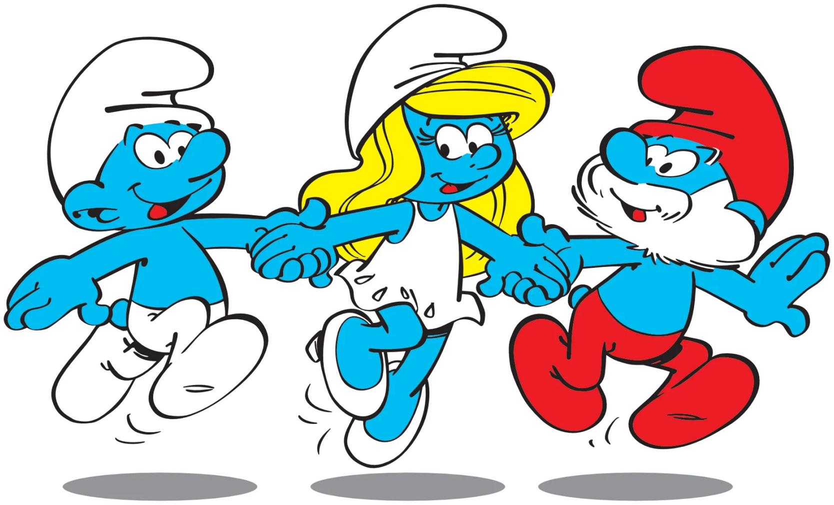
UCLouvain

LELEC2870

Heart failure



© Peyo

Brieux Kaczmarczyk	6444-17-00
Laura Vermeren	3335-19-00

1 Data Engineering

We initiate this project by creating boxplots to illustrate the distribution of variables.

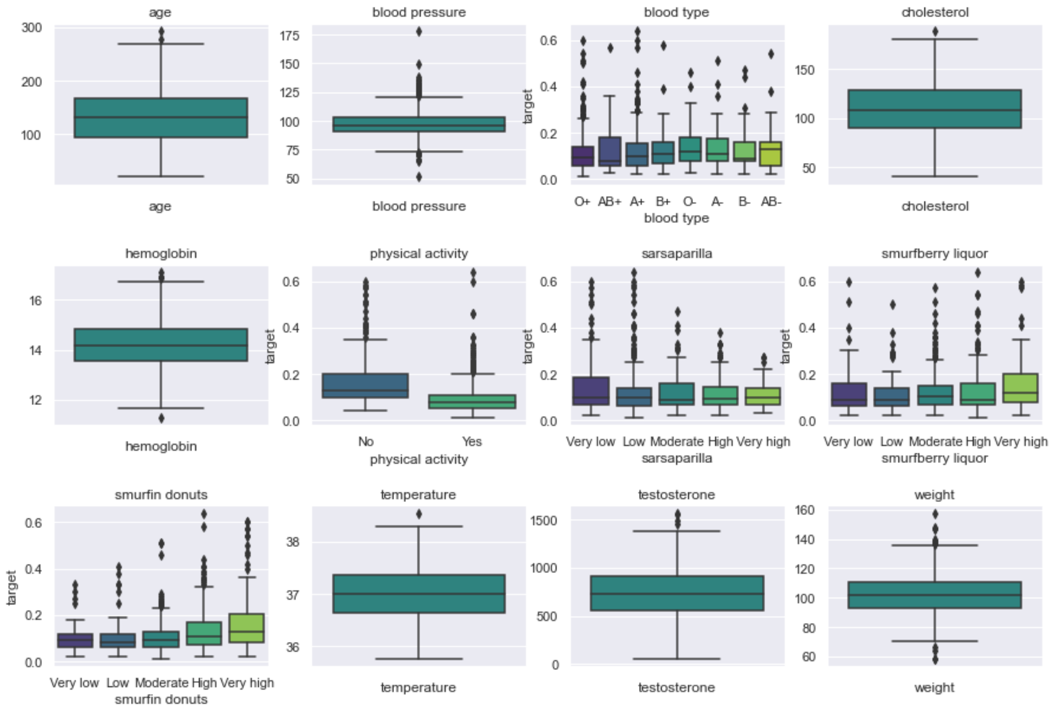


FIGURE 1 – Data description

In the boxplot of the categorical variable, the variables *Blood type* and *Sarsaparilla* do not appear to exert any influence on the target variable. Conversely, the boxplots of *Smurfberry liquor* and *Smurfin Donuts* show that higher consumption is associated with an increased risk of heart failure. Additionally, the graph illustrating *Physical activity* indicates that engaging in sports significantly reduces the risk of heart failure.

