Afonso Henrique, 98555 - afonso.henrique@tecnico.ulisboa.pt
Cátia Oliveira, 100142 - catia.oliveira@tecnico.ulisboa.pt
Link to the GitHub repository: `https://github.com/c-radicallis/Sis_Inteligentes_1st_Assignment.git`

1 Hair Dryer Dataset

In this exercise, we modelled a single input single output (SISO) fuzzy rule-based system using a hair dryer dataset, where the input is the voltage over the heating device, and the output is a voltage related to the outlet air temperature.
When the output temperature rises to a specified threshold, the input voltage is set to low, and when the output temperature lowers to another specified threshold, the input voltage is set to high.
No preprocessing of the data was necessary although the data file was modified to include a name for each variable separated by a comma on the first line.

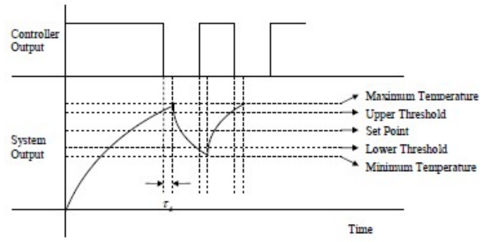

Figure 1: Closed loop response with on-off control

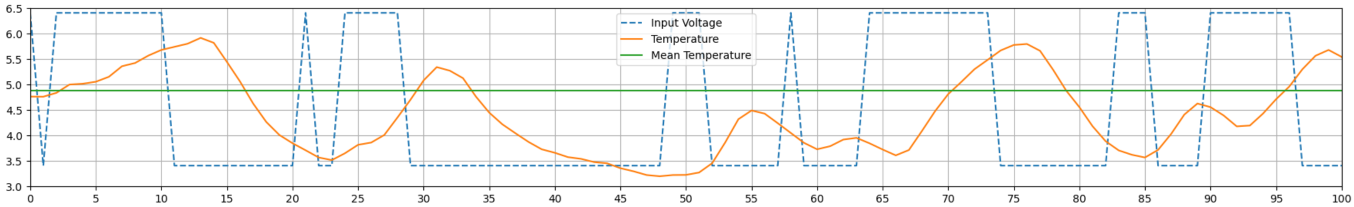

Figure 2: Input and temperature voltages over the first 100 time samples. Mean temperature (voltage) over the 1000 time samples is also plotted.

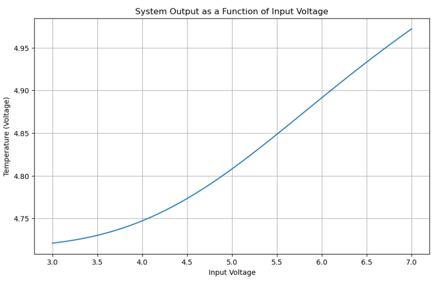

Figure 3: System output as a function of input voltage for 2 clusters.

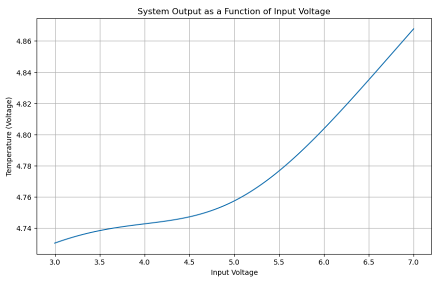

Figure 4: System output as a function of input voltage for 5 clusters.

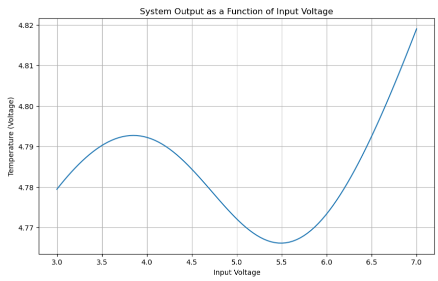

Figure 5: System output as a function of input voltage for 10 clusters.

