

Exploration des données à l'aide des arbres de décisions

Ecole thématique Explo-SHS

Grégoire Le Campion / Hugues Pécourt

Mercredi 14 octobre 2020

Contenu de l'Atelier

I- Présentation des principes généraux d'un arbre de décision

A- Présentation théorique

B- Avantages et limites

C- Fonctionnement des arbres

II- Exemple et décryptage d'un arbre

III-BaobARD : application shiny pour réaliser ces premières analyses

IV- Travaux pratiques

V- Un peu de code... éventuellement

Présentation des arbres de décision



Les arbres de décision c'est quoi ? I

Les arbres de décisions font partis du champs très à la mode du machine learning.

Très simplement le machine learning à pour objectif de chercher des patterns et d'effectuer des prédictions à partir de données en se basant sur des statistiques.

Le machine learning regroupe un ensemble de méthodes qui se divise en 2 grandes catégories : les méthodes d'apprentissage non-supervisé et les méthodes d'apprentissage supervisé.

Les arbres de décision font partis des méthodes d'apprentissage supervisé

L'objectif des méthodes d'apprentissage supervisé est d'inférer la relation entre différentes variables à partir d'un échantillon d'apprentissage.

Ainsi par exemple, si j'ai un ensemble d'information comme l'âge, le poid, le sexe et la taille, je peux mettre en oeuvre un modèle d'apprentissage supervisé pour prédire le poid en fonction des autres variables.

Les arbres de décision c'est quoi ? II

Très concrètement l'arbre de décision c'est :

Un outil permettant de prédire ou expliquer les valeurs prises par une variable, que vous aurez choisie, en fonction d'un ensemble d'autres variables que vous aurez sélectionnées.

Par ailleurs :

- ▶ Si la variable que vous souhaitez prédire est qualitative on parlera d'**arbre de classification**.
- ▶ En revanche, si la variable étudiée est quantitative alors on parlera d'**arbre de régression**.

Pourquoi les arbres de décisions ?

L'objectif d'un arbre de décision est donc de prédire les valeurs prises par une variable. Mais il existe d'autres tests statistiques plus connus qui ont le même objectif, comme notamment les tests de régressions.

Comme pour les arbres de décision, la régression cherche à prédire une variable à l'aide d'un ensemble d'autres variables et voit parmi ces prédicteurs ceux qui ont le plus d'effet sur notre variable cible.

Pourquoi alors ne pas utiliser simplement les techniques de régressions ? Notamment car ces méthodes sont assorties de conditions qu'il faut remplir pour être considérées comme fiables :

Par exemple : distribution normale des résidus, Homogénéité des résidus, non multicollinéarité...

Tenter de remplir les différentes conditions requiesnt ressemble très souvent à un parcours du combattant :



Les avantages de l'arbre de décision

L'arbre de décision possède donc l'avantage certain de ne pas avoir à se préoccuper des conditions citées précédemment.

Par ailleurs il possède également d'autres avantages...

1. Il peut servir à expliquer une variable qualitative (il s'agit d'un l'arbre de classification) aussi bien que quantitative (il s'agit d'un arbre de régression).
2. Il peut s'inscrire dans une approche mixte, on peut mélanger dans nos prédictors des variables continues, ordinales et binaire.
3. On peut introduire un grand nombre de variables dans notre modèle sans craindre de perturbations des variables sans effet. L'algorithme de production de l'arbre va sélectionner les meilleurs prédictors possible pour expliquer notre variable.
4. Gestion efficace des données manquantes qui même en nombre relativement important ne posent pas de problèmes majeurs.
5. L'algorithme de production n'est pas gourmand en ressources, peu de risque de faire planter l'ordinateur.
6. Produit une visualisation simple d'interprétation et connue bien au delà du monde de la data science qui permet de rendre compte des éventuelles interactions complexes entre les variables de notre base de données.

Les avantages de l'arbre de décision... et ses limites

L'arbre de décision offre donc une solution alternative plutôt intéressante et séduisante aux problèmes que peuvent poser les méthodes "classiques".

Toutefois il n'est bien sûr pas parfait :

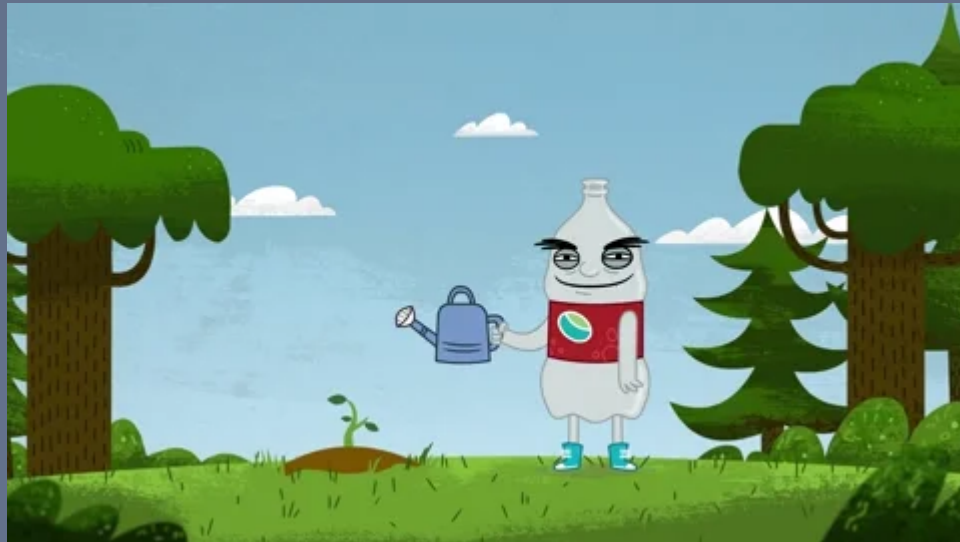
- ▶ Dans un monde idéal avec des données idéales (grand échantillons, 0 données manquantes, prédictors forts et non corrélés, distribution normale des résidus...) l'arbre de décision sera moins bon en prédiction que les méthodes multivariées (régression etc.)
- ▶ Sa production peut s'avérer assez simple mais son interprétation complexe.
- ▶ Vous n'aurez pas d'indicateurs ou de coefficients vous donnant l'importance ou le niveau des effets de tel ou tel prédictors.

En bref

Ce qu'il faut retenir :

L'arbre de décision est donc une méthode "tout-terrain" lorsque l'on doit faire de la prévision avec des données de qualité moyenne qui sont globalement la norme en SHS. Et il est surtout particulièrement adapté à **l'exploration des données**.

L'arbre de décision comment ça pousse ?



Petit traité d'arboriculture I

L'idée générale est que l'algorithme de production des arbres de décision va obéir à un principe de partitionnement récursif.

Le but de l'arbre va être de créer des groupes d'individus les plus homogènes possible entre eux par rapport à la variable étudiée.

Pour ce faire l'algorithme va "poser" des questions binaires (dont la réponse est oui/non) en lien avec les variables que vous aurez définies comme prédicteurs.

