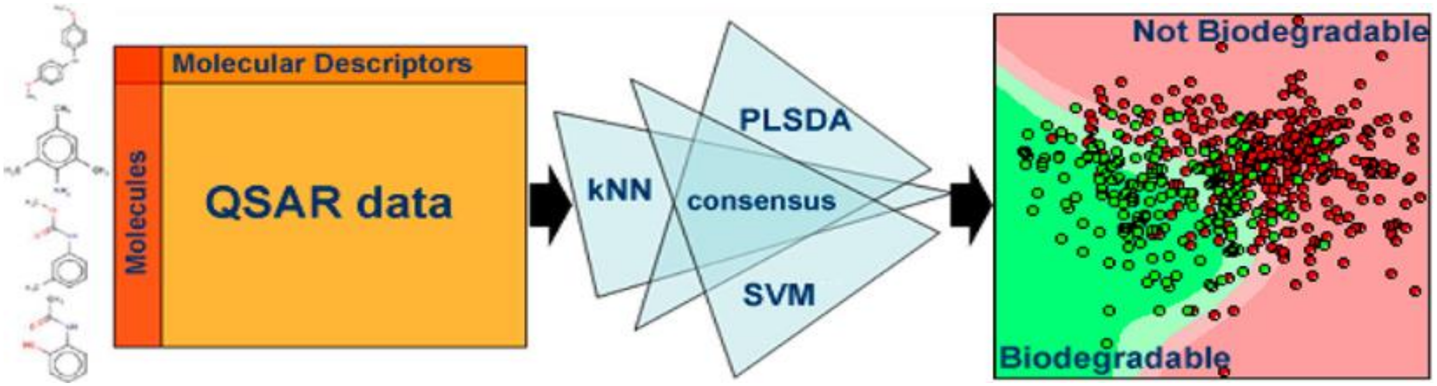
PREDICTING BIODEGRADABILITY

Jules Deplanchon
Mathys Bronnec
Lisa Charuel

How to predict if a compound is biodegradable from the study of the relationships between chemical structure and biodegradation of molecules ?



