Aprendizagem Automática 2022/2023

# Second Home Assignment

**Group number**: 36

**Student names and ids**:

Luís Pires 54471,

Inês Martins 54462,

Miłosz Włoch 59437

**How many hours each student contributed to the project**:

Luís Pires: 6 hours

Inês Martins: 6 hours

Miłosz Włoch: 6 hours

---

## Aims of the assignment:

- Students should provide the best possible classification models using whatever method covered in class
- Variable to Classify is Biodegradable
- Models should examine different hyperparameters and select the best one [Remember: everything else being similar, the simplest models should be prefered]
- Students should do Simple Cross Validation (testing=25% of all data) for evaluating the models, but the same data partitions must be used for all models

## Dataset description:

An augmented and edited version of the Dataset QSAR biodegradation Data Set containing 41 columns with both categorical and discrete values.

## Pre Data Processing Variable Selection:

Before doing any data processing it was necessary to make a preliminary selection of variables. For this we needed to have all features encoded which was the first step.

In this section we made two types of variable selection: selection of features based on missing data and correlation with results; and selection of features based on variance and the presence of the same value in various samples.

For the first one we started by showing a heatmap which indicated the features with missing data. After that we used an OLS regression to check the p-values for each feature. For each feature that had missing data we checked its p-value and if it was

above a threshold ($p > 0.05$), it would be removed because that meant that the feature had low correlation to the results likely due to missing data.

For the second type, we eliminate features that have a low variance relative to their scale ($\sigma < (max\_value - min\_value)/100$) and that have a high presence (> 80%) of the same value in multiple samples.

## Data Processing:

Dataset was not normalized and had some missing values as well. That is why before we could train any of the models to be able to classify column “Biodegradable” we needed to clean and standardize the data. First we normalized the data and after we handled missing values. Normalization was done using Standard Scaler and Power Transformer. We plotted every significant variable to see if they are skewed. Based on plots we decided to choose the following columns as skewed (not stable) [nHM, F01, F04, NssssC, nCb, C, nCp, HyWi_B, Mi, nN_N, nArNO2, nCRX3, SpPosA_B, B01, N_073] .We applied power transforming to these columns while standard scaling for others. Power Transformer deals with outliers very well. That is why it was used for fluctuating variables. Scaling is important because some classification models are sensitive to outliers and cannot handle them in a proper way. As for replacing null values with some numerical ones, we made our minds to choose Simple Imputer. We chose median strategy for imputation function over mean strategy because it gave us slightly better results for the second variable selection although both of the strategies made a lot of sense as our data were already normalized.

## Post Data Processing Variable selection:

After having the data scaled and the imputation done, a second round of variable selection was done. For this variable selection, two methods were compared and the best was chosen. The first method was using random forests to decide the feature selection. The feature selection used SelectFromModel function to determine the most significant attributes among others. Testing it with a knn classifier with sklearn default hyperparameters, it decided to choose only 9 columns out of 26 keeping an accuracy of about 94%.

The second method used was the stepwise method, more specifically the backward selection approach where the model starts with all features and one is removed to check whether or not there is improvement and from the best group of features the process is repeated until no more increase in the results is found. The results using sklearn’s knn classifier with hyperparameters was 17 columns out of 26 with an accuracy of about 95%.

Given the similarity in the results from both methods, we decided to go for the first one due to the reduced number of features.

## Model results:

- **A summary of the performed work**

  We decided to try 5 different types of models and choose the best of each and then compare them to choose the best model of all.

  For each model we did a grid search over a few of their hyperparameters to find the best ones. to evaluate the performance of the values for the hyperparameters we chose  to use accuracy except for SVC where we used the difference between accuracy of a prediction with the train set and a prediction with the test set. This was done because SVC is very prone to overfitting.

  For most of these models we searched for the best values over two hyperparameters. For the decision tree classifier and the random forest we would first search over one of the hyperparameters, find the best and use that as we search over the other hyperparameter.

  For the SVC and KNN we used two nested loops to iterate through all combinations of values due to the lower number of reasonable values each hyperparameter could take.

  As GaussianNB doesn't have hyperparameters to change the values, only one result exists for it.

  **A table with models tested and relevant statistics**

| Model | Best parameters | Accuracy | F1 | Recall | Precision |
|---|---|---|---|---|---|
| **Decision Tree Classifier** | **min_samples_split = 2, max_depth = 4** | **0.9238** | **0.9545** | **0.9550** | **0.9540** |
| **Random Forest** | **n_estimators = 100, max_depth = 23** | **0.9334** | **0.9596** | **0.9455** | **0.9741** |
| **Gaussian NB** | **N/A** | **0.9150** | **0.9485** | **0.9361** | **0.9613** |
| **KNN** | **weights = distance, n_neighbors = 7** | **0.9509** | **0.9710** | **0.9801** | **0.9620** |

| | | | | | |
|---|---|---|---|---|---|
| **SVC** | **C = 0.1, gamma = 0.1** | **0.9273** | **0.9568** | **0.9623** | **0.9513** |

- **Discussion and conclusion**

  In this Home Assignment we understood the importance of variable selection in simplifying the problem. We also understood the impact that scaling and imputation can have on the results on different models and that models can be more or less sensitive to outliers.
  The best model we found was KNN with parameters weights = distance and n_neighbors = 7, since it presents the largest accuracy, recall and f1, despite the fact that random forest has the largest precision.
  One reason why the best model was KNN may be the type of scaling of data that was made before model selection because KNN and SVC are more sensitive to outliers whereas decision trees and random forest aren't affected as much.
  Accuracy: 0.9509
  F1: 0.9710
  Recall: 0.9801
  Precision: 0.9620