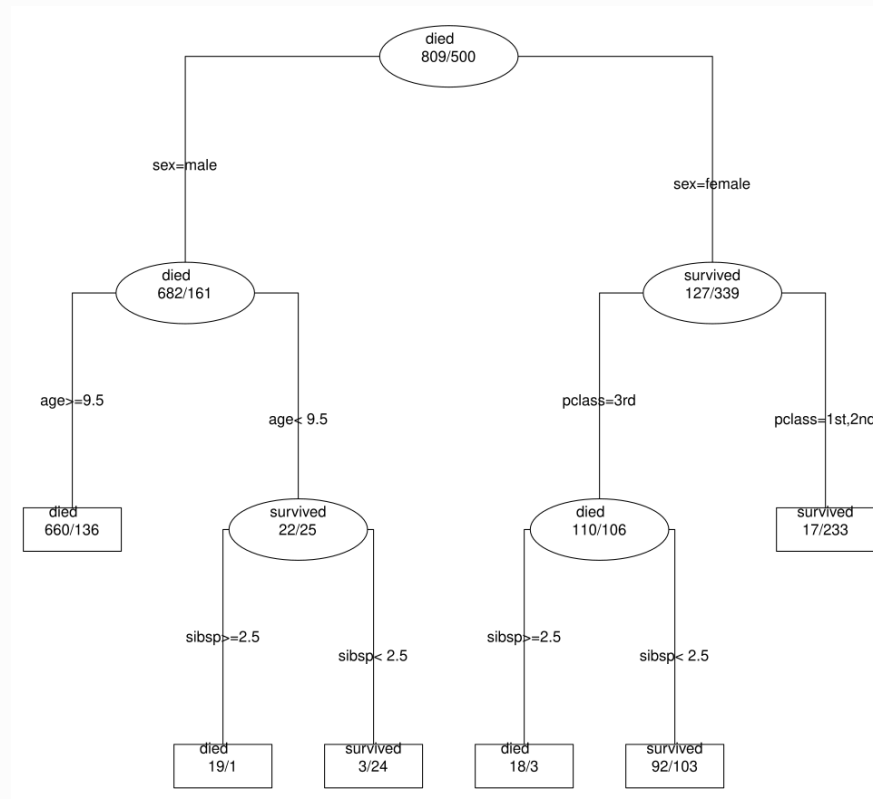
Ce sont les réponses à ces questions binaires qui constitueront les différentes ramifications de l'arbre.

Nous l'avons évoqué rapidement précédemment mais l'algorithme de l'arbre de décision va choisir les meilleurs prédicteurs possibles parmi l'ensemble des variables prédictrices que vous aurez choisi pour expliquer votre variable.

Petit traité d'arboriculture II

Pour y voir plus clair prenons un exemple et segmentons le processus de "pousse" en plusieurs étapes :

Voici un arbre de décision qui cherche à prédire la survie des passagers du Titanic



Etape 1

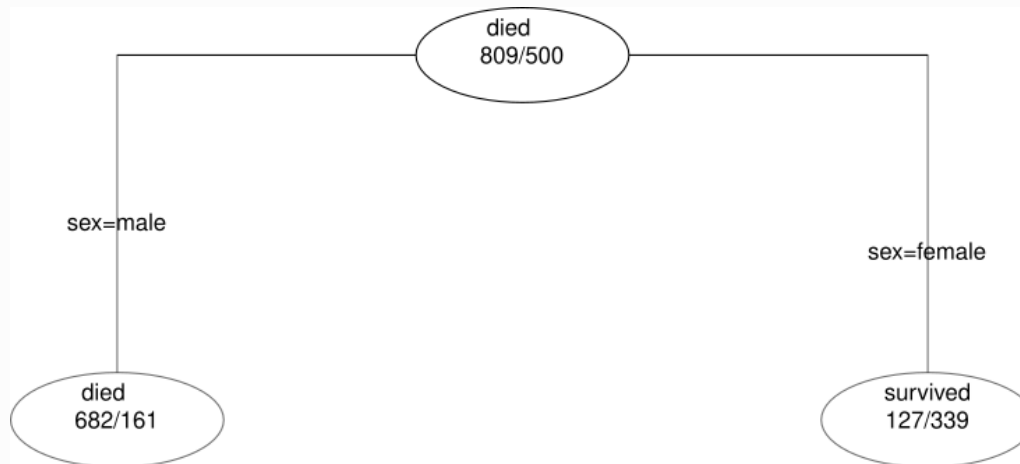
1- L'arbre pousse à partir de sa base constitué de l'ensemble des individus qui composent votre jeu de données. On parle de la racine de l'arbre.



Etape 2

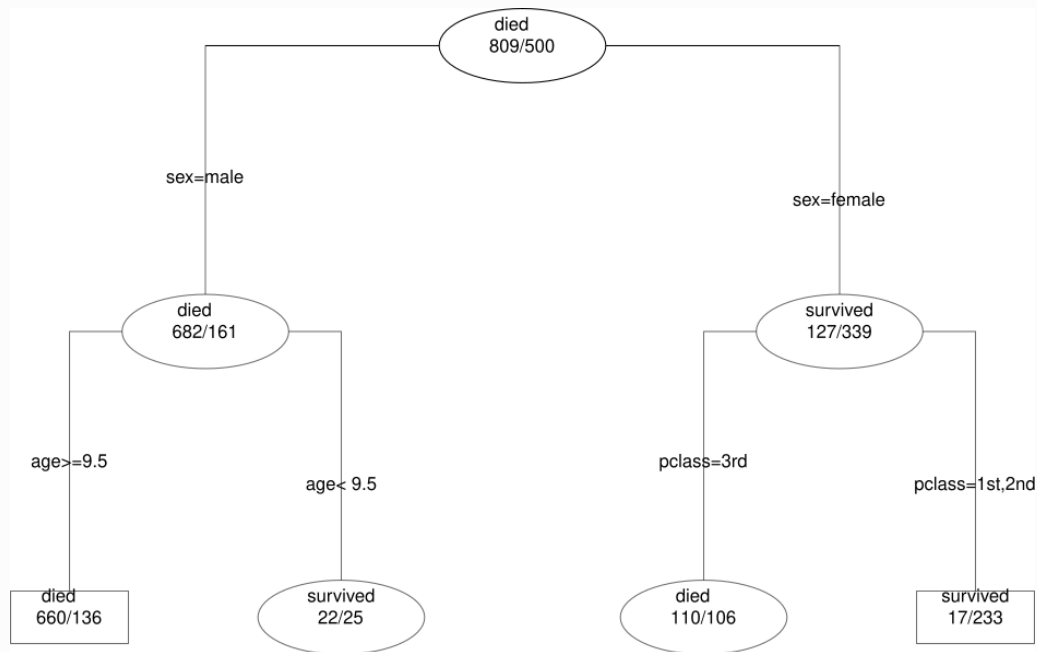
2- Une première question relative à l'une des variables prédictives sépare notre jeu de données en deux groupes.

Le choix de la question est fait de façon à ce que la réponse à la question permette d'obtenir 2 groupes les plus homogènes possible. C'est à dire le plus homogène possible en leur sein et différents l'un de l'autre.



Etape 3

3- Chacun des deux groupes obtenus peut à son tour être séparé en deux en choisissant à nouveau la meilleure question possible à poser au meilleur prédicteur possible pour constituer deux nouveaux groupes.



Etape 4 et 5

4- Le processus est répété récursivement !

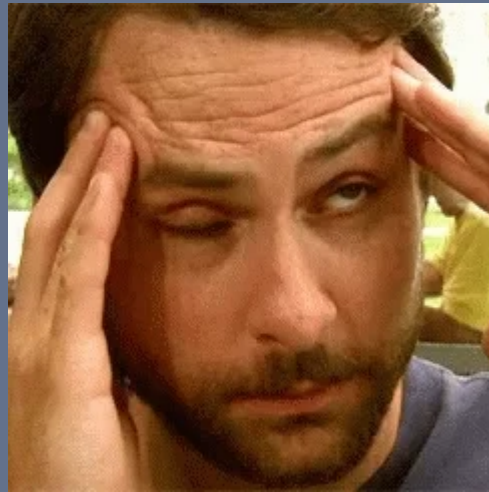
5- Lorsque l'arbre a atteint sa taille optimale, c'est à dire lorsque les divisions ne permettent plus d'obtenir des groupes suffisamment homogènes et différents les uns des autres, alors le processus s'arrête. Ces groupes d'individus finaux constituent les feuilles de l'arbre.

