Exploration des données à l'aide des arbres de décisions

Ecole thématique Explo-SHS

Grégoire Le Campion / Hugues Pécout

UMR 5319 Passages

fédération de recherche CIST

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

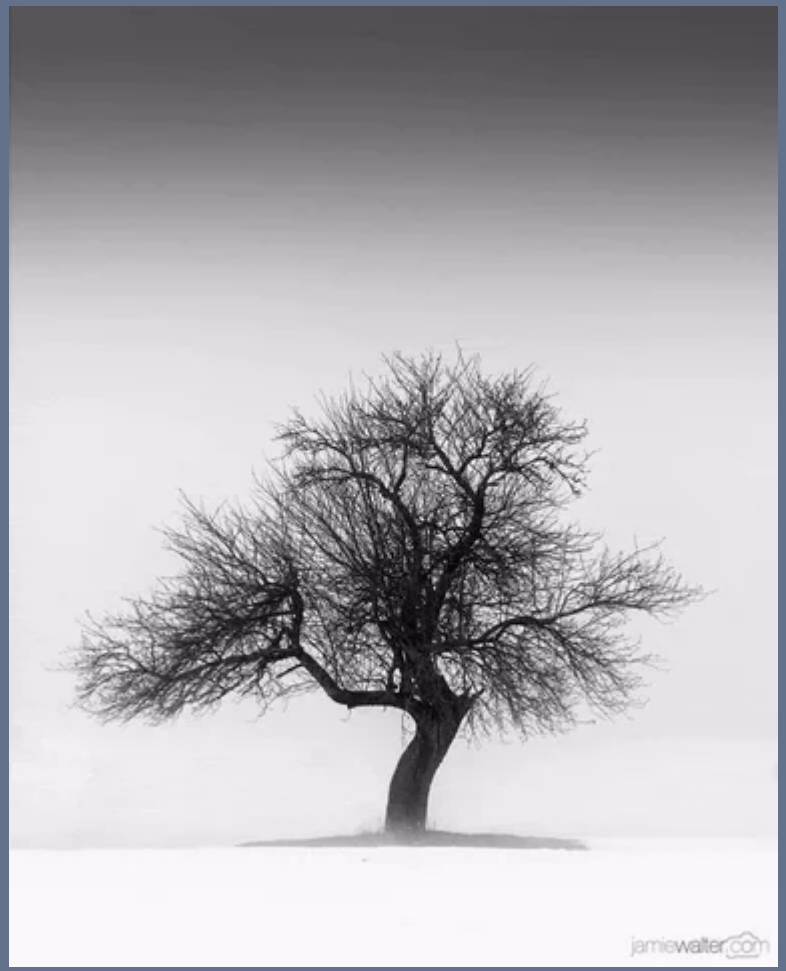
II- Exemple et décryptage d'un arbre

III-BaobARD : application shiny pour réaliser ses premières explorations

IV- Travaux pratiques

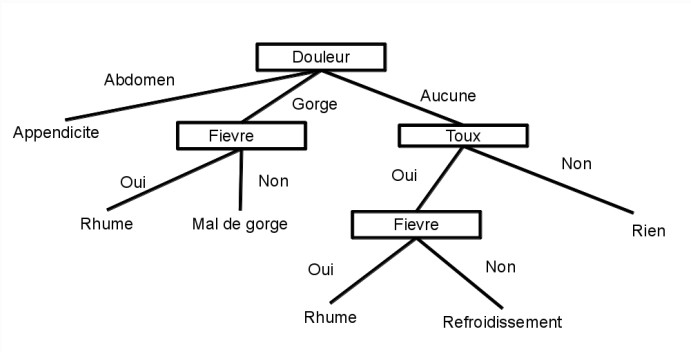
V- Un peu de code... éventuellement

Présentation des arbres de décision



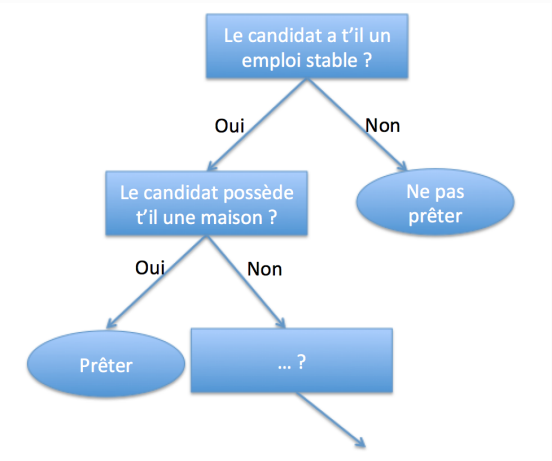
Les arbres de décision c'est quoi ? !

ça

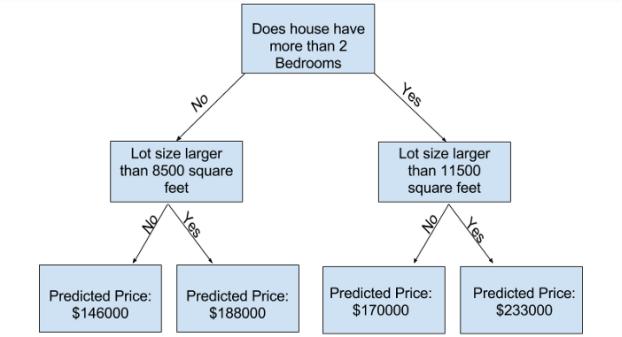


ou ça

ça aussi



ah oui et encore ça...



Les arbres de décision c'est quoi ? I

Les arbres de décision font partie du champs très à la mode du machine learning.

> Très simplement, le machine learning à pour objectif de chercher des schémas dans les données et d'effectuer des prédictions 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 partie 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 poids, le sexe et la taille, je peux mettre en oeuvre un modéle d'apprentissage supervisé pour prédire le poids 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 prisent par une variable, que vous aurez choisie, en fonction d'un ensemble d'autres variables que vous aurez sélectionnées.

► 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 prisent par une variable. Mais il existe d'autres tests statistiques plus connus qui ont le même objectif, comme notamment les tests de régression.

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

Pourquoi alors ne pas utiliser simplement les techniques de régression ?

> Notamment car ces méthodes sont assorties de conditions qu'il faut remplir pour être considérées comme fiable :

Par exemple : distribution normale des résidus, homogénéité des résidus, non multicolinérarité...

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

Les avantages de l'arbre de décision

1. L'arbre de décision ne se préoccupe pas des conditions citées précédemment.
2. Il peut servir à expliquer une variable qualitative (il s'agit d'un arbre de classification) aussi bien que quantitative (il s'agit d'un arbre de régression).
3. Il peut s'inscrire dans une approche mixte, on peut mélanger dans nos prédicteurs des variables continues, ordinales et binaires.
4. 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 séléctionner les meilleurs prédicteurs possibles pour expliquer notre variable.
5. Gestion efficace des données manquantes qui, même en assez grand nombre, ne posent pas de problèmes majeurs.
6. L'algorithme de production n'est pas gourmand en ressources, peu de risque de faire planter l'ordinateur.
7. 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 (grands échantillons, 0 données manquantes, prédicteurs 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.

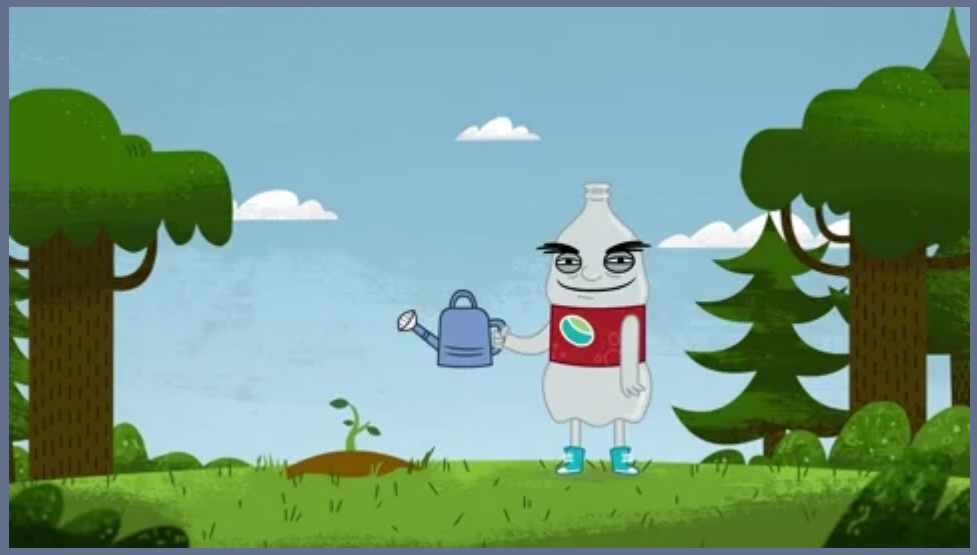
► Vous n'aurez pas d'indicateurs ou de coefficients vous donnant l'importance ou le niveau des effets de tel ou tel prédicteur.

