Technical Report 2:
Heart Failure Risk Prediction in the Smurf Society

Brainy Smurf   Doctor Smurf   Papa Smurf

In this technical report, we describe how we predict the risk of heart failure using a nonlinear model. We improve on our previous pipeline by replacing the linear regressor by a support vector machine (SVM) regressor. In addition, we combine the feature selection step and the model selection step for a more robust approach.

Data pre-processing We use the same pre-processing steps as in our previous work. Note that the encoding labels for the ”profession” variable are assigned randomly, while the encoding labels for the 3 other categorical variables are ordered according to the progression ”very low - low - average - high - very high”.

Feature selection We select the k features which are the most correlated with the target. The number k is treated as a hyper-parameter of the model and is thus tuned together with other hyper-parameters in the model selection step.

Model selection and implementation We use the `SVR` model from the python library `scikit-learn`. We choose to tune the following hyper-parameters: the kernel type, the regularization coefficient C, and the margin of tolerance ϵ (no penalty in the loss if error is smaller than ϵ). In order to find the best values for this set of hyper-parameters, we use a grid search with a 5-fold cross validation . We define the search space in Table 1.

Hyper-parameter	**Allowed values**
k	5, 7, 9, 11, 13
kernel	”poly”, ”rbf”, ”sigmoid”
C	0.01, 0.05, 0.1, 0.5, 1
ϵ	0.01, 0.05, 0.1, 0.5, 1

Table 1: Search space for the grid search.

The best parameters are $\{k$: 9, kernel: ”rbf”, C: 0.1, ϵ: 0.05$\}$. The mean validation RSME for the best model is 0.072, with a standard deviation of 0.0053. Results were consistent across runs, albeit small variations for the value of C and/or ϵ.

Results analysis We retrain a model with the best set of hyper-parameters on the whole training set and then obtain a RMSE of 0.073 on the test set. There are no signs of over-fitting. However, the improvement over linear models is much less than expected. We should really get to work with the heart scans, they must contain some useful information!