Petit traité d'arboriculture III

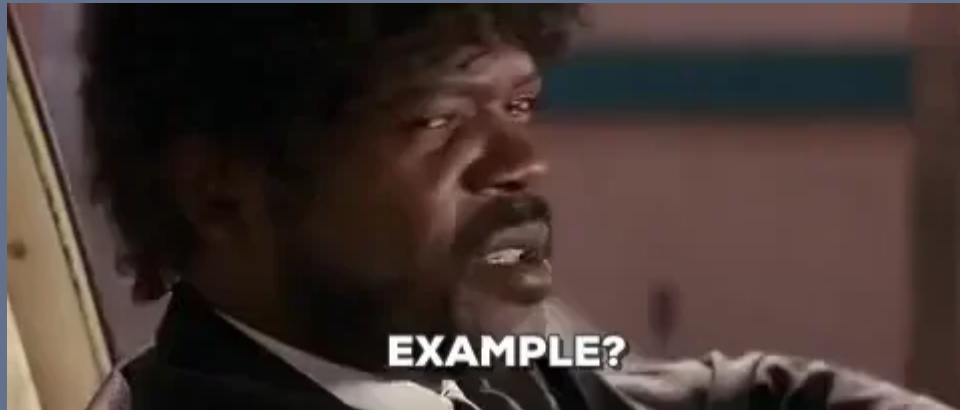
Il reste encore cependant à répondre à 3 questions fondamentales !

- ▶ Comment sont choisis les meilleures questions possible pour constituer nos branches de l'arbre ? Et donc l'utilisation de tel ou tel prédicteur.
- ▶ Comment est décider la taille optimale de l'arbre ?
- ▶ Comment interprete t-on ces groupes finaux d'individus ? Une fois nos groupes définitifs d'individus produits quel genre de modèle est crée pour chacun d'entre eux.

Est ce que ça va ?

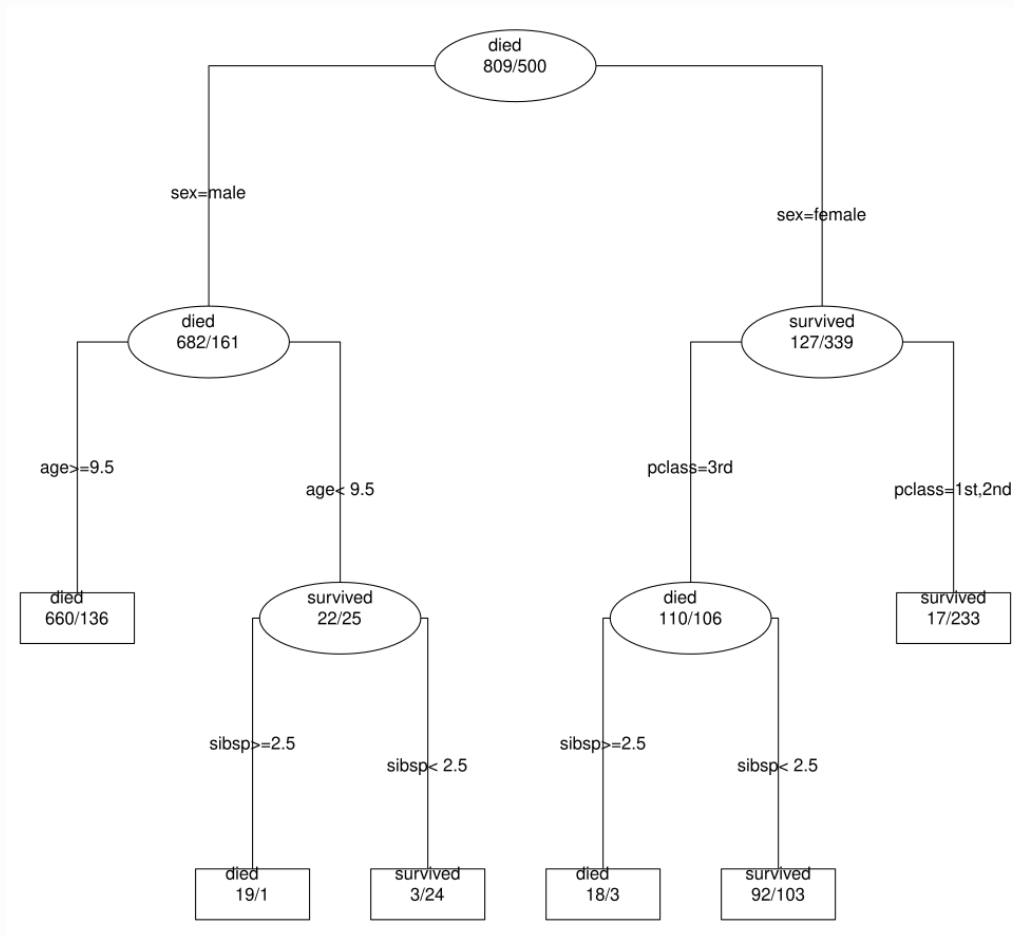


Utilisons un exemple !



Exemple 1 : un arbre de classification

Pour ce 1er exemple, reprenons notre cas sur les survivant du Titanic !



Décryptage

Nous cherchons à expliquer ce qui prédirait la survie de notre échantillon de Passagers. Nos prédicteurs : la classe économique de la cabine, l'âge, le genre, le nombre de parents et/ou enfants à bord, et le nombre de frères/soeurs/conjoints à bord.

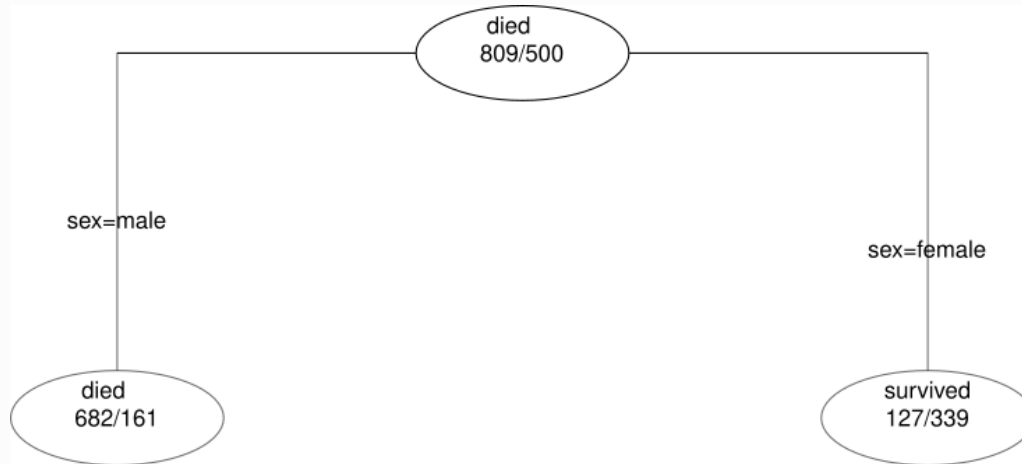
1- Tout en haut c'est la racine de l'arbre qui regroupe donc toutes nos données, c'est le point de départ.



Chaque groupe et sous-groupe d'individus de notre arbre de classification se lit de la même manière. On a d'abord la modalité la plus représentée dans le groupe, suivie du nombre d'individus possédant les différentes modalités de la variable que nous cherchons à prédire.

