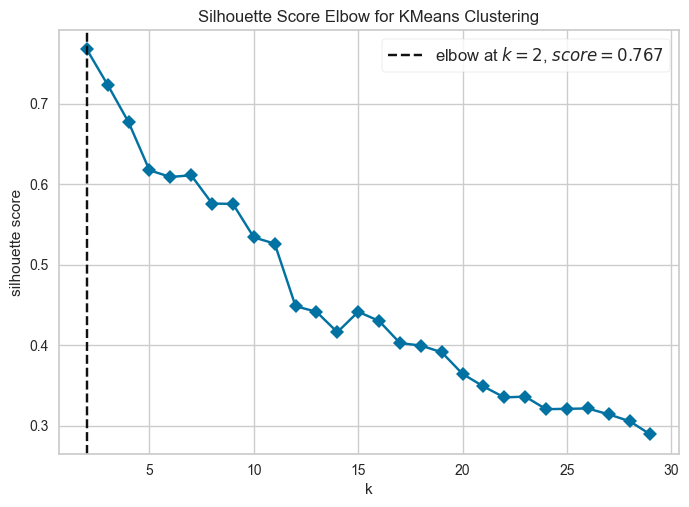
Confusion matrix of SVC is

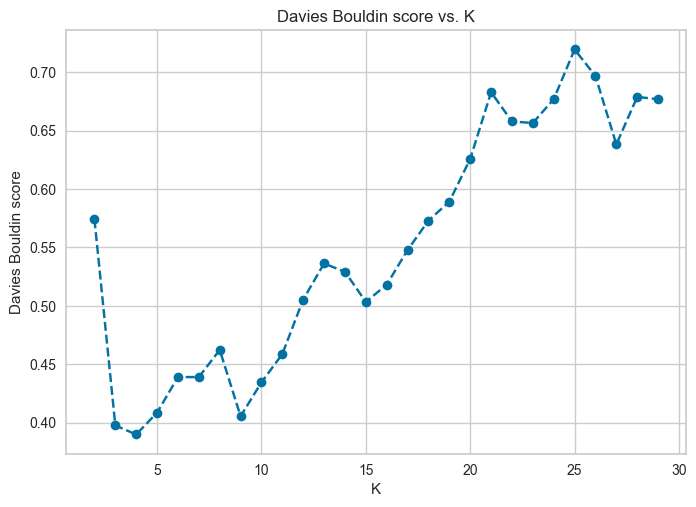
|  |  |
| --- | --- |
| 44 (true negative) | 49 (false positive) |
| 25 (false negative) | 78 (true positive) |

and for KNeighbors

|  |  |
| --- | --- |
| 59 | 34 |
| 21 | 82 |

Clustering was also performed. Task was to group our set of patients in such a way that objects in the same group are more similar (in some sense) to each other than to those in other groups. We opted for the elbow method using 3 metrices: silhouette score, distortion score, and Davies-Bouldin index. Using the elbow method we tried to find the optimal number of clusters (more than one cluster per class), but the results were inconsistent across different clustering metrics (k=2 for silhouette score, k=6 for distortion score, and k = 4 for Davies-Bouildin score.

