Université Paris Descartes **Master 1e année**
Année 2019-2020 **Classification**

Devoir maison.

Exercice 1. Les données suivantes ont été récoltées :

Jour	X_1 (Ensoleillement)	X_2 (Température)	X_3 (Humidité)	X_4 (Vent)	Y (Jouer)
1	Soleil	Chaud	Humide	Faible	Non
2	Soleil	Chaud	Humide	Fort	Non
3	Couvert	Chaud	Humide	Faible	Oui
4	Couvert	Chaud	Humide	Faible	Oui
5	Soleil	Frais	Sec	Faible	Oui
6	Couvert	Frais	Sec	Fort	Non
7	Couvert	Frais	Sec	Fort	Oui
8	Soleil	Chaud	Humide	Faible	Non
9	Soleil	Frais	Sec	Faible	Oui
10	Couvert	Chaud	Sec	Faible	Oui
11	Soleil	Chaud	Sec	Fort	Oui
12	Couvert	Chaud	Sec	Fort	Oui
13	Couvert	Chaud	Humide	Fort	Non
14	Couvert	Chaud	Sec	Faible	Oui

1. Transformer le tableau en un tableau de variables binaires $0-1$, en précisant la notation choisie.
2. À partir de ces données, construire selon la méthode CART l'arbre de classification, que l'on appellera T. On ne segmentera pas les nœuds purs ni les nœuds composés de 4 ou de moins de 4 individus. Justifier soigneusement chacune des étapes.
3. Combien de nœuds purs et de feuilles contient T ? Attacher à chaque feuille un label.
4. Donner la règle de classification $\hat{t}_{14}$ induite par l'arbre T. Calculer l'erreur empirique

$$\hat{R}(\hat{t}_{14}) = \frac{1}{14}\sum_{j=1}^{14} \mathbf{1}_{\hat{t}_{14}(X_j)\neq Y_j}.$$

5. Au jour 15, les conditions climatiques sont les suivantes

$$(X_1 = \text{Soleil}, X_2 = \text{Frais}, X_3 = \text{Sec}, X_4 = \text{Fort}).$$

   En utilisant le classifieur $\hat{t}_{14}$ déterminer si le joueur va ou non jouer.

Exercice 2. Nous allons utiliser la librairie `rpart` de `R`.

1. L'algorithme CART est implémenté dans la fonction `rpart`. Regarder l'aide de cette fonction et l'utiliser sur les données `tennis`.

2. Pour visualiser le résultat renvoyé par la fonction `rpart`, sauvegardé dans `arbre`, on peut soit utiliser un outil graphique (`plot(arbre)` et `text(arbre)`, ou bien la commande unique `prp(arbre)` avec le package `rpart.plot`), soit afficher directement la sortie `arbre` qui renvoie une écriture textuelle de l'arbre. Tester toutes ces fonctions.
3. Utiliser la fonction `predict` pour faire de la prédiction à partir du classifieur `arbre` : classer les données d'entrée elles-mêmes puis en déduire le taux d'erreur d'apprentissage.

Exercice 3. On reprend la base de données `iris`. L'algorithme des forêts aléatoires est implémenté dans la librairie `randomForest` de la librairie `randomForest`.

1. Séparer les données `iris` en une base `train` et une base `test`.
2. Appliquer la méthode CART sur les données `train`, puis calculer le taux d'erreur obtenu sur les données `test`.
3. Regarder l'aide de la fonction `randomForest` et l'utiliser pour construire un classifieur sur les données `train`. Examiner et comprendre les informations renvoyées en sortie. Utiliser la fonction `predict` pour calculer le taux d'erreurs sur les données de test et comparer avec le taux obtenu précédemment.

Nom : LOUIS
Prénom : James Kelson
Cours : Classification
N° etudiant : 21907145

- Exercice 1 -

1) Transformons le tableau en un tableau de variables binaires 0-1.

on prend la convention :

$X_1 = 1$ si X_1 = Soleil et $X_1 = 0$ si X_1 = couvert
$X_2 = 1$ si X_2 = Chaud et $X_2 = 0$ si X_2 = Frais
$X_3 = 1$ si X_3 = Sec et $X_3 = 0$ si X_3 = Humide
$X_4 = 1$ si X_4 = Fort et $X_4 = 0$ si X_4 = Faible
$Y = 1$ si Y = oui et $Y = 0$ si Y = Non.

Jour	X_1 Ensoleillement	X_2 Température	X_3 Humidité	X_4 Vent	Y Jouer	$\hat{Y}$
1	1	1	0	0	0	0
2	1	1	0	1	0	0
3	0	1	0	0	1	1
4	0	1	0	0	1	1
5	1	0	1	0	1	1
6	0	0	1	1	0	1
7	0	0	1	1	1	1
8	1	1	0	0	0	0
9	1	0	1	0	1	1
10	0	1	1	0	1	1
11	1	1	1	1	1	1
12	0	1	1	1	1	1
13	0	1	0	1	0	1
14	0	1	1	0	1	1
15	1	0	1	1		1

- on commence par la racine de l'arbre, contenant tous les individus. Nous avons 4 questions possibles à poser :

"$X_1 = 1$?", "$X_2 = 1$?", "$X_3 = 1$?", "$X_4 = 1$?"

