Saâd Qriouet M1 MINT 20171683

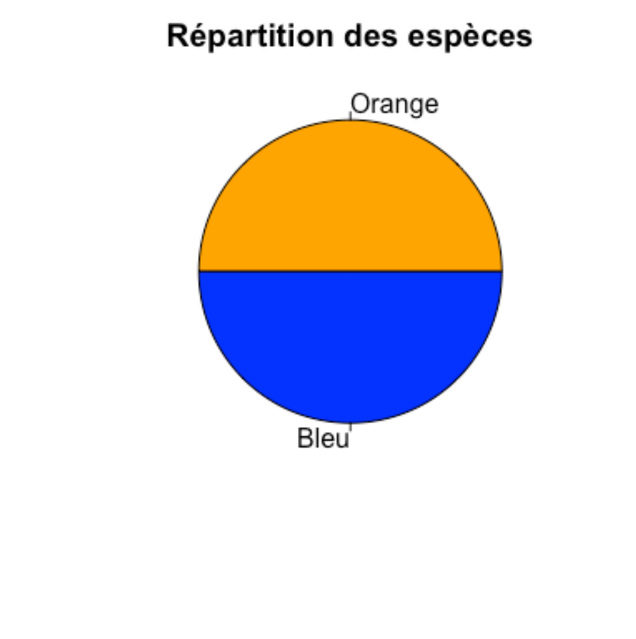
PRÉSENTATION PROJET MACHINE LEARNING

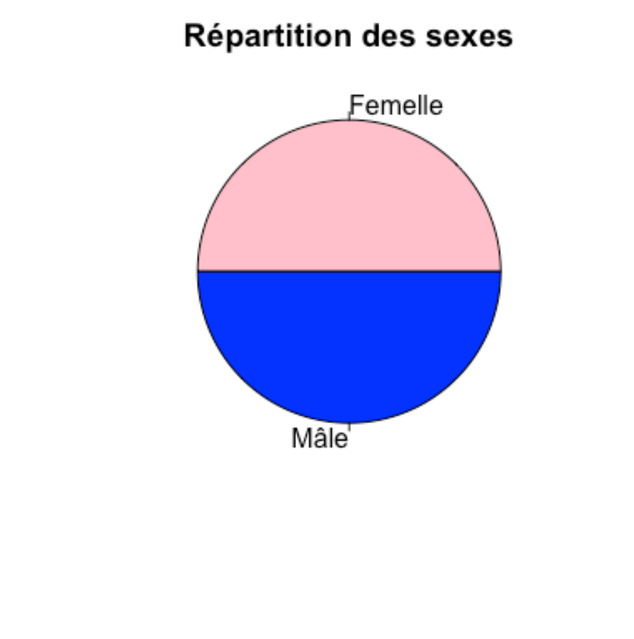
Présentation du jeu de données : Crabs

- 5 mesures morphologiques sur 200 crabes de deux couleurs et sexes différents
- Espèce Leptograpsus variegatus, collectées à Fremantle en Australie occidentale
- 8 variables :
  - La variable identifiante "index"
  - 5 variables quantitatives : FL, RW, CL, CW et BD
  - 5 variables qualitatives : sp et sex

Brève description du jeu de données :

- Variables qualitatives : Tableau d'effectif et fréquences + diagramme circulaire
  - Sp : Il y a autant de crabes bleus que de crabes oranges
  - Sex : Il y a autant de crabes mâles que de crabes femelles





