# LELEC2870 Machine Learning Project: Report 1

Baptiste Pierlot
28082000

Group 4

Cyril Bousmar
63401800

## I. Introduction

This report follows the previously made Technical Report (Brainy Smurf et al., 2024) on Heart Failure Risk Prediction in the Smurf Society using a *Linear Regression (LR)* model. Section II highlights the flaws within the previous method. Section III provides taken considerations to better improve the prediction and presents their results. Section IV summarizes and discusses our findings and suggests possible future work.

## II. Previous Pipeline Analysis

Arbitrarily choosing the number of features leads here to under-performing system. But *Pearson* correlation performs very well in this case as seen in Section III-A, and is a good choice.

The lack of expressiveness from the profession feature impacts negatively the precision, and `LabelEncoder` transformer should only be use on target and not input data as prescribed by the library.

There is a risk of over-fitting the model as the data is not splitted into training, validation, and test sets. Moreover, not trying to use another model despite poor results is dangerous as patients predictions will be off.

## III. Methods and Results

Python library `scikit-learn` is used extensively for implementation. Previous LR model from (Brainy Smurf et al., 2024) and its *Root Mean Squared Error (RMSE)* metric at $0.08$ are used as benchmark for all improvement methods. *KFold* and *RepeatedKFold* (10 repeats) cross-validators (5 folds := 80-20% training/validation sets in our case) are used in Sections III-B & III-C for robust results, to prevent over-fitting, and to allow hyper-parameters fine tuning.

### A. Pre-processing

Features can be separated into 3 categories: normally distributed, ordinal, and professions. Professions contains qualitative information and thus is transformed into multiple binary quantitative features through *One-Hot Encoding*. *Ordinal Encoding* is applied to ordinal features to express them quantitatively. Finally, `StandardScaler` is applied to standardize all the data. Other scalers and norms were considered but had no impact difference. The *Z-score* method is used to measure the impact of outliers (the lower, the better). Cholesterol, relatively impacting feature (cf. Figure 3) had a big outlier (index 992). Switching its value to the median had a very bad result but removing it improved the RMSE from $0.0777$ to $0.0776$ and cleaned the distribution.

Due to better expressiveness of feature's nature (as shown in figure 2), *Z-score* dropped from $3.0\%$ to $2.0\%$, and benchmark RMSE, Section III, drops to $0.0776$.

### B. Feature selection

Multiple methods where considered as possible candidate to provide the best set of features: *Correlation Filtering (CF)*, *Mutual Information*, *Maximum Relevance and Minimum Redundancy*, *Forward Wrapper*, *Backward Wrapper*, $n$ most important features from *Decision Tree*, and *Recursive Feature Elimination*. Alongside RMSE, *Mean Absolute Error* ensures the model is not subject to RMSE bias due to outliers, and $R^2$ regression score function evaluates model's performance.

Measurements in Table I show that CF is the best selection method. It drops Section III-A RMSE to $0.0774$, and improves the model expressiveness with $R^2$ from $0.3110$ to $0.3156$. The selection of $11^{th}$ first features in Figure 3 has the best result.

### C. Model selection

We explored several linear models provided by sklearn, namely: LR, Lasso, Ridge, ElasticNet and Poisson Regressor. To evaluate these models, we applied cross-validation, measuring the performance of each model using the RMSE metric on the different folds.

Once the best model had been identified, we tested it on the final test set to validate its performance. To complement this evaluation, we used GridSearchCV to optimize the hyper-parameters of regularized models (such as Ridge, Lasso and ElasticNet). This approach enabled us to test different values of the regularization parameters (alpha, l1_ratio, etc.), guaranteeing an optimal compromise between bias and variance.

At the end of these analyses, we concluded that the Ridge model with its default hyperparameters performed best in our case. It enabled us to achieve an RMSE of $0.0770$, representing an improvement of 0.003 over the basic model proposed by Smurfs.

## IV. Discussion and conclusions

While this study achieved some improvement in heart failure prediction by addressing previous shortcomings, the modest gains in accuracy (RMSE reduction from $0.08$ to $0.0770$), the negative predictive values and Figure 4 show that it remains suboptimal.

Investigating nonlinear feature transformations like *Polynomial Feature Expansion*, methods like *SMOTE*, or advanced models like *Random Forest* or *Neural Networks* is worth trying to achieve more robust and actionable predictions.

# Appendix

TABLE I
Feature selection metrics results

| Selection | RMSE | MAE | $R^2$ |
|---|---|---|---|
| Correlation Filtering | **0.077385** | **0.056731** | **0.315622** |
| Mutual Information | 0.077626 | 0.056826 | 0.311351 |
| Maximum Relevance and Minimum Redundancy | 0.077655 | 0.056826 | 0.310824 |
| Forward Wrapper | 0.077596 | 0.057068 | 0.311880 |
| Backward Wrapper | 0.077596 | 0.057068 | 0.311880 |
| Decision Tree | 0.077627 | 0.056825 | 0.311337 |
| Recursive Feature Elimination | 0.077558 | 0.056857 | 0.312549 |

Fig. 1. Correlation matrix showing how correlated or not each feature is in regards of another one (diagonal being itself). No feature have noticeably high correlation, even weight and smurfin donuts.

