Random Forests

Objectifs

- Connaitre les méthodes ensemblistes
- Bagging vs Boosting
- Comprendre le Bagging
- Introduction aux Random Forests

Plan

- Arbres de décision et haute variance
- Méthodes d’ensemble
- Bagging
- Random Forests

Limites des arbres de décision

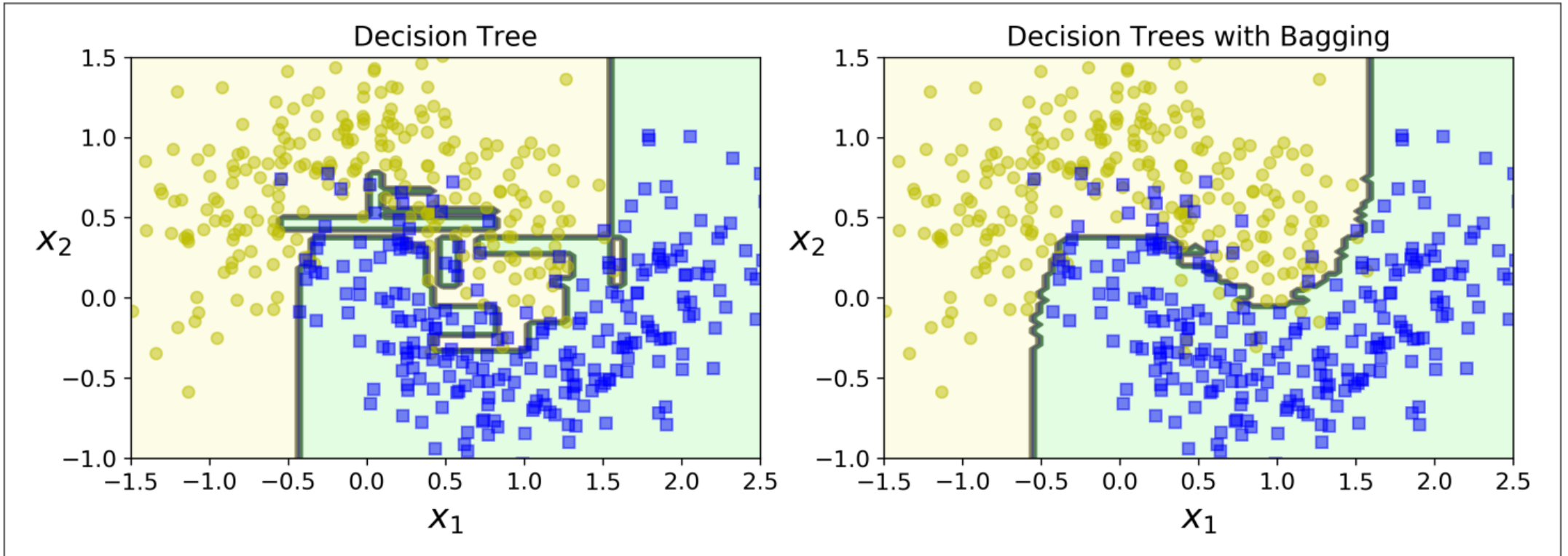
- Les arbres de décision sont très flexibles et peuvent s'adapter à des modèles complexes.
- Mais ce sont aussi des modèles à haute variance ➔ petits changements dans les données = grands changements dans la structure.
- Cette instabilité conduit à un sur-apprentissage (grande précision d'entraînement, mauvaise généralisation).

Les arbres de décision ont biais faible mais une variance élevée.
Ils capturent des détails importants dans les données —mais aussi du bruit.

Limites des arbres de décision

- Nous pouvons voir (à gauche) un arbre de décision qui s'adapte parfaitement aux données (limite de décision irrégulière)

- Un seul arbre s'adapte au bruit et fluctue avec de légères variations des données



Combiner des modèles pour réduire la variance

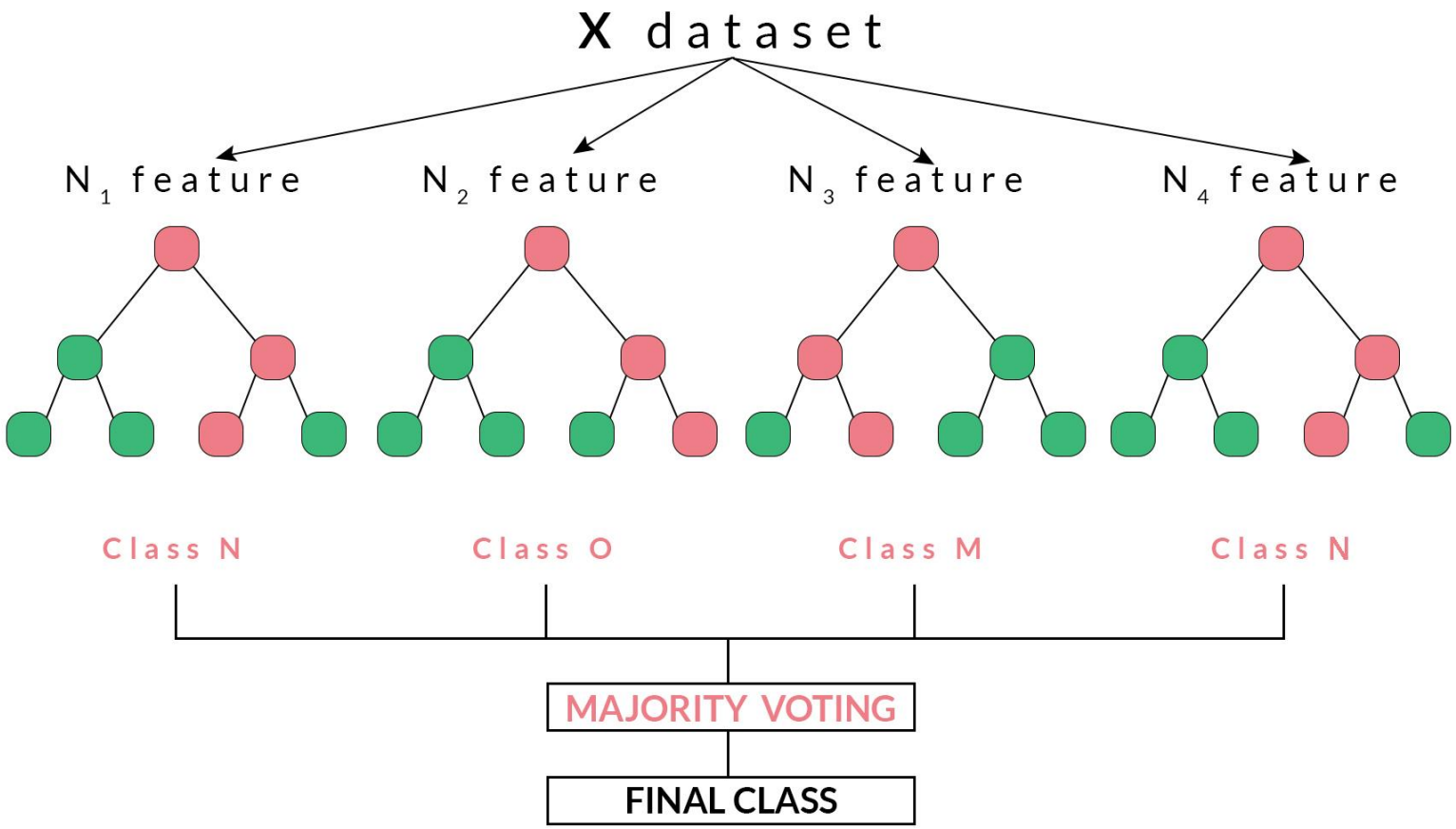
Est-ce que plusieurs modèles instables peuvent-ils devenir plus stables dans un ensemble ?