SUMMARY

I - Data pre-processing

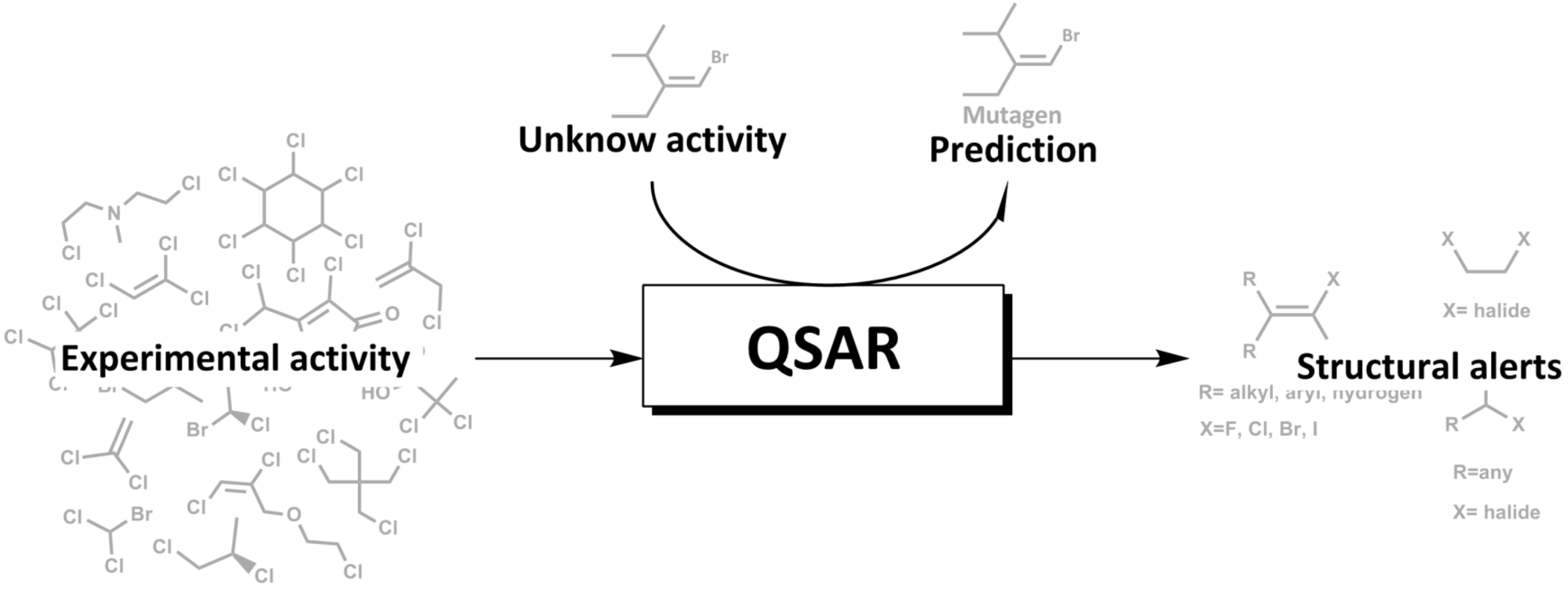
- Import
- Cleaning
- Normalization

II - Data visualization

- Matplotlib
- Correlation
- Seaborn

III – Modeling

- Random
- Logistic regression
- KNN
- Final model



THE DATA

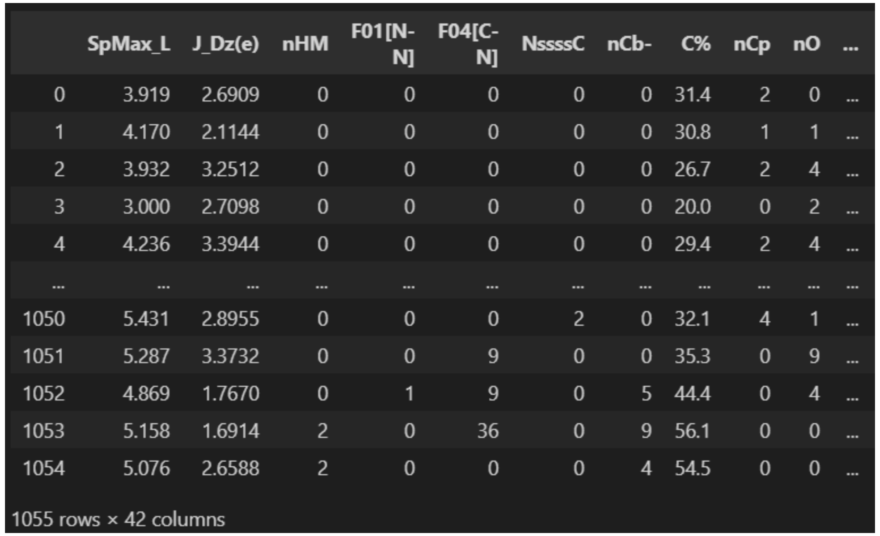
Dataset used : QSAR biodegradation

Data set containing values for **41 attributes** (molecular descriptors) used to classify 1055 chemicals into **2 classes** (ready and not ready biodegradable).

The data have been used to develop QSAR (**Quantitative Structure Activity Relationships**) models for the study of the relationships between chemical structure and biodegradation of molecules.

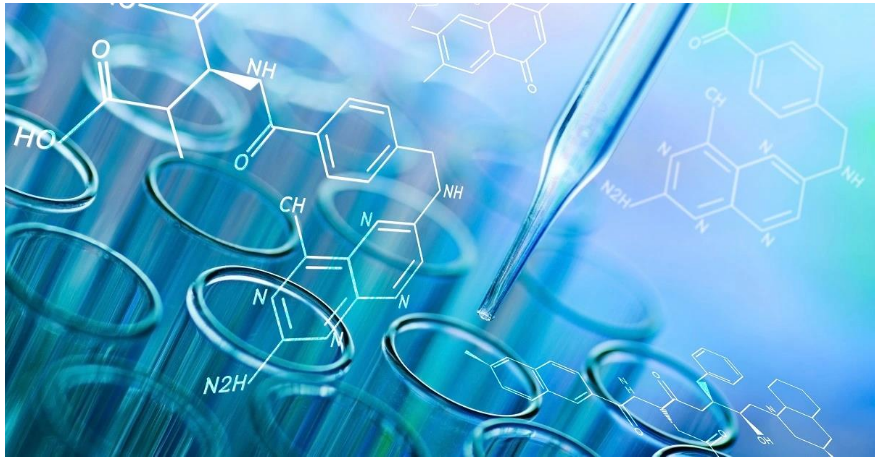
What the database allows :