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.

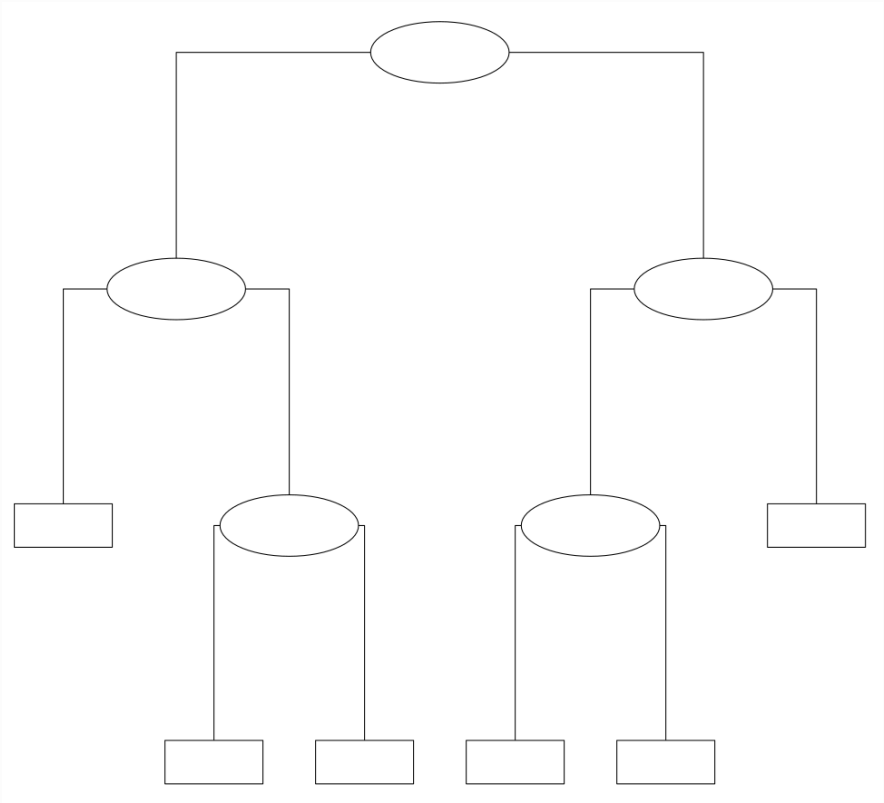
Nous l'avons évoqué précédemment l'algorithme de l'arbre de décision va choisir les meilleurs prédicteurs possibles parmi l'ensemble des variables prédictrices que vous aurez choisies.

Le corollaire est qu'il est possible que certains des prédicteurs que vous avez sélectionnés ne soient pas retenus par l'algorithme de l'arbre.

Petit traité d'arboriculture II

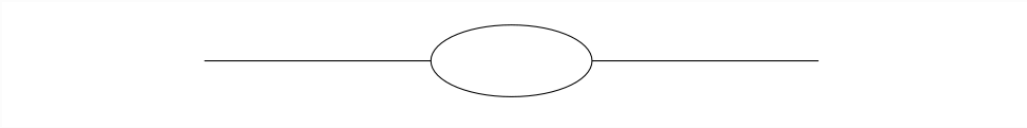
Pour y voir plus clair prenons un exemple et segmentons le processus de "pousse" en plusieurs étapes :

Voici un arbre de décision vierge d'indications



Etape 1

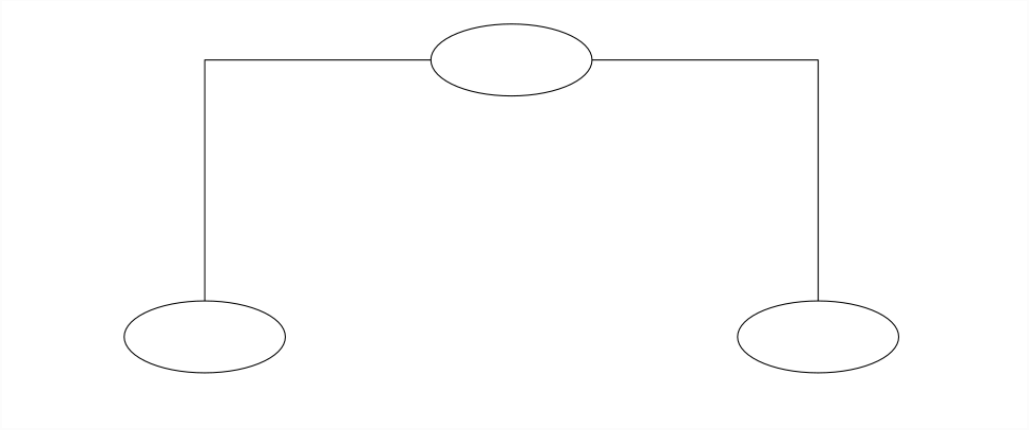
1- L'arbre pousse à partir de sa base constituée de l'ensemble des individus qui composent votre jeu de données. On parle de la racine de l'arbre.



Etape 2

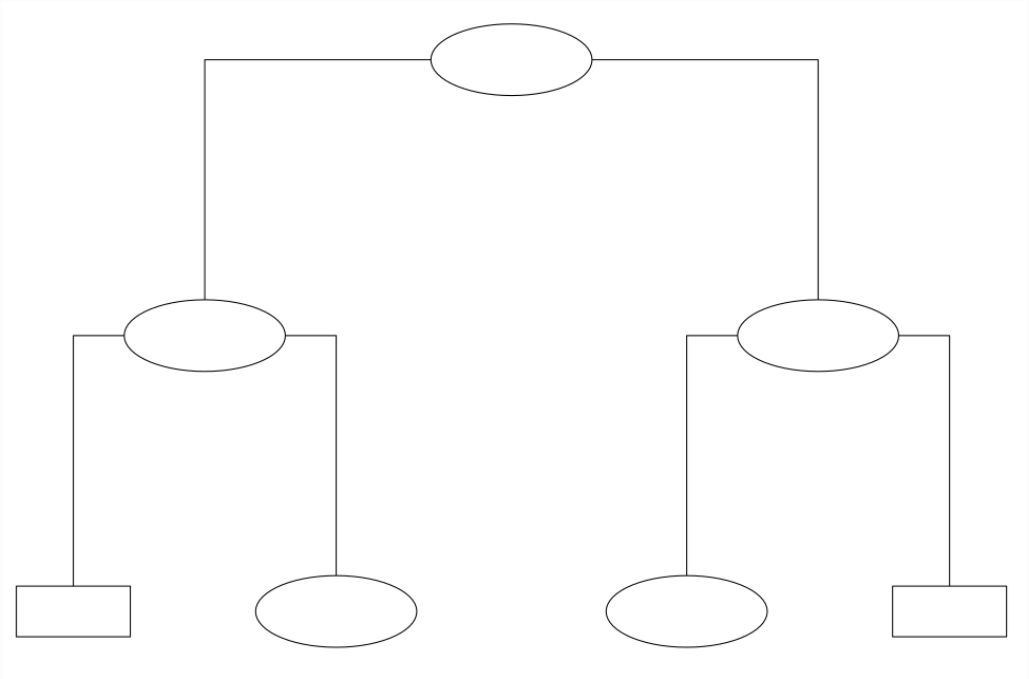
2- Une première question relative à l'une des variables prédictives sépare notre jeu de données en deux groupes.

> 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 possibles en leur sein et différents l'un de l'autre.



Etape 3

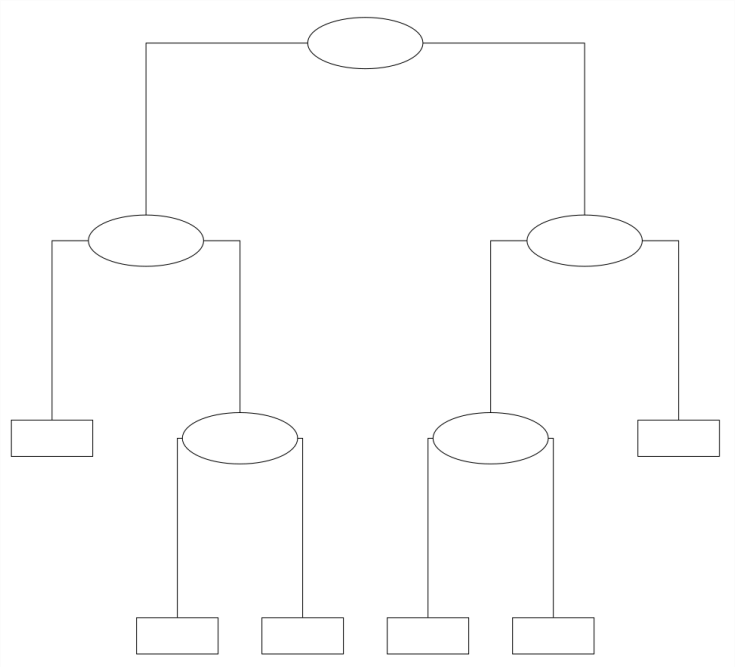
3- Chacun des deux groupes obtenus peut à son tour être séparé en deux en choisissant à nouveau la meilleure question possible à poser à l'aide du meilleur prédicteur possible. La question va varier en fonction des sous-groupes d'individus.