Décryptage

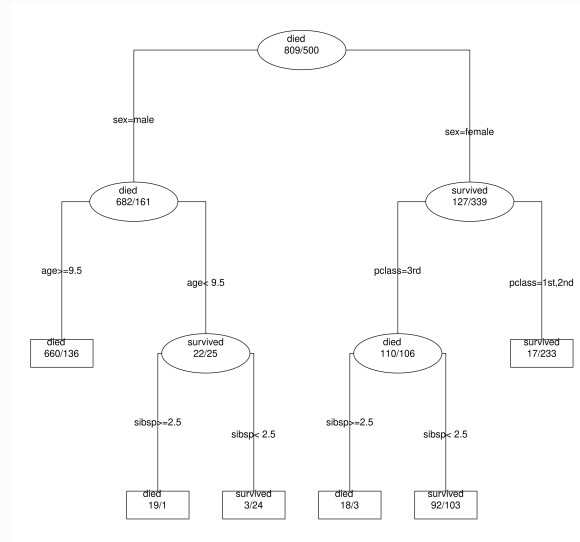
2- Un premier découpage est fait selon la meilleure variable possible pour séparer nos données en deux groupes homogènes.



Ici c'est la variable **sex** qui est d'abord la plus discriminante, les individus satisfaisant la condition "male" iront à gauche et constitueront un sous groupe où la modalité décès est plus importante (avec 682 hommes qui perdront la vie durant le naufrage et 161 qui en réchapperont). Les individus satisfaisant la modalité "female" formeront un autre sous-groupe où la modalité survie est la mieux représentée avec 339 femmes qui survivront et 127 ne périront durant le naufrage.

Décryptage

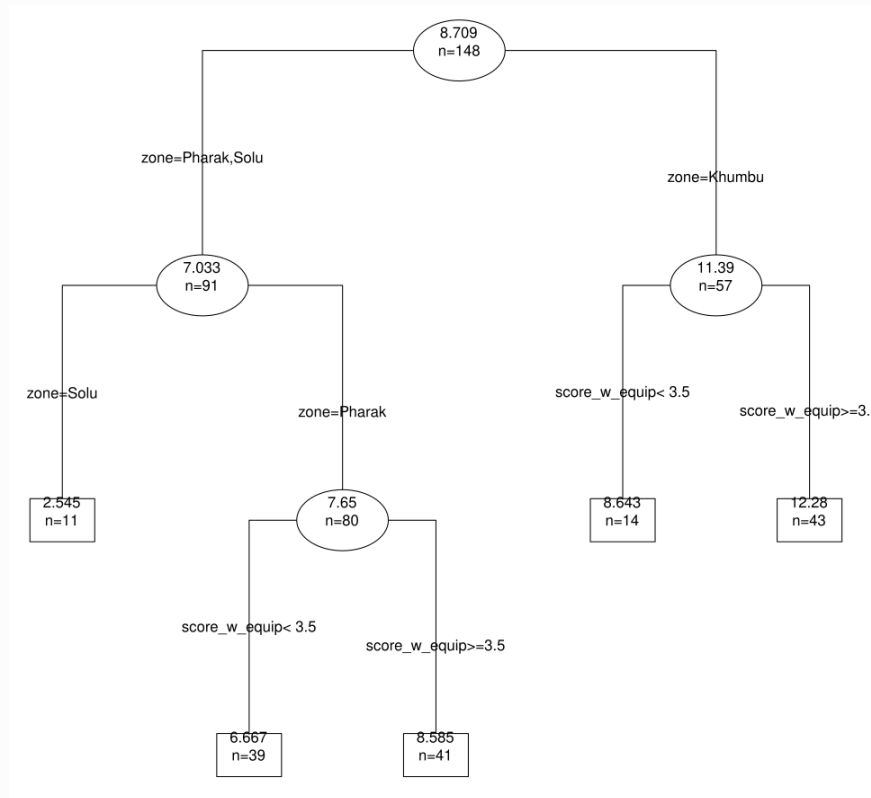
3- De chaque côté le découpage se poursuit de manière récursive jusqu'à ce que l'arbre est atteint sa taille optimale.



Dans cette modalité de représentation nous savons que l'arbre a atteint sa taille optimale quand les feuilles de l'arbre représentant nos sous-groupes d'individus prennent une forme rectangulaire. Ainsi, les individus qui satisfont successivement les critères "sex" = "male" et "age">= "9.5" appartiendront à un sous-groupe ayant d'avantage de chance de ne pas survivre au naufrage avec 660 individus qui sont morts et 136 qui ont réussi à survivre.

Exemple 2 : un arbre de régression

Nous allons maintenant tenter prédire le nombre d'équipements électriques de lodge situé au Népal.



Décryptage

Pour ce faire nous intégrons à notre arbre 4 prédicteurs : la zone du Népal dans laquelle se situe notre lodge, la caste du propriétaire, le genre du propriétaire et le nombre d'équipement raccordés à l'eau.

L'arbre de régression se lit exactement de la même manière qu'un arbre de classification !

En revanche l'information fournie pour chaque sous-groupe d'individu ne va pas être tout à fait identique.

Dans un arbre de régression on obtiendra pour sous groupe le nombre d'individus composant le sous-groupe ($n=$) et la valeur moyenne de la variable que nous cherchons à prédire pour le sous-groupe.

Dans cet exemple on note que les lodges situés dans le Solu seront ceux ayant le nombre d'équipement électrique le plus faible, 2.5 en moyenne et cela correspond à 11 individus.

Comment sont choisis les prédicteurs ?

Cette question du prédicteurs renvoie au découpage en sous-groupe d'individus de notre arbre de décision, à la meilleure question possible.

L'objectif est que chaque découpage réduise l'erreur de prédiction.

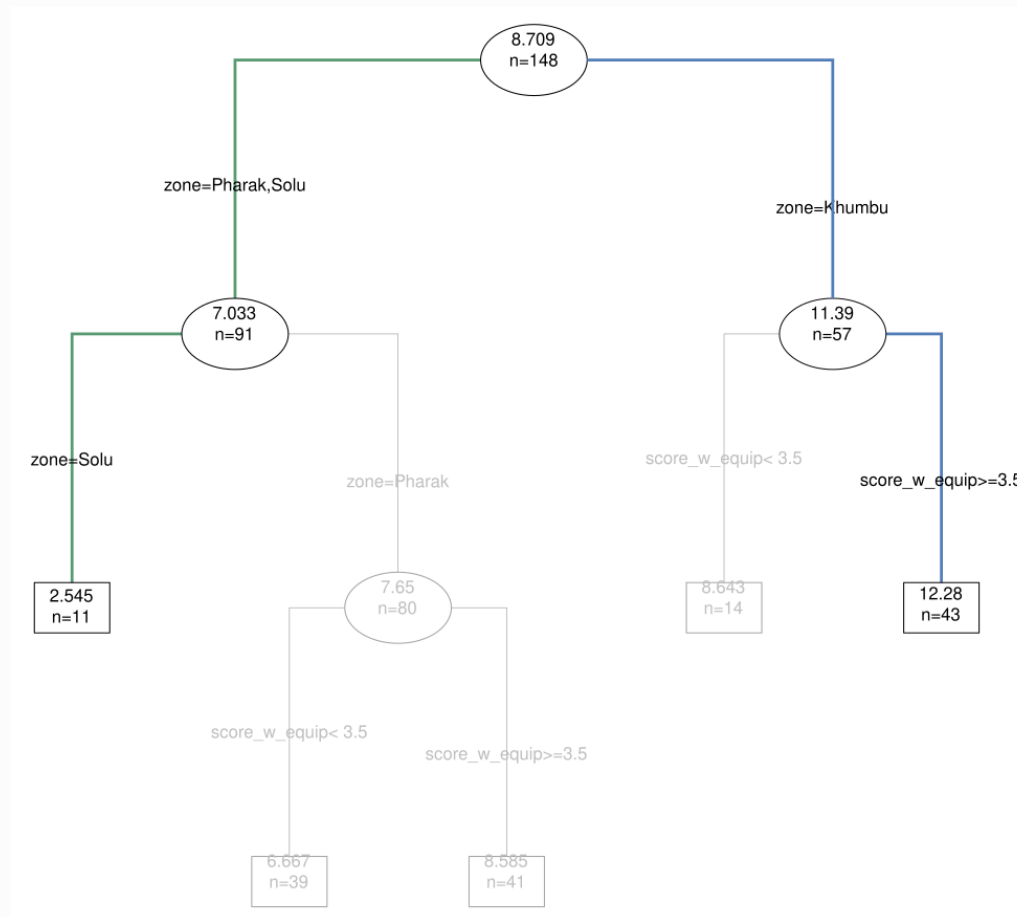
Pour trouver la variable permettant de réduire le plus cette erreur de prédiction l'algorithme va tout simplement essayer tous les découpages possibles.

