FORÊTS ALÉATOIRES, SÉLECTION DE VARIABLES ET SÉCURITÉ AÉRIENNE

PHILIPPE SAINT PIERRE

EN COLLABORATION AVEC B. GREGORUTTI, B. MICHEL

IMT, Equipe de Statistique et Probabilités

Université Paul Sabatier

SÉMINAIRE MIAT

Plan

APPRENTISSAGE STATISTIQUE

Apprentissage automatique + statistique

APPRENTISSAGE AUTOMATIQUE

- Observation d'un phénomène
- Construction d'un modèle adapté à ce phénomène
- Analyse et prédiction du phénomène à partir du modèle

$\rightarrow$ Processus automatique (pas d'intervention humaine)

STATISTIQUE

- Formalisation
- Evaluer la qualité du modèle

Objectifs : explorer, expliquer, prévoir, sélectionner des variables.

APPRENTISSAGE NON SUPERVISÉ (CLUSTERING)

TYPE DE DONNÉES

- On observe un ensemble de variables $\mathbf{X}_i$ pour chaque individu $i = 1, \ldots, n$.

$$\{\mathbf{X}_i, i = 1, \ldots, n\}.$$

OBJECTIF

Rechercher des classes d'individus homogènes dans un échantillon

- les individus similaires sont associés au même groupe,
- les individus considérés comme différents se retrouvent dans des groupes distincts.

$\rightarrow$ le nombre de groupes et la nature des groupes sont inconnus.

Exemple : Identifier des groupes de malades et chercher à les expliquer.

APPRENTISSAGE SUPERVISÉ

TYPE DE DONNÉES

- On observe un ensemble de variables $\mathbf{X}_i$ pour chaque individu $i = 1, \ldots, n$.
- On observe une sortie Y_i pour chaque individu $i = 1, \ldots, n$.

$$\{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}.$$

OBJECTIFS

1. Apprendre un modèle pour expliquer la sortie à partir d'une base d'apprentissage.
2. Utiliser ce modèle pour prédire la sortie d'un nouvel individu.

QUELQUES MÉTHODES D'APPRENTISSAGE SUPERVISÉ

- Régression linéaire et régression logistique
- Machine à vecteurs de support
- Méthode des k plus proches voisins
- Arbre de décision et forêts aléatoires
- ...

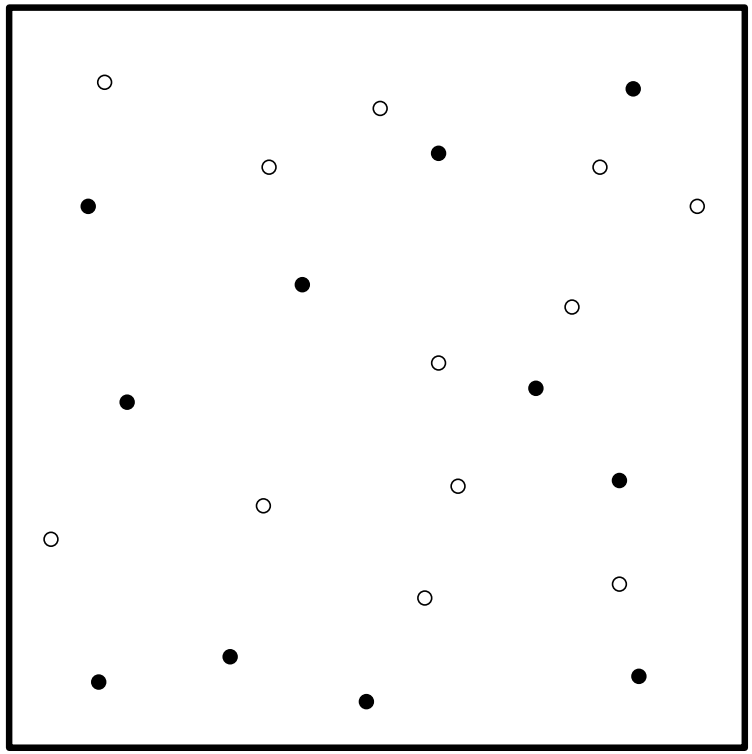
AUTRES CATÉGORIE DE MÉTHODES D'APPRENTISSAGE

- Apprentissage semi-supervisé
- Apprentissage actif
- ...

CLASSIFICATION

Expliquer le statut malade ou non malade.

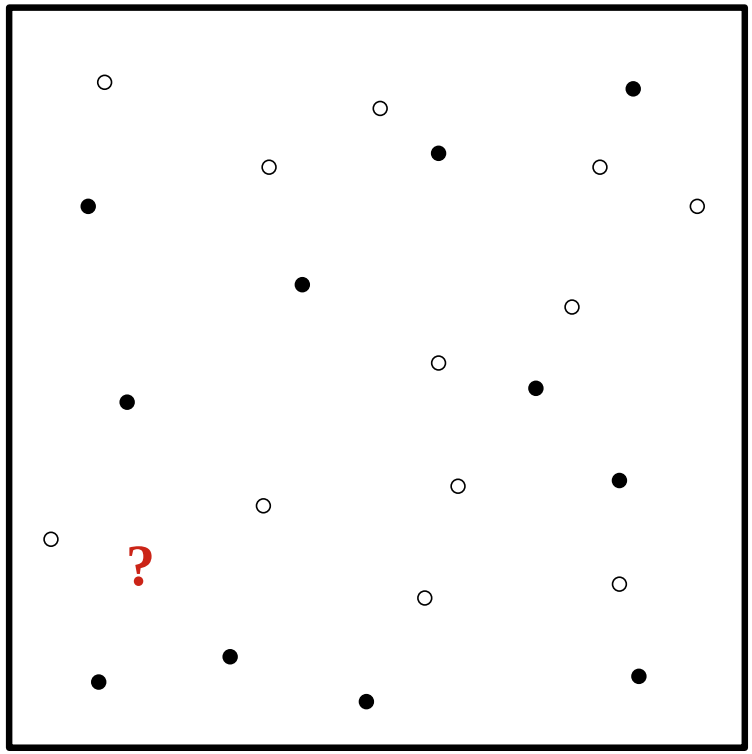
Prédire le statut d'un nouvel individu (malade - non malade).



CLASSIFICATION

Expliquer le statut malade ou non malade.

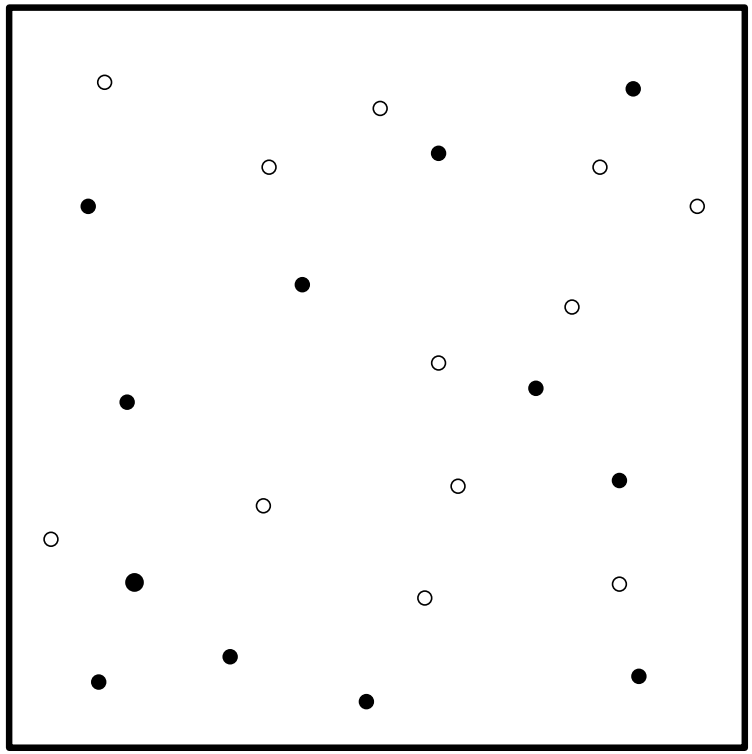
Prédire le statut d'un nouvel individu (malade - non malade).



CLASSIFICATION

Expliquer le statut malade ou non malade.

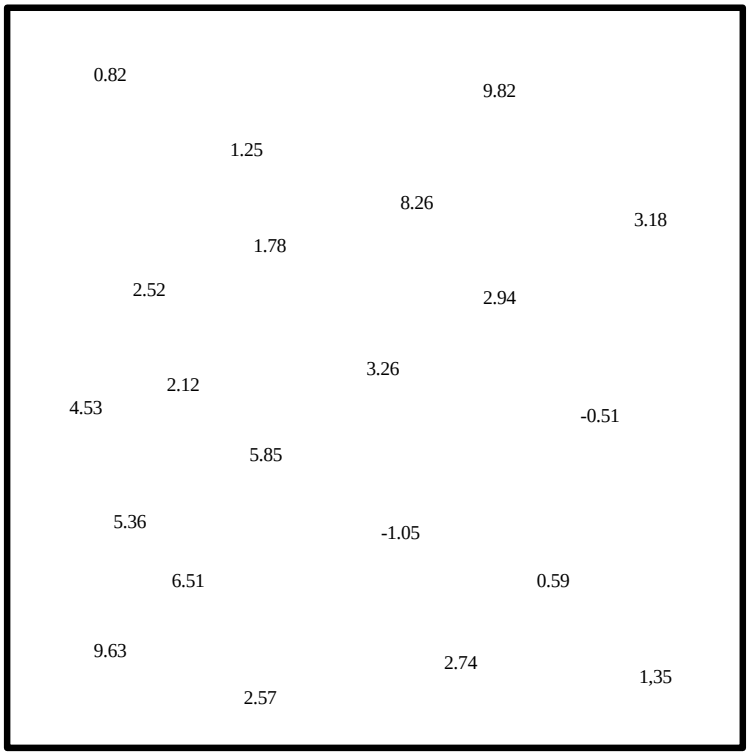
Prédire le statut d'un nouvel individu (malade - non malade).



RÉGRESSION

Expliquer le taux d'Ozone.

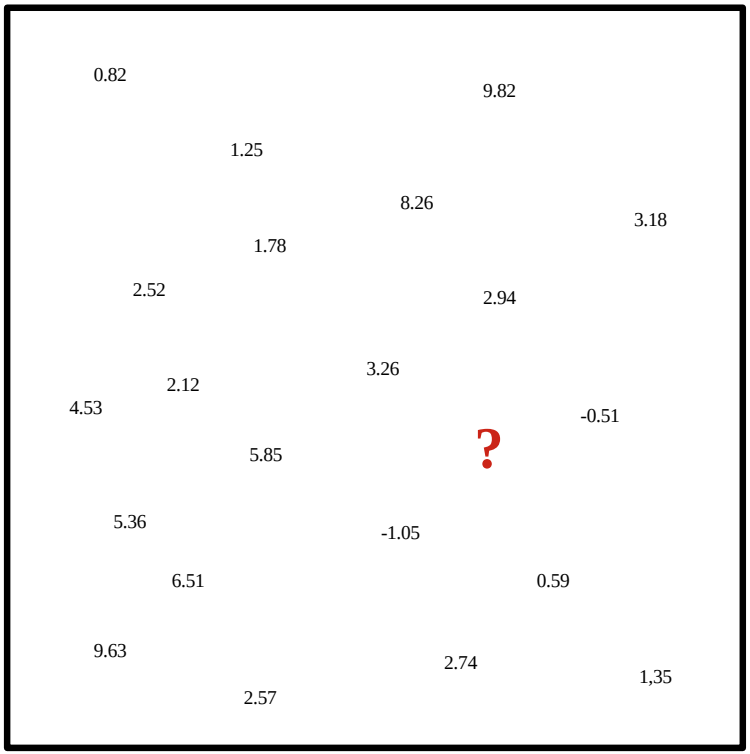
Prédire un taux d'Ozone.



RÉGRESSION

Expliquer le taux d'Ozone.

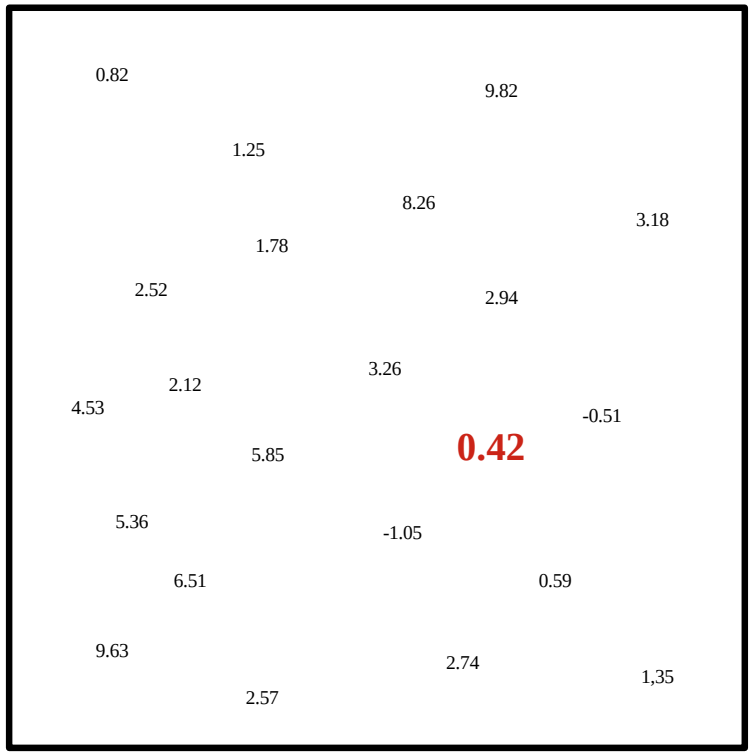
Prédire un taux d'Ozone.



RÉGRESSION

Expliquer le taux d'Ozone.

Prédire un taux d'Ozone.



On suppose qu'on observe un échantillon d'apprentissage

$$\mathscr{D}_n = (\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)$$

$(\mathbf{X}_i, Y_i) \in \mathbb{R}^p \times \mathscr{Y}$ sont $i.i.d.$ de loi $(\mathbf{X}, Y)$

$\mathbf{X} = (X_1, ..., X_p)$ un vecteur de covariables

Y la variable d'intérêt

Classification : $Y \in \{-1, 1\}$

- Estimer $f(\mathbf{x}) = \mathbb{P}[Y = 1 | \mathbf{X} = \mathbf{x}]$ à partir de $\mathscr{D}_n$

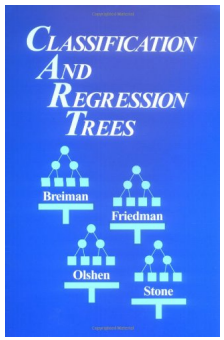
Régression : $Y \in \mathbb{R}$

- Estimer $f(\mathbf{x}) = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}]$ à partir de $\mathscr{D}_n$

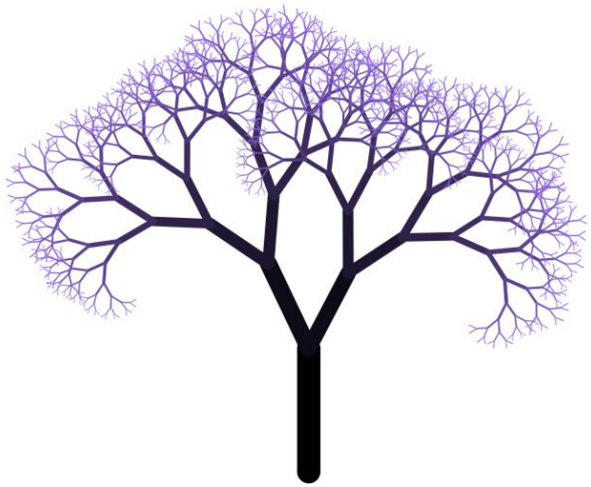
On ne cherche pas à estimer la distribution des données $(\mathbf{X}, Y)$

CLASSIFICATION AND REGRESSION TREE

- Breiman et al. (1984)

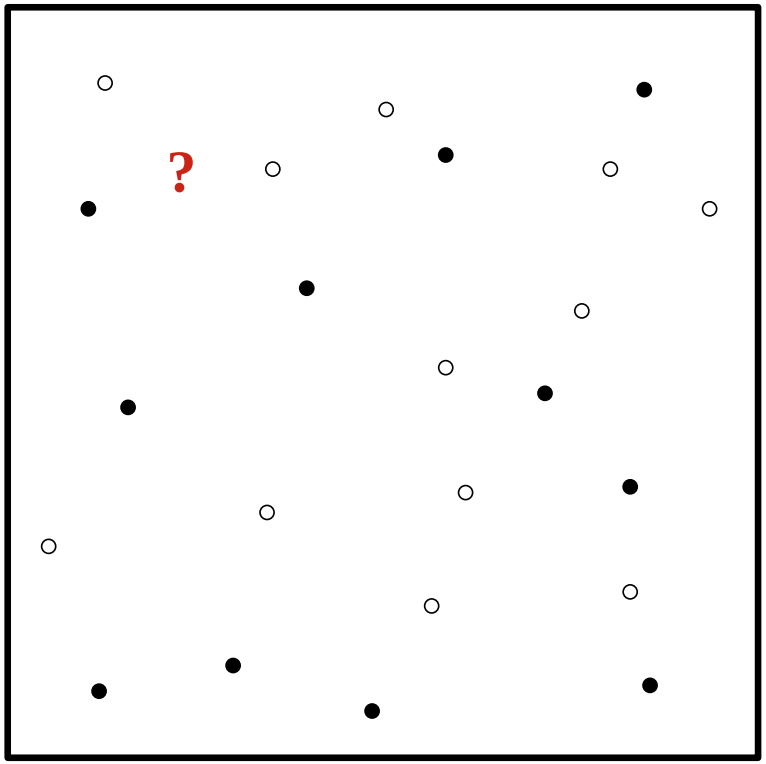


- Arbre binaire : construit de manière récursive en découpant chaque feuille en deux noeuds fils jusqu'à l'obtention d'un critère d'arrêt.



