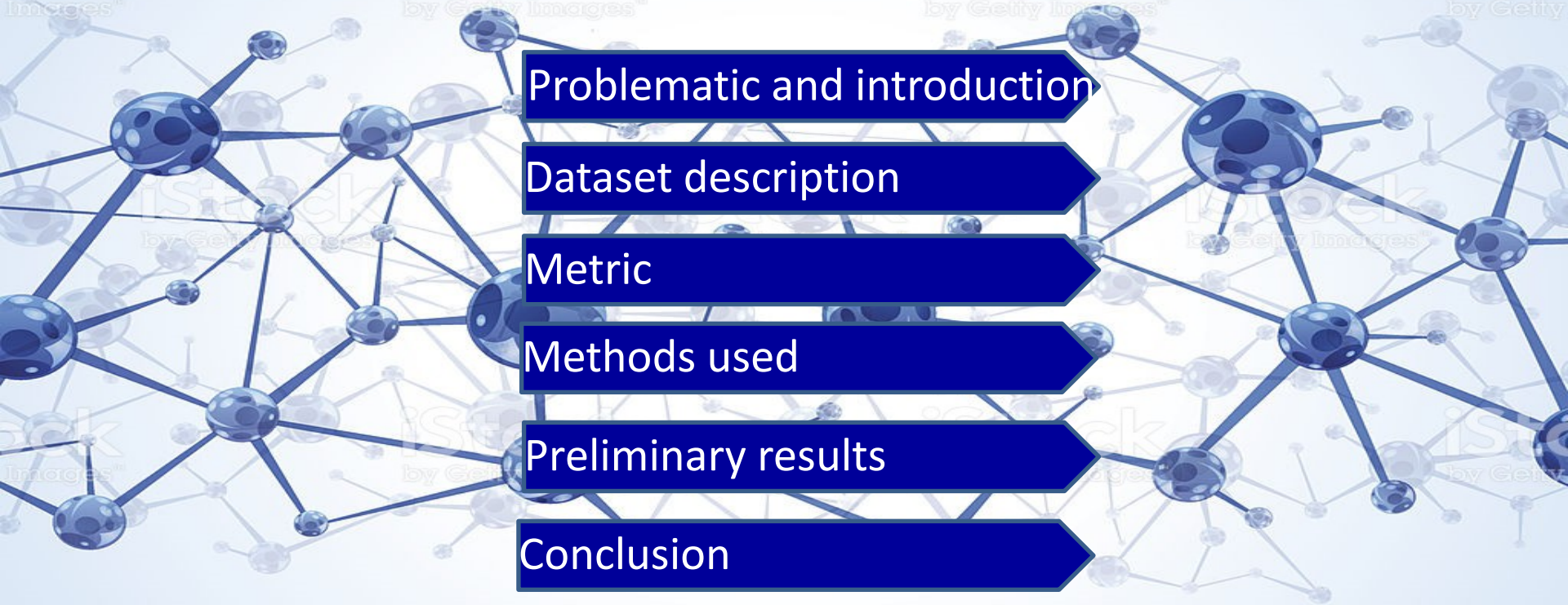
The Marine Team

Predict molecule effect on HIV infection

Summary

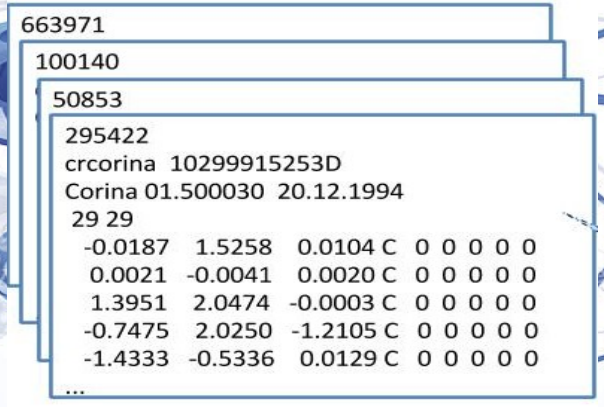


- Problematic and introduction
- Dataset description
- Metric
- Methods used
- Preliminary results
- Conclusion

