PYTHON FOR DATA ANALYSIS

PROJECT

CASSIE DOGUET - MAEL DARNAUD - MATHURIN DE CRECY

30/11/2023

SUMMARY

01 PRESENTATION OF THE DATASET

02 THE INS AND OUTS AND QUESTIONS

03 DATA PREPROCESSING

04 DATA VISUALISATION

05 MODELING

06 TRANSFORMATION INTO API

07 CONCLUSION

PRESENTATION OF THE DATASET

QSAR BIODEGRADATION

QSAR : Quantitative Structure-Activity Relationship

Predict the biological activity of chemical products using its molecular structure

Disponible on **UCI Machine Learning Repository**

PRESENTATION OF THE DATASET

	SpMax_L	J_Dz(e)	nHM	F01[N-N]	F04[C-N]	NssssC	nCb-	C%	nCp	nO	...	C-026	F02[C-N]	nHDon	SpMax_B(m)	Psi_i_A	nN	SM6_B(m)	nArCOOR	nX	experimental class
0	4.170	2.1144	0	0	0	0	0	30.8	1	1	...	0	0	0	3.315	1.967	0	7.257	0	0	RB
1	3.932	3.2512	0	0	0	0	0	26.7	2	4	...	0	0	1	3.076	2.417	0	7.601	0	0	RB
2	3.000	2.7098	0	0	0	0	0	20.0	0	2	...	0	0	1	3.046	5.000	0	6.690	0	0	RB
3	4.236	3.3944	0	0	0	0	0	29.4	2	4	...	0	0	0	3.351	2.405	0	8.003	0	0	RB
4	4.236	3.4286	0	0	0	0	0	28.6	2	4	...	0	0	0	3.351	2.556	0	7.904	0	0	RB

Chimical structure

Physico-chemical properties

Indicate presence of atoms

Cycles

Hydrogen

Oxygen

Nitrogen

Carbon

PRESENTATION OF THE DATASET

Prédiction on biodegrability

Predict biodegradability of new cheminal compounds before production or release into environment

Chemical regulation

Regulatory agencies can refer to the dataset to evaluate biodegradability Ensure that they meet specific environmental standards

Environmental impact

Evaluation of the environmental impact of chemical substances. Gide decision making regarding regulation of chemicals

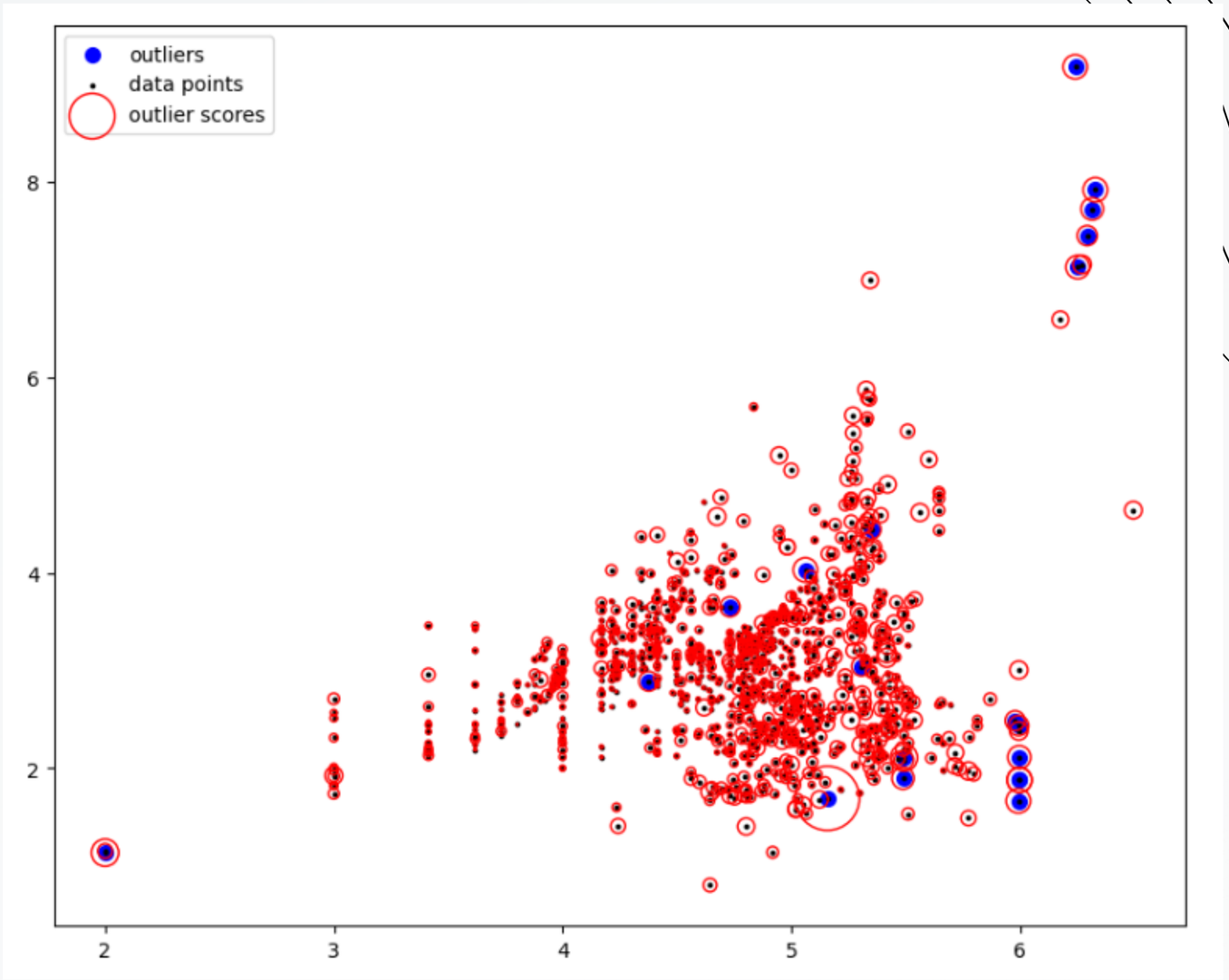
Chemical design and optimization

Resource for designing and optimizing chemical structure with improved biodegradability.
Make greener and more sustainable chemical compounds

DATA PREPROCESING

Removing Outliers :