Concrètement l'algorithme va systématiquement essayer de découper chaque nouveau groupe d'individus à l'aide de chaque prédicteur que vous aurez définis, et retenir uniquement celui permettant de réduire le plus cette erreur de prédiction.

C'est donc l'algorithme de l'arbre qui va choisir le prédicteur à utiliser, parmi les prédicteurs que vous aurez définis, pour découper un groupe d'individus en deux nouveaux sous-groupe

Comment sont choisis les prédicteurs ? II

Reprenons notre arbre de regression sur les équipement électriques des lodges:



Comment sont choisis les prédicteurs ? III

Pour définir le groupe d'individus ayant le nombre d'équipements électrique le plus faible (en vert) seul le prédicteur "zone" suffit.

C'est ce prédicteur qui utilisé 2 fois de suite qui permettra la meilleure réduction de l'erreur et donc de prédire les lodges ayant le score le plus faible

En revanche pour prédire le groupe des individus ayant le nombre d'équipement électrique le plus élevé (en bleu) deux prédicteurs sont nécessaire la zone et le nombre d'équiment raccordé à l'eau.

C'est la combinaison de ces 2 prédicteurs qui me permettra de définir le groupe avec le score le plus élevé.

On notera que ni le genre ni la caste ne sont utilisés comme variable discriminante pour constituer nos sous-groupe.

Quand l'arbre de décision a-t-il atteint sa taille optimale

L'objectif du découpage optimale est à la fois de réduire l'erreur de prédiction tout en évitant un surajustement où chaque feuille terminale serait composée d'un unique individu.

Pour stopper l'arbre il existe 4 critères différents.

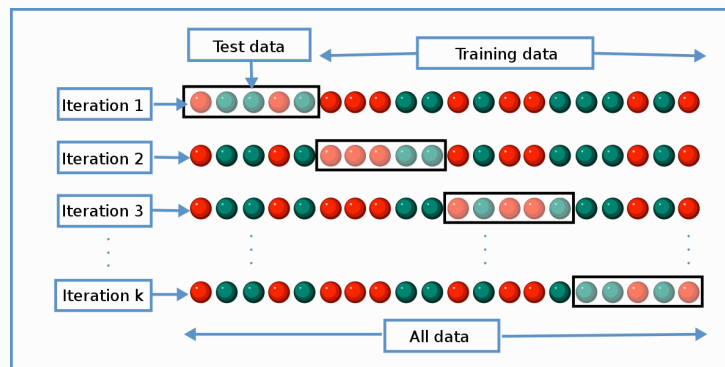
1. Obliger chaque feuille terminale à contenir un nombre minimal d'individus .
2. Ne pas dépasser un nombre de feuilles fixé à l'avance.
3. Retenir le nombre de feuilles permettant de minimiser l'erreur de prédiction à l'aide de la validation croisée (sur un autre jeu de données de validation, ou en leave-one-out cross validation dans les données d'apprentissage).
4. Interrompre le processus lorsqu'une division supplémentaire n'aboutirait pas à une diminution "sensible" de l'erreur de prédiction.

Les deux premiers critères sont relativement proches et il n'y a pas de règles les concernant hormis le bon sens et votre connaissance des données. C'est en général le 3ème critère qui est privilégié. Quant au critère numéro 4, il est quasiment impossible à définir à l'avance avant la réalisation de l'arbre. Nous ne l'aborderons pas ici

Point sur le critère 3 : la validation croisée

La validation croisée (cross-validation en anglais) est sujet complexe que nous aborderons pas ici. Mais ce qu'il faut retenir :

La validation croisée désigne un processus qui permet de tester la qualité de prédiction d'un modèle. Il existe plusieurs méthode mais la plus populaire est la validation croisée à k blocs (k-fold cross-validation en anglais). On va diviser notre échantillons total avec d'un côté une partie qui servira pour entrainer le modèle et une autre partie sur laquelle sera testé le modèle. On fait ça x fois avec des échantillons de tests de même taille mais sélectionner aléatoirement et des échantillons d'entrainement qui devront également avoir la même taille et sélectionné aléatoirement.



Point sur le critère 3 : la validation croisée II

Notre but est donc de réduire le plus possible l'erreur de prédiction en évitant un surajustement qui rendrait notre arbre inefficace à prédire notre variable étudiée sur un échantillon d'individus dont nous connaissons les prédicteurs mais pas la variable étudiée.

C'est ce compromis que permet la validation croisée

Classiquement, l'erreur de prédiction diminue constamment lorsque le nombre de feuilles augmente. Alors que l'erreur de prédiction obtenue en validation croisée va diminuer puis réaugmenter.

C'est à partir de ce seuil que nous définirons le nombre de découpage optimale et donc la taille optimale de l'arbre.

Point sur le critère 3 : la validation croisée II

Ce calcul du seuil de coupure optimal peut également apparaître sous la dénomination de calcul de complexité.

Il existe deux manières pour le visualiser :

- ▶ Avec un tableau
- ▶ Avec un graphique

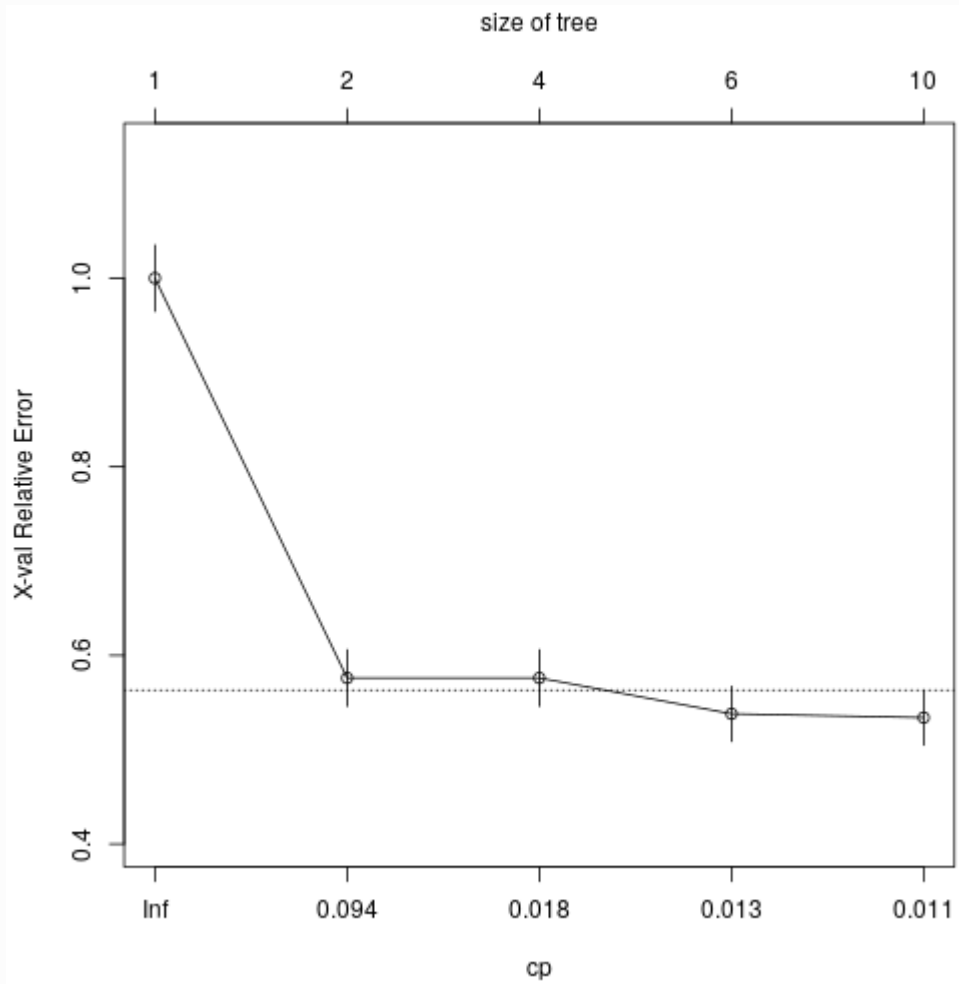
Il existe des formules pour récupérer le seuil exact à partir du tableau ou du graphique

