Forêts aléatoires et données de survie

Philippe Saint Pierre

IMT, Equipe de Statistique et Probabilités

Université Paul Sabatier - Toulouse III

Ecole CIMPA, Lomé, Septembre 2018

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 $i = 1, ..., n$.
- On observe une sortie Y_i pour chaque individu $i = 1, ..., n$.

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

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
- ...

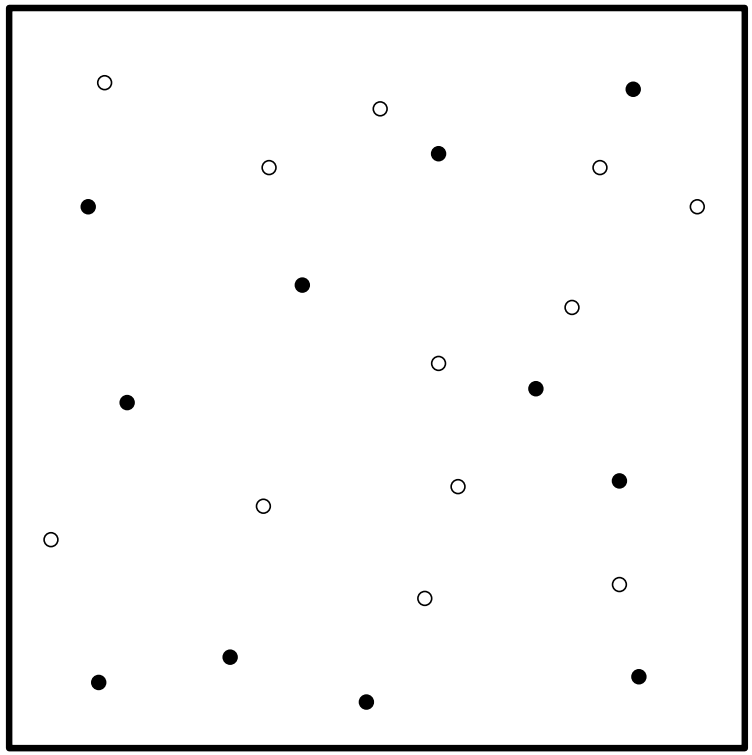
AUTRES CATÉGORIE DE MÉTHODES D'APPRENTISSAGE

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

CLASSIFICATION

Expliquer le statut malade ou non malade.

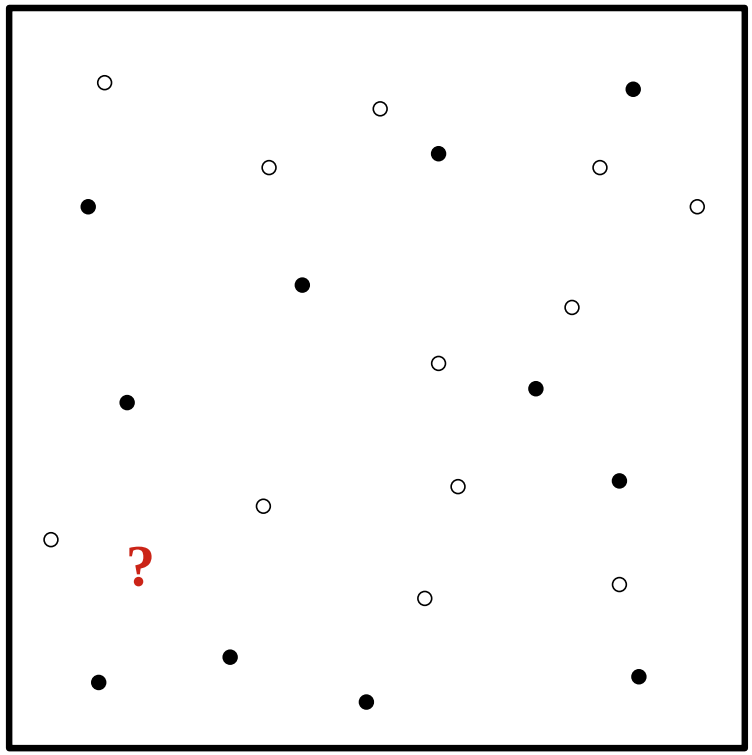
Prédire le statut d'un nouvel individu (malade - non malade).



CLASSIFICATION

Expliquer le statut malade ou non malade.

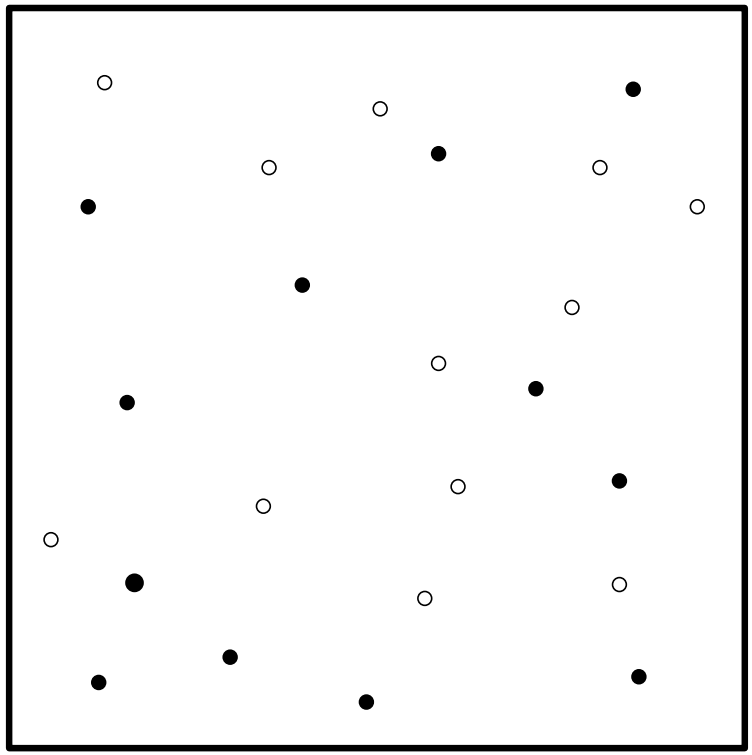
Prédire le statut d'un nouvel individu (malade - non malade).



CLASSIFICATION

Expliquer le statut malade ou non malade.

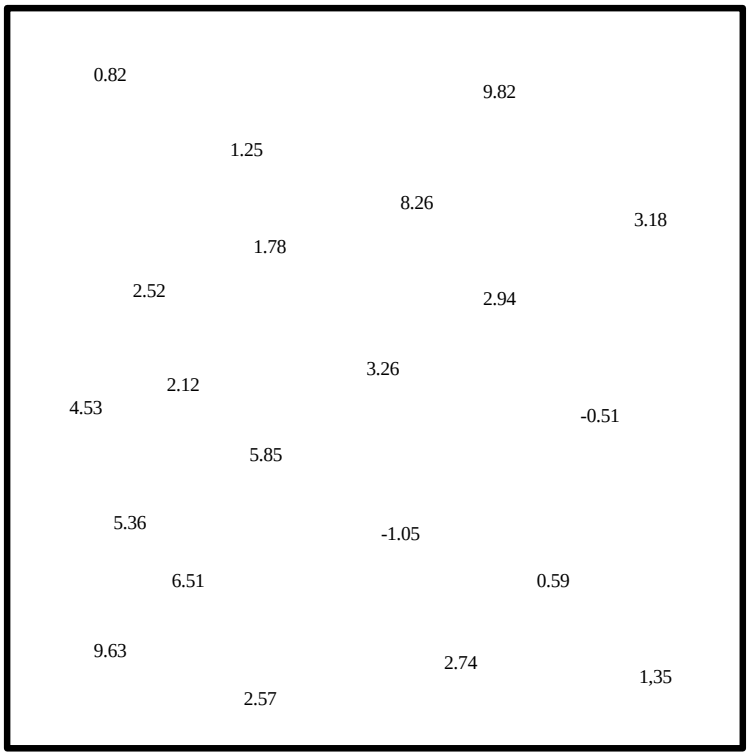
Prédire le statut d'un nouvel individu (malade - non malade).



RÉGRESSION

Expliquer le taux d'Ozone.

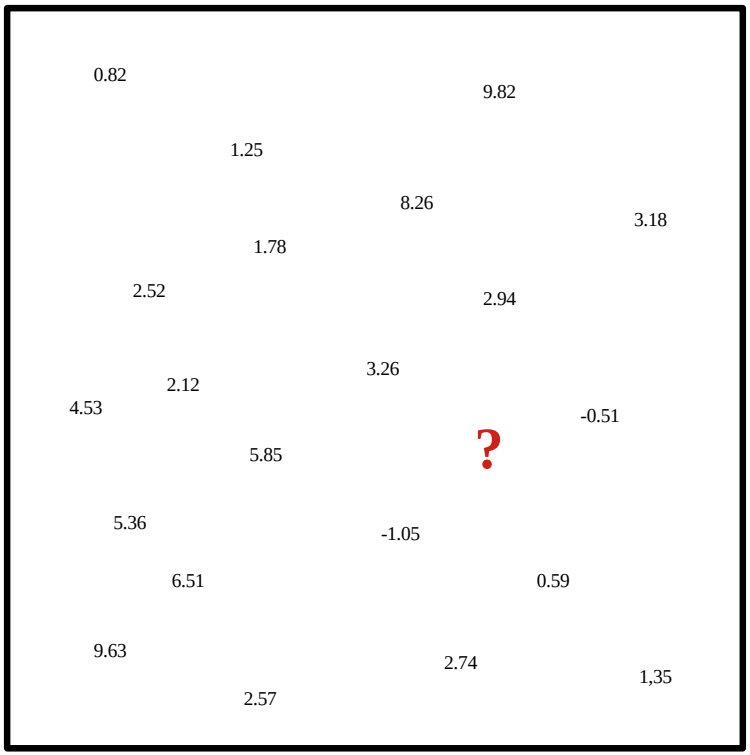
Prédire un taux d'Ozone.



RÉGRESSION

Expliquer le taux d'Ozone.

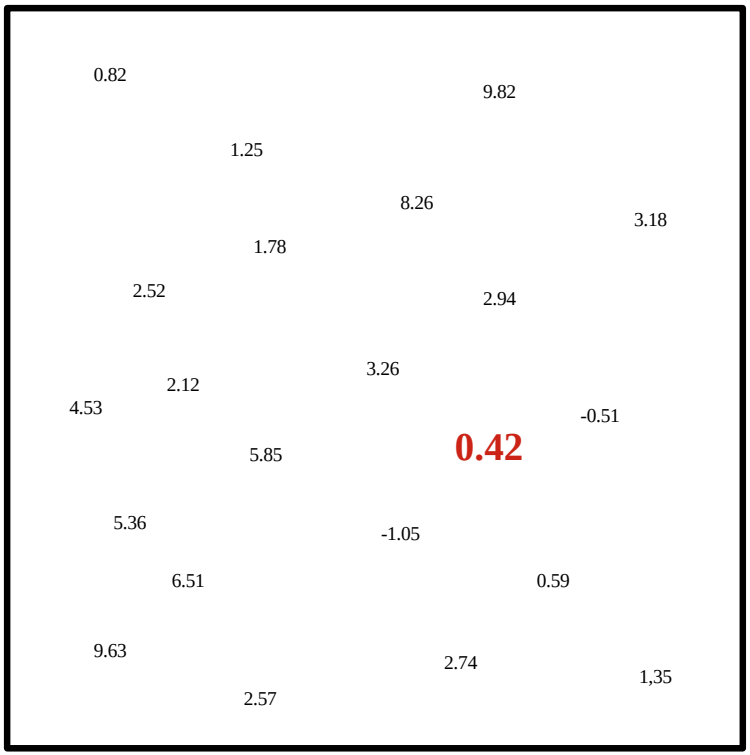
Prédire un taux d'Ozone.