- Au lieu d'un seul modèle, nous entraînons plusieurs modèles indépendants et faisons la moyenne de leurs prédictions.
- La moyenne réduit la variance.

Wisdom of the Crowd :

Un seul modèle peut se tromper, cependant si de nombreux modèles différents donnent leurs avis,
La décision globale sera plus robuste et plus stable.

Combiner des modèles pour réduire la variance



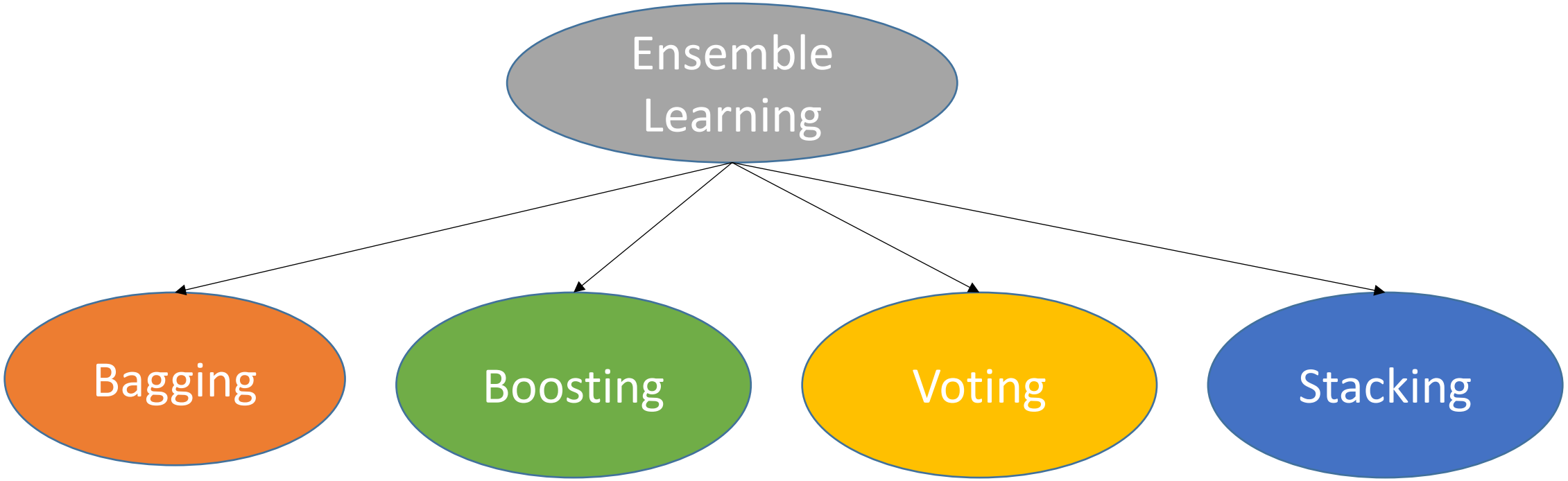
Méthodes d’ensemble

- Les méthodes d’ensemble combinent plusieurs modèles
- Les modèles de combinaison sont appelés base learners or weak learners
- L’ensemble produira un modèle plus robuste et plus précis.

Un groupe de modèles faibles peut surpasser un seul modèle fort s’ils sont combinés judicieusement.