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 suffisament 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

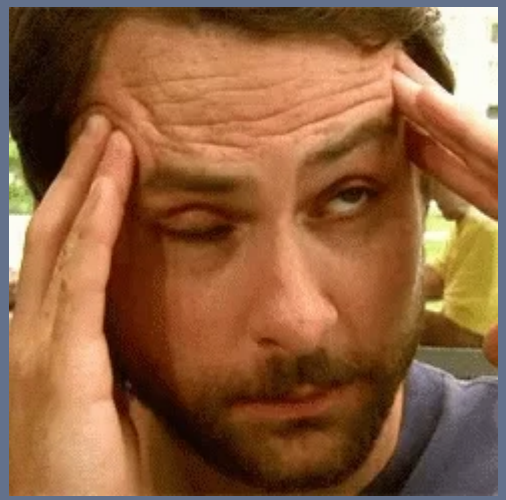
Il reste encore cependant à répondre à 3 questions fondamentales !

► Comment sont choisies les meilleures questions possibles pour constituer nos branches de l'arbre ? Et donc l'utilisation de tel ou tel prédicteur.

► Comment est décidée la taille optimale de l'arbre ?

► Comment interprète 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, nous chercherons à prédire la survie ou non d'un échantillon de passagers du Titanic.

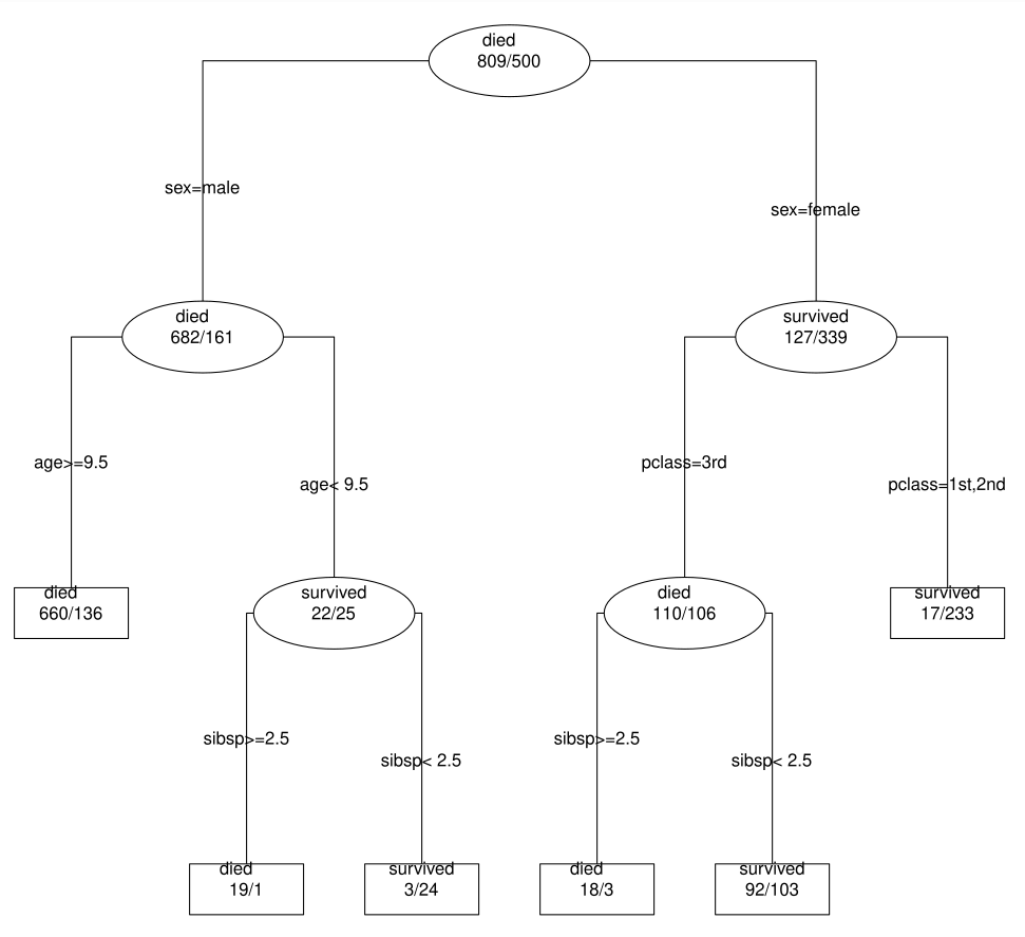
Les variables de notre jeu de données: la classe économique de la cabine (pclass), la survie au naufrage (survived), le genre (sex), l'âge (age), le nombre de frères/soeurs/conjoints à bord (sibsp) et le nombre de parents/enfants à bord (parch).

Voici un aperçu des données:

	pclass	survived	sex	age	sibsp	parch
1	1st	survived	female	29	0	0
2	1st	survived	male	0.916700006	1	2
3	1st	died	female	2	1	2
4	1st	died	male	30	1	2
5	1st	died	female	25	1	2
6	1st	survived	male	48	0	0

Exemple 1 : un arbre de classification

Voici à quoi ressemble notre arbre de classification cherchant à prédire la survie des passagers :



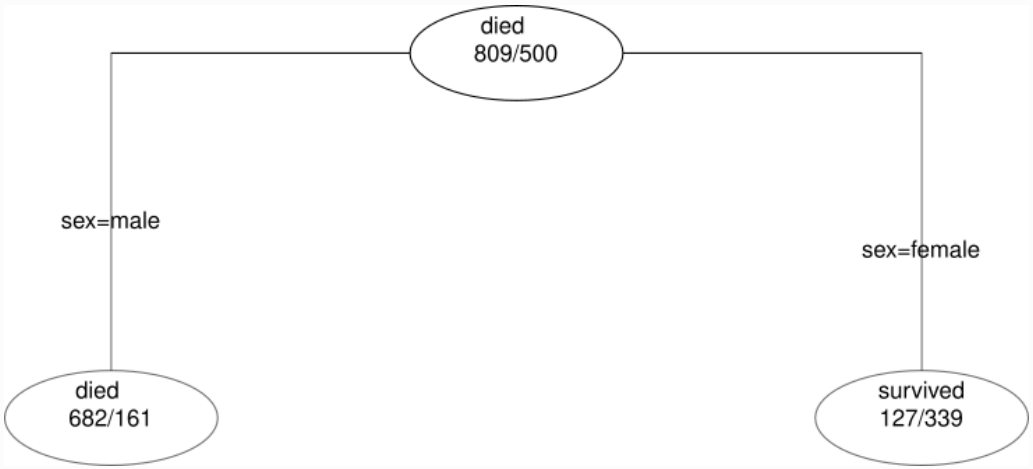
Décryptage

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

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

Décryptage

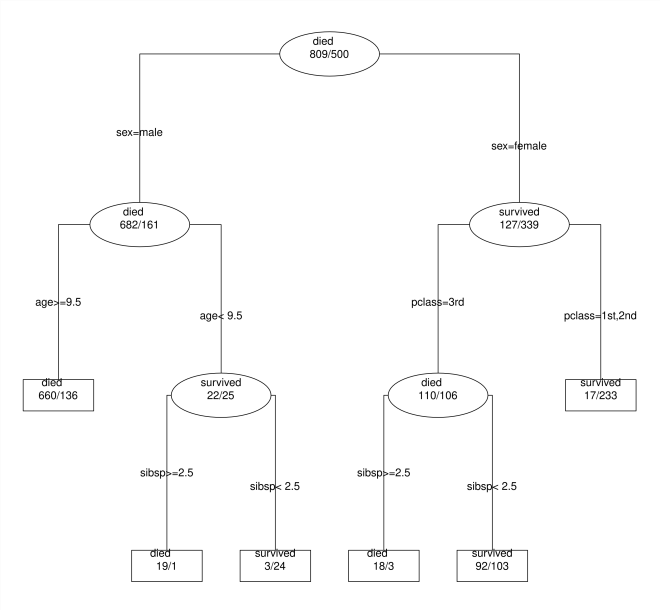
2- Un premier découpage est fait selon la meilleure variable possible pour séparer nos données en deux groupes homogènes : ici la variable "sex".