sklearn.neighbors.LocalOutlierFactor



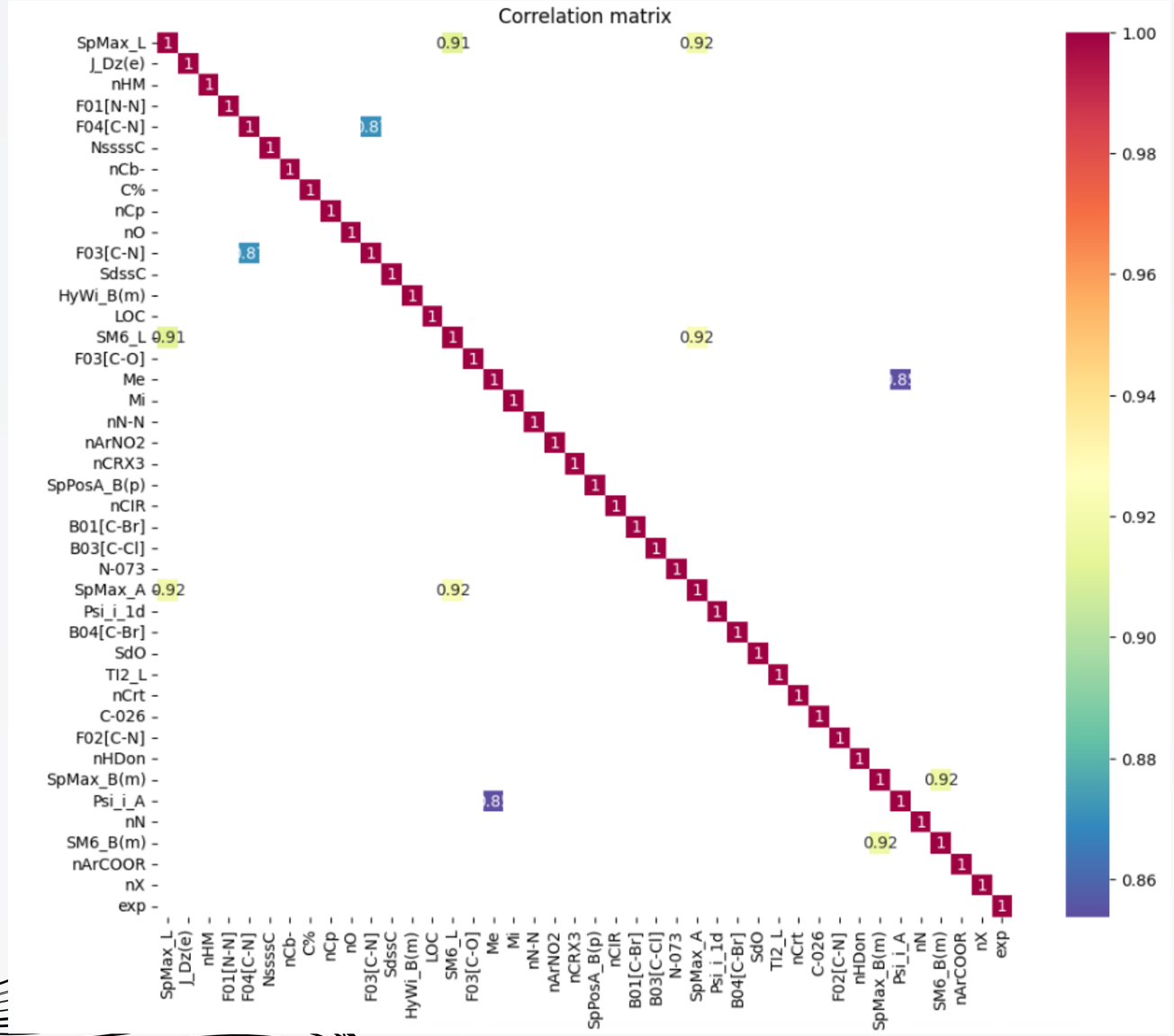
DATA PREPROCESING

Normalization :

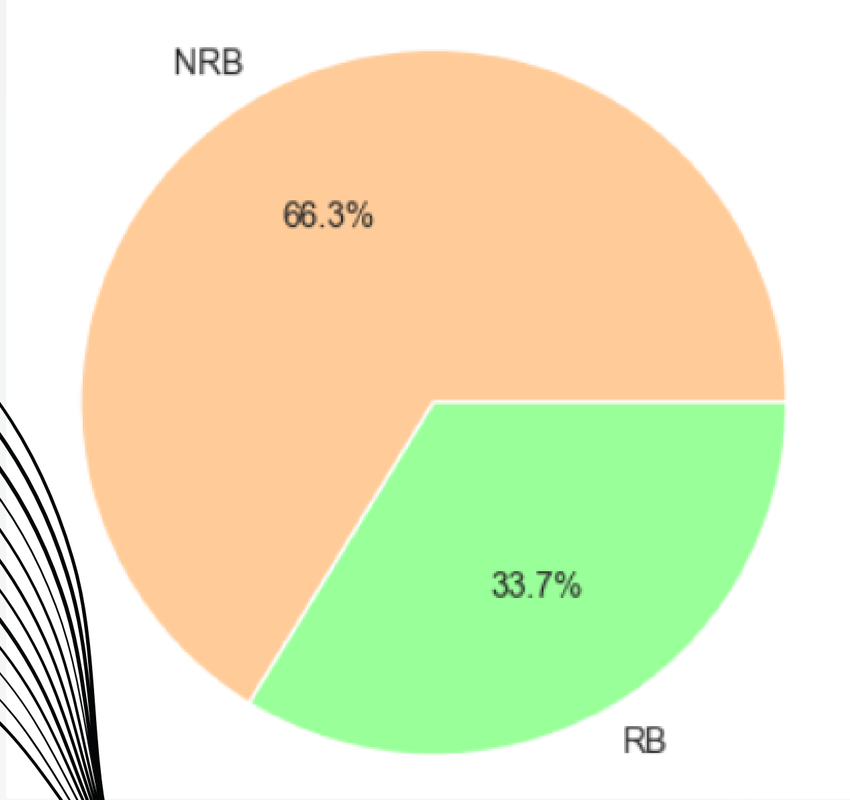
sklearn.preprocessing.MinMaxScaler

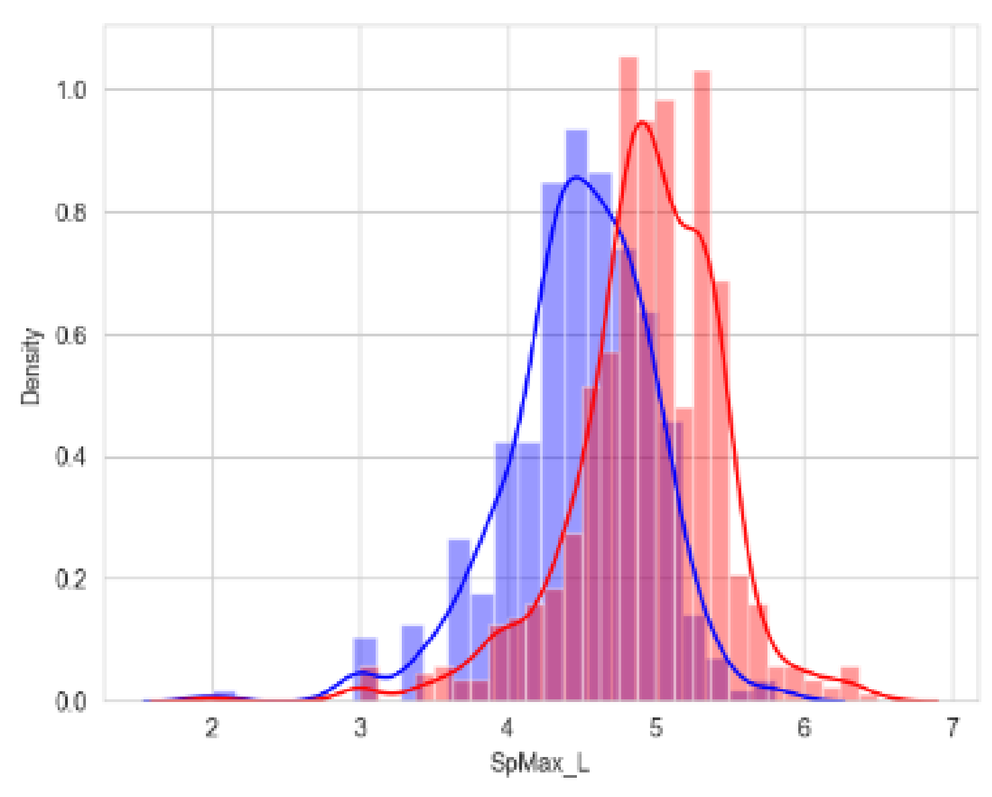
	SpMax_L	J_Dz(e)	nHM	F01[N-N]	F04[C-N]	NssssC	nCb-	C%	nCp	nO	F03[C-N]	SdssC	HyWi_B(m)	LOC
0	0.426824	0.297133	0.000000	0.000000	0.000000	0.000000	0.000000	0.432171	0.111111	0.000000	0.000000	0.526759	0.375815	0.567802
1	0.482651	0.206355	0.000000	0.000000	0.000000	0.000000	0.000000	0.420543	0.055556	0.083333	0.000000	0.526759	0.220130	0.310176
2	0.429715	0.385359	0.000000	0.000000	0.000000	0.000000	0.000000	0.341085	0.111111	0.333333	0.000000	0.526759	0.417572	0.575596
3	0.222420	0.300109	0.000000	0.000000	0.000000	0.000000	0.000000	0.211240	0.000000	0.166667	0.000000	0.526759	0.132995	0.204409
4	0.497331	0.407908	0.000000	0.000000	0.000000	0.000000	0.000000	0.393411	0.111111	0.333333	0.000000	0.499599	0.458605	0.613004

