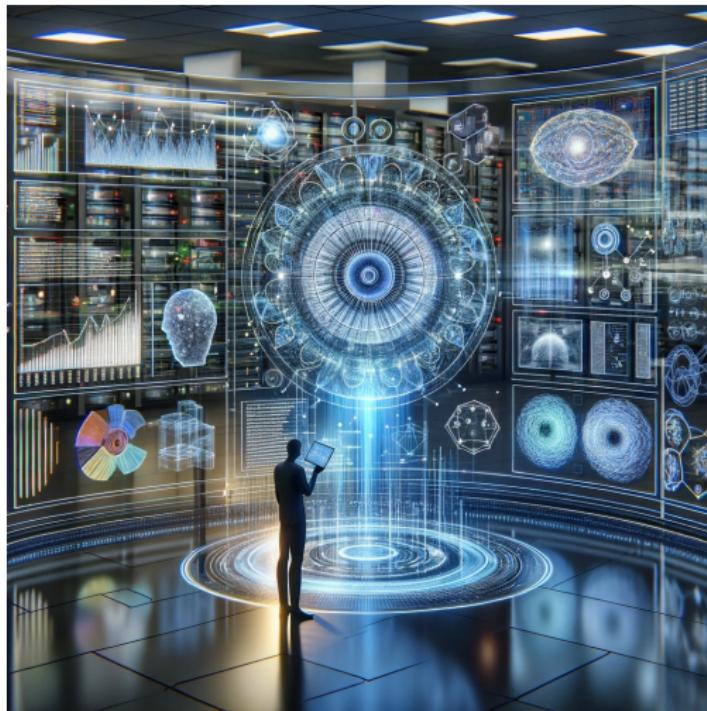


# Soutenance projets statistique en grande dimension

Ivanhoé Botcazou

22 janvier 2024

# Introduction



**Problématique :** Comment utiliser l'apprentissage automatique pour résoudre des tâches complexes sur des grands jeux de données ?

1 La base de données MNIST

2 Classification de cancers

3 Algorithmes de Boosting avec des arbres

4 Data challenge

# KNN avec la base de données MNIST

Ceci est un 5 Ceci est un 0 Ceci est un 4 Ceci est un 1 Ceci est un 9 Ceci est un 2 Ceci est un 1 Ceci est un 3

5 0 4 1 9 2 1 3

Ceci est un 1 Ceci est un 4 Ceci est un 3 Ceci est un 5 Ceci est un 3 Ceci est un 6 Ceci est un 1 Ceci est un 7

1 4 3 5 3 6 1 7

Ceci est un 2 Ceci est un 8 Ceci est un 6 Ceci est un 9 Ceci est un 4 Ceci est un 0 Ceci est un 9 Ceci est un 1

2 8 6 9 4 0 9 1

Ceci est un 1 Ceci est un 2 Ceci est un 4 Ceci est un 3 Ceci est un 2 Ceci est un 7 Ceci est un 3 Ceci est un 8

1 2 4 3 2 7 3 8

Ceci est un 6 Ceci est un 9 Ceci est un 0 Ceci est un 5 Ceci est un 6 Ceci est un 0 Ceci est un 7 Ceci est un 6

6 9 0 5 6 0 7 6

Ceci est un 1 Ceci est un 8 Ceci est un 7 Ceci est un 9 Ceci est un 3 Ceci est un 9 Ceci est un 8 Ceci est un 5

1 8 7 9 3 9 8 5

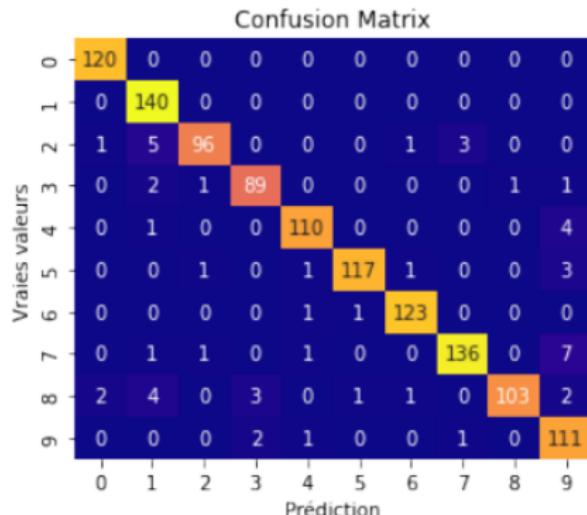
Ceci est un 9 Ceci est un 3 Ceci est un 3 Ceci est un 0 Ceci est un 7 Ceci est un 4 Ceci est un 9 Ceci est un 8