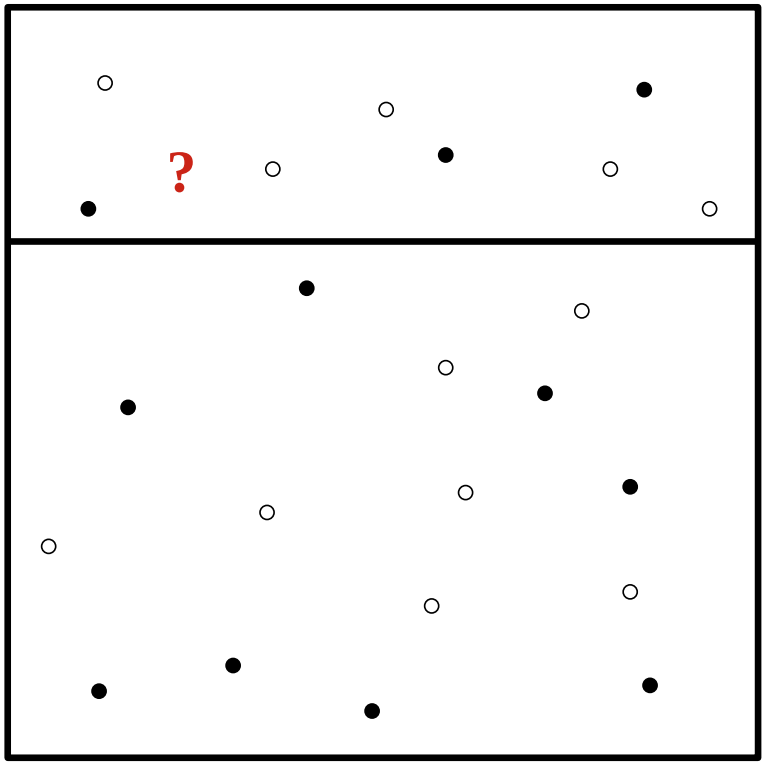
Exemple en classification

On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



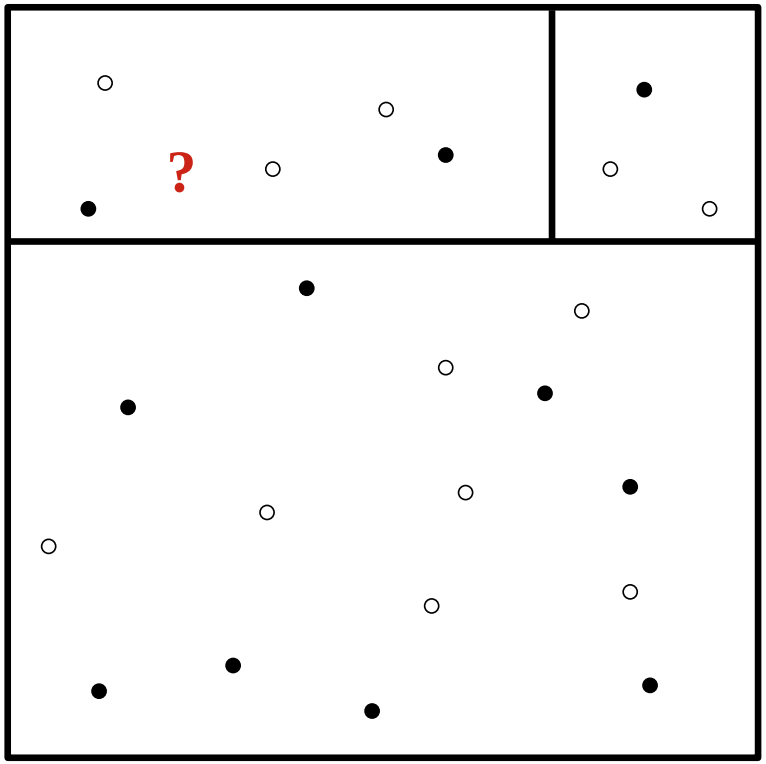
Exemple en classification

On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



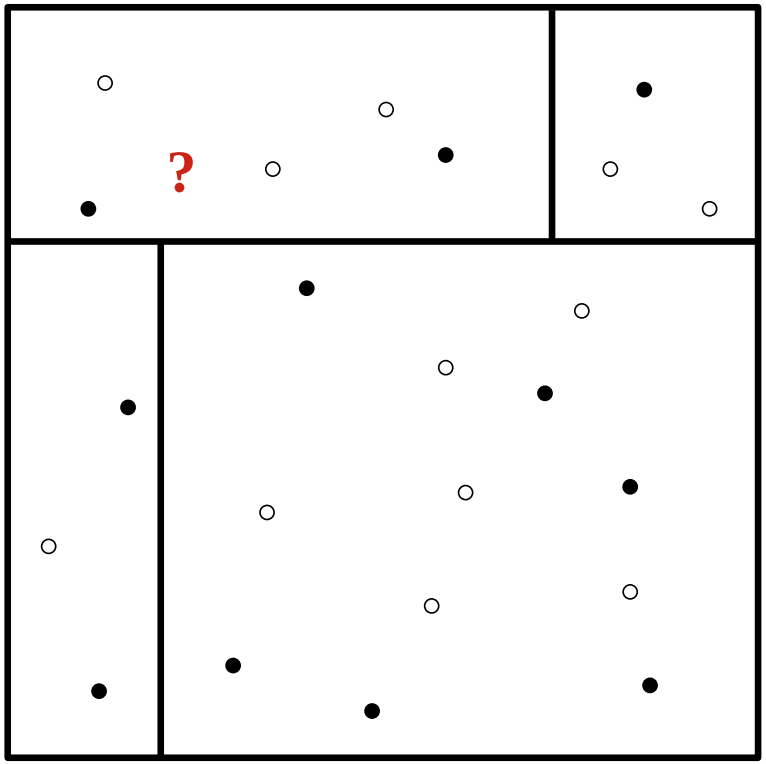
Exemple en classification

On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



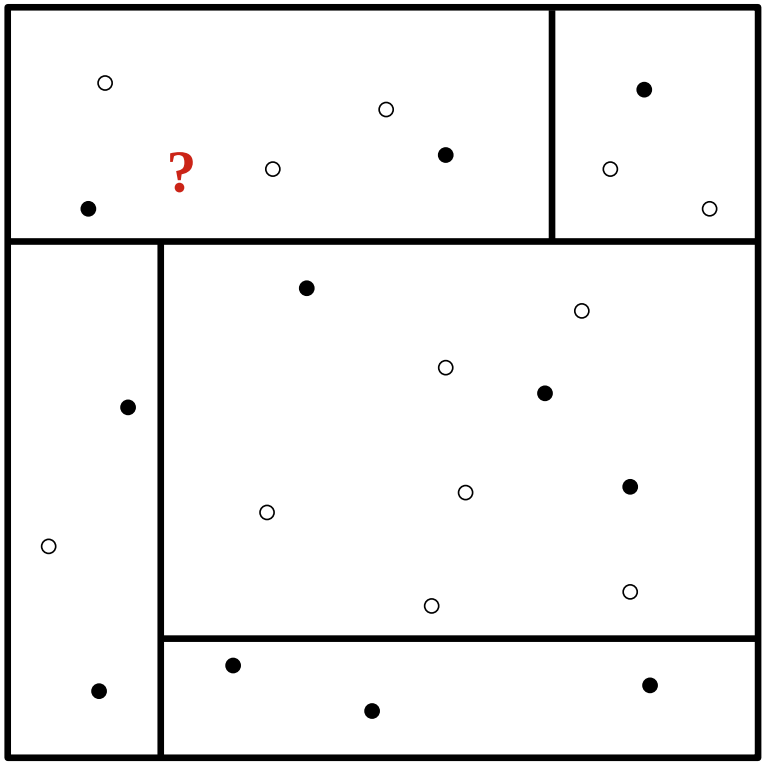
Exemple en classification

On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



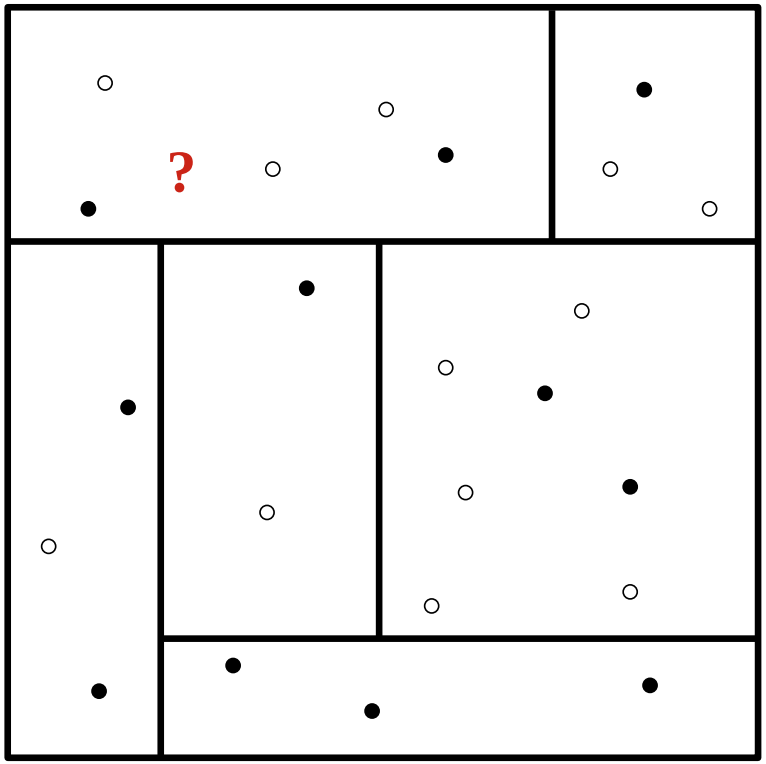
Exemple en classification

On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



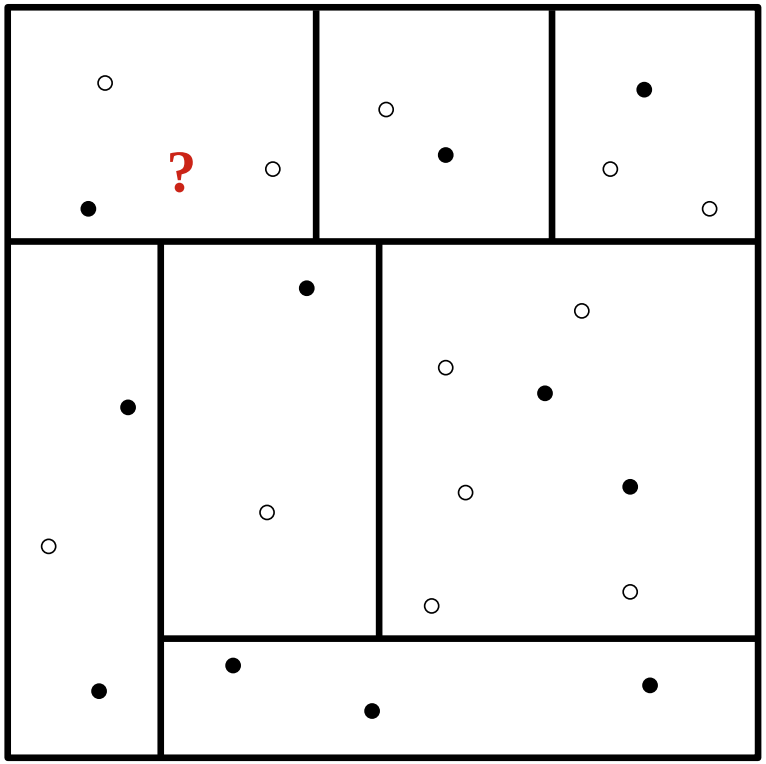
Exemple en classification

On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



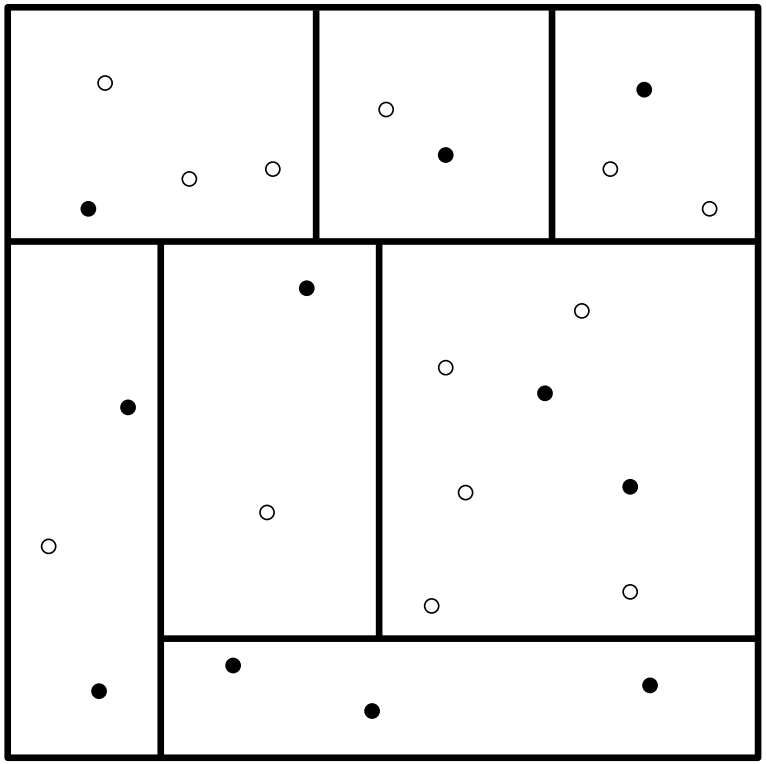
Exemple en classification

On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



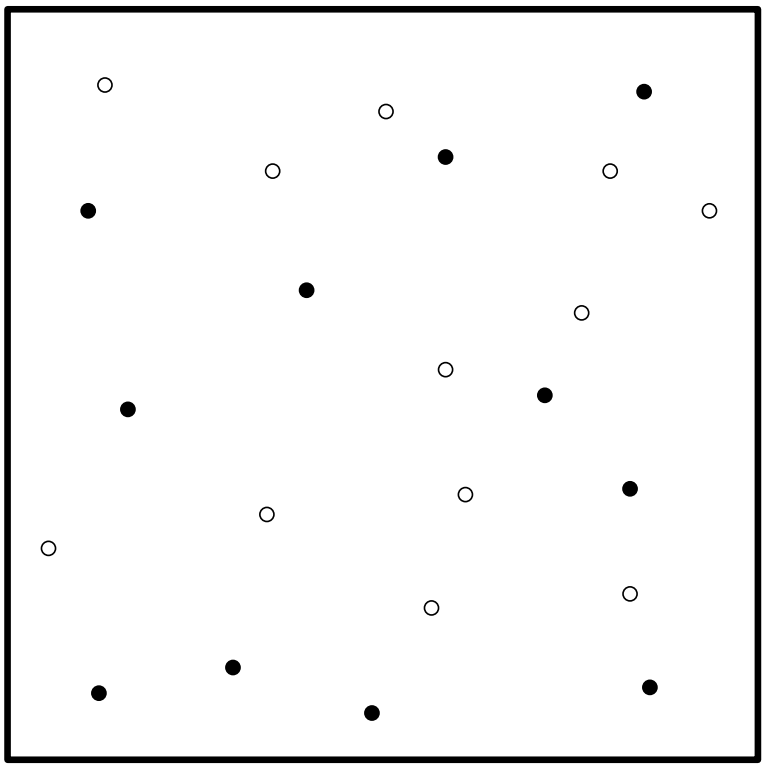
Exemple en classification

On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



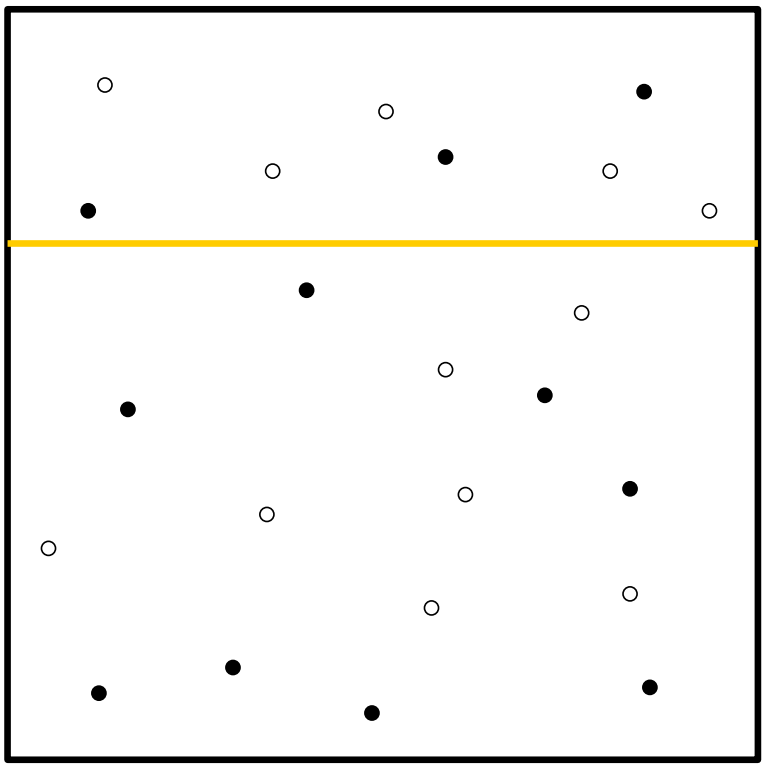
Exemple en classification

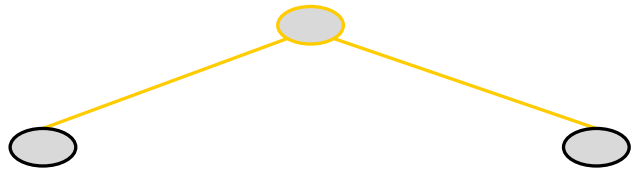
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



