QSAR BIODEGRADATION

DATA SET

Ulrich MAMA DIA4

TABLE OF CONTENTS

01

PROBLEM & SOLUTION

02

OUR PROCESS

03

SOLUTION

What is the RQSA?

- Structure-activity relationship (SAR) and quantitative structure-activity relationship (QSAR) models - collectively referred to as (Q)SAR - are theoretical models that can be used quantitatively or qualitatively to predict the physico-chemical, biological (e.g. a toxicological effect) and environmental fate of compounds based on knowledge of their chemical structure.
- An RSA is a qualitative relationship between a (sub)structure and the presence or absence of a propriété or activité under consideration.
- A QSAR is a mathematical model that combines one or more quantitative parameters derived from the chemical structure with a quantitative measurement from a propriété or activité.

Description of the dataset

- The QSAR biodegradation dataset was designed within the Chemometrics and QSAR research group in Milan (University of Milan). The data were used to develop QSAR (Quantitative Structure Activity Relationships) models for the study of the relationships between chemical structure and biodegradation of molecules.
- Experimental biodegradation values of 1055 chemicals were collected to compose this dataset.

Description of the dataset

We have a dataset of molecules with the different characteristics of which they are composed.
There are 41 variables helping to describe each chemical product as well as 2 classes RB (Ready Biodegradable) and NRB (Not Ready Biodegradable), these classes help to know if the chemical product is biodegradable or not.

Biodegradable: A substance is said to be biodegradable if, under the action of living organisms external to its substance, it can break down into various elements, "with no harmful effect on the natural environment" (according to French legislation), carbon dioxide CO2, water, methane.

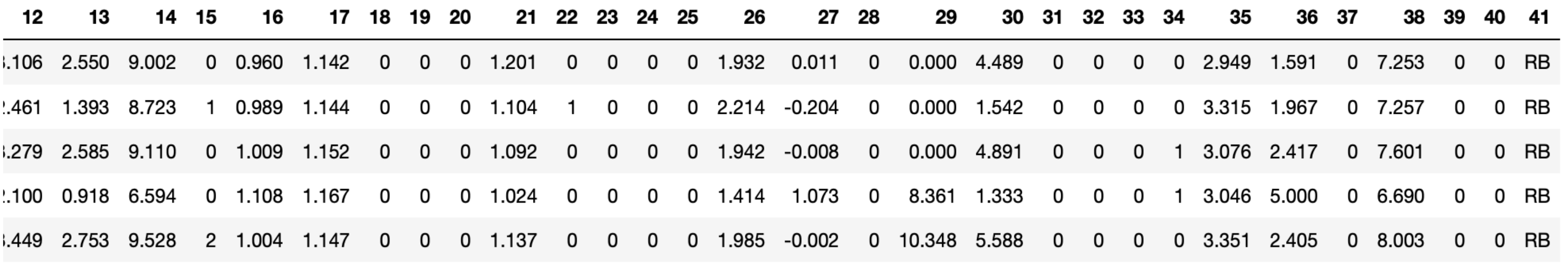
Description of the dataset

There are no labels for the different columns (no column name), we just have the 2 classes to know if the product is biodegradable or not. Thus, we are facing a classification problem.

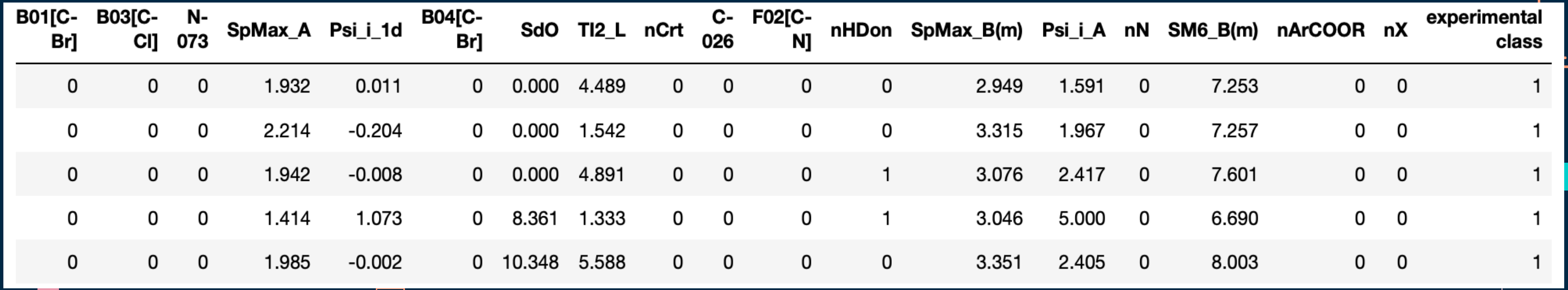
MAIN PURPOSE

Thus, the main objective of this problem is to use the dataset data in order to build the most reliable QSAR possible for the future chemicals to be studied.

