CY TECH	**ING2-MI** **EXAMEN DE DATAMINING 2** **2022-2023**
Durée : 2h	Examen papier Calculatrice autorisée 4 feuilles R/V

Le barème est donné à titre indicatif mais est susceptible d'évoluer.

Exercice 1 : 3 points

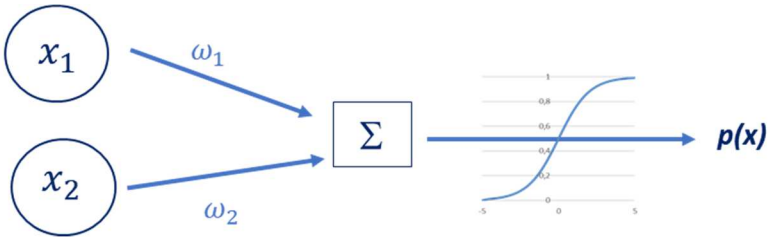
Soit la matrice de confusion suivante obtenue sur une base de 125 observations

		Predicted values		
		A	B	C
True values	A	51	2	1
	B	3	58	0
	C	2	3	5

1) Quel est le taux de bien classés ?
2) Que pouvez-vous dire sur la performance du modèle ?

Exercice 2 : 5 points

Soit y une variable cible binaire. Considérons le réseau de neurones sans biais suivant,



où Σ est la combinaison linéaire à laquelle on applique la fonction d'activation sigmoïde, $f(x) = 1/(1 + e - x)$.

1) Que représente $p(x)$ la sortie du réseau de neurones ?
2) Ecrire la sortie du réseau de neurones en fonction des entrées x_1 et x_2.

Dans l'algorithme de rétropragation du gradient, on initialise les poids à $w_1 = w_2 = 1$ et on considère la formule de mise à jour des poids

$$\Delta w_i = \alpha[y - p(\boldsymbol{x})]x_i$$
$$w_i \leftarrow w_i + \Delta w_i$$

avec un taux d'apprentissage $\alpha = 0.9$.

Soit la base d'apprentissage suivante,

	x_1	x_2	y
1	0.5	0.6	1
2	2.6	1.2	0

3) Quelle est la sortie du réseau de neurones pour la ligne 1 (arrondi à 10^{-2})?
4) Que deviennent les poids à l'issue de la 1ère itération (ligne 1) ?
5) Que deviennent les poids à l'issue de la 2ème itération (ligne 2) ?

Exercice 2 : Etude de cas 12 points

Afin de construire un modèle pour prédire le cancer de la prostate, on considère une base d'apprentissage de 100 patients caractérisés par 8 variables numériques (Radius, Texture, Perimeter, Area, Smoothness, Compactness, Symmetry, Fractal dimension) et une variable cible binaire représentant la présence ou l'absence

de cellules cancéreuses (Y=diagnostic_result). La classe positive (Y=1) est notée « M » pour Malignant et la classe négative (Y=0) est notée « B » pour Begnin.