Exemple en classification

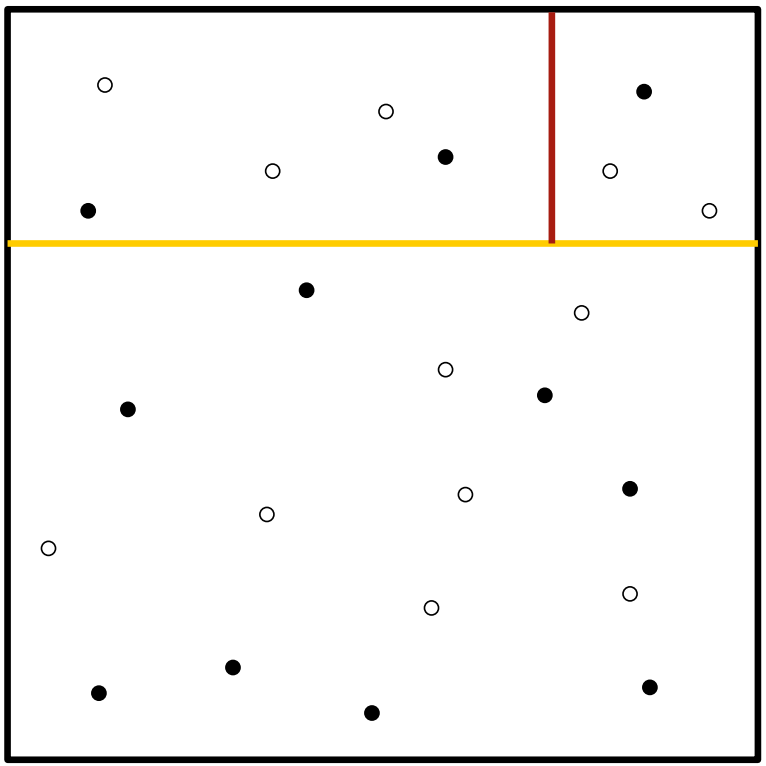
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

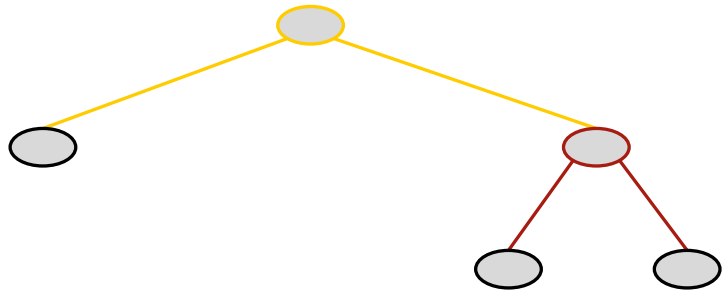




Exemple en classification

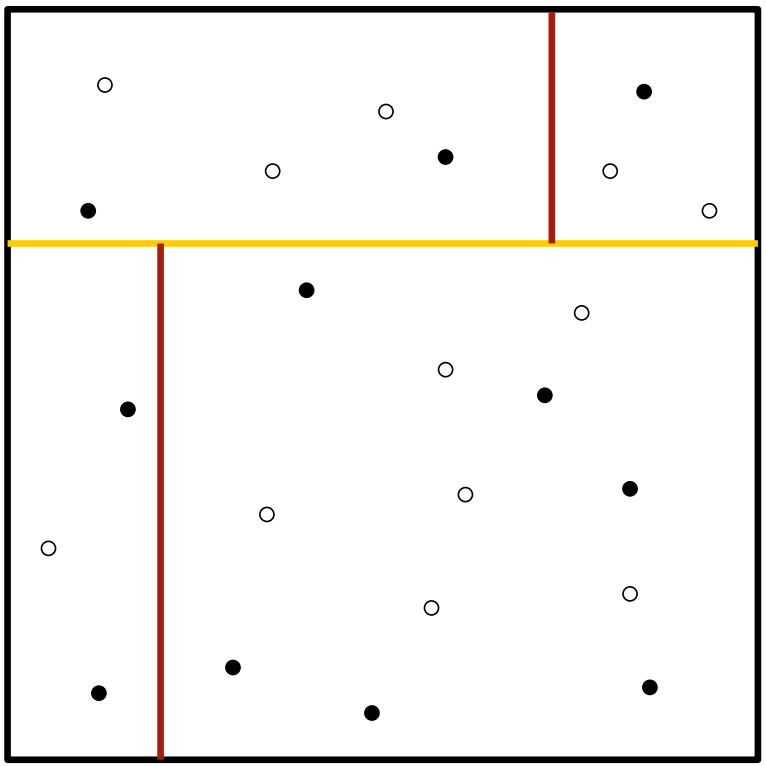
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

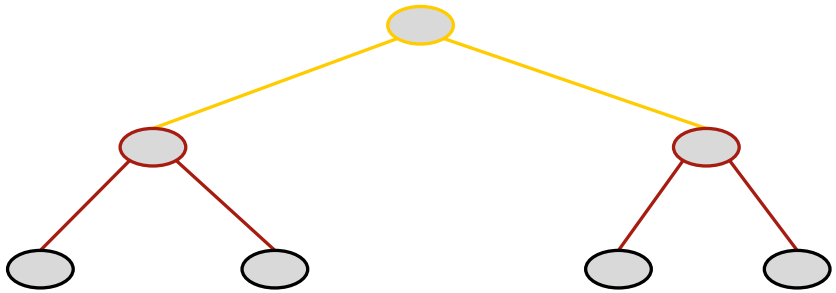




Exemple en classification

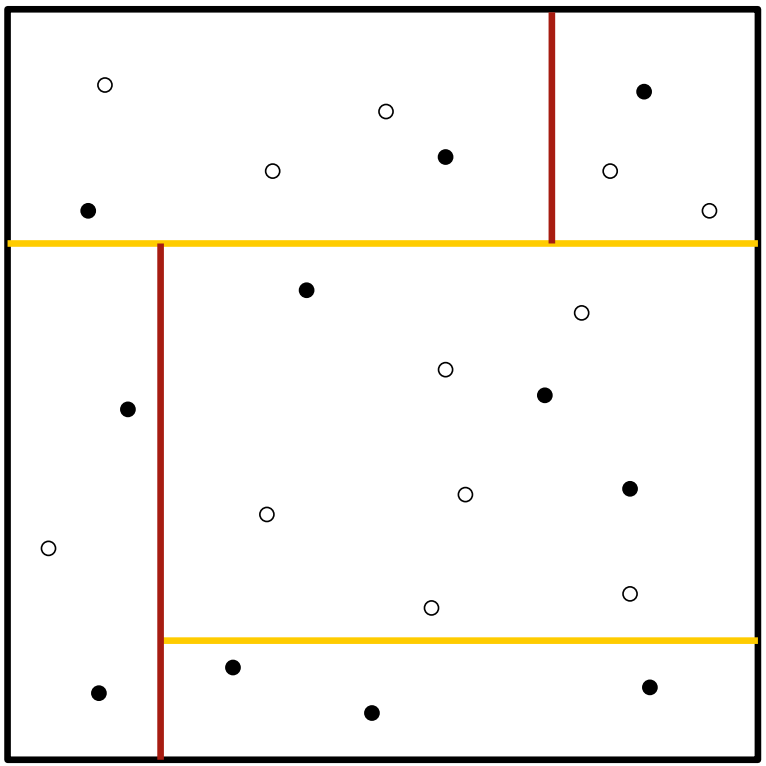
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

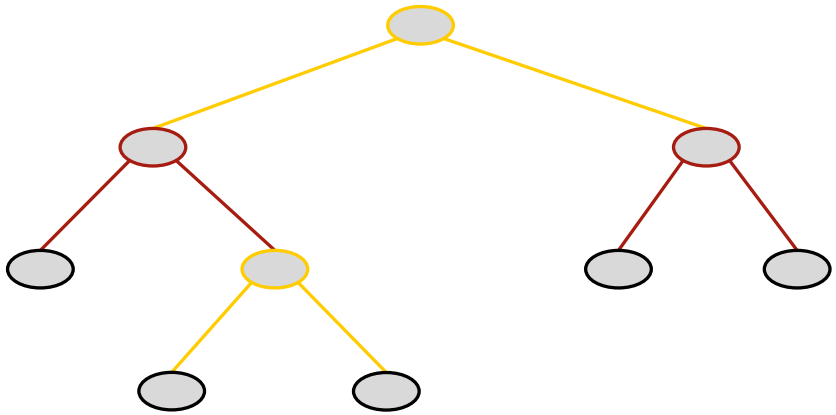




Exemple en classification

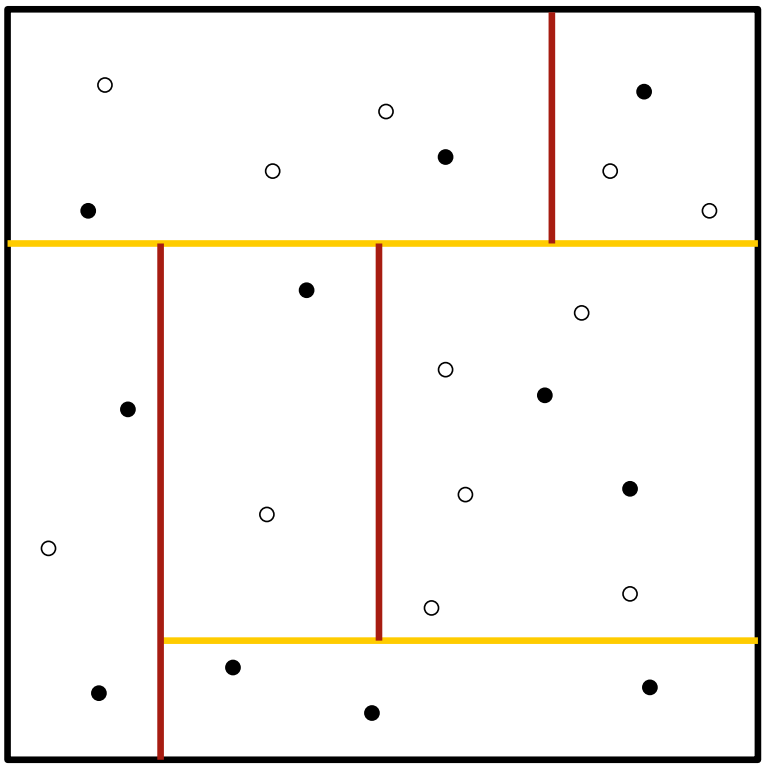
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

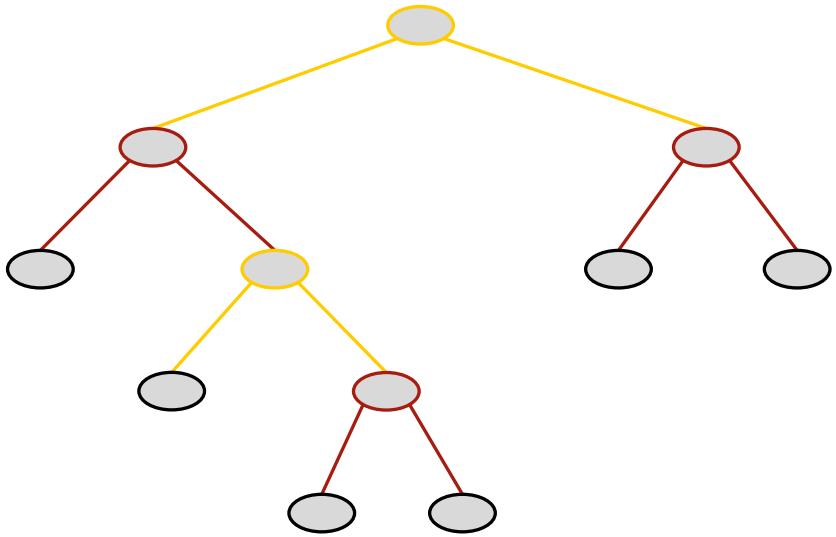




Exemple en classification

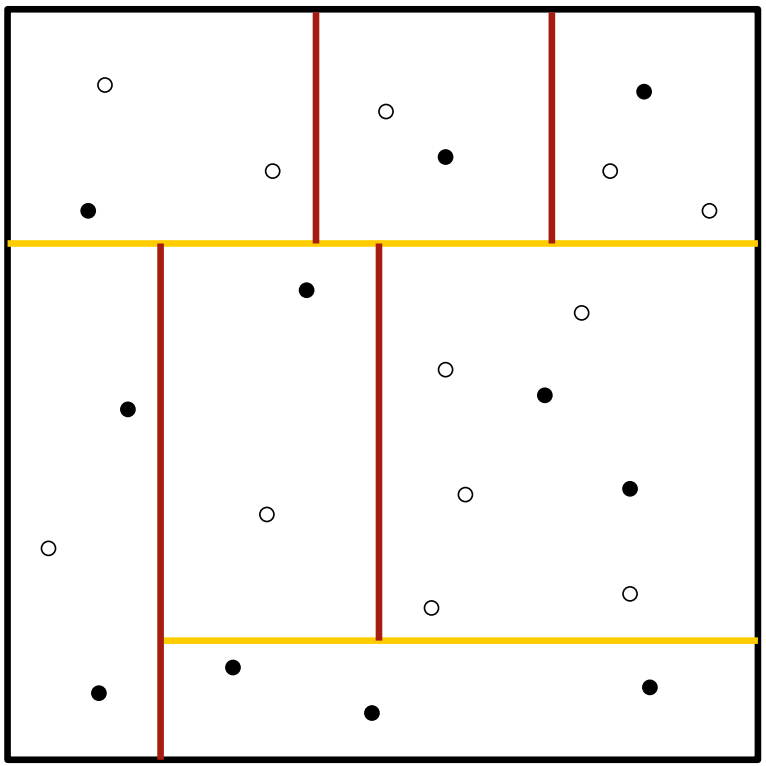
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

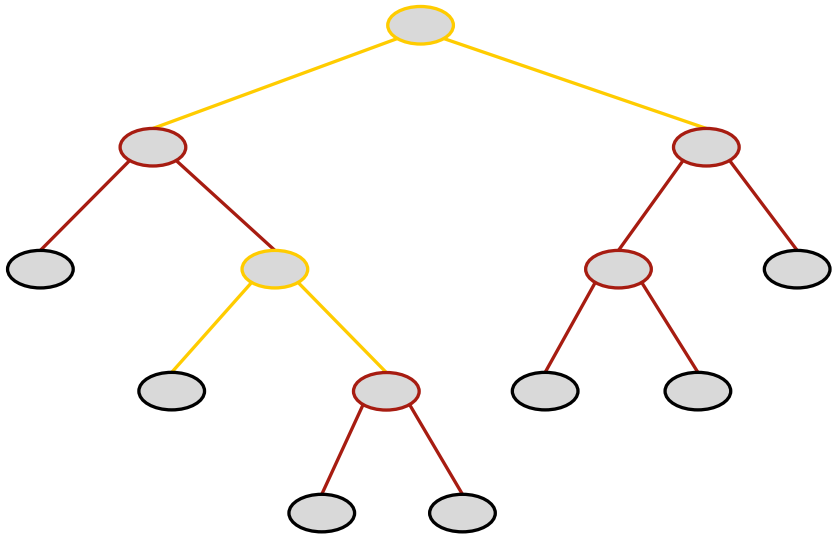




Exemple en classification

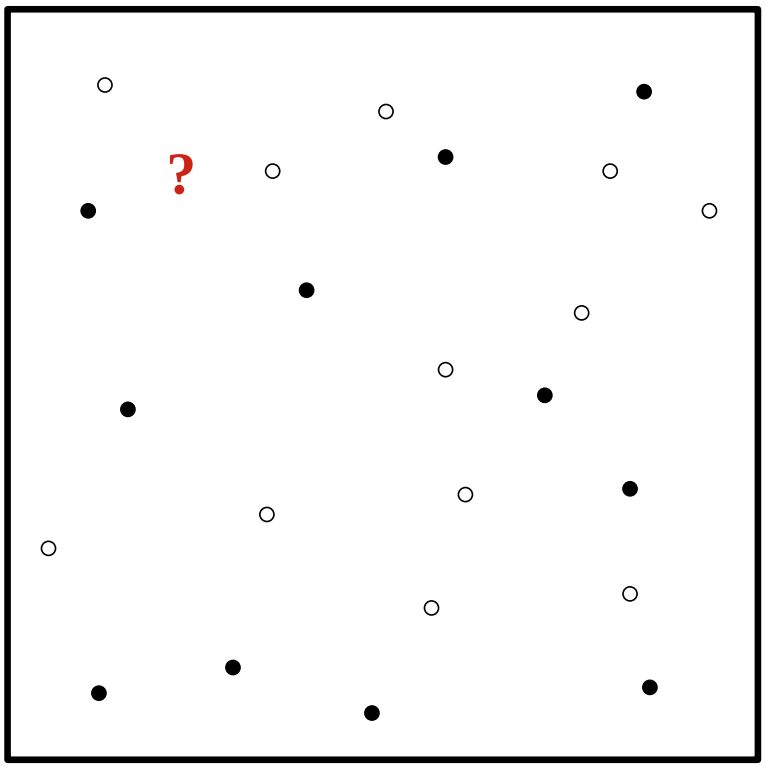
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.