Introduction

Problematic and introduction

- HIV is one of the a major global health issues. more than 34 million people are living with a HIV including 1.8 million children and 1,1 million people died of HIV-related illnesses worldwide
- High-throughput screening (HTS) is a method for scientific experimentation used in drug discovery but very expensive Machine Learning might by the solution



Dataset

Data Description

Raw DATA

```
663971
crcorina 10299918033D
Corina 01.500030  20.12.1994
 42 44
   -0.0171    1.4123    0.0098 C   0 0 0 0 0
    0.0021   -0.0041    0.0020 C   0 0 0 0 0
    1.2394   -0.6807   -0.0129 C   0 0 0 0 0
    2.4119    0.0423   -0.0197 C   0 0 0 0 0
    2.3919    1.4335   -0.0121 C   0 0 0 0 0
...
```

A block of the raw data

Preprocessed DATA

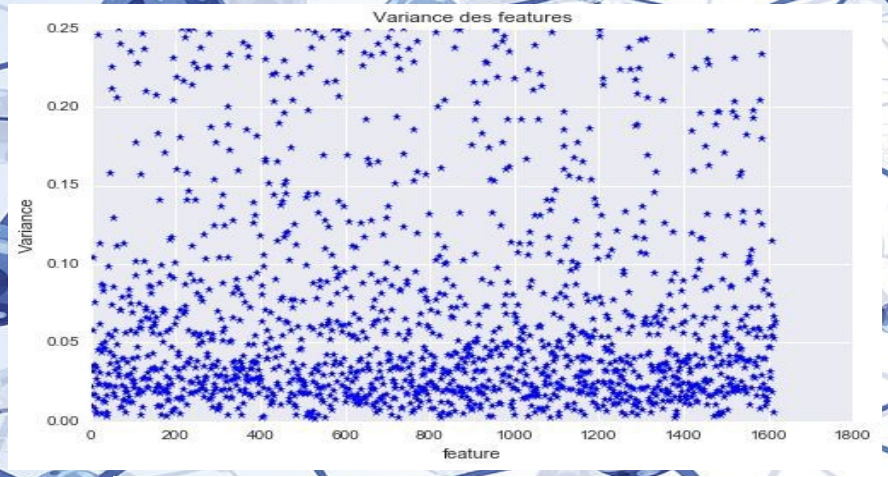
	Positive ex.	Negative ex.	Total
Training set	135	3710	3845
Validation set	14	370	384
Test set	1354	37095	38449
All	**1503**	**41175**	**42678**