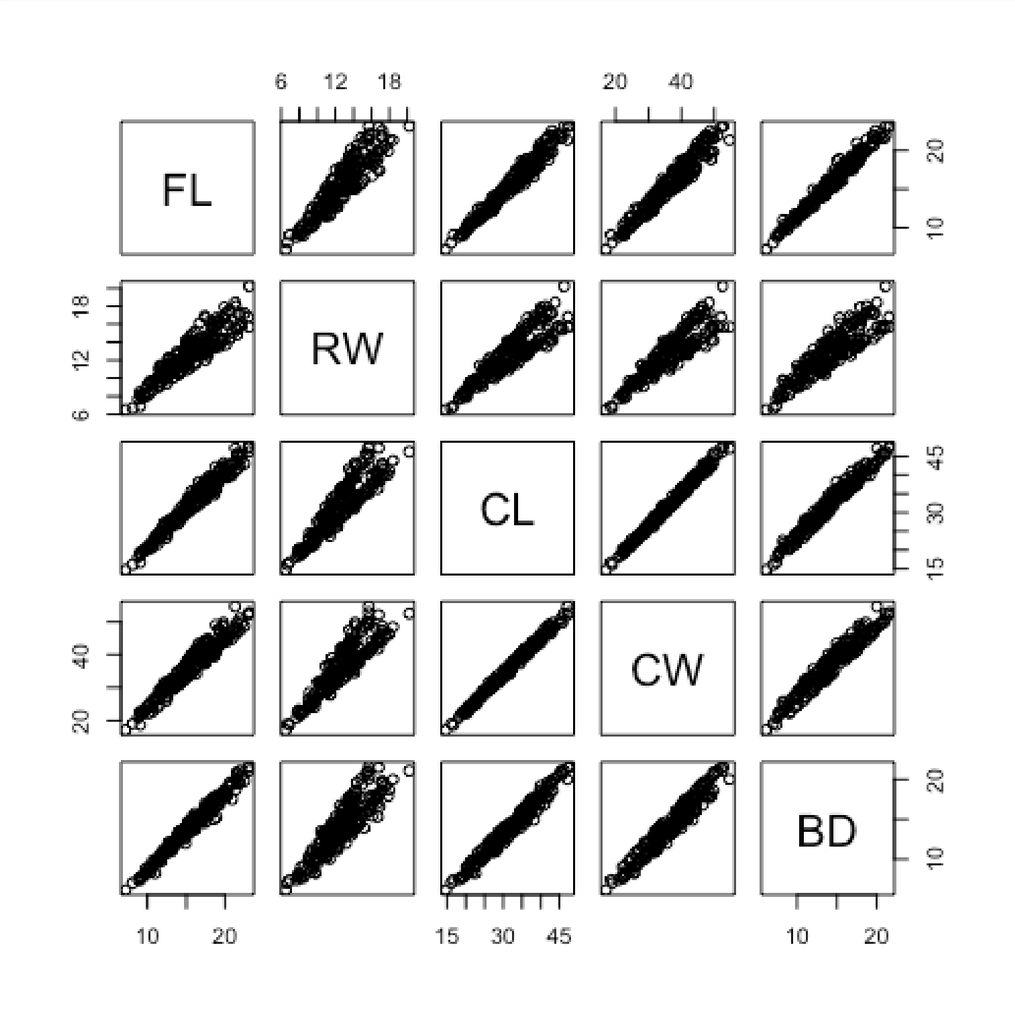
Brève description du jeu de données :

- Variables quantitatives : Calculs (moyenne + variance), tableau effectifs + frequences + frequences cumulées, histogramme et Boxplot
  - FL : Moyenne = 15.583 + valeurs allant de 1 à 50 + beaucoup points abérrants sur le Boxplot
  - RW : Moyenne = 12.7385 + valeurs allant de 7.2 à 23.1 + beaucoup points abérrants sur le Boxplot
  - CL : Moyenne = 32.1055 + valeurs allant de 14.7 à 47.6 + beaucoup points abérrants sur le Boxplot
  - CW : Moyenne = 36.4145 + valeurs allant de 17.1 à 54.6 + beaucoup de points abérrants sur le Boxplot
  - BD : Moyenne = 14.0305 + valeurs allant de 6.1 à 21.6 + beaucoup points abérrants sur le Boxplot