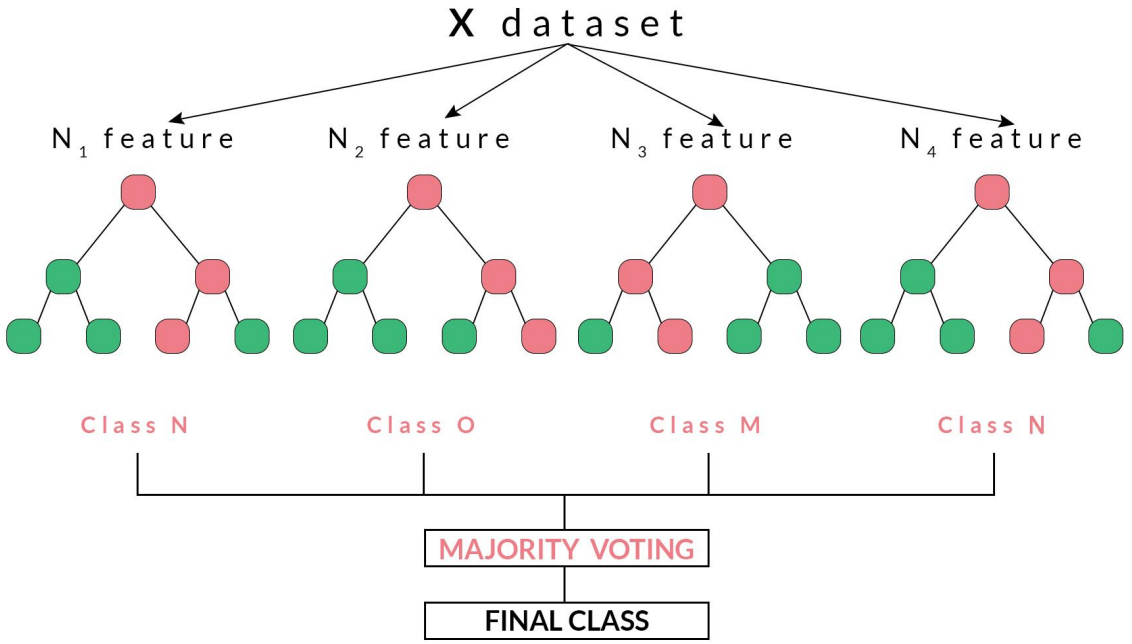
Méthodes d’ensemble

Nous pouvons diviser l’apprentissage d’ensemble comme suit:



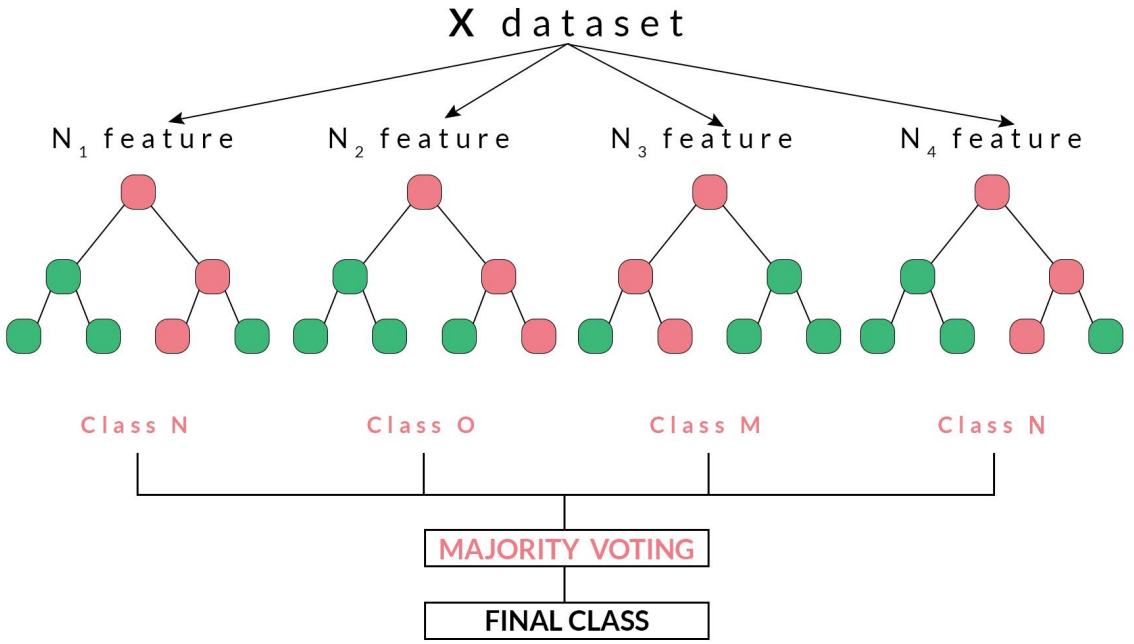
Méthodes d’ensemble : Bagging

- Crée différents sous-ensembles d’apprentissage par échantillonnage avec remplacement(bootstrap).
- Entraîne des modèles indépendants sur chaque sous-ensemble.
- Agrégats des prédictions : vote majoritaire (classification) ou moyenne (régression).
- Exemple: Random Forest