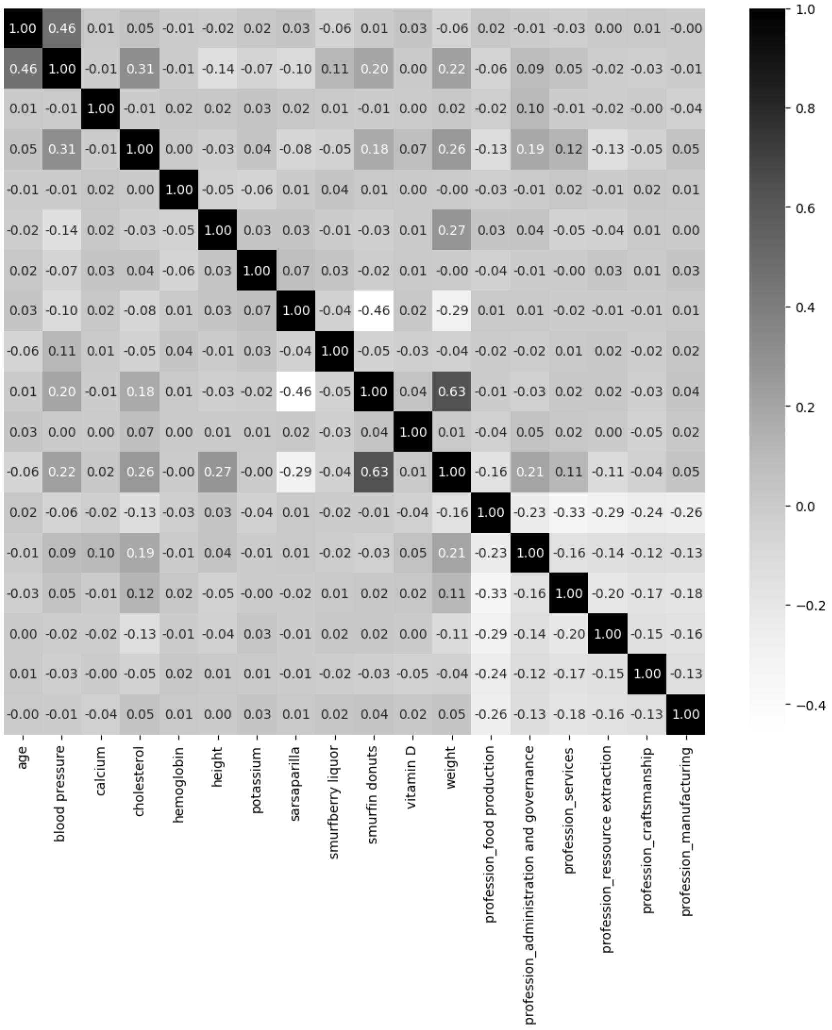


Fig. 2. Distribution for each feature after transformation but before standardization. Professions are splitted in binary features, ordinal features are quantified, and the outlier in cholesterol is removed.

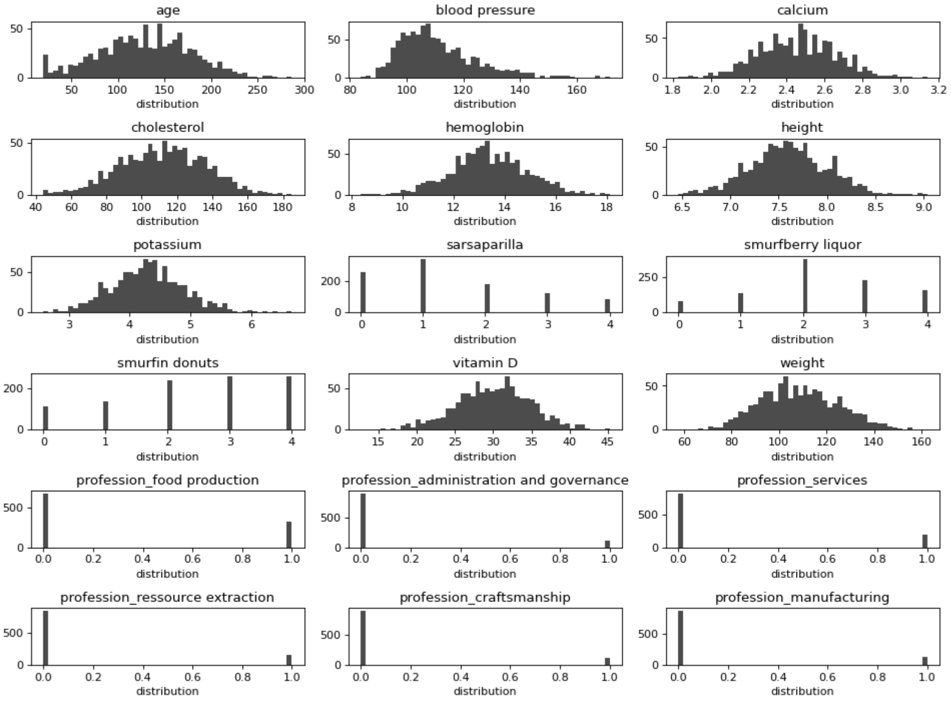


Fig. 3. Correlation between features and target after transformation and standardization of all features. Blood pressure as significantly more impact than any other feature.