DATA PREPROCESING

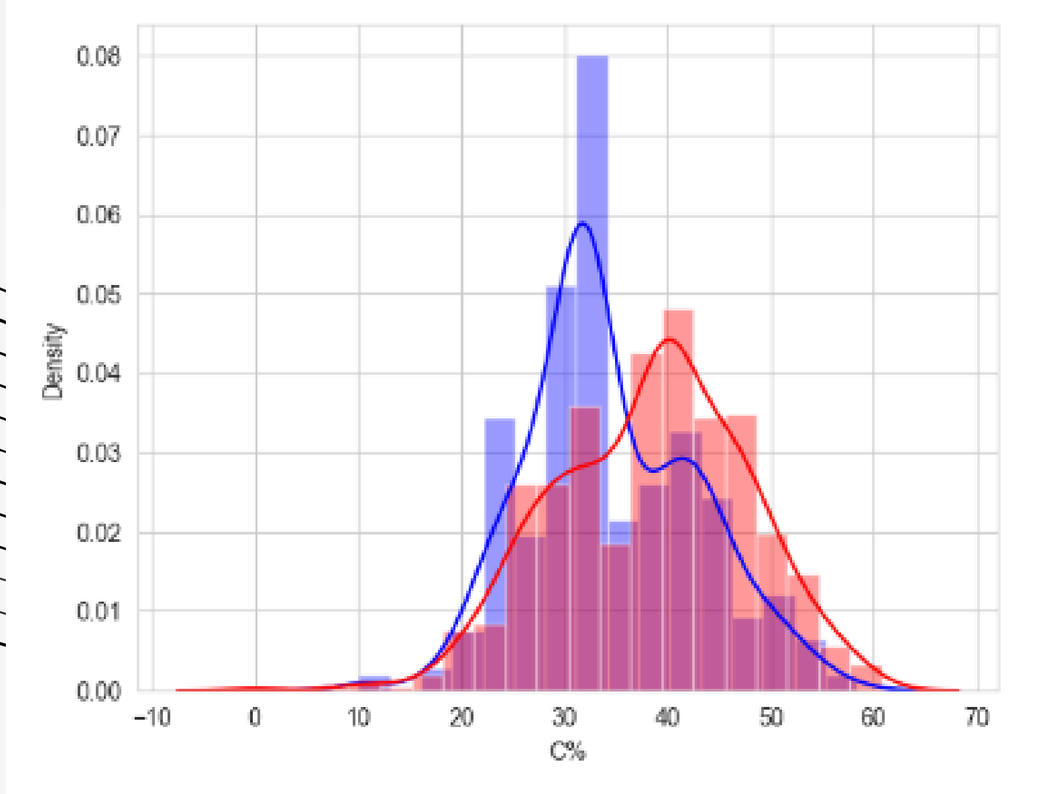


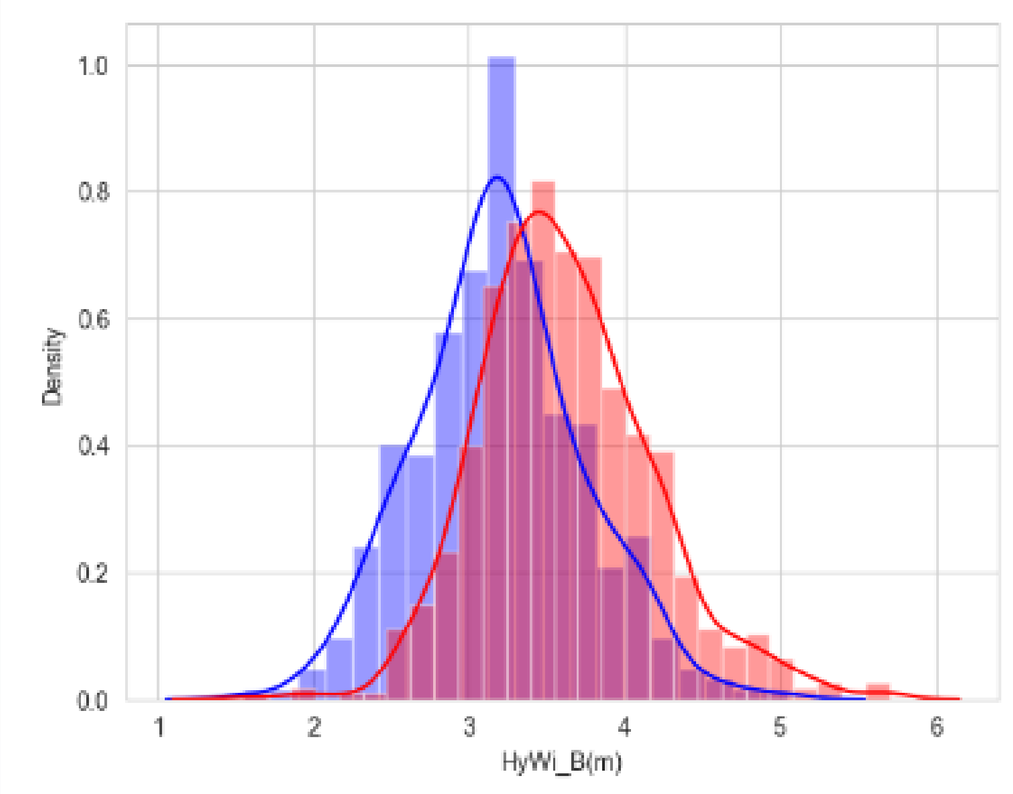
DATA VISUALISATION



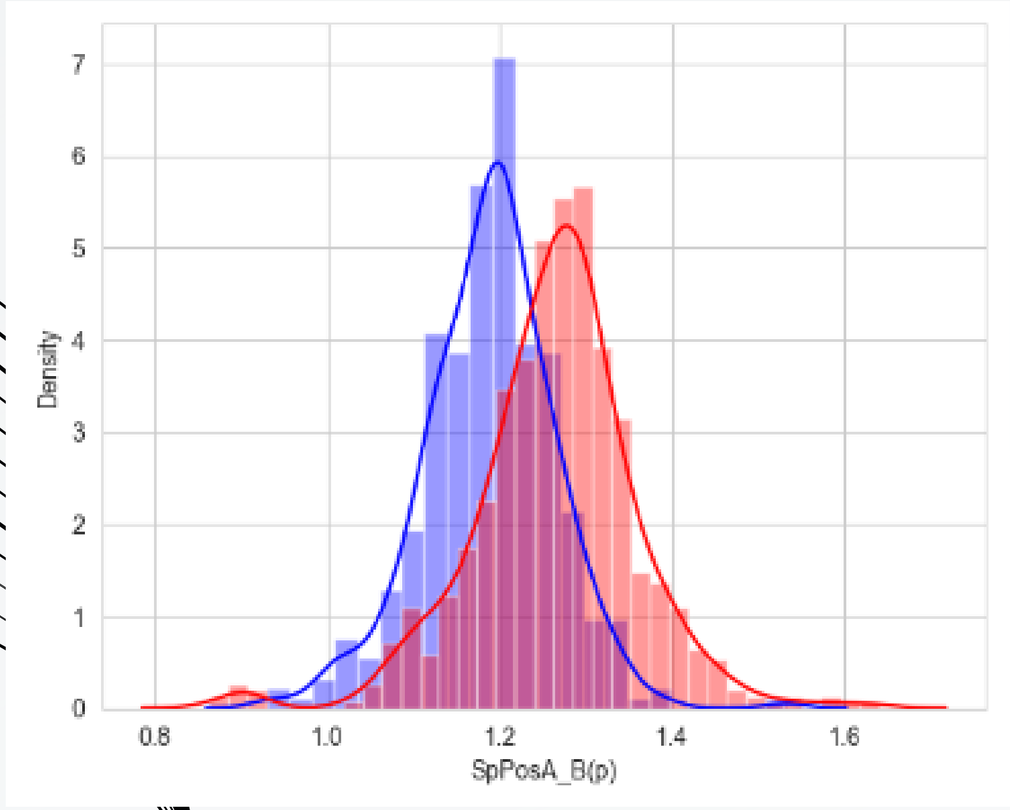


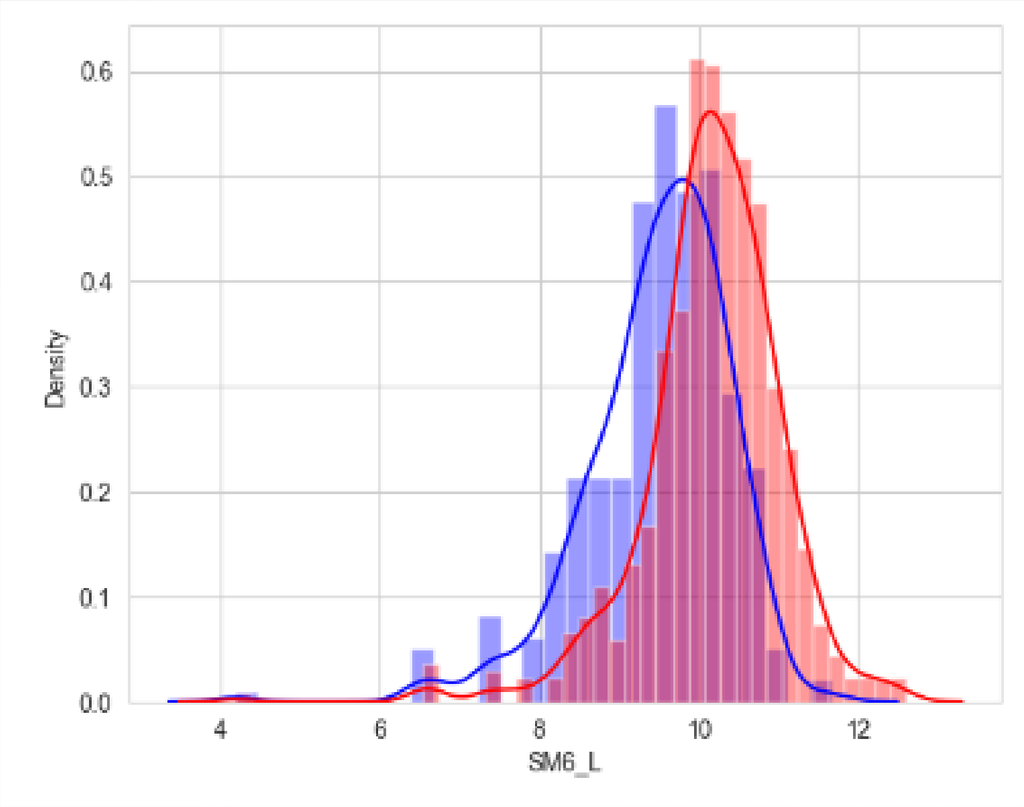
DATA VISUALISATION





