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

**Examen de Data Mining**

**Session principale 2024**

**2 ème année du cycle de formation d'ingénieurs**

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

Le traitement du cancer de la prostate change si le cancer a atteint ou non les noeuds lympatiques entourant la prostate. Pour éviter une investigation lourde un certain nombre de variables sont considérées comme explicatives de la variable $Y$ : $Y = 0$ si le cancer n'a pas atteint le réseau lympatique et $Y = 1$ sinon. Le but de cette étude est donc d'expliquer et de prédire $Y$ par les variables suivantes :

- `age` : âge du patient au moment du diagnostic;
- `acide` : le niveau d'acide phosphate sérique;
- `rayonx` : le résultat d'une analyse par rayon X, 0= négatif et 1= positif;
- `taille` : la taille de la tumeur, 0= petite et 1= grande;
- `grade` : l'état de la tumeur déterminé par biopsie, 0= moyen et 1= grave;
- `log.acid` : le logarithme népérien du niveau d'acidité;

On dispose de la base de données `cancer_prostate` (fichier cancerprostate.txt) constituée de 53 individus. Chacun des 53 individus est décrit par les 6 variables prédictives présentées ci-dessus ainsi que par sa valeur sur la variable $Y$. Ci-dessous les statistiques descriptives des données :

```
> cancer_prostate<-read.table("cancerprostate.txt",sep=";",header=T)
> summary(cancer_prostate)
      age            acide         rayonx taille grade  Y         log.acid
 Min.   :45.00   Min.   :0.4000   0:38   0:26   0:33   0:33   Min.   :-0.9163
 1st Qu.:56.00   1st Qu.:0.5000   1:15   1:27   1:20   1:20   1st Qu.:-0.6931
 Median :60.00   Median :0.6500                               Median :-0.4308
 Mean   :59.38   Mean   :0.6942                               Mean   :-0.4189
 3rd Qu.:65.00   3rd Qu.:0.7800                               3rd Qu.:-0.2485
 Max.   :68.00   Max.   :1.8700                               Max.   : 0.6259
```

## Partie 1

Afin d'expliquer $Y$, on a réalisé une régression logistique à l'aide du logiciel $R$.

Dans une première étape nous avons obtenu les résultats suivants :

```
> modele1 = glm(Y ~., family=binomial,cancer_prostate)
> summary(modele1)

Call:
glm(formula = Y ~ ., family = binomial, data = cancer_prostate)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0960  -0.6102  -0.2863   0.4834   2.2000

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.08672    7.83450   1.287   0.1979
age         -0.04289    0.06166  -0.696   0.4867
acide       -8.48006    7.63305  -1.111   0.2666
rayonx       2.06673    0.85469   2.418   0.0156
taille       1.38415    0.79546   1.740   0.0819
grade        0.85376    0.81247   1.051   0.2933
log.acid     9.60912    6.21652   1.546   0.1222
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 44.768  on 46  degrees of freedom
AIC: 58.768
```

Par la suite, nous avons effectué une régression logistique pas à pas *backward* sur cette base pour obtenir ce que nous appelons modele2. Les résultats sont donnés ci-dessous :

```
> modele2<-step(?,?)
Start:  AIC=58.77
Y ~ age + acide + rayonx + taille + grade + log.acid

           Df Deviance    AIC
- age       1   45.259 57.259
- grade     1   45.883 57.883
- acide     1   46.560 58.560
<none>          44.768 58.768
- taille    1   47.949 59.949
- log.acid  1   48.126 60.126
- rayonx    1   51.368 63.368

Step:  AIC=57.26
Y ~ acide + rayonx + taille + grade + log.acid

           Df Deviance    AIC
- grade     1   46.425 56.425
<none>          45.259 57.259
- acide     1   47.776 57.776
- taille    1   48.300 58.300
- log.acid  1   49.615 59.615
- rayonx    1   51.742 61.742

Step:  AIC=56.43
Y ~ acide + rayonx + taille + log.acid
```

```
          Df Deviance    AIC
<none>          46.425 56.425
- acide     1   48.986 56.986
- log.acid  1   50.660 58.660
- taille    1   51.246 59.246
- rayonx    1   53.707 61.707

> predict(modele2, cancer_prostate[1:10,])
         1          2          3          4          5
-3.30721429 -2.49149192 -3.07950459 -2.86846373 -3.07950459
         6          7          8          9         10
-3.19119121 -1.45961312  0.06972952 -2.49149192 -0.48704273
```

1) Remplacer les " ?" de la commande **step** par les paramètres adéquats pour effectuer la régression logistique pas à pas.