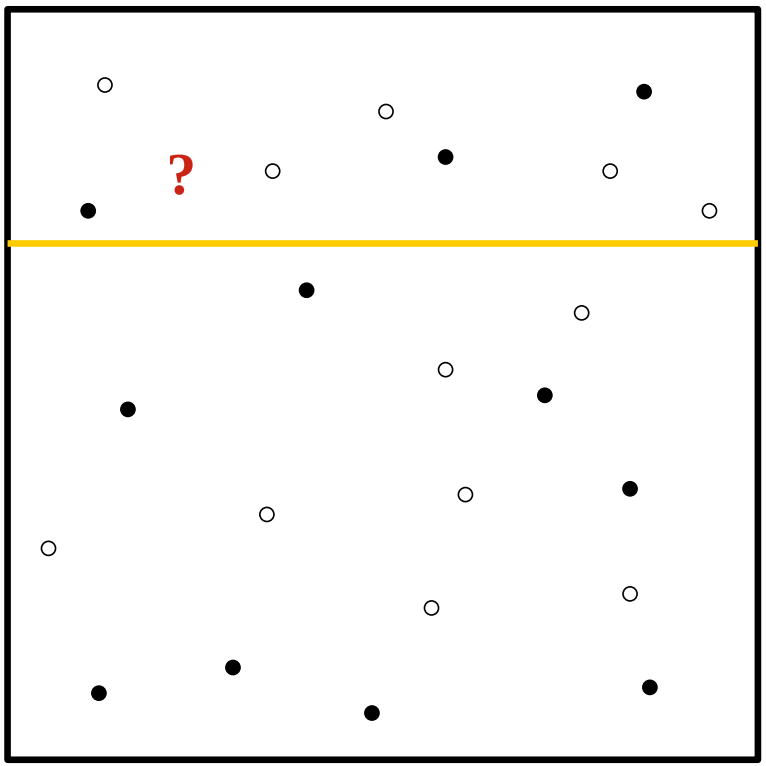
Exemple en classification

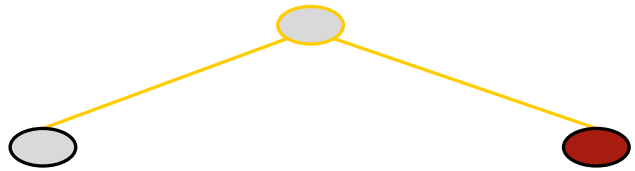
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



Exemple en classification

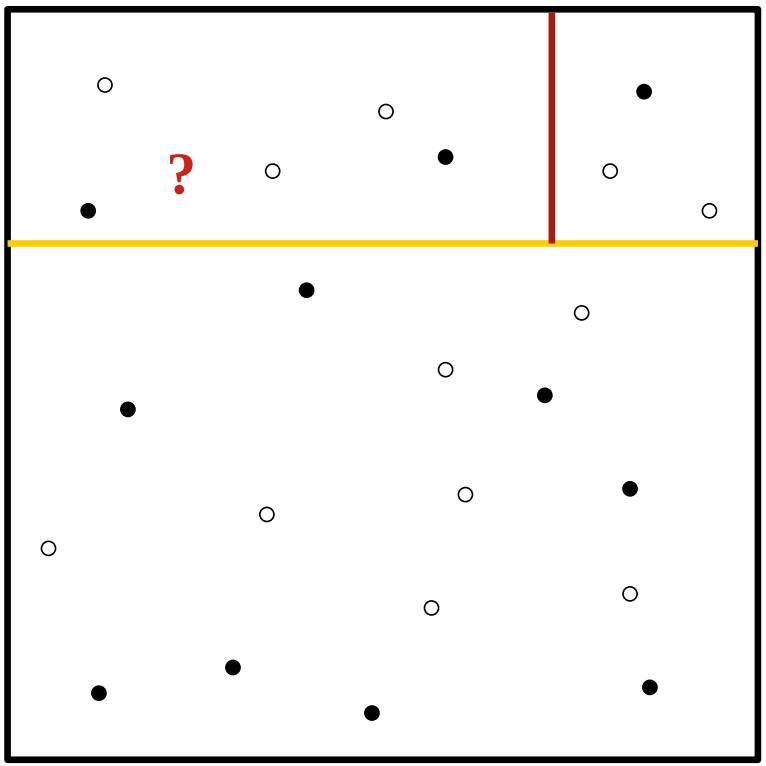
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

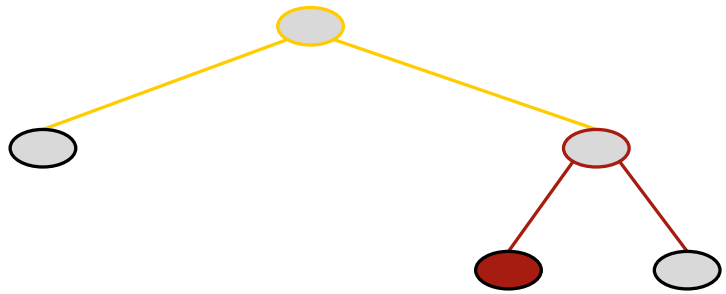




Exemple en classification

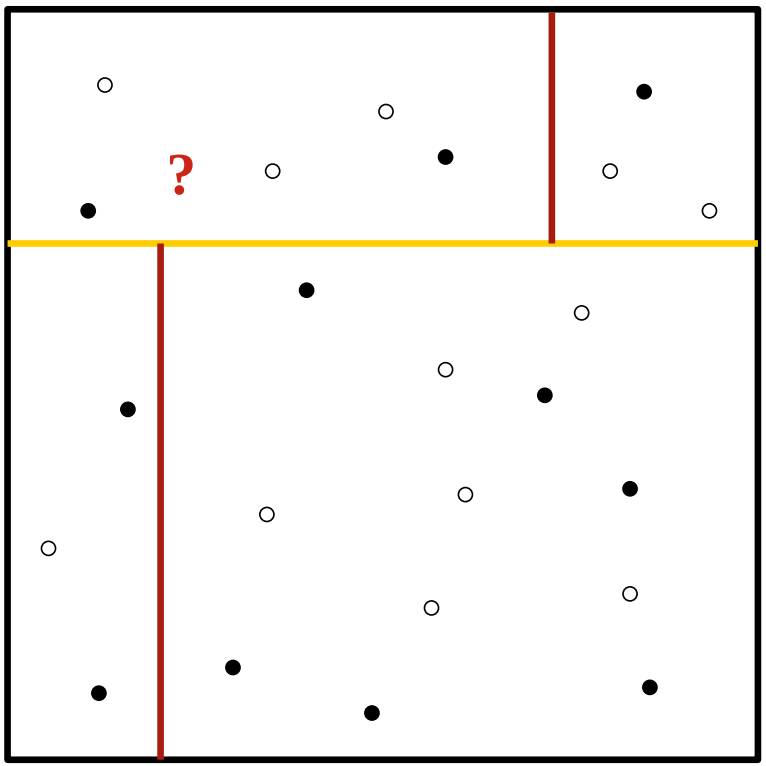
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

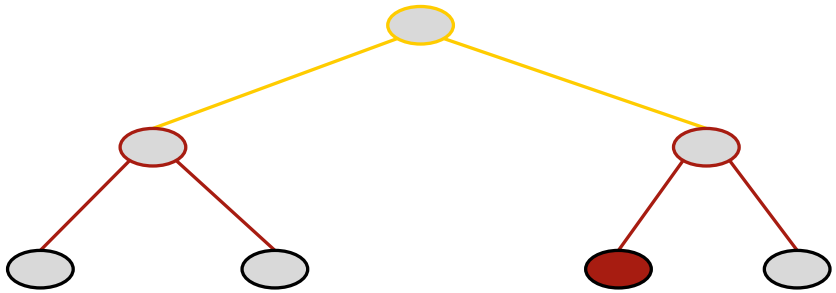




Exemple en classification

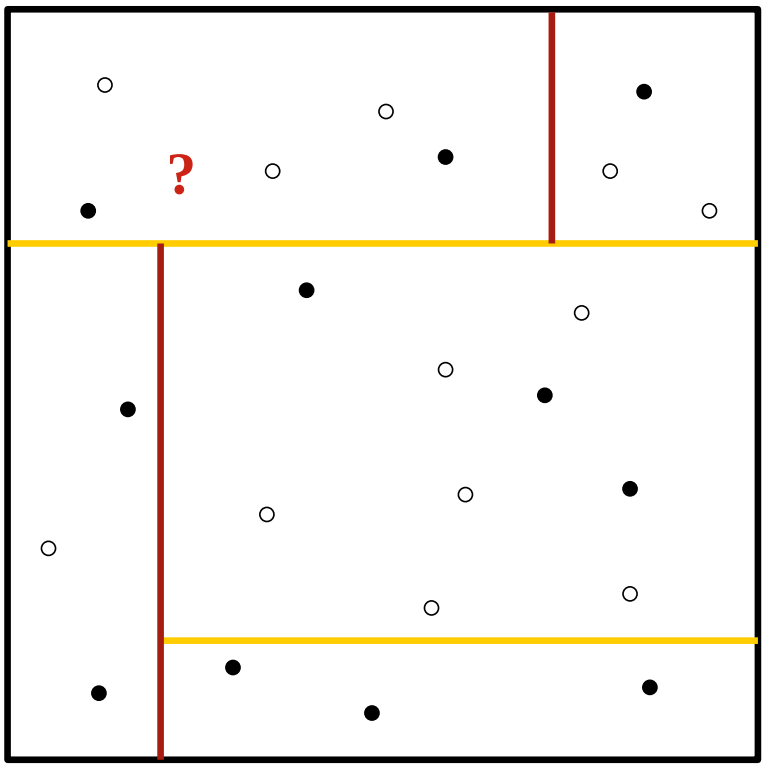
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

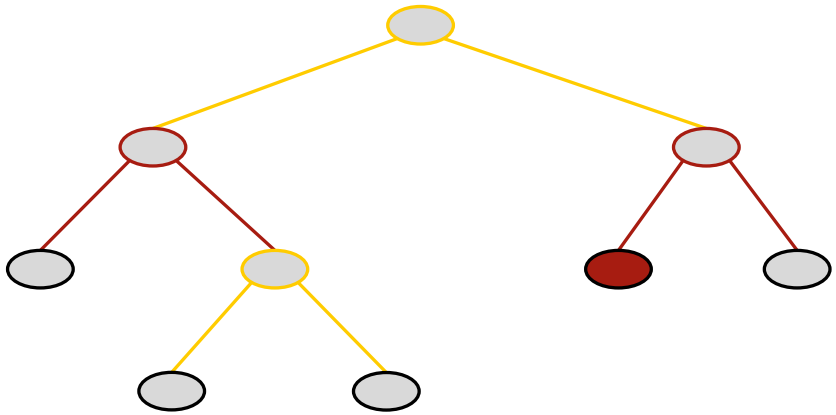




Exemple en classification

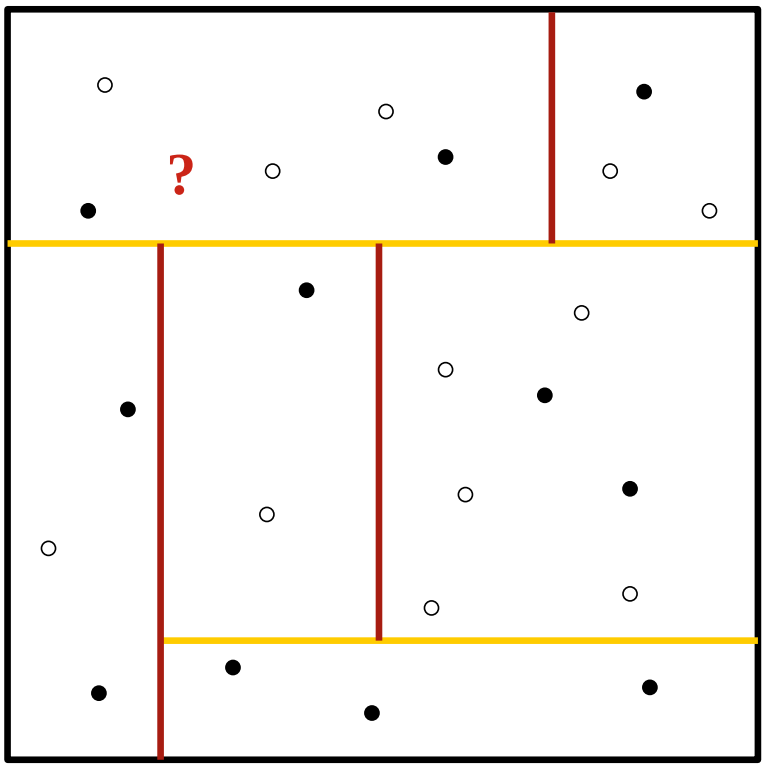
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

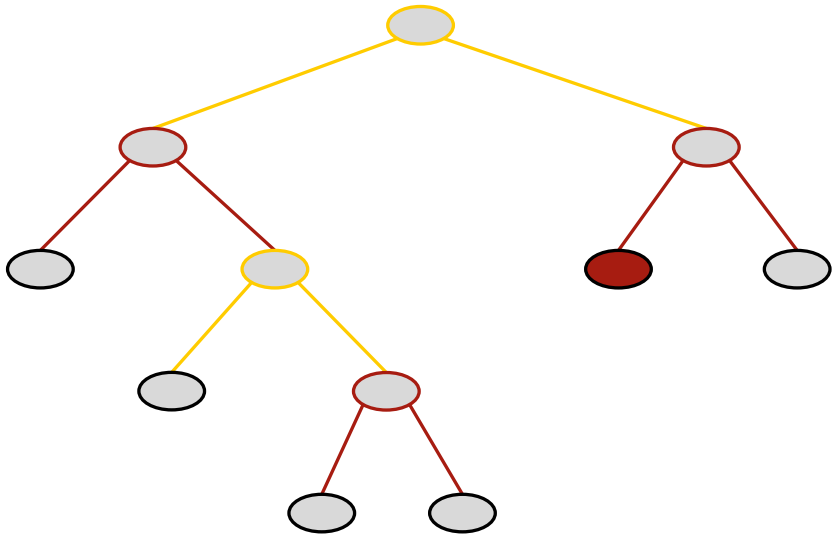




Exemple en classification

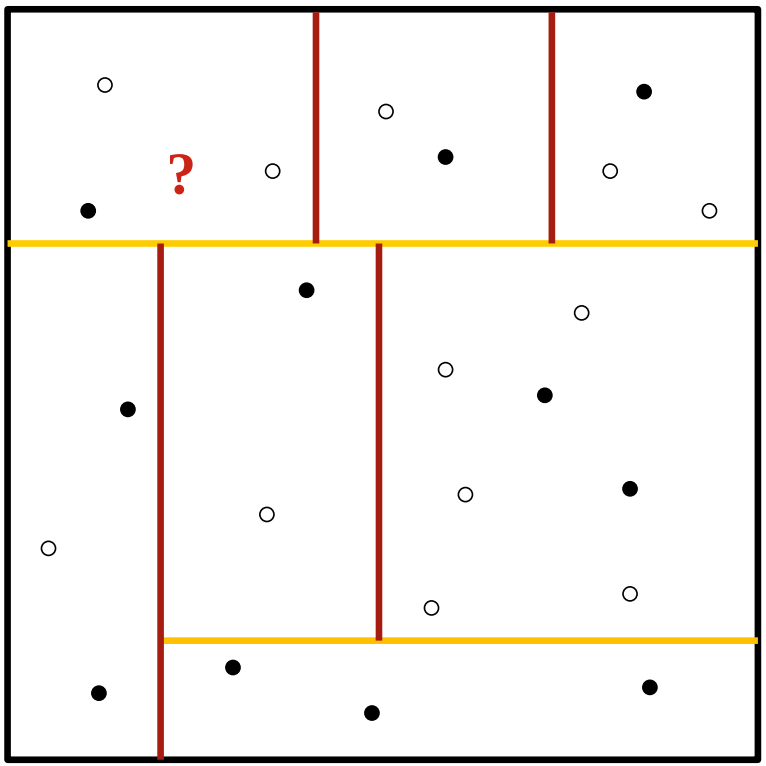
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

