Dr Aude Sportisse
Chercheuse dans l'équipe Maasai, Inria Sophia Antipolis
Enseignante à EFELIA

aude.sportisse@inria.fr

Introduction à l'Intelligence Artificielle

Appliquée à la biologie

Septembre-Décembre 2023

Programme du jour

Première partie de séance (8h-9h45):

1. Arbre de décision
2. Données manquantes
3. TP: imputation de données manquantes à l'aide de forêts aléatoires

Deuxième partie de séance (10h15-12h):

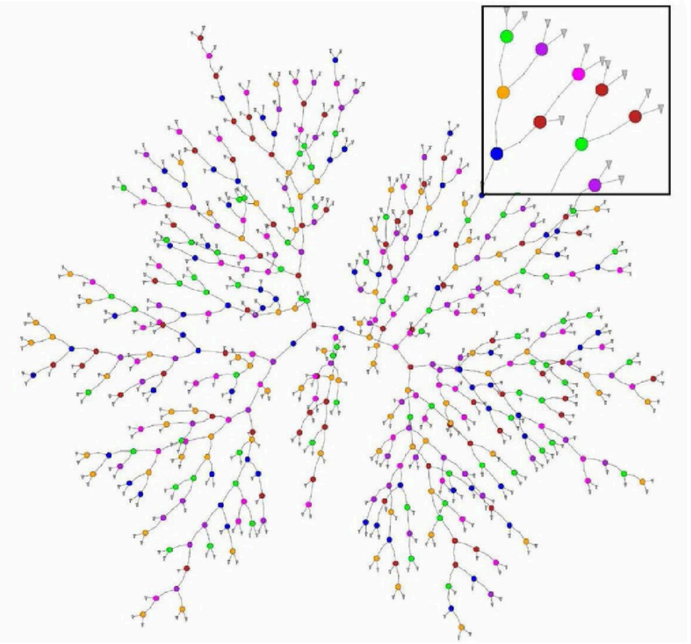
1. Oraux sur les articles
2. Retour sur le cours

1. Arbre de décision

Arbre de décision

- Début dans les années 60-80 (systèmes experts)
- Elément fondamental de deux méthodes, dites *ensemblistes*, très performantes, qui cherchent à apprendre plusieurs arbres de décision pour les combiner:
  - « random forests » (forêts aléatoires) & « gradient boosting trees ».
- Un arbre de décision va apprendre des règles de décision sur les variables (attributs) des données.

["arbre de décision" par un médecin]



Dans un arbre de décision, à chaque noeud (en couleur), une séparation a lieu en fonction de la réponse donnée à une question, jusqu'à atteindre une feuille (un triangle). -

Pour en savoir plus Gérard Biau et Erwan Scornet, « A Random Forest Guided Tour », *TEST,* 25, 197, 2016. Christophe Giraud, *Introduction to High-Dimensional Statistics,* CRC Press/Chapman and Hall, 2014.

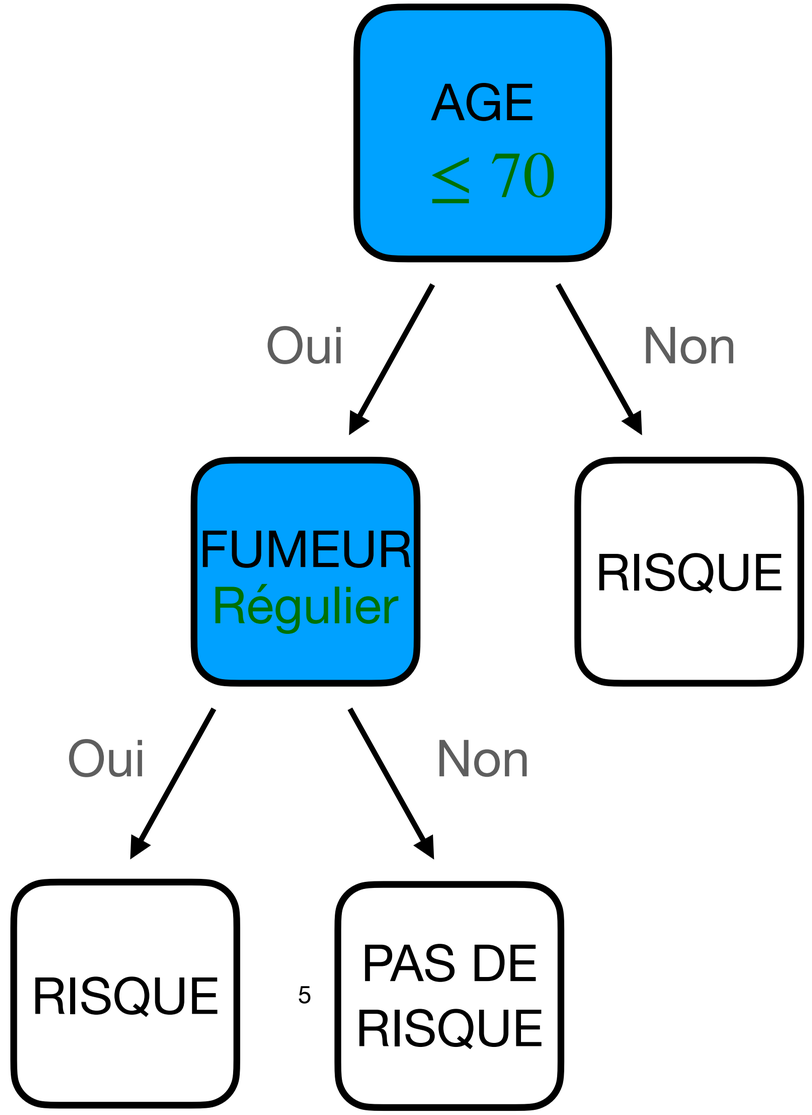
Exemple d'arbre de décision

Variables

Règles de décisions

Pour prédire si un patient est à risque de développer une certaine maladie, un arbre de décision peut être le suivant:

- Si l'âge de la personne est supérieur à 70, il est directement classé à risque.
- Sinon, il n'est classé à risque que s'il est fumeur régulier.



