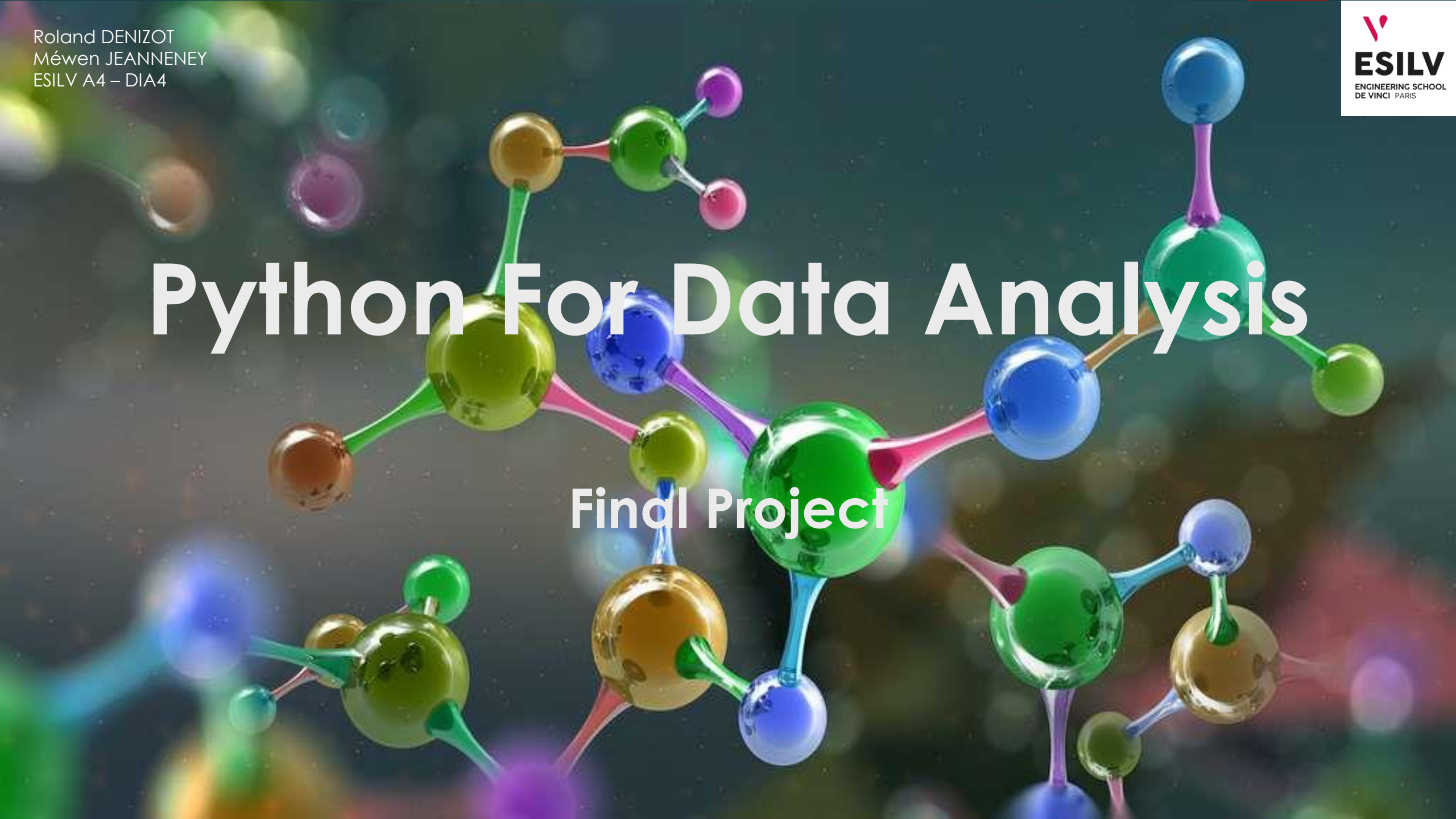
Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS



Python For Data Analysis

Final Project

Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Ins and Outs

Our data set named « QSAR Biodegradation Data Set » has been used to develop Quantitative Structure Activity Relationships models for the study of the relationships between chemical structures and biodegradation of molecules.
It contains experimental values of 1055 chemicals with 41 molecular descriptors for each of them and 1 experimental class. These data come from the webpage of the National Institute of Technology and Evaluation of Japan (NITE).