Les individus satisfaisant la condition "male" formeront la ramification de gauche et constitueront un sous groupe ou la modalité décès est plus importante (avec 682 hommes qui perdront la vie durant le naufrage et 161 qui en réchapperont). Les individus satisfisant la modalité "female" formeront un autre sous-groupe ou la modalité survie est la mieux représentée avec 339 femmes qui survivront et 127 périront durant le naufrage.

Décryptage

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



Ainsi, les individus qui satisfont successivement les critères "sex" = "male" et "age">= "9.5 appartiendront à un sous-groupe ayant davantage de chance de ne pas survivre au naufrage avec 660 individus qui sont morts et 136 qui ont réussi à survivre. On note que les ramifications d'un arbre de décision peuvent être de longueurs différentes

Exemple 2 : un arbre de régression

Nouvel exemple : prédire le nombre d'équipements électriques de lodges (hébergements touristiques) situés au Népal.

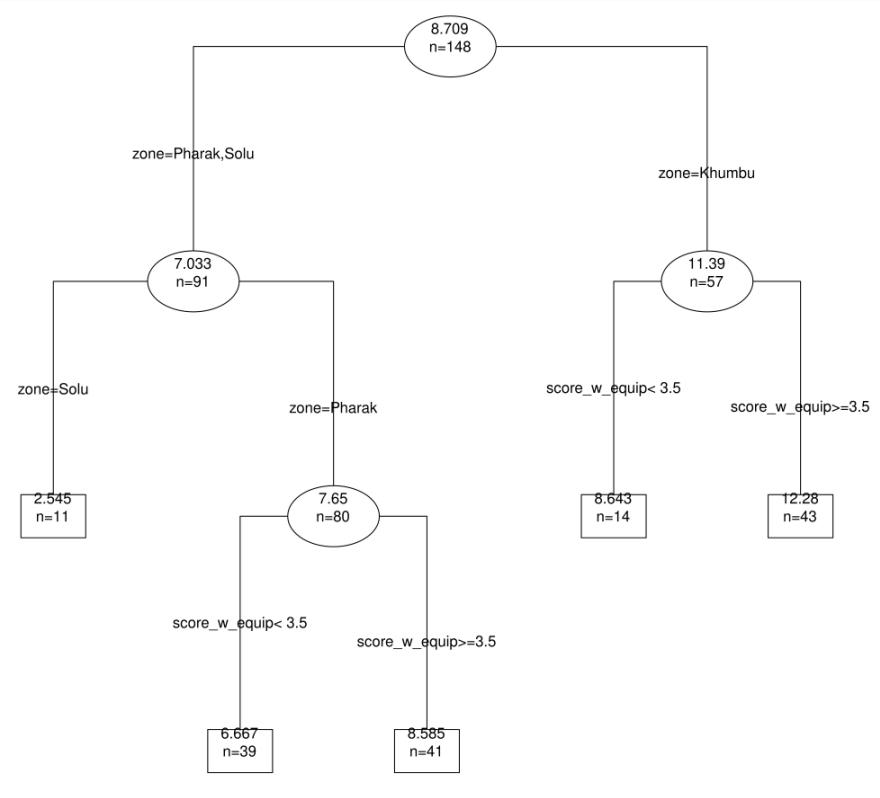
Nos variables : la zone du Népal (zone), la caste du propriétaire (caste), le genre du propriétaire (gender), le nombre d'équipements raccordés à l'eau (score_w_equip) et le nombre d'équipement électrique (score_e_equip).

Voici un aperçu des données :

	zone	caste	gender	score_w_equip	score_e_equip
1	Khumbu	Sherpa	female	4	7
2	Khumbu	Sherpa	female	4	16
3	Khumbu	Rai	female	4	10
4	Khumbu	Sherpa	male	4	14
5	Khumbu	Sherpa	female	4	13
6	Khumbu	Sherpa	male	4	12

Exemple 2 : un arbre de régression

Voici l'arbre de régression :



Décryptage

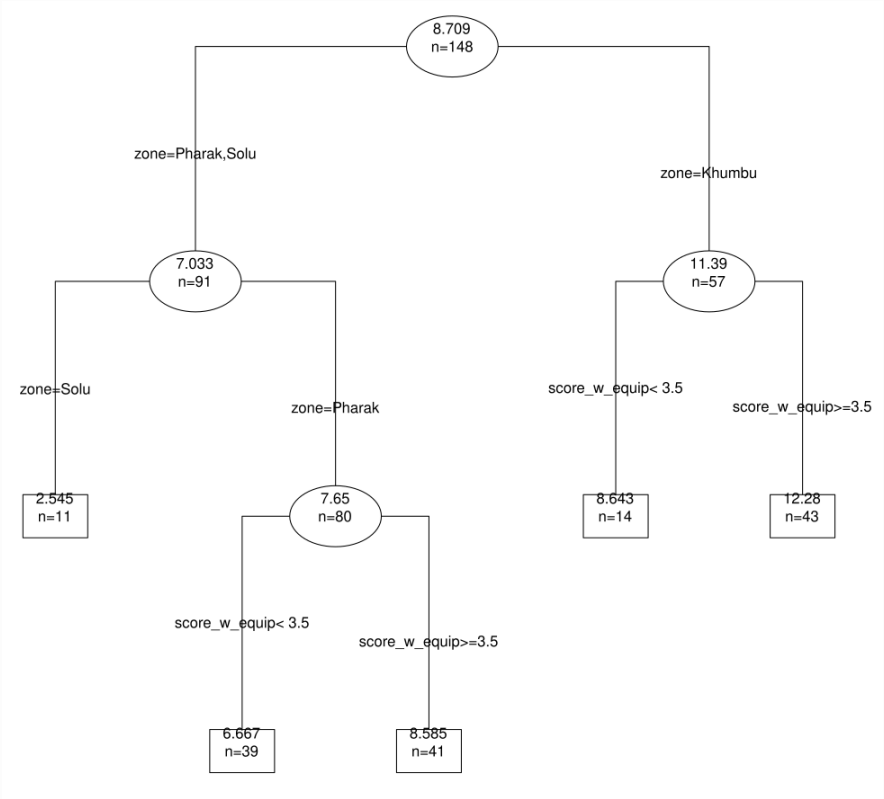
> L'arbre de régression se lit exactement de la même manière qu'un arbre de classification !

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 chaque 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.

Décryptage

Dans cette exemple on note que les lodges situés dans le Solu seront ceux ayant le nombre d'équipements électriques le plus faible, 2.5 en moyenne et cela correspond à 11 individus. Alors que ceux du Pharak avec un score d'équipements raccordés à l'eau supérieur ou égal à 3.5, auront un nombre moyen d'équipements électriques de 8.5



Quel modèle pour nos sous-groupes 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 l'interprétation est lié au principe même de création des arbres de décision : subdiviser nos données en sous-groupe les plus homogènes possibles 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.

► Pour les arbres de régression ce sera le plus souvent la moyenne de la variable que nous cherchons à prédire.

► Pour les arbres de classification, le mode (c'est à dire la modalité la mieux représentée) de la variable que nous cherchons à prédire.

Comment sont choisis les prédicteurs ?

Cette question du prédicteur renvoie au découpage en sous-groupe d'individus de notre arbre de décision.

L'objectif de chaque découpage est de réduire l'erreur de prédiction de l'arbre de décision.

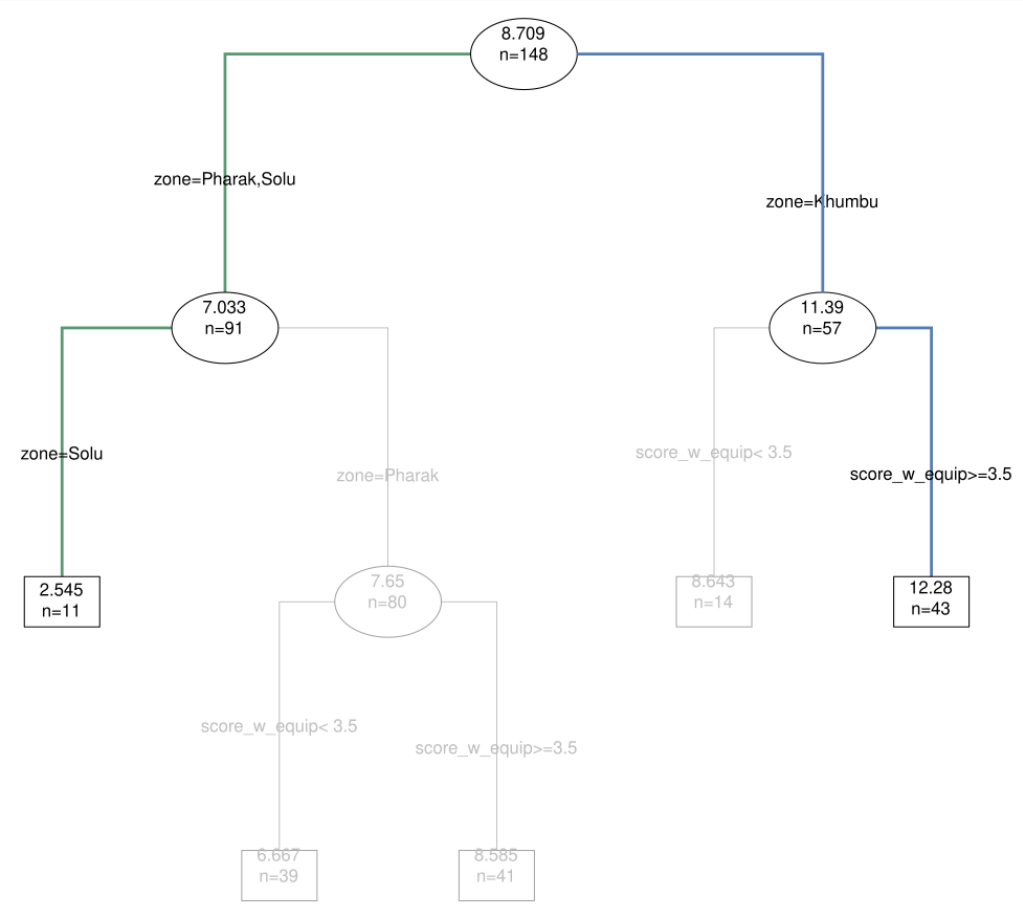
Ainsi, l'algorithme va chercher le prédicteur qui permet de réduire le plus cette erreur de prédiction