BEFORE AND AFTER CLEANED THE DATASET

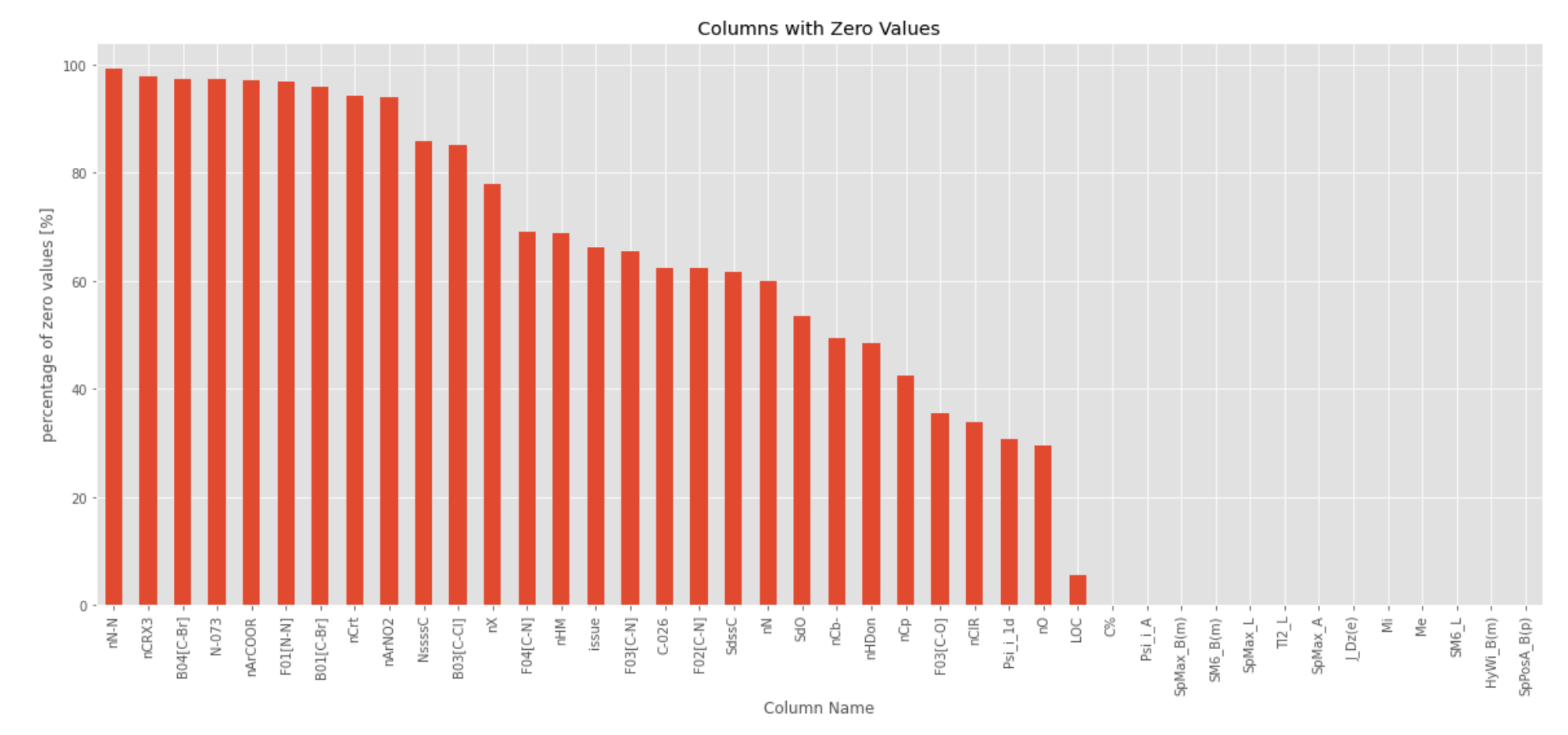


12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
.106	2.550	9.002	0	0.960	1.142	0	0	0	1.201	0	0	0	0	1.932	0.011	0	0.000	4.489	0	0	0	0	2.949	1.591	0	7.253	0	0	RB
.461	1.393	8.723	1	0.989	1.144	0	0	0	1.104	1	0	0	0	2.214	-0.204	0	0.000	1.542	0	0	0	0	3.315	1.967	0	7.257	0	0	RB
.279	2.585	9.110	0	1.009	1.152	0	0	0	1.092	0	0	0	0	1.942	-0.008	0	0.000	4.891	0	0	0	1	3.076	2.417	0	7.601	0	0	RB
.100	0.918	6.594	0	1.108	1.167	0	0	0	1.024	0	0	0	0	1.414	1.073	0	8.361	1.333	0	0	0	1	3.046	5.000	0	6.690	0	0	RB
.449	2.753	9.528	2	1.004	1.147	0	0	0	1.137	0	0	0	0	1.985	-0.002	0	10.348	5.588	0	0	0	0	3.351	2.405	0	8.003	0	0	RB