DATA VISUALISATION



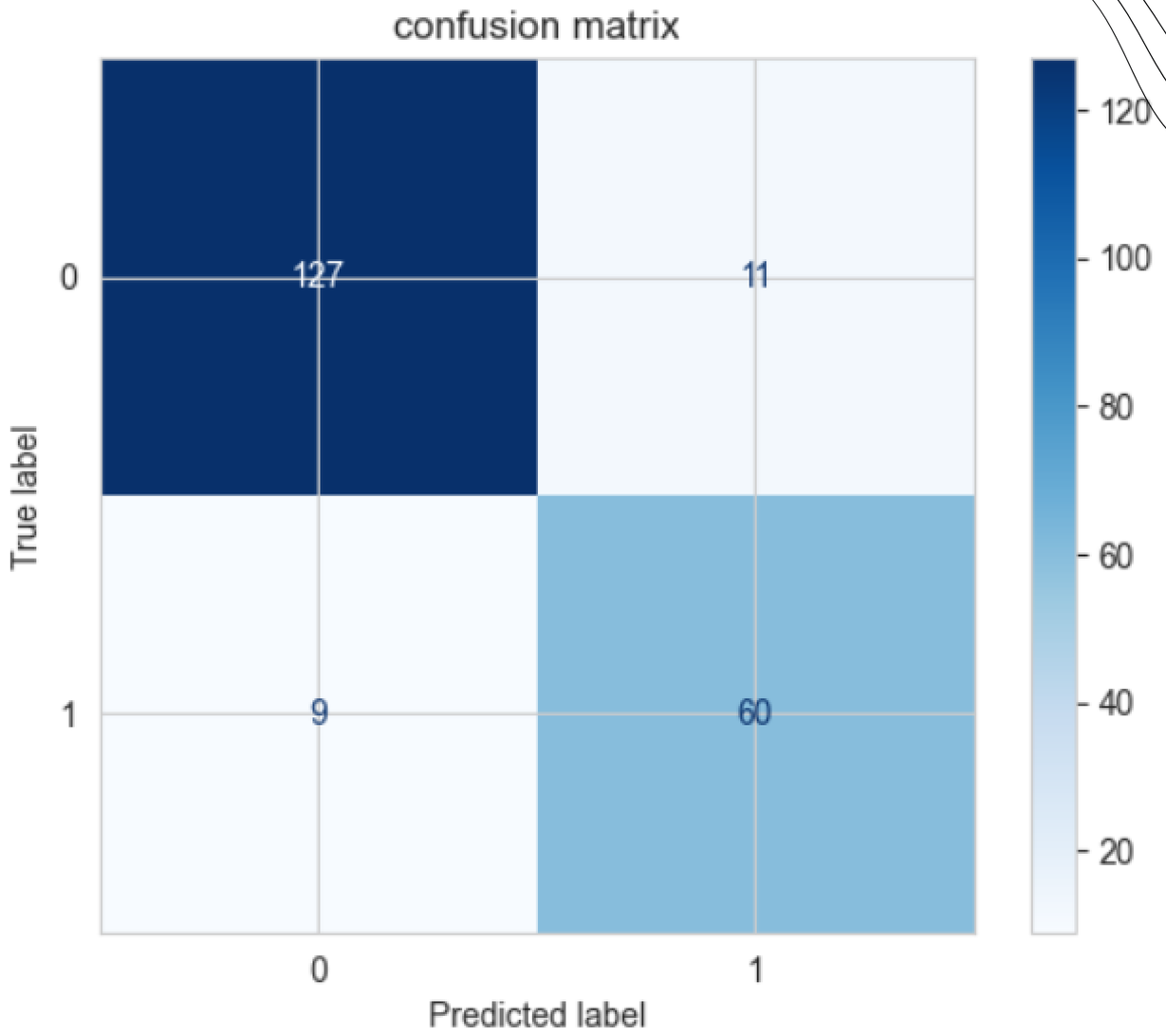


MODELING: KNN

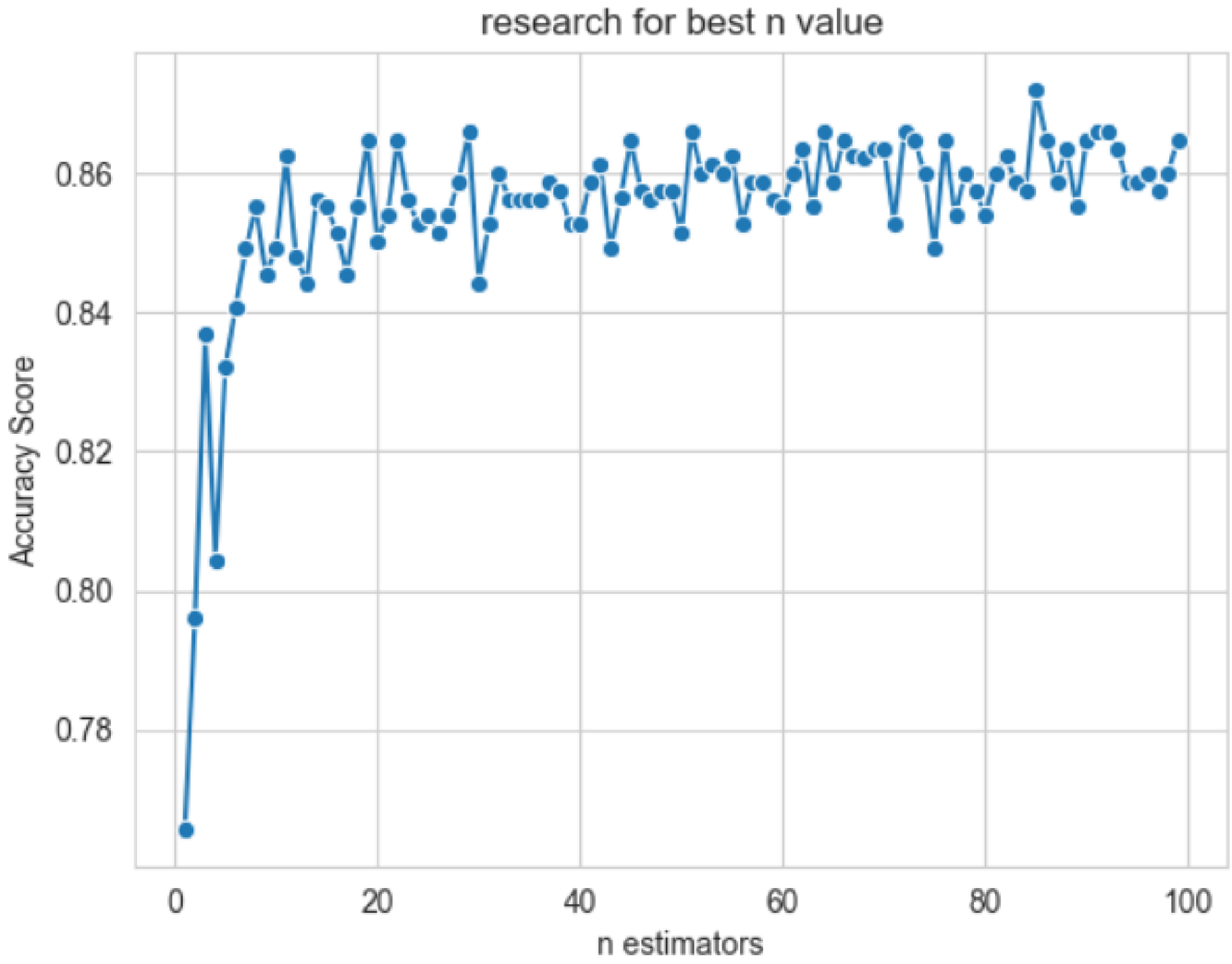
MODELING: KNN

Mean squarred error: 0.311
F1 score: 0.845
Accuracy: 0.869
Recall: 0.857
Precision: 0.903

(Run time : 38.8s)