> Pour ce faire l'algorithme va tout simplement essayer tous les découpages possibles avec tous les prédicteurs.

C'est donc l'algorithme de l'arbre qui va choisir le prédicteur à utiliser, parmi la liste des prédicteurs, pour découper un groupe d'individus en deux nouveaux sous-groupes

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 électriques le plus faible (en vert) l'algorithme va chercher à scinder nos individus en sous-groupes.

Pour ce faire il va tester toutes nos variables prédictrices et ne retenir que celles étant les plus efficaces pour réduire l'erreur de prédiction et constituer les groupes les plus homogènes.

> Pour prédire les lodges ayant le score le plus faible c'est la variable zone qui utilisée 2 fois de suite permettra la meilleure réduction de l'erreur de prédiction.

Ce principe s'applique pour chaque ramification de l'arbre de décision.

> Pour prédire le groupe des individus ayant le nombre d'équipements électriques le plus élevé (en bleu) la combinaison de deux prédicteurs sera nécessaire : la zone et le nombre d'équipement raccordé à l'eau.

On notera que ni le genre ni la caste ne sont utilisés comme variables discriminantes pour constituer nos sous-groupes.

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

> L'objectif du découpage optimal est à la fois de réduire l'erreur de prédiction tout en évitant un surajustement où chaque groupe d'individus est composée d'un unique individu.

Pour stopper l'arbre il existe 4 critères.

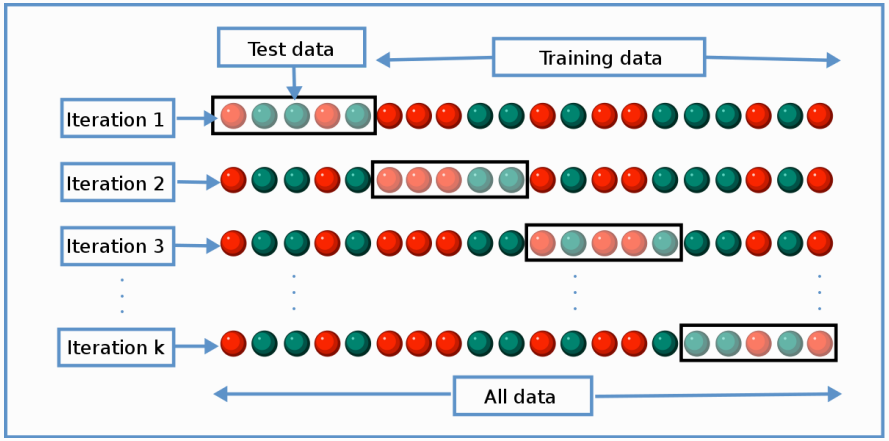
1. Obliger chaque groupe terminal à contenir un nombre minimal d'individus .
2. Ne pas dépasser un nombre de groupes fixé à l'avance.
3. Retenir le nombre de groupes permettant de minimiser l'erreur de prédiction à l'aide de la validation croisée.
4. Interrompre le processus lorsqu'une division supplémentaire n'aboutit 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 definir 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 un sujet complexe que nous ne ferons que survoler ici. 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 méthodes mais la plus populaire est la validation croisée à k blocs (k-fold cross-validation en anglais).Son principe : on divise notre échantillon avec d'un côté une partie qui servira pour entraîner le modèle et une autre partie sur laquelle sera testé le modèle. On répète cette opération x fois avec des échantillons de tests de même taille mais sélectionnés aléatoirement parmi nos individus et des échantillons d'entraînement qui devront également avoir la même taille et être sélectionnés aléatoirement.



Point sur le critère 3 : la validation croisée II

Notre modèle va donc tourner autant de fois que nous avons décidé de scinder notre groupe d'individus en groupes test VS entraînement.

> A l'issue du processus de validation croisée l'algorithme arrivera à un compromis entre la diminution de l'erreur de prédiction et le sur-ajustement.

Classiquement, l'erreur de prédiction diminue constamment lorsque le nombre de sous-groupes augmente. Alors que l'erreur de prédiction obtenue en validation croisée va diminuer puis ré-augmenter.

> C'est ce seuil où l'erreur obtenue en validation croisée remonte qui fournit le fameux compromis qui définit le nombre de découpage optimale et donc la taille optimal de l'arbre.

Point sur le critère 3 : la validation croisée III

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

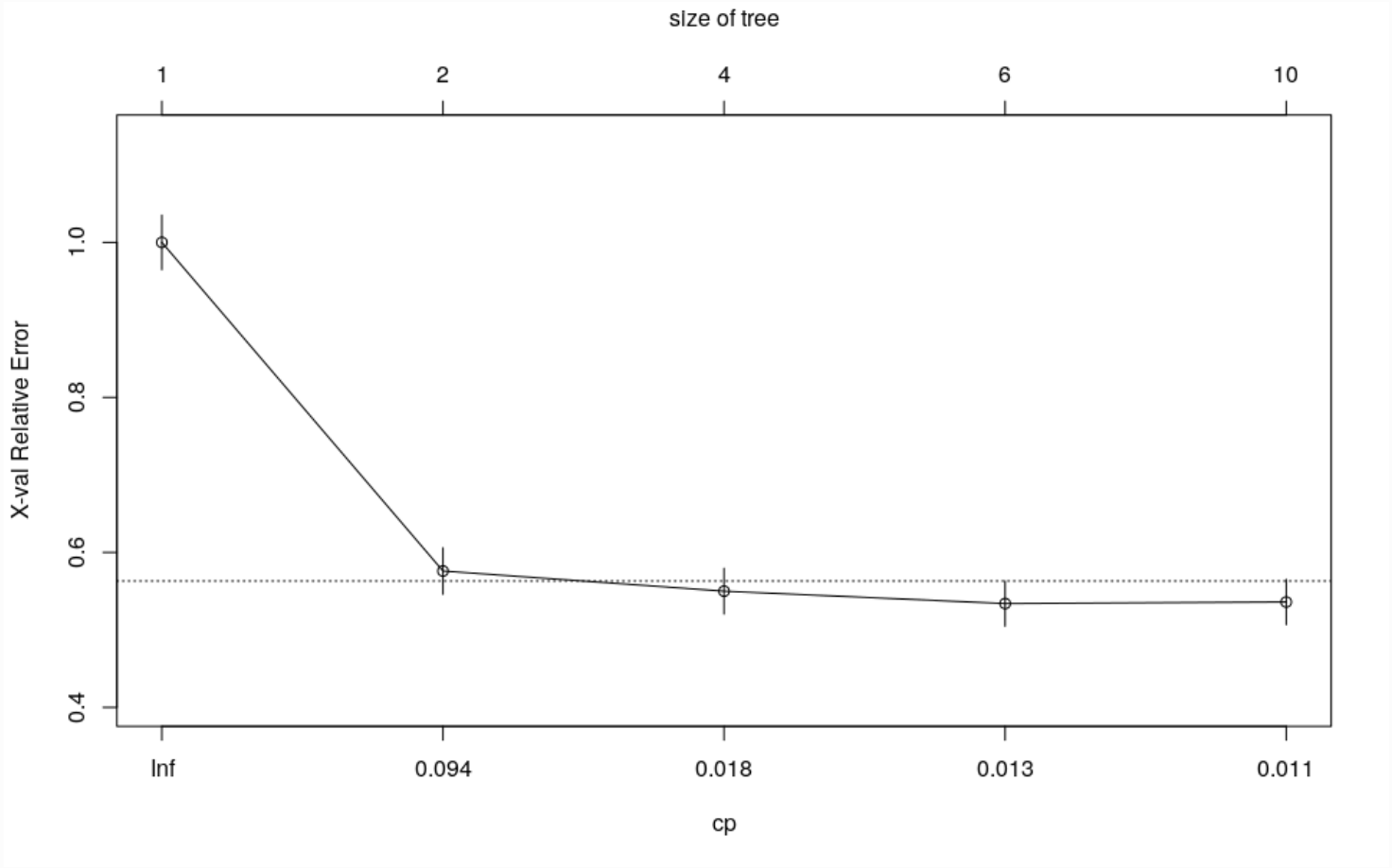
Il existe dans R des moyens pour récupérer le seuil exact à partir du tableau ou du graphique

