Challenge 1: QSAR

Giovanni Filomeno

Artificial Intelligence in Life Science

JOHANNES KEPLER UNIVERSITY LINZ

Problem overview

The aim is to predict a molecule's biological activity or toxicity based on its chemical structure. Given a dataset of molecules together with activities, a model should be trained to predict the activities of new molecules based on the selected model.

In the training data: +1=active, 0=unknown (alias unlabeled), -1=inactive

Model

Random Forest Classifier:

- Chosen for its robustness and ability to handle high-dimensional, sparse fingerprint data. Random Forests can capture complex relationships and typically perform well in QSAR tasks.
- A set of 11 classifier (one per task).
- Hyperparameter tuning with BayesSearchCV (Bayesian Optimization) for parameters such as:
  - n_estimators, max_depth, min_samples_split, etc.
  - Uses ROC-AUC as the primary scoring metric.

Pre-processing

Molecule Standardization:

- Using RDKit's MolStandardize module.

Fingerprint Generation:

- Morgan Fingerprint (ECFP): radius=2, 1024 bits (*).
- MACCS Keys (166 bits).
- Concatenation of Morgan + MACCS for richer molecular descriptors.

Data Splitting:

- train_test_split (5% test set).

* A search is done a priori. The mean of the molecule is around 300 bits. Therefore, 1024 would be enough to cover similarities and so on

Results

#	AUC mean	AUC Task1	AUC Task2	AUC Task3	AUC Task4	AUC Task5	AUC Task6	AUC Task7	AUC Task8	AUC Task9	AUC Task10	AUC Task11
1	0.766	0.843	0.632	0.926	0.712	0.898	0.892	0.511	0.708	0.911	0.751	0.644

> 0.765

Possible
Improvement area