La voie littérale (avec R)

L'erreur de prédiction en validation croisée correspond au **xerror**

```
##
## Classification tree:
## rpart(formula = survived ~ ., data = ptitanic)
##
## Variables actually used in tree construction:
## [1] age      parch  pclass sex      sibsp
##
## Root node error: 500/1309 = 0.38197
##
## n= 1309
##
##          CP nsplit rel error xerror      xstd
## 1 0.424000      0      1.000  1.000 0.035158
## 2 0.021000      1      0.576  0.576 0.029976
## 3 0.015000      3      0.534  0.538 0.029238
## 4 0.011333      5      0.504  0.522 0.028911
## 5 0.010000      9      0.458  0.522 0.028911
```

La voie graphique (avec R)



Quel modèle pour nos sous-groupe d'individus

Cette question du modèle créé pour chaque groupe d'individus renvoie simplement à comment interpréter ces groupes.

Quelle valeur attribuer à chaque groupe d'individus

Le coeur de cette interprétation est lié au principe même de création des arbres de décision :

A savoir l'objectif de subdiviser nos données en sous-groupe les plus homogènes possible en leur sein (concernant la variable étudiée) et différents les uns des autres

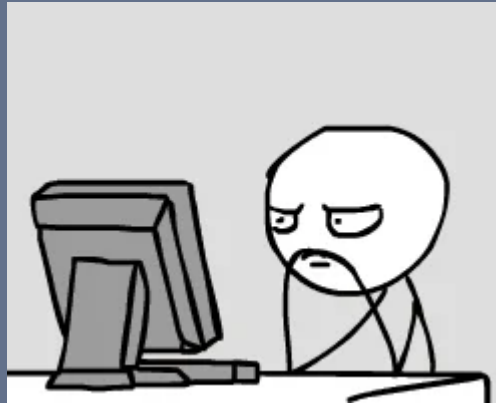
Comme les sous-groupes d'individus sont supposées suffisamment homogènes ont fait très généralement le choix d'un ajustement par une constante:

- La moyenne le plus souvent pour les arbres de régression.
- La modalité la mieux représentée (donc la probabilité d'obtenir cette modalité) pour les arbres de classification.

Ouf ! Finis pour la théorie !



Maintenant la pratique !



BaobARD, une application shiny pour faire des arbres sans coder !

— https://frama.link/BaobARD_01

— https://frama.link/BaobARD_02

En suivant le lien vous arriverez sur cette première page qui donne quelques informations sommaires sur l'application et sur les arbres de décision.

BAOBARD GRÉGOIRE LE CAMPION

A propos

Import des données

Arbres de décision

Bienvenue sur l'application BaobARD une appli qui vous permet de créer votre propre Arbre de décision !

L'**application BaobARD** vous permet de créer votre propre arbre de régression ou de classification à partir d'un simple fichier csv !

Si vous n'avez aucune idée de ce que peuvent être des arbres de régression ou de classification ou pourquoi les utiliser un très bref rappel sera effectué juste ci dessous. Mais je vous conseille fortement de vous rendre sur le tuto et la page Arbre de décision sur le site web **OUVRIR** .

Outre cette première page, cette application contient 2 autres onglets.

Le premier vous permet de charger et visualiser les données que vous souhaitez analyser. Vous pouvez importer uniquement des fichiers au format csv.

Le deuxième onglet vous permettra de construire et visualiser un arbre de décision. Il vous reviendra de déterminer la variable que vous souhaitez prédire et les variables explicatives. Une fois satisfait vous pourrez télécharger votre visualisation aux formats png, pdf ou svg.

Pourquoi un arbre de décision ?

Parce que vous voulez enfin pouvoir faire plus qu'une corrélation ! Parce que la liste des pré-requis pour faire une régression ou tout autre test statistique cherchant à étudier la causalité est plus longue que le bras ! Et surtout parce que jamais vos données n'ont pu remplir une seule de ces conditions !

L'arbre de décision est **un outil tout terrain** parfaitement adapté aux données en SHS.

C'est un outil statistique d'exploration des données et de prédiction. Il peut servir à expliquer aussi bien une variable qualitative (on parle alors d'arbre de classification) qu'une variable quantitative (arbre de régression). Par rapport à d'autres méthodes classiques (analyse factorielle, régression, réseau de neurones...) l'arbre de décision possède de nombreux avantages :

- les données d'entrée peuvent être « mixtes », c'est-à-dire qu'un même arbre peut utiliser simultanément des variables prédictives qualitatives, ordinales et continues ;
- la gestion des données manquantes est particulièrement efficace ;
- la construction d'un arbre est rapide et peu gourmande en ressources ;
- l'arbre de décision est relativement simple à interpréter.

La page import données vous permet de charger vos données et de les visualiser.

BAOBARD GRÉGOIRE LE CAMPION

A propos

Import des données

Arbres de décision

Charger un fichier CSV

Browse...

penguins.csv

Upload complete

Le poids des fichiers est limité à 30Mb

Charger des fichiers au format UTF-8

Ajuster les options suivantes en fonction de votre fichier Importé

☒ 1ere ligne comme en-tête

Séparateur de champ

☒ Comma

☐ Semicolon

☐ Tab

Séparateur de texte

☐ Aucun

☒ Guillemet double

☐ Guillemet simple

Visualiser

☒ Uniquement les 1eres lignes

☐ Ensemble des données

X	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.10	18.70	181	3750	male	2007
2	Adelie	Torgersen	39.50	17.40	186	3800	female	2007
3	Adelie	Torgersen	40.30	18.00	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.70	19.30	193	3450	female	2007
6	Adelie	Torgersen	39.30	20.60	190	3650	male	2007

C'est donc ici que les analyses vont se faire, diverses options sont disponibles. Mais surtout le plus important c'est là qu'il faudra choisir la variable à étudier et les variables qui feront office de prédicteurs.