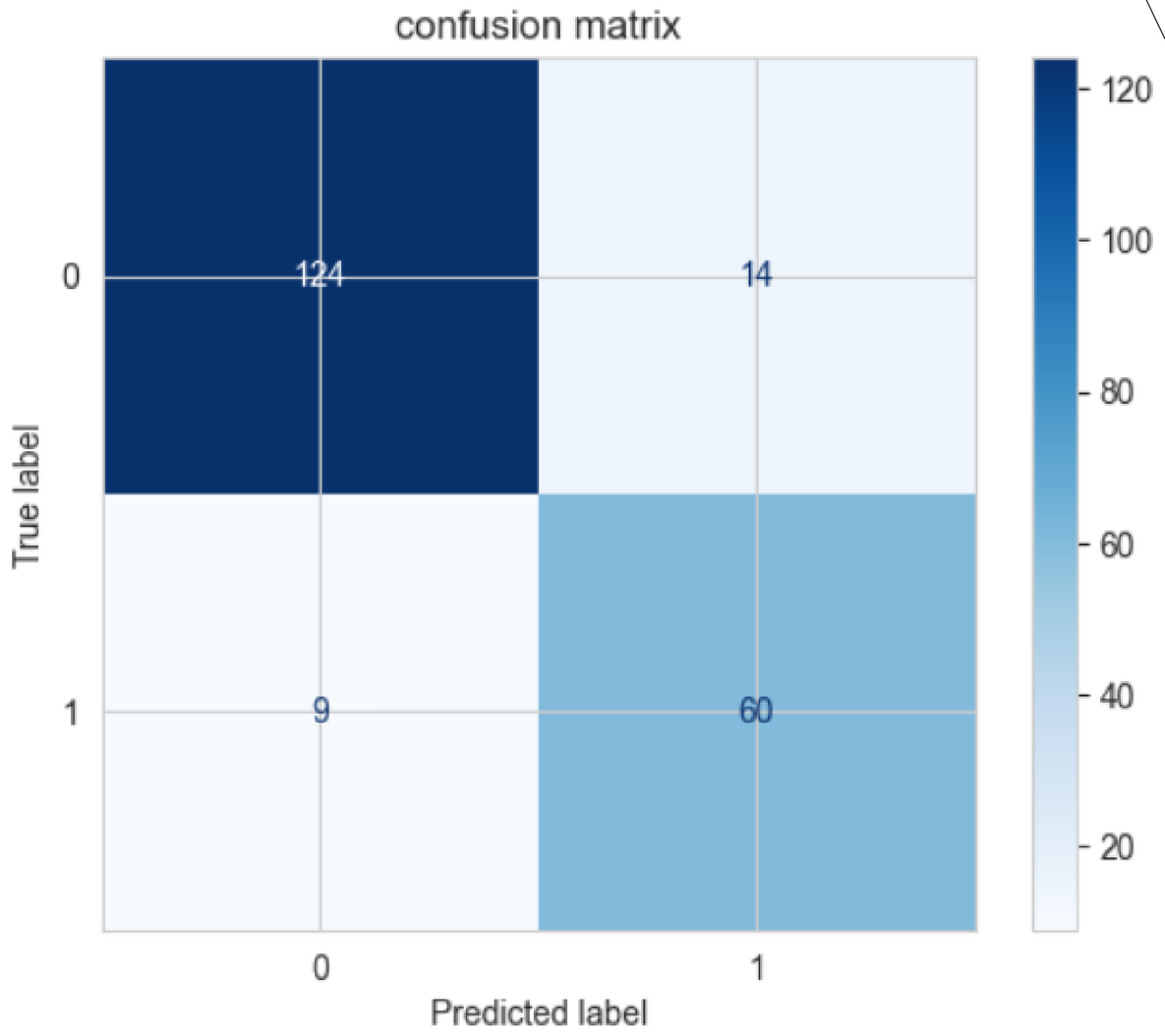
MODELING: RANDOM FOREST



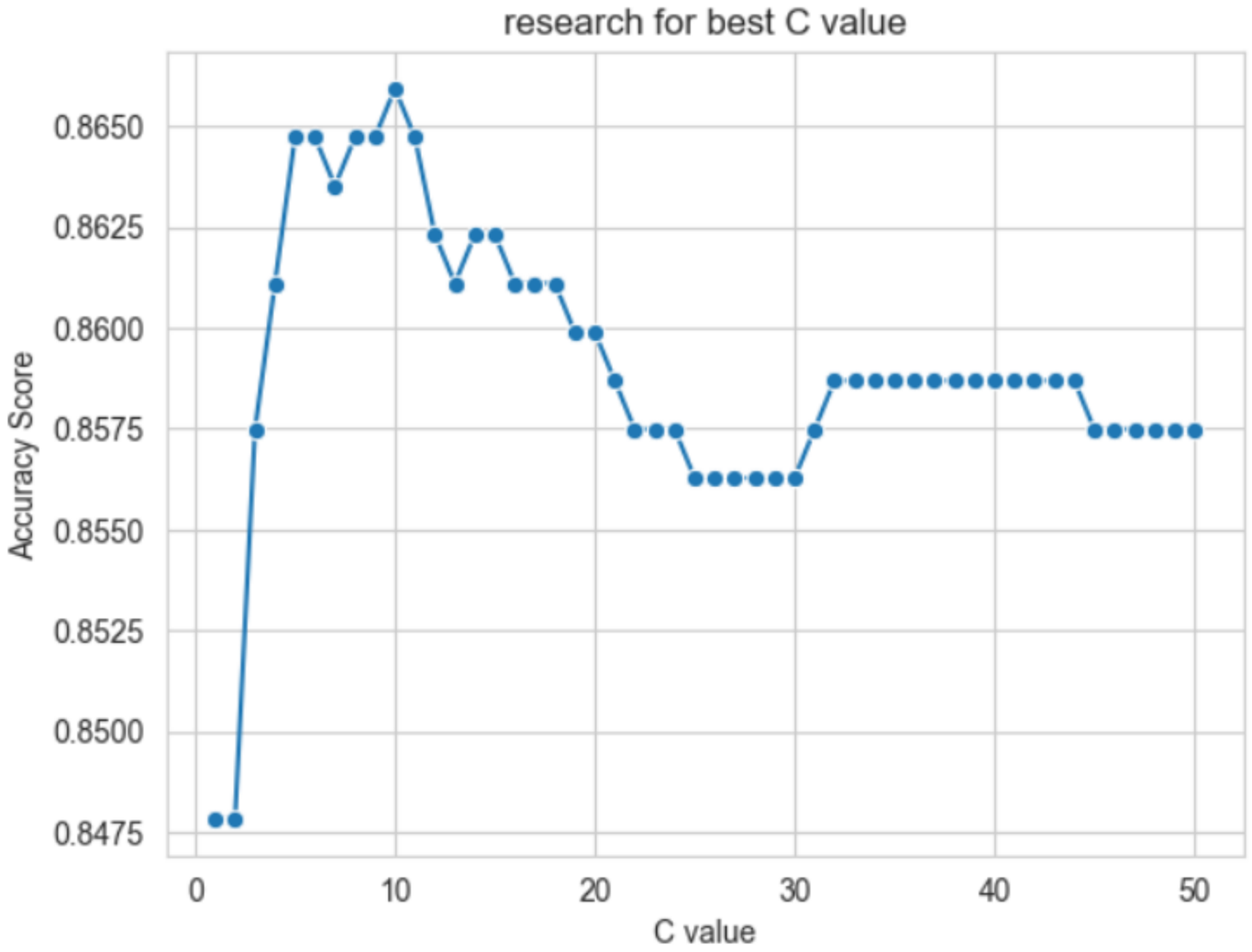
MODELING: RANDOM FOREST

Mean squared error: 0.333
F1 score: 0.811
Accuracy: 0.869
Recall: 0.839
Precision: 0.888

(Run time : 128.5s)