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 $\mathscr{D}_n$

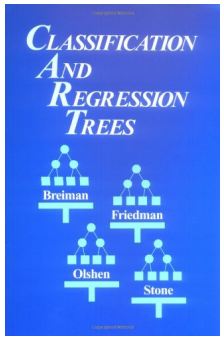
Régression : $Y \in \mathbb{R}$

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

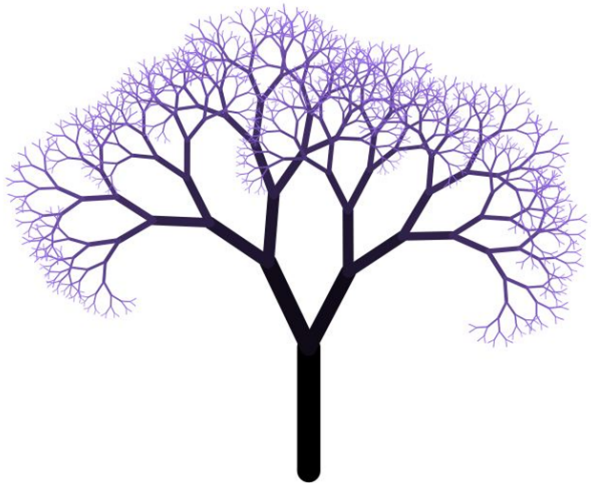
On ne cherche pas à estimer la distribution des données $(\mathbf{X}, Y)$

CLASSIFICATION AND REGRESSION TREE

- Breiman et al. (1984)

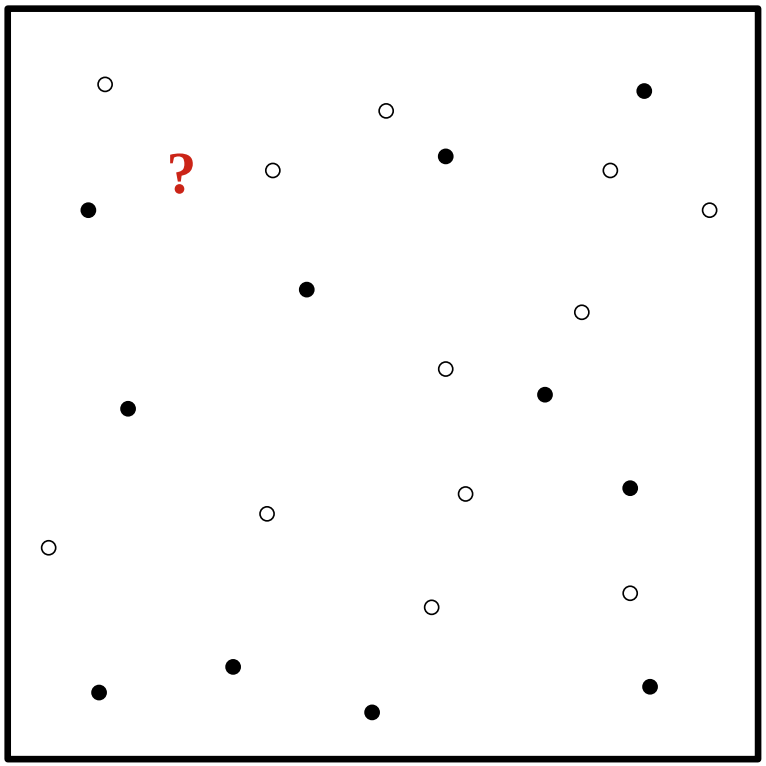


- Arbre binaire : construit de manière récursive en découpant chaque feuille en deux noeuds fils jusqu'à l'obtention d'un critère d'arrêt.