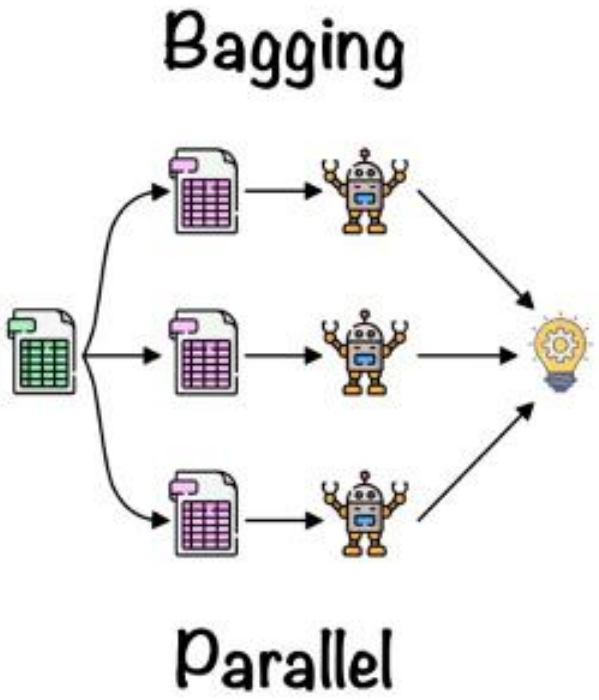
Méthodes d'ensemble: Boosting

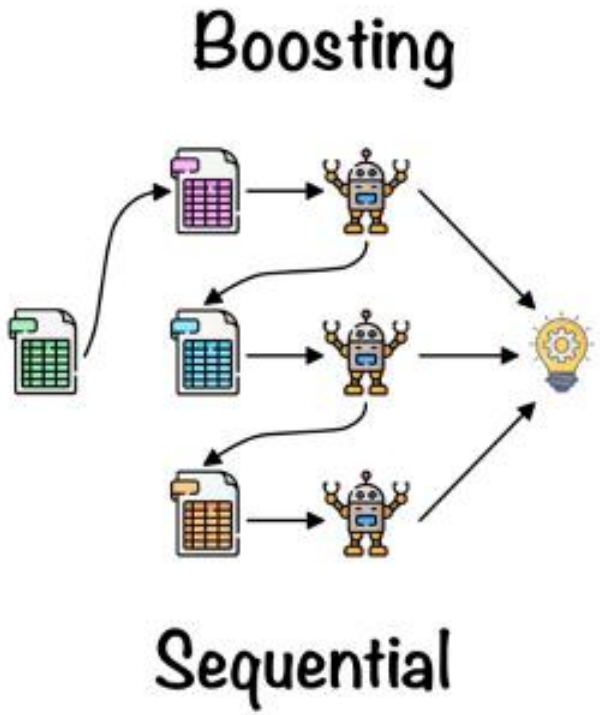
- Entraine les weak learners de manière séquentielle, chacun se concentrant sur les erreurs précédentes.
- Combine les weak learner pour produire un modèle final solide avec une précision améliorée.
- Exemple: AdaBoost, XGBoost…



Bagging vs Boosting

- Bagging entraine les base learners en parallèle
- Boosting entraine les base learners de manière séquentielle





Pourquoi utiliser le Bagging?

- Stratégie d’ensemble pour améliorer la stabilité et la précision des modèles prédictifs.
- Cible la réduction de la variance : rend les modèles à forte variance (par exemple, les arbres de décision profonds) plus fiables.
- Utile lorsqu’un seul modèle fait de l’overfitting sur des données d’entraînement mais présente un faible biais
- Utilisation courante dans le monde réel: Random Forests

Qu’est-ce que le Bagging ?

Bagging = Bootstrap Aggregation

➔ Entrainer plusieurs base learners sur différents échantillons bootstrapés de l’ensemble d’apprentissage, PUIS faire une moyenne de leurs prédictions.

Bootstrap Aggregation ?

Echantillons Bootstrap

- Sous-ensembles bootstrap avec remplacement à partir de l'ensemble de données d'origine ;
- Chaque taille d'échantillon est généralement égale à la taille de l'ensemble de données d'origine.

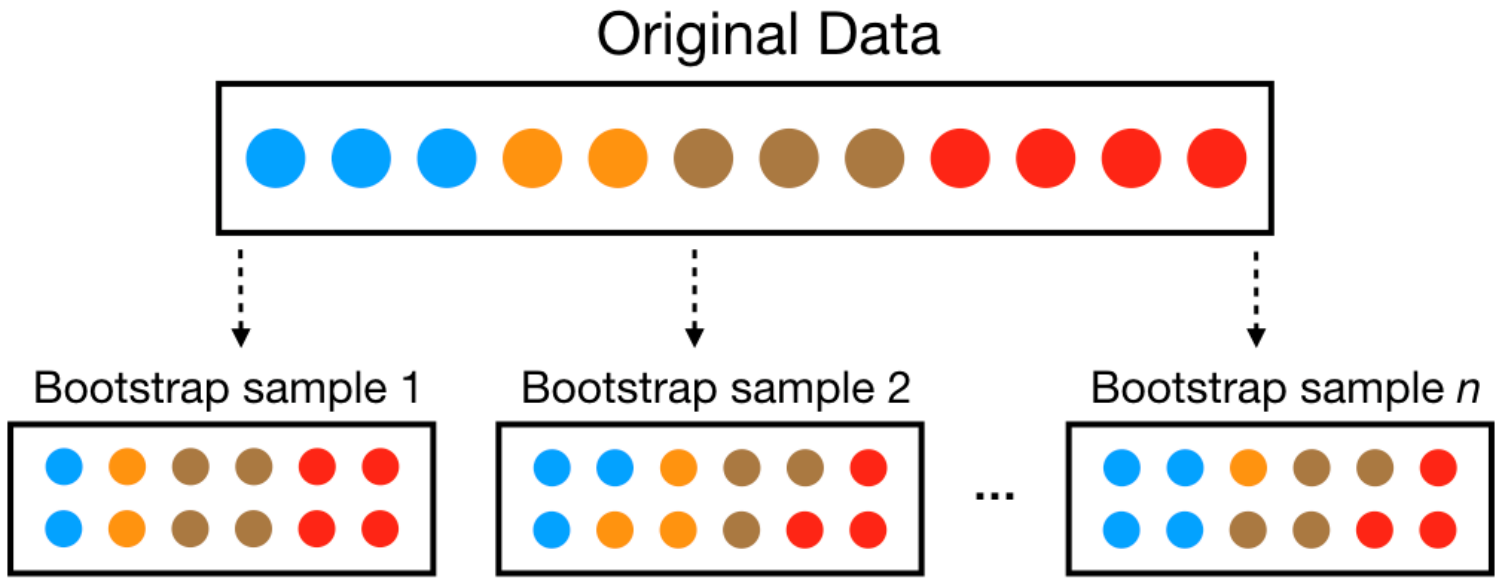
Agrégation:

- **vote majoritaire pour la classification, moyenne arithmétique pour la régression.**