We validate these visual observations through ANOVA analyses on the categorical variables and obtain the results in Table 1. The null hypothesis of ANOVA was not rejected for *Blood type* and *Sarsaparilla*. We can conclude that the means of these variables are the same for

One way ANOVA			
Dependent variable	Independant variable	F - statistics	P - value
Risk of heart failure (Y1)	Blood type	0.452	0.869
Risk of heart failure (Y1)	Physical activity	119.361	<0.001
Risk of heart failure (Y1)	Sarsaparilla	1.627	0.165
Risk of heart failure (Y1)	Smurfin liquor	3.315	0.014
Risk of heart failure (Y1)	Smurfin donuts	12.344	<0.001

TABLE 1 – Results of One Way Anova

all categories of the variables, implying that the risk of heart failure is the same regardless of the variable category. Therefore, we decide to remove those variables for the rest of the project.

Note : the conditions for performing an ANOVA have been verified : see appendix.

1.1 Treatment of categorical variables

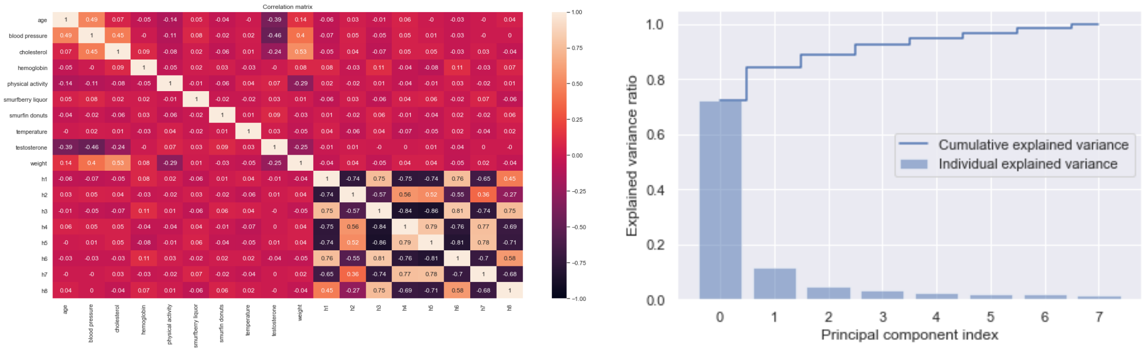
Since no information were provided concerning the levels of consumption, (e.g. *Very low* is four times smaller than *Low*) the label *Very low* is set to 1, *Low* to 2, and so on until *Very high* is set to 5. For physical activities, *No* corresponds to 1 and *Yes* to 2.

1.2 Missing value and Outliers

The dataset is complete with no missing values. Regarding outliers, some are visible in the box plots in Fig. 1 (a table of the percentiles for each variable can be found in the appendix). Due to the absence of information on data collection methods or data dispersion among Smurfs, it is challenging to determine whether these values are encoding errors or legitimate extreme data points. Consequently, we choose not to remove any values for our analyses.

1.3 Feature transformation

We have not yet discussed features derived from the heart scans. In this section, we examine the correlation among all variables, as depicted in Figure 2 (a) [1].



(a) Correlation matrix

(b) PCA on the h_i

FIGURE 2 – Feature transformation

Figure 2 (a) illustrates a considerable amount of correlation among the features extracted from the heart scan (h_1,h_2, h_3, ..., h_8). It suggests a multicollinearity issue, which can pose challenges for linear regression. To solve this problem, we scale the h_i and we conduct Principal Component Analysis (PCA) on these scaled features. We decide to retain only two components since they collectively account for more than 80% of the variance (see figure 2 (b)).

[1] A larger version of the correlation matrix is available in the appendix.

2 Features selection

Having removed irrelevant variables from our model, we can start the first variable selection using methods independent of our model, known as filters. Before that, we split our database into a train and a test set and then scale them. To ensure a meticulous selection process, we chose to base our selection on the three filtering methods discussed in class. This approach provides a robust set of variables for implementing our models. While implementing the models, we will conduct a second round of variable selection based on our model, referred to as wrapper methods, simultaneously with the hyperparameter selection of the model. This procedure streamlines the code execution and reduces the complexity associated with feature and hyperparameter selection.