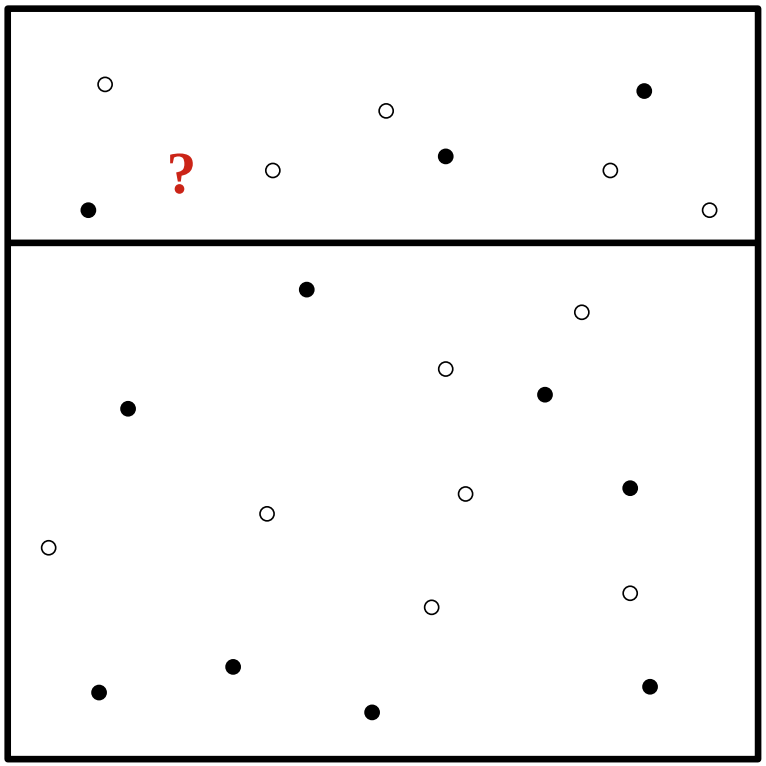
Exemple en classification

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



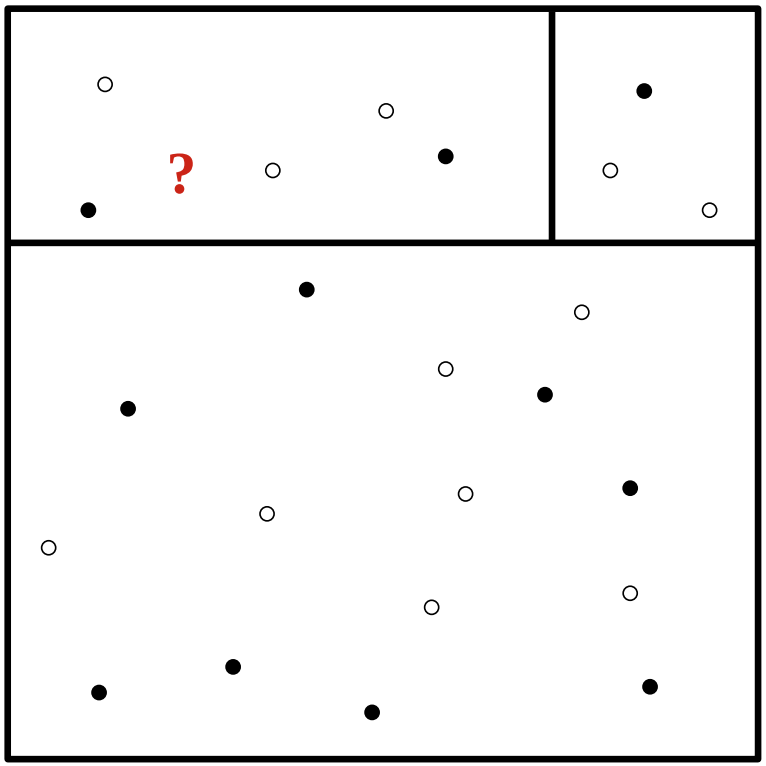
Exemple en classification

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



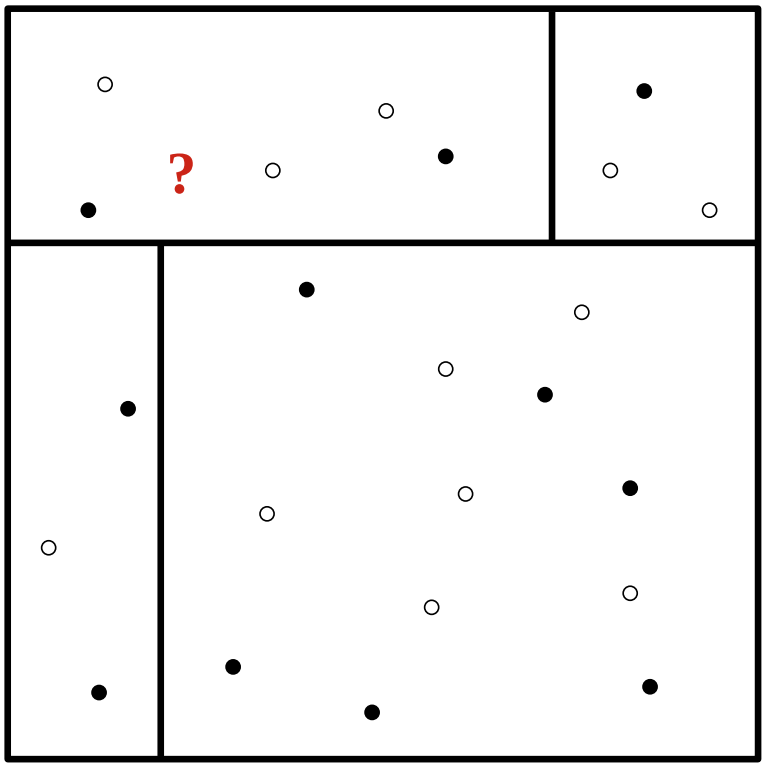
Exemple en classification

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



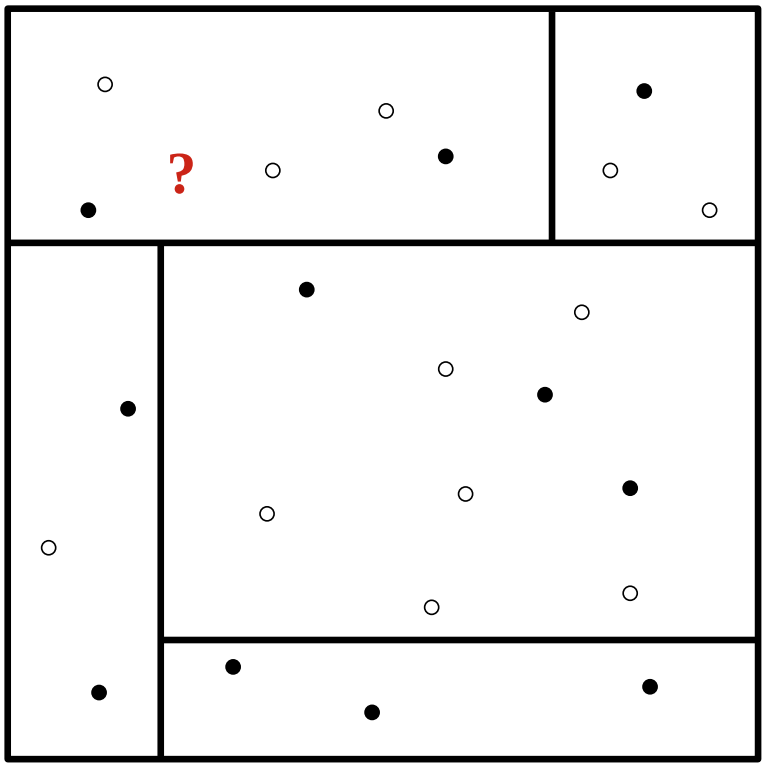
Exemple en classification

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



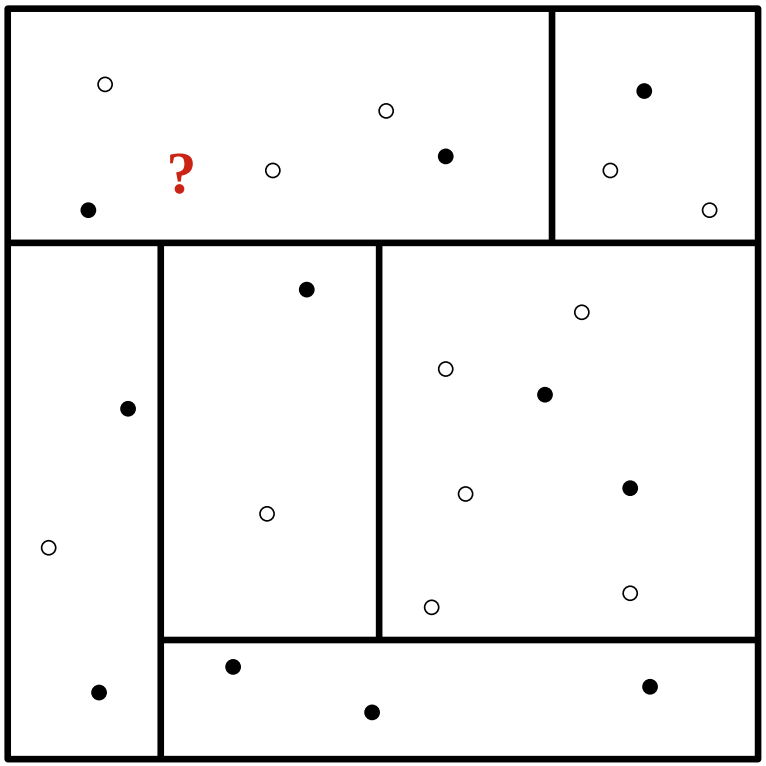
Exemple en classification

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



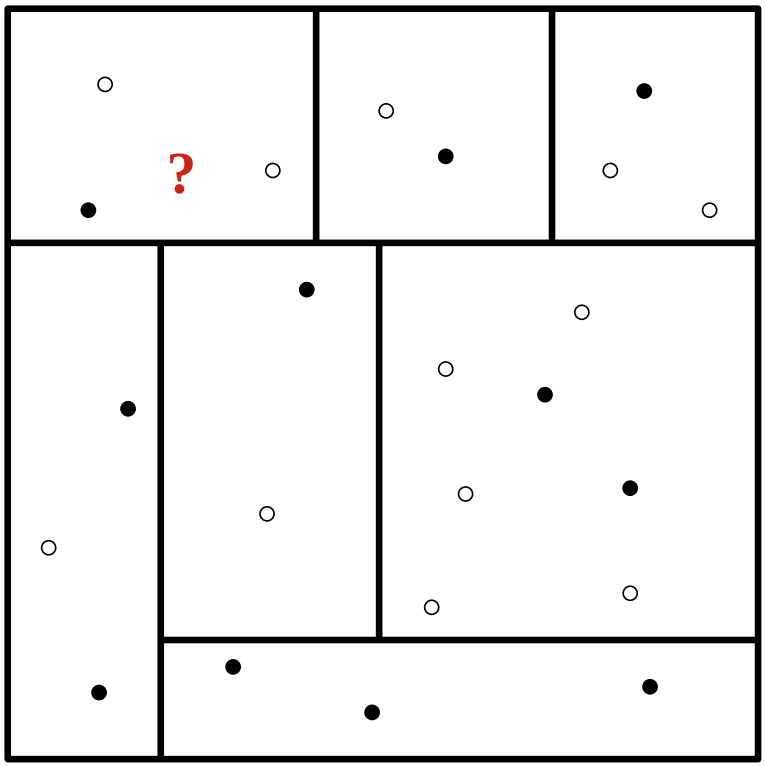
Exemple en classification

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



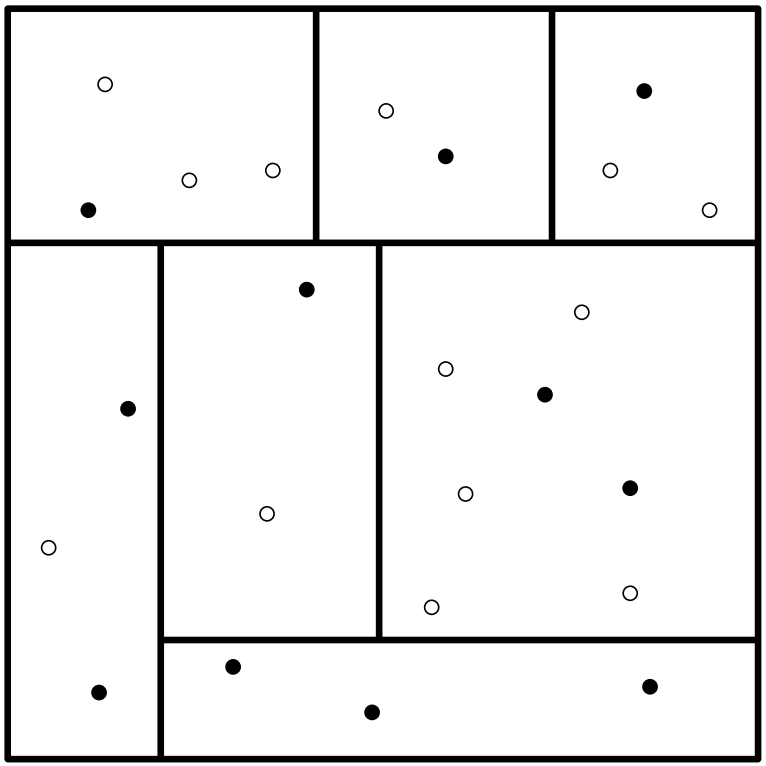
Exemple en classification

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



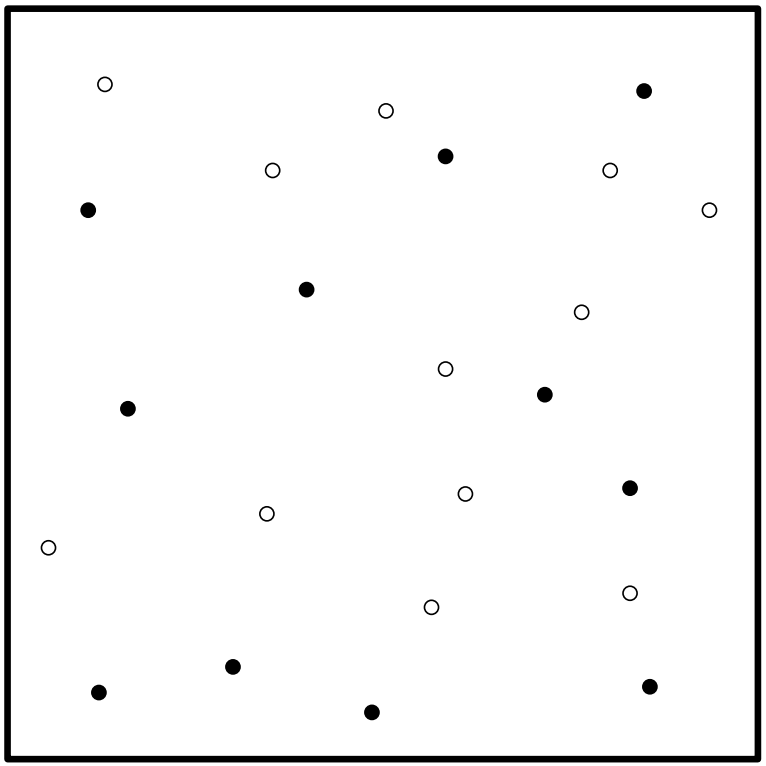
Exemple en classification

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



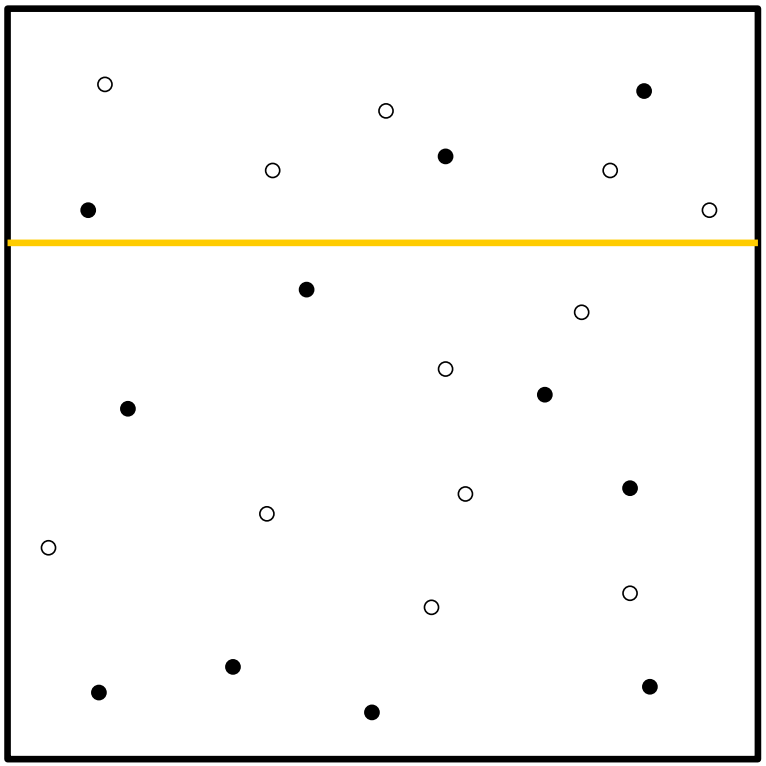
Exemple en classification

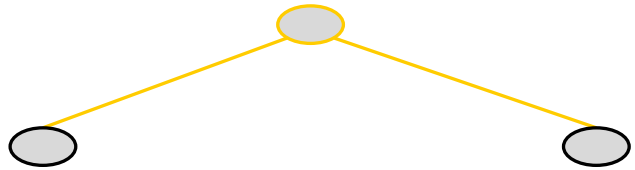
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



Exemple en classification

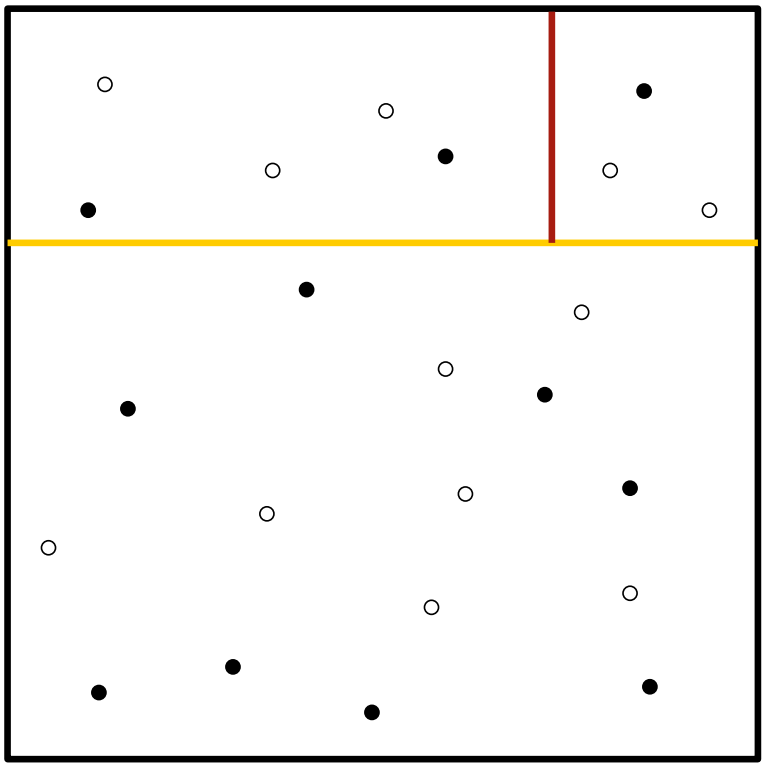
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

