CY TECH	**ING2-MI** **EXAMEN DE DATAMINING 2** **2021-2022**
Durée : 2h	Examen papier Calculatrice autorisée 4 feuilles R/V

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

Exercice 1 : Questions rapides — 6 points

1) Parmi les méthodes de classification suivantes, quelle(s) est (sont) celle(s) qui sont stochastiques ?
   - Naïve Bayes
   - Arbres de décision
   - Forêts aléatoires
   - Régression logistique
   - Réseaux de neurones

2) Soit un problème de classification avec une variable cible binaire. On construit un modèle de prévision sur une base d'apprentissage et on obtient la matrice de confusion suivante sur une base test.

		Valeurs prédites	
		0	1
Vraies valeurs	0	2	2
	1	3	93

   a) Quel est le taux de bien classés ?
   b) Peut-on conclure que le modèle est bon ?

3) On considère un ensemble d'apprentissage $D=\{(x_i,y_i)\}$ tel que x_i est décrit par p variables explicatives $X_1,\ldots,X_p$ et y_i appartient {C1, C2, C3}. Nous souhaitons utiliser D pour créer un classifieur bayésien.

   a) Quelles sont les probabilités qui doivent être calculées pour définir le classificateur ?
   b) Quelles seront les entrées et les sorties de ce classifieur une fois défini ?
   c) Pour valider le classifieur, nous utilisons un ensemble de test $T=\{(x_j,y_j)\}$. Comment procéder pour réaliser cette validation ? On précisera en particulier une métrique utilisée pour cela.
   d) Dans certains cas, le résultat de cette validation n'est pas satisfaisant pour la raison suivante : certaines probabilités calculées par l'algorithme sont nulles. Comment expliquez-vous cela ? Comment remédier à ce problème ?

Exercice 2 : Etude de cas — 10 points

The dataset is designed to use geographical (Fire location), temporal (Month and Day), Fire Weather variables (FFMC, DMC,DC,ISI) and weather variables (RH, Temp, Rain, Wind) to predict the area burned by forest fires. Data were obtained from the UCI Machine Learning databaseand contain details for 517 fires found in the Montesinho Natural Park in Portugal.

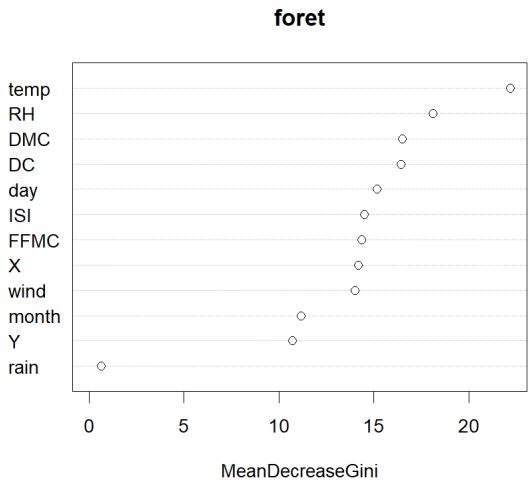
1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6

8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. fire - "yes" if the area is burned and "no" elseif (270 yes – 247 no)

1) Random Forest

Ci-dessous les sorties R obtenues avec le package randomForest.

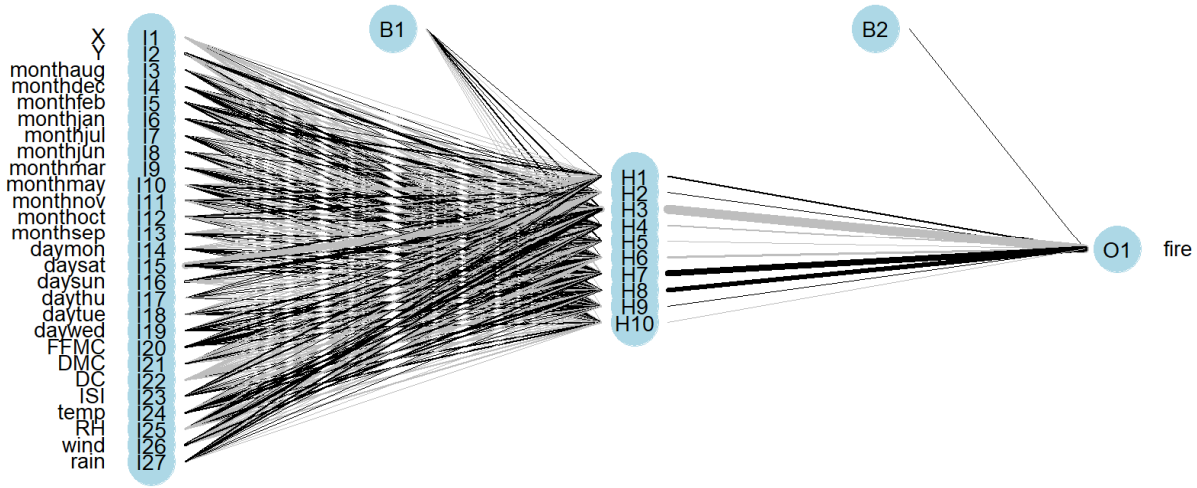
```
> library(randomForest)
> foret=randomForest(fire~.,data=train)
> print(foret)

Call:
 randomForest(formula = fire ~ ., data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 46.22%
Confusion matrix:
    no yes class.error
no  86  84   0.4941176
yes 75  99   0.4310345
> varImpPlot(foret)
```



a) A quoi correspond la ligne : `No. of variables tried at each split: 3`? Comment est choisi ce nombre ?
b) Expliquez comment est calculée l'erreur OOB (Out Of Bag).
c) Calculer l'erreur sur l'ensemble d'apprentissage à partir de la matrice de confusion. Conclusion.
d) Quelles variables contribuent le plus dans cette forêt aléatoire ?