- Si l'on veut découper la racine de l'arbre selon la variable X_1, on crée deux nouveaux nœuds :

  - Dans le nœud $X_1 = 1$, on aura les observations $e_1, e_2, e_5, e_8, e_9, e_{11}$ (donc 6 observations). Dont 3 ont la modalité $Y=0$ (e_1, e_2, e_8) et 3 ont la modalité $Y=1$ (e_5, e_9, e_1). Par conséquent,

$$I(X_1 = 1) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0,5.$$

  - Dans le nœud $X_1 = 0$, on aura les observations $e_3, e_4, e_6, e_7, e_{10}, e_{12}, e_{13}, e_{14}$. (soit 8 observations). Parmi elles, 2 ont la modalité $Y=0$ (e_6, e_{13}) et les autres ont la modalité $Y=1$.
  Par conséquent,

$$I(X_1 = 0) = 1 - \left(\frac{2}{8}\right)^2 - \left(\frac{6}{8}\right)^2 = 0,375$$

Pour ce choix de questions, la somme pondérée vaut

$$SP_1 = \frac{6}{14} \times \left(I(X_1 = 1)\right) + \frac{8}{14} \times I(X_1 = 0)$$

$$\underline{SP_1 = 0,4285714}$$

- Si on veut découper la racine de l'arbre selon la variable X_2. ($X_2 = 1$?)
  - Dans le Nœud $X_2 = 1$ nous avons les observations $e_1, e_2, e_3, e_4, e_8, e_{10}, e_{11}, e_{12}, e_{13}, e_{14}$ (10 observations) dont 4 ont la modalité $Y = 0$ (e_2, e_8, e_1, e_{13}).

$$I(X_2 = 1) = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 0,48$$

  - Dans le Nœud $X_2 = 0$, nous aurons les observations e_5, e_6, e_7, e_9 (4 observations) dont 1 a la modalité $Y = 0$ (e_6)

$$I(X_2 = 0) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0,375$$

La somme pondérée Vaut :

$$SP_2 = \frac{10}{14} \times 0,48 + \frac{4}{14} \times 0,375$$

$$\underline{SP_2 = 0,45}$$

- Si on veut découper la racine de l'arbre selon la variable X_3 ($X_3 = 1$?)
  - Dans le nœud $X_3 = 1$ nous aurons les observations : $e_5, e_6, e_7, e_9, e_{10}, e_{11}, e_{12}, e_{14}$ dont 1 a la modalité $Y = 0$ (e_6).

$$I(X_3 = 1) = 1 - \left(\frac{1}{8}\right)^2 - \left(\frac{7}{8}\right)^2 = 0,21875$$

  - Dans le nœud $X_3 = 0$ nous aurons les observations : $e_1, e_2, e_3, e_4, e_8, e_{13}$ (6 observations) dont 2 ont la modalité $Y = 1$ et les autres ($Y = 0$).

$$I(X_3 = 0) = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = 0,444$$

La somme pondérée est :

$$SP_3 = \frac{8}{14} \times 0,21875 + \frac{6}{14} \times 0,444$$

$$\underline{SP_3 = 0,315476}$$

- Si on veut découper la racine de l'arbre selon la variable $X_4 = 1$ ($X_4 = 1$?)
  - Dans le nœud $X_4 = 1$, nous aurons les observations $e_2, e_6, e_7, e_{11}, e_{12}, e_{13}$ (6 observations) dont 3 ont la modalité $Y = 0$ (e_2, e_6, e_{13}) et 3 ont la modalité $Y = 1$ (e_7, e_{11}, e_{12})

$$I(X_4 = 1) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0,5$$

  - Dans le nœud $X_4 = 0$, nous aurons les observations $e_1, e_3, e_4, e_5, e_8, e_9, e_{10}, e_{14}$ (8 observations) dont 2 ont la modalité $Y = 0$ (e_1, e_8) et les autres ($Y = 1$).

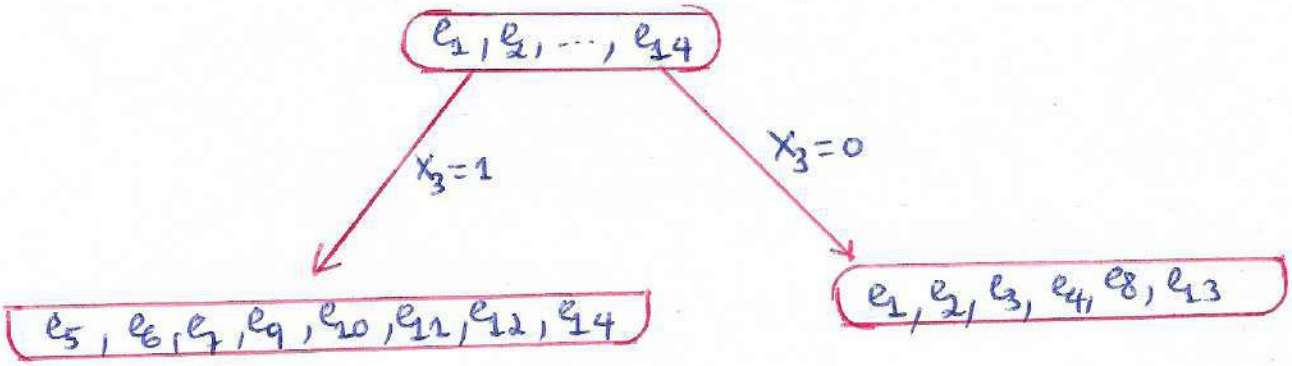
$$I(X_4 = 0) = 1 - \left(\frac{2}{8}\right)^2 - \left(\frac{6}{8}\right)^2 = 0,375$$