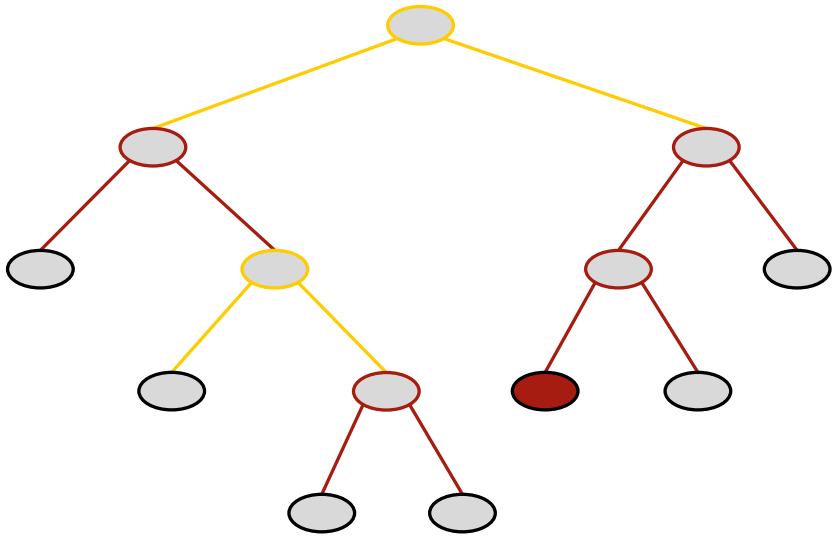




Exemple en classification

On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.





CROISSANCE DE L'ARBRE

- Choix de la variable X_j à utiliser pour la découpe du noeud.
- Choix de coupure k pour la variable : $\{X_j \leq k\} \cup \{X_j > k\}$.
- A chaque noeud on teste toutes les variables et toutes les coupures possibles.
- Choix de la variable et de la coupure qui minimisent un certain critère.

Critère en régression

- Minimiser la variance empirique des noeuds fils A et B

$$\mathbb{V}_A = \frac{1}{n_A} \sum_{i \in A} (Y_i - \bar{Y}_A)^2.$$

- A chaque noeud on minimise

$$\frac{n_A}{n} \times \mathbb{V}_A + \frac{n_B}{n} \times \mathbb{V}_B.$$

Critère en classification

- On cherche à augmenter l'homogénéité des groupes.
- Minimiser l'indice de Gini (mesure d'impureté) des noeuds fils

$$G_A = \sum_{c=1}^{L} \hat{p}_A^c (1 - \hat{p}_A^c).$$

- Autres critères existent.

ELAGAGE

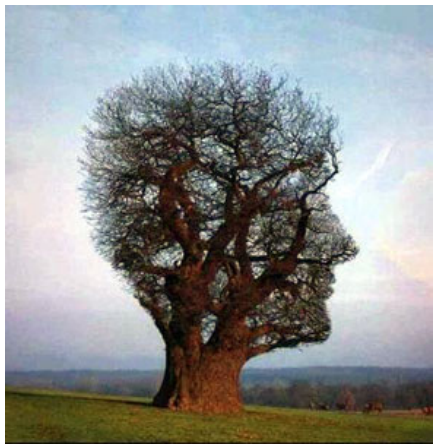
- Arbre développé jusqu'à atteindre une règle d'arrêt.

  $\rightarrow$ Grande variance et biais faible de l'arbre maximal.

- Elagage : chercher le meilleur sous-arbre de l'arbre maximal.
- Compromis entre erreur de prédiction et nombre de feuilles.

  $\rightarrow$ critère pénalisé : $R_\alpha(T) = R(T) + \alpha \tilde{T}$.

- Pas d'élagage avec les forets.


- Avantages de CART
  - Méthode non paramétrique.
  - Classification et régression.
  - Variable explicatives : qualitatives et quantitatives.
  - Coût numérique faible en regard des performances.
  - Interprétation.
  - Problèmes complexes, données de grande dimension.

- Limites de CART
  - Structure d'arbre (optimum local), découpe binaire.
  - Instable : supprimer 1 observation peut fortement modifier l'arbre.

    $\Rightarrow$ peu robuste aux observations erronées ou atypiques.

FORÊTS ALÉATOIRES

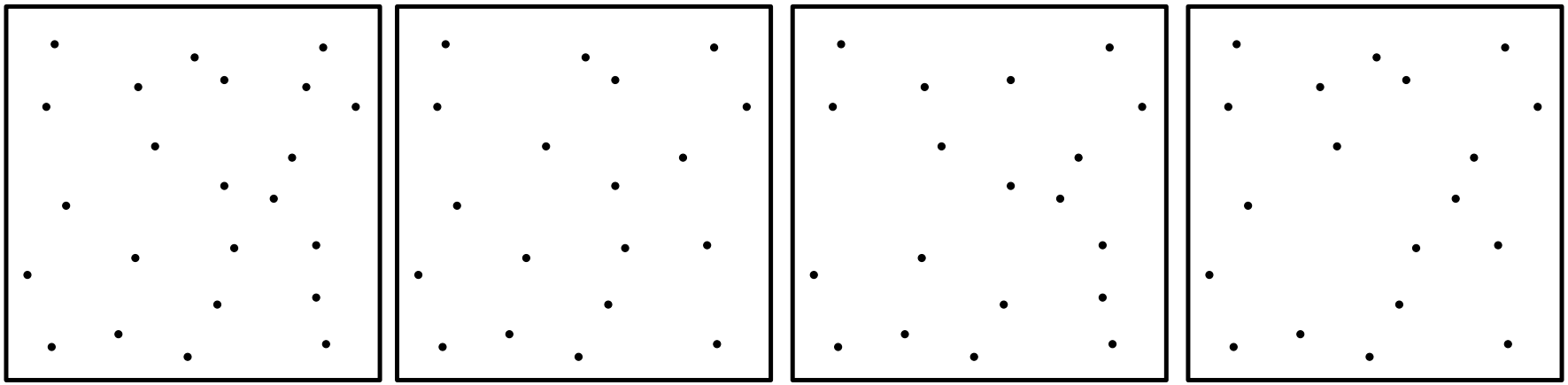
- Introduite par Breiman (2001).
- Idée : faire la moyenne de plusieurs arbres afin d'obtenir des classifieurs plus performants.
- Les arbres sont simples (pas optimisés) et randomisés (différents).

ARBRES RANDOMISÉS

- Inspirée de la méthode random subspace (Ho, 1998)
- Randomisation dans la construction de l'arbre.
  - A chaque noeud, sélection aléatoire de m_{try} variables parmi les p.
  - Choix de la meilleure découpe en minimisant le critère CART parmi les m_{try} variables.
  - L'arbre est pleinement développé.

BAGGING

- Méthode d'ensemble (Breiman, 1996).
- Plusieurs échantillons bootstrap (avec remise).
- 3 qualités
  - Aléa supplémentaire (diversité).
  - Moins de données à traiter (en préservant la distribution).
  - Données Out-of-bag.



AGRÉGATION

- Construction d'un ensemble d'arbres de type CART.
- Prédictions obtenues par agrégation des prédictions individuelles.
- Agrégation rapide et facilement parallélisable.




ERREUR OUT-OF-BAG

$\mathscr{D}_n$ données d'apprentissage,

$\mathscr{D}_n^m, m = 1, \ldots, M$ échantillons bootstrap,

$\bar{\mathscr{D}}_n^m = \mathscr{D}_n \setminus \mathscr{D}_n^m$ est l'échantillon out-of-bag.

Estimation de l'erreur de prédiction d'un arbre avec l'échantillon OOB

$$\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^m) = \frac{1}{|\bar{\mathscr{D}}_n^m|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathscr{D}}_n^m} (Y_i - \hat{f}_m(\mathbf{X}_i))^2$$

$\hat{f}_m$ estimateur de l'arbre m.

$\rightarrow$ Permet d'estimer l'erreur sur des données qui n'ont pas servi à construire l'arbre (proche de la validation croisée).

IMPORTANCE PAR PERMUTATION

- Importance par permutation (Breiman, 2001).
- Plusieurs mesures d'importance.
- Idée : augmentation de l'erreur en cassant le lien entre X_j et Y.

MESURE D'IMPORTANCE EMPIRIQUE

$$\hat{I}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{mj}) - \hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{m}) \right]$$

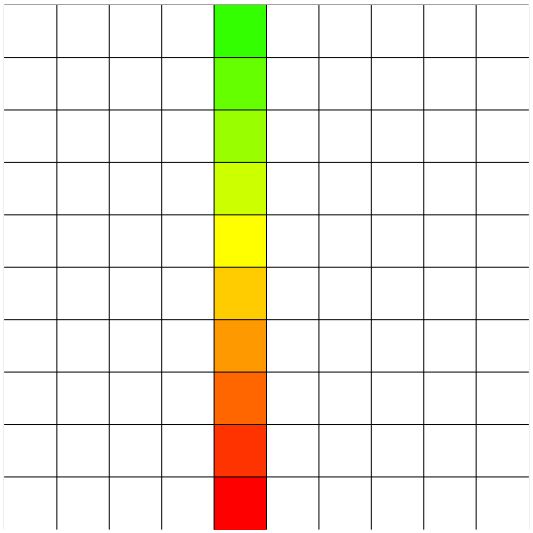
$\hat{R}$ est le risque empirique,

$\bar{\mathcal{D}}_n^m$ échantillon out-of-bag.

IMPORTANCE PAR PERMUTATION

MESURE D'IMPORTANCE EMPIRIQUE

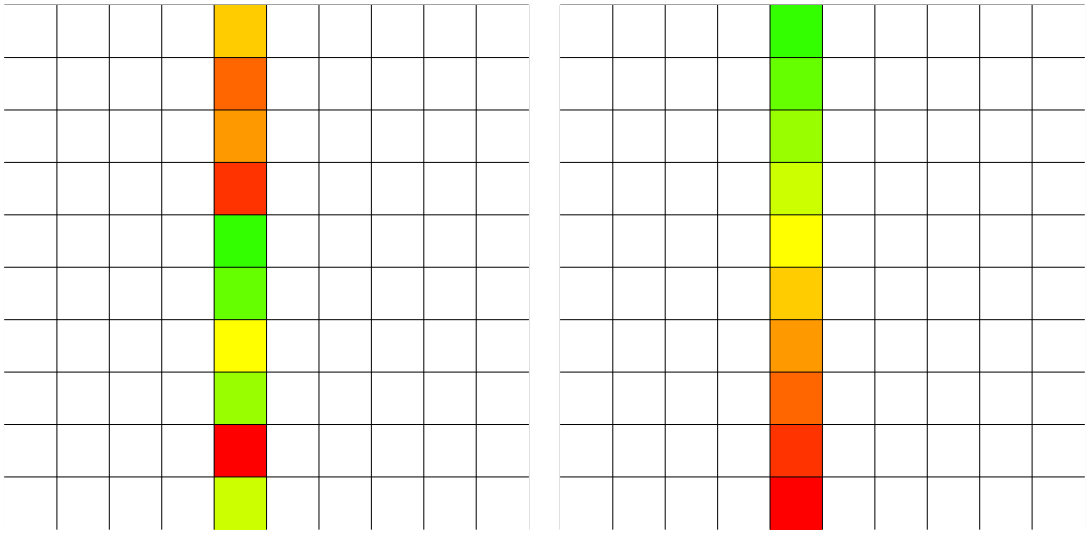
$$\hat{I}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mj}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m}) \right]$$



IMPORTANCE PAR PERMUTATION

MESURE D'IMPORTANCE EMPIRIQUE

$$\hat{I}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mj}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m}) \right]$$



IMPORTANCE PAR PERMUTATION

MESURE D'IMPORTANCE EMPIRIQUE

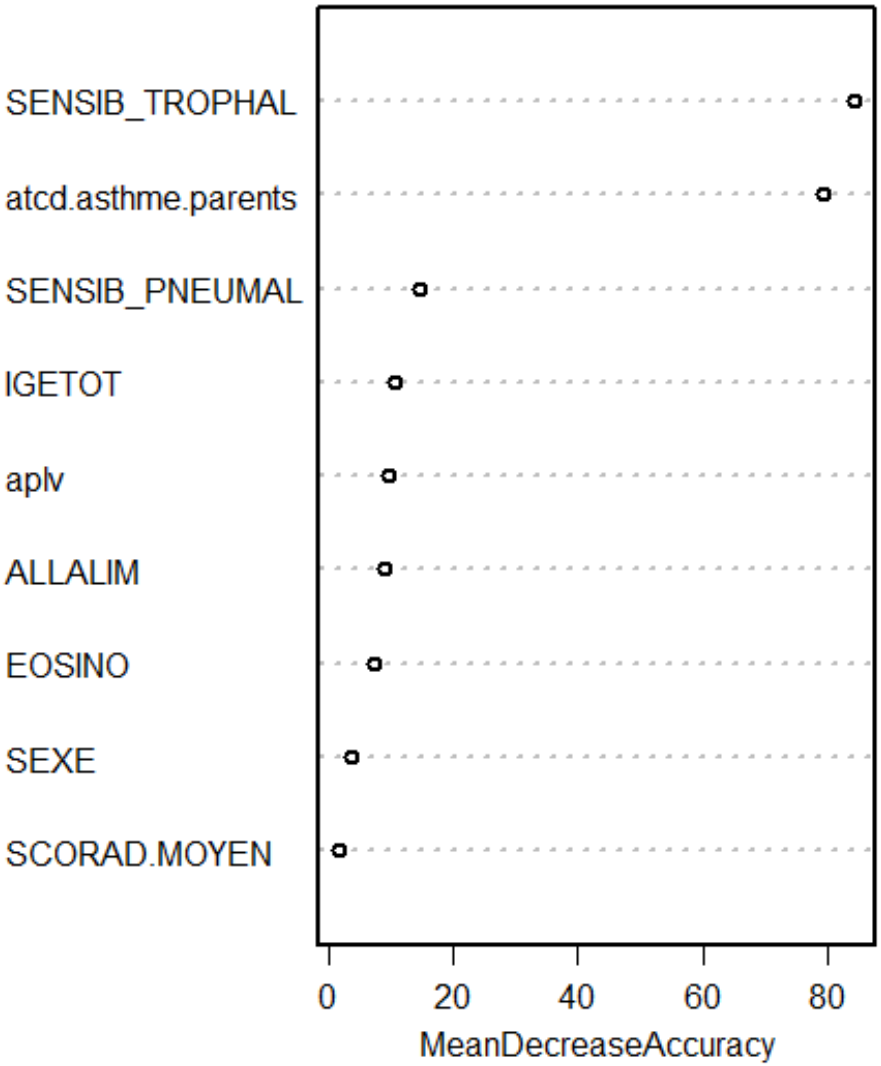
$$\hat{I}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mj}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m}) \right]$$

