Université de Maroua
Faculté des Sciences
Département de Maths-Info

Parcours : Data Science
Niveau : M2
Année : 2022 – 2023

EXAMEN DE WEB MINING

Code UE : **DSI 553**

EPREUVE THEORIQUE

Par : Pr Kaladzavi / M. Touza Isaac

Durée : 3h

Exercice 1 : Web Content Mining — 7pts

Le tableau suivant contient des données sur différentes pages web. Chaque page est décrite par plusieurs attributs tels que le temps de chargement (en secondes), le nombre de visites, le taux de rebond et la catégorie. L'objectif est de construire un modèle d'arbre de décision à partir des données du tableau afin de prédire la catégorie des pages web en fonction de leurs caractéristiques. Les lignes de 1 à 12 sont utilisées comme données d'apprentissage pour construire le modèle, tandis que les lignes restantes (de 13 à 16) servent de données de tests pour évaluer les performances du modèle.

N°	Page Web	Temps de chargement (en secondes)	Nombre de visites	Taux de rebond	Catégorie
1	Page A	2.3	100	20%	Bon
2	Page B	4.1	80	40%	Excellent
3	Page C	3.5	120	15%	Moyen
4	Page D	1.8	150	10%	Bon
5	Page E	5.2	60	55%	Faible
6	Page F	3.9	90	30%	Excellent
7	Page G	2.6	110	18%	Moyen
8	Page H	1.5	180	5%	Bon
9	Page I	2.8	70	25%	Excellent
10	Page J	4.3	130	12%	Moyen
11	Page K	3.2	95	35%	Bon
12	Page L	1.9	200	8%	Faible
13	Page M	4.7	50	60%	Excellent
14	Page N	3.4	115	17%	Moyen
15	Page O	2.1	140	9%	Bon
16	Page P	1.6	175	6%	Faible

1. Définir arbre de décision **0,5pt**
2. On souhaite utiliser l'algorithme ID3 pour construire l'arbre de décision en utilisant les données d'apprentissage de ce tableau.
   2.1. Montrez toutes les étapes de calcul pour la construction l'arbre de décision puis dessinez-le. **3pts**

2.2. Donner la catégorie d'une page web Q dont le temps de chargement est de 2 secondes et ayant été consultée par 150 utilisateurs et le taux de rebond est de 29%. **0,5pt**

2.3. Quels sont les résultats de test de l'arbre obtenu sur les données des lignes de 13 à 16 ? Ces résultats correspondent-ils à ceux du tableau ci-dessus ? **1pt**

2.4. Construire la matrice de confusion pour ce modèle de classification **1pt**

2.5. Déduire le taux d'erreur et de succès de l'arbre de décision construit. **1pt**

NB : subdiviser les valeurs de chacune des variables : Temps de chargement, Taux de rebond et Nombre de visites en deux groupes selon les critères spécifiés :

- Temps de chargement : <= 2.8 et > 2.8
- Taux de rebond <= 12% et > 12%
- Nombre de visites <= 90 et > 90

Exercice 3 : Le Web Structure Mining **5pts**

Considérons un site web contenant 5 pages, représentées par les identifiants de page suivants : P1, P2, P3, P4 et P5. Les liens entre ces pages sont les suivants :

- P1 contient des liens vers P2 et P3.
- P2 contient un lien vers P4.
- P3 contient un lien vers P4.
- P4 contient des liens vers P2 et P5.
- P5 contient un lien vers P3.

1. Dessinez un graphe où chaque page est représentée par un nœud, et les liens entre les pages sont représentés par des arêtes. Utilisez les informations fournies pour connecter les nœuds correspondant aux pages concernées. **1pt**
2. Quel nom donne-t-on à cette représentation ? **0,5pt**
3. Construisez la matrice d'adjacence A en utilisant la représentation du graphe. La matrice d'adjacence est une matrice carrée où chaque élément A[i, j] indique s'il existe un lien de la page i à la page j (1 pour un lien existant, 0 sinon). **1pt**
4. Utilisez une méthode de clustering, pour regrouper les pages en fonction de leurs liens. Choisissez le nombre de clusters à 2. **2,5pts**

Exercice 3 : Web Usage Mining **8pts**

Vous travaillez pour une entreprise de commerce électronique qui souhaite améliorer l'expérience utilisateur sur son site web. Votre tâche consiste à analyser les données des utilisateurs afin de comprendre le comportement des utilisateurs et d'identifier des opportunités d'optimisation du site. On met donc à votre disposition un ensemble de données contenant des journaux de navigation des utilisateurs. Chaque entrée du journal contient des informations telles que l'adresse IP de l'utilisateur, l'URL visitée, la date et l'heure de la visite, ainsi que d'autres informations pertinentes.

IP	URL	Date et heure
192.168.0.1	/home	2023-06-01 08:15
192.168.0.2	/products	2023-06-01 09:32
192.168.0.3	/home	2023-06-01 10:05
192.168.0.4	/products	2023-06-01 10:15
192.168.0.1	/cart	2023-06-01 11:25
192.168.0.5	/home	2023-06-01 12:10
192.168.0.2	/products	2023-06-01 12:40
192.168.0.3	/checkout	2023-06-01 13:20
192.168.0.6	/home	2023-06-01 14:05
192.168.0.4	/products	2023-06-01 14:30
192.168.0.5	/products	2023-06-01 15:15
192.168.0.1	/home	2023-06-01 16:00
192.168.0.3	/products	2023-06-01 16:45
192.168.0.6	/cart	2023-06-01 17:20
192.168.0.2	/products	2023-06-01 17:50
192.168.0.4	/home	2023-06-01 18:35
192.168.0.5	/products	2023-06-01 19:10
192.168.0.3	/checkout	2023-06-01 19:40
192.168.0.6	/home	2023-06-01 20:25
192.168.0.1	/products	2023-06-01 20:50

1. En se servant de l'adresse IP, identifiez les utilisateurs de ce site en précisant les ressources consultées. **1,5pt**
2. En considérant le seuil temporel $\Delta t = 2h$, déterminer pour chaque utilisateur ses visites. **1,5pt**
3. Donner la représentation horizontale de l'ensemble des données obtenues à la question 1. **1,5pt**
4. On veut utiliser l'algorithme Apriori pour déterminer les règles d'associations entre les pages :
   - 4.1. Définir règle d'association **0,5pt**
   - 4.2. Calculer le nombre d'itemsets fréquents qu'on peut former avec toutes ces pages. **0,5pt**
   - 4.3. Déterminer l'ensemble des pages fréquentes, sachant que le support minimal est de 2 pages **1pt**
   - 4.4. Déduire les règles d'association entre ces pages fréquents avec une confiance minimale de 80%. **1pt**
   - 4.5. Donner une interprétation d'une règle de votre choix. **0,5pt**