9 3 3 0 7 4 9 8

Ceci est un 0 Ceci est un 9 Ceci est un 4 Ceci est un 1 Ceci est un 4 Ceci est un 4 Ceci est un 6 Ceci est un 0

0 9 4 1 4 4 6 0

## Visualisation des données



## Matrice de confusion KNN\_10

# Mauvaises prédictions du modèle

Un 2 prédit pour un 7



Un 9 prédit pour un 7



Un 7 prédit pour un 4



Un 8 prédit pour un 3



Un 8 prédit pour un 3



Un 8 prédit pour un 1



Un 5 prédit pour un 4



Un 2 prédit pour un 1



Un 5 prédit pour un 2



Un 3 prédit pour un 2



Un 2 prédit pour un 7



Un 3 prédit pour un 1



Un 2 prédit pour un 1



Un 8 prédit pour un 5



Un 9 prédit pour un 4



Un 7 prédit pour un 9



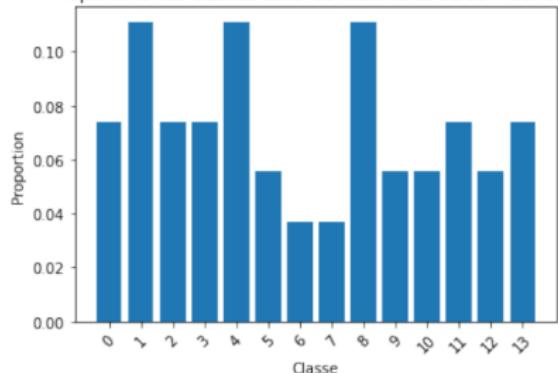
# Classification des cancers avec des arbres de décisions



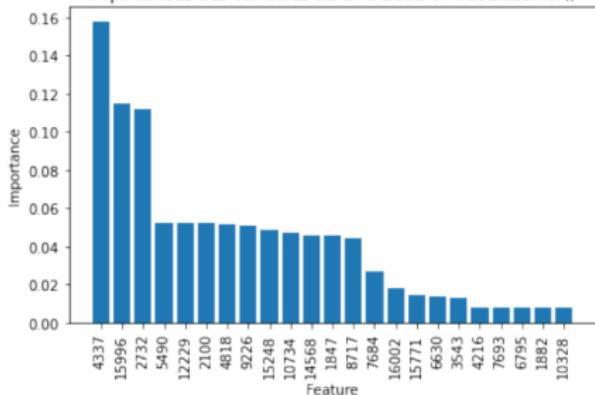
## Remarque

*Classification de cancer en fonction de ses caractéristiques génétiques, couvrant un grand nombre de gènes. La difficulté réside dans le fait qu'il y a peu d'individus (143) pour un nombre très élevé de colonnes (16 063).*

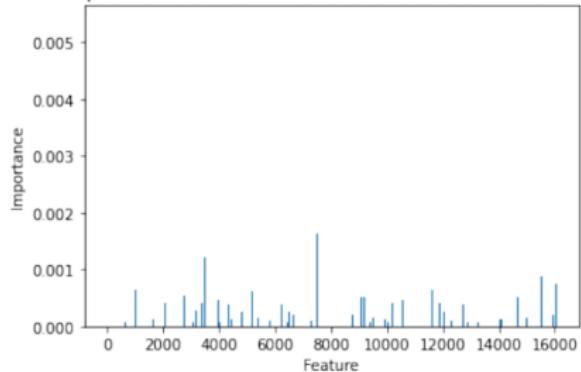
Proportions des classes dans les ensembles de données test



Importances des variables dans le DecisionTreeClassifier()



Importances des variables dans le RandomForestClassifier()