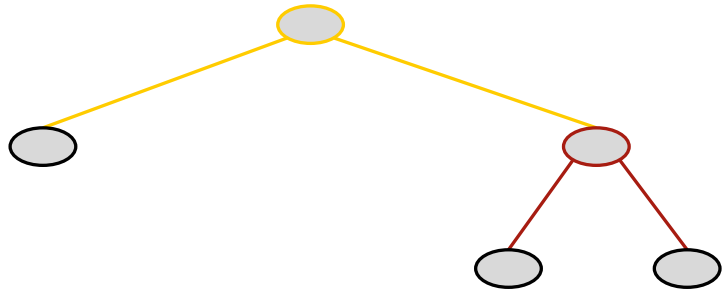




Exemple en classification

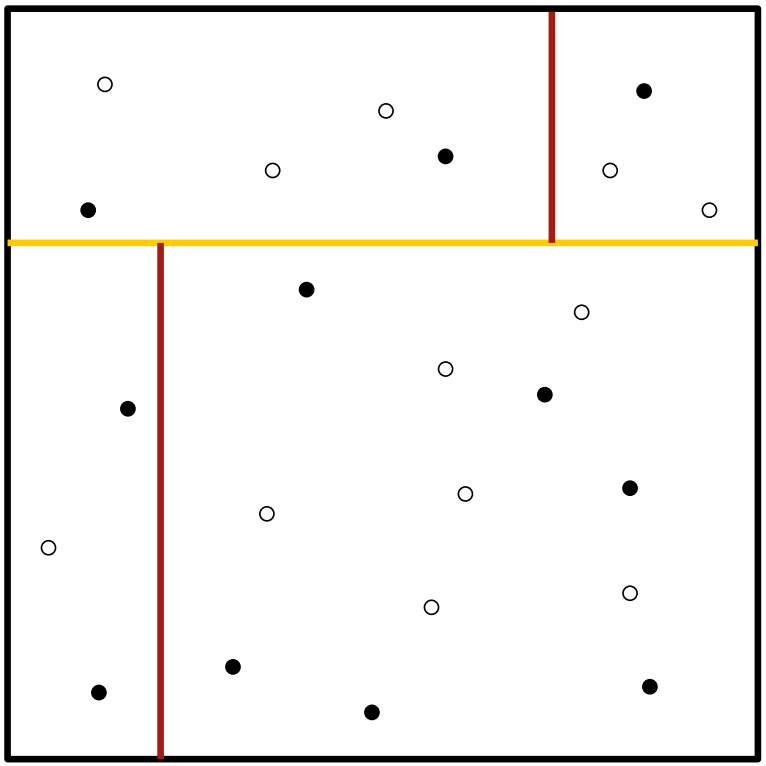
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

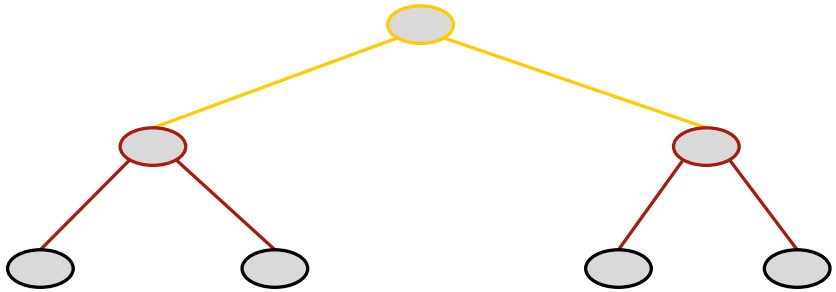




Exemple en classification

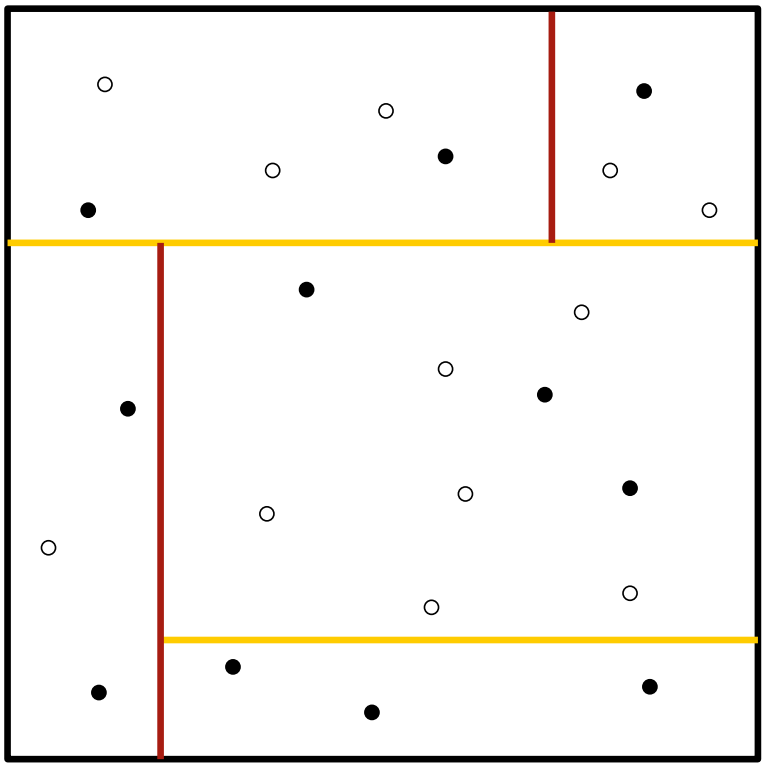
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

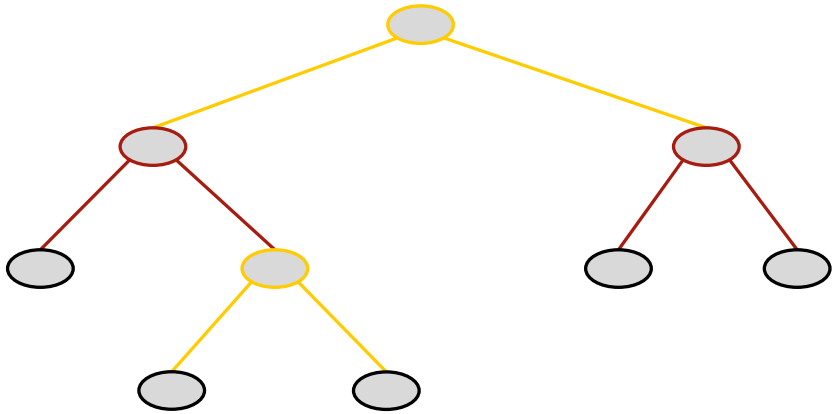




Exemple en classification

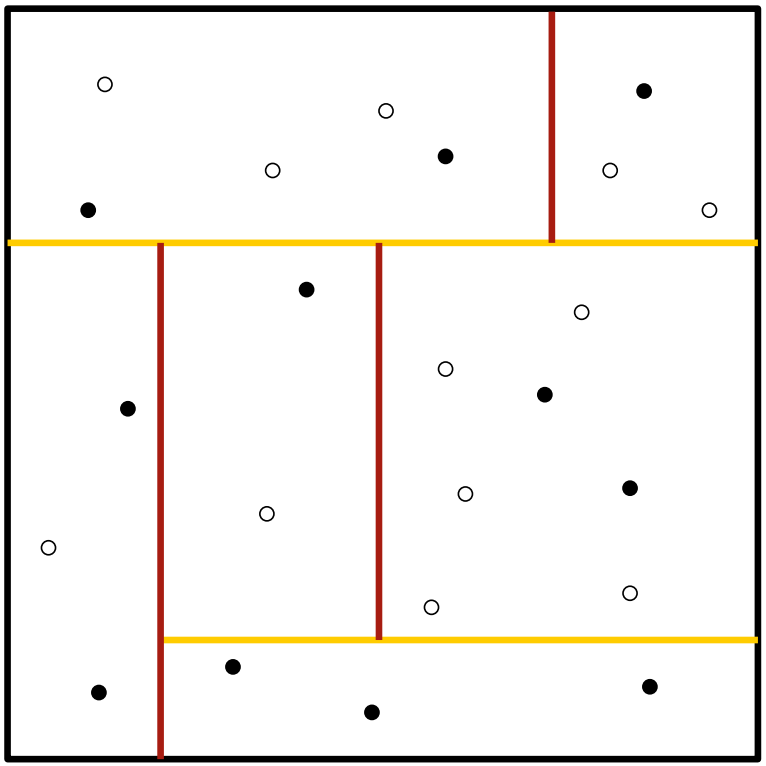
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

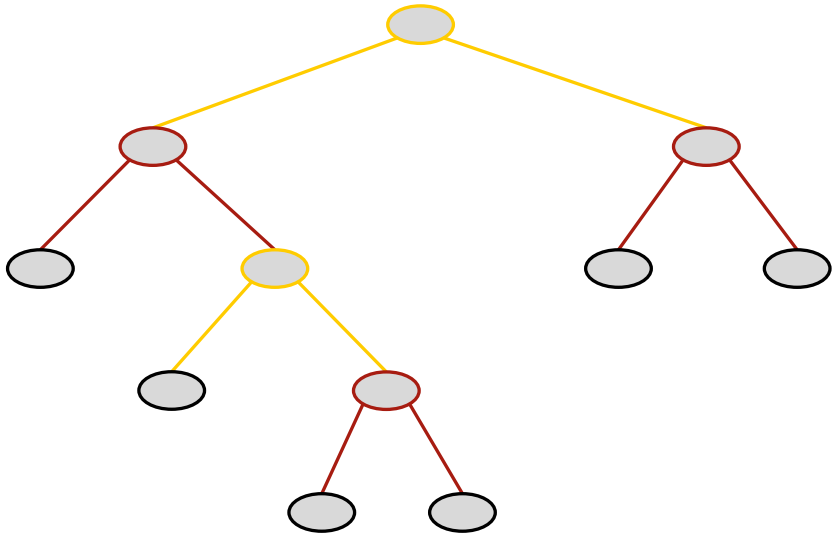




Exemple en classification

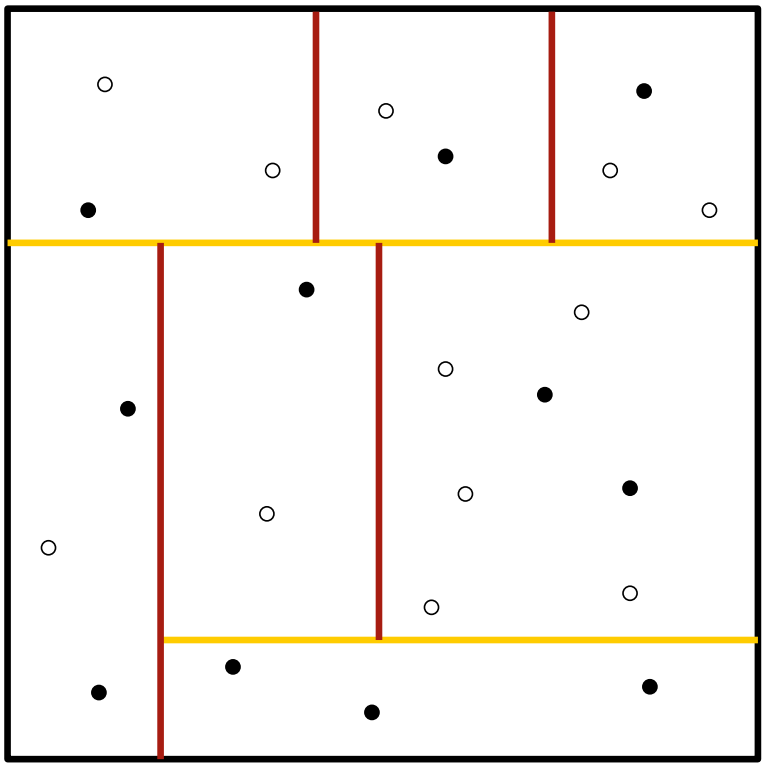
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