- Binary classification
- Mathematical computation of properties



	SpMax_L	J_Dz(e)	nHM	F01[N-N]	F04[C-N]	NssssC	nCb-	C%	nCp	nO	...
0	3.919	2.6909	0	0	0	0	0	31.4	2	0	...
1	4.170	2.1144	0	0	0	0	0	30.8	1	1	...
2	3.932	3.2512	0	0	0	0	0	26.7	2	4	...
3	3.000	2.7098	0	0	0	0	0	20.0	0	2	...
4	4.236	3.3944	0	0	0	0	0	29.4	2	4	...
...	...	...	...	...	...	...	...	...	...	...	...
1050	5.431	2.8955	0	0	0	2	0	32.1	4	1	...
1051	5.287	3.3732	0	0	9	0	0	35.3	0	9	...
1052	4.869	1.7670	0	1	9	0	5	44.4	0	4	...
1053	5.158	1.6914	2	0	36	0	9	56.1	0	0	...
1054	5.076	2.6588	2	0	0	0	4	54.5	0	0	...

1055 rows × 42 columns

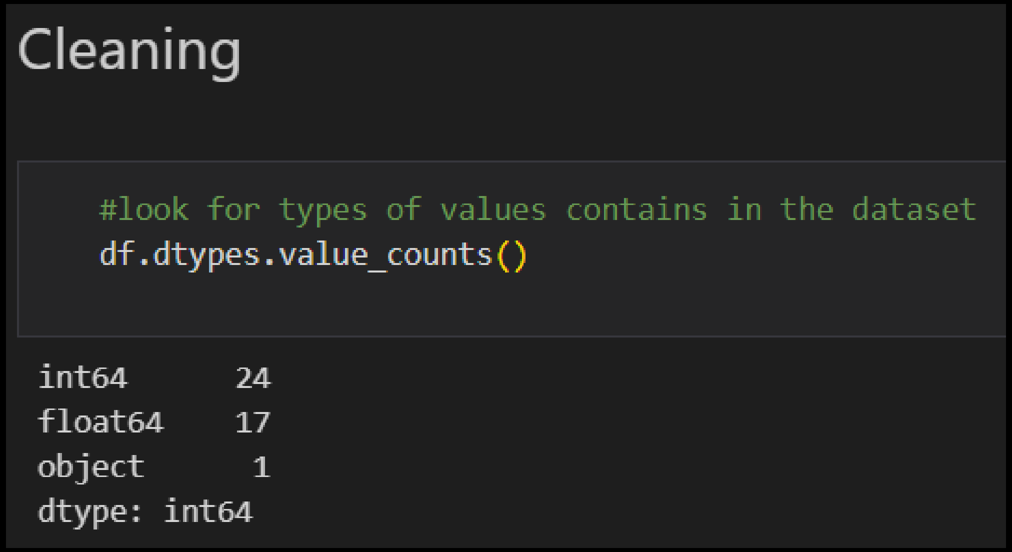


DATA PRE-PROCESSING

The DataFrame is missing column descriptions. These have to be added from the source website : [QSAR biodegradation - UCI Machine Learning Repository]

Only one object column : experimental class.
This is our target, we renamed it *classD*.

There is no missing value in the dataset.

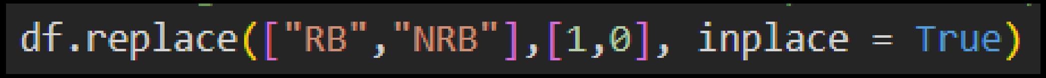


Cleaning

```
#look for types of values contains in the dataset
df.dtypes.value_counts()
```