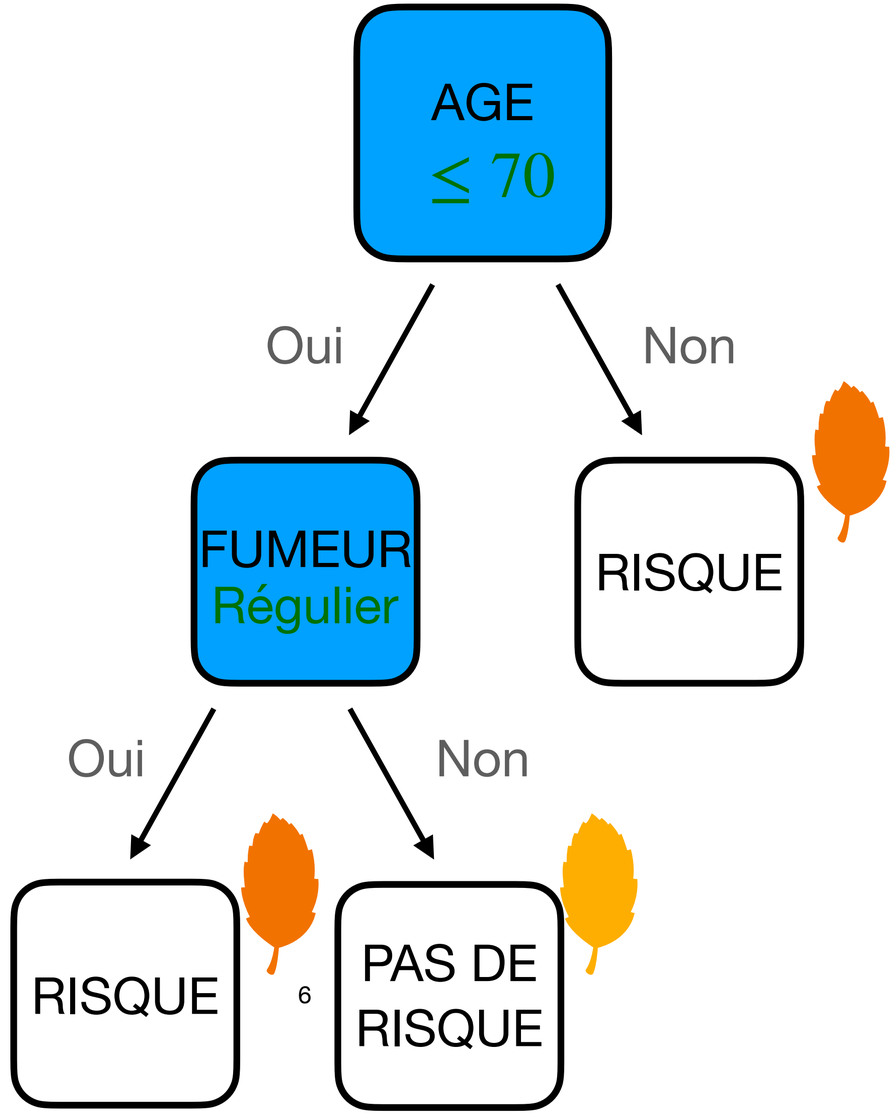
***Définition* [arbre de décision]:**

ensemble de structure hiérarchique, basé sur des règles de décisions sur une variable à la fois.

Un peu de vocabulaire

On peut différentier:

- Les **noeuds de décision** qui coupe en deux parties les données. La **racine** est le premier noeud de décision.
- Les **feuilles de l'arbre** (noeuds finaux): ils contiennent la prédiction (ici: patient à risque / patient non à risque).



Arbre de décision en apprentissage statistique

On peut en fait très bien construire un arbre de décision *à la main*.

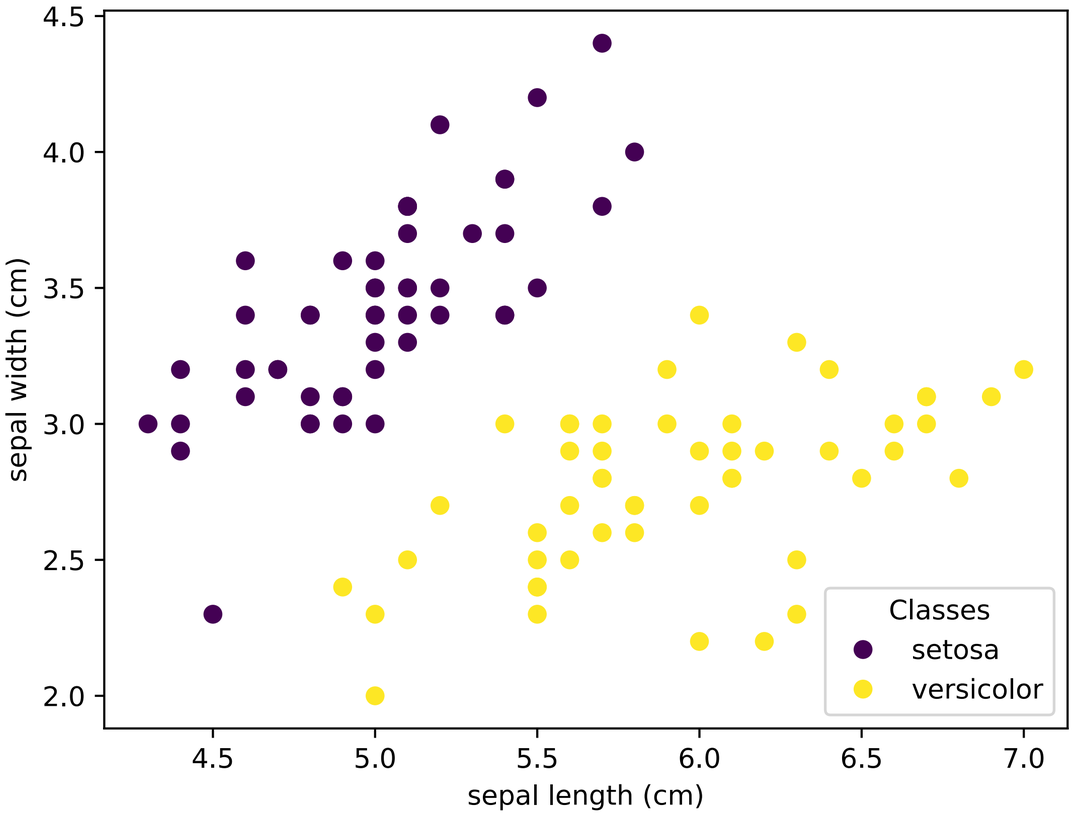
Le but en apprentissage statistique est d'entrainer un modèle capable d'**apprendre automatiquement**:

- Les **variables** qui forment les noeuds décisionnels;
- Les valeurs de seuil (**règles de décision**).
- Les **valeurs de prédictions** dans les feuilles