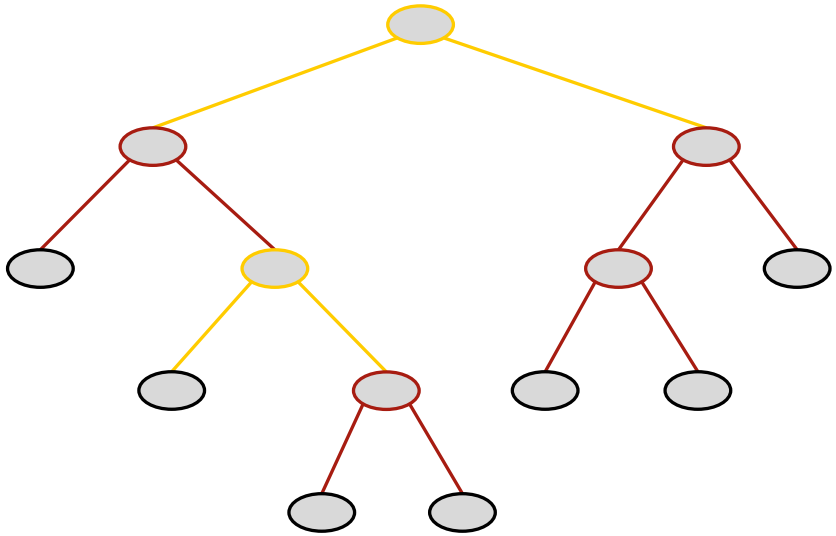




Exemple en classification

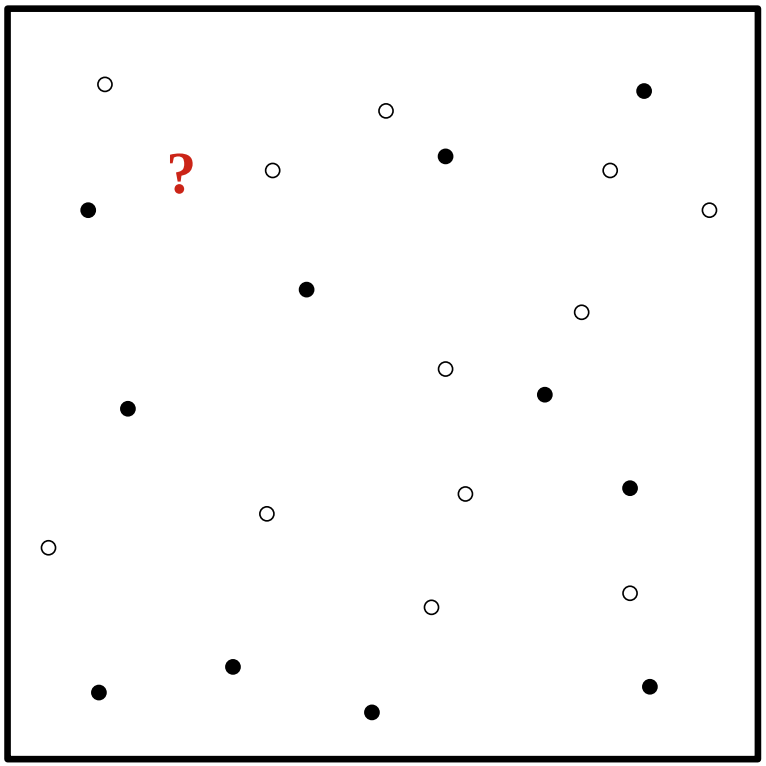
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.





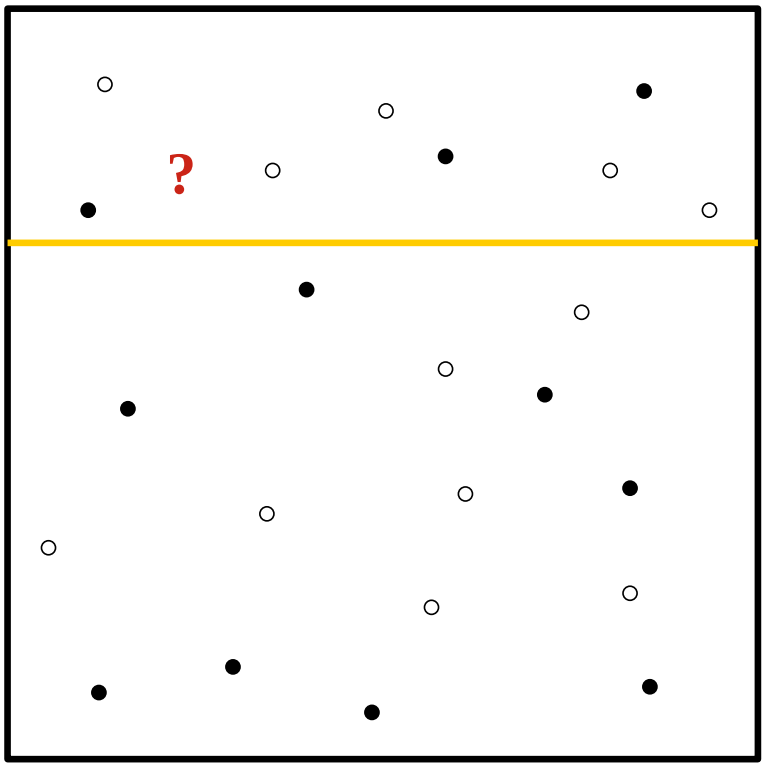
Exemple en classification

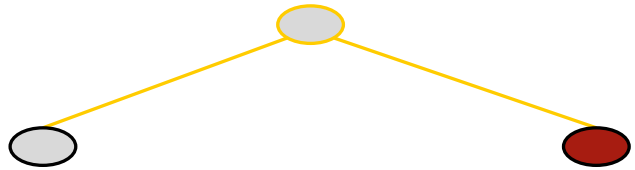
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.



Exemple en classification

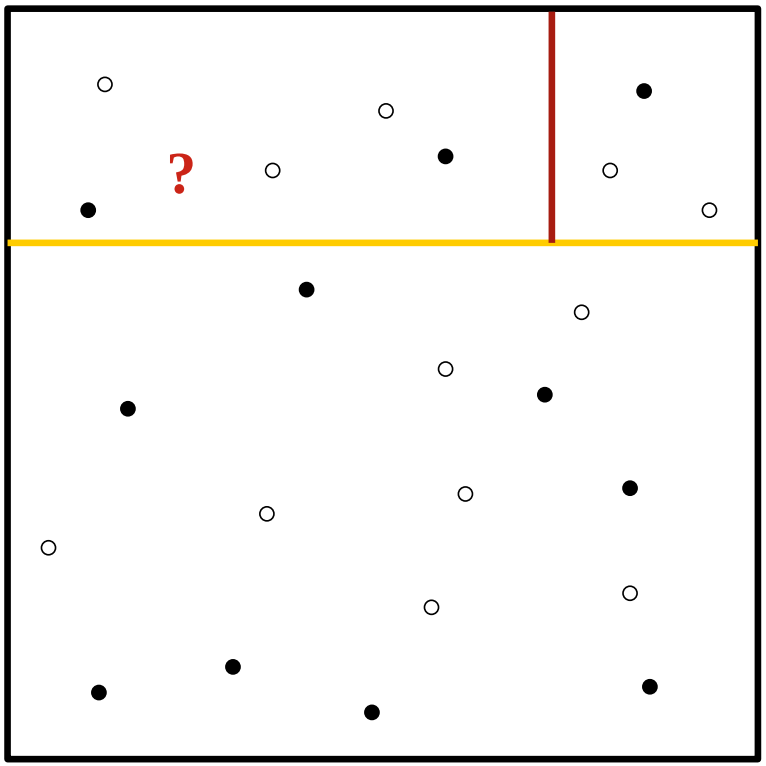
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

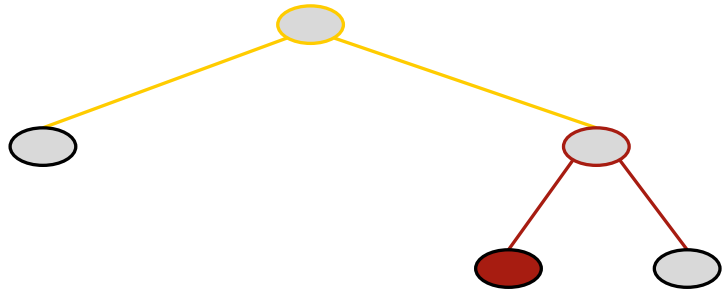




Exemple en classification

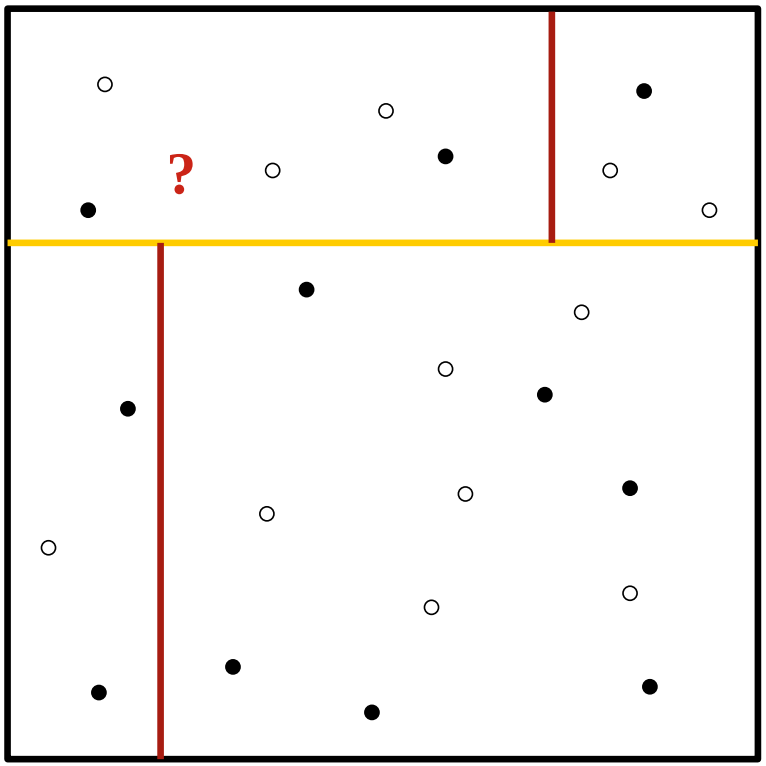
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