La somme pondérée est :

$$\underline{SP_4 = 0,4285714}$$

Le choix de question donnant la plus petite somme pondérée est la question portant sur X_2 (valeur : 0,315476)

Nous pouvons commencer à construire l'arbre.



on continue avec le découpage dans chaque noeud.

Pour le noeud $X_3 = 1$, on peut envisager 3 questions possibles "$X_1 = 1$?", "$X_2 = 1$?", "$X_4 = 1$?".

- Si on découpe ce noeud suivant la valeur de X_1, alors.
  - dans la branche $X_1 = 1$ on aura e_5, e_9, e_{11}. ces observations ont toutes la modalité $Y = 1$,

$$I(X_1 = 1) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

  - dans la branche $X_1 = 0$, on aura $e_6, e_7, e_{10}, e_{12}, e_{14}$ dont e_6 a la modalité $Y = 0$

$$I(X_1 = 0) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0,32$$

La somme pondérée est alors

$$SP_1' = \frac{3}{8} \times 0 + \frac{5}{8} \times 0,32$$

$$\underline{SP_1' = 0,2}$$

Si on découpe ce noeud selon la valeur de X_2, alors

- dans la branche $X_2 = 1$ on aura : $e_{10}, e_{11}, e_{12}, e_{14}$ qui ont toutes la modalité $Y = 1$.

$$I(X_2 = 1) = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

- dans la branche $X_2 = 0$ on aura : e_5, e_6, e_7, e_9 dont e_6 a la modalité $Y = 0$

$$I(X_2 = 0) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0,375$$

La somme pondérée est :

$$SP_2' = \frac{4}{8} \times 0 + \frac{4}{8} \times 0,375$$

$$\underline{SP_2' = 0,1875}$$

Si on découpe le noeud selon la question $X_4 = 1$? alors

dans la branche $X_4 = 1$ on aura : e_6, e_7, e_{11}, e_{12} seul e_6 a la modalité $Y = 0$

$$I(X_4 = 1) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0,375$$

dans la branche $X_4 = 0$ on aura e_5, e_9, e_{10}, e_{14}

$$I(X_4 = 0) = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

La somme pondérée est :

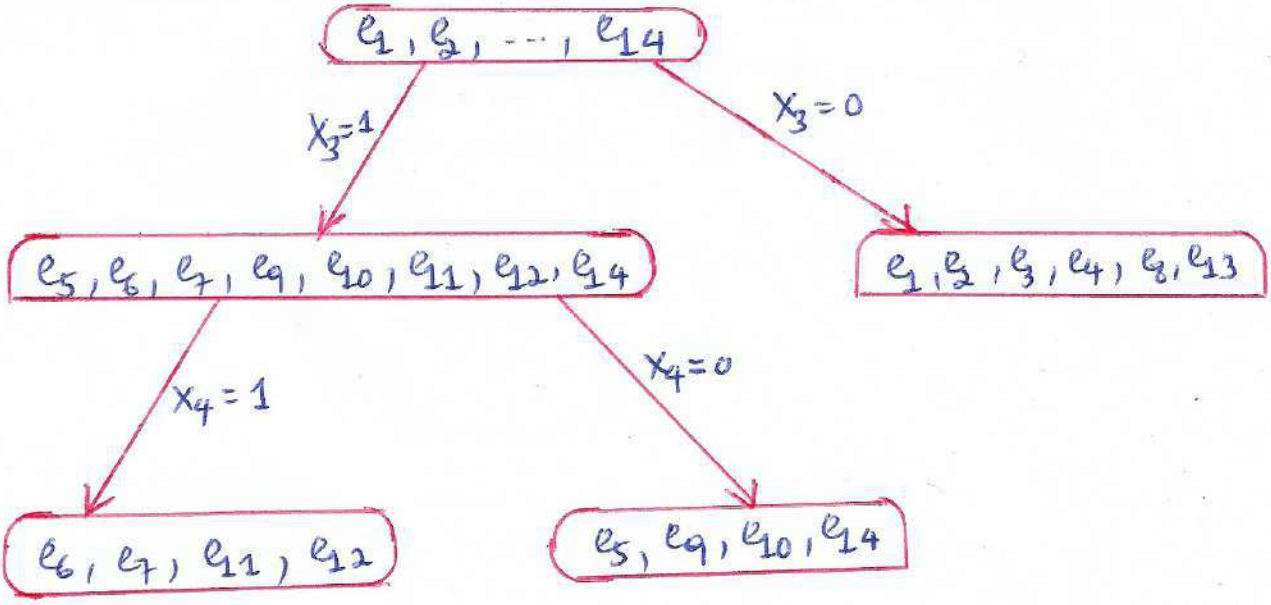
$$SP_3' = \frac{4}{8} \times 0,375 + \frac{4}{8} \times 0$$

$$\underline{SP_3' = 0,1875}$$

La plus petite valeur correspond aux choix de questions "$X_2 = 1$?" ou "$X_4 = 1$?"

on choisit "$X_4 = 1$?"