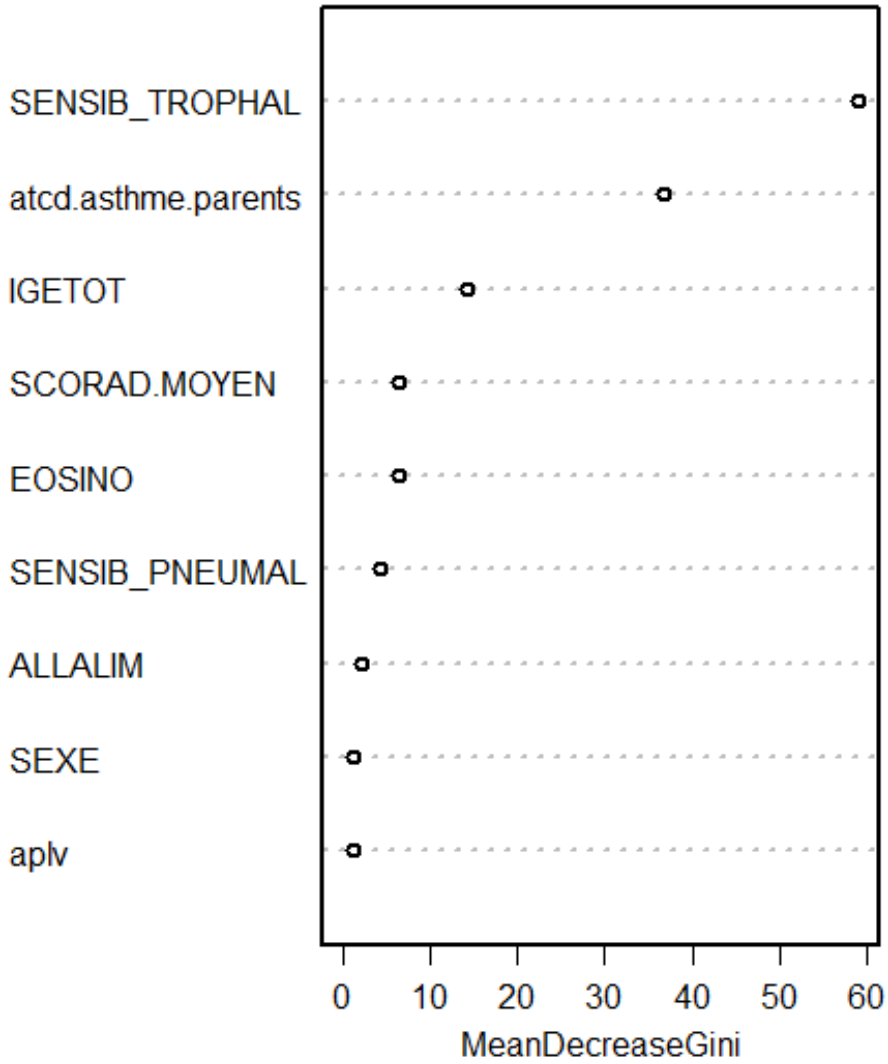
MESURE D'IMPORTANCE THÉORIQUE

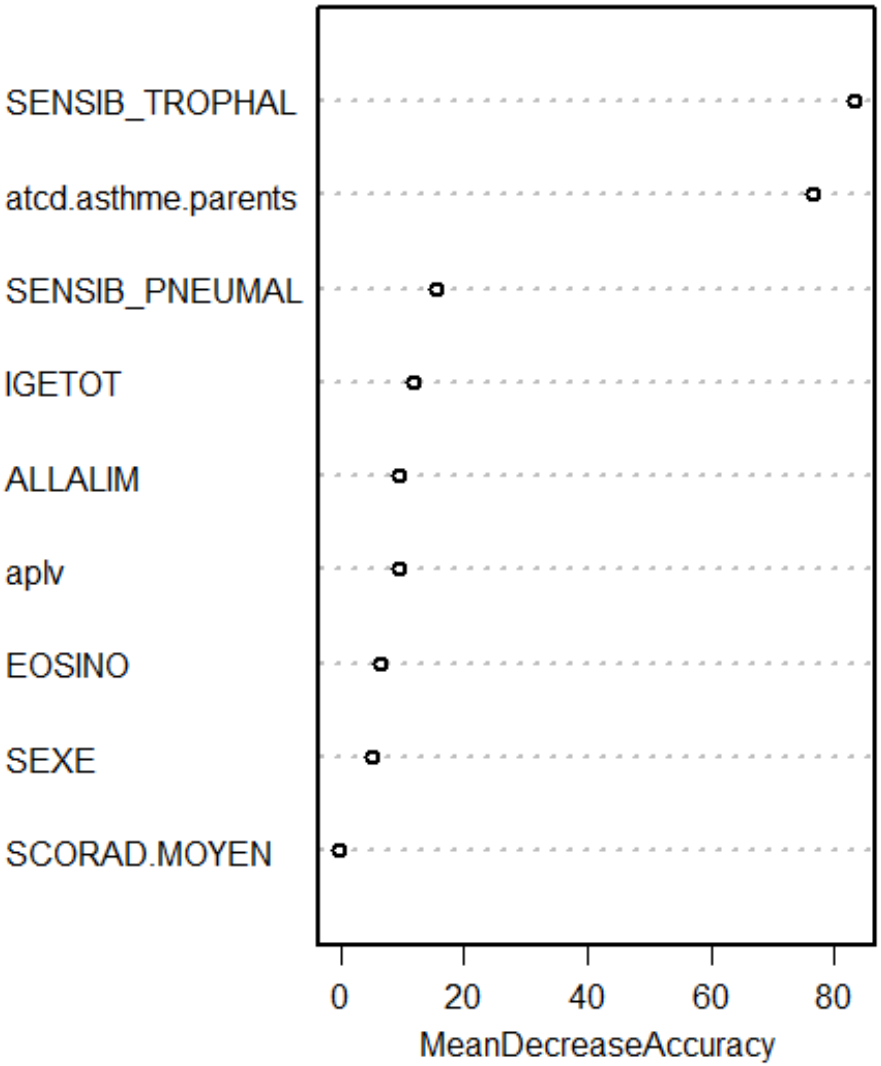
$$I(X_j) = \mathbb{E}\left[\left(Y - f(\mathbf{X}_{(j)}) \right)^2 \right] - \mathbb{E}\left[(Y - f(\mathbf{X}))^2 \right]$$

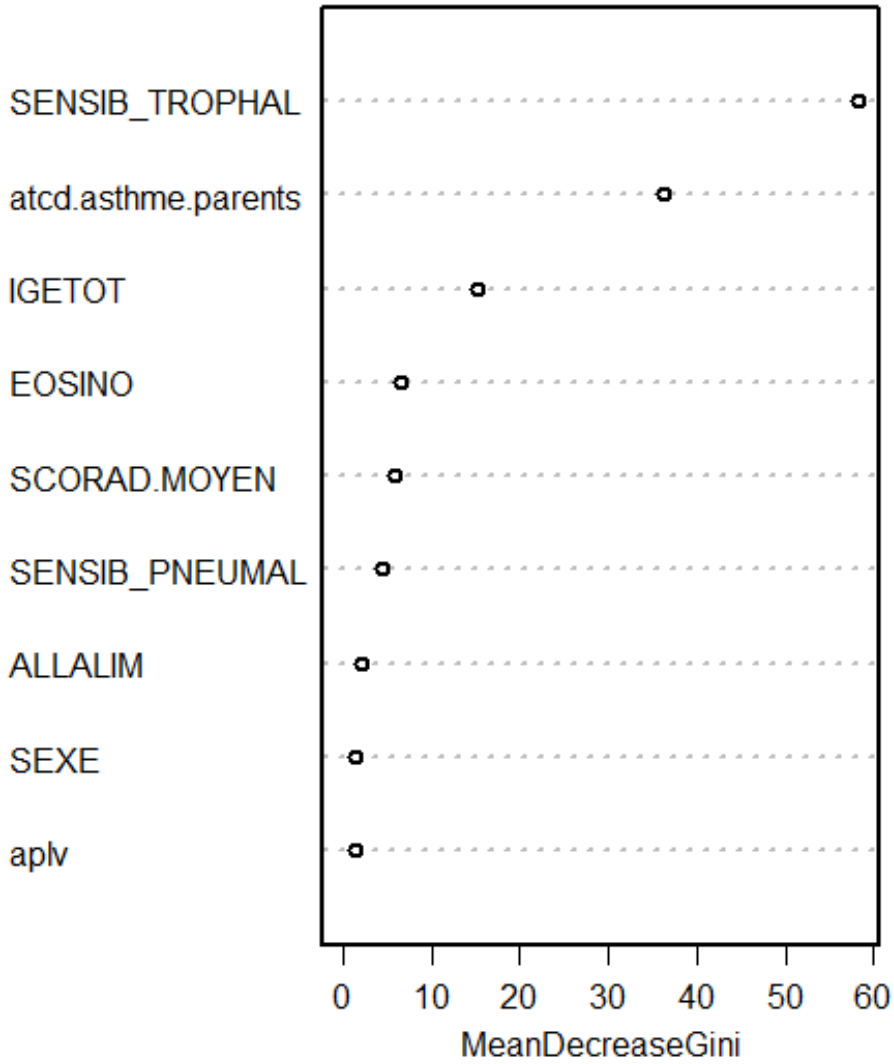
$\mathbf{X}_{(j)} = (X_1, \ldots, X'_j, \ldots, X_p)$,

X'_j est une réplication indépendante de X_j .


SENSIB_TROPHAL
atcd.asthme.parents
SENSIB_PNEUMAL
IGETOT
aplv
ALLALIM
EOSINO
SEXE
SCORAD.MOYEN
0
20
40
60
80
MeanDecreaseAccuracy

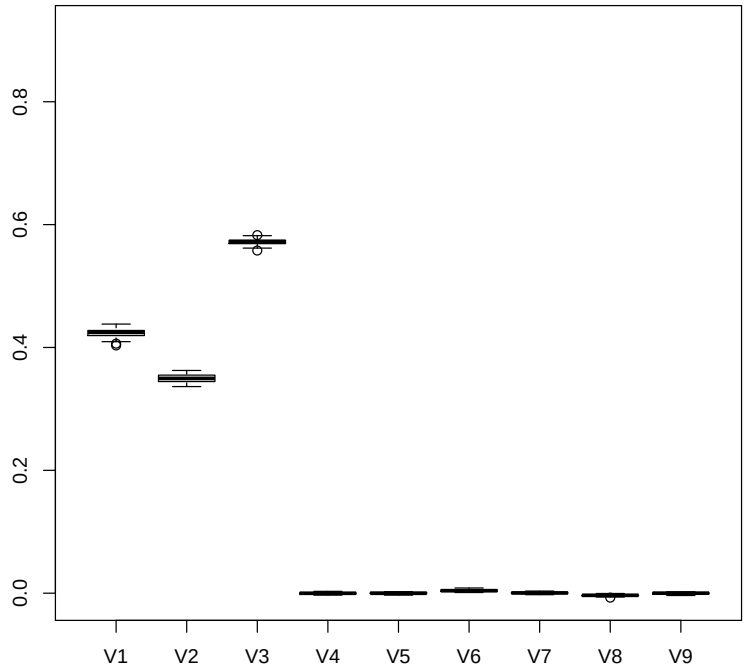

SENSIB_TROPHAL
atcd.asthme.parents
IGETOT
SCORAD.MOYEN
EOSINO
SENSIB_PNEUMAL
ALLALIM
SEXE
aplv
0
10
20
30
40
50
60
MeanDecreaseGini


SENSIB_TROPHAL
atcd.asthme.parents
SENSIB_PNEUMAL
IGETOT
ALLALIM
aplv
EOSINO
SEXE
SCORAD.MOYEN
0
20
40
60
80
MeanDecreaseAccuracy


SENSIB_TROPHAL
atcd.asthme.parents
IGETOT
EOSINO
SCORAD.MOYEN
SENSIB_PNEUMAL
ALLALIM
SEXE
aplv
0
10
20
30
40
50
60
MeanDecreaseGini

CORRÉLATION

- Variables les plus pertinentes $\not\Leftrightarrow$ importances les plus grandes.
- L'importance dépend de
  - la corrélation entre les variables,
  - du nombre de variables corrélées.



RÉSULTATS

Modèle de régression additif

$$Y = \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \tag{1}$$

PROPOSITION

- On montre que
$$I(X_j) = 2\mathbb{V}[f_j(X_j)].$$
- Si $\mathbb{E}[f_j(X_j)] = 0$ alors
$$I(X_j) = 2\mathbb{C}[f_j(X_j), Y] - 2\sum_{k \neq j} \mathbb{C}[f_j(X_j), f_k(X_k)].$$

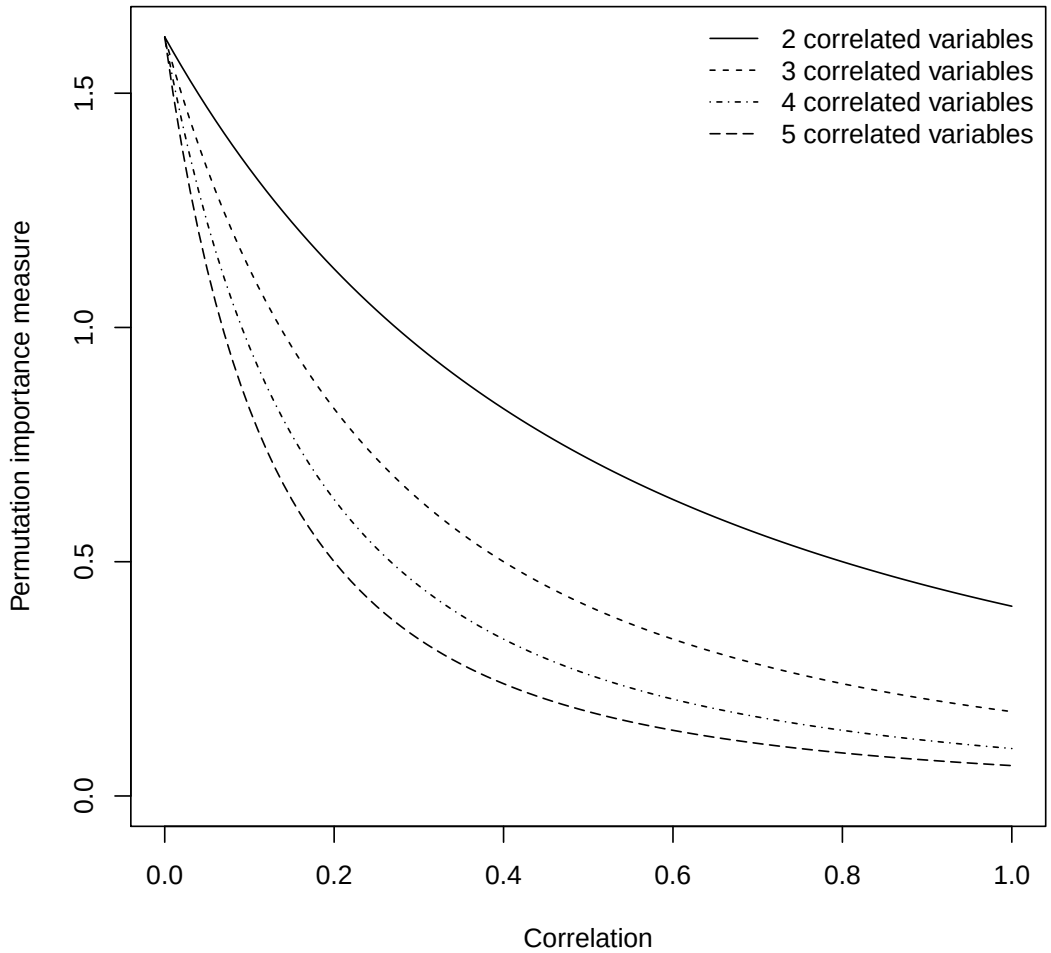
CAS GAUSSIEN

$$(\mathbf{X}, Y) \sim \mathcal{N}_{p+1}\left(0, \begin{pmatrix} C & \tau \\ \tau^t & \sigma_y^2 \end{pmatrix}\right) \text{ et } Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon.$$

PROPOSITION

Si $C = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$ et $\tau = (\tau_0, \ldots, \tau_0)^t \in \mathbb{R}^p$ alors

$$I(X_u) = 2\left(\frac{\tau_0}{1-\rho+p\rho}\right)^2.$$


2 correlated variables
3 correlated variables
4 correlated variables
5 correlated variables
Permutation importance measure
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
Correlation

SÉLECTION DE VARIABLES

- Sélection de variables : parcimonie, prédiction, ...
- On ne peut pas évaluer les performances de l'ensemble des sous-modèles.
- Utilisation d'algorithmes de sélection basés sur l'importance des variables.

ALGORITHME NON RÉCURSIF

1. Ranger les variables en fonction de l'importance.
2. Construire une forêt et calculer l'erreur.
3. Éliminer la variable la moins importante (backward).
4. Répéter 2-3 tant qu'il reste des variables disponibles.

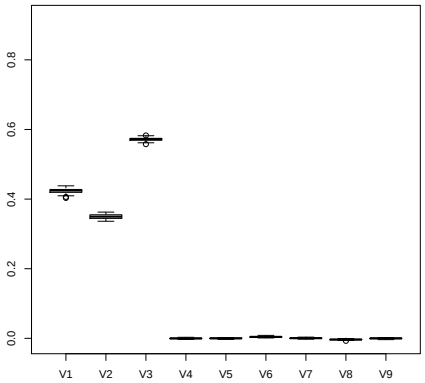
ALGORITHME RFE

- Algorithme récursif pour corriger le biais de corrélation.
- Algorithme adapté de SVM-RFE (Guyon et al., 2002).

