Université de Carthage
Ecole Supérieure de la Statistique et de l'Analyse de l'Information

Examen de Data Mining

3 ème année du cycle de formation d'ingénieurs

Durée de l'épreuve : 1 heure 30 - Documents non autorisés
Nombre de pages : 4 - Date de l'épreuve : 24 janvier 2020

On considère le jeu de données contenu dans le fichier "ronfle.txt" portant sur un échantillon de 100 patients pour lesquels on a relevé les mesures suivantes : AGE (en années), ALCOOL (en nombre de verres bus), SEXE (F ou H), RONFLE (ronflement : O = ronfle ; N = ne ronfle pas), TABAC (tabac : O = fumeur ; N = non fumeur).
Les statistiques descriptives de ce jeu de données sont résumées dans le tableau suivant :

```
> don<-read.table(file ="ronfle.txt", header = T)
> summary(don)
      AGE             ALCOOL       SEXE   RONFLE TABAC
 Min.   :23.00   Min.   : 0.00   F:25   N:65   N:36
 1st Qu.:43.00   1st Qu.: 0.00   H:75   O:35   O:64
 Median :52.00   Median : 2.00
 Mean   :52.27   Mean   : 2.95
 3rd Qu.:62.25   3rd Qu.: 4.25
 Max.   :74.00   Max.   :15.00
```

On cherche à expliquer/prédire la variable RONFLE à l'aide des autres variables.

Partie I

Dans cette partie, on a expliqué la variable RONFLE à l'aide d'un arbre de décision.
Les résultats obtenus sont présentés ci-dessous :

```
> modele_AD<- rpart(RONFLE ~ AGE+ALCOOL+SEXE+TABAC, data = don, minsplit=15)
> print(modele_AD)
n= 100

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 100 35 N (0.65000000 0.35000000)
   2) AGE< 63.5 77 21 N (0.72727273 0.27272727)
     4) ALCOOL< 1.5 34  4 N (0.88235294 0.11764706) *
     5) ALCOOL>=1.5 43 17 N (0.60465116 0.39534884)
      10) AGE>=57.5 5  0 N (1.00000000 0.00000000) *
      11) AGE< 57.5 38 17 N (0.55263158 0.44736842)
        22) AGE< 49.5 24  8 N (0.66666667 0.33333333)
          44) ALCOOL< 5.5 12  1 N (0.91666667 0.08333333) *
          45) ALCOOL>=5.5 12  5 O (0.41666667 0.58333333) *
```

```
       23) AGE>=49.5 14  5 0 (0.35714286 0.64285714) *
   3) AGE>=63.5 23  9 0 (0.39130435 0.60869565)
     6) ALCOOL< 3.5 14  6 N (0.57142857 0.42857143) *
     7) ALCOOL>=3.5 9  1 0 (0.11111111 0.88888889) *
> printcp(modele_AD)

Classification tree:
rpart(formula = RONFLE ~ AGE + ALCOOL + SEXE + TABAC, data = don,
    method = "class", minsplit = 15)

Variables actually used in tree construction:
[1] AGE    ALCOOL

Root node error: 35/100 = 0.35

n= 100

        CP nsplit rel error xerror    xstd
1 0.142857      0   1.00000 1.0000 0.13628
2 0.057143      1   0.85714 1.0857 0.13868
3 0.038095      2   0.80000 1.1429 0.13997
4 0.010000      6   0.62857 1.0286 0.13714
> pred1 <- predict(modele_AD, newdata = don, type = "class")
> MC1 <- table(?, ?)
> print(MC1)
   pred1
     N  0
  N 54 11
  0 11 24
```

1. Indiquer la signification du paramètre `minsplit`.

2. Indiquer le nombre de règles générées par cet arbre.

3. En utilisant l'arbre obtenu, classer l'individu ayant les caractéristiques suivantes en donnant la règle qui a permis de le classer :
(AGE=57 ; ALCOOL= 5 ; SEXE = F ; TABAC=N)

4. A partir du tableau fourni par la commande "`printcp(modele_AD)`" déterminer, en justifiant votre réponse, le nombre de noeuds terminaux de l'arbre optimal.

5. Compléter la fonction `table` par les paramètres adéquats afin d'obtenir la matrice de confusion `MC1` évaluant les prédictions de `modele_AD` sur les données.

6. On aurait voulu utiliser la méthode Random Forest pour expliquer la variable RONFLE. Donner les 3 paramètres à préciser afin d'effectuer cette méthode.

Partie II

On a aussi effectué une régression logistique afin d'expliquer la variable RONFLE. Les résultats obtenus sont présentés ci-dessous :

```
> modele_RL1<- glm(RONFLE ~ AGE+ALCOOL+SEXE+TABAC, family=binomial,don)
> modele_RL1

Call:  glm(formula = RONFLE ~ AGE + ALCOOL + SEXE + TABAC, family = binomial,
    data = don)

Coefficients:
(Intercept)          AGE       ALCOOL         SEXEH       TABACO
   -4.48413      0.06258      0.23373      0.64018     -1.17352

Degrees of Freedom: 99 Total (i.e. Null);  95 Residual
Null Deviance:      129.5
Residual Deviance: 109.7        AIC: 119.7

> TC1<-table(predict(modele_RL1, don,type="response")>0.5,don[,4])
> TC1

        N  O
  FALSE 54 21
  TRUE  11 14
```

7. En utilisant modele_RL1, classer l'individu ayant les caractéristiques suivantes : (AGE=42 ; ALCOOL= 0 ; SEXE = F ; TABAC=N)

8. Donner les commandes à executer pour évaluer le modèle donné par la régression logistique par une validation croisée.

9. Nous avons effectué une sélection pas à pas " forward ". Les résultats obtenus sont présentés ci-dessous. Expliquer le principe de la sélection pas à pas " forward ".

```
> modele_simple <- glm(RONFLE ~  1, data = don,"binomial")
> modele_RL2<-step(modele_simple, scope = ~ AGE+ALCOOL+SEXE+TABAC,dir = "forward")
Start:  AIC=131.49
RONFLE ~ 1

          Df Deviance    AIC
+ AGE      1   123.51 127.51
+ ALCOOL   1   124.00 128.00
+ SEXE     1   125.97 129.97
<none>         129.49 131.49
+ TABAC    1   128.40 132.40

Step:  AIC=127.51
RONFLE ~ AGE

          Df Deviance    AIC
+ ALCOOL   1   114.80 120.80
+ SEXE     1   120.25 126.25
<none>         123.51 127.51
+ TABAC    1   122.86 128.86
```

```
Step:  AIC=120.8
RONFLE ~ AGE + ALCOOL

        Df Deviance    AIC
+ TABAC  1   110.66 118.66
<none>       114.80 120.80
+ SEXE   1   114.52 122.52

Step:  AIC=118.66
RONFLE ~ AGE + ALCOOL + TABAC

       Df Deviance    AIC
<none>      110.66 118.66
+ SEXE  1   109.72 119.72
> TC2<-table(predict(modele_RL2, don,type="response")>0.5,don[,4])
> TC2

         N  O
  FALSE 56 21
  TRUE   9 14
```

10. Comparer les variables sélectionnées par l'arbre de décision et celle de la sélection pas à pas " forward " de la régression logistique.

11. Evaluer la qualité des 3 modèles `modele_AD`, `modele_RL1` et `modele_RL2`.