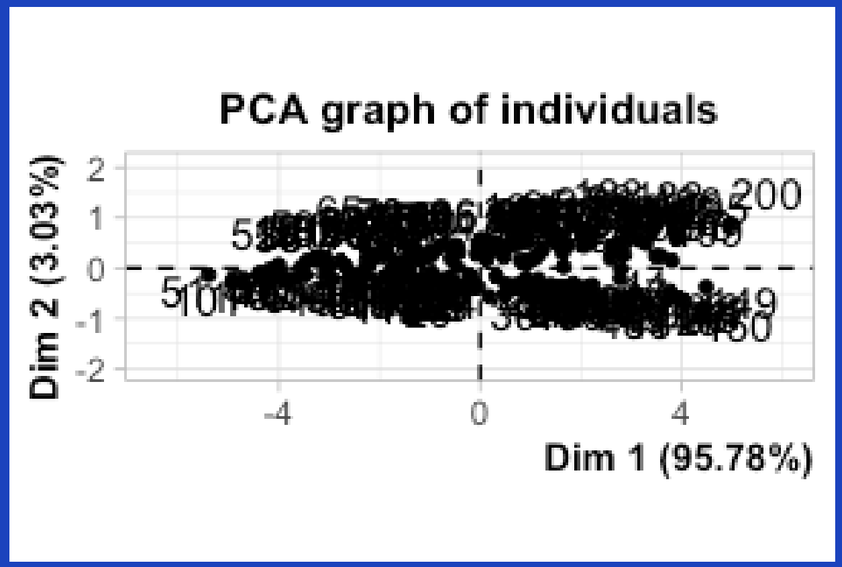
REALISATION DE L'ACP

- Scatterplot
- Première ACP :
  - Affichage + problèmes observés
  - Choix des dimmensions + controle de la pertinence de l'ACP
- ACP améliorée