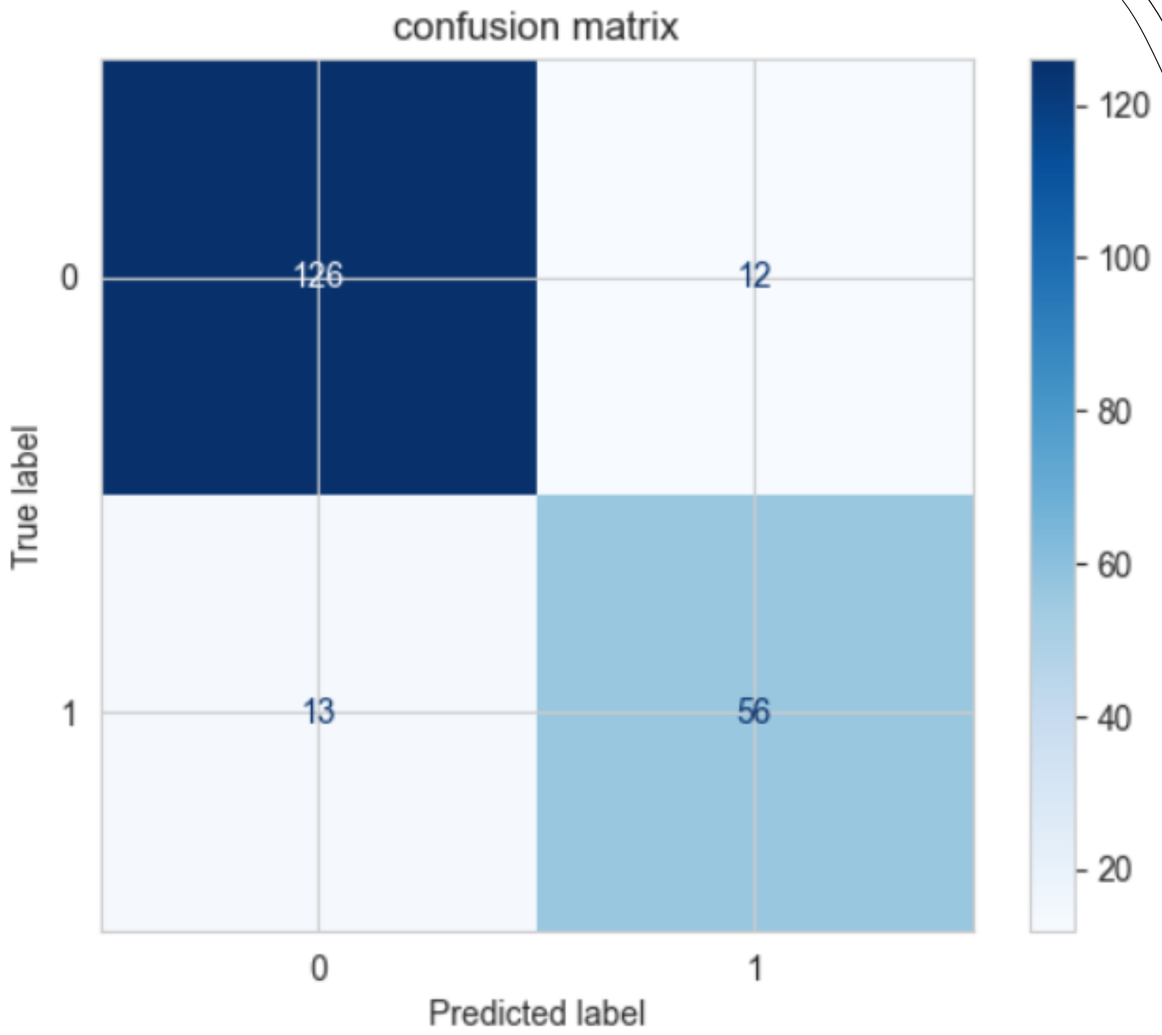
MODELING: SVC



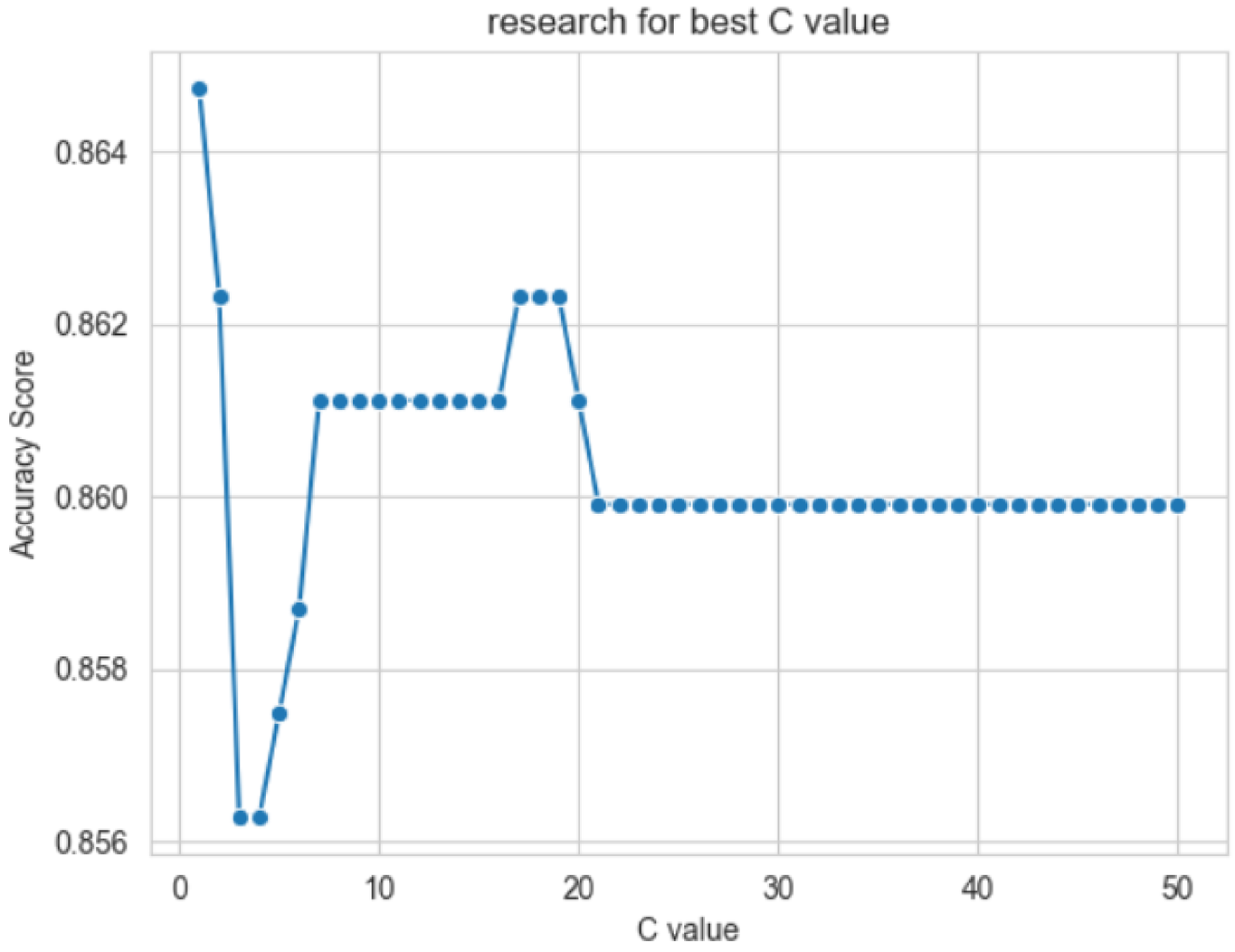
MODELING: SVC

Mean squared error: 0.348
F1 score: 0.823
Accuracy: 0.812
Recall: 0.818
Precision: 0.879

(Run time : 218.4s)