L'arbre devient :



Pour la branche $X_3 = 0$, on peut envisager trois questions possibles : "$X_1 = 1$?", "$X_2 = 1$?", "$X_4 = 1$?"

Si on découpe selon la valeur de X_1 on aura dans la branche $X_1 = 1$: e_1, e_2, e_8 qui ont pour modalité $Y = 0$

$$I(X_1 = 1) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$$

dans la branche $X_1 = 0$ on aura : e_3, e_4, e_{13}

dont e_{13} a pour modalité $Y = 0$

$$I(X_1 = 0) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0,444$$

La somme pondérée est :

$$SP_1'' = \frac{3}{6} \times 0 + \frac{3}{6} \times 0,444$$

$$\underline{SP_1'' = 0,222}$$

Si on découpe selon la valeur de X_2 ; toutes les observations se trouvent dans la branche $X_2 = 1$ avec e_1, e_2 e_8 et e_{13} qui ont pour modalité $Y = 0$

$$I(X_2 = 1) = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 = 0,444$$

$$I(X_2 = 0) = 1 - 0 - 0 = 1$$

La somme pondérée est :

$$SP_2'' = \frac{6}{6} \times 0,444 + \frac{0}{6} \times 1$$

$$\underline{SP_2'' = 0,444}$$

Si on découpe selon la valeur de X_4, dans la branche $X_4 = 1$ on aura : e_2, e_{13} qui ont la modalité $Y = 0$

$$I(X_4 = 1) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

dans la branche $X_4 = 0$ on aura : e_1, e_3, e_4, e_8

$$I(X_4 = 0) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0,5$$

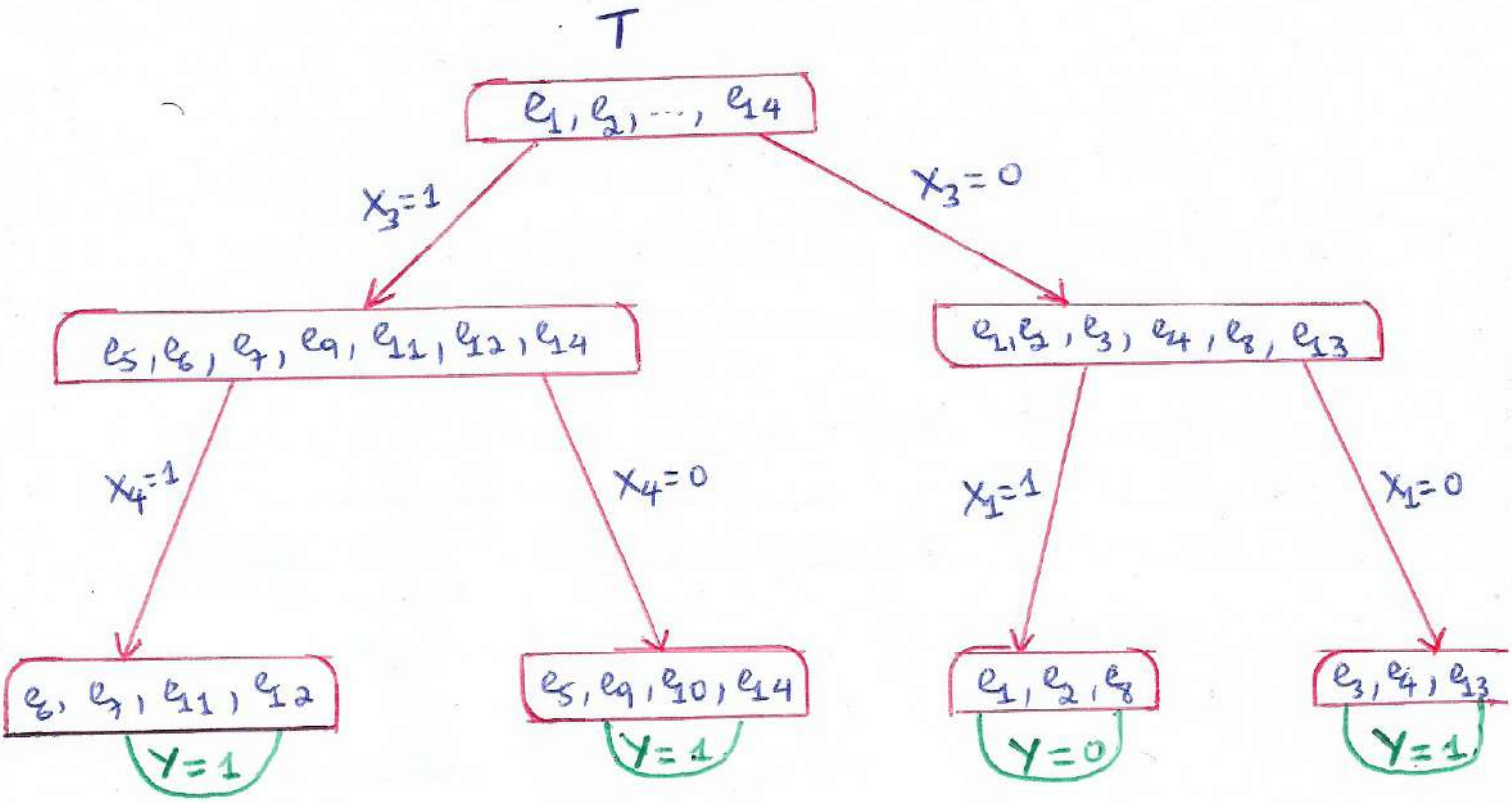
la somme pondérée est :