The objective is, by analyzing these features, to classify these chemicals into two classes: Ready and Not Ready Biodegradable.

Here are the first lines of the dataset:

	SpMax_L	J_Dz(e)	nHM	F01[N-N]	F04[C-N]	NssssC	nCb-	C%	nCp	nO	F03[C-N]	SdssC	HyWi_B(m)	LOC	SM6_L	F03[C-O]	Me	Mi	nN-N	nArNO2	nCRX3	SpPosA_B(p)	nCIR	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	3.919	2.6909	0	0	0	0	0	31.4	2	0	0	0.000	3.106	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
1	4.170	2.1144	0	0	0	0	0	30.8	1	1	0	0.000	2.461	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
2	3.932	3.2512	0	0	0	0	0	26.7	2	4	0	0.000	3.279	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
3	3.000	2.7098	0	0	0	0	0	20.0	0	2	0	0.000	2.100	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
4	4.236	3.3944	0	0	0	0	0	29.4	2	4	0	-0.271	3.449	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
5	4.236	3.4286	0	0	0	0	0	28.6	2	4	0	-0.275	3.313	2.522	9.383	1	1.014	1.149	0	0	0	1.119	0	0	0	0	1.980	-0.008	0	10.276	4.746	0	0	0	0	3.351	2.556	0	7.904	0	0	RB

Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

Thoughts on the asked question

In order to better predict the class, we must first understand the meaning of each variables in the dataset.
After that, we have to measure the relevancy of each variable, in order to focus first on the most important variables and make a good prediction from the start of the study.
41 Variables may seem like a lot to us, there might have some of them creating more noise than help to well visualise the class of each chemical.
This is why we focused on the most important variables to plot and visualise the link betwwen each variables and the link between the variables and the target.

Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Variables Created

First of all, we created a first variable named « full_table_non_triee », this is the raw data of the dataset.
Here are the first rows of the raw DataFrame:

	0	1	2	3	4	5	6	7	8	9	10	11	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
0	3.919	2.6909	0	0	0	0	0	31.4	2	0	0	0.000	3.106	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
1	4.170	2.1144	0	0	0	0	0	30.8	1	1	0	0.000	2.461	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
2	3.932	3.2512	0	0	0	0	0	26.7	2	4	0	0.000	3.279	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
3	3.000	2.7098	0	0	0	0	0	20.0	0	2	0	0.000	2.100	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
4	4.236	3.3944	0	0	0	0	0	29.4	2	4	0	-0.271	3.449	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

Then, after a preprocessing, we obtained a cleaned DataFrame and named it simply « df ».
Here are the first rows of the cleaned DataFrame:

	SpMax_L	J_Dz(e)	nHM	F01[N-N]	F04[C-N]	NssssC	nCb-	C%	nCp	nO	F03[C-N]	SdssC	HyWi_B(m)	LOC	SM6_L	F03[C-O]	Me	Mi	nN-N	nArNO2	nCRX3	SpPosA_B(p)	nCIR	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	3.919	2.6909	0	0	0	0	0	31.4	2	0	0	0.000	3.106	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	1
1	4.170	2.1144	0	0	0	0	0	30.8	1	1	0	0.000	2.461	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	1
2	3.932	3.2512	0	0	0	0	0	26.7	2	4	0	0.000	3.279	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	1
3	3.000	2.7098	0	0	0	0	0	20.0	0	2	0	0.000	2.100	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	1
4	4.236	3.3944	0	0	0	0	0	29.4	2	4	0	-0.271	3.449	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	1

Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

Variables Created

For all the visualizations, we used to create some variables whose names was made to be as clear as possible.
For example « fig » refers to the figure we initialized, « model » refers to the model we chose to make predictions, all the names with « data » in it refers to a subset of the DataFrame to make some specific visualizations, « ax » refers to the creation of a plot, « x » refers to the abscissa and « y » to the ordinate of a graph.

Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

Context of the problem in our studies

This problem is twice linked to our studies:

- First, we study programmation since 4 years and python since 2 years, so it seems normal to us to do a data analysis using this programmation language. Python is the most suitable to analyse a data set because of its integrated libraries like Pandas, Matplotlib and Scikit Learn.
- Second, we are in the Data and Artificial Intelligence major, that's why it is accurate to make predictions using different models of Machine Learning in a part of this project.

Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

Reasonning

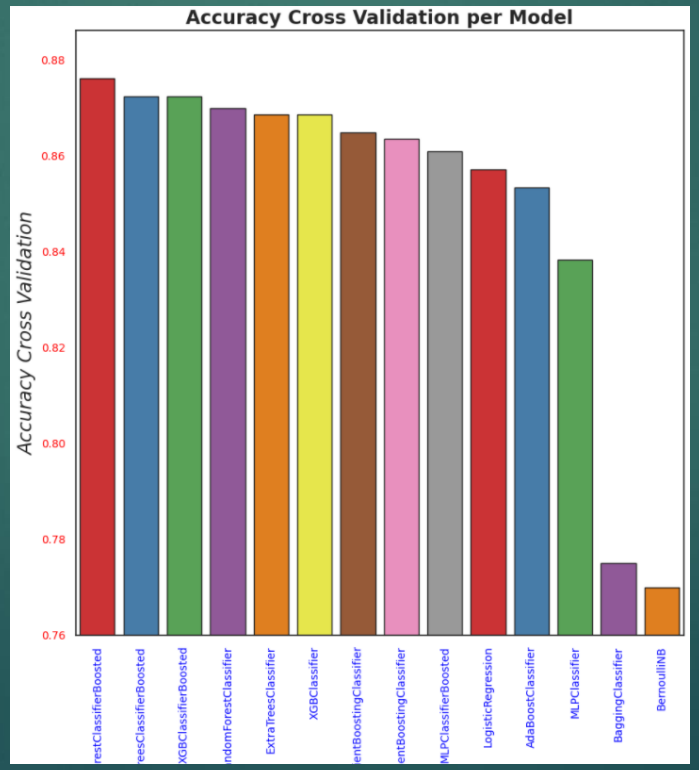
The execution of this project is made by several steps:

- We first called essential libraries that already exist in python (Numpy, Pandas, Matplotlib, Scikit Learn, Seaborn)
- Then we have imported the dataset from its original link to run the algorithm from everywhere without the need to include the csv file into the folder.
- Then, we pre-processed the data and we cleaned it because we have found that the name of the columns was not included in the file for example.
- After that, we started visualization and we started by calculating the most important variables because there were too much variables to make relevant visualizations.
- One these most important variables known, we started to make visualizations and compare variables between each of them and the target.
- After visualizating the most important variables, we could consider that we understood the purpose of the problem and we could start modeling and machine learning. We tried about ten models and we measured the accuracy of each of them with cross validation to find the best one.
- Then, we took the best models and we customized the hyperparameters in order to improve again these models and obtain the best accuracy possible.

Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

Reasonning

- We wanted to notify that we tried several modifications of the variables like scaling for example but we didn't find better results with these manipulations so we decided not to keep them in the final project but we kept them in commentary to show our tries.
- The results might change if we re-run the algorithm because there is a split into training and test sets but the top models doesn't change very much.



Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

Django API

Due to several problems with django, I couldn't be possible to launch the app or create a Django Project.
Moreover, we still created a file based on a Django project that we did last year while modifying it for this topic. In theory, therefore, it works but it is impossible to verify it due to the bug we are encountering with our Django application.
The Django API contains 5 parts. 4 of these are visualisations with interactions with the users and the last one is the modelisation of ten machine learning models and the performance of each of them.

Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Difficulties

As the subject is very specific and we are not specialized in it, the first difficulty was to well understand all the variables.
After that, we had to discover a lot of new python functions and libraries, so it could take a long time to be comfortable with them but at the end, we could consider that we succeed to understand all of them.
Another difficulty was for modeling because the code of each classifier is different and it is always a different method to calculate hyperparameters etc.
The last difficulty was to make a Django API because it didn't run well on our PC so we started from a project we've done on Year 3 and we adapted the file to the actual subject without a way to check our work, so it might be not perfect but we still made the maximum on this API.

Roland DENIZOT
Méwen JEANNENEY
ESILV A4 – DIA4

Team

- Méwen JEANNENEY
- Roland DENIZOT