Technical Report 1:
Heart Failure Risk Prediction in the Smurf Society

Brainy Smurf Doctor Smurf Papa Smurf

With the rise of cardiovascular disease among Smurfs and the destruction of the magical crystal ball, there is a pressing need for new predictive tools. Leveraging medical data collected by Doctor Smurf, we developed a pipeline based on linear regression to predict heart failure risk. In this technical report, we describe how we prepare the data, we detail the pipeline with its feature selection step, and investigate model error distribution.

Data pre-processing We split the data into a training set (1000 samples) and a test set (500 samples). The same pre-processing steps are applied to both sets : we get rid of the `"img_filename"` variable, we use label encoding to transform all categorical variables into a numerical form, and finally, we standardize the data (mean and standard deviation are computed only on the training set).

Feature selection We compute the Pearson correlation coefficient between each feature and the target. We keep the 7 features with the highest correlation coefficient in absolute value (Papa Smurf insisted that we keep a least half the features). Among these features, we can find "Blood pressure", "cholesterol" and "weight".

Model selection and implementation We use the linear regression model from the python library `scikit-learn`. There are no parameters to tune.

Results analysis We obtain a RMSE of 0.08 on the test set. To help us gain more insights, we investigate error distribution by plotting true target values against predicted values (cf. Fig 1).

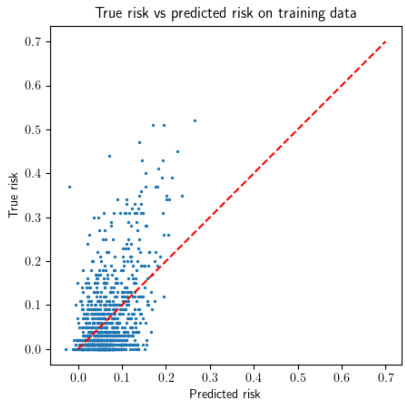

Figure 1: True risk versus Predicted risk. Red dotted line indicates perfect predictions.

We do not really know how to interpret all this but it does not seem so good... Further works will surely need to involve non-linear models and a way to integrate the heart scans.