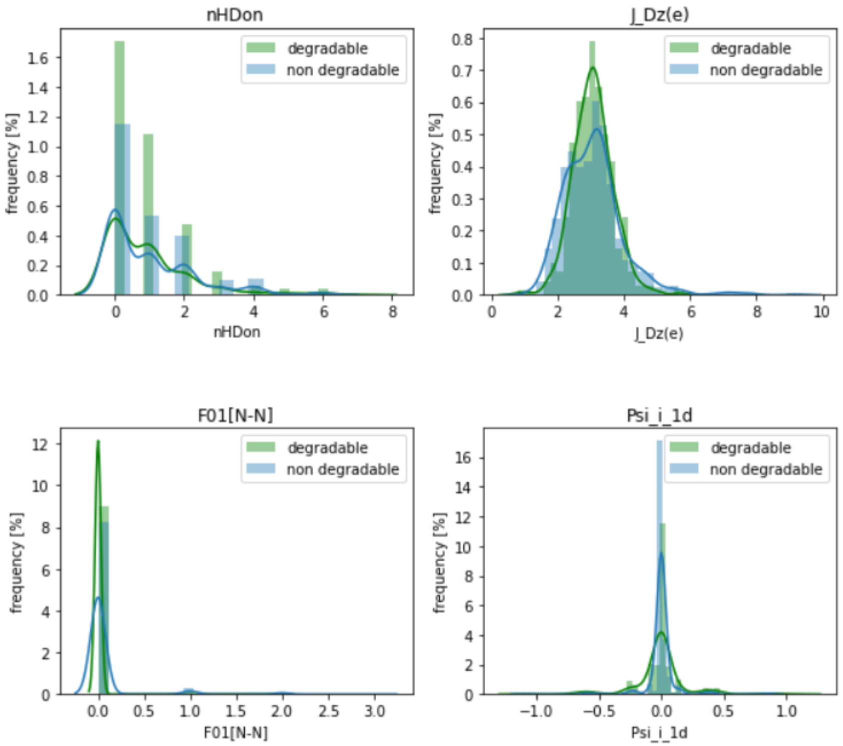
```
int64      24
float64    17
object      1
dtype: int64
```

```
The target class will be expressed as integers :
NRB = 0 --> Not Biodegradable
RB = 1 --> Biodegradable
```

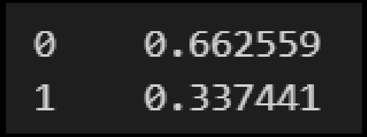


```
df.replace(["RB","NRB"],[1,0], inplace = True)
```

DATA VISUALIZATION



The plots shows how good the features are able to divide the degradable and nondegradable substances.

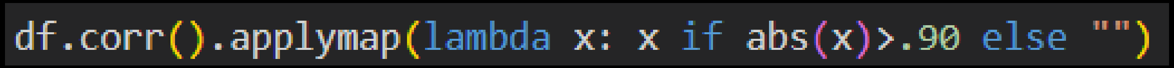


```
0    0.662559
1    0.337441
```

→ NRB

→ RB

These percentages provide insights into the class distribution in the "degradable" column of the dataset => shows an imbalance.



```
df.corr().applymap(lambda x: x if abs(x)>.90 else "")
```

This operation is useful for highlighting strong correlations between variables while removing highly correlations (|r| > 0.90).
=> "SM6_L","SpMax_A","SM6_B(m)“ dropped

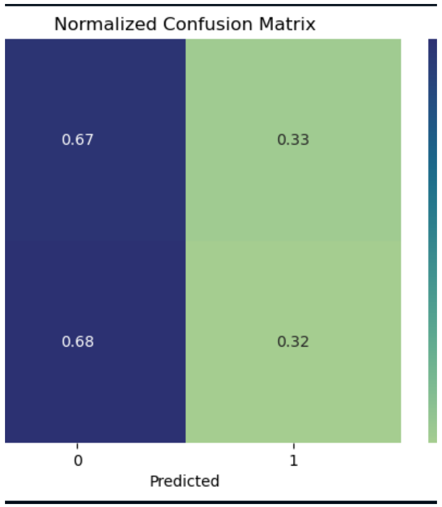
The distribution show that the best prediciting features for class seperation are:

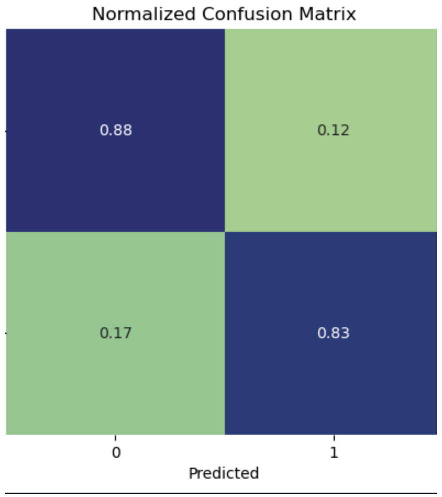
- SpPosA_B(p)
- HyWi_B(m)
- C%
- SpMax_B(m)
- SpMax_L