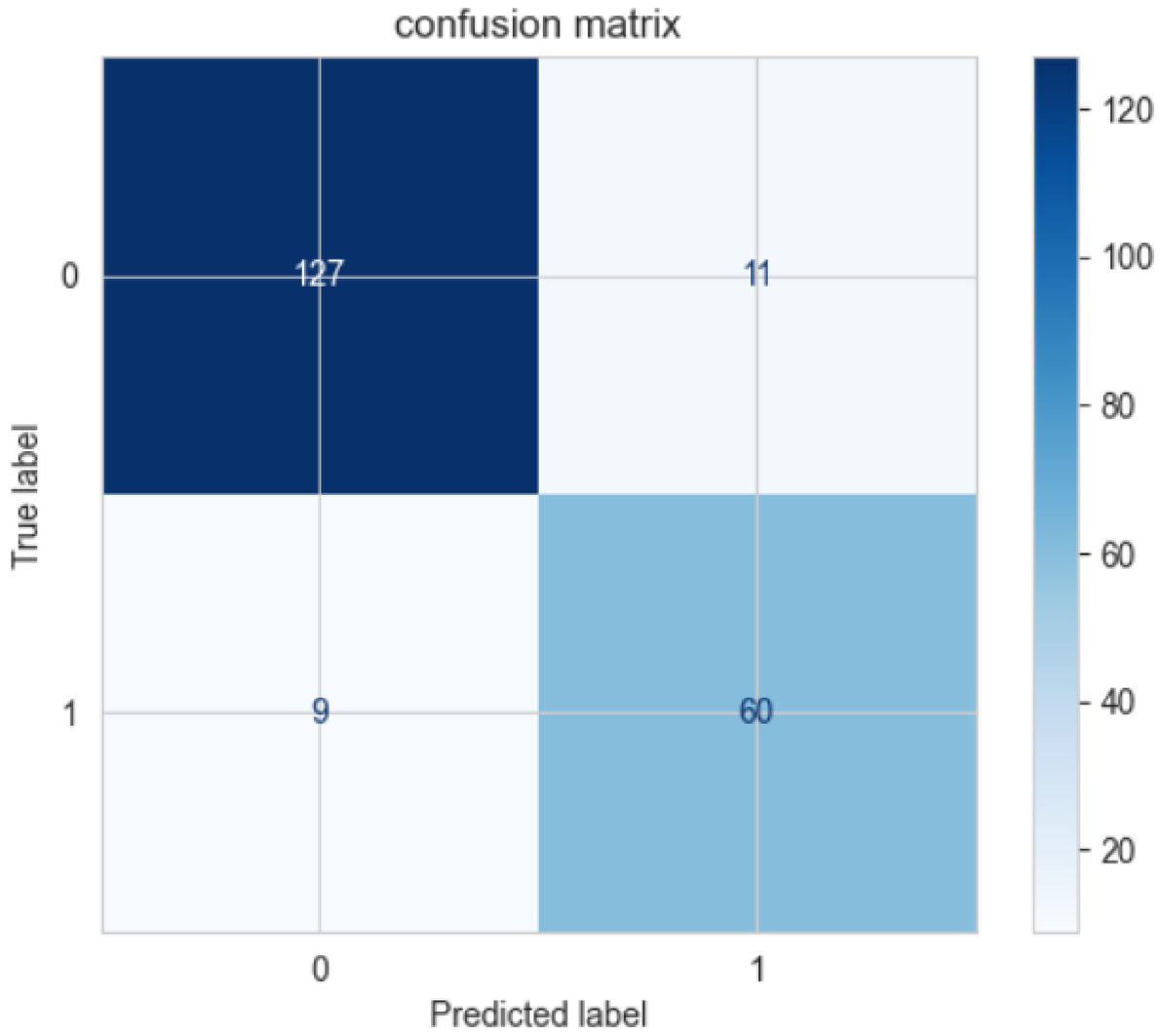
MODELING: LOGISTIC REGRESSION



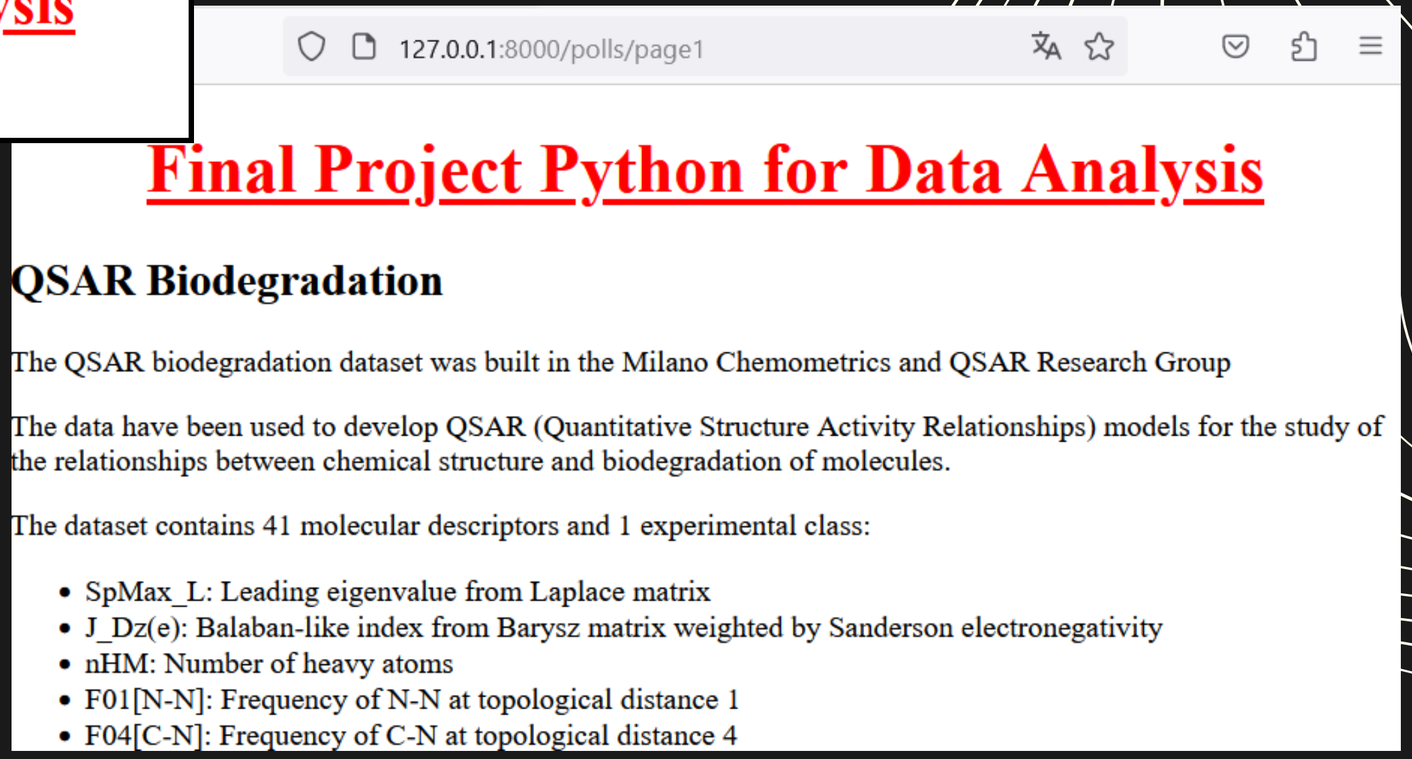
MODELING: LOGISTIC REGRESSION

Mean squarred error: 0.311
F1 score: 0.845
Accuracy: 0.869
Recall: 0.857
Precision: 0.903