Réalisation de l'ACP : Scatterplot

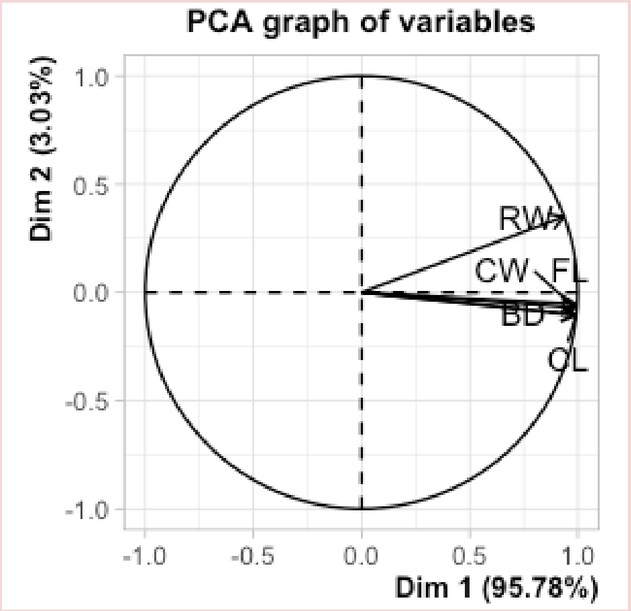


Réalisation de l'ACP : Première ACP

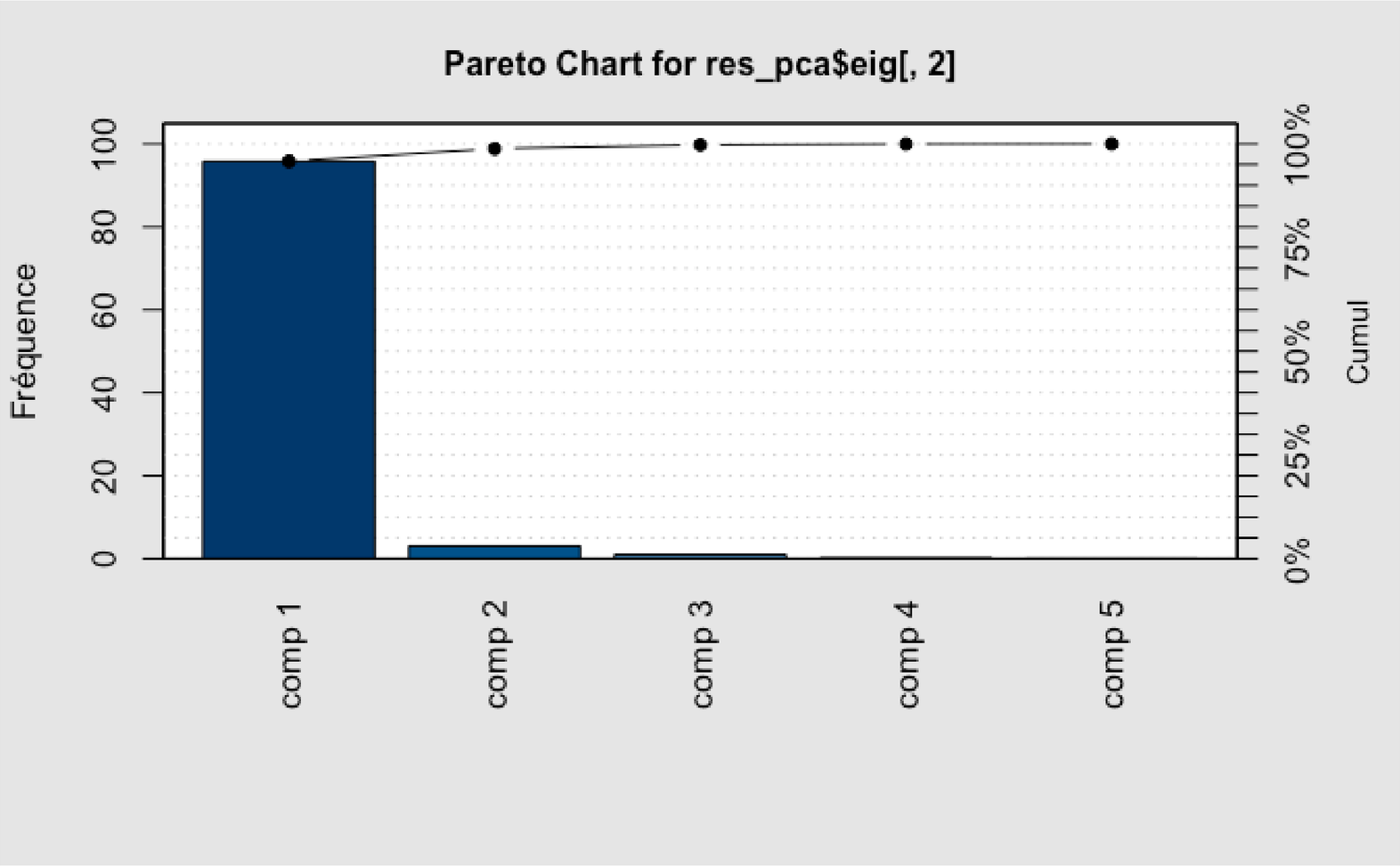


GRAPHIQUES ILLISIBLES :