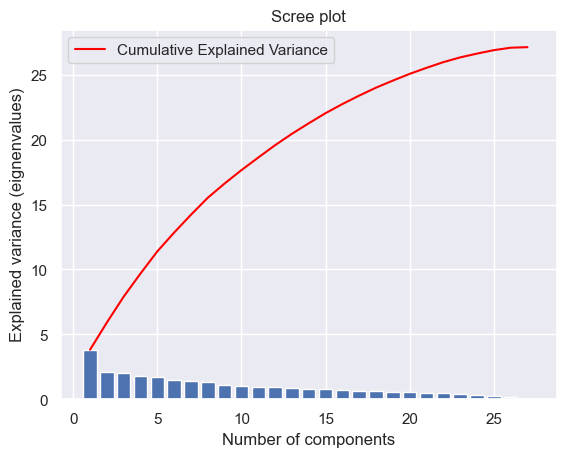


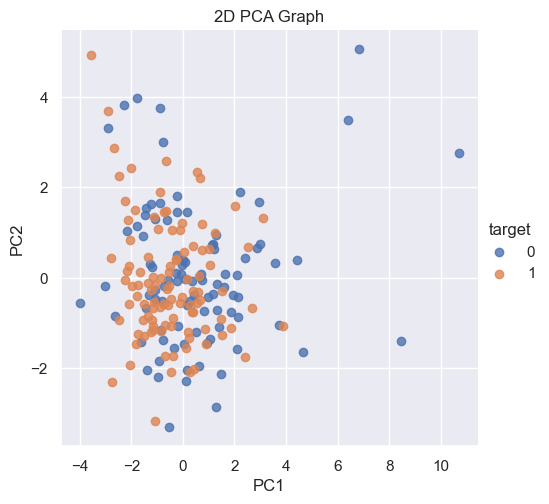


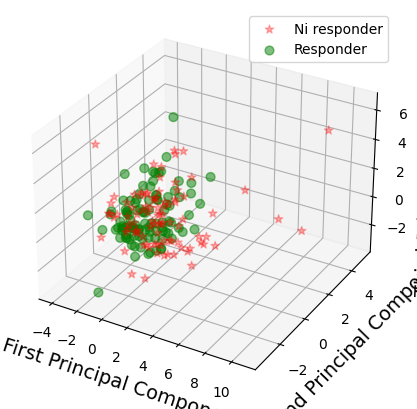
Dimensionality reduction for the purpose of visualization:

For the purpose of visualization of the data, we implemented two dimensionality reduction techniques, hoping to identify and discard features that are less useful to our cause. A dimensionality reduction techniques are used for reducing the dimensionality of data. This makes it easy to plot the data on 2D or 3D graphs. The dimensionality reduction reduces the number of variables by transforming them into into 2 (or 3) variables in a way that preserves as much information as possible. After plotting the reduced data, no real pattern was seen, which also explains the unclear clustering results of the previous step.

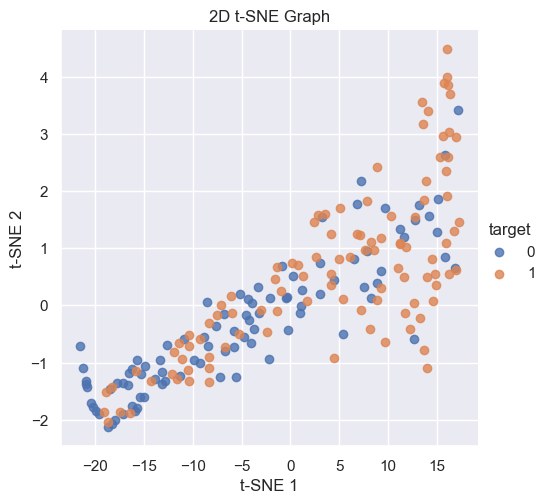
PCA:

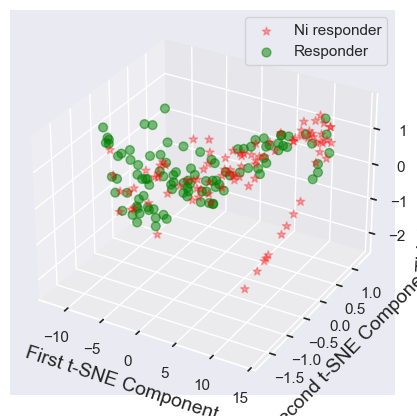






t-SNE





Questions for UKC

* What would be our baseline for evaluating whether models are successful or not? Generally, the dummy baseline is a prediction model that always predicts the average of the seen (training) data, or the majority in this case, as we have a classification problem. Can we construct a simple baseline that is a bit more meaningfull than just taking the majority? Can a doctor with his medical expertise make a good (or even better) prediction judging solely on the data/variables?
* Are certain biomarkers correlated? Furthermore, are there any features in general that are/should be correlated?