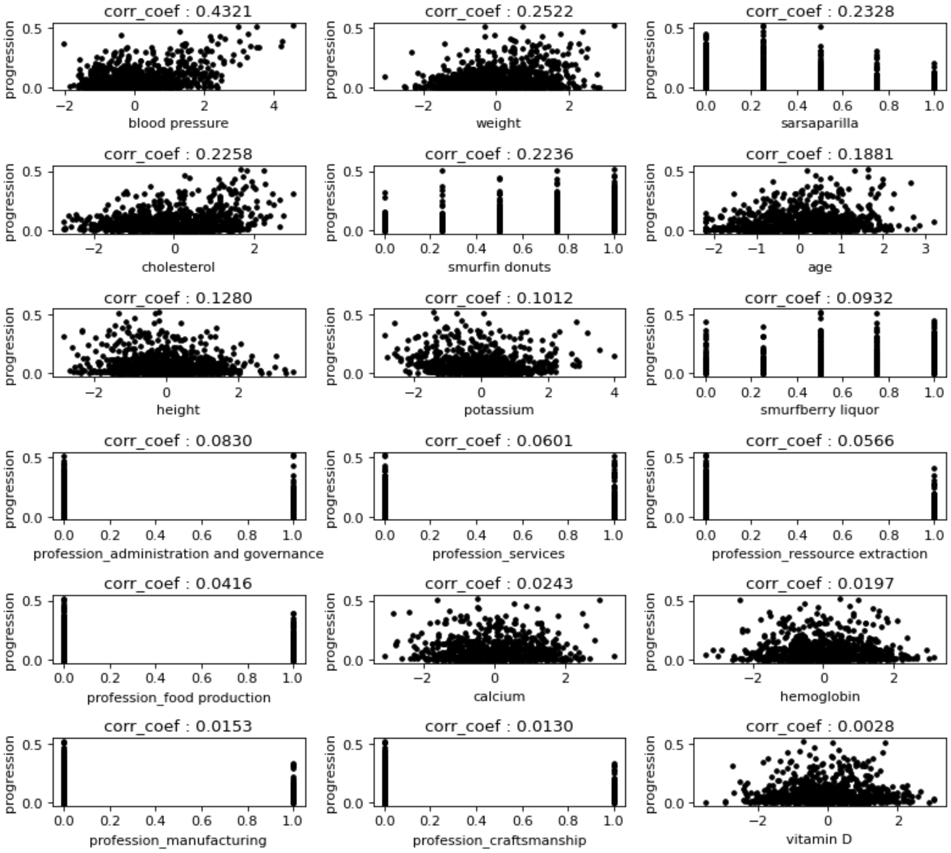
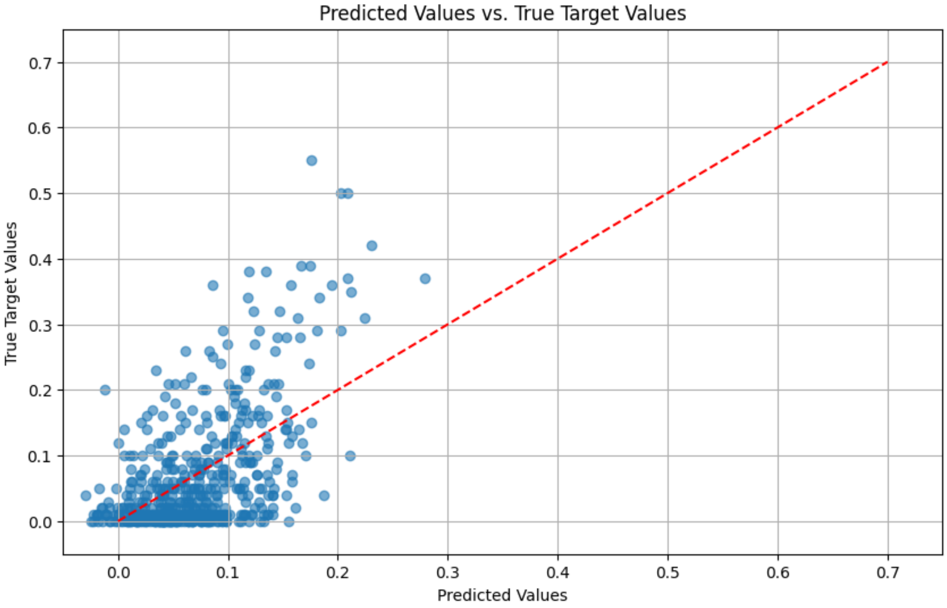


Fig. 4. Comparisons of the new model predictions and to the ground truth values using the test set.


Predicted Values vs. True Target Values
True Target Values
Predicted Values