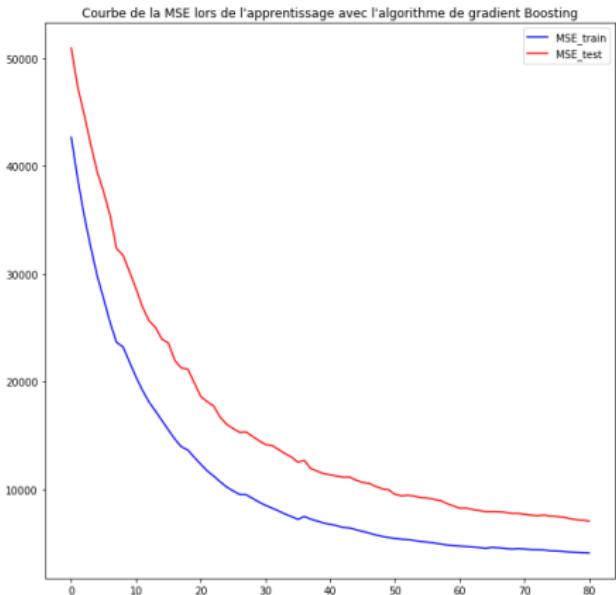
		Real Label		
		Positive	Negative	
Predicted Label	Positive	True Positive (TP)	False Positive (FP)	Precision = $\frac{\sum TP}{\sum TP + FP}$
	Negative	False Negative (FN)	True Negative (TN)	
				Recall = $\frac{\sum TP}{\sum TP + FN}$
				Accuracy = $\frac{\sum TP + TN}{\sum TP + FP + FN + TN}$

# Gradient Boosting pour un problème de régression

Le but de cet algorithme est de construire un estimateur  $F \in \text{Lin}(\mathcal{F})$ , où  $\mathcal{F}$  est l'ensemble des arbres de décision de profondeur  $p$  et  $\text{Lin}(\mathcal{F})$  est l'ensemble des combinaisons linéaires d'éléments de  $\mathcal{F}$ . Considérons la fonction de coût

$$\Psi(z, y) = \frac{(z - y)^2}{2}$$

L'objectif est de trouver  $F^*$  qui minimise  $C(F) = E[\Psi(F(X), Y)]$

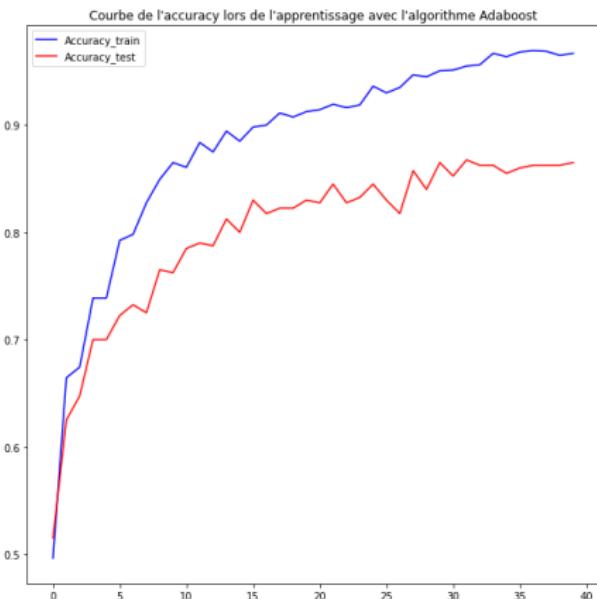


# Adaboost pour un problème de classification binaire

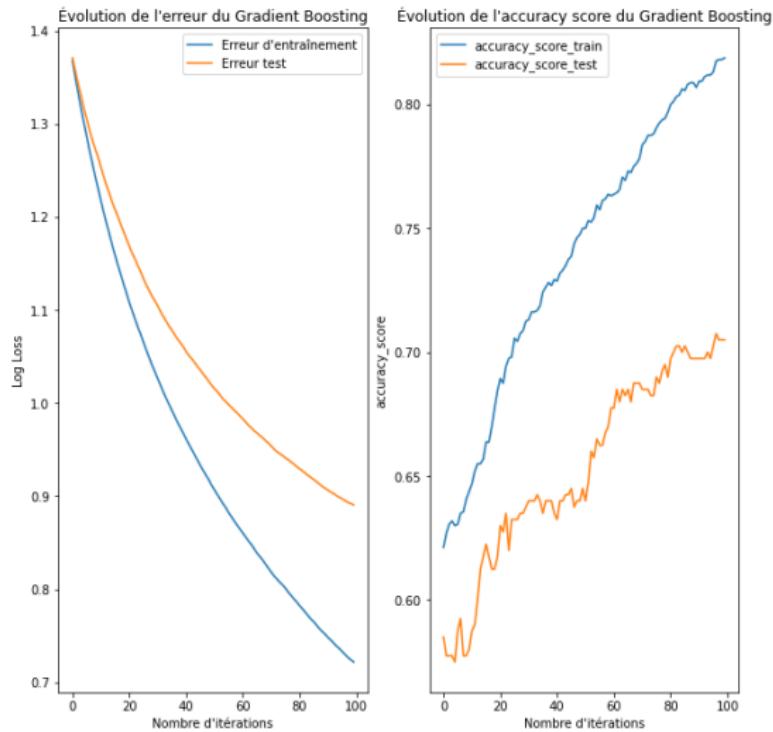
## Algorithm 10.1 AdaBoost.M1.

