Université Mohamed Khider-Biskra
Faculté des sciences exactes et des sciences de la nature et de la vie
Département d'informatique

Master SIOD — Fouille de Données Avancée
07 Mars 2021 — 10:40-11:40, S3,S4

Examen

On dispose de la base d'exemples suivante représentant les résultats d'analyse de 15 patients concernant l'infection par le corona virus COVID19 :

N°	AGE	DR	FV	TX	FAT	COVID19
1	V	Oui	Non	Oui	Oui	+
2	J	Oui	Oui	Non	Non	+
3	V	Non	Oui	Non	Non	-
4	E	Non	Non	Oui	Oui	+
5	J	Oui	Non	Oui	Non	-
6	E	Oui	Oui	Non	Oui	-
7	V	Non	Non	Oui	Oui	+
8	V	Oui	Oui	Non	Non	-
9	J	Non	Oui	Oui	Non	-
10	J	Non	Non	Oui	Oui	+
11	V	Oui	Oui	Non	Non	+
12	J	Oui	Non	Oui	Oui	-
13	V	Non	Oui	Oui	Non	-
14	E	Non	Non	Oui	Oui	+
15	J	Oui	Oui	Non	Non	-

- N° : Numéro du patient
- DR : Difficulté respiratoire (Oui, Non)
- TX : Toux (Oui, Non)
- AGE : l'âge du patient (J :Jeune, V :Vieux, E :enfant)
- FV : Fièvre (Oui, Non)
- COVID 19 : Résultat du test PCR (infecté par le virus COVID19 : + (oui), - (non))

On vous demande de :

1. Règles d'association
   (a) Construire un modèle de décision en utilisant la méthode OneR sur les exemples de 1 à 10.
   (b) Calculer sa précision et son rappel sur la base d'entrainement (composée des exemples de 1 à 10) et sur la base de test (composée des exemples de 11 à 15).
   (c) Commenter le résultat.
2. Arbre de décision
   (a) Construire, en utilisant les exemples de 1 à 10, l'arbre de décision permettant de prédire l'infection par le virus COVID-19 en utilisant l'algorithme ID3 avec le Gain d'information pour le choix des attributs.
   (b) Calculer la précision et le rappel du modèle sur la base d'entrainement puis sur la base de test composée des exemples de 11 à 15.
   (c) Commenter le résultat.
3. Comparer les deux modèles.

Bonne Chance

Pr A.Djeffal

Corrigé type

1. Règles d'association **(7 pt)**
   (a) Construire un modèle de décision en utilisant la méthode OneR sur les exemples de 1 à 10.
   - Modèle OneR **(2.5 pts)**

AGE	+	-
V	2	2
J	2	2
E	1	1
Err		**5**

DR	+	-
Oui	2	3
Non	3	2
Err		**4**

FV	+	-
Oui	1	4
Non	4	1
Err		**2**

TX	+	-
Oui	4	2
Non	1	3
Err		**3**

FAT	+	-
Oui	4	1
Non	1	4
Err		**2**

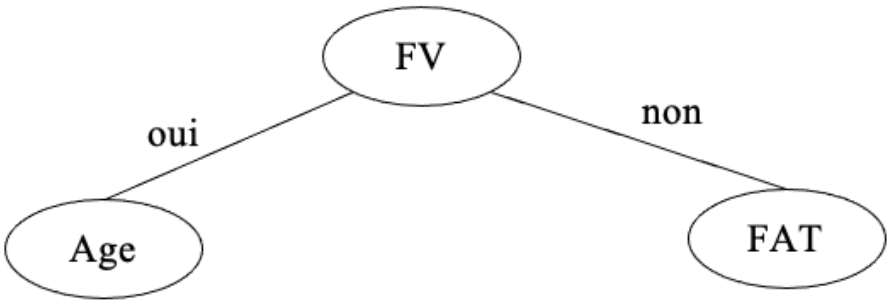
   - **Le modèle :** **(2 pts)**
     - **Si FV = Oui Alors COVID19 = -**
     - **Sinon COVID19 = +**

   (b) Précision
   - Sur la base d'entrainement = 8/10 = 80% **(0.25 pt)**
   - Sur la base de test = 3/5 = 60% **(0.25 pt)**

   (c) Rappel
   -
   - Sur la base d'entrainement = CP/(CP + FN) = 4/(4+1) = 80% **(0.5 pt)**
   - Sur la base de test = CP/(CP + FN) = 1/(1+1) = 50% **(0.5 pt)**

   (d) Commenter le résultat : Le modèle représente une modèle surentrainé (sur-apprentissage) **(1 pt)**

2. Arbre de décision **(12 pts)**
   (a) Construction de l'arbre de décision
   - H(covid)=H(5,5)=1 **(0.5 pt)**
     H(Age)=4/10 H(V) + 4/10 H(J) +2/10 H(E) = 4/10 H(2,2) + 4/10 H(2,2) +2/10 H(1,1)
     = 0.4 + 0.4 + 0.2 = 1
     Gain(Age) = 1 - 1 = 0 **(0.5 pt)**
     H(DR) = 5/10 H(oui) + 5/10 H(non) = 5/10 H(2.3) + 5/10 H(2.3) = 0.971
     Gain(DR) = 1 - 0.971 = 0.029 **(0.5 pt)**
     H(FV) = 5/10 H(oui) + 5/10 H(non)
     = 5/10 H(4.1) + 5/10 H(4.1) = 0.722 **Variable de segmentation**
     Gain(FV) = 1 - 0.722 = 0.278 **(0.5 pt)**
     H(TX) = 6/10 H(oui) + 4/10 H(non) = 6/10 H(4.2) + 4/10 H(3.1) = 0.875
     Gain (TX) = 1 - 0.875 = 0.125 **(0.5 pt)**
     H(FAT) = 5/10 H(oui) + 5/10 H(non) = 5/10 H(4.1) + 5/10 H(4.1) = 0.722
     Gain(FAT) = 0.278 **(0.5 pt)**