Arbre de décision pour la classification

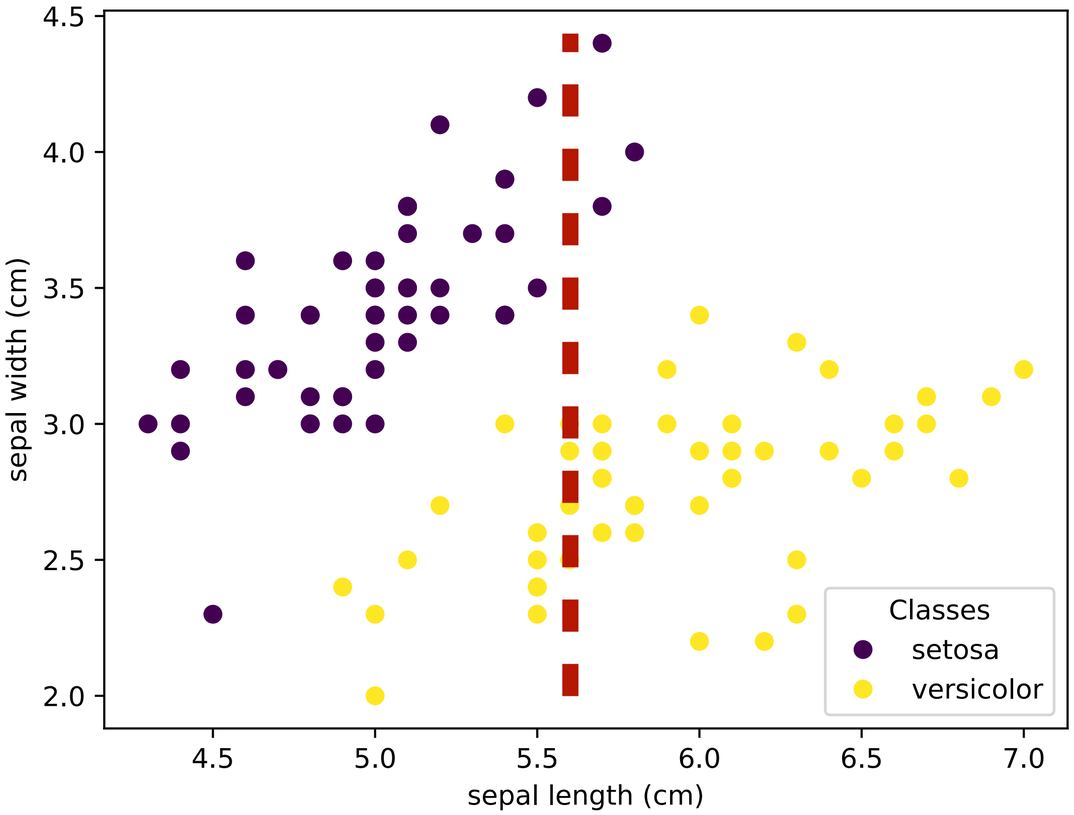
- Rappel en classification: on essaie de prédire une **variable catégorielle**. Exemple: classifier les données en **iris setosa (violet)** ou **iris versicolor (jaune)**.

- **Cadre de l'apprentissage supervisé:** on construit l'arbre sur le jeu d'entrainement où nous avons accès aux étiquettes (on sait quel point est iris setosa / iris versicolor).



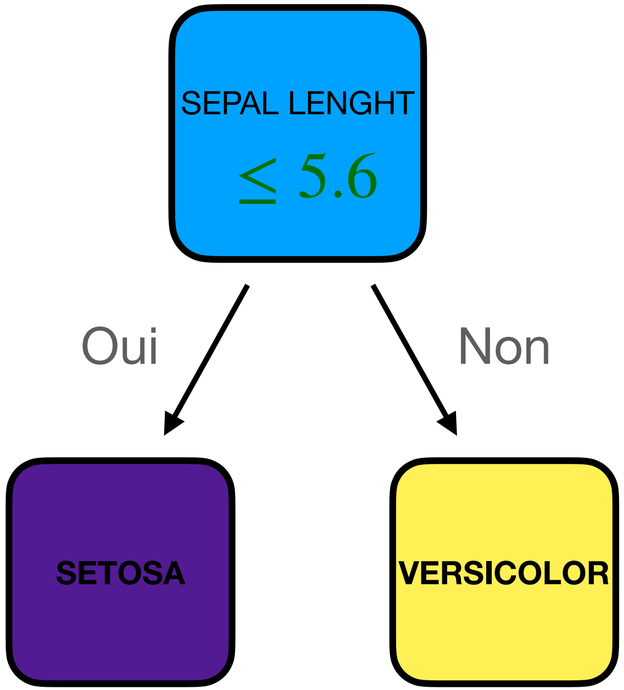
Arbre de décision pour la classification

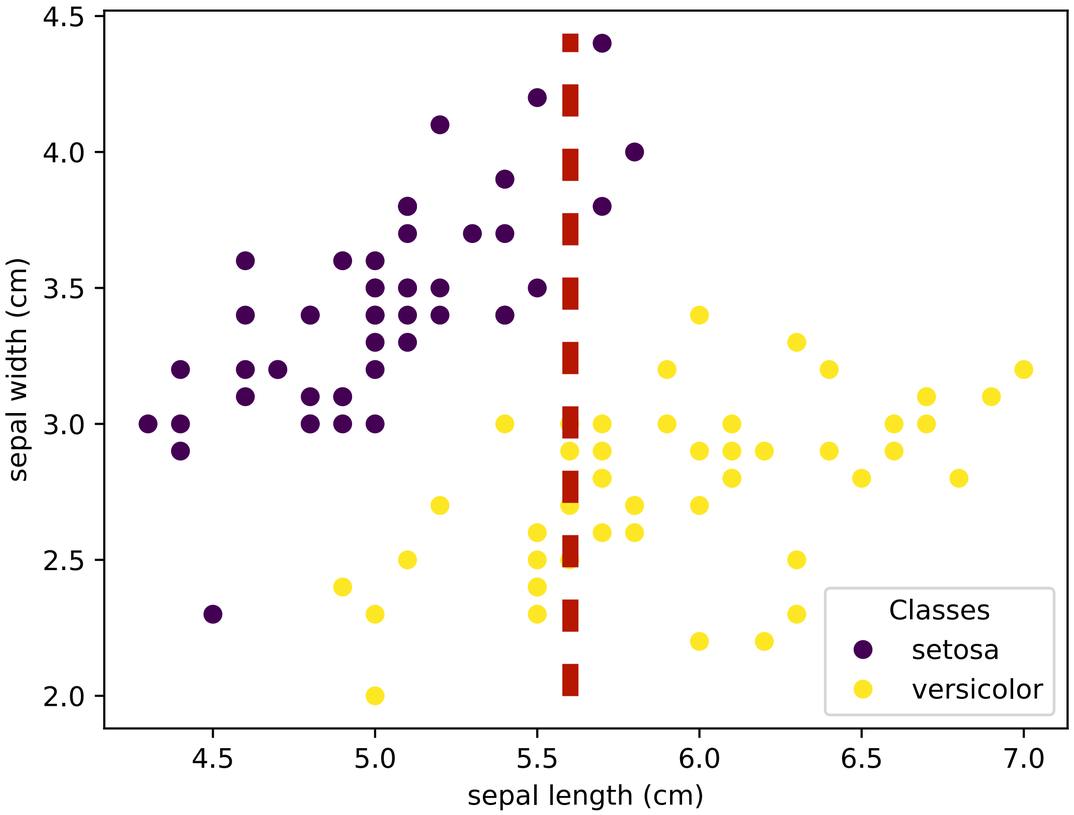
- La racine de l'arbre va effectuer une **division verticale du jeu de données**
- L'algorithme choisit une valeur de la variable **sepal length** qui sépare le jeu de données en deux parties.