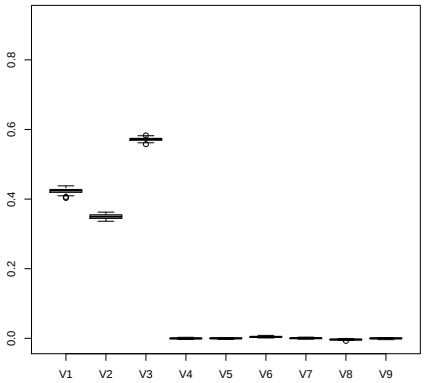
RECURSIVE FEATURE ELIMINATION (RFE)

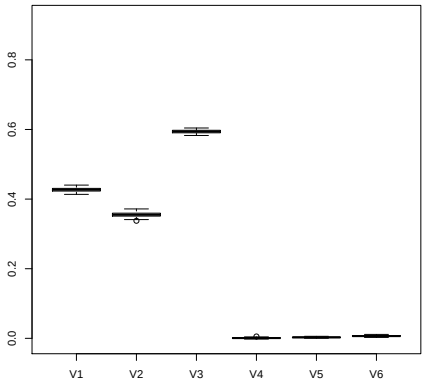
1. Construire une forêt et calculer l'erreur.
2. Calculer la mesure d'importance pour chaque variable.
3. Éliminer la variable la moins importante.
4. Répéter 1-3 tant qu'il reste des variables disponibles.

ALGORITHME RFE

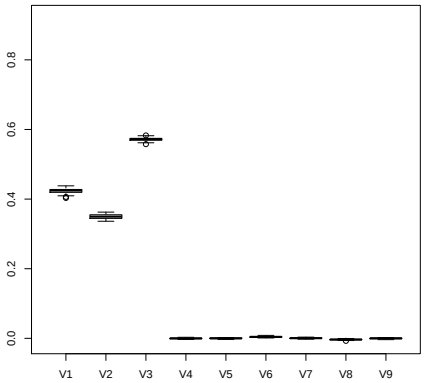


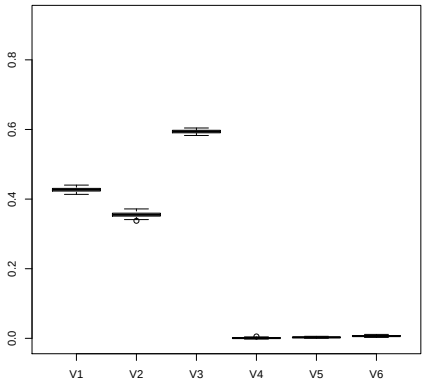
ALGORITHME RFE

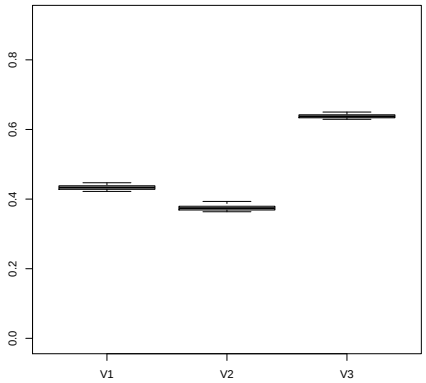




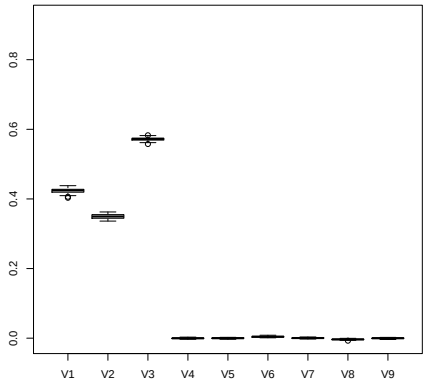
ALGORITHME RFE

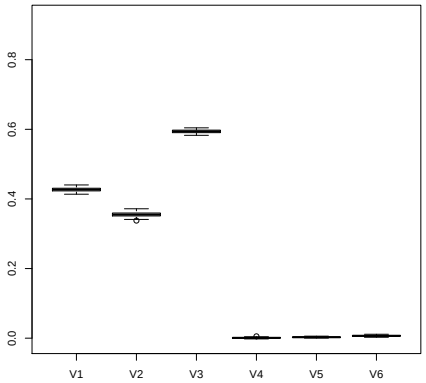


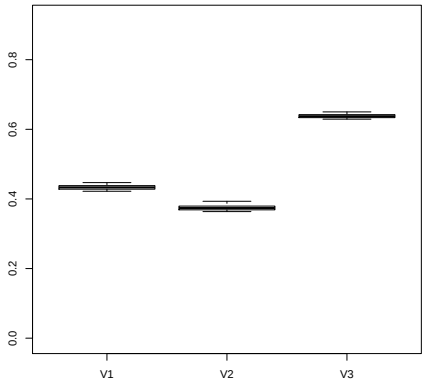


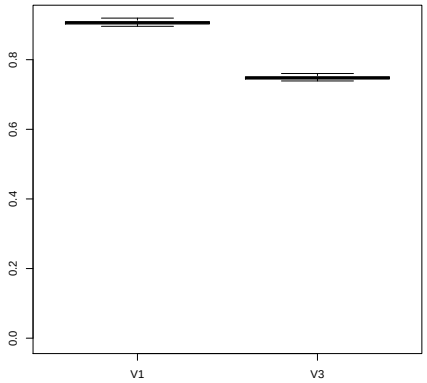


ALGORITHME RFE









MESURE D'IMPORTANCE GROUPÉE

Changement de contexte :

- Supposons que $\{1,\ldots,p\}$ est formé de K groupes
- Sélection de variables : trouver les groupes de variables les plus prédictifs

IMPORTANCE GROUPÉE EMPIRIQUE

Soit $J = (j_1,\ldots,j_k) \subset \{1,\ldots,p\}$ et $\mathbf{X}_J = (X_{j_1},\ldots,X_{j_k})$

$$\hat{I}(\mathbf{X}_J) = \frac{1}{M}\sum_{m=1}^{M}\left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mJ}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m})\right],$$

$\hat{R}$ = risque empirique,

$\bar{\mathscr{D}}_n^m = \mathscr{D}_n \setminus \mathscr{D}_n^m$ (échantillon out-of-bag).

MESURE D'IMPORTANCE GROUPÉE

Changement de contexte :

- Supposons que $\{1,\ldots,p\}$ est formé de K groupes
- Sélection de variables : trouver les groupes de variables les plus prédictifs

IMPORTANCE GROUPÉE EMPIRIQUE

Soit $J = (j_1,\ldots,j_k) \subset \{1,\ldots,p\}$ et $\mathbf{X}_J = (X_{j_1},\ldots,X_{j_k})$

$$\hat{I}(\mathbf{X}_J) = \frac{1}{M}\sum_{m=1}^{M}\left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mJ}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m})\right],$$

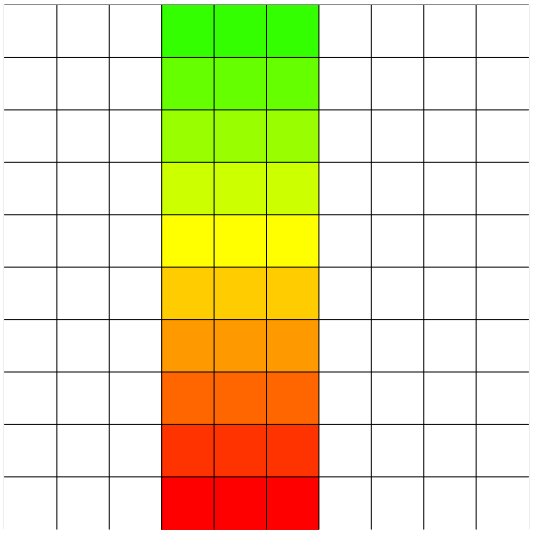
$\hat{R}$ = risque empirique,

$\bar{\mathscr{D}}_n^m = \mathscr{D}_n \setminus \mathscr{D}_n^m$ (échantillon out-of-bag).

MESURE D'IMPORTANCE GROUPÉE

IMPORTANCE GROUPÉE EMPIRIQUE

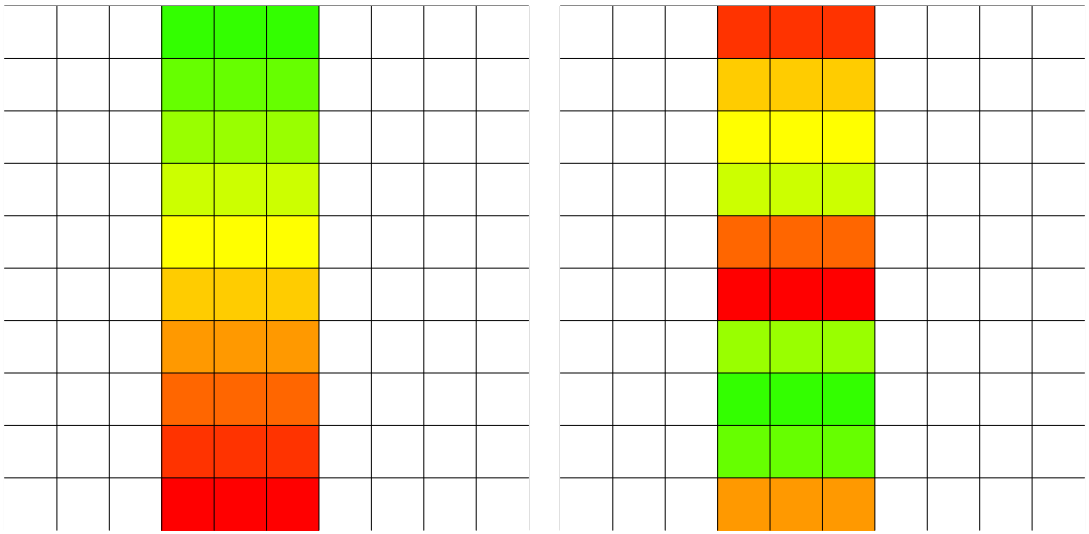
$$\hat{I}(\mathbf{X}_J) = \frac{1}{M}\sum_{m=1}^{M}\left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mJ}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m})\right]$$



MESURE D'IMPORTANCE GROUPÉE

IMPORTANCE GROUPÉE EMPIRIQUE

$$\hat{I}(\mathbf{X}_J) = \frac{1}{M}\sum_{m=1}^{M}\left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mJ}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m})\right]$$



MESURE D'IMPORTANCE GROUPÉE

IMPORTANCE GROUPÉE EMPIRIQUE

$$\hat{I}(\mathbf{X}_J) = \frac{1}{M}\sum_{m=1}^{M}\left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mJ}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m})\right]$$

IMPORTANCE GROUPÉE THÉORIQUE

$$I(\mathbf{X}_J) = \mathbb{E}\left[Y - f(\mathbf{X}_{(J)})\right]^2 - \mathbb{E}\left[Y - f(\mathbf{X})\right]^2$$

$\mathbf{X}_{(J)} = (\mathbf{X}'_J, \mathbf{X}_{\bar{J}})$,

$\mathbf{X}'_J$ est une copie indépendante du vecteur $\mathbf{X}_J$.

RÉSULTATS

$$Y = f_J(\mathbf{X}_J) + f_{\bar{J}}(\mathbf{X}_{\bar{J}}) + \varepsilon$$

PROPOSITION

$$I(\mathbf{X}_J) = 2\mathbb{V}(f_J(\mathbf{X}_J))$$

COROLLAIRE

1. Si $f_J(\mathbf{x}_J) = \sum_{j\in J} f_j(x_j)$ et si $(X_j)_{j\in J}$ sont indépendantes, alors

$$I(\mathbf{X}_J) = 2\sum_{j\in J} \mathbb{V}\left(f_j(X_j)\right) = \sum_{j\in J} I\left(X_j\right)$$

2. Si $f_J(\mathbf{x}_J) = \sum_{j\in J} \alpha_j x_j$, alors

$$I(\mathbf{X}_J) = 2\alpha_J^\top \mathbb{C}(\mathbf{X}_J)\alpha_J,\ \alpha_J = (\alpha_j)_{j\in J}$$

MESURE D'IMPORTANCE GROUPÉE

Points importants :

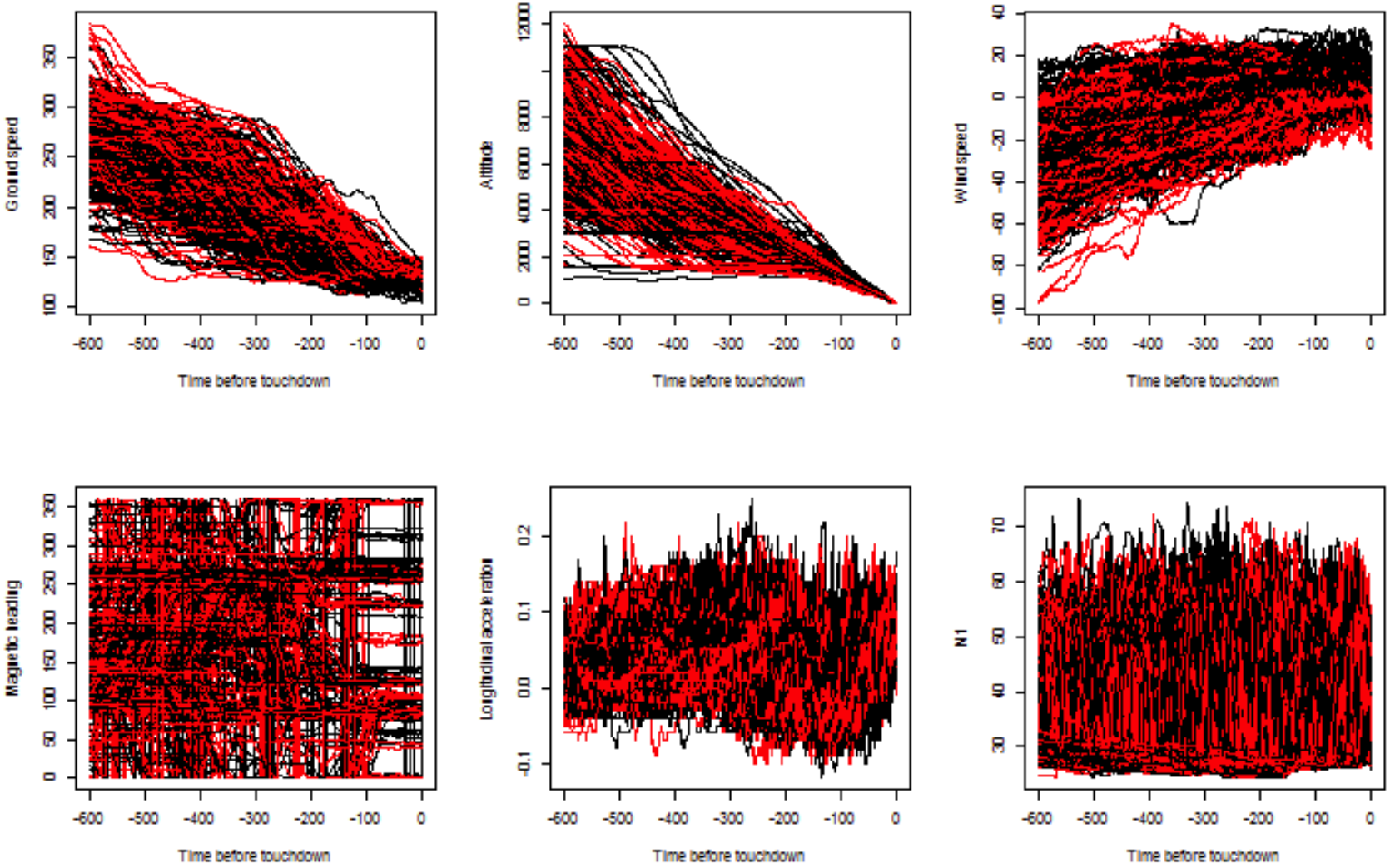
- Structure additive de f + indépendance :

$$I(\mathbf{X}_J) = \sum_{j \in J} I(X_j).$$

- Résultat plus valable en cas de corrélation.

