Contrôle TD – Apprentissage

1h. **Polycopiés de cours autorisés.**

Prenez soin de lire tous les exercices avant de commencer. La notation est donnée à titre indicatif.

Exercice 1 : (8 pts)

classe prédite / classe réelle	A	B	C	Total
A	35	1	2	**38**
B	2	30	0	**32**
C	3	2	25	**30**
Total	**40**	**33**	**27**	**100**

La *précision* pour une classe donnée mesure le taux d'exemples corrects parmi les exemples prédits dans cette classe. Le *rappel* mesure le taux d'exemples corrects parmi les exemples de la classe. Le taux de *faux positifs* d'une classe mesure le nombre d'objets positifs parmi ceux n'appartenant pas à la classe. Le taux de *vrais positifs* d'une classe mesure le nombre d'objets positifs parmi les vrais objets de la classe. La *sensibilité* est la probabilité qu'un test soit positif si l'objet appartient à la classe. La *spécificité* est la probabilité qu'un test soit négatif si l'objet n'appartient pas à la classe.

Soit la matrice de confusion ci-dessus, obtenu par application d'un modèle M1 de fouille de données. Après avoir **rappelé la formule de calcul**, déterminer :

1- Le taux d'erreur en généralisation
2- La précision pour la classe A.
3- Le rappel pour la classe B.
4- Le taux de faux positifs (FP rate) pour la classe C.
5- Le taux de vrais positifs (TP rate) pour la classe A.
6- La sensibilité pour la classe B
7- La spécificité pour la classe C
8- La F-mesure de la classe A.

Exercice 2 : **Classification supervisée** (12 pts)

On considère la base « essai » décrite dans le fichier donné en annexe 1 dans lequel l'attribut de classe est « Class ». Cette base est découpée en 2 sous-ensembles D1 (valeur **paire** pour Customer_ID) et D2 (valeur **impaire** pour Customer_ID). **D1** est utilisé en **apprentissage**, et **D2** est l'ensemble de **test**.

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Pour chaque question, expliquer la démarche avant d'effectuer la résolution.

Travail demandé.

1- L'attribut Customer_ID est-il pertinent pour discriminer les deux classes C0 et C1 ? **Pourquoi** ?
2- On choisit de ne pas tenir compte de l'attribut Customer_ID pour la suite de cet exercice. Utiliser l'index de GINI pour construire un modèle d'arbre de décision permettant d'affecter une classe à un étudiant.
3- On va évaluer le modèle que vous avez construit sur l'ensemble D2. Construire la matrice de confusion sur l'ensemble D2.
4- Indiquer l'erreur apparente en test de votre modèle.

ANNEXE 1

```
@relation essai

@attribute Customer_ID    integer
@attribute Gender         {M, F}
@attribute Car_Type       {Family, Sports, Luxury}
@attribute Shirt_Size     {S, M, L, XL}
@attribute Class          {C0, C1}

@data
%
% 20 instances
%
1, M, Family, S, C0
3, M, Sports, M, C0
5, M, Sports, XL, C0
7, F, Sports, S, C0
9, F, Sports, M, C0
11, M, Family, L, C1
13, M, Family, M, C1
15, F, Luxury, S, C1
17, F, Luxury, M, C1
19, F, Luxury, M, C1
%
%
%
2, M, Sports, M, C0
4, M, Sports, L, C0
6, M, Sports, XL, C0
8, F, Sports, S, C0
10, F, Luxury, L, C0
12, M, Family, XL, C1
14, M, Luxury, XL, C1
16, F, Luxury, S, C1
18, F, Luxury, M, C1
20, F, Luxury, L, C1
```