La voie littérale (avec R)

Vous avez ici le nombre de découpage et la réduction de l'erreur de prédiction correspondante. L'erreur de prédiction en validation croisée correspond au **xerror**

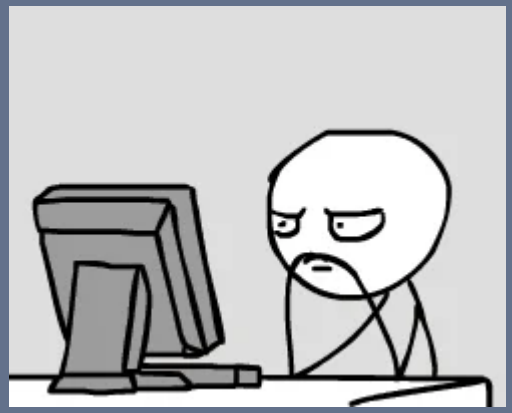
```
         CP nsplit rel error xerror     xstd
1 0.424000      0     1.000  1.000 0.035158
2 0.021000      1     0.576  0.576 0.029976
3 0.015000      3     0.534  0.550 0.029477
4 0.011333      5     0.504  0.534 0.029157
5 0.010000      9     0.458  0.536 0.029198
```

La voie graphique (avec R)



Ouf ! Fini 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 :

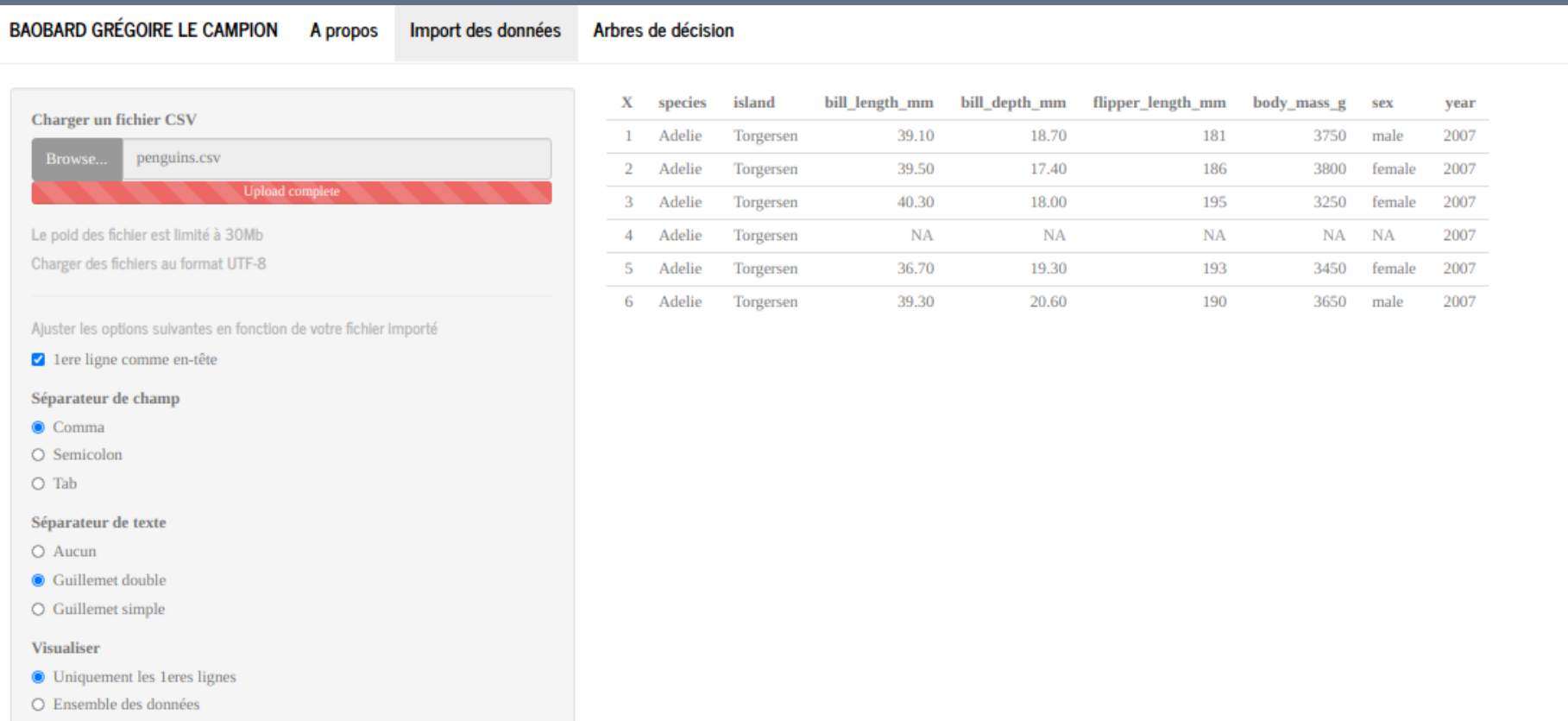
- les données d'entrée peuvent être « mixtes », c'est-à-dire qu'un même arbre peut utiliser simultanément des variables prédictives qualitatives, ordinales et continues ;

- la gestion des données manquantes est particulièrement efficace ;

- la construction d'un arbre est rapide et peu gourmande en ressources ;

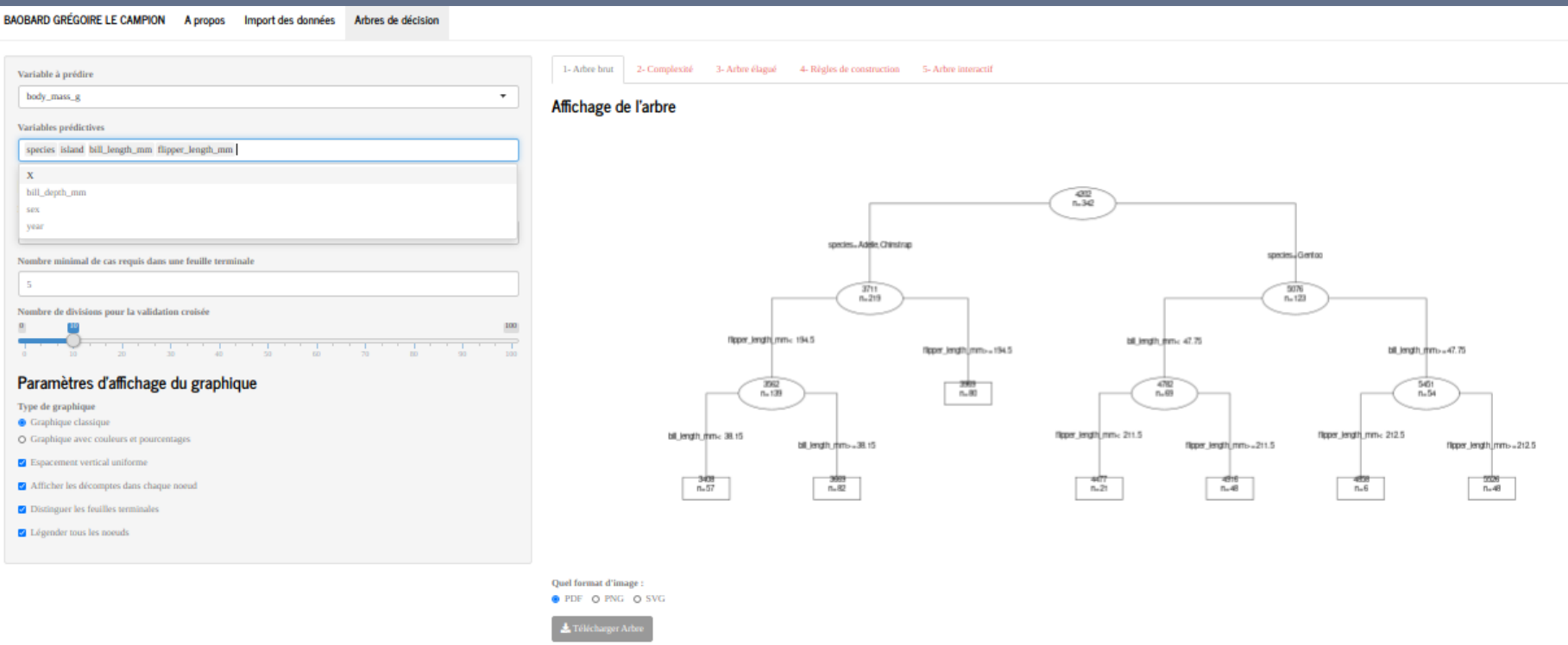
- l'arbre de décision est relativement simple à interpréter.