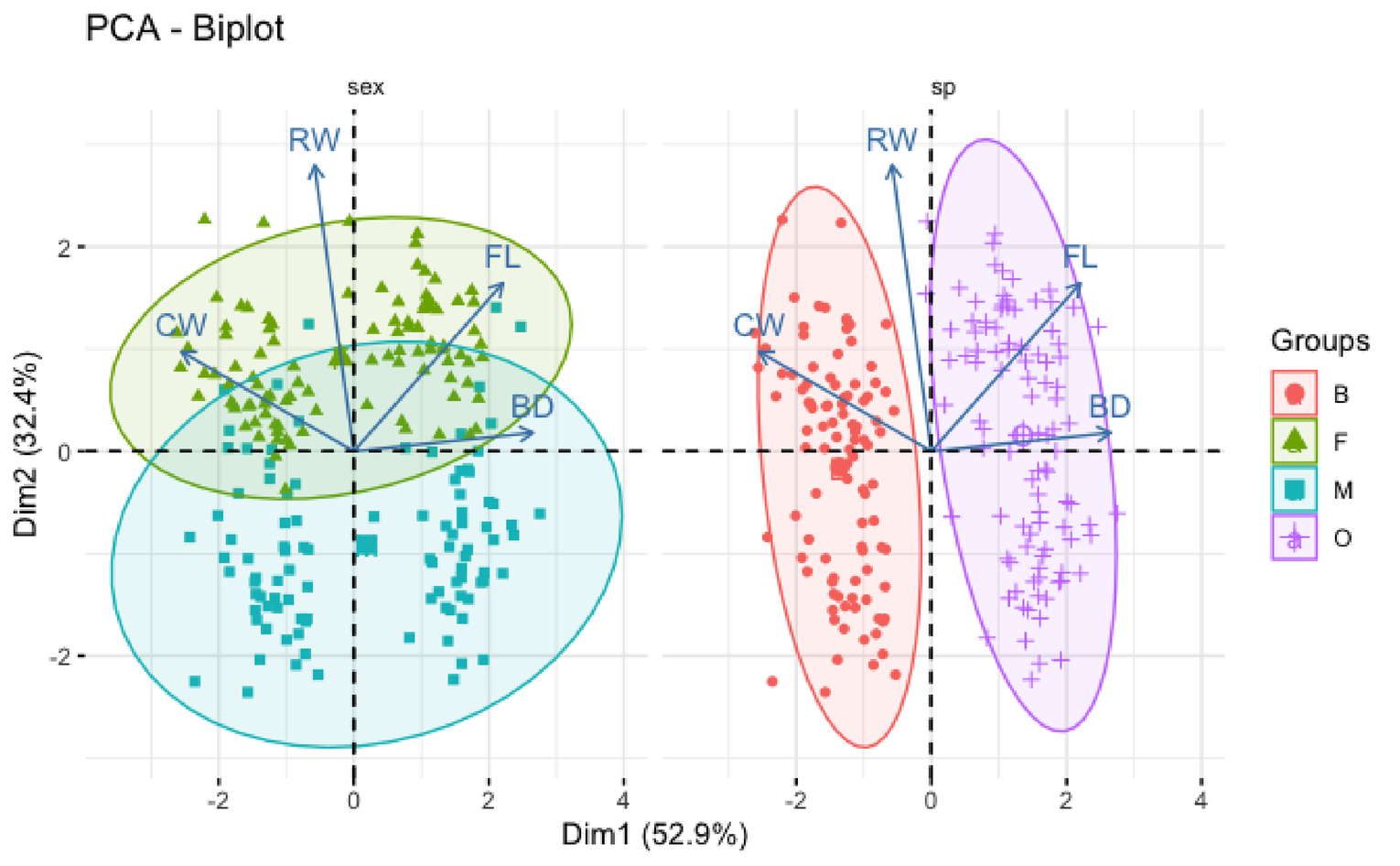
- Graphique des individus illisible
- Cercle des corrélations : pas d’infos sur les fleches



Réalisation de l'ACP : Première ACP



Réalisation de l'ACP : ACP Améliorée



Cercle de corrélations :

- Dimension 1 : Différencie les crabes selon sp :
  - Valeur élevée : crabes Oranges
  - Valeur faible : crabes Bleus
  - Oranges : rapport BD/CL élevé + rapport CW/CL faible
- Dimension 2 : Différencie les crabes selon sex :
  - Valeur élevée : crabes Femelles
  - Valeur faible : crabes Mâles (moins flagrant)
  - Femelle : rapport RW/CL élevé