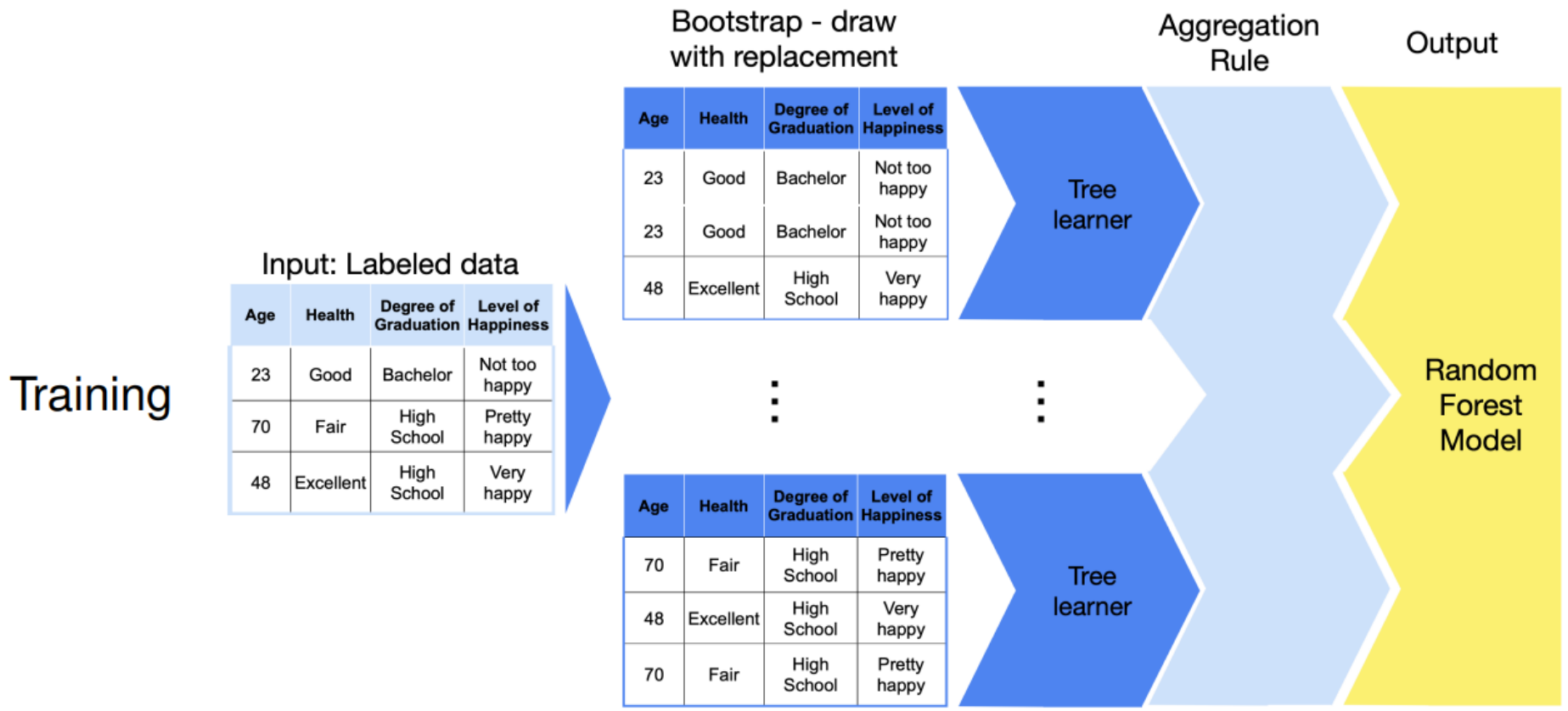
Echantillons Bootstrap

- Des échantillons aléatoires sont tirés de l'ensemble de données d'origine
- Les échantillons bootstrap sont généralement de la même taille que l'ensemble de données d'origine

- Bootstrapping avec **remplacement**
- **Le remplacement** signifie qu'un échantillon peut être tiré une ou plusieurs fois, tandis que d'autres ne seront pas tirés du tout