Variable à prédire
body_mass_g

Variables prédictives
species island bill_length_mm flipper_length_mm

X
bill_depth_mm
sex
year

Nombre minimal de cas requis dans une feuille terminale
5

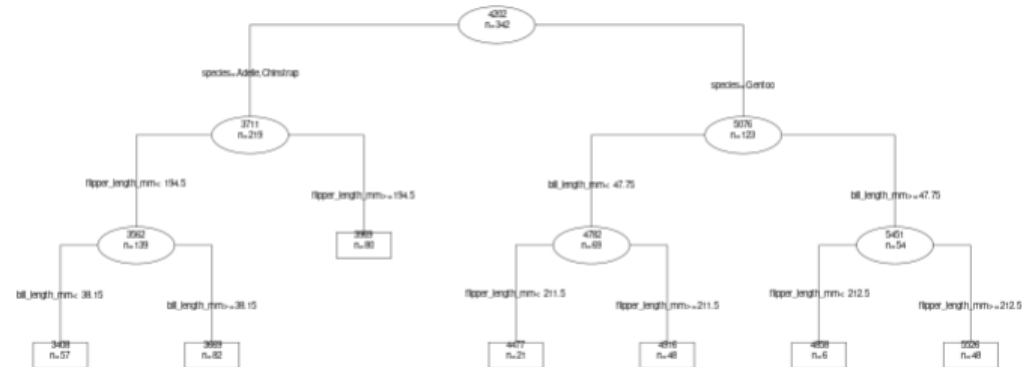
Nombre de divisions pour la validation croisée
0 100

Paramètres d'affichage du graphique

Type de graphique
☒ Graphique classique
☐ Graphique avec couleurs et pourcentages

☒ Espacement vertical uniforme
☒ Afficher les décomptes dans chaque noeud
☒ Distinguer les feuilles terminales
☒ Légender tous les noeuds

Affichage de l'arbre



Quel format d'image :

☒ PDF ☐ PNG ☐ SVG

Télécharger Arbre

Le 1er onglet est "Arbre brut", il s'agit de la première visualisation brute de votre arbre, c'est à dire avec les paramètres par défaut. Un export en png, pdf et svg est possible.

BAOBARD GRÉGOIRE LE CAMPION A propos Import des données Arbres de décision

Variable à prédire
body_mass_g

Variables prédictives
species island bill_length_mm flipper_length_mm

X
bill_depth_mm
sex
year

Nombre minimal de cas requis dans une feuille terminale
5

Nombre de divisions pour la validation croisée
0 100

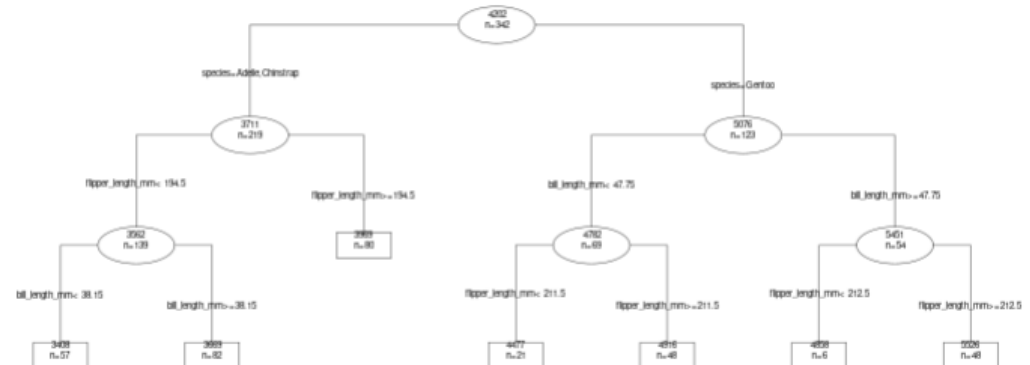
Paramètres d'affichage du graphique

Type de graphique
☒ Graphique classique
☐ Graphique avec couleurs et pourcentages

☒ Espacement vertical uniforme
☒ Afficher les décomptes dans chaque nœud
☒ Distinguer les feuilles terminales
☒ Légender tous les nœuds

1- Arbre brut 2- Complexité 3- Arbre élagué 4- Règles de construction 5- Arbre interactif

Affichage de l'arbre



Quel format d'image :

☒ PDF ☐ PNG ☐ SVG

Télécharger Arbre

Le 2ème onglet "Complexité", permet de visualiser le graph de l'erreur de prédiction en validation croisée. Baobard vous fournira le niveau de coupure optimale de l'arbre issue du calcul de complexité.

BAOBARD GREGOIRE LE CAMPION A propos Import des données Arbres de décision

Variable à prédire

body_mass_g

Variables prédictives

species island bill_length_mm flipper_length_mm

Paramètres de l'analyse

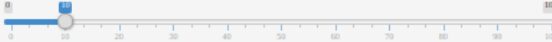
Nombre minimal de cas requis dans un noeud pour tenter un split

5

Nombre minimal de cas requis dans une feuille terminale

5

Nombre de divisions pour la validation croisée

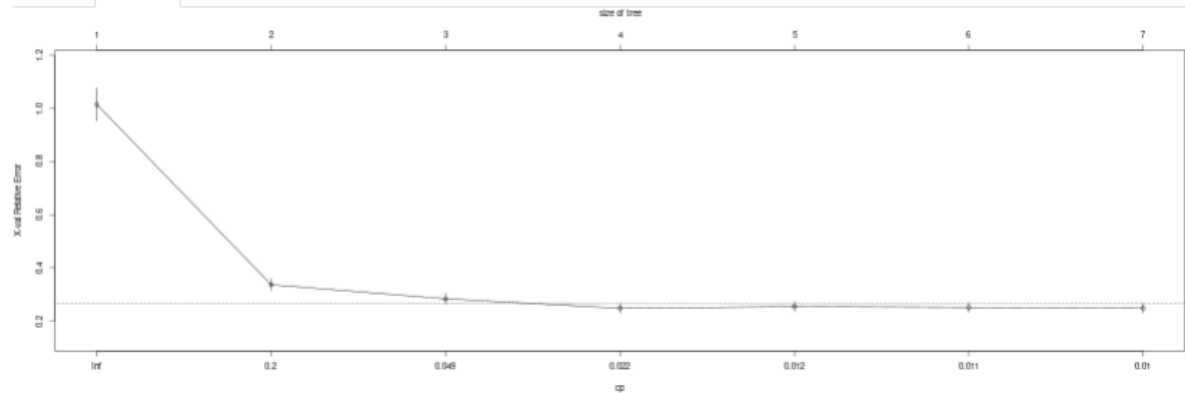


Paramètres d'affichage du graphique

Type de graphique

- ☒ Graphique classique
- ☐ Graphique avec couleurs et pourcentages
- ☒ Espacement vertical uniforme
- ☒ Afficher les décomptes dans chaque noeud
- ☒ Distinguer les feuilles terminales
- ☒ Légender tous les noeuds

1- Arbre brut 2- Complexité 3- Arbre élagué 4- Règles de construction 5- Arbre interactif

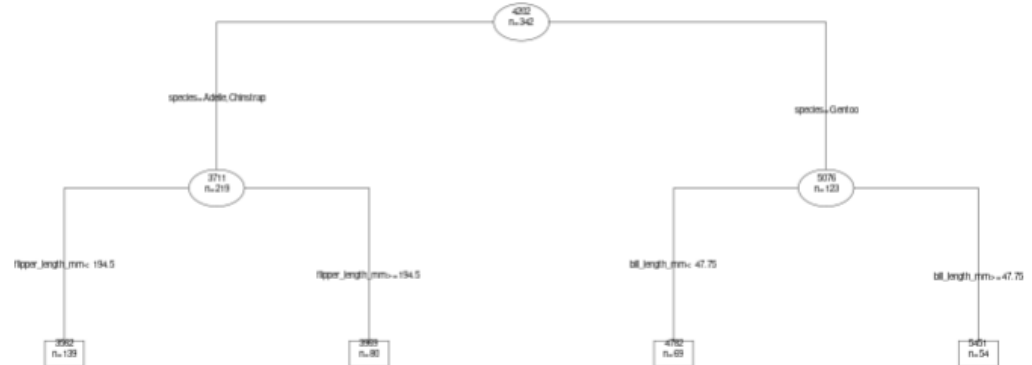


Affichage du cp optimal, c'est à dire le niveau d'élagage idéal !