Correlation filter

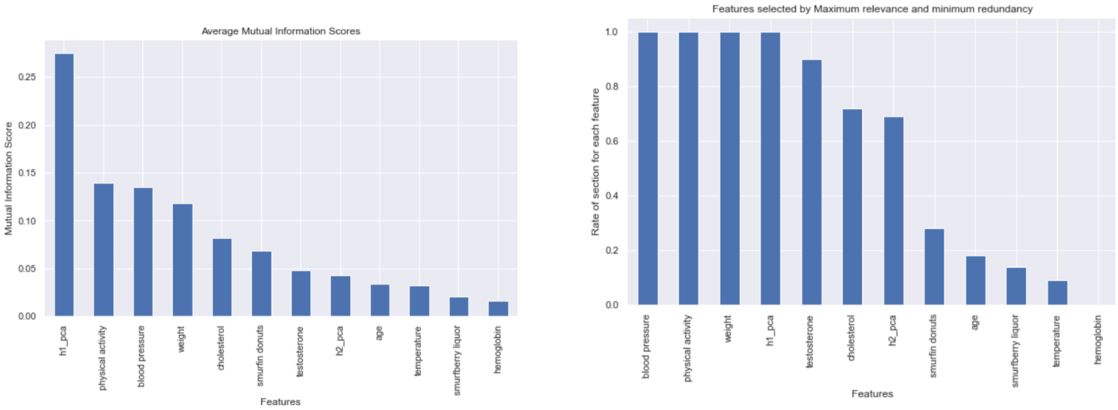
We applied a correlation filter to our data. Instead of opting for a fixed number of features, we chose to select features with a correlation above a specific threshold. In the literature [2], a threshold of 0.3 for the absolute value of correlation is considered "fair" in medicine, and a correlation (in absolute value) above 0.2 is deemed non-negligible in other fields. However, we found that a threshold of 0.3 was somewhat restrictive for our dataset and therefore decided to set the threshold at 0.2. The features selected using this method are : *age, blood pressure, cholesterol, physical activity, testosterone, weight, h1_pca.*

Mutual Information filter

Given that correlation only captures linear dependence, we need additional methods to select features. Determining a threshold for mutual information is not feasible, as mutual information have no superior bounds. As a solution, we plot the MI of the variablesto see at which point the MI becomes no longer relevant. By repeating the code several times, we observed variability in the obtained results ; the last features selected changed from one iteration to another. This variability stems from the fact that some features have similar mutual information. To address this issue, we conducted 100 iterations of mutual information calculations and averaged the results (see figure 3 (a)).

Based on the graph, we see gaps in the MI between the first and the second variables, the fourth and fifth and the sixth and seventh. We decide that the first six are the selected ones with this method. A comparison with the features selected by the correlation filter reveals differences ; *testosterone* and *age* were not chosen. It may be due to the fact that, as we see in Figure 2 (a), the correlation between these two variables and *blood pressure* are quite high (and so they give the same information).

[2] Akoglu, H. (2018). User's guide to correlation coefficients. Turkish journal of emergency medicine, 18(3), 91-93.



(a) Average mutual information after 100 iterations

(b) Rate of selected features after 100 iteration of MRMR criterion

FIGURE 3 – Feature selection

Maximum relevance and minimum redundancy (MRMR)

Within the chosen subset, redundancy among features may exist, especially if they are highly correlated with each other. To address this, we implement the Maximum Relevance and Minimum Redundancy approach. The first filter selects 7 features, and the second selects 6. We decide to implement MRMR with 7 features to select. Similar to the approach used for mutual information, we conduct 100 iterations and create a plot to identify features that were consistently selected. Four features were always selected, and 7 were selected 60% of the time (see Figure 3 (b)). We have chosen to proceed with these 7 features for the remainder of our analysis since we will perform a wrapper to perfect our selection.

3 Model selection

The data of all the plots have been scaled up for better visualization of the results.