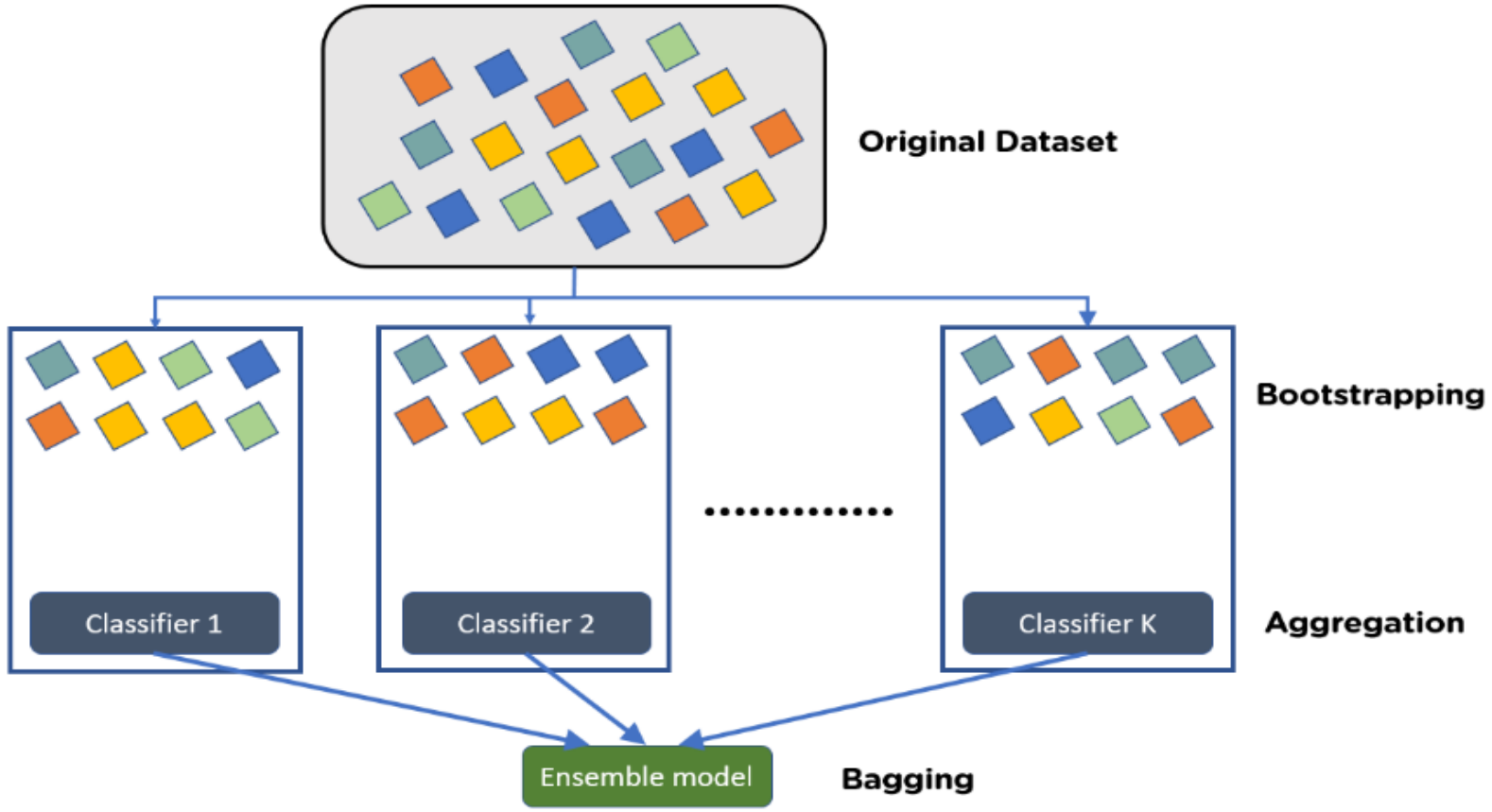


Bootstrap samples

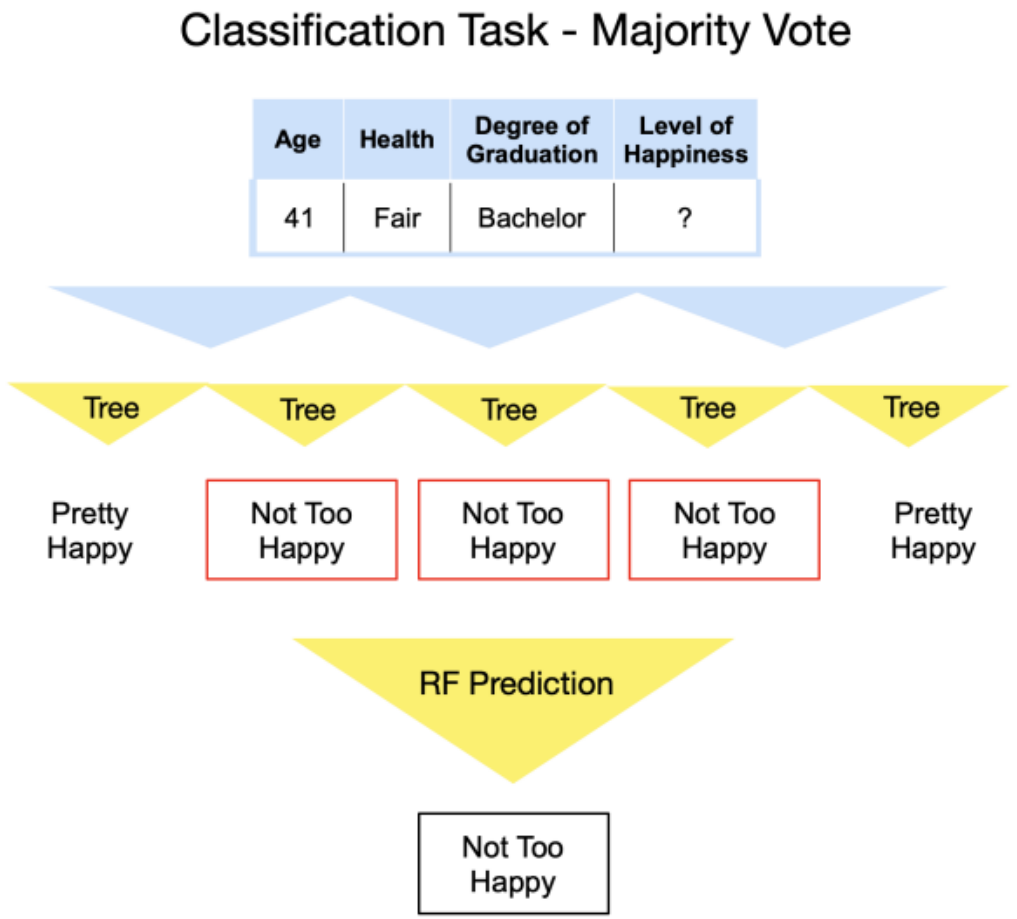


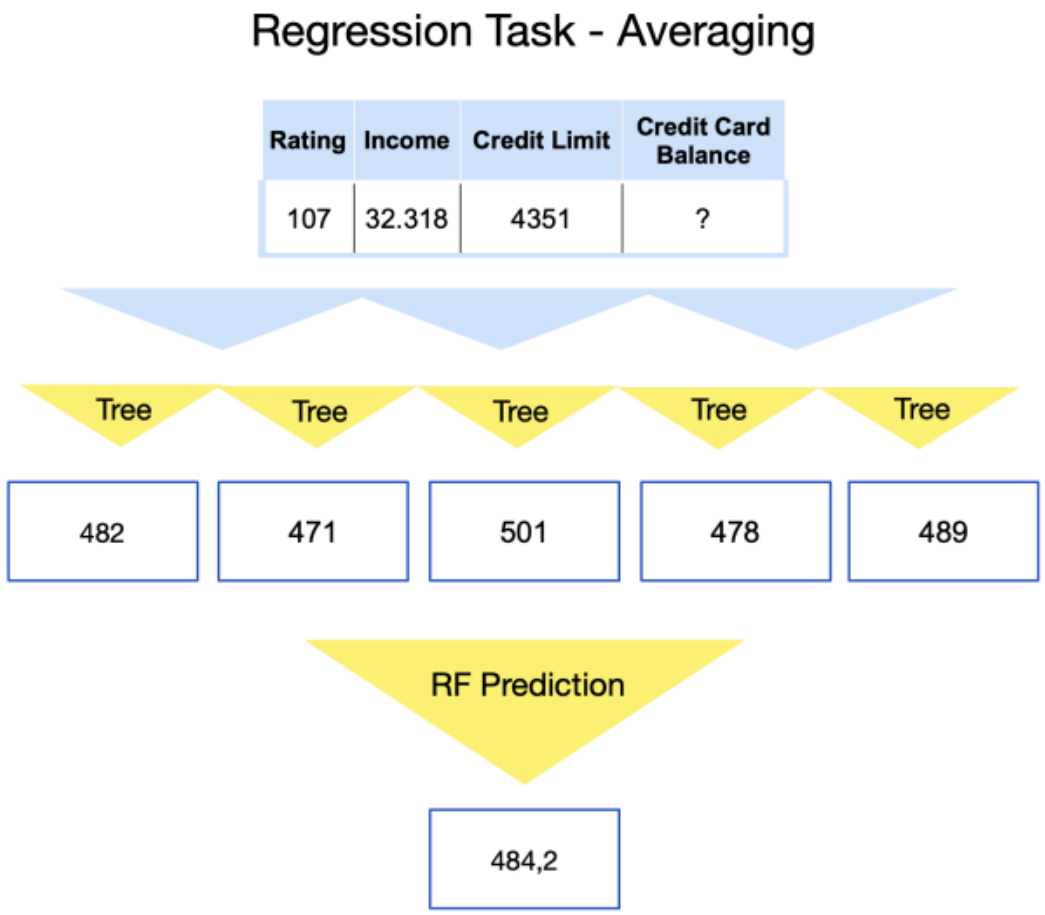
Agrégation

- Chaque Base learner est entrainé sur un échantillon bootstrap différent
- La prédiction finale est la moyenne des prédictions de chaque base learner



Agrégation


Classification Task - Majority Vote
Age
Health
Degree of Graduation
Level of Happiness
41
Fair
Bachelor
?
Tree
Tree
Tree
Tree
Tree
Pretty Happy
Not Too Happy
Not Too Happy
Not Too Happy
Pretty Happy
RF Prediction
Not Too Happy


Regression Task - Averaging
Rating
Income
Credit Limit
Credit Card Balance
107
32.318
4351
?
Tree
Tree
Tree
Tree
Tree
482
471
501
478
489
RF Prediction
484,2

Pourquoi le bootstrap fonctionne

- Chaque échantillon bootstrap contient
  - **Des échantillons originaux uniques**
  - **Échantillons répétés**