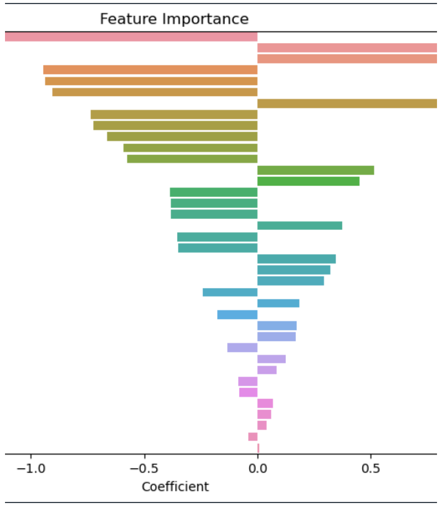
MODELING

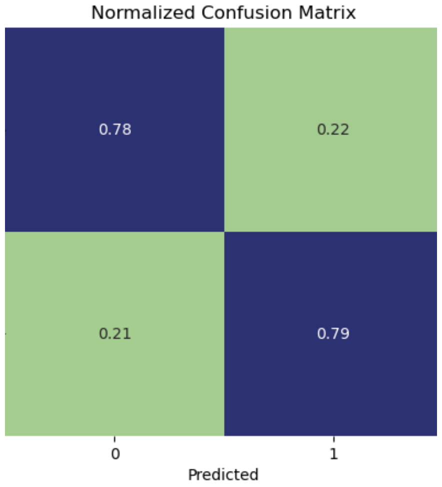
In the context of modeling, we have decided to use different algorithms:

- Random
- Linear Regression
- KNN (k-Nearest Neighbors)







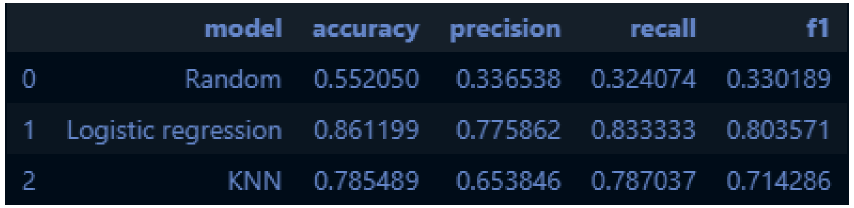


CONCLUSION

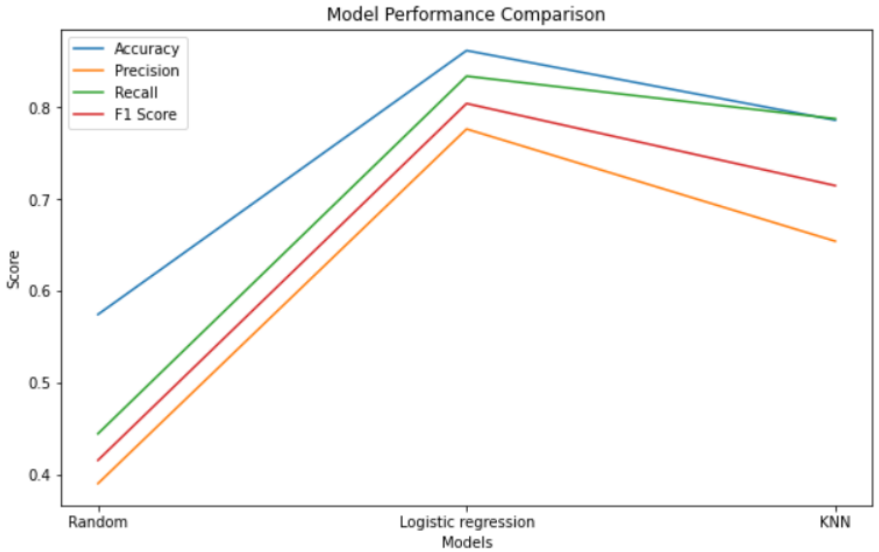
To explain the results of the above model classification, we need to analyze the performance metrics of each model.

The model with higher accuracy, precision, recall, and F1 score is considered better in terms of classification performance.

We can see that the best model tested for this classification problem is the Logistic regression model.



	model	accuracy	precision	recall	f1
0	Random	0.552050	0.336538	0.324074	0.330189
1	Logistic regression	0.861199	0.775862	0.833333	0.803571
2	KNN	0.785489	0.653846	0.787037	0.714286



THANK YOU FOR YOUR ATTENTION

Jules Deplanchon
Mathys Bronnec
Lisa Charuel

SOURCES

https://archive.ics.uci.edu/dataset/254/qsar+biodegradation

https://www.neuraldesigner.com/learning/examples/qsar-biodegradation/#DataSet

https://pubs.acs.org/doi/10.1021/ci4000213