$$SP_3'' = \frac{2}{6} \times 0 + \frac{4}{6} \times 0,5$$

$$\underline{SP_3'' = 0,3333}$$

La plus petite valeur correspond à la question "$X_1 = 1$?"

l'arbre devient :



On s'arrête là, car on avait comme critère d'arrêt :

Nombres d'individus par noeuds terminal doit être inférieur ou égal à 4.

3) T contient 2 noeuds purs $N_2(e_5, e_9, e_{10}, e_{14})$ et $N_3(e_1, e_2, e_8)$.

T contient 4 feuilles.

- Attachons un label à chaque feuille.

- La feuille N_1 : $\{e_6, e_7, e_{11}, e_{12}\}$ a 3 individus qui ont la modalité $Y=1$ et 1 individu avec $Y=0$ donc on attache $Y=1$ à N_1.

- La feuille N_2 étant un nœud pur, 4 individus avec la modalité $Y=1$ donc on l'associe à $Y=1$.

- La feuille N_3 est un nœud pur, 3 individus avec la modalité $Y=0$ donc on l'associe à $Y=0$

- La feuille N_4 : (e_3, e_4, e_{13}) a deux individus avec la modalité $Y=1$ et un individu avec $Y=0$ donc on l'associe à $Y=1$.

4) Donnons la règle de classification $\hat{t}_{14}$ induite par l'arbre T.

- Si $X_3=1$ et $X_4=1$, alors $Y=1$
- Si $X_3=1$ et $X_4=0$, alors $Y=1$
- Si $X_3=0$ et $X_1=1$, alors $Y=0$
- Si $X_3=0$ et $X_1=0$, alors $Y=0$

Calculons l'erreur empirique.

$$\hat{R}(\hat{t}_{14}) = \frac{1}{14} \sum_{j=1}^{14} \mathbb{1}_{\hat{t}_{14}(X_j) \neq Y_j}$$

$$= \frac{1}{14} (0+0+0+0+0+1+0+0+0+0+0+0+1+0)$$

$$= \frac{1}{14} \times 2$$

$$= \frac{1}{7}$$

$$\hat{R}(\hat{t}_{14}) = 0,1428$$

5) Au Jour 15, les conditions climatiques sont les suivantes : (X_1: Soleil, X_2: Frais, X_3: sec, X_4: Fort)

Déterminons si le joueur va ou non jouer.

$$X_1 = 1 \ , \ X_2 = 0, \ X_3 = 1 \ , \ X_4 = 1$$

puisque $X_3 = 1$ et $X_4 = 1$ on associe le label $Y = 1$ donc on déduit que le joueur va jouer.

Devoir de maison/ Cours Classification

James Kelson LOUIS

5/20/2020

Exercice 2

1.Utilisons la fonction CART sur les données “tennis”

a) Chargement du jeu de données

```
load('tennis.RData')
head(tennis,5)
```

```
##         Ciel Temperature Humidite   Vent Jouer
## 1 Ensoleille        27.5       85 Faible   Non
## 2 Ensoleille        25.0       90   Fort   Non
## 3    Couvert        26.5       86 Faible   Oui
## 4      Pluie        20.0       96 Faible   Oui
## 5      Pluie        19.0       80 Faible   Oui
```

Importation de librairie

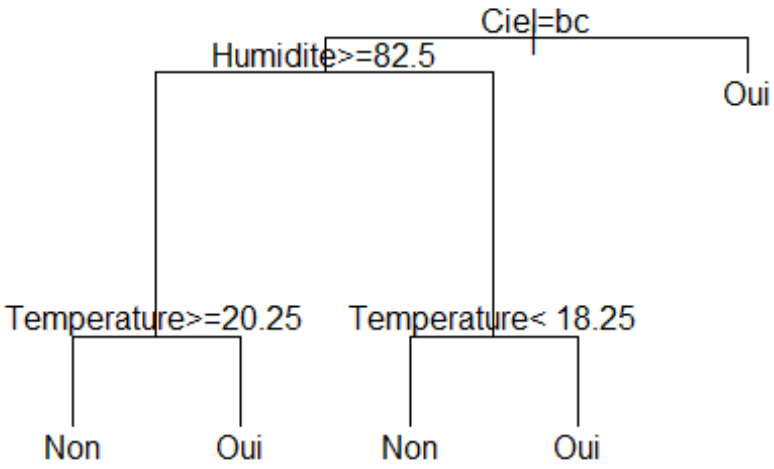
```
attach(tennis)
library(rpart)
library(rpart.plot)

arbre <- rpart(Jouer ~ ., data = tennis, minsplit = 1, cp = 0)
```