[1] 0.01279351

Le 3ème onglet "Arbre élagué", permet de visualiser votre arbre à sa taille optimale à l'aide du niveau de "coupe" obtenu dans l'onglet "Complexité". Il se peut que l'arbre élagué est la même forme que l'arbre brut. Un export au format pdf, png et svg est possible.

Affichage de l'arbre élagué



Quel format d'image :

☒ PDF ☐ PNG ☐ SVG

Télécharger Arbre

Le 4ème onglet "Règle de construction", permet une lecture littérale de votre arbre.

Variable à prédire

body_max_g

Variables prédictives

species island bill_length_mm flipper_length_mm

Paramètres de l'analyse

Nombre minimal de cas requis dans un noeud pour tenter un split

5

Nombre minimal de cas requis dans une feuille terminale

5

Nombre de divisions pour la validation croisée



Paramètres d'affichage du graphique

Type de graphique

- ☒ Graphique classique
- ☐ Graphique avec couleurs et pourcentages

- ☒ Espacement vertical uniforme
- ☒ Afficher les décomptes dans chaque noeud
- ☒ Distinguer les feuilles terminales
- ☒ Légender tous les noeuds

Vous trouverez ici les règles de décision et construction de l'arbre élagué

```
n=342 (2 observations deleted due to missingness)

node), split, n, deviance, yval
* denotes terminal node

1) root 342 219387700 4201.754
 2) species=Adelie,Chinstrap 219 41488530 3710.731
 4) flipper_length_mm< 194.5 139 20571660 3561.871 *
 5) flipper_length_mm>=194.5 80 12484970 3969.375 *
 3) species=Gentoo 123 31004250 5076.016
 6) bill_length_mm< 47.75 69 9126377 4782.246 *
 7) bill_length_mm>=47.75 54 8314271 5451.389 *
```

Chaque ligne correspond à un noeud ou à une feuille de l'arbre. On commence par la racine (1) qui contient les n individus de notre population. A l'étape suivante l'arbre découpe la population en fonction de la variable déterminée et crée les noeuds 2) et 3) et ainsi de suite jusqu'à arriver aux feuilles terminales.

Le 5ème onglet "Arbre interactif", permet de visualiser votre arbre élagué sous forme interactive, ce qui permet éventuellement de faciliter la lecture et l'analyse de votre arbre.

Variable à prédire

body_mass_g

Variables prédictives

species island bill_length_mm flipper_length_mm

Paramètres de l'analyse

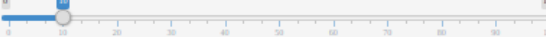
Nombre minimal de cas requis dans un nœud pour tenter un split

5

Nombre minimal de cas requis dans une feuille terminale

5

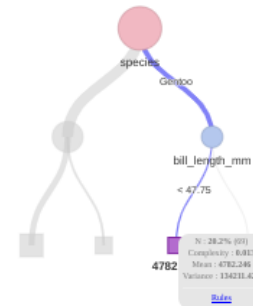
Nombre de divisions pour la validation croisée



Paramètres d'affichage du graphique

Type de graphique

- ☒ Graphique classique
- ☐ Graphique avec couleurs et pourcentages
- ☒ Espacement vertical uniforme
- ☒ Afficher les décomptes dans chaque nœud
- ☒ Distinguer les feuilles terminales
- ☒ Légender tous les nœuds



Un peu de R ?

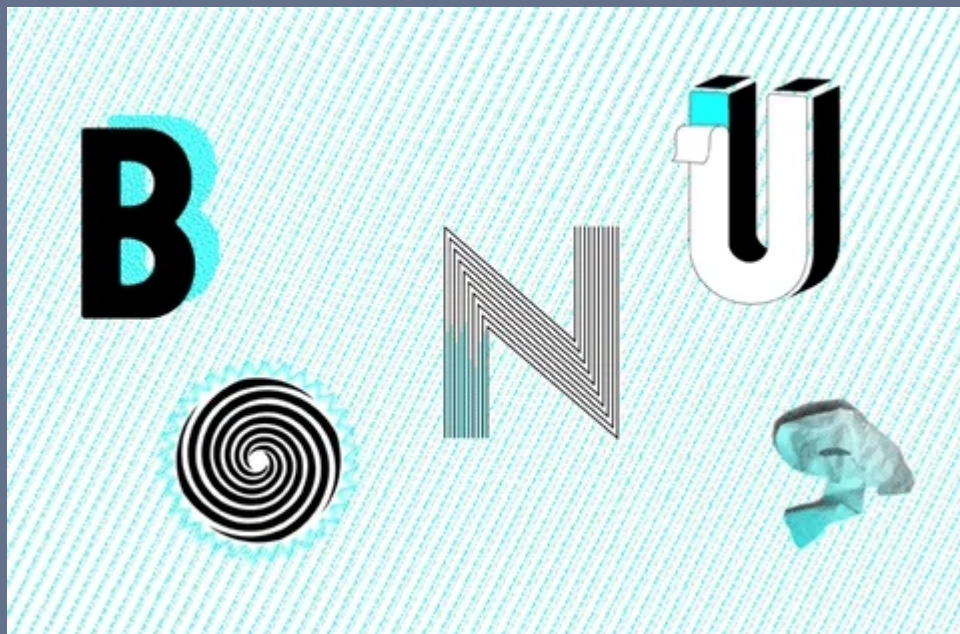


Un peu de code

Pour les aventuriers qui veulent se lancer sur R, voici les lignes de codes qui traduisent au minimum ce dont nous venons de parler.

```
install.packages(c("rpart", "rpart.plot")) # pour installer les de  
library(rpart) #charger la librairie dans R  
library(rpart.plot)  
  
data(ptitanic) #charger les données exemple sur le Titanic (conten  
View(ptitanic) #visualiser les données  
  
arbre1 <- rpart(survived ~ sex + age + pclass + sibsp + parch, dat  
plot(arbre1, compress=TRUE, margin=0.09, uniform=TRUE) #produit le  
text(arbre1, pretty=1, fancy=TRUE, all=TRUE, use.n=TRUE) # rajoute
```

Un petit bonus



Survol très superficiel des forêts aléatoires

Les forêts aléatoires (ou random forest) sont une extension des arbres de décisions visant à être encore plus efficace sur de la prédiction.

Le principe sous jacent est plutôt simple, une multitude de modèle faible et simple une fois combinés formeront un modèle robuste!

On est dans la famille des algorithmes qui font de l'aggrégation de modèle, l'algorithme va construire une “forêt” d'arbre de décision, c'est à dire plusieurs centaines voire milliers, construit de manière aléatoire. Finalement absolument tout est dans le nom!

Survol très superficiel des forêts aléatoires II

La partie aléatoire du random forest concerne la construction de chaque arbre de décision de notre forêt.

Pour chaque arbre conçu dans notre modèle un échantillon aléatoire d'individus est sélectionné et la construction d'un sous-groupe se fait sur un sous-ensemble de variables lui aussi sélectionné aléatoirement.

Une fois le modèle élaboré, tous les arbres de décision vont tourner sur les données. Et on aura donc des prédictions différentes pour chaque individu !

L'estimation finale sera :

- ▶ La modalité la plus fréquente si le modèle est basé sur des arbres de classification.
- ▶ La moyenne des valeurs prédites si le modèle est basé sur des arbres de régression.