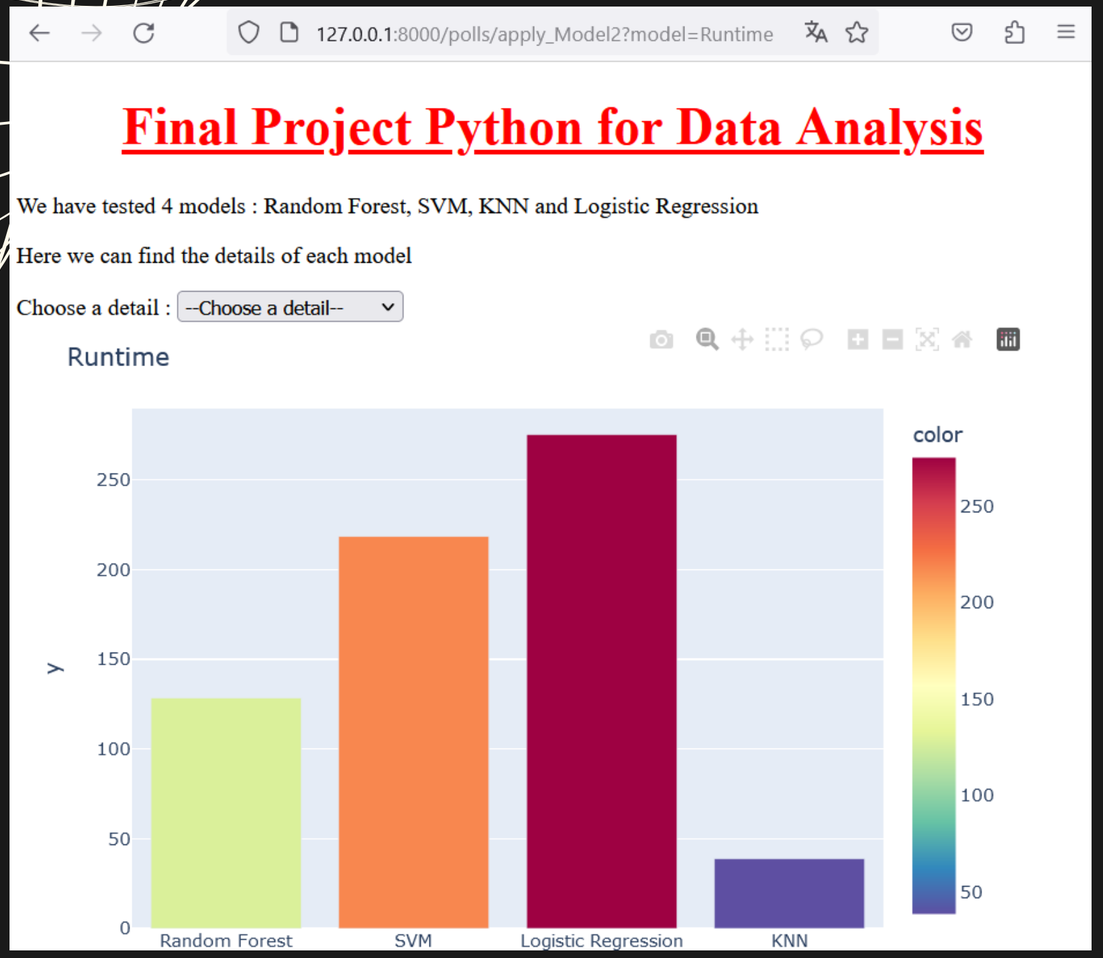
(Run time : 275.2s)

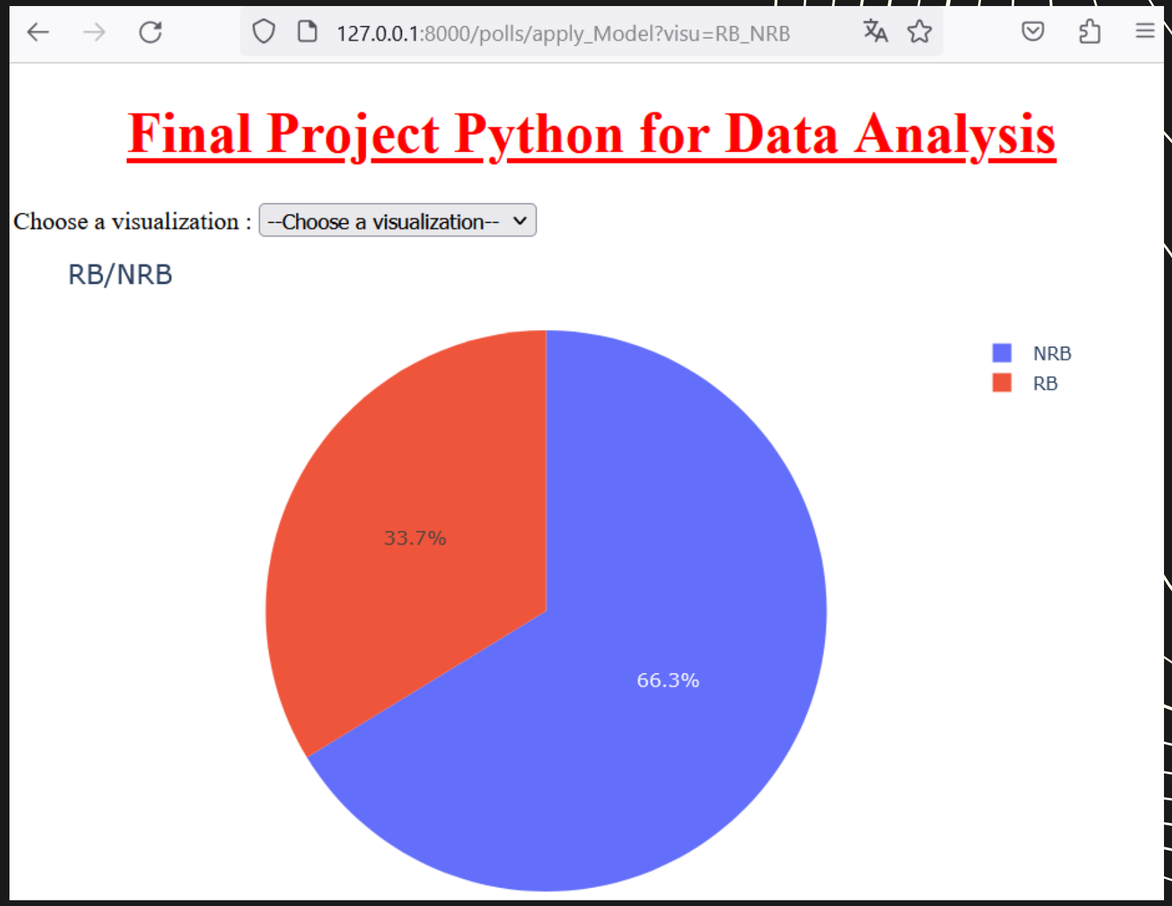


TRANSFORMATION INTO AN API



TRANSFORMATION INTO AN API





THANK'S FOR LISTENING