Arbre de décision pour la classification

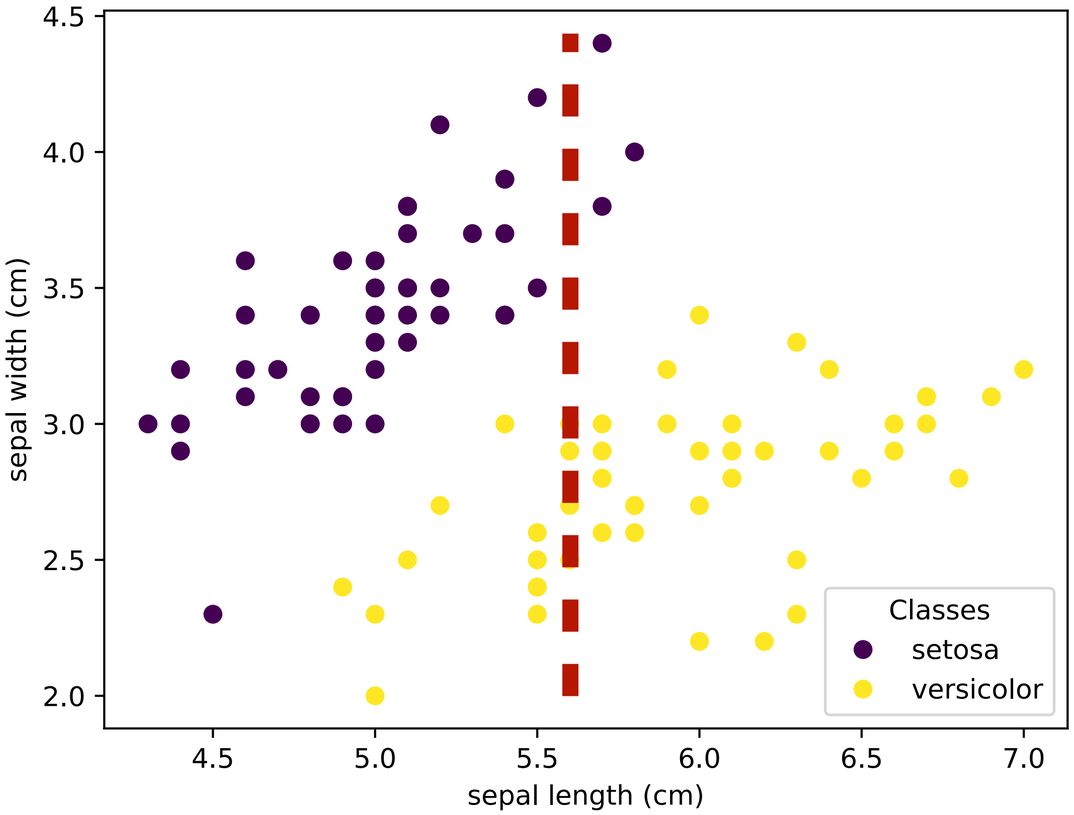
- Il effectue ensuite un « vote majoritaire » pour assigner à gauche la prédiction **iris setosa** (+ de points violets) et à droite la prédiction **iris versicolor** (+ de points jaunes)





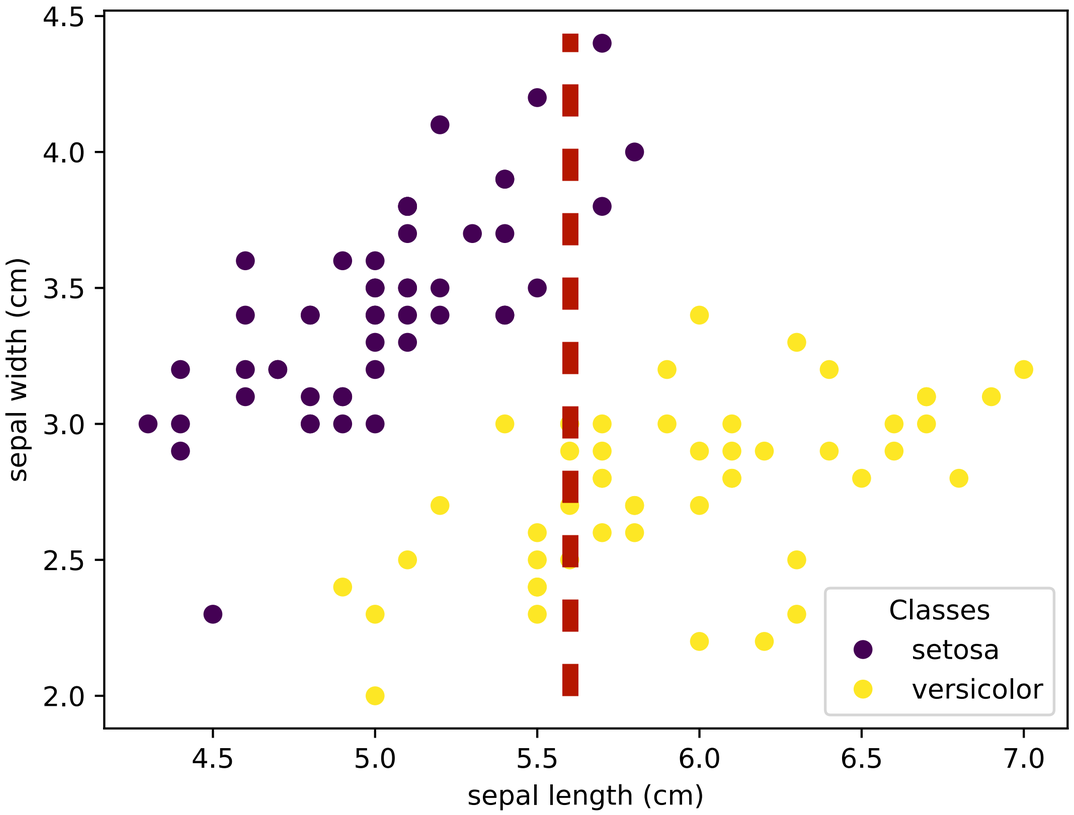
Arbre de décision pour la classification

- On peut regarder les probabilités pour un point d'être vraiment un iris setosa à gauche et d'être vraiment un iris versicolor à droite en calculant:
  - À gauche: $\frac{\sharp \text{ iris setosa}}{\sharp \text{ points total}}$
  - À droite: $\frac{\sharp \text{ iris versicolor}}{\sharp \text{ points total}}$
- Ici, on a pas des probabilités très proches de 1: en faisant cette division verticale, on ne sépare pas bien les setosa des versicolor.