3.1 Linear regression

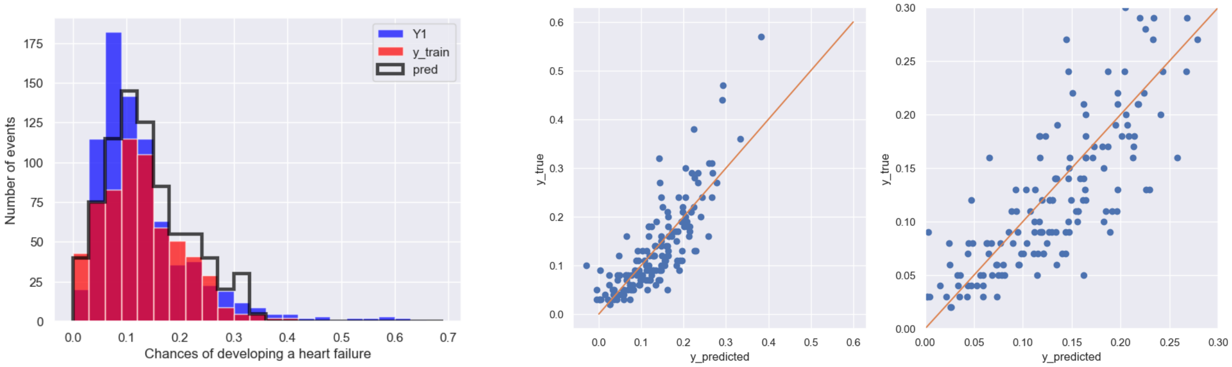


FIGURE 4 – Prediction visualization - Linear Regression

The mean *Root Mean Squared Error* (RMSE) over 10 runs on the testing set is : 0.05606. The execution time is : 0.02012 seconds. With a forward wrapper, we are able to determinate the optimal number of feature to select in order to get a better fit. In this case, the number of selected features is 6, which are : 'blood pressure', 'physical activity', 'weight', 'h1 pca', 'cholesterol', 'h2 pca'.

3.2 Multi Layer Perceptron

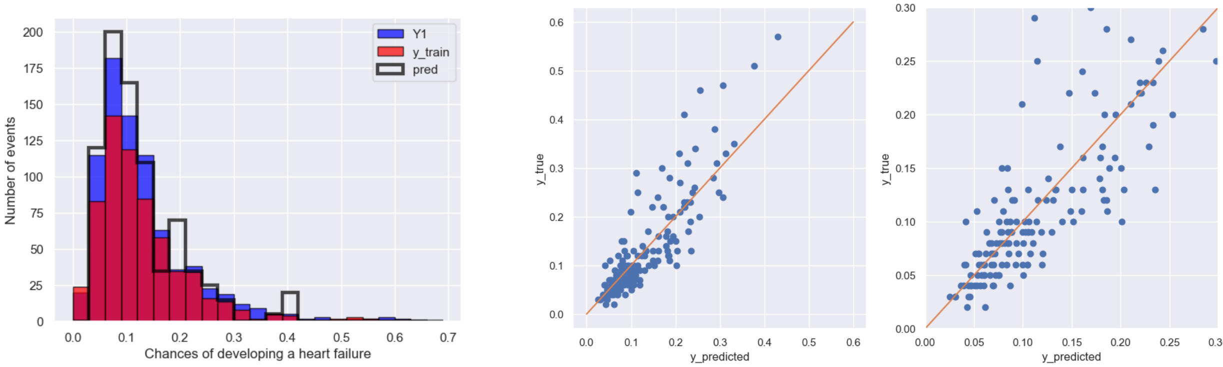


FIGURE 5 – Prediction visualization - MLP

The different hyperparameters used here are : number of hidden layers set to 2 layers of 128 neurons each, learning rate type, here : *squared error* , initial learning rate value set to 0.001 and maximum iterations set to 128.The mean RMSE over 10 runs on the testing set is : 0.05153.The execution time is : 1.5731 seconds. The most optimal number of features for this model is 4, which are : 'blood pressure', 'physical activity', 'weight', 'h1 pca'.

3.3 K-nearest neighbors

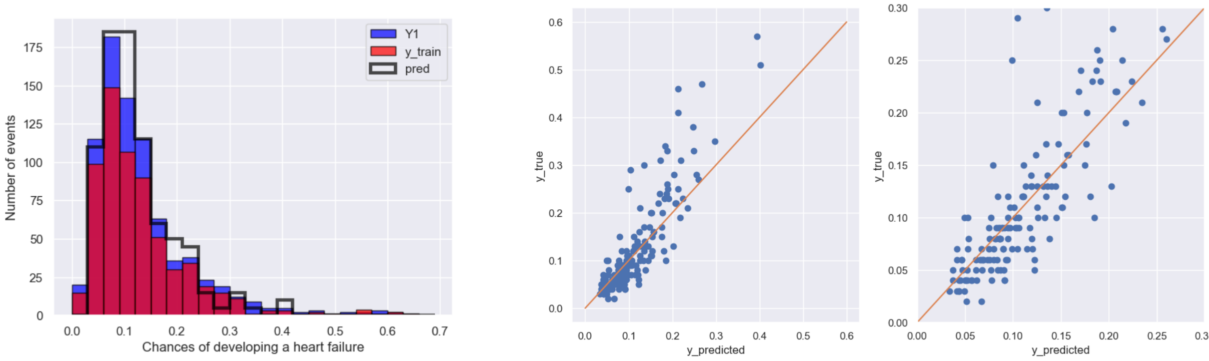


FIGURE 6 – Prediction visualization - KNN

For the KNN, the different hyperparameters tuned were : number of nearest neighbors set to 10 and the weight function, here : *distance*. The mean RMSE over 10 runs on the testing set is : 0.05433. The execution time is : 0.03124 seconds. The most optimal number of features for this model is 6, which are : 'blood pressure', 'physical activity', 'weight', 'h1 pca','cholesterol', 'h2 pca'.