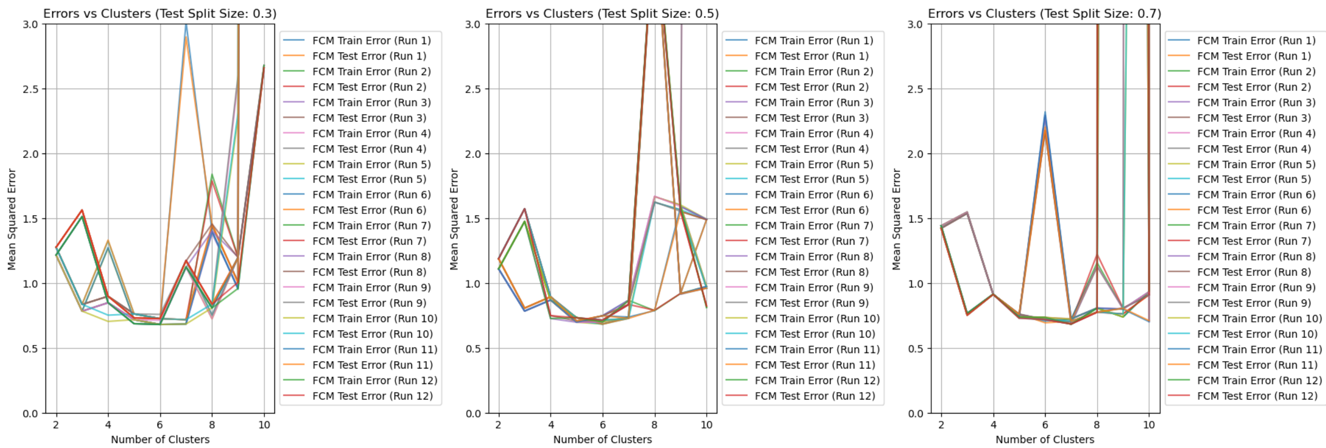

Figure 6: Mean squared error for different size train-test splits. For each size split the model was trained and tested 12 times.

The methods used for clustering were the Fuzzy c-means (fcm) and the Gustafson-Kessel method (gk), although the latter method yielded very inconsistent results most of the time we believe were due to it's sensitivey to the initial placement of cluster centers, which lead to convergence to local minima, resulting in suboptimal clustering, and as such we opted not to present them in the results. The global fit option was used in the estimation of the consequent parameters.

2 Wisconsin Breast Cancer Original Dataset

2.1 Goal and pre-process of the dataset

The goal is to develop a fuzzy rule-based model for classifying the classes where it is able to predict if a person has cancer (1) or does not have cancer(0), by using a set of measures relevant for cancer diagnosis. For simplicity, while writing the code, the measures and outcome were given simpler names. This conversion is shown in Table 7.

The first step was to clean the data and make sure that any 'NAN' values were removed from the dataset. To choose which parameters help obtain the best model possible, a process similar to grid-search was made where the parameters were varied and some results obtained are shown in Subsection 2.2.

Figure 7: Names of the features in the code.

Measures	Name in code
Clump Thickness	x_1
Uniformity of Cell Size	x_2
Uniformity of Cell Shape	x_3
Marginal Adhesion	x_4
Single Epithelial Cell Size	x_5
Bare Nuclei	x_6
Bland Chromatin	x_7
Normal Nucleoli	x_8
Mitoses	x_9
Outcome	y

The parameter tuning was made with the following parameter grid:

$$\text{Number of clusters/rules: } ncl = \{2, 3, 5, 8, 10\}$$

$$\text{Method used for clustering: } mtd = \{'fcm', 'gk'\}$$

$$\text{Optimization: } fit = \{True, False\}$$

The methods used for clustering were the Fuzzy c-means (fcm) and the Gustafson-Kessel method (gk). As for the Optimization parameter, when this parameter is true, a global fit is being used, otherwise, if false, a local fit is being made.

The membership functions were all made with a Gaussian distribution function.

2.2 Results and Conclusions

The metrics used to evaluate the performance of the classification models were accuracy, recall, precision, F1-score and Cohen's kappa. Figures 8 to 11 show the accuracy and F1-score in the function of the number of clusters for global and local fit.

