Python Project

UNDERWOOD Theo and TOUBIANA Benjamin

DIA 7

2021-2022

Problem : Determine if a chemical is ready biodegradable

What : **QSAR biodegradation Data Set**

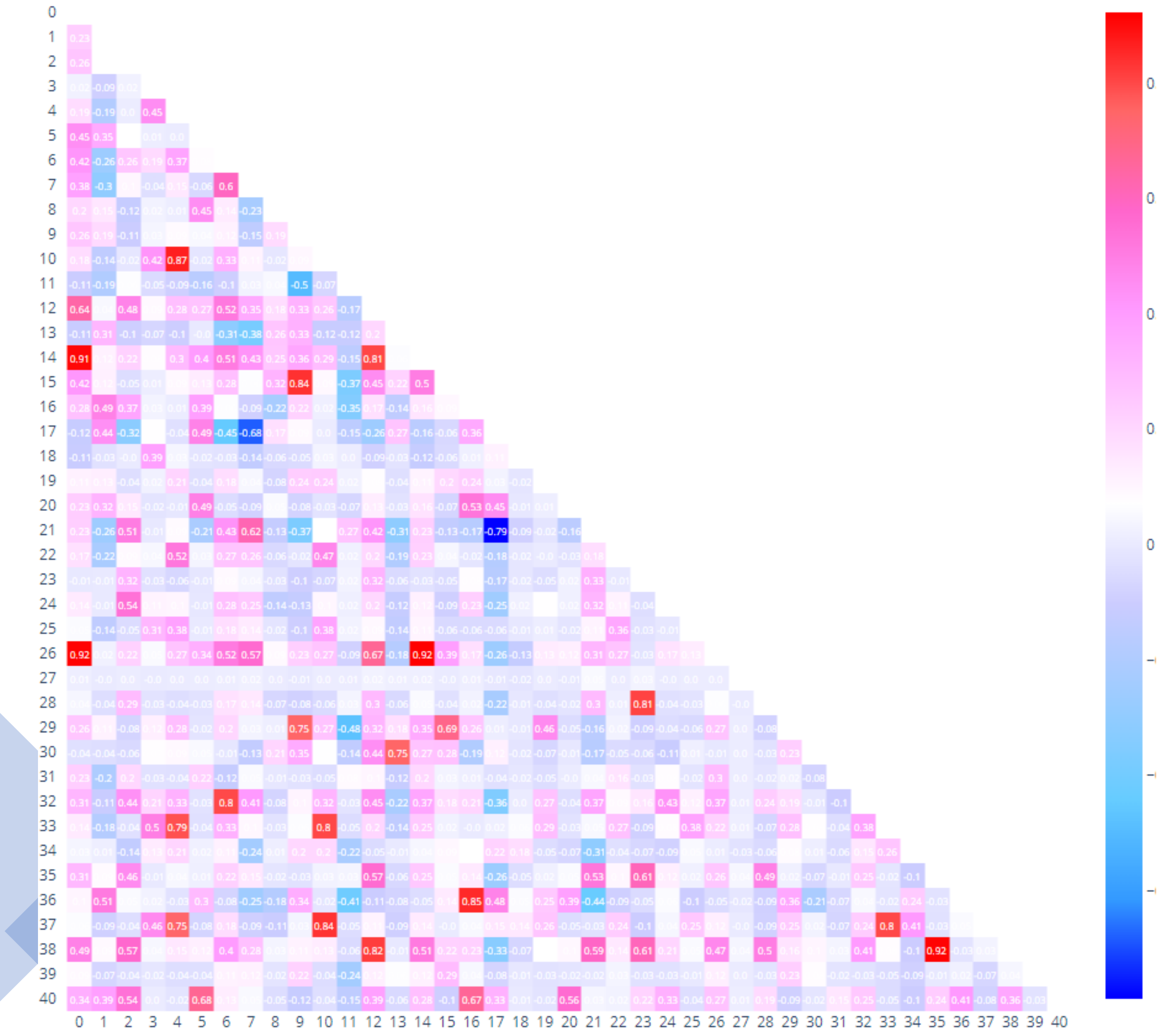
How: Molecular characteristics of each chemical