Prédiction avec CART

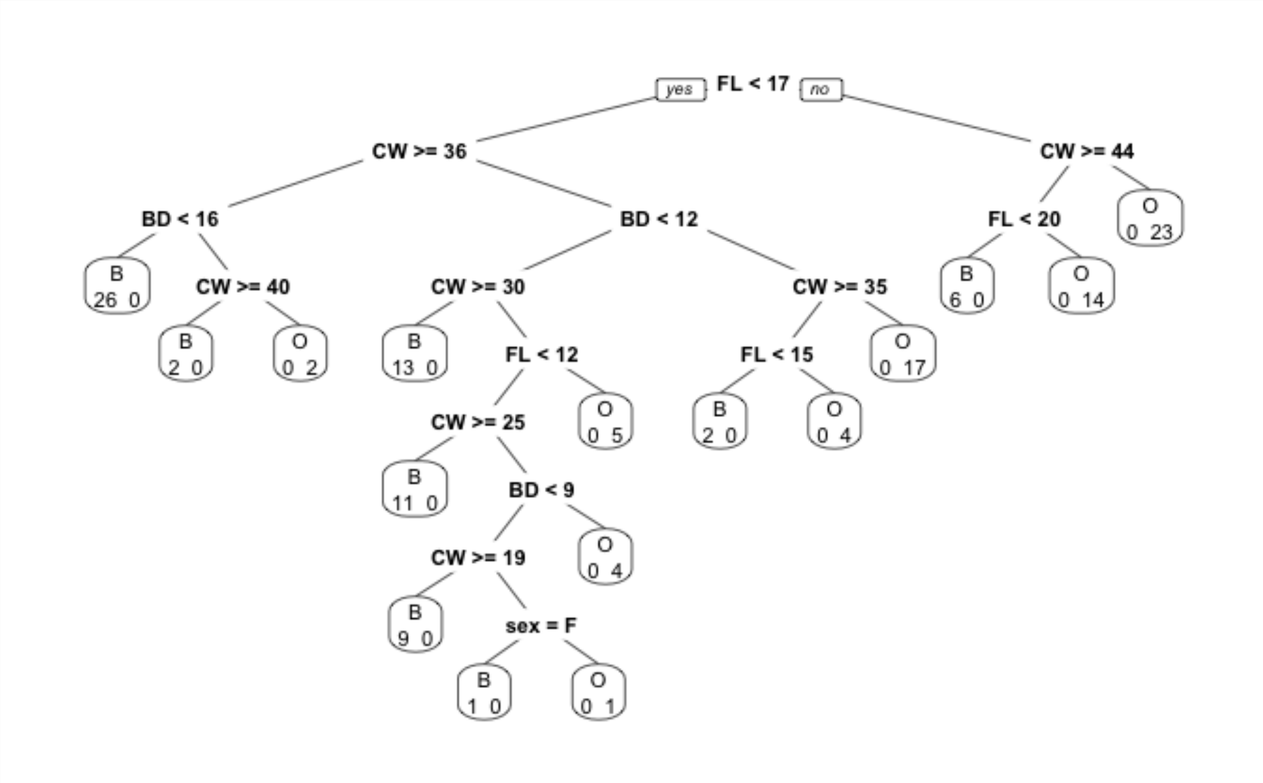
- Arbre de classification complet
  - Création d'échantillons
  - Arbre complet
  - Calcul compléxité optimale
- Arbre de simplification
- Performances de l'arbre simplifié

Arbre de classification :

- But : prédire la variable sp, variable qualitative
  => arbre de classification
- Subdivision du jeu de données en deux : un pour l'apprentissage (construction de l'arbre de décision) , et un pour le test (prédictions et évaluation des performances du modèle)

