Éléments de correction CC Apprentissage -2017

Exercice 1 :

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**

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

1. **(1 pt)**

- Le taux d'erreur en généralisation : **10%**

- accuracy rate : **90%**

2. La précision pour la classe A ; **(1 pt)**

P_A = (A,A) / (.,A) = 35/40

3. Le rappel pour la classe B ; (**1 pt)**

R_B = (B,.B) / (B, .) = 30/32

4. Le taux de faux positifs (FP rate) pour la classe C ; **(1 pt)**

FP_C = ((A,C) + (B,C)) / ((A,.) + (B,.)) = (2 + 0) /(38 + 32) = 2/70

5 . Le taux de vrais positifs (TP rate) pour la classe A ; **(1 pt)**

VP_A = (A,A) / (A,.) = 35/38

6. La sensibilité pour la classe B ; **(1 pt)**

$Sensi_B$ = VP_B = (B,B) / (B,.) = 30/32

7. La spécificité pour la classe C ; **(1 pt)**

$Speci_C$ = VN_C = ((A,A) + (B,B)) / ((A,.) + (B,.)) = (35+30)/(38+32) = 65/70

8. La F-mesure de la classe A. **(1 pt)**

$F\text{-}mesure_A$ = (2 * R_A * P_A) / (R_A + P_A) = (2 * 35/38 * 35/40) / (35/38 + 35/40)

Exercice 2 :

1. L'attribut **Custmer_ID** n'est pas pertinent car sa valeur est unique pour chaque enregistrement (**Overfitting**) **(1 pt)**

2. Détermination de la racine de l'arbre de décision

	Gendre		**Car_Type**			**Shirt_Size**			
	M	**F**	**Family**	**Sports**	**Luxury**	**S**	**M**	**L**	**XL**
C0	3	2	0	4	1	1	1	2	1
C1	2	3	1	0	4	1	1	1	2
GINI	**0,48**	**0,48**	**0**	**0**	**0,32**	**0,5**	**0,5**	**0,44**	**0,44**
	0,48		**0,16**			**0,464**			

$GINI_{Gendre}(M) = 1 - (3/5)^2 - (2/5)^2$; $GINI_{Gendre}(F) = 1 - (2/5)^2 - (3/5)^2$

$GINI_{split}$(Gendre) = 5/10 * (1 – (3/5)² – (2/5)²) + 5/10 * (1 – (2/5)² – (3/5)²) = 0,48 **(1 pt)**

GINI(Gendre) = 1 – (5/10)² – (5/10)² = 0,5 **(0,5 pt)**

$GINI_{Car\_Type}(Family) = 1 - (0/1)^2 - (1/1)^2 = 0$; $GINI_{Car\_Type}(Sports) = 1 - (4/4)^2 - (0/4)^2 = 0$;

$GINI_{Car\_Type}(Luxury) = 1 - (1/5)^2 - (4/5)^2 = 0,32$

$GINI_{split}$(Car_Type) = 0 + 0 + 5/10 * (1 – (1/5)² – (4/5)²) = 0,16 **(1 pt)**

GINI(Car_Type) = 1 – (5/10)² – (5/10)² = 0,5 **(0,5 pt)**

$GINI_{Shirt\_Size}(S) = 1 - (1/2)^2 - (1/2)^2$; $GINI_{Shirt\_Size}(M) = 1 - (1/2)^2 - (1/2)^2$;

$GINI_{Shirt\_Size}(L) = 1 - (2/3)^2 - (1/3)^2$; $GINI_{Shirt\_Size}(XL) = 1 - (1/3)^2 - (2/3)^2$

$GINI_{split}$(Shirt_Size) = 2/10 * (1 – (1/2)² – (1/2)²) + 2/10 * (1 – (1/2)² – (1/2)²) + 3/10 * (1 – (2/3)² – (1/3)²) + 3/10 * (1 – (1/3)² – (2/3)²) ≈ 0,464 **(1 pt)**

GINI(Shirt_Size) = 1 – (5/10)² – (5/10)² = 0,5 **(0,5 pt)**

- L'attribut **Car_Type** a le plus petit index GINI, on choisit **Car_Type** comme la racine de l'arbre de décision.
- Partitionnement avec les attributs non-choisis **(1 pt)**

En partant de nœud **Car_Type**, on obtient les étudiants {2,4,6,8} qui sont de classe C0 en prenant la branche « Sports », {12} qui est de classe C1 en prenant la branche «Family» et {10,14,16,18,20} qui ne sont pas de même classe en prenant la branche «Luxury»,

Nous devons traiter les lignes suivantes :

Ligne	**10**	**14**	**16**	**18**	**20**
Gendre	F	M	F	F	F
Shirt_Size	L	XL	S	M	L
Classe	C0	C1	C1	C1	C1

Nous observons les attributs Gendre et Shirt_Size :

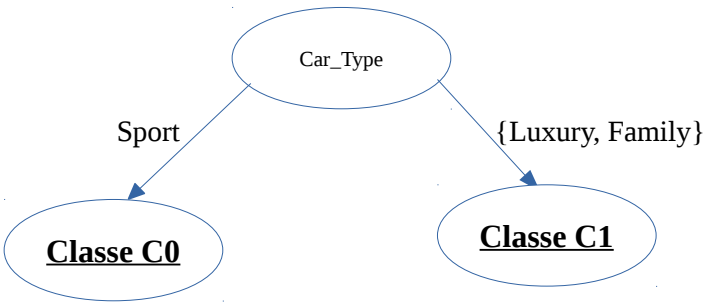
	Gendre		**Shirt_Size**			
	M	**F**	**S**	**M**	**L**	**XL**
C0	0	1	0	0	1	0
C1	1	3	1	1	1	1
GINI	**0**	**0,375**	**0**	**0**	**0,5**	**0**
	0,3		**0,2**			

GINI(Gendre) = 0,32 ; $GINI_{split}$(Gendre) = 0,3 ;

GINI(Shirt_Size) = 0,32 ; $GINI_{split}$(Shirt_Size) = 0,2

- L'attribut **Shirt_Size** a le plus petit index de GINI, cependant, ces différentes branches contiennent très peu d'éléments. Il est donc moins pertinent.
- En choisissant l'attribut **Gendre,** le gain n'est pas vraiment important après la division en deux branches (« M » et « F »), car **GINI(Gendre) ≈ $GINI_{split}$(Gendre).** **(1 pt)**

On obtient l'arbre de décision suivant :



3 - Évaluation du modèle M1 sur D2. **(2 pts)**

Customer_ID	Classe réelle	M1
1	C0	C1 (faux)
3	C0	C0
5	C0	C0
7	C0	C0
9	C0	C0
11	C1	C1
13	C1	C1
15	C1	C1
17	C1	C1
19	C1	C1

3 – Matrice de confusion sur D2 **(2 pts)**

	Classes prédites			
Classes réelles		**C0**	**C1**	**Total**
	C0	4	1	**5**
	C1	0	5	**5**
	Total	**4**	**6**	**10**

4. Erreur apparente : $Erreur_{App}(M1) = 0,1$ **(1pt)**