1. Initialize the observation weights  $w_i = 1/N$ ,  $i = 1, 2, \dots, N$ .
2. For  $m = 1$  to  $M$ :
  - (a) Fit a classifier  $G_m(x)$  to the training data using weights  $w_i$ .
  - (b) Compute
$$\text{err}_m = \frac{\sum_{i=1}^N w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i}.$$
  - (c) Compute  $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$ .
  - (d) Set  $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))]$ ,  $i = 1, 2, \dots, N$ .
3. Output  $G(x) = \text{sign} \left[ \sum_{m=1}^M \alpha_m G_m(x) \right]$ .

Algorithme extrait du livre :  
"The elements of statistical learning"



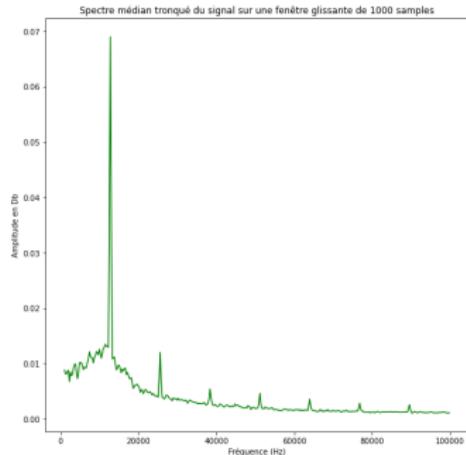
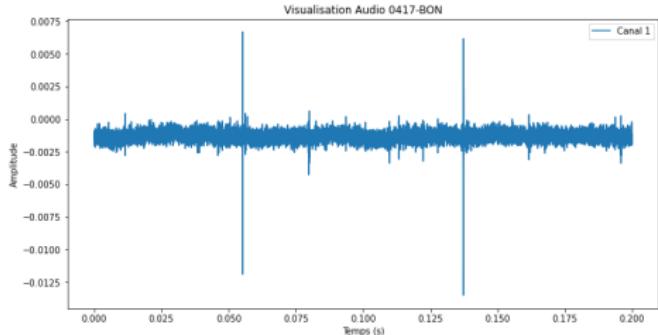
# Gradient Boosting pour un problème de classification multi-classe



# Biosonar - Détection de clics d'Odontocètes

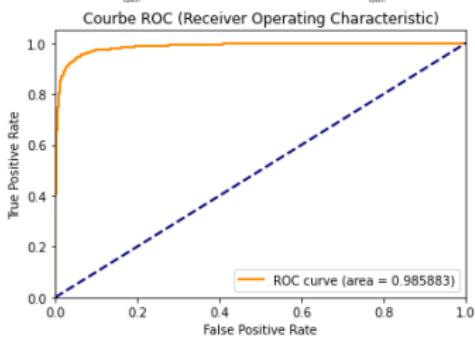
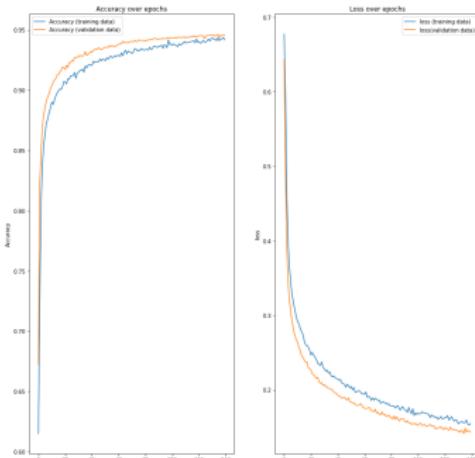
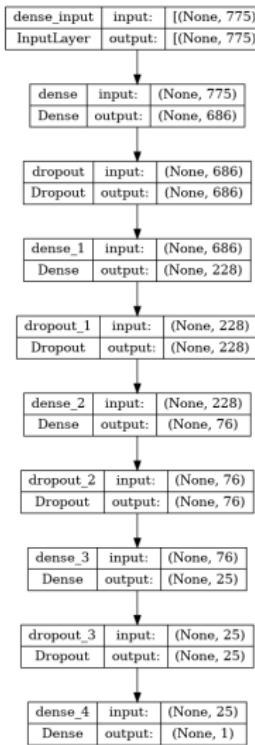