Visualisation des données

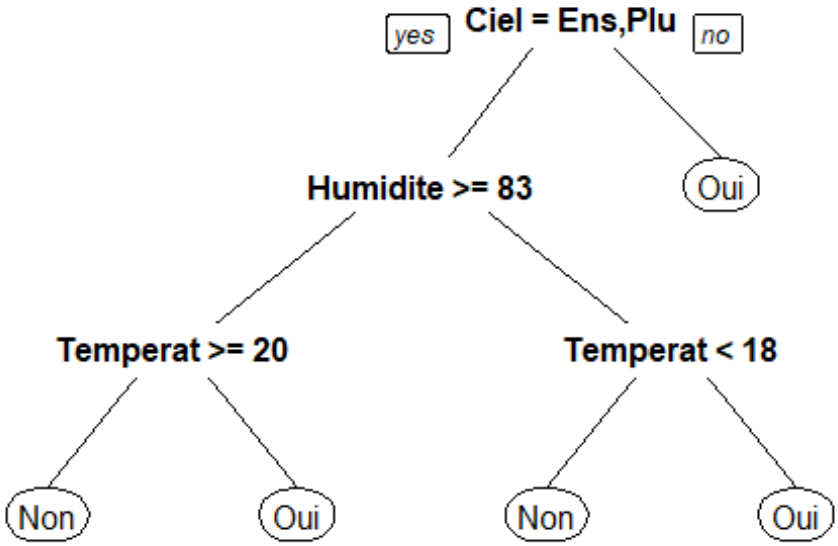
1ère méthode

```
par(xpd = NA)
plot(arbre)
text(arbre)
```



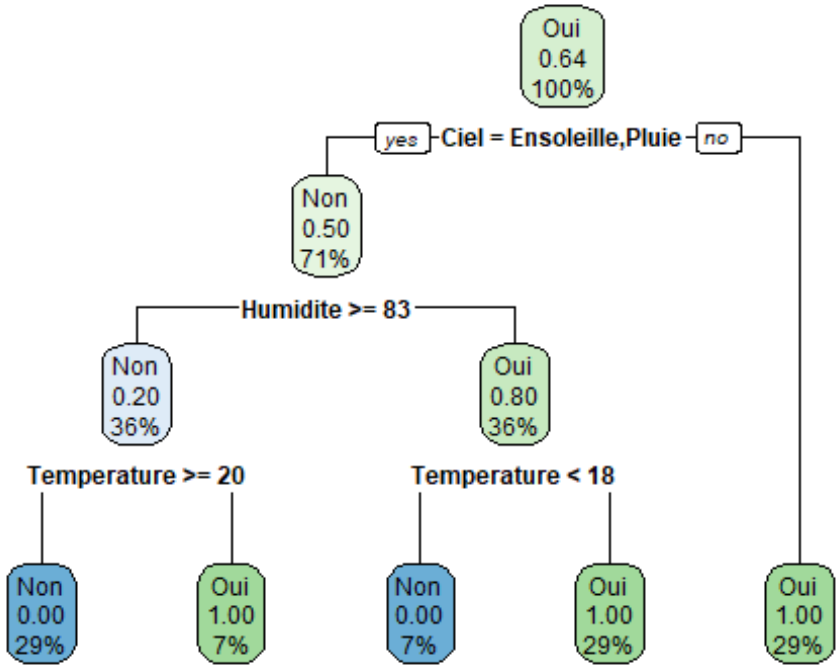
2ème méthode

```
prp(arbre)
```



3ème méthode

```
rpart.plot(arbre)
```



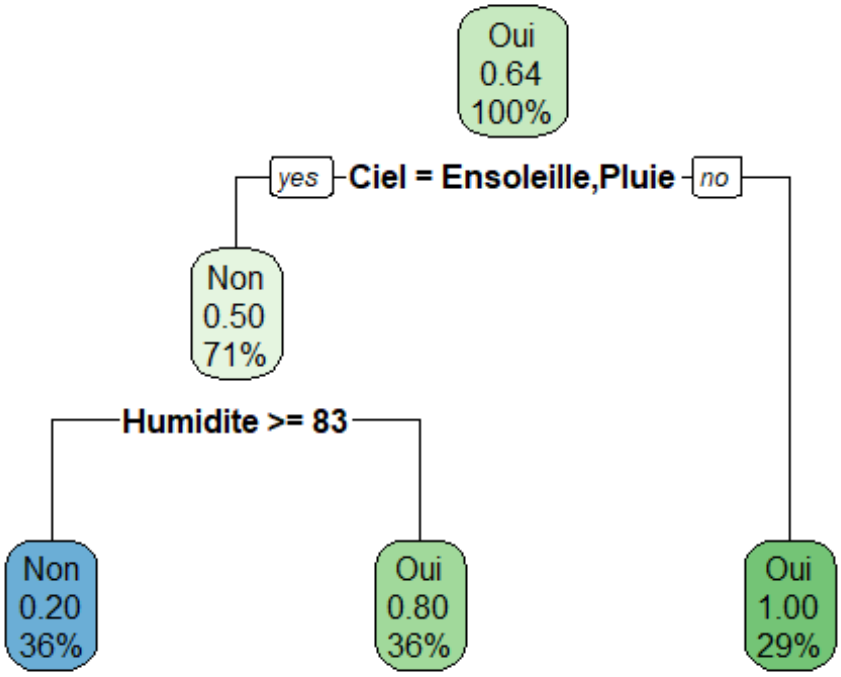
4ème méthode

```
arbre
```

```
## n= 14
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 14 5 Oui (0.3571429 0.6428571)
##    2) Ciel=Ensoleille,Pluie 10 5 Non (0.5000000 0.5000000)
##      4) Humidite>=82.5 5 1 Non (0.8000000 0.2000000)
##        8) Temperature>=20.25 4 0 Non (1.0000000 0.0000000) *
##        9) Temperature< 20.25 1 0 Oui (0.0000000 1.0000000) *
##      5) Humidite< 82.5 5 1 Oui (0.2000000 0.8000000)
##       10) Temperature< 18.25 1 0 Non (1.0000000 0.0000000) *
##       11) Temperature>=18.25 4 0 Oui (0.0000000 1.0000000) *
##    3) Ciel=Couvert 4 0 Oui (0.0000000 1.0000000) *
```