3.4 Random forest

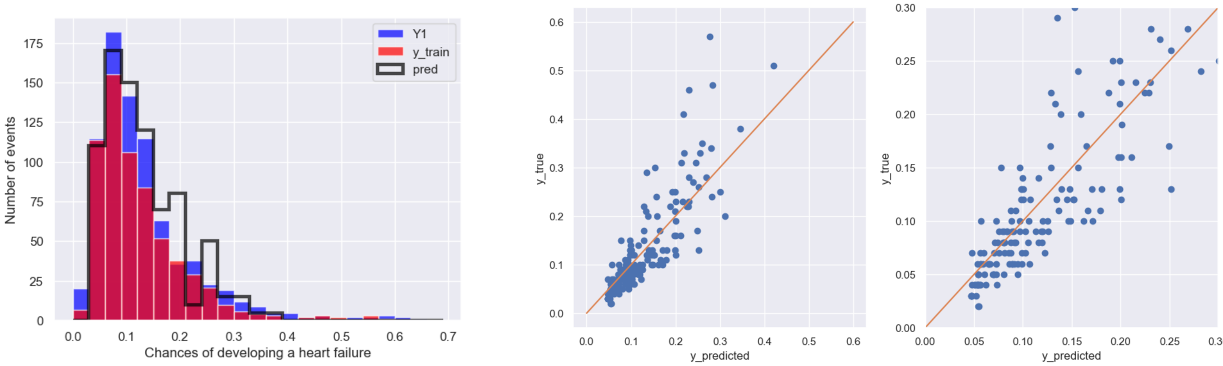


FIGURE 7 – Prediction visualization - Random Forest

For the random forest, the different hyperparameter tuned were : the number of estimators set to 50, the criterion, here : *absolute error* ; and the maximum depth set to *None* . The mean RMSE over 10 runs on the testing set is : 0.05231. The execution time is : 0.40436 seconds. The most optimal number of features for this model is 5, which are : 'blood pressure', 'physical activity', 'weight', 'h1 pca', 'h2 pca'.

3.5 Model comparison

The metric used to determine the best model is the *Root Mean Squared Error*. Hyperparameters tuning has been realized with *for loops*, selecting a different number of features as well as different values for several hyperparameters. For each combination of hyperparameters and features, a model is trained, and the combination linked to the lowest RMSE is kept.
The KNN model achieved the best result with an RMSE of 0.04295. However, reproducing this result consistently is not possible. Generally, the MLP emerges as the best estimator, with a mean RMSE of 0.05153. It is the model used for predicting Y2.
Despite higher RMSE and mediocre visualization plots, linear regression and stochastic gradient descent are methods that require fewer computational resources compared to models such as MLP or Random Forest, especially when these have a high number of neurons or estimators. One can conclude that the usefulness of these models is situational, but their result-over-computational-power ratio is decent. As mentioned beforehand, the random forest and the MLP models require more power but yield more accurate results when it comes to the first dataset.
Two common characteristics shared by all our models that we can highlight are the difficulty in predicting cases with a high chances of heart failure and very similar results for cases with a low chances. As seen in all the visualization plots, the data points tend to be in the upper part of the plot for values of $y_{predicted}$ higher than 0.3. We can then conclude that our models are more effective in predicting cases with a relatively low chance ($< 30\%$) of developing a heart failure.

Appendix

Appendix 1 : Verification of anova application conditions

The conditions for performing the ANOVA are the following : normality, which is checked using boxplots, data independence which is assumed according to the project guidelines, and homoscedasticity which is verified using a Levene test (see table 2). Homoscedasticity was violated for three of the variables, so we performed a kruskal wallis test for these variables (see table 3) . As the conclusions were the same as for the anova, we decided to leave them in the main text, see 1.

