

# Assignment report,

Machine Learning

MSc Applied Bioinformatics,

Marie Schmit

# Abstract

Data exploratory and analysis

1. Enose samples are separated by their sensory scores. Sample 10F9 can be considered as an outlier. HPLC samples are less clearly clustered by samples, but three close groups still emerged. There are some potential outliers like 0F12 or 5F6.

Table PCA scatter plots of enose and HPLC grouped by sensors class

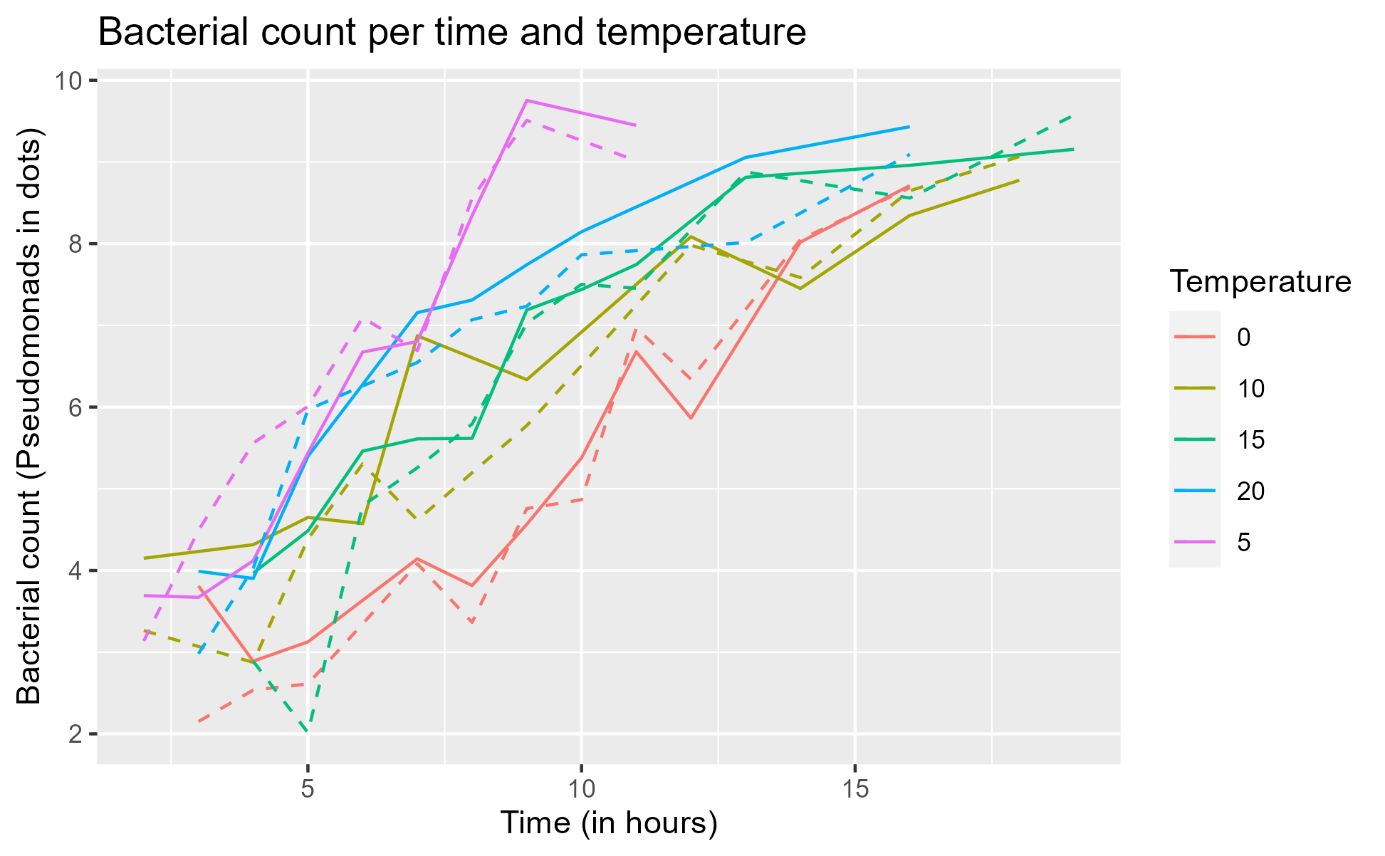
|  |  |
| --- | --- |
|  |  |

We try other analytical method to find better separation. First, we display the PCA plot with other methods (in 3D, in a biplot). We also use HCA. With this method, three clusters emerge, but they do not necessarily correspond to their sensory value. The previous outlier for enose (10F9) is no longer an outlier: it is now in the same cluster as F1a, which does not correspond to its sensory value. The same goes for HPLC, for which the clusters 2 and 3 are often mixed. This analysis is not the best.

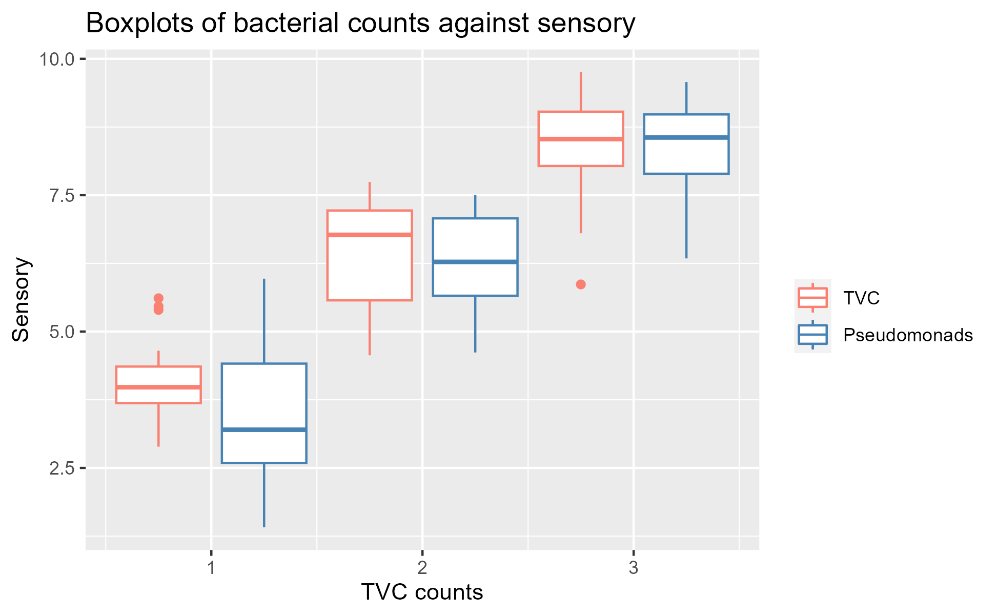
|  |  |
| --- | --- |
|  |  |

We decide that the data that seem to stand out are not sufficiently so to be considered as outliers. We choose not to remove them to avoid the risk of introducing errors if they are significant.

1. The number of bacteria, for both type of bacteria, increases with time. There are more bacteria for higher temperatures. The count of bacteria is abnormally high for 5 degrees. We also have less samples for this temperature. The number of bacteria seems to grow with time and temperatures increase.



1. The number of bacteria is for both types TVC and Pseudomonias higher when the sensory score is higher. Rotted meat has thus more bacteria than fresh one. TVC bacteria are slightly more numerous than pseudomonas. Also, TVC has three outliers.



Classification

We partition the data using the function createDataPartition, that ensure a balanced representation of our train and test sets. We make the partition reproducible by using the function “seed” and use the created index to create the datasets.

We first train the model using k nearest neighbour methods. We want to find the best fit and k parameter. To find the best model, we test different methods of scaling (center, auto scale and range scale), for different values of k (from 1 to 20). We evaluate the accuracy of the model for each of those scales and k values and we select the parameters leading to the best accuracy.

For HPLC data, the results are the following with one partition. Without scaling, the best accuracy is 0.9 for k =10. With scaling, the best combination is method “center” with various k values (different k give the same accuracy, we choose k = 10). Scaling was not necessary here.