Number of examples

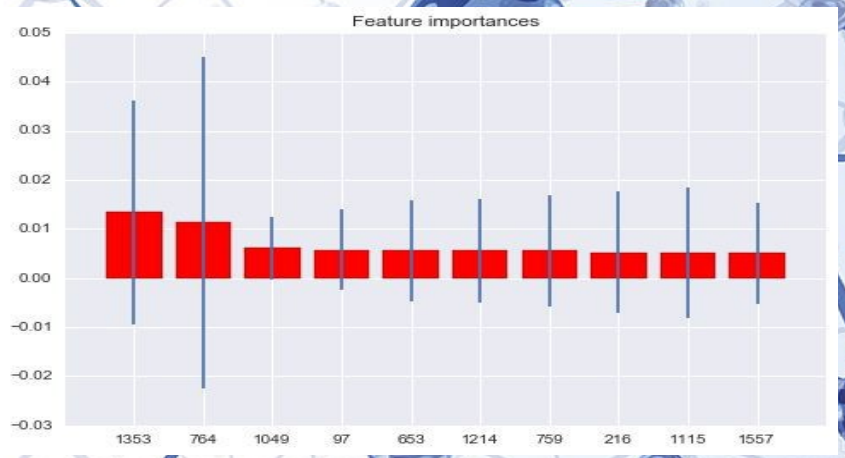
Dataset

Data description

Variance of the features



Importance of the features

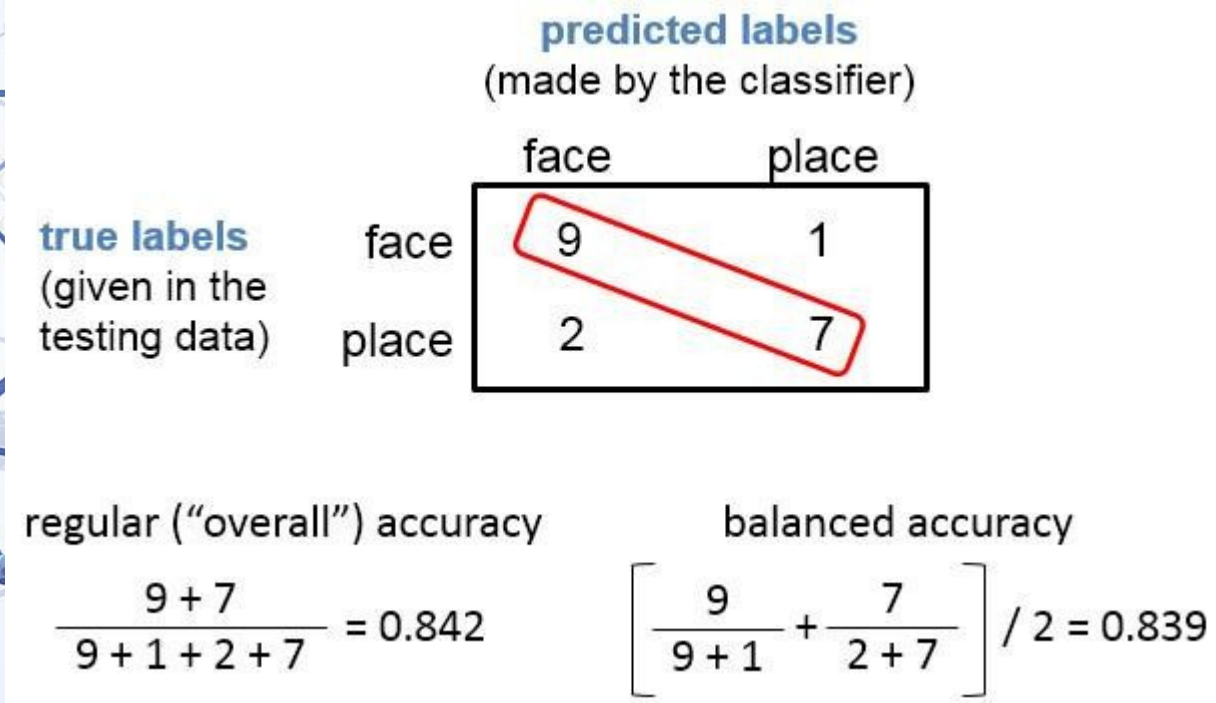


	1	2	3	4	5	6	7	8	9
Feature	1353	764	1049	97	653	1214	759	216	1115
Score	0.013403	0.011273	0.006161	0.005701	0.005559	0.005557	0.005552	0.005233	0.005189

Random Forest → Features Ranking

Metric

Balanced accuracy



regular ("overall") accuracy

$$\frac{9+7}{9+1+2+7} = 0.842$$

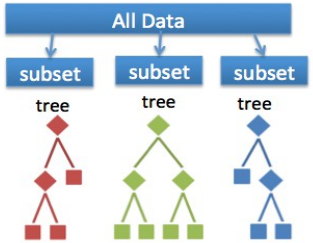
balanced accuracy

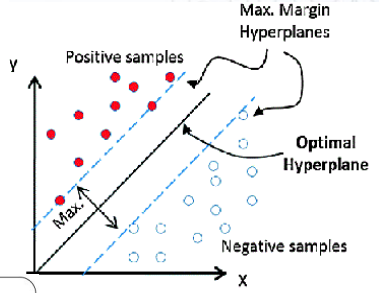
$$\left[\frac{9}{9+1} + \frac{7}{2+7}\right] / 2 = 0.839$$