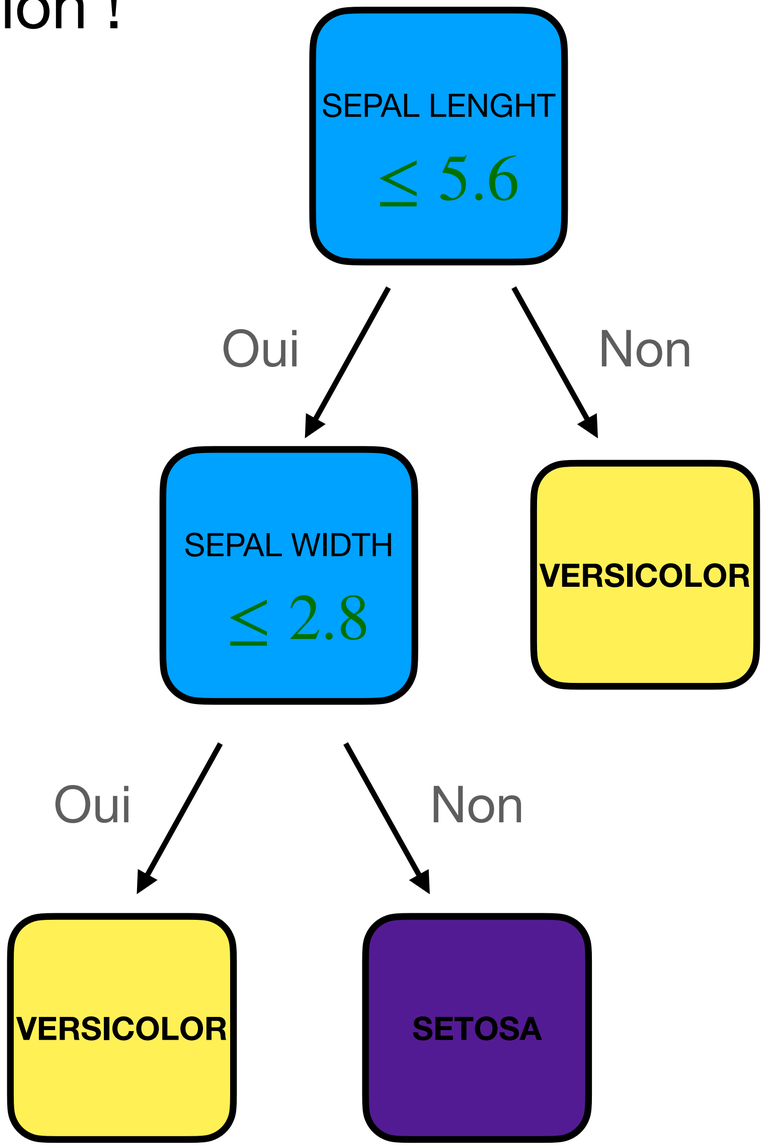


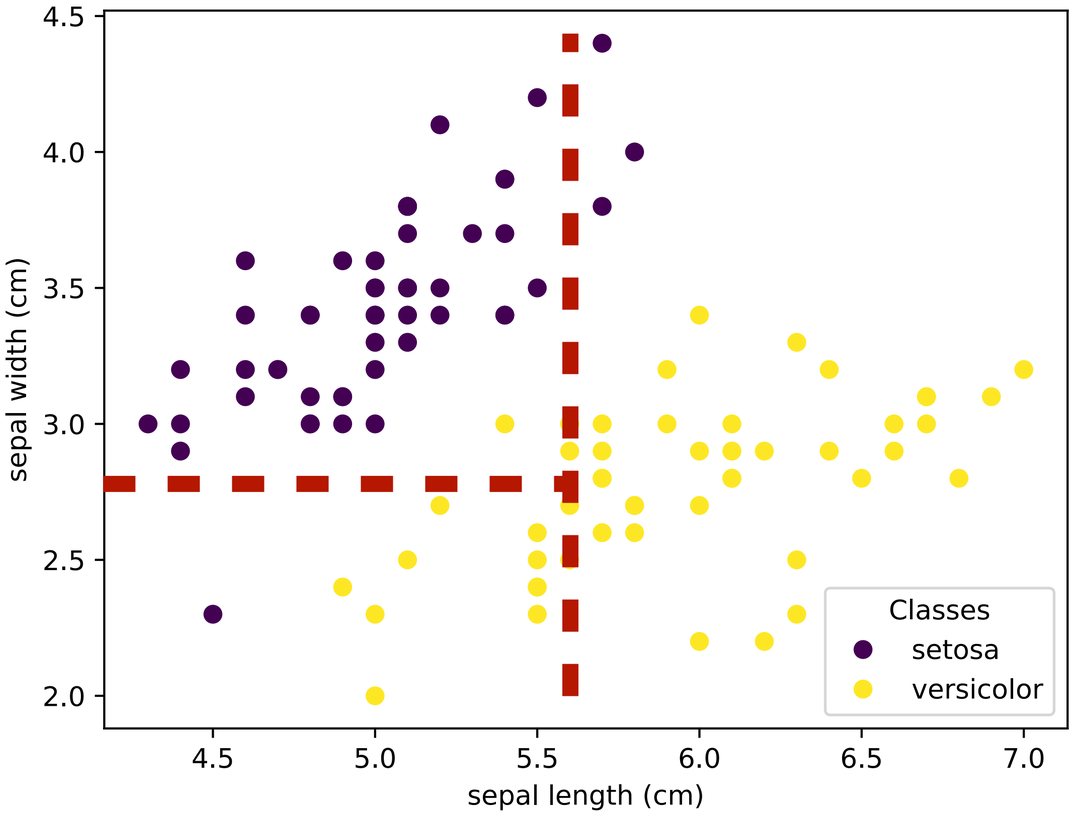
Arbre de décision pour la classification

- On peut regarder les probabilités pour un point d'être vraiment un iris setosa à gauche et d'être vraiment un iris versicolor à droite en calculant:
  - À gauche: $\frac{\sharp \text{ iris setosa}}{\sharp \text{ points total}}$
  - À droite: $\frac{\sharp \text{ iris versicolor}}{\sharp \text{ points total}}$
- Ici, on a pas des probabilités très proches de 1: en faisant cette division verticale, on ne sépare pas bien les setosa des versicolor.



- Solution: rajouter un noeud de décision !