2) Expliquer le principe de la sélection pas à pas utilisée ci-dessus.

3) Comparer les résultats du modele2 à ceux du modele1.

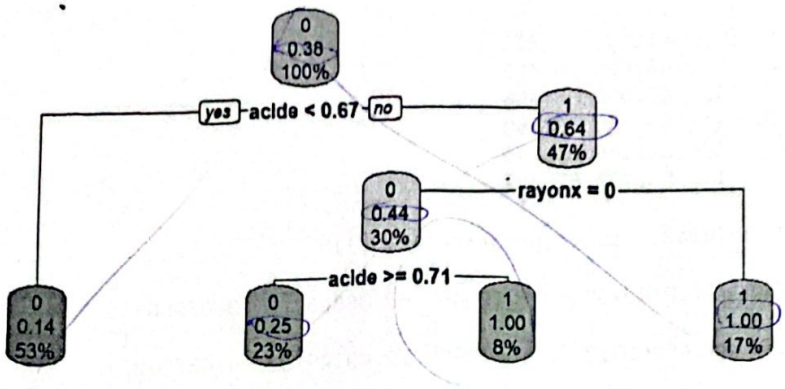
4) A partir des résultats des modèles 1 et 2, déterminer la classe d'affectation de l'individu 1 sachant qu'il a les caractéristiques suivantes : `age=66 ; acide=0.48 ; rayonx= 0 ; taille= 0 ; grade = 0 ; log.acid= -0.73`.

## Partie 2

Afin d'expliquer $Y$, nous avons aussi effectué un arbre de classification sur le logiciel R. Les résultats sont présentés ci-dessous :

```
> modele3 <- rpart(Y~ ., data = cancer_prostate, method = "class",minsplit=5)
> printcp(modele3)

Classification tree:
rpart(formula = Y ~ ., data = cancer_prostate, method = "class",
    minsplit = 5)

Variables actually used in tree construction:
[1] acide  age    rayonx

Root node error: 20/53 = 0.37736

n= 53

     CP nsplit rel error xerror    xstd
1 0.350      0      1.00   1.00 0.17644
2 0.150      1      0.65   1.00 0.17644
3 0.050      3      0.35   0.65 0.15662
4 0.025      5      0.25   0.85 0.16991
5 0.010      7      0.20   0.85 0.16991
```

5) A partir de ces résultats, donner la commande qui permet d'obtenir l'arbre optimal à partir de l'arbre du modele3.




6) On considère l'arbre donné par la figure ci-dessus. Commenter cet arbre puis donner les règles qui en découlent.

7) A partir de cet arbre, déterminer la classe d'affectation de l'individu 1 de la question 4).

## Partie 3

Afin d'expliquer $Y$, nous avons enfin effectué un *random forest* sur le logiciel R. Les résultats sont présentés ci-dessous :

```
> modele5 <- randomForest(Y~.,data=cancer_prostate, mtry= 3,ntree=500)
> modele5$confusion
   0  1 class.error
0 26  7   0.2121212
1  9 11   0.4500000
> imp <- importance(modele5)
> imp
         MeanDecreaseGini
age              3.710594
acide            6.488794
rayonx           3.852525
taille           2.078109
grade            2.005427
log.acid         6.238672
```

8) Expliquer le lien entre le choix de la valeur du paramètre `mtry` et le taux d'erreur réel du modèle donné par le *random forest*.

9) Expliquer comment a-t-on obtenu la matrice de confusion donnée par `modele5$confusion` ?

10) Commenter les résultats de `importance(modele5)`.