2. On classe les données d'entrées.

```
arbre_2 <- rpart(Jouer ~ ., data = tennis, control = rpart.control(minsplit =
5))
rpart.plot(arbre_2)
```



Prédiction

```
prediction <- predict(arbre_2, tennis, type = "class")
prediction
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14
## Non Non Oui Non Oui Oui Oui Non Oui Oui Oui Oui Oui Non
## Levels: Non Oui
```

Déduisons le taux d'erreur d'apprentissage.

```
taux_erreur_app <- round((sum(prediction!=Jouer)/length(Jouer))*100,2)
cat("Le taux d'erreur d'apprentissage est de:",taux_erreur_app,"%" )
```

```
## Le taux d'erreur d'apprentissage est de: 14.29 %
```

Exercice 3

Importation de la librairie randomForest

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.6.2
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

Séparons les données iris en une base train et une base test.

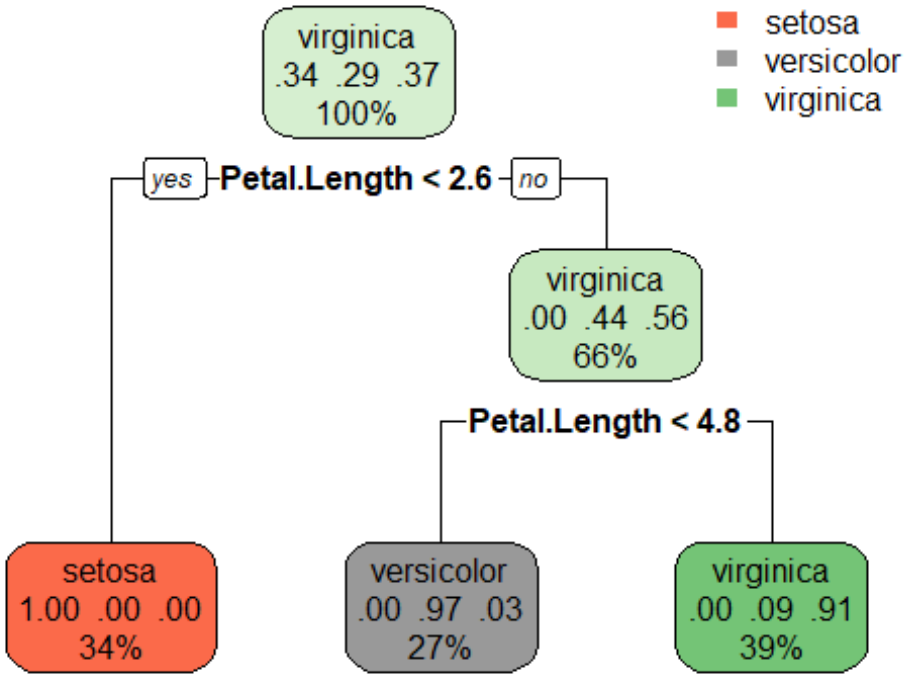
```
data(iris)
ind <- sample(1:nrow(iris),round(0.8*nrow(iris)))
length(ind)

## [1] 120

attach(iris)
X_train <-iris[ind,]
X_test <- iris[-ind,]
```

2. Appliquons la méthode CART sur les données train, puis calculons le taux d'erreur obtenu sur les données test.

```
model <- rpart(Species ~., data = X_train)
par(xpd = NA)
rpart.plot(model)
```




```
(pred1 <- predict(model,X_test, type = "class"))

##         10         12         14         16         17         21
32
##     setosa     setosa     setosa     setosa     setosa     setosa
setosa
##         42         48         52         56         63         66
77
##     setosa     setosa versicolor versicolor versicolor versicolor
virginica
```

```
##         81         83         84         85         88         90
93
## versicolor versicolor  virginica versicolor versicolor versicolor
versicolor
##         94         95         99        108        110        118
125
## versicolor versicolor versicolor  virginica  virginica  virginica
virginica
##        137        147
##  virginica  virginica
## Levels: setosa versicolor virginica

err_cart <- round(sum(pred1!=X_test$Species)/nrow(X_test)*100,2)
cat("Le taux d'erreur de prédiction est de: ", err_cart,"%")

## Le taux d'erreur de prédiction est de:  6.67 %
```