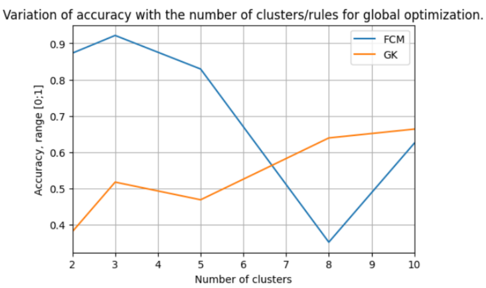

Figure 8: Variation of accuracy with the number of clusters/rules for global optimization.

Figure 9: Variation of F1-Score with the number of clusters/rules for global optimization.

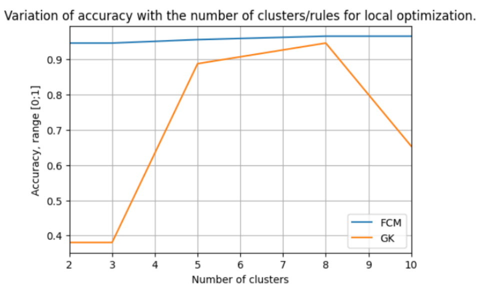

Figure 10: Variation of accuracy with the number of clusters/rules for local optimization.

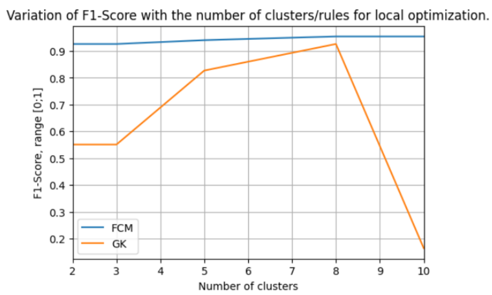

Figure 11: Variation of F1-Score with the number of clusters/rules for local optimization.

From Figures 8 and 9, it can be concluded that the best model obtained by global optimization is the model made with the following parameters:

$$\{ncl, mtd, fit\} = \{2, 'fcm', True\}$$

As for the models made with the local fit, from Figures 10 and 11, for the method of Fuzzy c-means, the models present similar accuracies and F1-scores for all the number of clusters chosen, however, the best model that was considered as the following parameters:

$$\{ncl, mtd, fit\} = \{8, 'fcm', False\}$$

The confusion matrices shown in Figures 12 and 13 are the results obtained using the parameters for the best models of global fit and local fit, respectively.

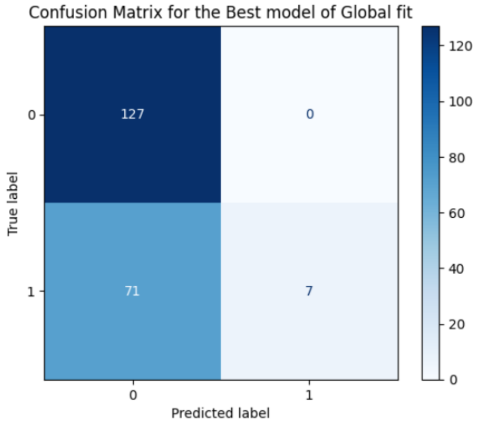

Figure 12: Confusion matrix for the best model of global fit..

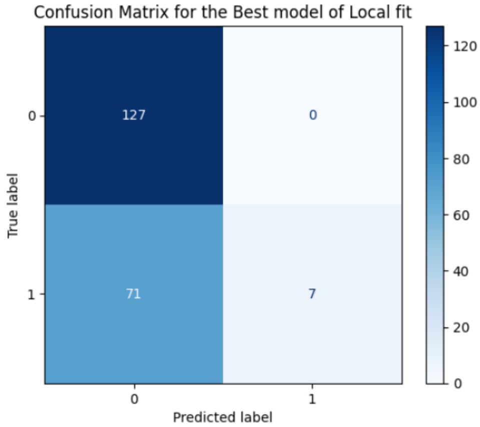

Figure 13: Confusion matrix for the best model of local fit.