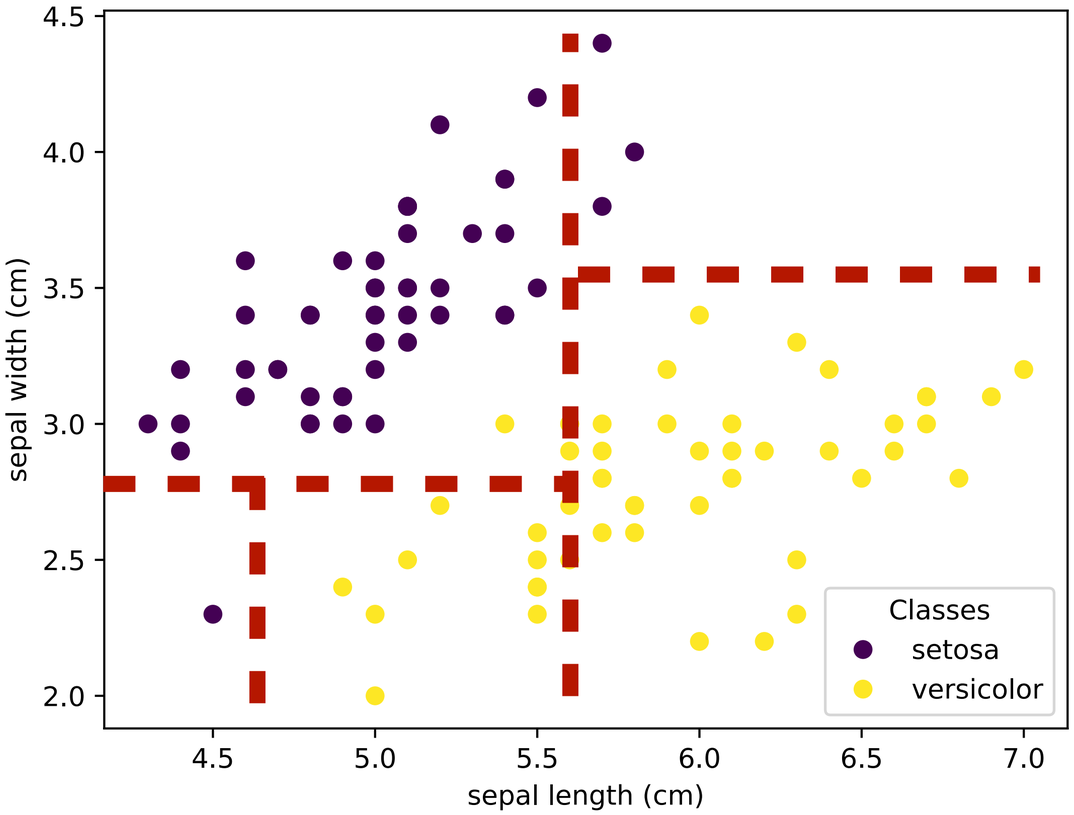




Arbre de décision pour la classification

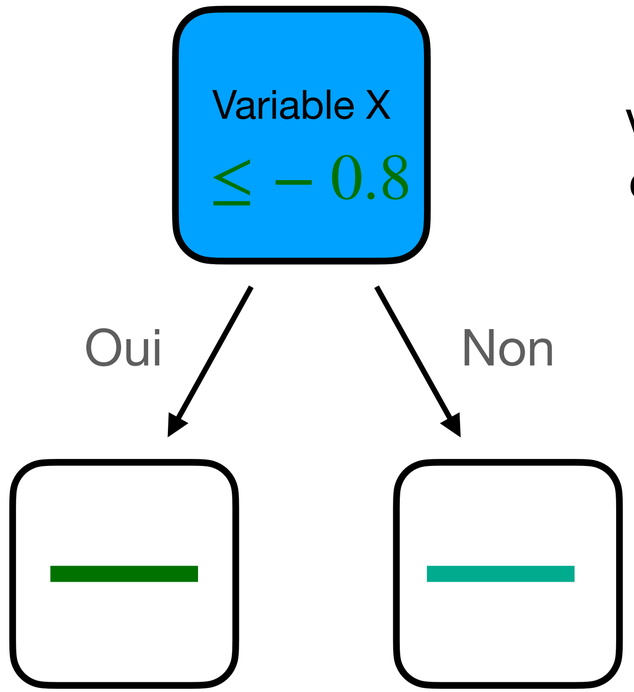
Et ainsi de suite, on peut ajuster les divisions, jusqu'à obtenir 0% d'erreur sur le jeu d'entrainement.

Automatiquement, l'algorithme sélectionne la variable et les valeurs de seuil qui permette le mieux de séparer les données à chaque étape.

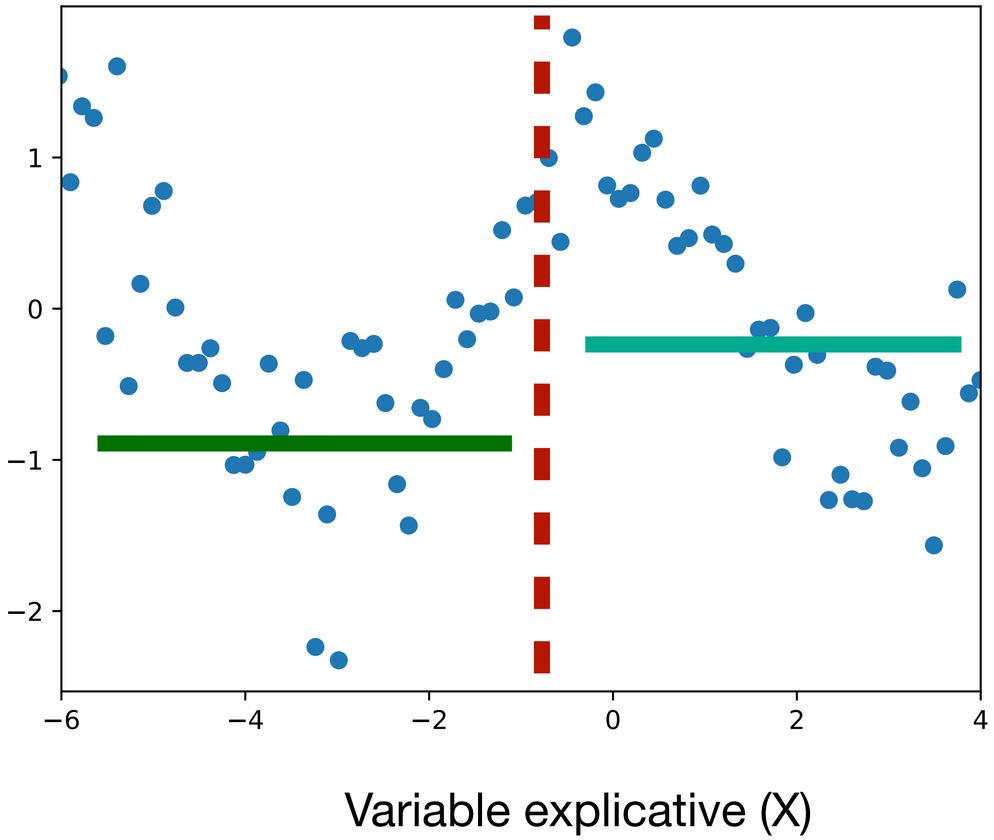


Arbre de décision pour la régression

- Rappel en régression: on essaie de prédire une **variable continue**.
- C'est le même principe !