Why : Classify chemicals by ready biodregradability

Data Exploration – Raw Data

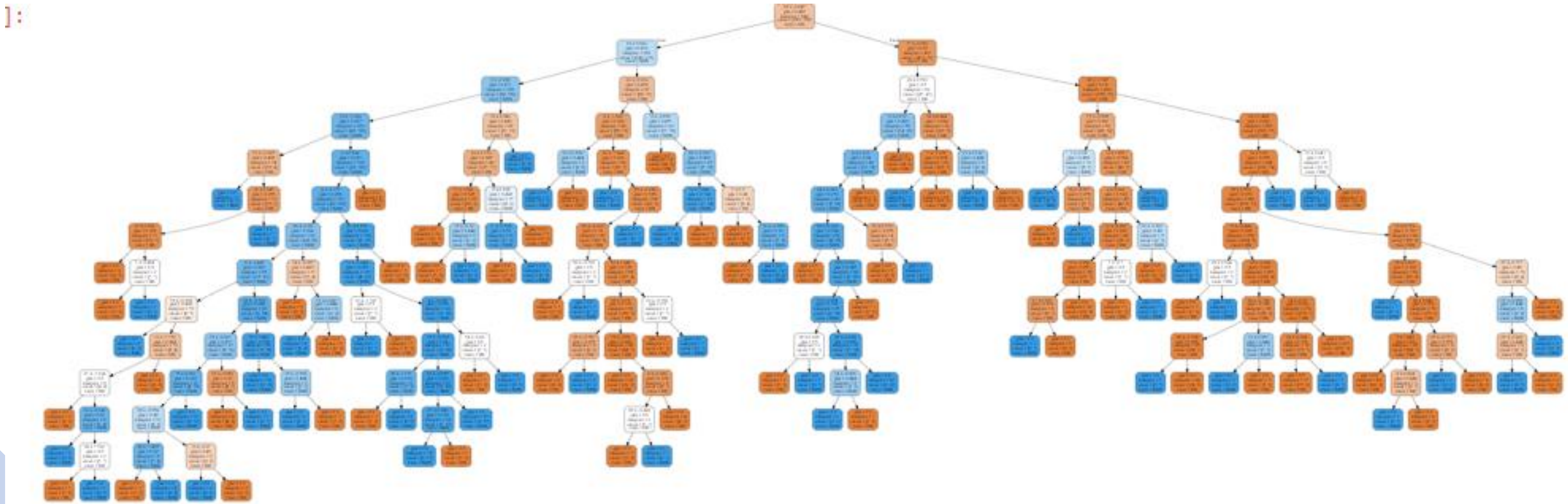
- File format : ’’.csv’’
- Individuals : 1055 Chemicals
- Features : 41 Molecular characteristics
- Targets : 2 possibilities – Ready Biodegradable, Non Ready Biodegradable