2) Neural network

Ci-dessous le réseau de neurones obtenu avec R pour une couche cachée et 10 neurones. La matrice de confusion est calculée sur la base d'apprentissage.



```
> print(MatConf)
     prev
       no yes
  no   33 137
  yes   4 170
```

a) Combien y-a-t-il de neurones en entrée ? Justifier.
b) Combien y-a-t-il de poids dans ce réseau de neurones ? Justifier.
c) Pour une nouvelle entrée, le réseau de neurones retourne 0,128. Qu'est-ce que cela signifie ? Quelle est la classe prédite ?
d) Calculer l'erreur sur l'ensemble d'apprentissage à partir de la matrice de confusion. Conclusion.

3) Régression logistique

Ci-dessous les résultats de la régression logistique après sélection des variables avec la procédure `step`.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.63249  -1.13912   0.00037   1.08156   1.72993

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.04014    0.92741  -1.122   0.2621
X              0.09280    0.04919   1.887   0.0592 .
monthaug      -0.39687    0.87846  -0.452   0.6514
monthdec      17.04207  977.54530   0.017   0.9861
monthfeb       0.58324    0.97869   0.596   0.5512
monthjan     -15.93027 2399.54487  -0.007   0.9947
monthjul      -0.08298    0.95153  -0.087   0.9305
monthjun      -0.42012    1.03421  -0.406   0.6846
monthmar      -0.97819    0.89951  -1.087   0.2768
monthmay      -0.02848    1.63621  -0.017   0.9861
monthnov     -16.56972 2399.54486  -0.007   0.9945
monthoct      -0.57547    1.05461  -0.546   0.5853
monthsep       0.28775    0.86712   0.332   0.7400
temp           0.04127    0.02652   1.556   0.1196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 476.70  on 343  degrees of freedom
Residual deviance: 443.38  on 330  degrees of freedom
AIC: 471.38
```

a) Ecrire le modèle.
b) Que prédit le modèle pour une position X=1 et une température de 20° au mois d'aout.
c) Quel est l'odds ratio de la température ? Qu'est-ce que cela signifie ?
d) Calculer l'erreur sur l'ensemble d'apprentissage à partir de la matrice de confusion. Conclusion.

```
> MatConf=table(train$fire,prev)
> print(MatConf)
     prev
        0   1
  no   97  71
  yes  53 123
```

e) Quelle sont les variables importantes dans ce modèle ? Est-ce cohérent avec les résultats obtenus avec la forêt aléatoire ?

Exercice 3 : Naive Bayes — 4 points

On cherche à prédire une variable cible binaire Y en fonction de deux variables continues X_1 et X_2.
L'objectif est donc de trouver

$$\max_{k\in\{0,1\}} P(Y = k|X_1 = x_1, X_2 = x_2)$$

1) Expliquer pourquoi et sous quelle(s) hypothèse(s), le problème est équivalent à

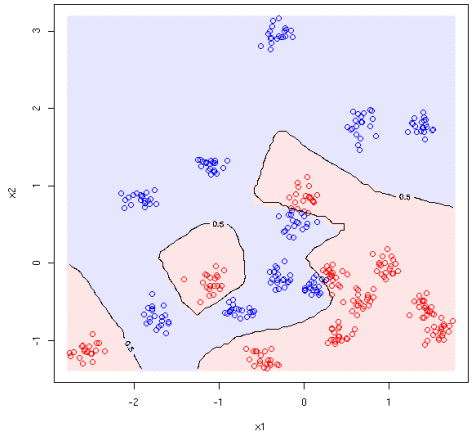
$$\max_{k\in\{0,1\}} P(Y = k) \times f_{X_1|Y=k}(x_1) \times f_{X_2|Y=k}(x_2)$$

où $f_{X_i|Y=k}$ est la fonction de densité de la variable X_i conditionnellement à la classe k.

Sur une base d'apprentissage de 100 individus, nous obtenons les résumés numériques du tableau ci-dessous.

		X_1		X_2	
	effectif	moyenne	variance	moyenne	variance
Y=0	40	1	4	-1	1
Y=1	60	0	4	2	9

1) Donnez une estimation de P(Y=k) pour k=0 et k=1
2) Si on suppose que $X_i|Y=k$ suit une loi normale (rappel de la fonction de densité de la loi normale à la fin de l'exercice), écrivez les fonctions de densité de X_1 et X_2 à l'intérieur de chaque classe.
3) Quelle classe sera prédite pour un nouvel individu tel que $x_1=0$ et $x_2=0$? Justifiez votre résultat.
4) La base d'apprentissage est représentée sur le graphique ci-dessous. Pensez-vous que la méthode de classification Naïve Bayes est adaptée au problème ? Si non, quelle(s) méthode(s) proposez-vous ? Justifiez votre réponse



N.B. Pour $X \sim N(\mu,\sigma^2)$ la fonction de densité est $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$