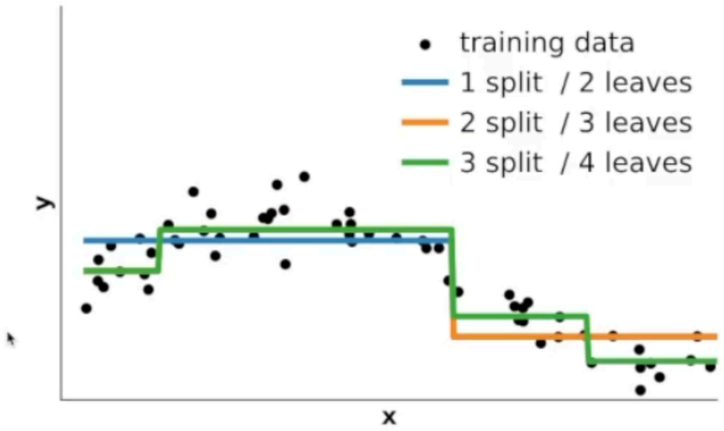


On prédit y avec la moyenne des valeurs à gauche / à droite

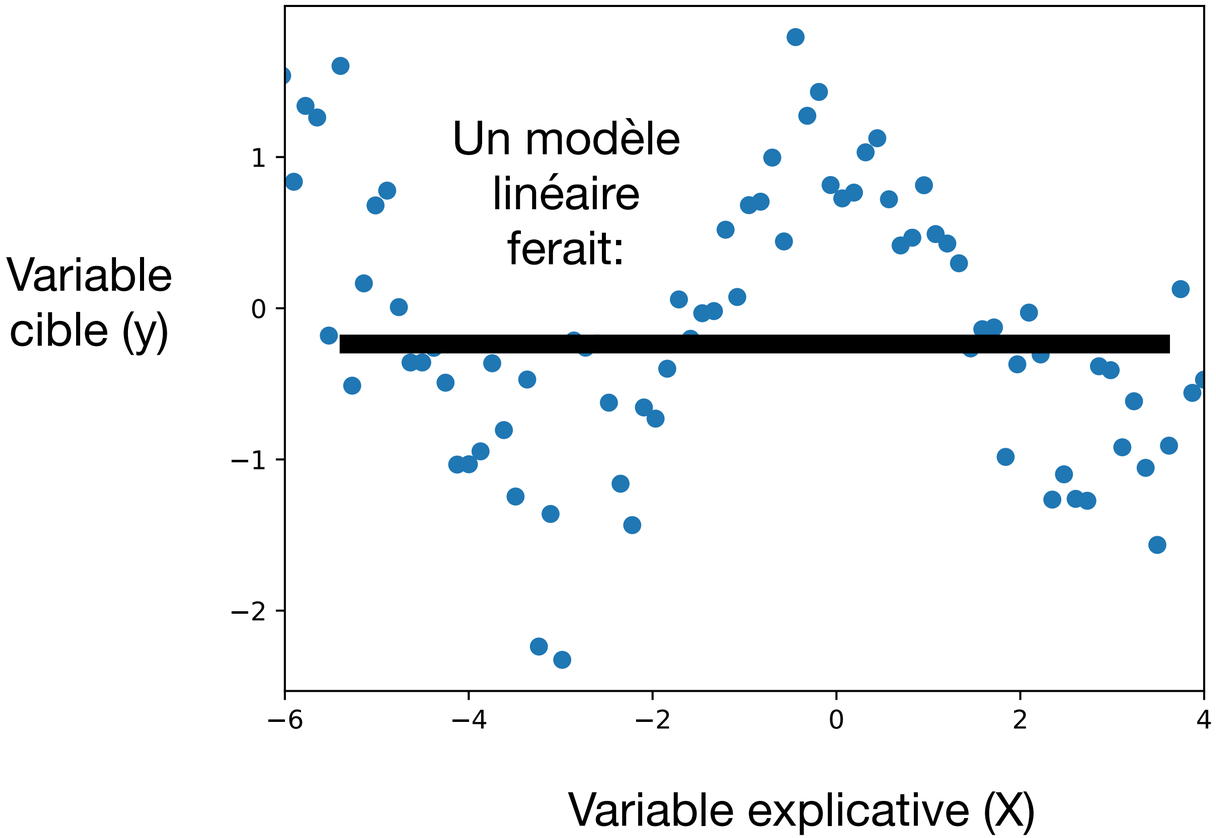


Arbre de décision pour la régression

- Ce n'est **pas un modèle linéaire** (qui tracerait une droite)
- Ici, il faut faire plus de divisions pour affiner le modèle et mieux prendre en compte la structure des données

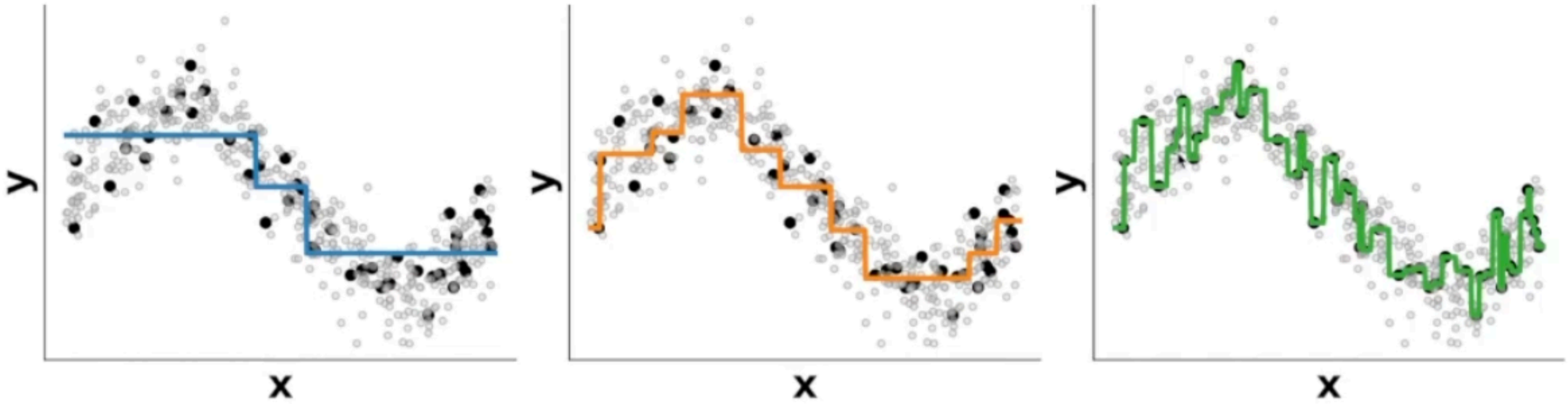


Source: Mooc scikit-learn



Attention au sur-apprentissage !

Il faut bien régler la profondeur de l'arbre (combien de noeuds décisionnels faire).



Source: Mooc scikit-learn

2. Données manquantes

Données manquantes

- Pourquoi des données manquantes ? Problèmes de capteurs, perte de données, non-réponse dans un sondage, agrégation de jeux de données (de deux hôpitaux différents par exemple), …
- **Plus il y a de données, plus il y a de données manquantes !**
- La présence des données manquantes **peut induire un biais dans l'analyse statistique**.

Pitie-Salpêtrière	88	0	No	3
Beaujon	103	0	NA	5
Bicêtre	NA	0	Yes	6
Bicêtre	NA	0	No	NA
Lille	62	0	Yes	6
Lille	NA	0	No	NA

Comment gérer les données manquantes ?

- **Gestion des données manquantes:**
  - On ne peut pas calculer NA+1
  - Objectif: avoir un **jeu de données complet (combler les *trous* formés par les données manquantes)**
  - Comment faire: **première idée naïve ?**

Comment gérer les données manquantes ?

- On aura nécessairement une **perte d'information**…
- Et souvent, faire l'analyse statistique seulement avec les points observés, cela peut induire un biais**.**

Pitie-Salpêtrière	88	0	No	3
~~Beaujon~~	~~103~~	~~0~~	~~NA~~	~~5~~
~~Bicêtre~~	~~NA~~	~~0~~	~~Yes~~	~~6~~
~~Bicêtre~~	~~NA~~	~~0~~	~~No~~	~~NA~~
Lille	62	0	Yes	6
~~Lille~~	~~NA~~	~~0~~	~~No~~	~~NA~~