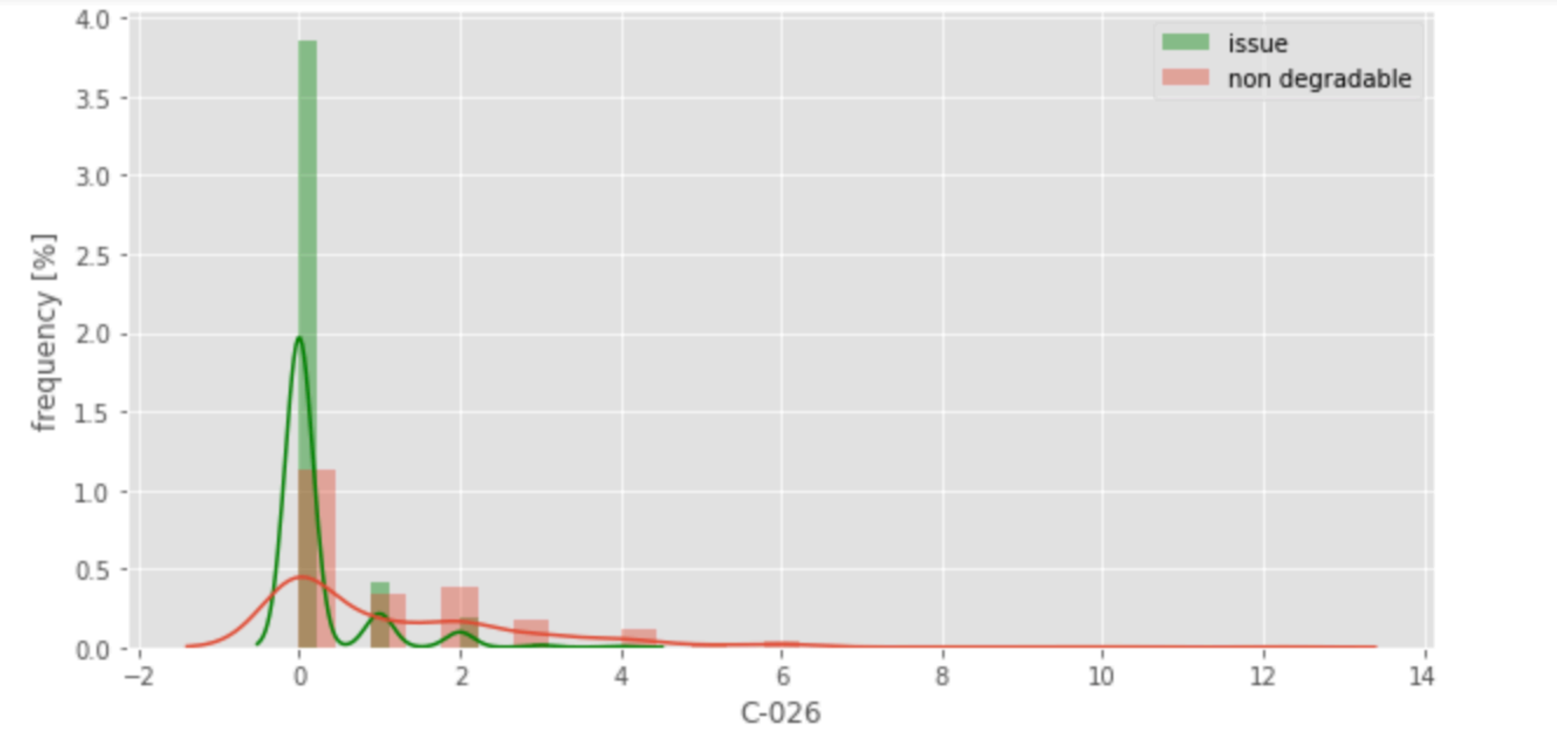


B01[C-Br]	B03[C-Cl]	N-073	SpMax_A	Psi_i_1d	B04[C-Br]	SdO	TI2_L	nCrt	C-026	F02[C-N]	nHDon	SpMax_B(m)	Psi_i_A	nN	SM6_B(m)	nArCOOR	nX	experimental class
0	0	0	1.932	0.011	0	0.000	4.489	0	0	0	0	2.949	1.591	0	7.253	0	0	1
0	0	0	2.214	-0.204	0	0.000	1.542	0	0	0	0	3.315	1.967	0	7.257	0	0	1
0	0	0	1.942	-0.008	0	0.000	4.891	0	0	0	1	3.076	2.417	0	7.601	0	0	1
0	0	0	1.414	1.073	0	8.361	1.333	0	0	0	1	3.046	5.000	0	6.690	0	0	1
0	0	0	1.985	-0.002	0	10.348	5.588	0	0	0	0	3.351	2.405	0	8.003	0	0	1