La page **Import des données** vous permet de charger vos données et de les visualiser.

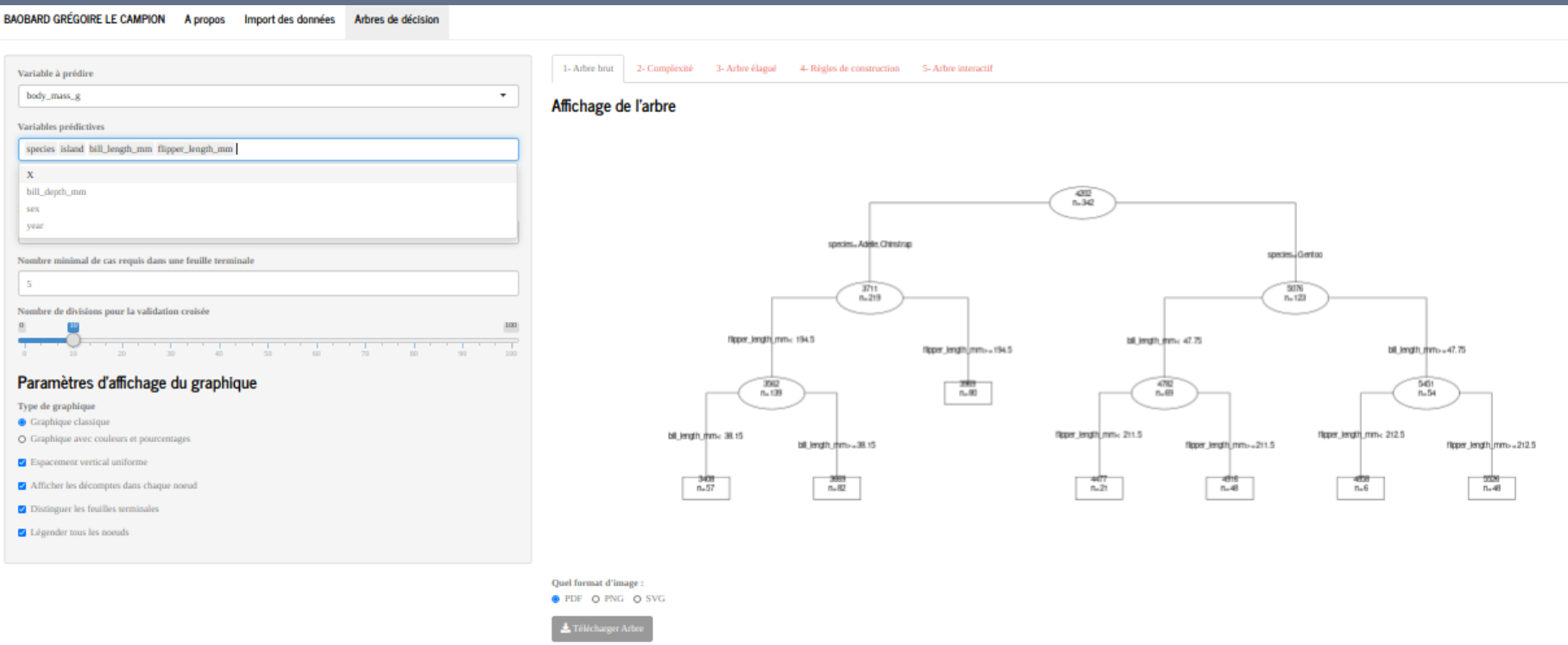


X	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.10	18.70	181	3750	male	2007
2	Adelie	Torgersen	39.50	17.40	186	3800	female	2007
3	Adelie	Torgersen	40.30	18.00	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.70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

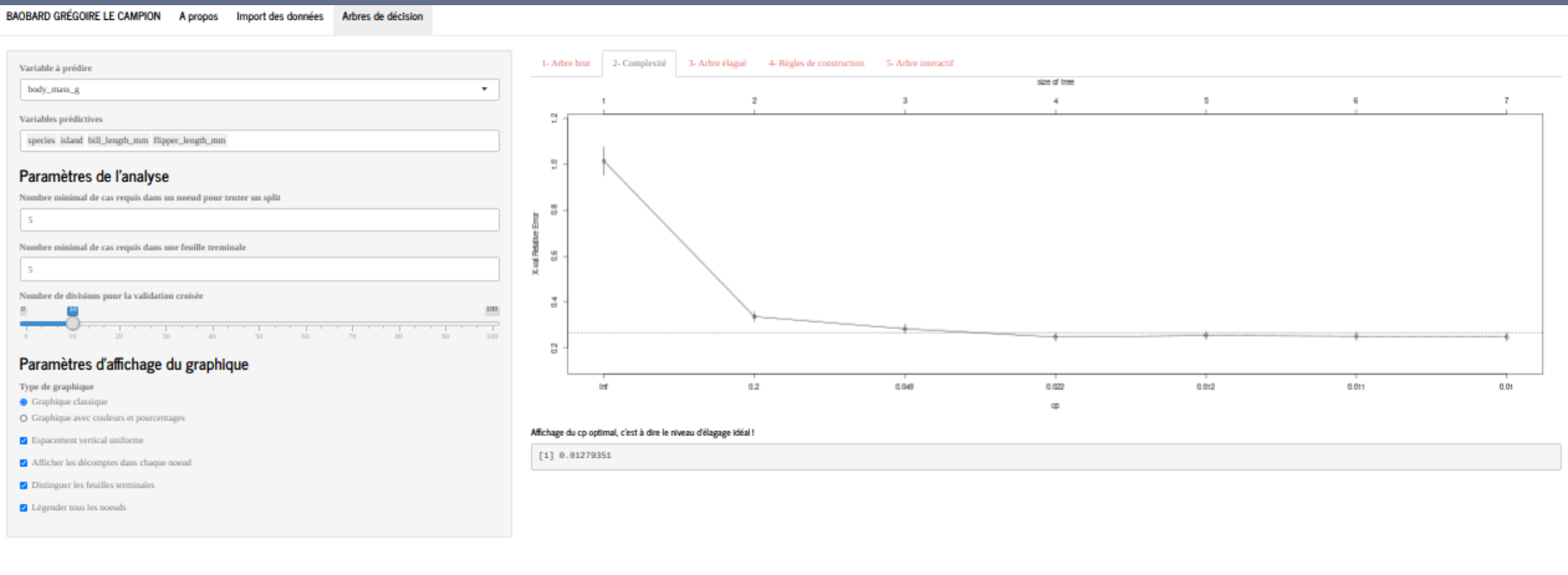
C'est donc ici que les analyses vont se faire, diverses options sont disponibles. Mais surtout le plus important c'est là qu'il faudra choisir la variable à étudier et les variables qui feront office de prédicteurs.