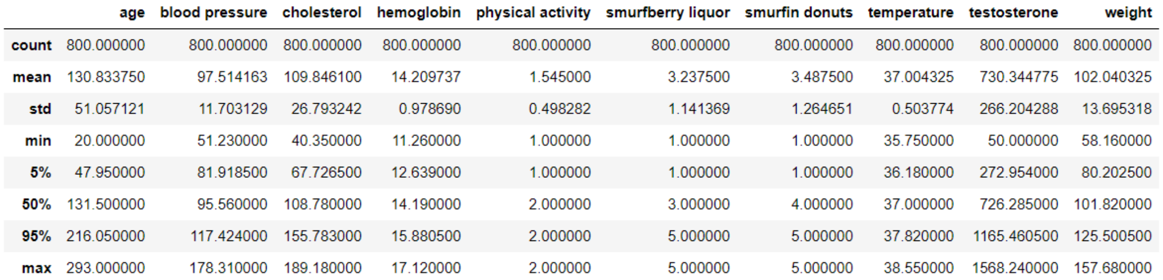
Levene Test for homoscedasticity		
Variable	p-value	Conclusion
Blood type	0.941	group variances are similar
Physical activity	<0.01	homoscedasticity is not respected.
Sarsaparilla	0.040	homoscedasticity is not respected.
Smurfin liquor	0.078	group variances are similar
Smurfin donuts	<0.01	homoscedasticity is not respected.

Table 2 – Levene test for homoscedasticity

Kruskal Wallis		
Variable	p-value	Conclusion
Physical activity	<0.01	Significative differences
Sarsaparilla	0.722	No significative differences
Smurfin donuts	<0.01	Significative differences

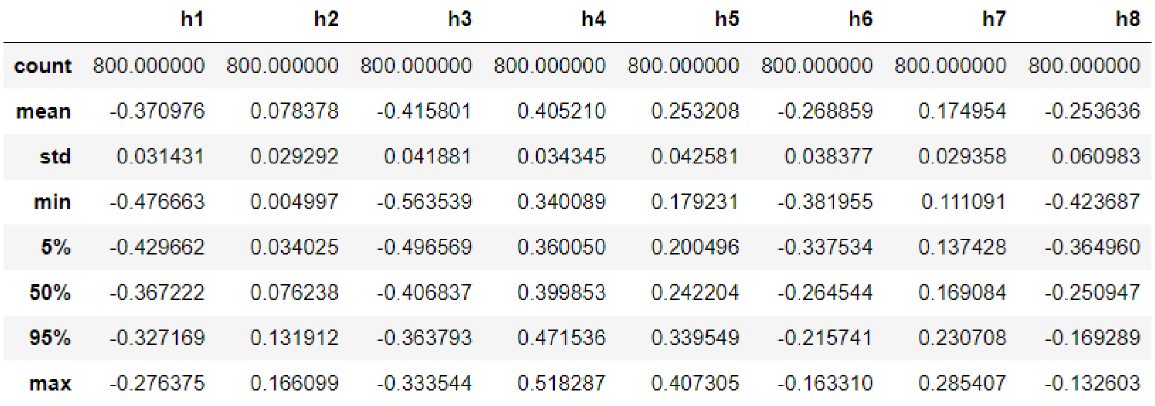
Table 3 – Kruskal Wallis for the variables for which homoscedasticity was not respected.

3.5.1 Appendix 2 : Percentile table



	age	blood pressure	cholesterol	hemoglobin	physical activity	smurfberry liquor	smurfin donuts	temperature	testosterone	weight
count	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000
mean	130.833750	97.514163	109.846100	14.209737	1.545000	3.237500	3.487500	37.004325	730.344775	102.040325
std	51.057121	11.703129	26.793242	0.978690	0.498282	1.141369	1.264651	0.503774	266.204288	13.695318
min	20.000000	51.230000	40.350000	11.260000	1.000000	1.000000	1.000000	35.750000	50.000000	58.160000
5%	47.950000	81.918500	67.726500	12.639000	1.000000	1.000000	1.000000	36.180000	272.954000	80.202500
50%	131.500000	95.560000	108.780000	14.190000	2.000000	3.000000	4.000000	37.000000	726.285000	101.820000
95%	216.050000	117.424000	155.783000	15.880500	2.000000	5.000000	5.000000	37.820000	1165.460500	125.500500
max	293.000000	178.310000	189.180000	17.120000	2.000000	5.000000	5.000000	38.550000	1568.240000	157.680000