Data Exploration – Visualization



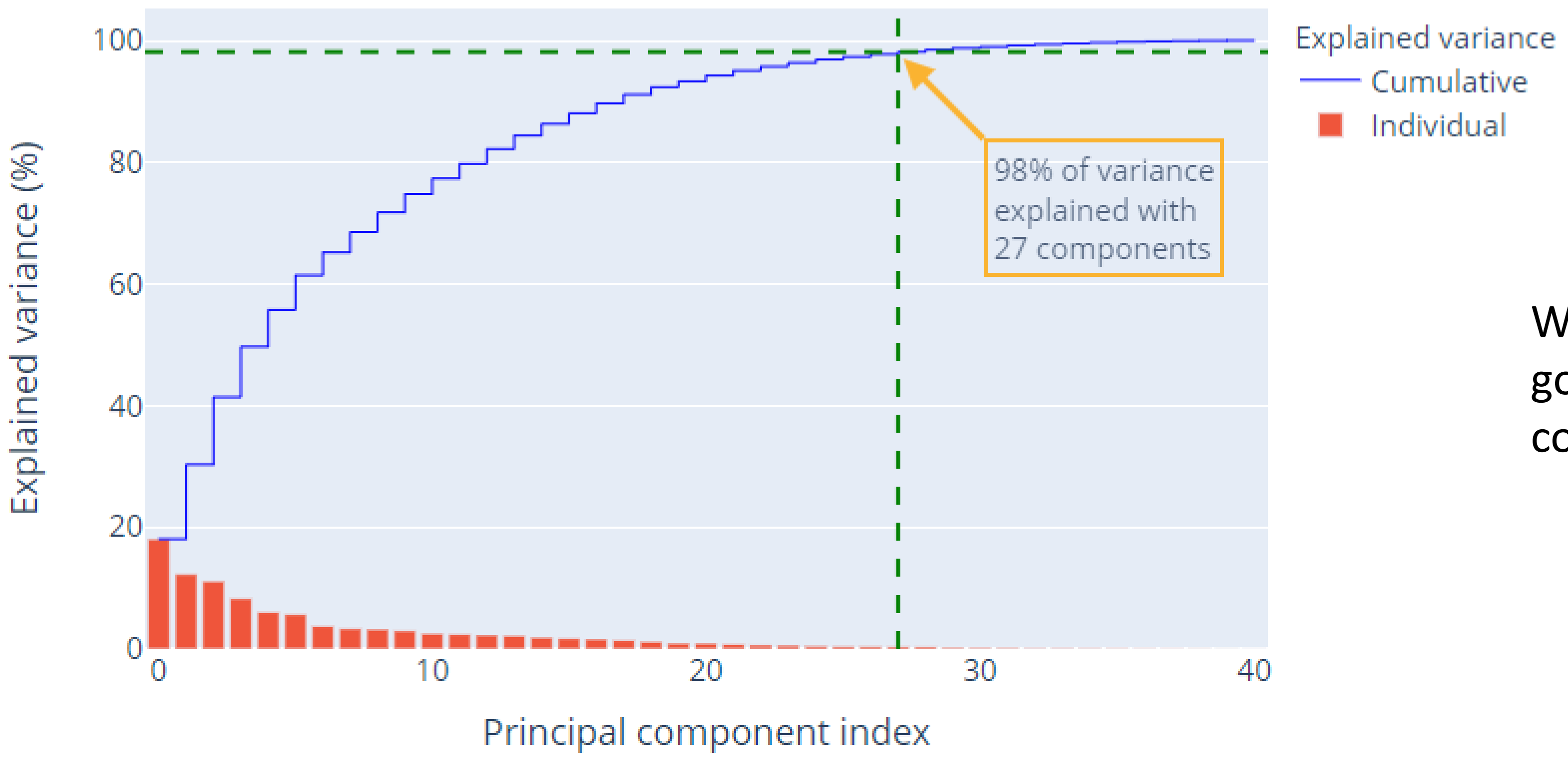
Since the variables are molecular features, we actually don't understand well what they represent and how they relate to each other. But here is the correlation matrix.

- Here is a plot of a decision tree to find out which features have the greatest impact on the predictions. It’s not very clear because of the many variables.



Data Pre-processing

- We did not change the base data since there isn’t missing values
- Standardization of the data (mean of 0 and standard deviation of 1)
- Splitting dataset into training and testing sets



