Université de Maroua
Faculté des Sciences
Département de Maths-Info

Parcours : Data Science
Niveau : M2
Année : 2023 – 2024

CONTRÔLE CONTINU DE WEB MINING

Code UE : **DSI 553**

EPREUVE PRATIQUE

Par : Pr Kaladzavi / M. Touza Isaac

Durée : 3h

Consignes de travail :

- L'épreuve est notée sur 20pts
- Le fichier final de votre travail doit porter votre nom
- Le code source doit être commenté
- Mentionner votre nom et matricule au début du document
- Le code source du script et les autres fichiers doivent être compressé et envoyer à l'adresse isaac_touza@outlook.fr

Travail à faire : **20pts**

1. Accédez au site suivant **https://www.lemonde.fr/sciences/** puis scraper à l'aide du navigateur les informations suivantes sur toutes les articles des différentes pages : le titre, l'image descriptive, le texte de l'article, date de publication ou de mise à jour (si elle existe), la nature de l'article (entretien, décryptages, récit, reportage, histoire, carte blanche, tribune, etc..) et le nom de l'auteur. Et exportez les données obtenues sous format CSV. **4pts**
2. Démarrer votre environnement de développement python puis charger les packages suivantes : Pandas, Numpy, Matplotlib et sklearn **1pt**
3. Charger dans une base des données nommées « articles_Sciences» les données de la question 1. **0,5pt**
4. Visualiser les 5 premiers articles. **0,5pt**
5. Afficher les statistiques d'articles publiés par auteur (utilisez un diagramme de votre choix si possible) **1pt**
6. Écrire un code python qui affiche les titres des articles publiés par **Nathaniel Herzberg** et leur date de publication. **1pt**
7. On souhaite créer un modèle de classification supervisée (On considère la nature des articles comme classe) afin de l'entrainer sur ces données. Former une base des données constituée uniquement des titres de l'article, date de publication, les noms des auteurs et de la nature des articles (Mettre la valeur « autres » pour les articles n'ayant pas de nature). **1pt**
8. Subdiviser l'ensembles des données en données d'entrainement et de test. Prendre 20% pour le test. **1pt**
9. Effectuer le prétraitement des valeurs de chaîne en les normalisant et en les convertissant en valeurs numériques. **2pts**
10. Construire deux modèles de classification, l'un bayésienne en utilisant la loi de Bernoulli et l'autre en utilisant l'algorithme KNN, puis entrainer vos modèles sur le jeu de données ci-dessus. **4pts**
11. Imprimer les rapports de classification puis évaluer les performances de chaque modèle utilisé. **2pts**
12. Générer les matrices de confusion pour ces deux modèles de classification. **2pts**