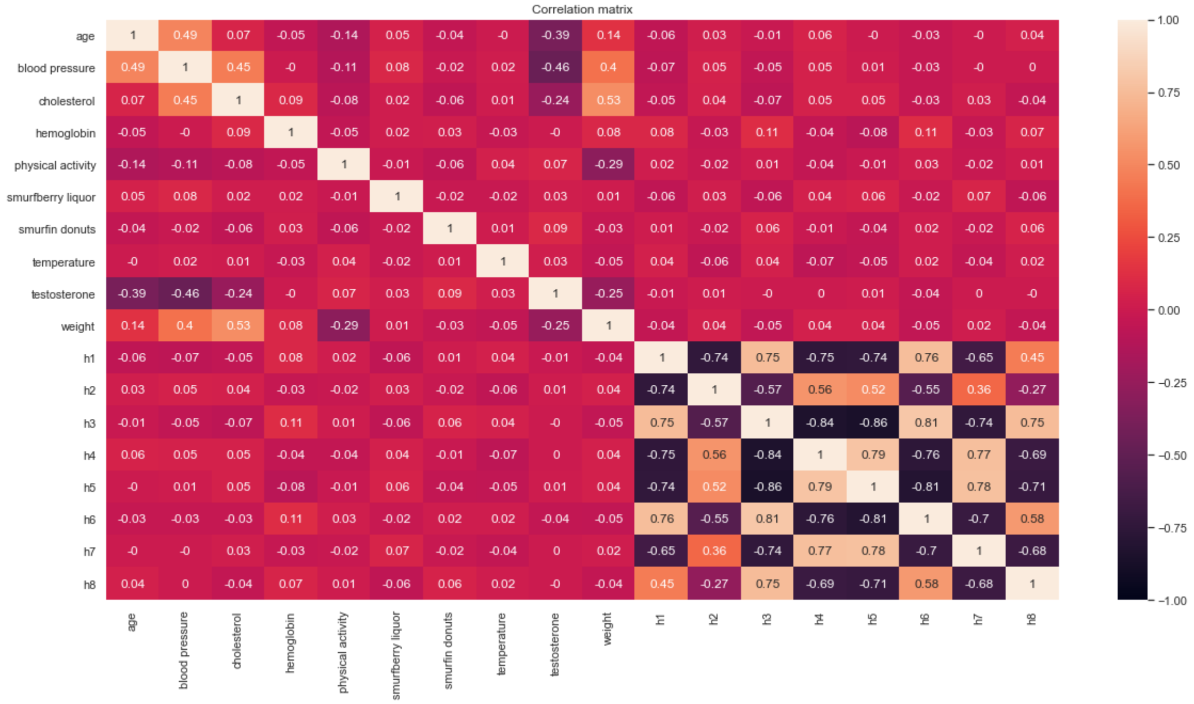
Figure 8 – Percentile of the features



	h1	h2	h3	h4	h5	h6	h7	h8
count	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000	800.000000
mean	-0.370976	0.078378	-0.415801	0.405210	0.253208	-0.268859	0.174954	-0.253636
std	0.031431	0.029292	0.041881	0.034345	0.042581	0.038377	0.029358	0.060983
min	-0.476663	0.004997	-0.563539	0.340089	0.179231	-0.381955	0.111091	-0.423687
5%	-0.429662	0.034025	-0.496569	0.360050	0.200496	-0.337534	0.137428	-0.364960
50%	-0.367222	0.076238	-0.406837	0.399853	0.242204	-0.264544	0.169084	-0.250947
95%	-0.327169	0.131912	-0.363793	0.471536	0.339549	-0.215741	0.230708	-0.169289
max	-0.276375	0.166099	-0.333544	0.518287	0.407305	-0.163310	0.285407	-0.132603