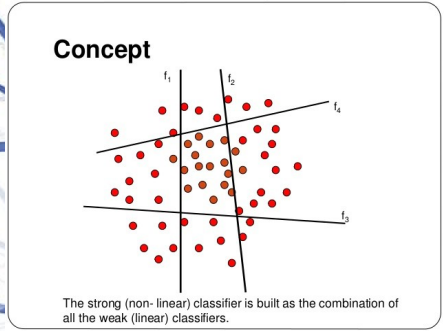
Methods used

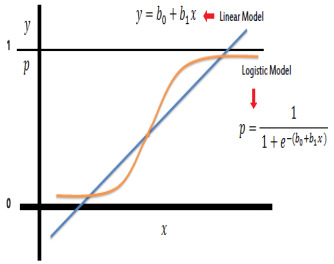
Baseline - sklearn

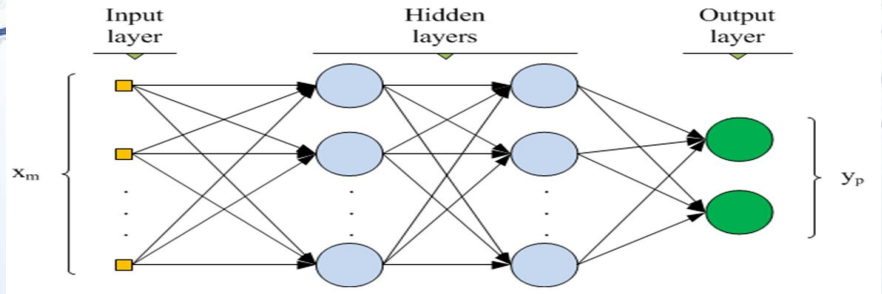
- Random forest
- SVM
- Ada boost
- Logistic regression
- Mutlilayer perceptron











Z Normalization

Znormalization :
Mean=0
Variaence=1

Results without normalization

Logistic regression

BAC : 0.5

		Predicted label		
		-1	1	Total
True label	-1	37094	0	37094
	1	1354	0	1354
	Total	38448	0	

Adaboost

BAC : 0.59

		Predicted label		
		-1	1	Total
True label	-1	36867	227	37094
	1	1113	241	1354
	Total	37980	468	

Random forest

BAC : 0.61

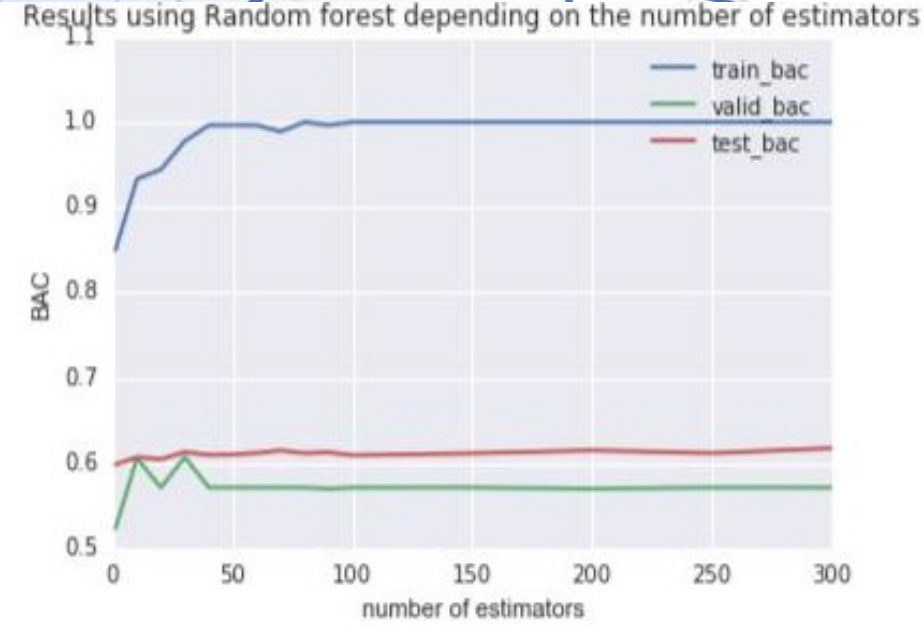
		Predicted label		
		-1	1	Total
True label	-1	36913	181	37094
	1	1042	312	1354
	Total	37955	493	

SVM

BAC : 0.51

		Predicted label		
		-1	1	Total
True label	-1	37087	7	37094
	1	1130	24	1154
	Total	38217	31	

Results without normalization



		Predicted label	
Bac= 0,4410		-1	1
True label	-1	34964	2130
	1	913	441
	Total	35877	2571

		Predicted label	
Bac=0,4492		-1	1
True label	-1	36628	466
	1	924	430
	Total	37552	896

Conclusion

- The purpose is to predict if a molecule can or not, be active against the HIV.
- A preprocessing of the dataset combined to an adequate binary classification method gave satisfying result
- Another configuration of neural network could improve the results.