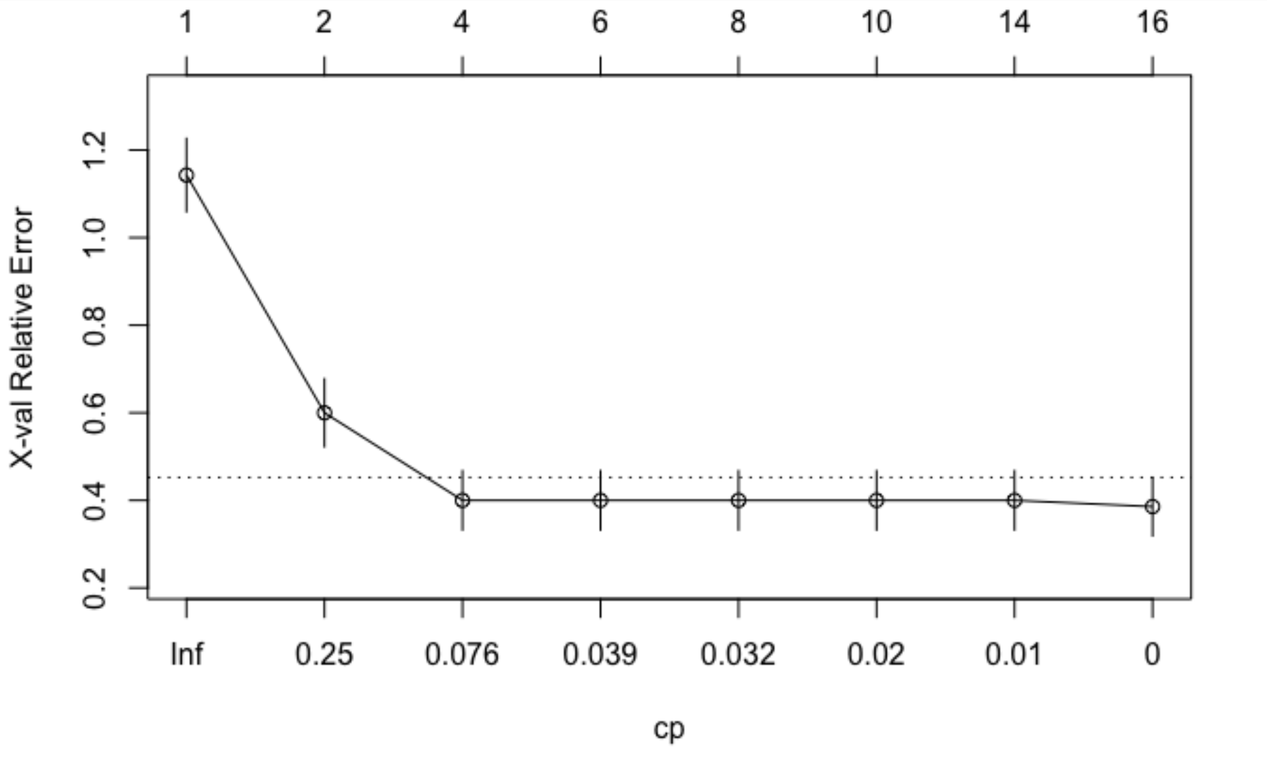
Arbre de classification :

- 2 observations dans chaque feuille et sans contrainte sur la qualité du découpage
- Arbre complet :