  - **Echantillons Out Of the Bag (OOB)**

- L'échantillonnage avec remplacement crée des distributions d'entraînement différentes pour chaque base learner ➔ **Les modèles font des erreurs différentes.**

- Calcul de la moyenne ou vote :
  - **Annule les erreurs indépendantes des modèles (bruit)**
  - **Réduit la variance tout en préservant le biais du base learner.**

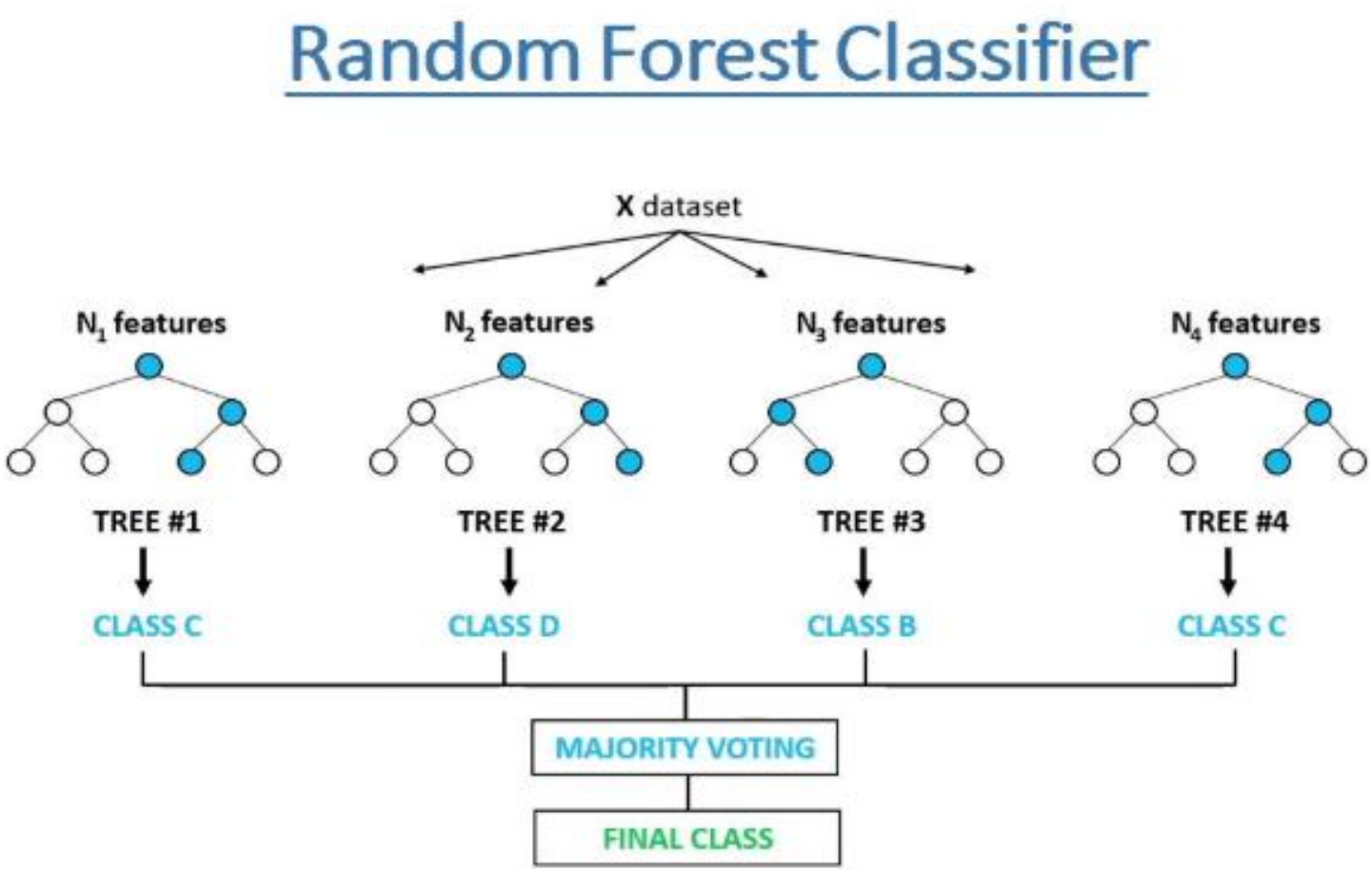
Algorithme du Bagging

1. **Given** training set $D = \{(x_i, y_i)\}_{i=1}^{N}$ and a base learning algorithm A.
2. For $b = 1, \ldots, B$:
   - Draw bootstrap sample D_b by sampling N examples from D **with replacement**.
   - Train base learner $h_b = A(D_b)$.
3. For a new input x:
   - Classification: predict $\hat{y} = \arg\max_c \sum_{b=1}^{B} \mathbf{1}\{h_b(x) = c\}$ (majority vote).
   - Regression: predict $\hat{y} = \frac{1}{B} \sum_{b=1}^{B} h_b(x)$ (average).