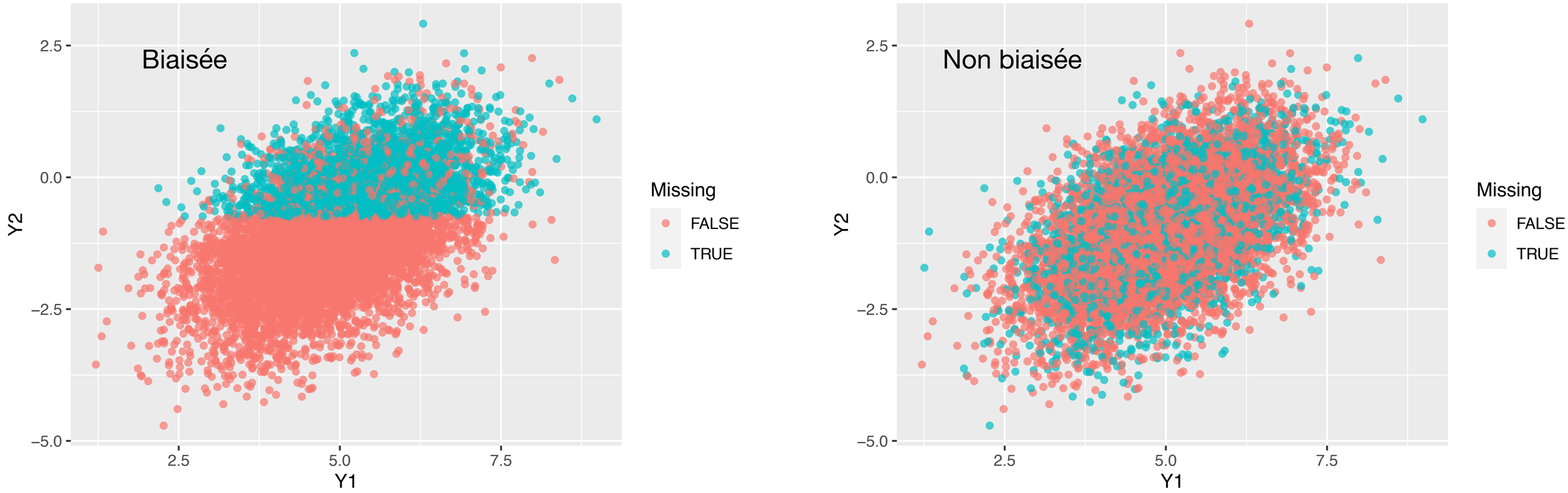
- **Et si on enlevait les données manquantes ?** (les lignes qui contiennent des valeurs manquantes)

Biais introduit par la présence des données manquantes

- Les points bleus sont manquants pour la variable Y2. Les points rouges sont observés pour les deux variables Y1 et Y2.
- Quelle sous-population de points observés va être biaisée ?

Biais introduit par la présence des données manquantes

- La sous-population de points observés est **biaisée à gauche**
- C-à-d cette sous-population de points observés **n'est pas représentative** de la population générale de points (observés+manquants)

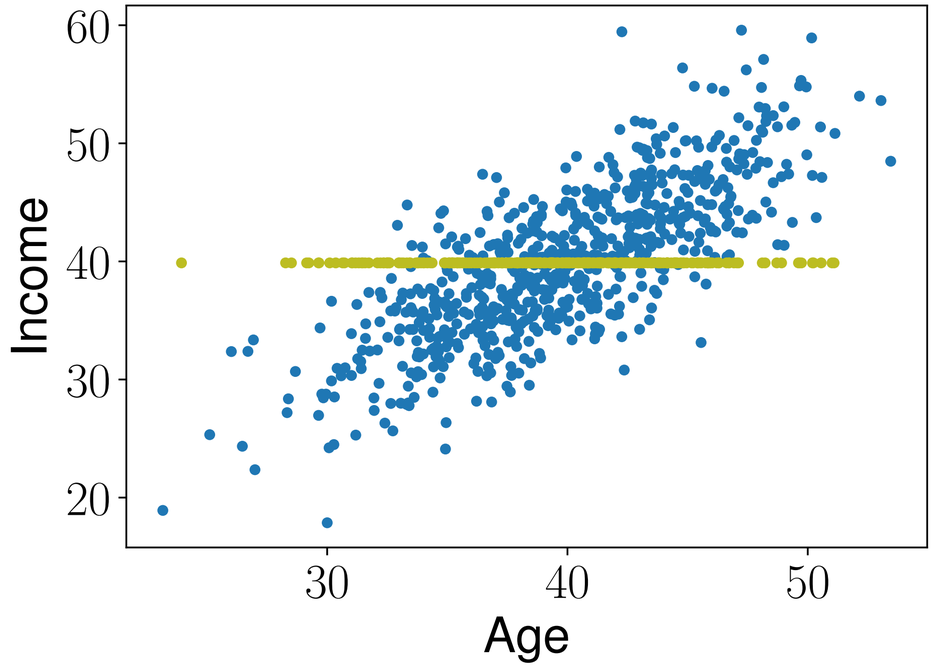


Imputation des données manquantes

Par la moyenne des variables

- Objectif: obtenir un jeu de données complet en *remplissant les trous formés par les données manquantes* avec des valeurs que l'on apprend
- On peut imputer les variables par la moyenne.

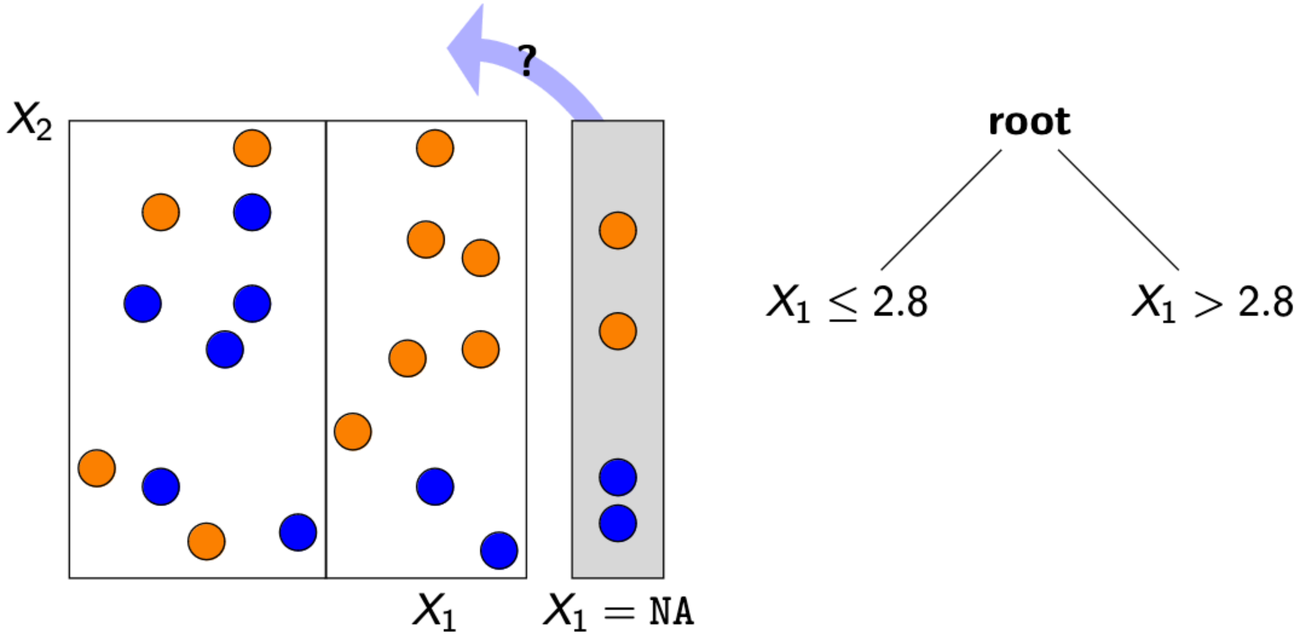
  - Si la variable **Revenu (Income)** est manquante, on peut imputer les valeurs manquantes par la moyenne de la variable **Income**
  - La qualité d'imputation n'est pas bonne



Imputation des données manquantes

En utilisant les forêts aléatoires

- De quel côté envoyer les données manquantes ?



Source: Nicolas Prost

3. TP: imputation de données manquantes

https://mybinder.org/v2/gh/AudeSportisse/Efelia-cours/HEAD?labpath=TP3_missing_values.ipynb