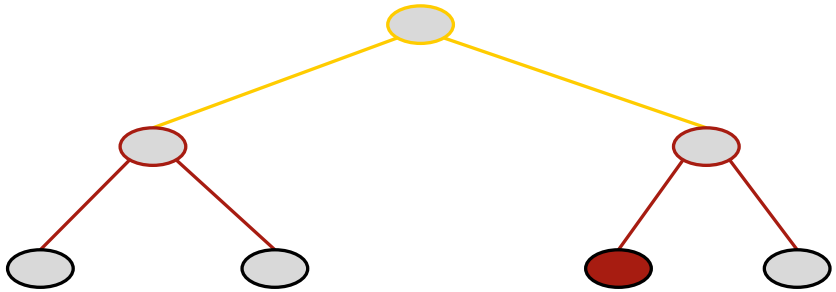




Exemple en classification

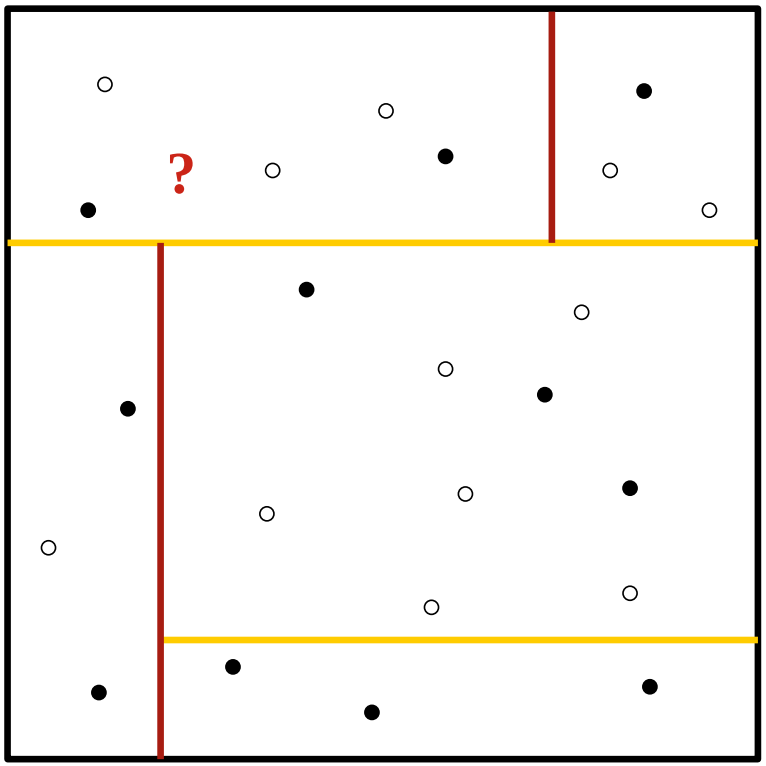
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

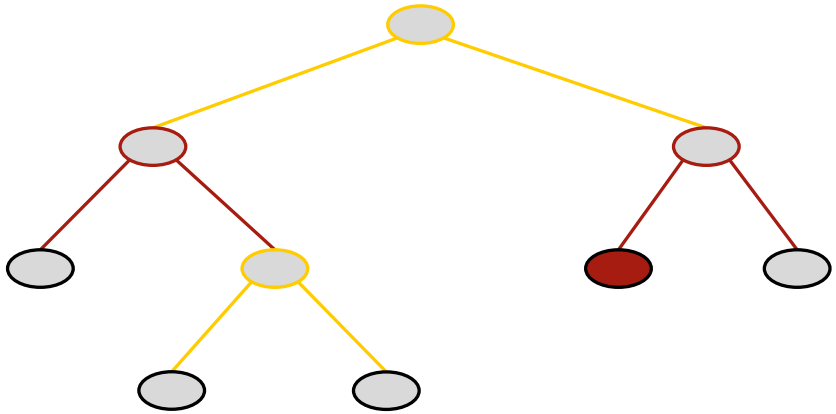




Exemple en classification

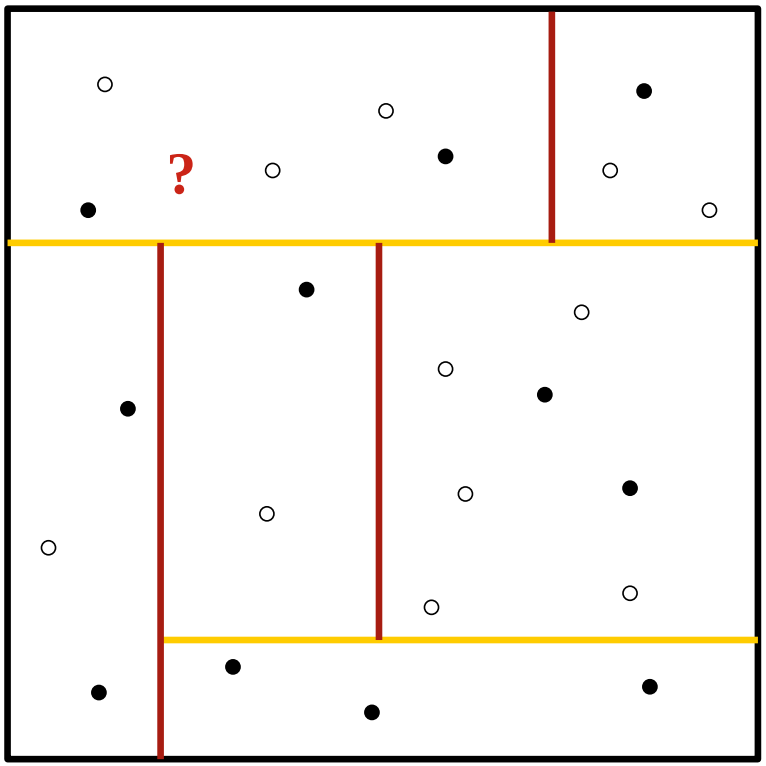
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

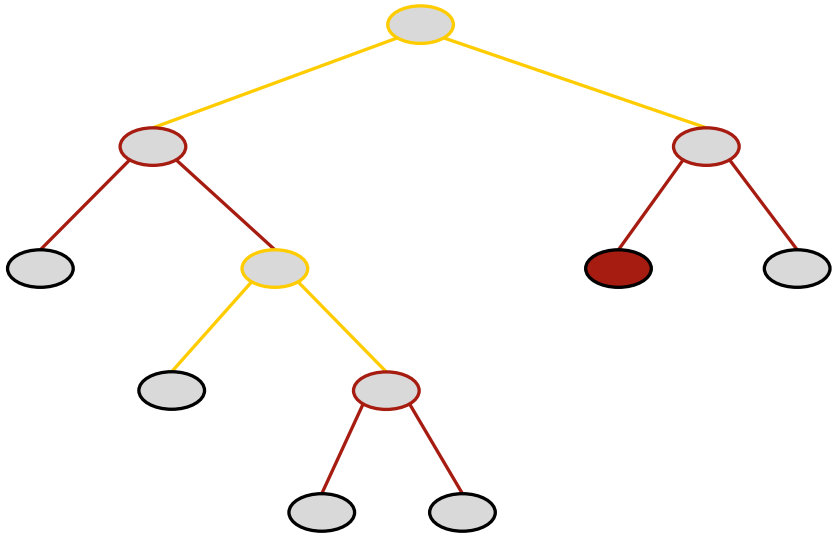




Exemple en classification

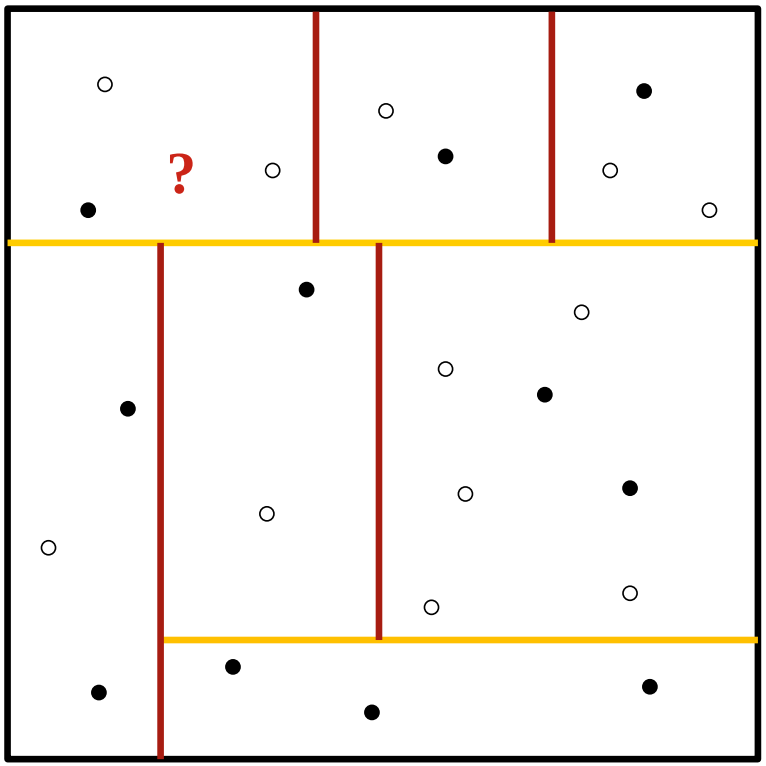
On observe 2 variables X_1 et X_2 et une sortie $Y = -1$ ou 1.

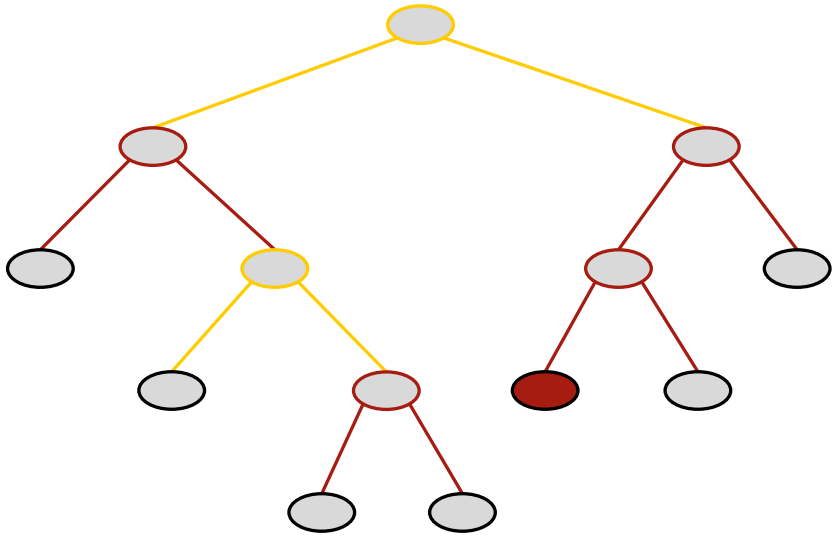




Exemple en classification

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

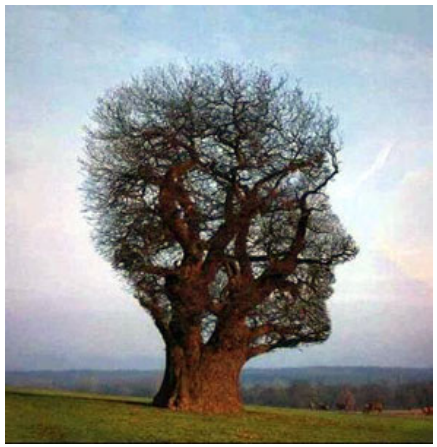
- Arbre développé jusqu'à atteindre une règle d'arrêt.

  $\rightarrow$ 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 : $R_\alpha(T) = R(T) + \alpha \bar{T}$.