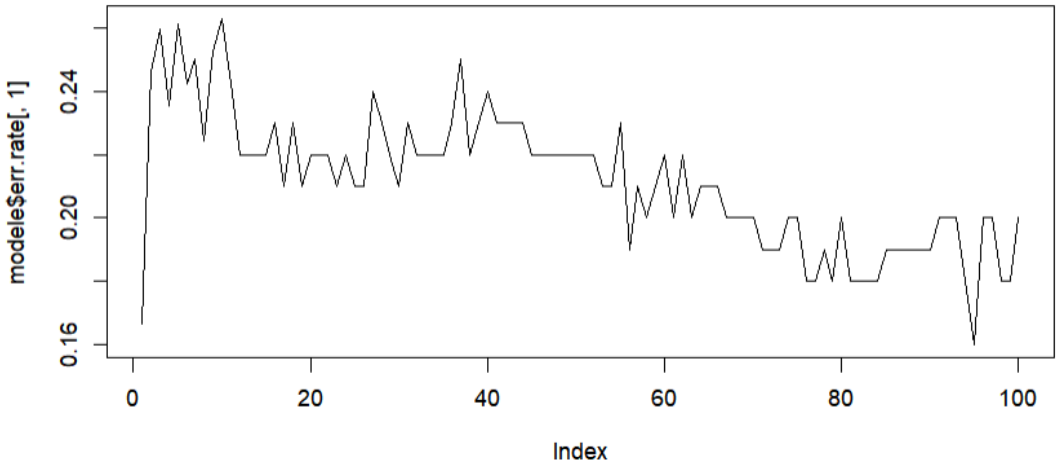
Partie 1. Foret aléatoire

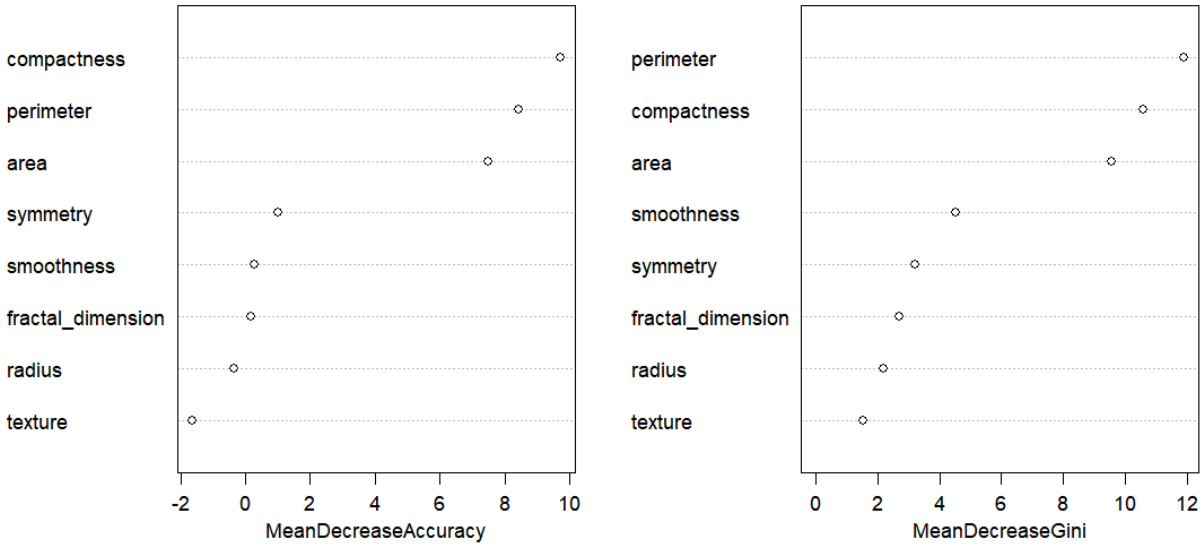
On ajuste une forêt aléatoire avec le package RandomForest de R dont les résultats sont donnés ci-dessous.

1) Combien y-a-t-il d'arbres dans la forêt ? Pensez-vous que cela est suffisant ou bien faut-il ajouter des arbres (justifiez votre réponse) ?
2) Combien de variables sont testées (mises en concurrence) à chaque nœud de chaque arbre ? Expliquez ce chiffre.
3) Quelles sont les variables les plus importantes pour prédire la présence ou l'absence de cellules cancéreuses ?
4) Expliquez pourquoi les forêts aléatoires sont des algorithmes stochastiques (aléatoires) ?
5) Donnez 3 hyperparamètres qui ont un impact sur le temps de calcul.
6) Calculer le taux de bien classés, la specificity (recall) et la precision.

```
Call:
 randomForest(formula = diagnosis_result ~ ., data = tab, importance = T,      ntree = 100)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 2

        OOB estimate of  error rate: 20%
Confusion matrix:
   B  M class.error
B 29  9   0.2368421
M 11 51   0.1774194
```





Partie 2. Régression logistique

On construit une régression logistique. Ci-dessous les résultats obtenus pour un modèle avec toutes les variables.