FIGURE 9 – Percentile of h_i

Appendix 3 : Correlation matrix



Correlation matrix

	age	blood pressure	cholesterol	hemoglobin	physical activity	smurfberry liquor	smurfin donuts	temperature	testosterone	weight	h1	h2	h3	h4	h5	h6	h7	h8
age	1	0.49	0.07	-0.05	-0.14	0.05	-0.04	-0	-0.39	0.14	-0.06	0.03	-0.01	0.06	-0	-0.03	-0	0.04
blood pressure	0.49	1	0.45	-0	-0.11	0.08	-0.02	0.02	-0.46	0.4	-0.07	0.05	-0.05	0.05	0.01	-0.03	-0	0
cholesterol	0.07	0.45	1	0.09	-0.08	0.02	-0.06	0.01	-0.24	0.53	-0.05	0.04	-0.07	0.05	0.05	-0.03	0.03	-0.04
hemoglobin	-0.05	-0	0.09	1	-0.05	0.02	0.03	-0.03	-0	0.08	0.08	-0.03	0.11	-0.04	-0.08	0.11	-0.03	0.07
physical activity	-0.14	-0.11	-0.08	-0.05	1	-0.01	-0.06	0.04	0.07	-0.29	0.02	-0.02	0.01	-0.04	-0.01	0.03	-0.02	0.01
smurfberry liquor	0.05	0.08	0.02	0.02	-0.01	1	-0.02	-0.02	0.03	0.01	-0.06	0.03	-0.06	0.04	0.06	-0.02	0.07	-0.06
smurfin donuts	-0.04	-0.02	-0.06	0.03	-0.06	-0.02	1	0.01	0.09	-0.03	0.01	-0.02	0.06	-0.01	-0.04	0.02	-0.02	0.06
temperature	-0	0.02	0.01	-0.03	0.04	-0.02	0.01	1	0.03	-0.05	0.04	-0.06	0.04	-0.07	-0.05	0.02	-0.04	0.02
testosterone	-0.39	-0.46	-0.24	-0	0.07	0.03	0.09	0.03	1	-0.25	-0.01	0.01	-0	0	0.01	-0.04	0	-0
weight	0.14	0.4	0.53	0.08	-0.29	0.01	-0.03	-0.05	-0.25	1	-0.04	0.04	-0.05	0.04	0.04	-0.05	0.02	-0.04
h1	-0.06	-0.07	-0.05	0.08	0.02	-0.06	0.01	0.04	-0.01	-0.04	1	-0.74	0.75	-0.75	-0.74	0.76	-0.65	0.45
h2	0.03	0.05	0.04	-0.03	-0.02	0.03	-0.02	-0.06	0.01	0.04	-0.74	1	-0.57	0.56	0.52	-0.55	0.36	-0.27
h3	-0.01	-0.05	-0.07	0.11	0.01	-0.06	0.06	0.04	-0	-0.05	0.75	-0.57	1	-0.84	-0.86	0.81	-0.74	0.75
h4	0.06	0.05	0.05	-0.04	-0.04	0.04	-0.01	-0.07	0	0.04	-0.75	0.56	-0.84	1	0.79	-0.76	0.77	-0.69
h5	-0	0.01	0.05	-0.08	-0.01	0.06	-0.04	-0.05	0.01	0.04	-0.74	0.52	-0.86	0.79	1	-0.81	0.78	-0.71
h6	-0.03	-0.03	-0.03	0.11	0.03	-0.02	0.02	0.02	-0.04	-0.05	0.76	-0.55	0.81	-0.76	-0.81	1	-0.7	0.58
h7	-0	-0	0.03	-0.03	-0.02	0.07	-0.02	-0.04	0	0.02	-0.65	0.36	-0.74	0.77	0.78	-0.7	1	-0.68
h8	0.04	0	-0.04	0.07	0.01	-0.06	0.06	0.02	-0	-0.04	0.45	-0.27	0.75	-0.69	-0.71	0.58	-0.68	1

FIGURE 10 – Correlation matrix

Appendix 4 : SGD

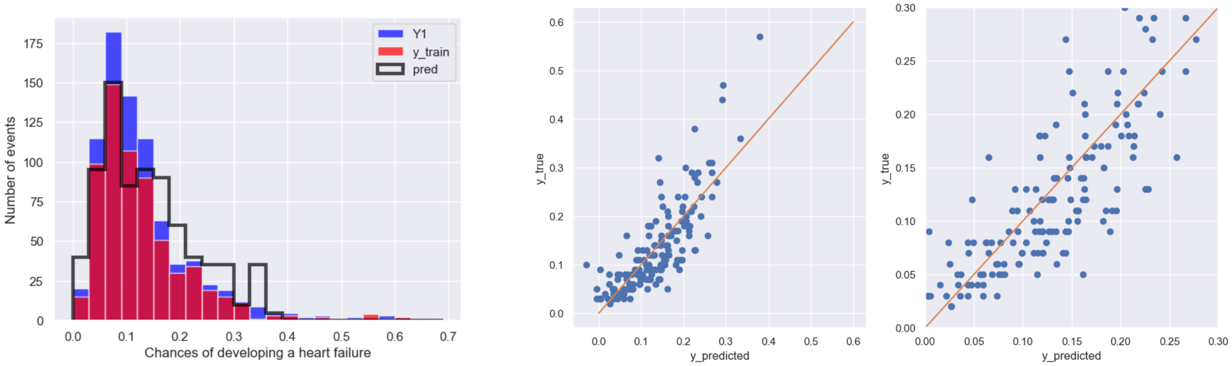


Figure 11 – Prediction visualization - SGD

For the SGD, the different hyperparameters tuned were : the number of iterations set to 200, the learning rate function, here : *adaptive* ; and the loss function : *squared error*.
The mean RMSE over 10 runs on the testing set is : 0.05607.
The execution time is : 0.02044 seconds.
The most optimal number of selected features for this model is 6, which are : 'blood pressure', 'physical activity', 'weight', 'h1 pca','cholesterol', 'h2 pca'.