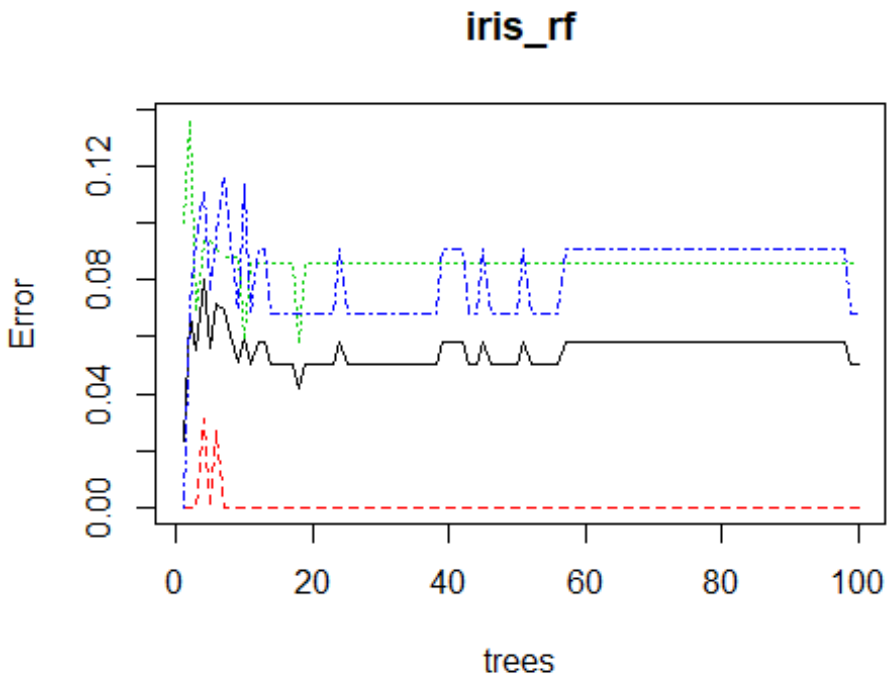
3. Utilisons la fonction randomForest pour construire un classifieur sur les données train.

```
iris_rf <- randomForest(Species~.,data=X_train,ntree=100,proximity=TRUE)

summary(iris_rf)

##                 Length Class  Mode
## call               5  -none- call
## type               1  -none- character
## predicted        120  factor numeric
## err.rate         400  -none- numeric
## confusion         12  -none- numeric
## votes            360  matrix numeric
## oob.times        120  -none- numeric
## classes            3  -none- character
## importance         4  -none- numeric
## importanceSD       0  -none- NULL
## localImportance    0  -none- NULL
## proximity      14400  -none- numeric
## ntree              1  -none- numeric
## mtry               1  -none- numeric
## forest            14  -none- list
## y                120  factor numeric
## test               0  -none- NULL
## inbag              0  -none- NULL
## terms              3  terms  call

plot(iris_rf)
```



3.b Utilisons la fonction predict pour calculer le taux d'erreurs sur les données de test.

```
(pred2<-predict(iris_rf,newdata=X_test))
```

```
##         10         12         14         16         17         21
32
##     setosa     setosa     setosa     setosa     setosa     setosa
setosa
##         42         48         52         56         63         66
77
##     setosa     setosa versicolor versicolor versicolor versicolor
versicolor
##         81         83         84         85         88         90
93
## versicolor versicolor  virginica versicolor versicolor versicolor
versicolor
##         94         95         99        108        110        118
125
## versicolor versicolor versicolor  virginica  virginica  virginica
virginica
##        137        147
##  virginica  virginica
## Levels: setosa versicolor virginica
```

```
err_rf <- round(sum(pred2!=X_test$Species)/nrow(X_test)*100,2)
cat("Le taux d'erreur de prédiction est de: ", err_rf,"%")
```

```
## Le taux d'erreur de prédiction est de:  3.33 %
```

Comparons avec le taux obtenu précédemment

```
table(pred1,X_test$Species)
```

```
##
## pred1        setosa versicolor virginica
##   setosa          9          0         0
##   versicolor      0         13         0
##   virginica       0          2         6
```

```
table(pred2,X_test$Species)
```

```
##
## pred2        setosa versicolor virginica
##   setosa          9          0         0
##   versicolor      0         14         0
##   virginica       0          1         6
```

```
table(pred1,pred2)
```

```
##             pred2
## pred1        setosa versicolor virginica
##   setosa          9          0         0
##   versicolor      0         13         0
##   virginica       0          1         7
```

```
cat("Avec la première classification il y a avait:",
sum(pred1!=X_test$Species), " classification(s) incorrecte(s) \n Avec la
deuxième classification il y a:",sum(pred2!=X_test$Species), "
classification(s) incorrecte(s)")
```

```
## Avec la première classification il y a avait: 2  classification(s)
incorrecte(s)
##  Avec la deuxième classification il y a: 1  classification(s)
incorrecte(s)
```