1) Dans la procédure de sélection de modèle pas-à-pas, quelle variable allez-vous supprimer du modèle à la première itération ?
2) Quand arrête-t-on la procédure pas-à-pas ?

```
Call:
glm(formula = diagnosis_result ~ ., family = "binomial", data = tab)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4909  -0.3416   0.1959   0.4555   1.5749

Coefficients:
                  Estimate  Std. Error z value Pr(>|z|)
(Intercept)       0.878336   14.700016   0.060   0.9524
radius           -0.020060    0.069691  -0.288   0.7735
texture           0.079146    0.069698   1.136   0.2561
perimeter         0.094814    0.205670   0.461   0.6448
area             -0.003468    0.013239  -0.262   0.7934
smoothness      -20.139671   29.154818  -0.691   0.4897
compactness      46.217251   23.240042   1.989   0.0467 *
symmetry         -4.738381   19.109754  -0.248   0.8042
fractal_dimension -161.459595  127.108181  -1.270   0.2040
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132.81  on 99  degrees of freedom
Residual deviance:  66.24  on 91  degrees of freedom
AIC: 84.24

Number of Fisher Scoring iterations: 6
```

Sur la figure ci-dessous, on a le modèle obtenu à l’issue de la procédure de sélection de modèle pas-à-pas.

```
Call:
glm(formula = diagnosis_result ~ compactness + fractal_dimension,
    family = "binomial", data = tab)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1308  -0.3953   0.1658   0.5044   1.6818

Coefficients:
                 Estimate Std. Error z value    Pr(>|z|)
(Intercept)        11.689      3.644   3.207    0.001339 **
compactness        64.141     13.025   4.925 0.000000845 ***
fractal_dimension -289.415     74.620  -3.878    0.000105 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132.813  on 99  degrees of freedom
Residual deviance:  69.894  on 97  degrees of freedom
AIC: 75.894

Number of Fisher Scoring iterations: 6
```

3) Est-ce que les variables importantes à la classification sont les mêmes que pour le modèle de foret aléatoire ?
4) Ecrire le modèle obtenu pour $p(x) = P(Y = 1|x)$ où Y est la variable cible (on rappelle que « Y=1 » équivaut à « diagnostic_result=M ») et $x = (x_1, \dots, x_p)$ sont les variables retenues par le modèle. En déduire la probabilité que les cellules soient malignes (M) pour les valeurs suivantes :

radius	texture	perimeter	area	smoothness	compactness	symmetry	fractal_dimension
23	12	151	954	0.143	0.278	0.242	0.079

5) On obtient la matrice de confusion suivante.

prev		
	B	M
B	31	7
M	6	56

Comparer les performances de la régression logistique et de la forêt aléatoire à l'aide du taux de bien classés.

Partie 3. Naive Bayes

On utilise maintenant la méthode Naive Bayes et on obtient les résultats suivants sur le modèle.

1) Quelles sont les deux hypothèses sur les variables $x = (x_1, \dots, x_p)$ que nous devons supposer pour appliquer le modèle Naive Bayes ?
2) A quoi correspond la ligne « A-priori probabilities » ?
3) Dans la parte « Conditional probabilities » à quoi correspondent les colonnes 1 et 2 des lignes B et M pour la variable radius ?

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
   B    M
0.38 0.62

Conditional probabilities:
   radius
Y       [,1]     [,2]
  B 17.94737 5.061499
  M 16.17742 4.678252

   texture
Y       [,1]     [,2]
  B 17.76316 5.185396
  M 18.51613 5.218950

   perimeter
Y       [,1]     [,2]
  B  78.5000 17.47856
  M 107.9839 19.71559

   area
Y       [,1]     [,2]
  B 474.3421 219.6037
```

4) Comparer la performance de ce modèle avec les précédents à l'aide de la matrice de confusion suivante,

	B	M
B	32	6
M	14	48