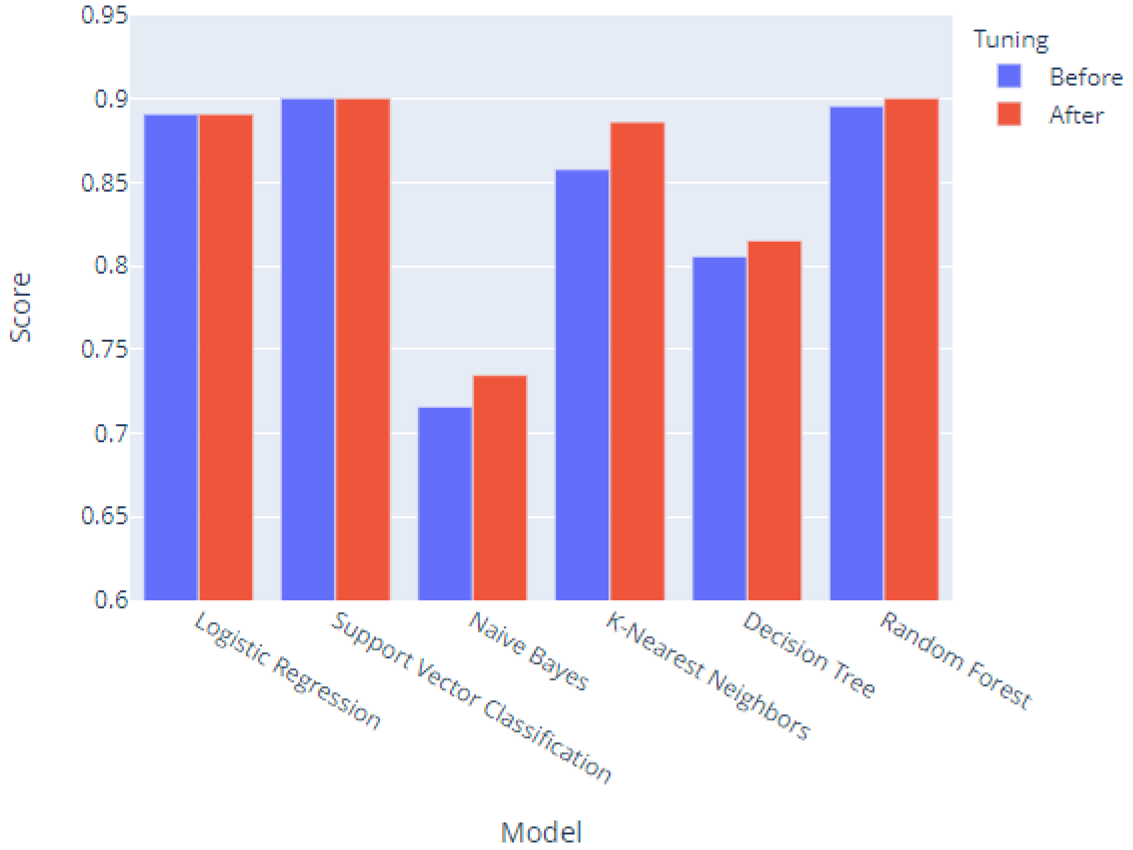
We made an ACP to find how to go about reducing the amount components we need.

We obtain a variance of 98% with 25 components, so we finally decided to keep all the variables because treating 27 or 41 is the same for us. We would have used PCA if we had obtained less than 10 components.

Modelization

Model	Accuracy before tuning	Accuracy after tuning
Logistic Regression	0.89	0.89
Support Vector Classification	0.90	0.90
Naive Bayes	0.72	0.73
K-Nearest Neighbors	0.86	0.89
Decision Tree	0.81	0.82
Random Forest	0.89	0.90