- 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

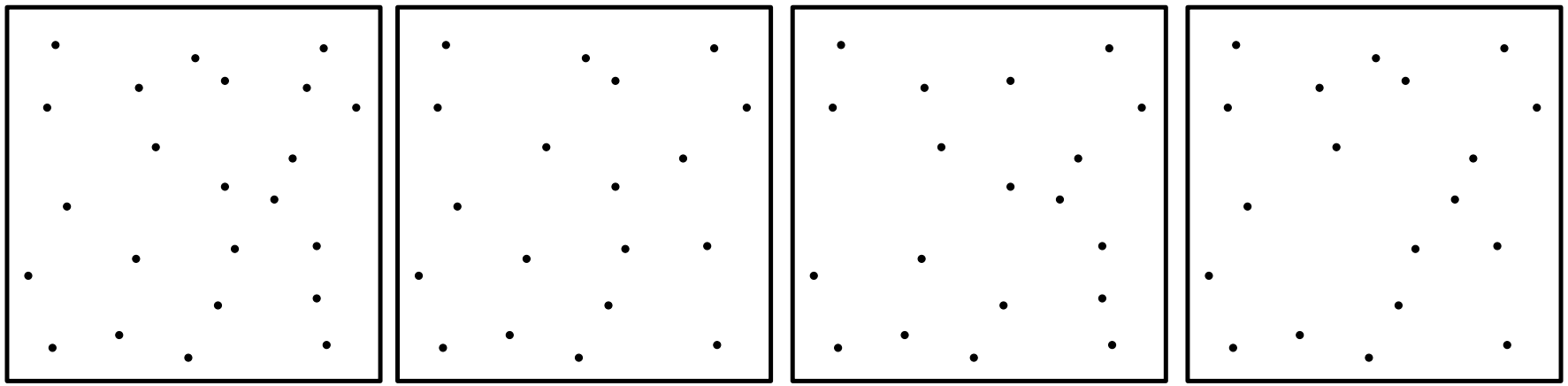
- Introduite par Breiman (2001).
- Idée : faire la moyenne de plusieurs arbres afin d'obtenir des classifieurs plus performants.
- Les arbres sont simples (pas optimisés) et randomisés (différents).

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 ($\simeq \frac{1}{3}$ des données).



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

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

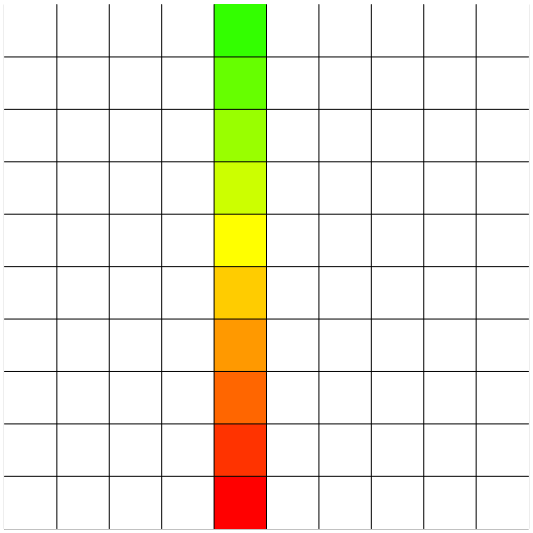
$\hat{R}$ est le risque empirique,

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

IMPORTANCE PAR PERMUTATION

MESURE D'IMPORTANCE EMPIRIQUE

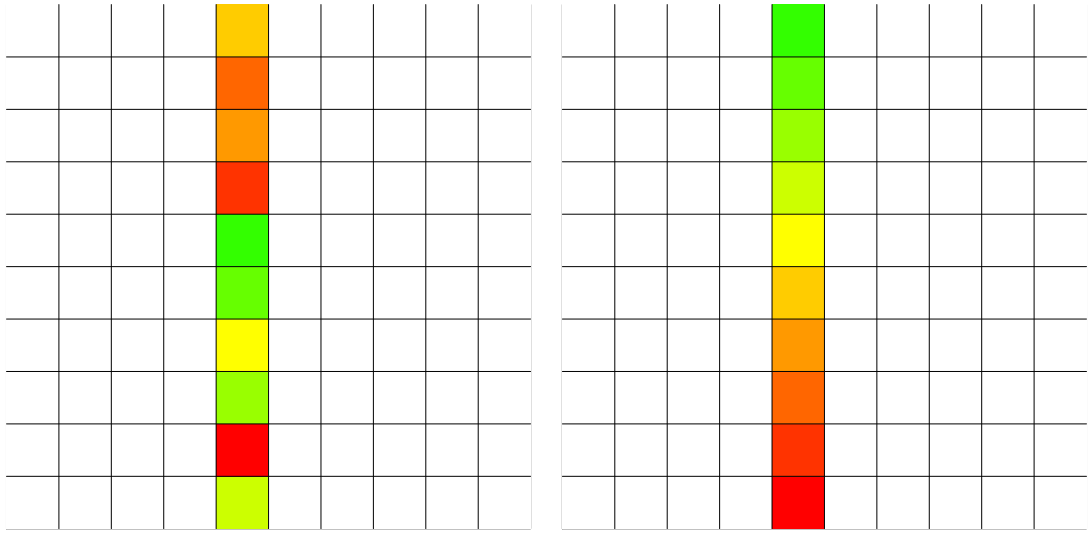
$$\hat{I}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mj}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m}) \right]$$



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

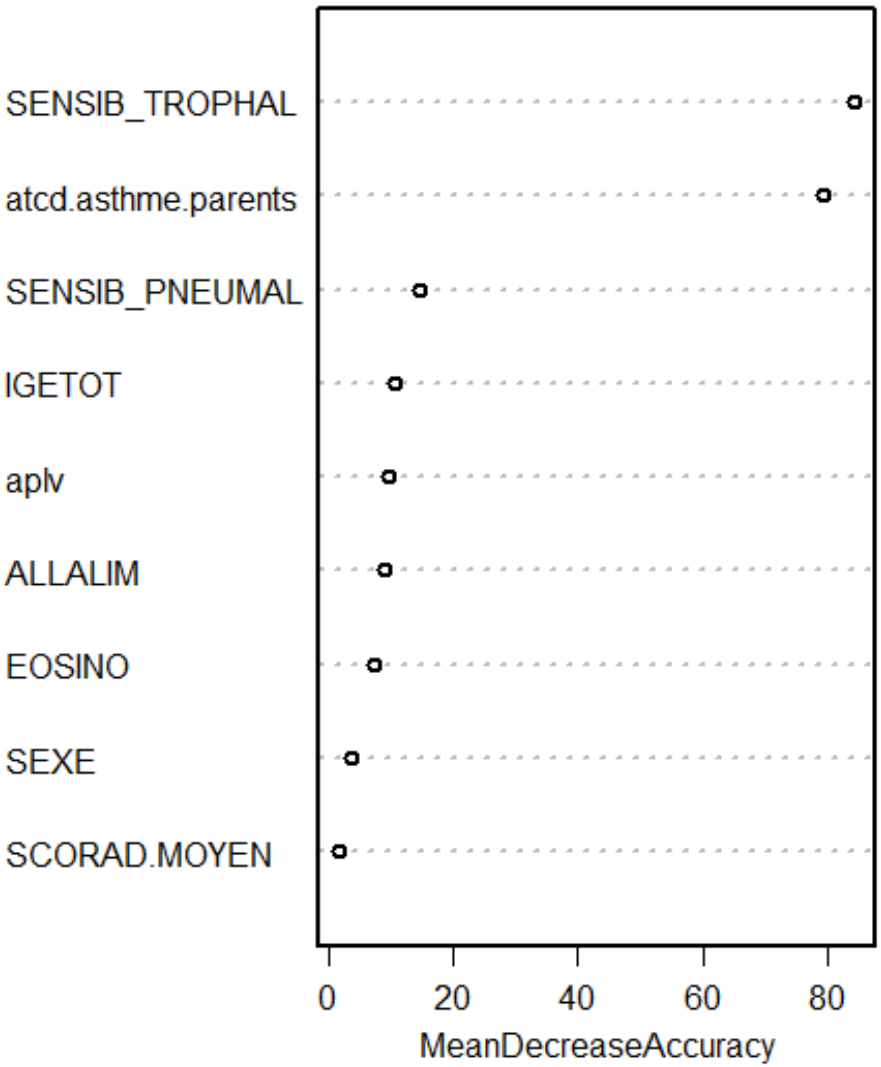
$$\hat{I}(X_j) = \frac{1}{M}\sum_{m=1}^{M}\left[\hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{mj}) - \hat{R}(\hat{f}_m, \bar{\mathscr{D}}_n^{m})\right]$$