# Visualisation des données et preprocessing

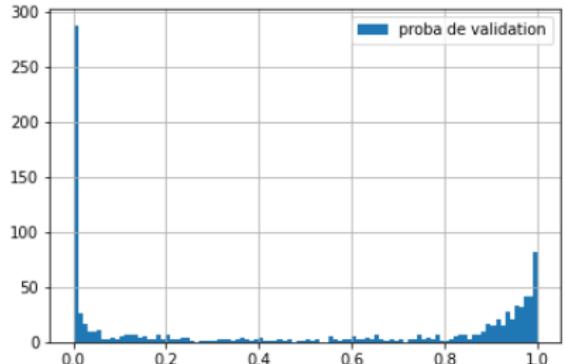
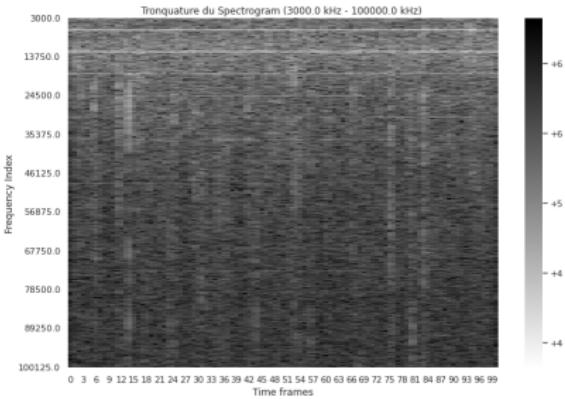
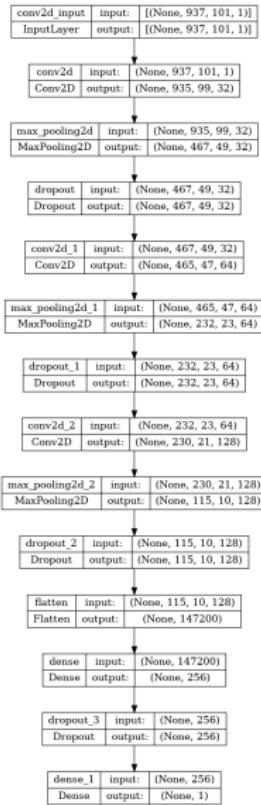


	5120.0	5632.0	6144.0	6656.0	7168.0	7680.0	8192.0	8704.0	9216.0	9728.0	...	sc.min	sc.max	sb.mean	sb.std	sb.min	sb.max	sb.mean	sb.std	sb.min	sb.max
14714	0.314585	0.567127	0.497398	0.632767	0.635438	0.343430	0.206735	0.359481	0.282247	0.064285	...	0.324120	0.843605	0.586475	0.013715	-0.196981	-0.029097	0.236200	-0.007373	0.1136434	-0.052309
22786	-0.279047	-0.150175	-0.141035	-0.103095	-0.158307	0.007323	-0.033879	0.402535	0.387660	0.219651	...	0.942626	-0.466895	0.722880	-0.000505	0.306931	0.393846	0.051222	0.062083	0.790364	0.282435
15055	0.342820	0.606136	0.643781	0.934973	0.784519	0.535906	0.519194	0.742698	0.545135	0.162142	...	-0.068900	1.255371	0.044605	-0.028070	0.096113	0.463660	0.092396	-0.134285	0.449815	-0.321972
12613	-0.6555397	-0.701079	-0.741630	-0.742450	-0.680815	-0.805033	-0.795378	-0.904526	-0.739233	-0.655454	...	0.831983	0.493039	0.440323	-0.235342	-0.258978	-0.593659	0.403702	0.073228	2.196431	0.569836
6286	-0.529578	-0.634100	-0.646681	-0.626867	-0.586255	-0.605662	-0.616391	-0.737111	-0.609664	-0.620415	...	0.642381	-0.468137	0.447916	-0.133331	0.326426	0.382031	-0.010296	-0.051283	0.056569	0.061842

# Modèle de deep learning type MLP



# Modèle de deep learning type CNN



# Merci pour votre écoute