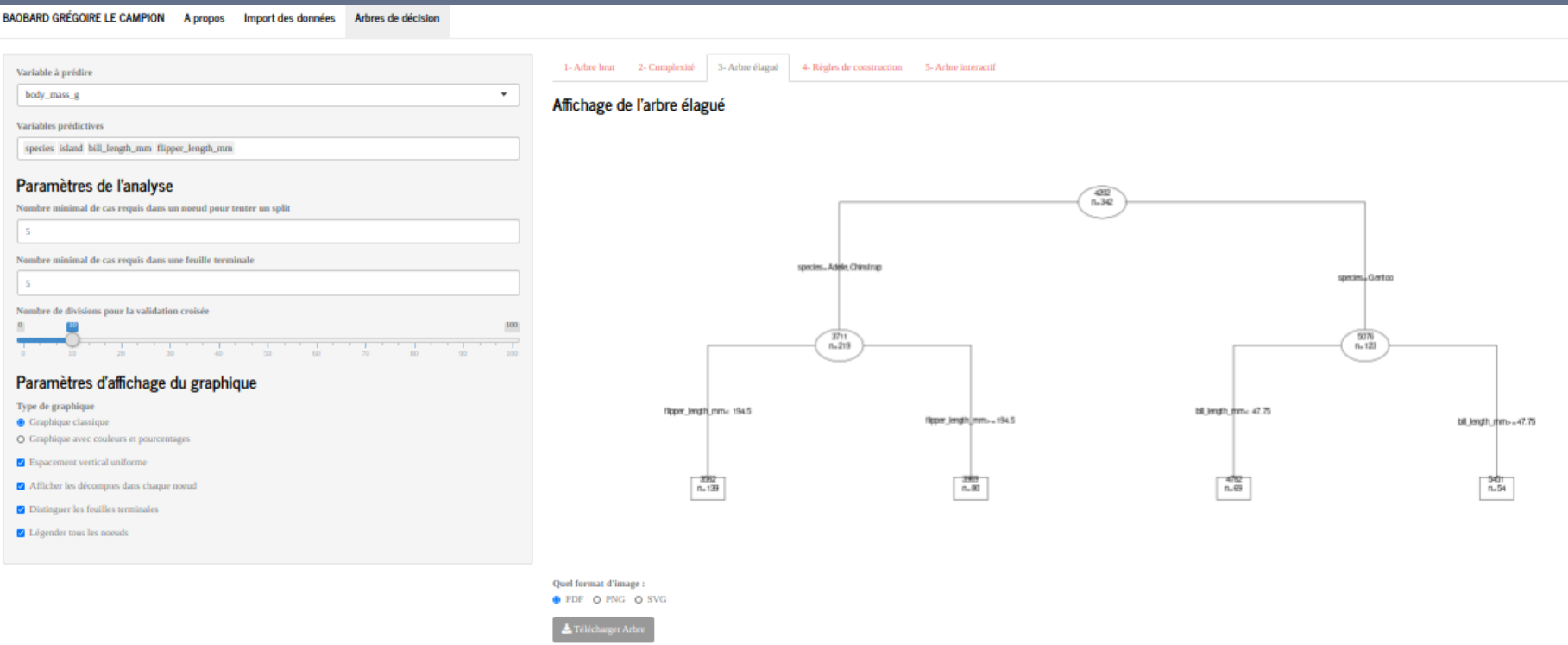
Le 1er onglet est **Arbre brut**, il s'agit de la première visualisation brute de votre arbre, c'est à dire avec les paramètre par défaut. Un export en png, pdf et svg est possible.



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



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 ait la même forme que l'arbre brut. Un export au format pdf, png et svg est possible.



Le 4ème onglet **Régle de construction**, permet une lecture littérale de votre arbre.

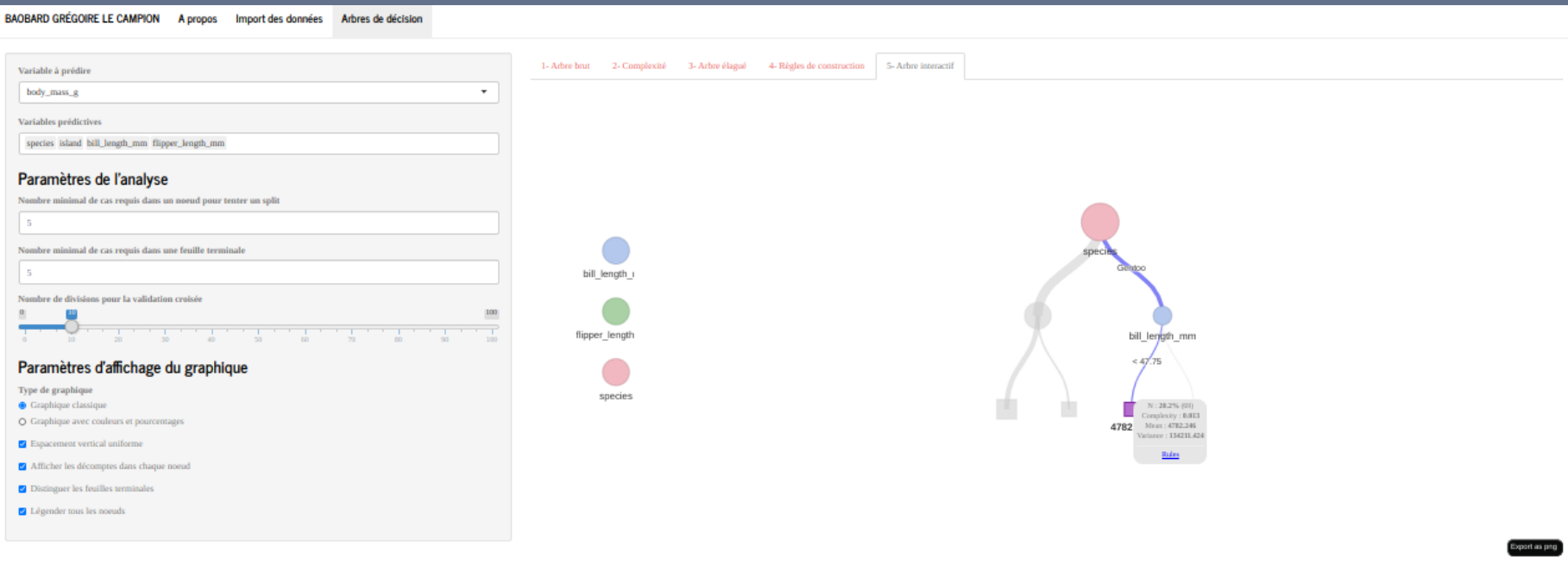
Comment les règles de constructions se lisent ? D'abord le numéro du noeud, puis le critère de division et le nombre total d'individus du groupe. Ensuite on aura une différence entre arbre de classification et arbre de régression.

Pour les arbres de classification viendra ensuite, le nombre des individus n'appartenant pas à la modalité prédite, la modalité prédite (c'est à dire majoritaire), et enfin entre parenthèse la proportion des individus bien et mal classés.

Pour les arbres de régression ce sera ensuite la somme des carrés des écarts à la valeur prédite pour les valeurs de toutes les instances du noeud. Puis la moyenne du groupe.

Si une ligne se termine par un * c'est qu'il s'agit d'une feuille terminale

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.



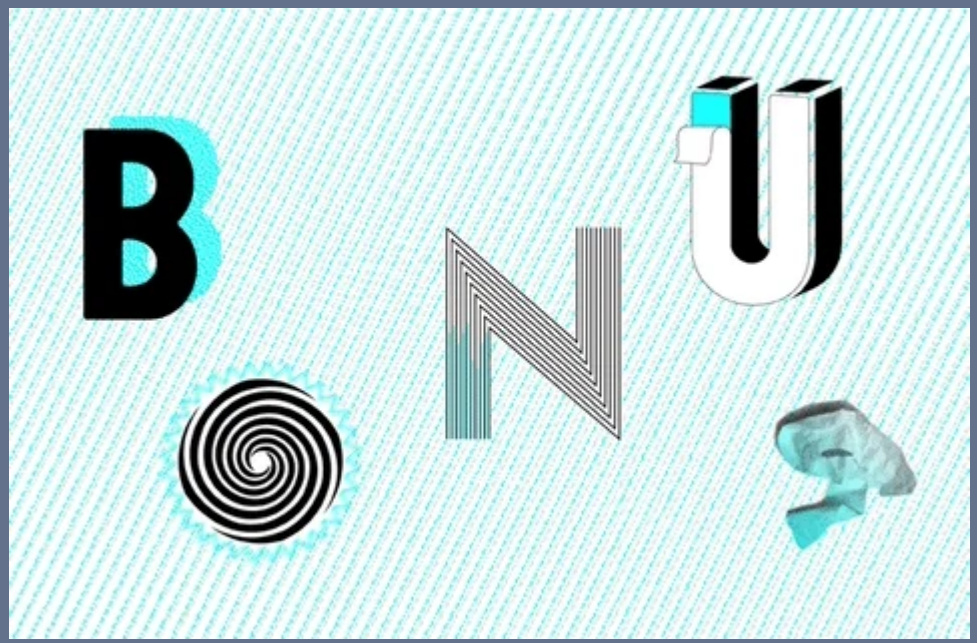
Un peu de R ?

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) #charger la librairie dans R

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) #produire l

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 modèles faibles et simples une fois combinés formeront un modèle robuste!

On est dans la famille des algorithmes qui font de l'agrégation de modèles, l'algorithme va construire une "forêt" d'arbre de décision, c'est à dire plusieurs centaines voire milliers, construite 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 selectionné et la construction d'un sous-groupe se fait sur un sous-ensemble de variables lui aussi selectionné aléatoirement.

Une des particularités dans la forêt aléatoire c'est qu'il n'y a pas "d'élagage" des arbres de décisions, c'est à dire que l'arbre pousse jusqu'au sur-ajustement.

Survol très superficiel des forêts aléatoires III

Une fois le modèle élaboré, tous les arbres de décision vont tourner sur les données qui n'auront pas servies à leur construction.

La prédiction finale correspondra à la modalité prédite le plus frequemment (dans le cas d'une variable à prédire qualitative) , ou au score moyen de tous les arbres de décisions (dans le cas d'une variable quantitative).

L'importance des variables explicatives sera fournie par une moyenne, réalisée sur l'ensemble des arbres de décision, de la diminution de l'erreur sur chaque ramification des arbres régit par la variable explicative.