(0.5 pt)

– **FV =Non**

H(COVID10)= H(4,1)= 0.722 **(0.5 pt)**

H(AGE) = 2/5H(J) + 2/5H(V) + 1/5H(E) = 2/5H(1,1) + 2/5H(2,0) + 1/5H(1,0) = 0.4

Gain(AGE) = 0.722 - 0.4 = 0.322 **(0.5 pt)**

H(DR) = 3/5H(Oui) + 2/5H(Non) = 3/5H(2,1) + 2/5H(2,0) = 0.55

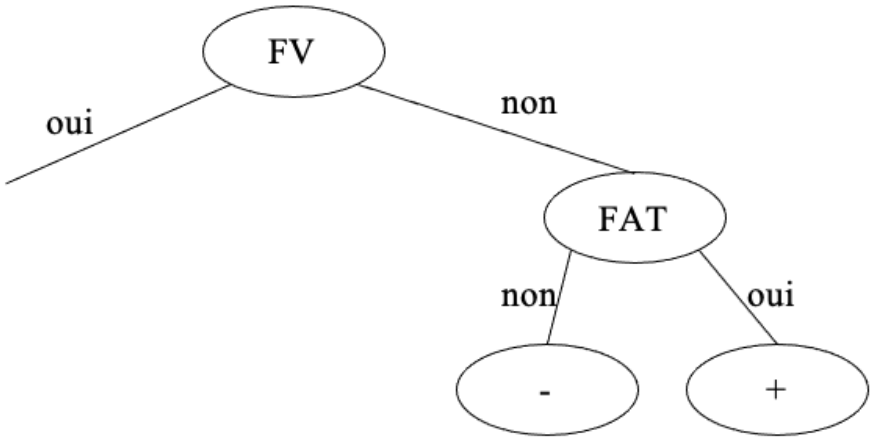
Gain(DR) = 0.722 - 0.55 = 0.17 **(0.5 pt)**

H(TX) = 4/5H(3,1) + 1/5H(1,0) = 0.64

Gain(TX) = 0.722 - 0.64 = 0.08 **(0.5 pt)**

H(FAT) = 4/5H(Oui) + 1/5H(Non) = 4/5H(4,0) + 1/5H(1,0) = 0

Gain(FAT) = 0.722 - 0 = 0.722 **Variable de segmentation** **(0.5 pt)**



– **FV =Oui**

H(COVID10)= H(4,1)=0.722 **(0.5 pt)**

H(AGE) = 2/5H(J) + 2/5H(V) + 1/5H(E) = 2/5H(1,1) + 2/5H(2,0) + 1/5H(1,0) = 0.4

Gain(AGE) = 0.322 **Variable de segmentation**

H(DR) = 3/5H(Oui) + 2/5H(Non) = 3/5H(2,1) + 2/5H(2,0) = 0.55

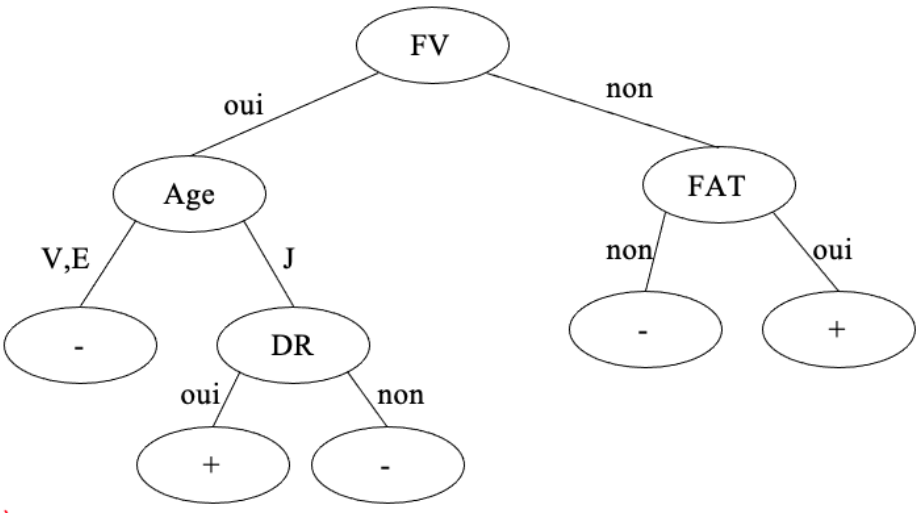
Gain(DR) = 0.171 **(0.5 pt)**

H(TX) = 4/5H(3,1) = 0.722

Gain(TX) = 0 **(0.5 pt)**

H(FAT) = 1/5H(Oui) + 5/5H(Non) = 1/5H(1,0) + 4/5H(3,1) = 0.64

Gain(FAT) = 0.0722 **(0.5 pt)**



(1.5 pt)

(b) Précision

- Sur la base d'entrainement = 10/10 = 100% **(0.25 pt)**
- Sur la base de test = 2/5 = 40% **(0.25 pt)**

(c) Rappel

- Sur la base d'entrainement = CP/(CP + FN) = 5/5 = 100% **(0.5 pt)**
- Sur la base de test = CP/(CP + FN) = 1/(1+1) = 50% **(0.5 pt)**

(d) Commenter le résultat : Le modèle représente une modèle surentrainé (sur-apprentissage) **(1 pt)**

3. Le deuxième modèle est plus précis que le premier. **(1 pt)**