En toute généralité, il est difficile d'identifier l'importance groupée à la somme des importances individuelles

ANALYSE DE DONNÉES FONCTIONNELLES MULTIVARIÉES



ANALYSE DE DONNÉES FONCTIONNELLES MULTIVARIÉES

APPROCHE PAR PROJECTION

1. Réduire la dimension : projeter chaque variable fonctionnelle sur une base de fonctions
2. Utiliser les coefficients de base comme nouvelles variables explicatives

Base d'ondelettes :

$$\mathscr{B} = \{\phi_{00}, \psi_{00}, \psi_{10}, \psi_{11}, \psi_{20}, \psi_{21}, \psi_{22}, \psi_{23}, \ldots\}$$

Décomposition en ondelettes :

$$X_i^u(t) = \zeta_i^u \phi_{00}(t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \xi_{ijk}^u \psi_{jk}(t)$$

ANALYSE DE DONNÉES FONCTIONNELLES MULTIVARIÉES

APPROCHE PAR PROJECTION

1. Réduire la dimension : projeter chaque variable fonctionnelle sur une base de fonctions
2. Utiliser les coefficients de base comme nouvelles variables explicatives

Base d'ondelettes :

$$\mathscr{B} = \{\phi_{00}, \psi_{00}, \psi_{10}, \psi_{11}, \psi_{20}, \psi_{21}, \psi_{22}, \psi_{23}, \ldots\}$$

Décomposition en ondelettes :

$$X_i^u(t) = \zeta_i^u \phi_{00}(t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \xi_{ijk}^u \psi_{jk}(t)$$

ANALYSE DE DONNÉES FONCTIONNELLES MULTIVARIÉES

Procédures de sélection dans le cas fonctionnel :

1. Sélection des groupes formés par l'ensemble des coefficients de base
2. Sélection de niveaux de fréquence
3. Sélection d'intervalles de temps

Variable fonctionnelle 1 : $\zeta^1, \xi^1_{00}, \xi^1_{10}, \xi^1_{11}, \xi^1_{20}, \xi^1_{21}, \xi^1_{22}, \xi^1_{23}, \cdots \Rightarrow$ Groupe 1

Variable fonctionnelle 2 : $\zeta^2, \xi^2_{00}, \xi^2_{10}, \xi^2_{11}, \xi^2_{20}, \xi^2_{21}, \xi^2_{22}, \xi^2_{23}, \cdots \Rightarrow$ Groupe 2

Variable fonctionnelle 3 : $\zeta^3, \xi^3_{00}, \xi^3_{10}, \xi^3_{11}, \xi^3_{20}, \xi^3_{21}, \xi^3_{22}, \xi^3_{23}, \cdots \Rightarrow$ Groupe 3

...

ANALYSE DE DONNÉES FONCTIONNELLES MULTIVARIÉES

Procédures de sélection dans le cas fonctionnel :

1. Sélection des groupes formés par l'ensemble des coefficients de base
2. Sélection de niveaux de fréquence
3. Sélection d'intervalles de temps

Pour une variable fonctionnelle u,

Niveau de fréquence 1 : ζ^u $\Rightarrow$ Groupe 1

Niveau de fréquence 2 : ξ^u_{00} $\Rightarrow$ Groupe 2

Niveau de fréquence 3 : ξ^u_{10}, ξ^u_{11} $\Rightarrow$ Groupe 3

Niveau de fréquence 4 : $\xi^u_{20}, \xi^u_{21}, \xi^u_{22}, \xi^u_{23}$ $\Rightarrow$ Groupe 4

...

ANALYSE DE DONNÉES FONCTIONNELLES MULTIVARIÉES

Procédures de sélection dans le cas fonctionnel :

1. Sélection des groupes formés par l'ensemble des coefficients de base
2. Sélection de niveaux de fréquence
3. Sélection d'intervalles de temps

Pour une variable fonctionnelle u et un intervalle $\mathscr{T}$,

$$\mathscr{S}(\mathscr{T}) = \{(j,k) : \psi_{jk}(t) \neq 0, \forall t \in \mathscr{T}\}$$

Intervalle $\mathscr{T}_1$: $\zeta^u, \{\xi^u_{jk} : (j,k) \in \mathscr{S}(\mathscr{T}_1)\}$ $\Rightarrow$ Groupe 1

Intervalle $\mathscr{T}_2$: $\zeta^u, \{\xi^u_{jk} : (j,k) \in \mathscr{S}(\mathscr{T}_2)\}$ $\Rightarrow$ Groupe 2

...

RÉDUCTION DE DIMENSION DE n PROCESSUS INDÉPENDANTS

Fixons $u \in \{1,\ldots,p\}$. Pour chaque courbe $i \in \{1,\ldots,n\}$:

$$X_i^u(t) = \zeta_i^u \phi_{00}(t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \xi_{ijk}^u \psi_{jk}(t)$$

Obtenir une représentation commune aux n observations

- Réduction simultanée de n processus indépendants
- Méthode de seuillage

RÉDUCTION DE DIMENSION DE n PROCESSUS INDÉPENDANTS

Fixons $u \in \{1,\ldots,p\}$. Pour chaque courbe $i \in \{1,\ldots,n\}$:

$$X_i^u(t) = \zeta_i^u \phi_{00}(t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \xi_{ijk}^u \psi_{jk}(t)$$

Obtenir une représentation commune aux n observations

- Réduction simultanée de n processus indépendants
- Méthode de seuillage

APPLICATION

- Package R "randomForest".
- Essentiellement 3 paramètres à choisir.
  - Choix des variables à chaque noeud mtry : $\sqrt{p}$ en classification et $\frac{p}{3}$ en régression.
  - Nombre d'arbres ntree.
  - Nombre minimum d'observations dans chaque feuille nodesize.

    $\Rightarrow$ Compromis entre la diversité des arbres, une bonne prédiction et le temps de calcul.
- Package R "RFgroove"
  - Importance groupée.
  - Sélection de variable pour des groupes de variables.

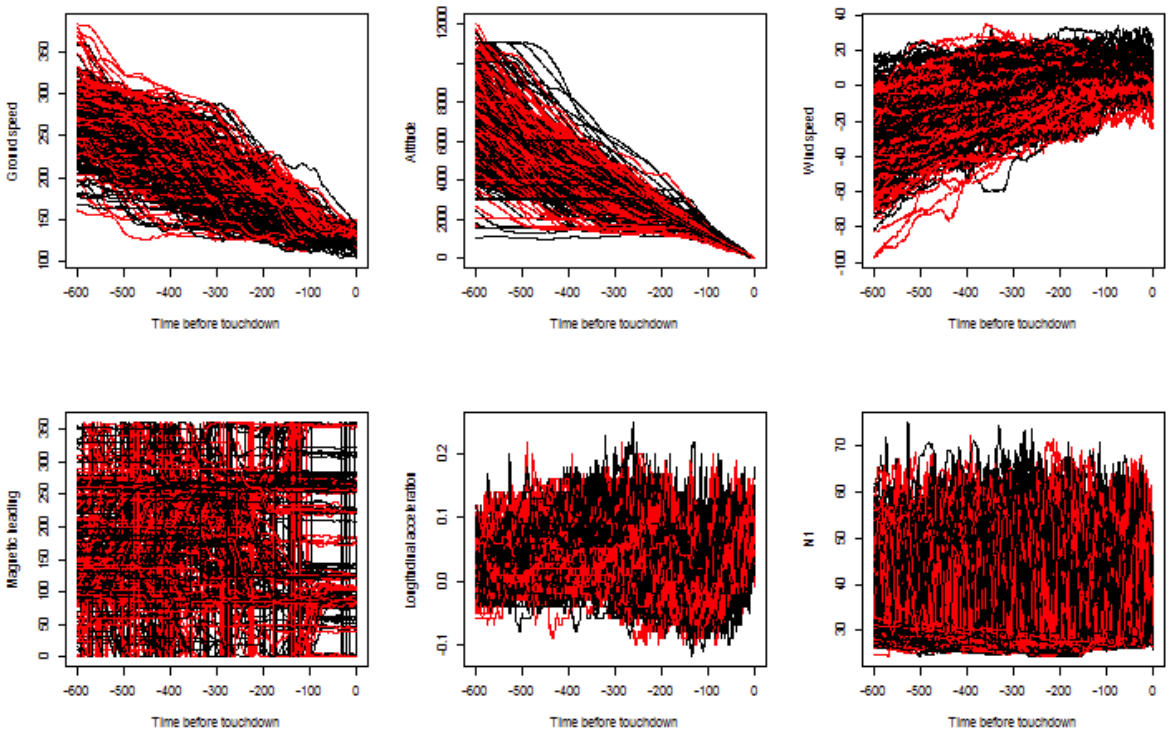
SÉCURITÉ AÉRIENNE (SAFETY LINE)

- Données issues de boîtes noires (enregistreurs de vols).

- Atterrissage long : dépasse 60% de la longueur de la piste.

ATTERRISSAGE LONG

- 254 vols, 37 variables conservées, 56% de classe 1.
- On s'intéresse aux 10 dernières minutes.



ATTERRISSAGE LONG

1. Réduction de la dimension en projetant sur une base d'ondelettes.

   Méthode de seuillage.

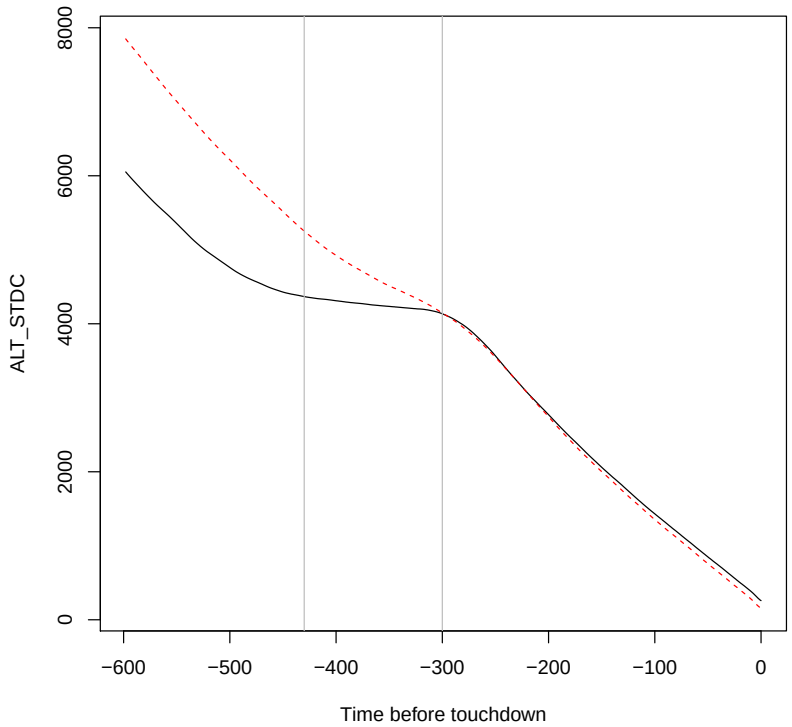
2. Selection de variables avec l'algorithme RFE.

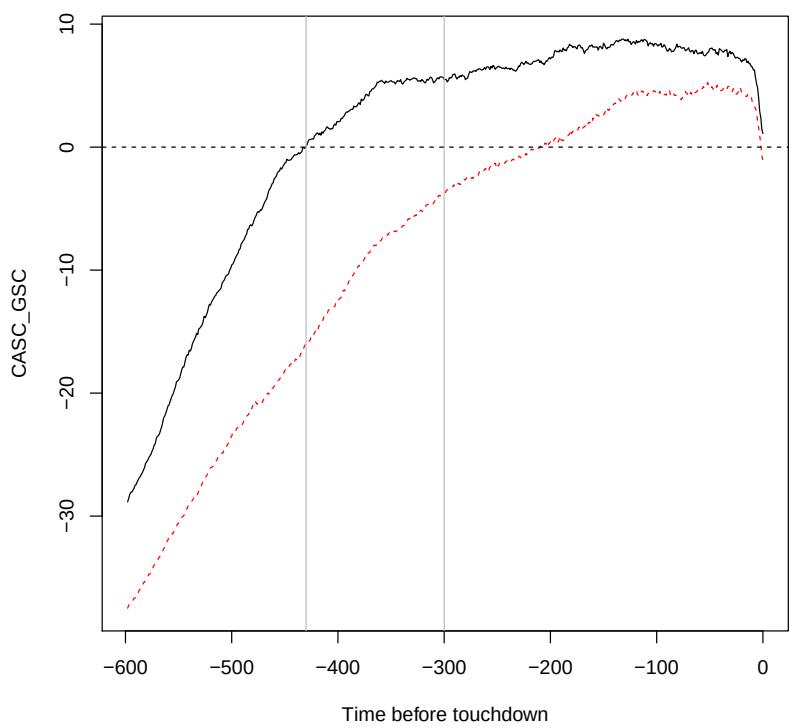
   $\rightarrow$ Altitude, vitesse du vent, masse de l'avion.

3. Estimation des probabilités a posteriori $\mathbb{P}(Y = 1 \mid X = x)$.

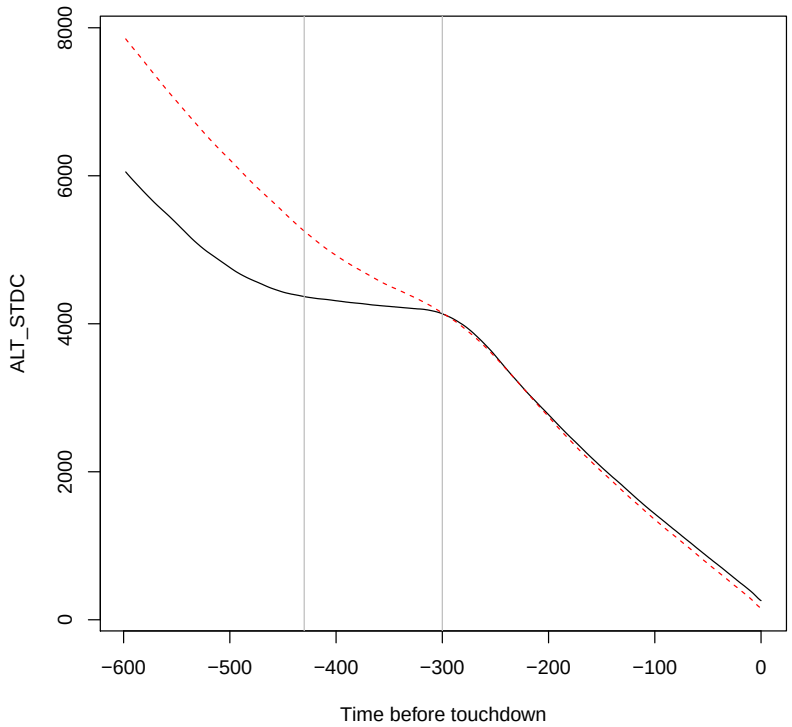
   $\rightarrow$ Profils types des vols à risque faible et élevé.

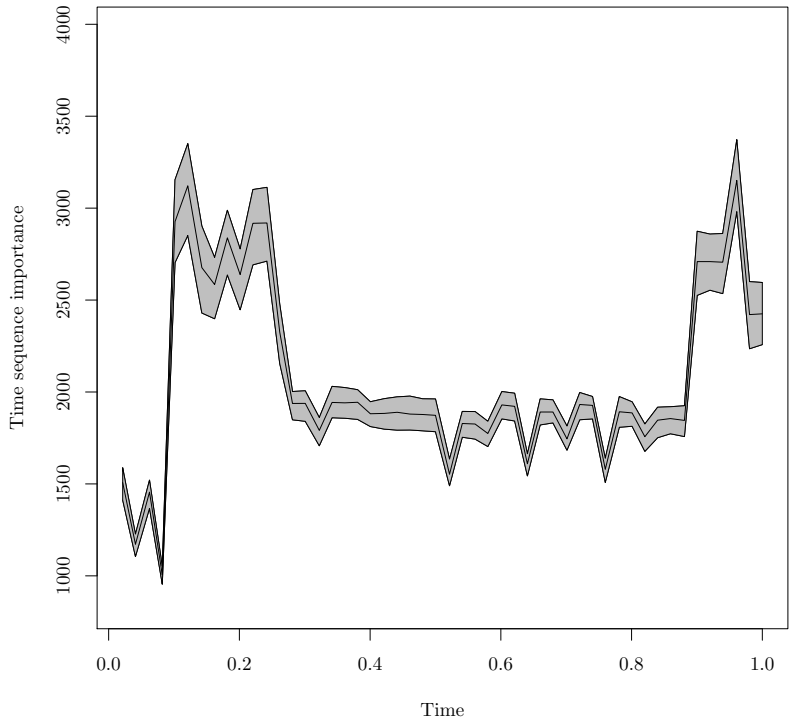
ATTERRISSAGE LONG





ATTERRISSAGE LONG





DISCUSSION

- Méthode non paramétrique (classification et régression).
- Algorithme rapide, facile à implémenter, performant dans des problèmes complexes (grande dimension, interactions, ...).
- Modèle de type "boîte noire" : interprétation délicate.
- Algorithme difficile à étudier d'un point de vue théorique.
- Sélection de variables basée sur l'importance.
- Effet de la corrélation sur l'importance.
- Importance groupée et analyse de données fonctionnelles.

Gregorutti, B., Michel, B. and Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multivariate functional data analysis. *Computational Statistics & Data Analysis*, 90 :15-35.

Gregorutti, B., Michel, B. and Saint-Pierre, P. (2015). Correlation and variable importance in random forests.