Random Forests

- Random Forests are un algorithme de Bagging
- Il s'agit d'un ensemble d'arbres de décision utilisant le Bagging + **une sélection aléatoire des variables.**
- Random Forests utilise des **Arbres de décision** comme base learner
- Chaque arbre est entraîné sur des échantillons bootstrap de l'ensemble de données
- Le résultat final est la prédiction de chaque arbre

Random Forests

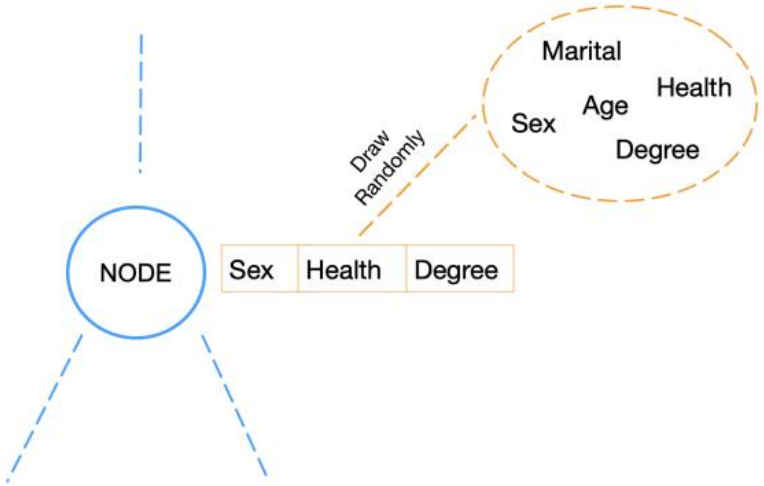


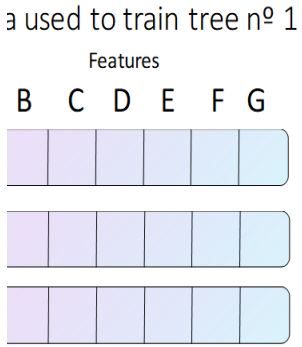
Sélection aléatoire des variables

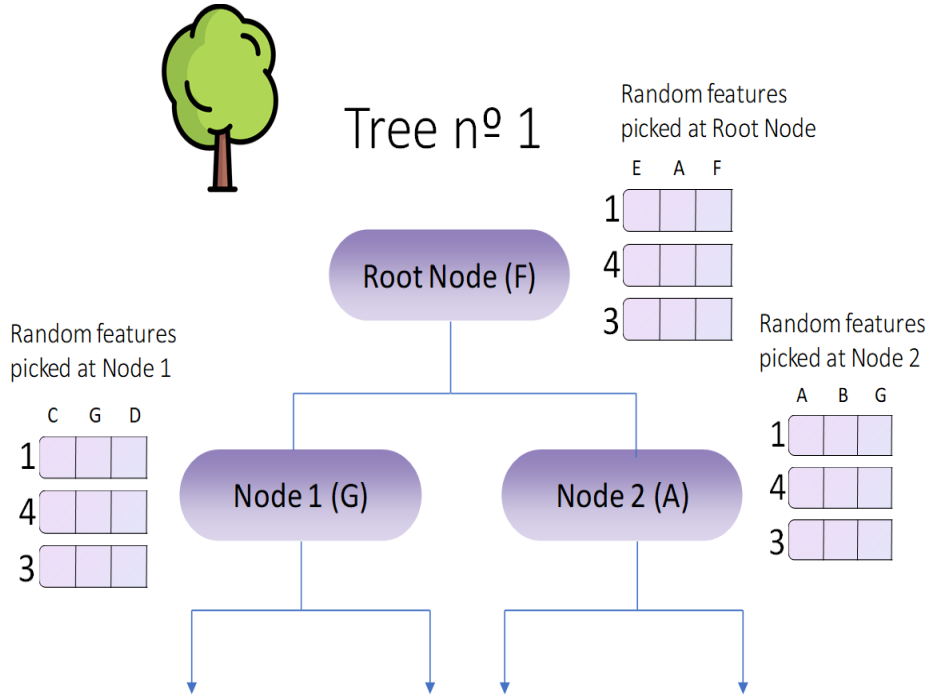
- La nouveauté des Random Forests est la sélection aléatoire des caractéristiques
- À **chaque division de l'arbre**, l'algorithme introduit une **sélection aléatoire de caractéristiques**
- Un **sous-ensemble aléatoire** des variables sera utilisé
- Le but ici est de **réduire la corrélation des arbres**

Sélection aléatoire des variables

- À chaque nœud, un sous-ensemble aléatoire des variables est choisi
- Ceci va décorrélé les arbres dans le Random Forest
- Chaque arbre verra des données différentes et sera entrainé à l'aide de variables différentes







Avantages des Random Forests

- **Réduit la variance** en faisant la moyenne de plusieurs arbres et nullifie les erreurs d'arbre individuelles.

- **Robuste au surapprentissage** dû à l'aggrégation des résultats des arbre, ceci malgré que certains arbres puissent être très complexes (profonds)

- Peut gérer efficacement des données de grande dimension et de grands ensembles de données.
- Fonctionne très bien avec les **données tabulaires**

- Fournit la **Feature Importance** qui explique quelles variables contribuent le plus aux prédictions.

Limites des Random Forests

- Moins interprétable qu'un seul arbre de décision :

  ➔Nécessitera d'afficher tous les arbres pour comprendre la prédiction finale ce qui n'est pas toujours facile (avec un grand nombre d'arbres)

- L'entrainement peut prendre énormément de temps surtout lorsqu'un grand nombre d'arbres est utilisé