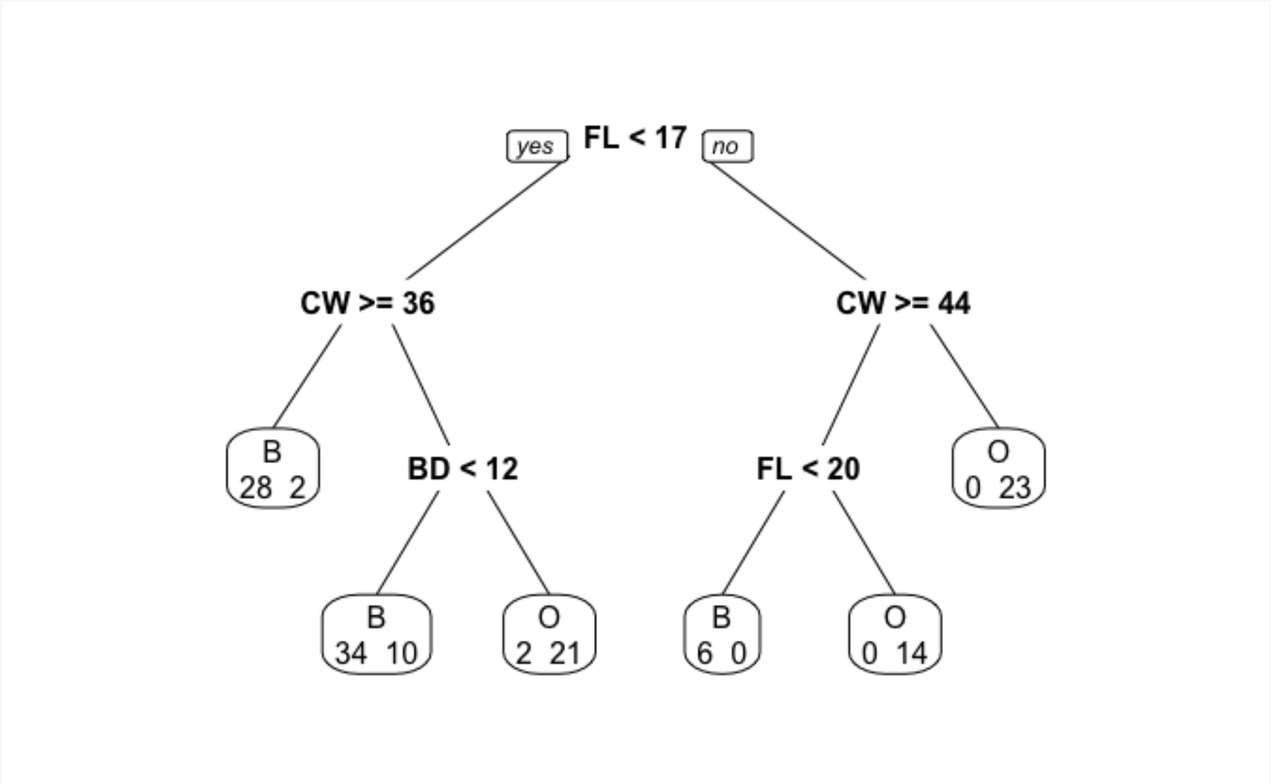
Arbre de classification :

- Calcul compléxité optimale : par validation croisée



Arbre simplifié :

- Automatisation du calcul de la complexité optimale
- Arbre simplifié :



Performances de l'arbre :

- Prédiction et distribution des espèces prédites sur l'échantillon test :

```
pred <- predict(modtree2, newdata = crabsTest, type = "class")
print(table(pred))
```

```
## pred
##  B  O
## 34 26
```

Performances de l'arbre :

Résultats :

- Qualité de prédiction dépend de l'espèce
- Taux de prédiction correcte total = 87%
- Oranges = 82%
- Bleus = 92%
- Modèle très efficace, réalise de bonnes prédictions sur le jeu d'apprentissage

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  O
##          B 28  6
##          O  2 24
##
##                Accuracy : 0.8667
##                  95% CI : (0.7541, 0.9406)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 2.603e-09
##
##                   Kappa : 0.7333
##
##  Mcnemar's Test P-Value : 0.2888
##
##             Sensitivity : 0.8000
##             Specificity : 0.9333
##          Pos Pred Value : 0.9231
##          Neg Pred Value : 0.8235
##              Prevalence : 0.5000
##          Detection Rate : 0.4000
##    Detection Prevalence : 0.4333
##       Balanced Accuracy : 0.8667
##
##        'Positive' Class : O
##
```

Affichage de la matrice de confusion et indicateurs d'évaluation