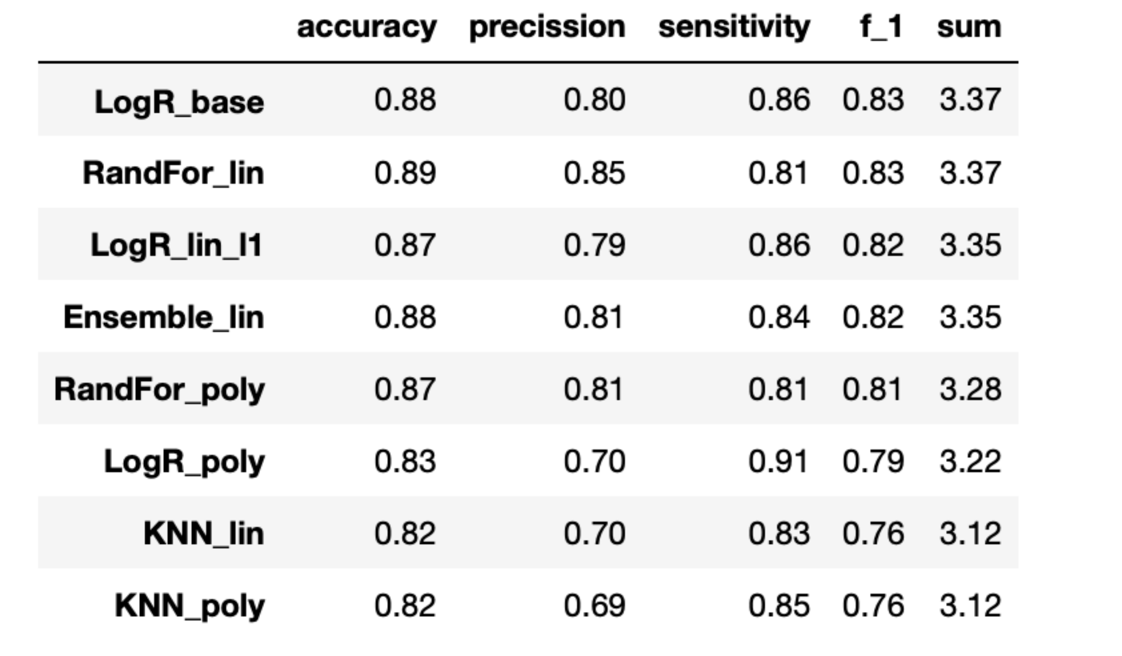
THERE IS A LOT OF COLUMN WITH THE NBUMBER 0



THERE IS A LOT OF PRODUCT DEGRAGABLE (GREEN)



We can see that when we add, all parameters, the best model is the Logistic Regression



	accuracy	precission	sensitivity	f_1	sum
LogR_base	0.88	0.80	0.86	0.83	3.37
RandFor_lin	0.89	0.85	0.81	0.83	3.37
LogR_lin_l1	0.87	0.79	0.86	0.82	3.35
Ensemble_lin	0.88	0.81	0.84	0.82	3.35
RandFor_poly	0.87	0.81	0.81	0.81	3.28
LogR_poly	0.83	0.70	0.91	0.79	3.22
KNN_lin	0.82	0.70	0.83	0.76	3.12
KNN_poly	0.82	0.69	0.85	0.76	3.12

THANK U