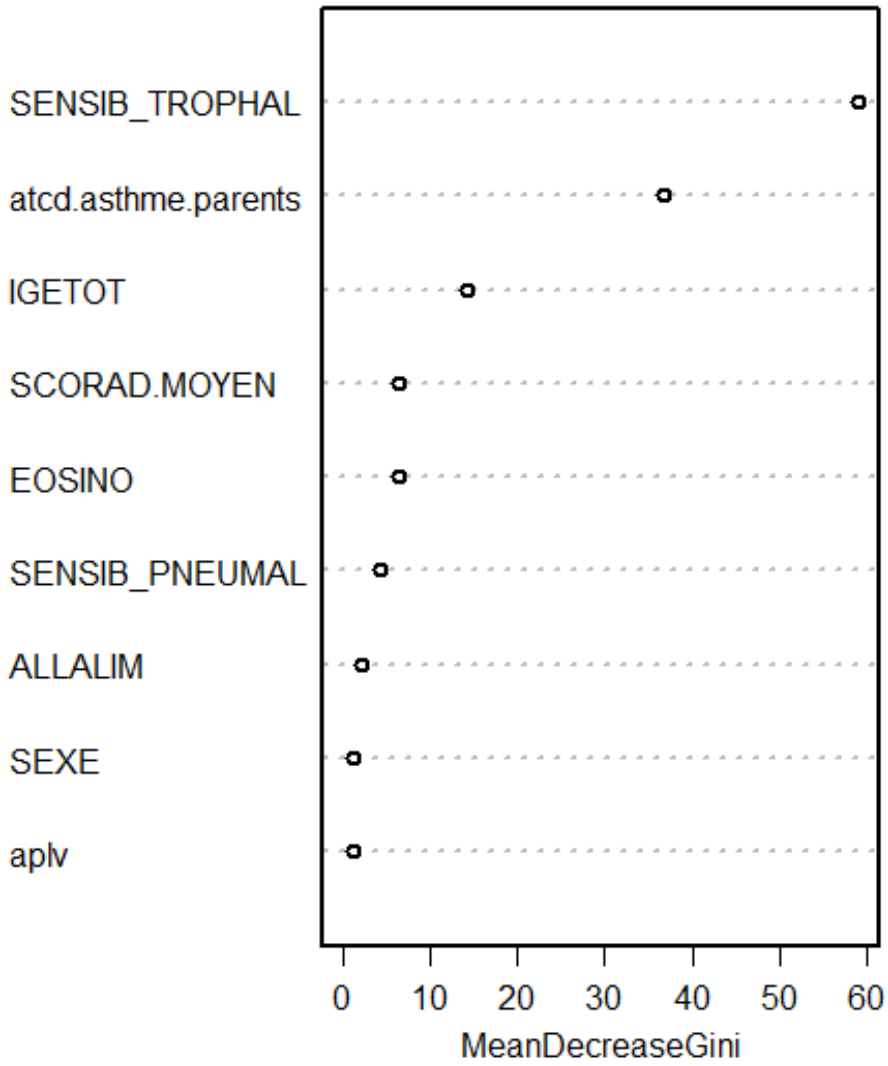
MESURE D'IMPORTANCE THÉORIQUE

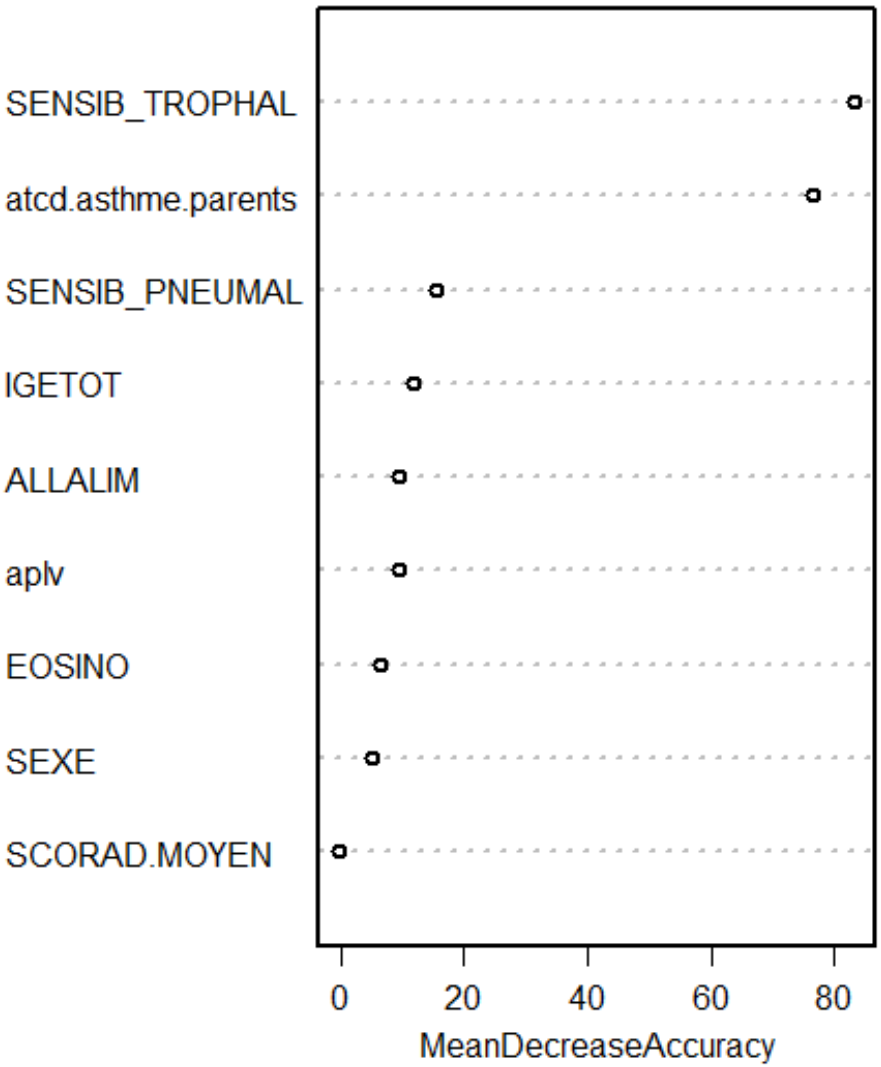
$$I(X_j) = \mathbb{E}\left[\left(Y - f(\mathbf{X}_{(j)})\right)^2\right] - \mathbb{E}\left[(Y - f(\mathbf{X}))^2\right]$$

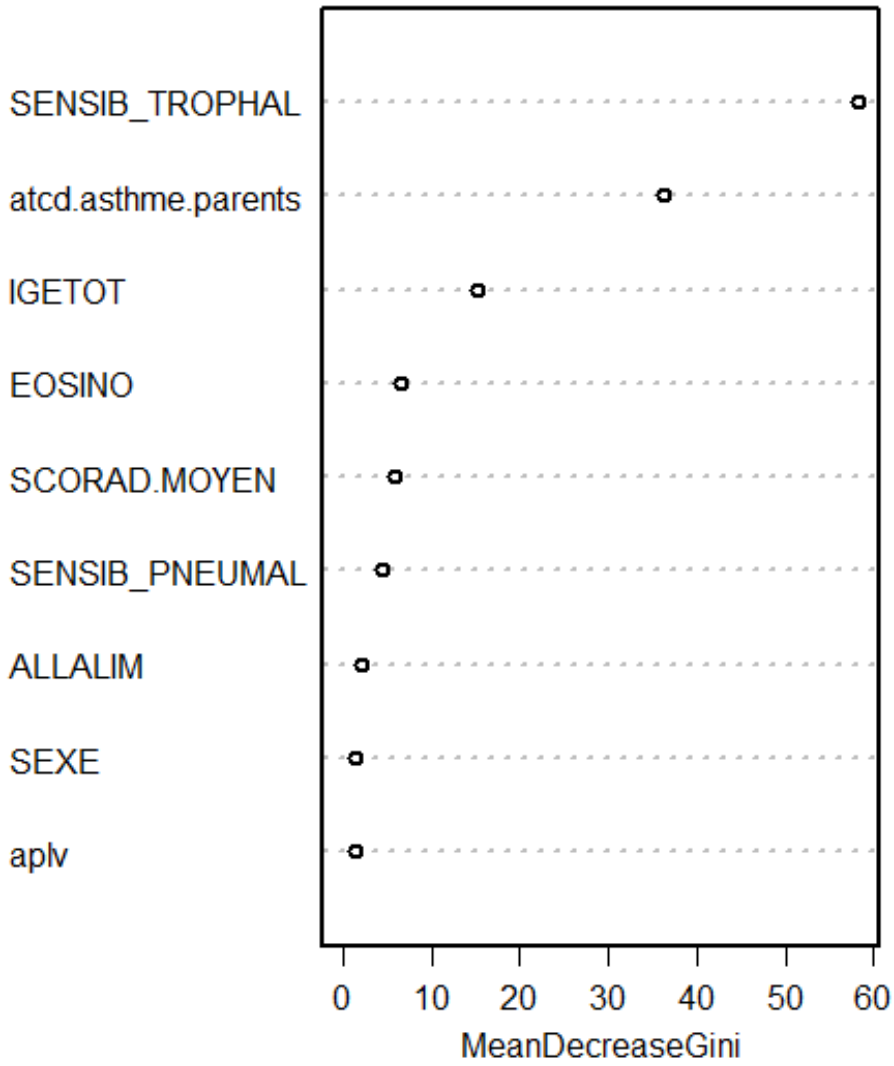
$\mathbf{X}_{(j)} = (X_1, \ldots, X'_j, \ldots, X_p)$,

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

SURVIVAL TREE

- Survival tree is similar to decision tree which is built by recursive splitting of tree nodes.
- For each node, examine every possible split on each predictor.
- Select the split which maximizes the survival difference (spliting criterion) between two children nodes.
- Idea :

$$RandomForests + SurvivalTree = RandomSurvivalForest$$

- References
  - Segal, M.R. Regression trees for censored data. Biometrics,1988.
  - LeBlanc, M. and Crowley, J. Survival Trees by Goodness of Split. Journal of the American Statistical Association, 1993.
  - H. Ishwaran, U. B. Kogalur, E. H. Blackstone and M. S. Lauer, Random Survival Forests. Annals of Applied Statistics, 2008.

SPLITTING RULES

- Log-rank splitting
  - The log-rank statistic for a given split at the value c for predictor X (two groups $X \leq c$ and $X > c$) is a measure of node separation.
  - The larger the value for log-rank statistic, the greater the difference between the two groups, and the better the split is.
  - The best split at node h is determined by finding the predictor X^* and split value c^* which maximize the log-rank statistic.
- The logrank test is most commonly used dissimilarity measure that estimates the survival difference between two groups.
- Others rules : Conservation of events splitting, Log-rank score splitting or Approximate logrank splitting

CUMULATIVE HAZARD FUNCTION ESTIMATION

- For a given tree, an estimate of the Cumulative Hazard Function (CHF) is derived as follows
  - The cumulative hazard estimate for each terminal node is obtained using the Nelson-Aalen estimator
  - If there are M terminal nodes in the tree, then there are M such estimates.
  - Each tree provides a sequence of such estimates.
- To compute the CHF for an individual i with predictor X_i, simply drop X_i down the tree. The terminal node for i yields the desired estimator.
- Note this value is computed for all individuals i in the data.

ENSEMBLE ESTIMATION OF THE CHF

- Each tree provides an estimation of the cumulative hazard given X_i.
- An ensemble estimate is simply obtained by averaging the different estimation of the cumulative hazard given X_i (one for each tree).
- In the context of Out-Of-Bag (OOB) obervations
  - Note that the ensemble estimator uses all the trees prediction to compute the mean CHF
  - The OOB estimator of the CHF $\hat{H}(t \mid X_i)$ is obtained by averaging over only those bootstrap samples in which i is excluded (i.e., those datasets in which i is an OOB value).
- The estimation of the cumulative hazard function is used to derive error rate performance.

CONCORDANCE INDEX

- Given the OOB estimator, the error rate can be easily derived using Harrell's concordance index (Harrell et al., 1982).
- Harrell's C-index does not depend on choosing a fixed time for evaluation of the model and specifically takes into account censoring of individuals.
- Concordance index estimation need to define what constitutes a worse predicted outcome
  - $t_1, \ldots, t_N$ denote all unique event times in the data
  - Individual i is said to have a worse outcome than j if

$$\sum_{k=1}^{N} \hat{H}(t_k \mid X_i) > \sum_{k=1}^{N} \hat{H}(t_k \mid X_j)$$

CONCORDANCE INDEX

The concordance error rate is computed as follows :

1. Form all possible pairs of observations over all the data.
2. Omit those pairs where
   - the shorter event time is censored,
   - or where event times are equal unless one of the observations is a death and the other a censored observation.
   - Let *Permissible* denote the total number of permissible pairs.
3. Let *Concordance* denote the total sum over all permissible pairs
   - Count 1 for each permissible pair in which the shorter event time had the worse predicted outcome.
   - Count 0.5 if the predicted outcomes are tied.

ERROR RATE (PREDICTION ERROR)

4 The concordance index C is defined by

$$C = \frac{Concordance}{permissible}$$

5 The error rate is defined by

$$Error = 1 - C$$

6 Note that

- $0 \leq Error \leq 1$.
- $Error = 0.5$ corresponds to a procedure doing no better than random guessing.
- $Error = 0$ indicates perfect accuracy.

ALGORITHM : RANDOM SURVIVAL FOREST

1. Draw $ntree$ bootstrap samples from the original data.
2. Grow a survival tree for each bootstrap sample.
   - At each node of the tree randomly select $mtry$ predictors .
   - Select the split which maximizes the survival difference between two children nodes.
3. Grow the tree to full size, each terminal node should have no less than $nodesize$ unique deaths.
4. Calculate a cumulative hazard function for each tree.
5. Using OOB data, calculate an OOB ensemble CHF by averaging information from the exigible trees. One estimate $\hat{H}(t \mid X_i)$ for each individual i in the data is calculated.
6. Compute the error rate using the concordance index and the estimation of the OOB cumulative hazard.

VARIABLE IMPORTANCE

For a given variable X_j, the importance variable is obtained as follows

- Compute the prediction error $Error_k(X_j)$ (defined using the C-index) for each tree k : $Error_k(X_j)$
- Permute the variable X_j
- Compute theprediction error for the permuted predictor $\bar{X}_j$ for each tree k : $Error_k(\bar{X}_j)$
- Permutation importance for variable X_j is obtained by averaging over the M exigible tree

$$\hat{I}(X_j) = \frac{1}{M} \sum_{k=1}^{M} [Error_k(\bar{X}_j) - Error_k(X_j)]$$

ILLUSTRATION WITH R

R packages

- randomSurvivalForest
- randomForestSRC

```
> library("randomSurvivalForest")
> data(veteran,package="randomSurvivalForest")
> ntree <- 1000
> v.out <- rsf(Survrsf(time,status) ~ karno,
        veteran, ntree=ntree, forest=T)
> print(v.out)

Call:
 rsf.default(formula = Survrsf(time, status)
      ~ karno, data = veteran, ntree = ntree)

                         Sample size: 137
                    Number of deaths: 128
                     Number of trees: 1000
          Minimum terminal node size: 3
       Average no. of terminal nodes: 8.437
No. of variables tried at each split: 1
              Total no. of variables: 1
                       Splitting rule: logrank
              Estimate of error rate: 36.28%
```

ILLUSTRATION WITH R