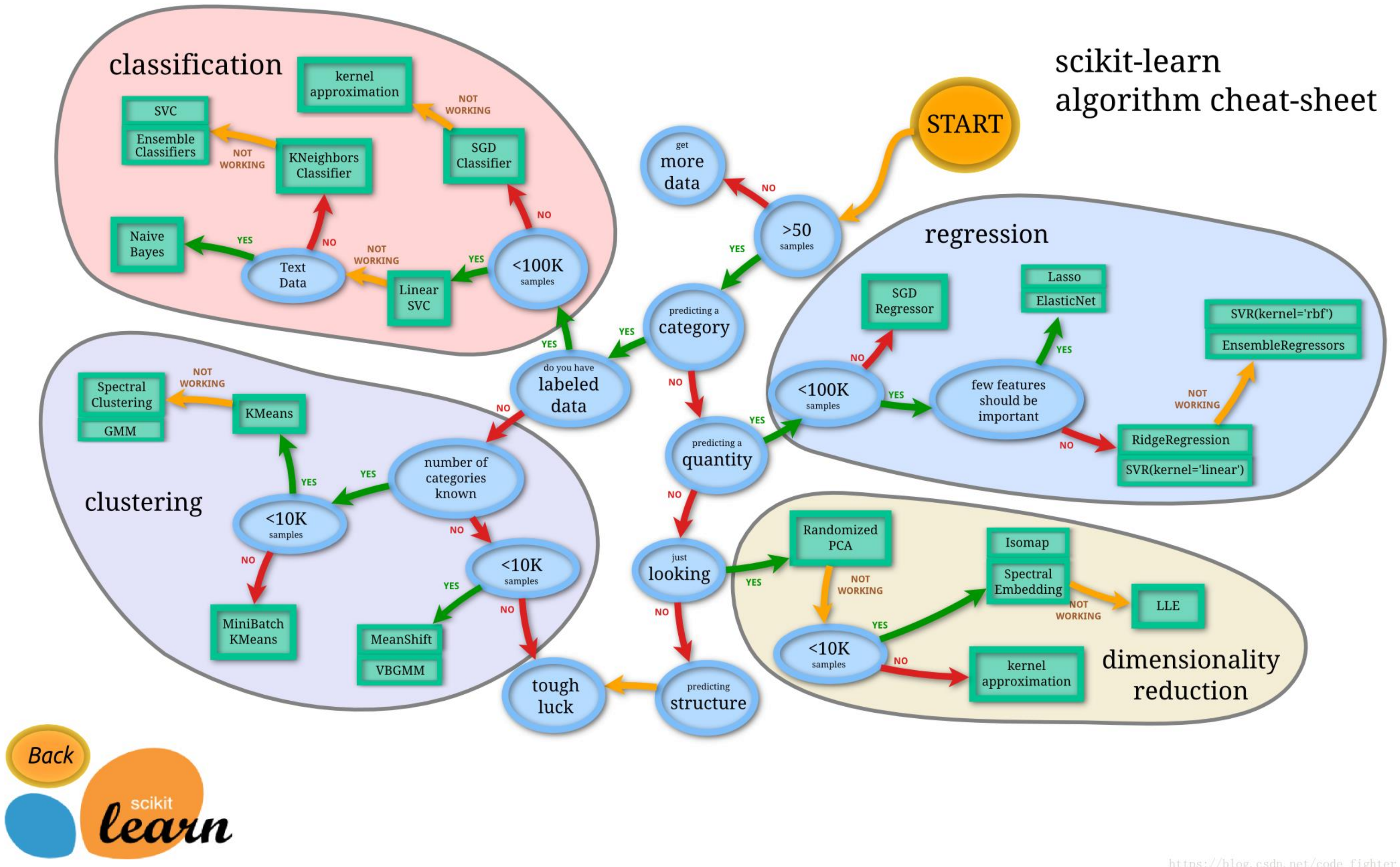
Conclusion



- The hyperparameters tuning has often improved the accuracy of the models even if sometimes it was the same.

- We notice that the ranking before and after the tuning remains the same, which confirms that some models are better than others in some situations.

- The type of data processed is numbers and the site of scikit-learn recommends SVC for this kind of dataset. That’s why it’s the best model.


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
NO
predicting a quantity
YES
regression
SGD Regressor
NO
<100K samples
YES
Lasso
ElasticNet
YES
few features should be important
NO
SVR(kernel='rbf')
EnsembleRegressors
NOT WORKING
RidgeRegression
SVR(kernel='linear')
clustering
NOT WORKING
Spectral Clustering
GMM
KMeans
YES
<10K samples
YES
number of categories known
NO
NO
MiniBatch KMeans
<10K samples
YES
MeanShift
VBGMM
NO
tough luck
predicting structure
NO
just looking
NO
YES
dimensionality reduction
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
Back
scikit learn
https://blog.csdn.net/code_fighter