

Université de Sherbrooke
Département d'informatique

IFT870/BIN710
Forage de données / Forage de données pour la bio-informatique

Hiver 2025

Examen intratrimestriel

Enseignant :
Davy Ouedraogo

Le samedi 01 mars 2025
de 09 h 00 à 17 h 00
À remettre sur turnin.dinf.usherbrooke.ca

Cet examen est à faire de façon individuelle. Lors de la correction, la note zéro (0) sera attribuée à tout travail pour lequel une preuve de plagiat est attestée. L'utilisation des outils d'intelligence artificielle générative est interdite. Toute source utilisée dans la réalisation de cet examen doit être mentionnée. Pour la soumission du travail, se connecter dans un navigateur au serveur <http://turnin.dinf.usherbrooke.ca>, puis choisir le cours IFT870 (BIN710) et le projet ExamenIntra. Charger le fichier `examenintra.ipynb` et le soumettre. Le nom du fichier de remise doit être exactement `examenintra.ipynb`. Le fichier doit contenir une section qui mentionne votre nom, votre prénom et votre matricule ou CIP.

Cet examen comporte 4 questions et 4 pages.

Question 1: /20 points

Question 2: /25 points

Question 3: /45 points

Question 4: /10 points

Total: /100 points

NOM : _____.

PRÉNOM : _____.

MATRICULE : _____.

SIGNATURE : _____.

Collecte du jeu de données :

On vous fournit des données contenant des informations sur des revues de publication en libre accès. Ces données ont été utilisées pour créer la ressource <http://flourishoa.org/>

Récupérer les trois (3) tables du jeu de données sur GitHub :

<https://github.com/FlourishOA/Data>

Question 1 : Exploration-Description (20 pts)

- a) Présenter un bref résumé de chacun des attributs des trois (3) tables à la suite de votre étude quantitative et de votre visualisation des statistiques descriptives au besoin. (20 pts)

Question 2 : Exploration-Prétraitement (25 pts)

- a) Effectuer un prétraitement des données pour supprimer les duplications et corriger les incohérences s'il y en a. (15 points)
- b) Y a-t-il une corrélation entre les catégories de journaux (attribut « category ») et les coûts de publication (attribut « price ») ? Justifier la réponse. (10 pts)

Question 3 : Représentation-Classification-Régression (45 pts)

- a) Construire un modèle pour prédire les valeurs de catégories de journaux manquantes de la façon la plus précise possible. Cela inclut :
- i. La sélection d'attributs informatifs;
 - ii. Le choix et le paramétrage d'un modèle de classification;
 - Le but est de déterminer la meilleure valeur du paramètre « n_neighbors » parmi les entiers 1, 3 et 5 pour le modèle « KNeighborsClassifier » tel qu'implémenté par la librairie scikit-learn. Le modèle de classification utilisé doit être basé sur la distance *dString* telle que définit dans l'annexe [**Annexe**] à la page 4.
 - iii. Le calcul du score du modèle (le modèle avec les meilleurs paramètres obtenus précédemment);
 - iv. L'application du modèle pour prédire les catégories manquantes.

Selon vous, pourquoi le meilleur paramètre « n_neighbors » déduit serait le plus adéquat pour ce jeu de données? Justifier tous les choix effectués. (20 pts)

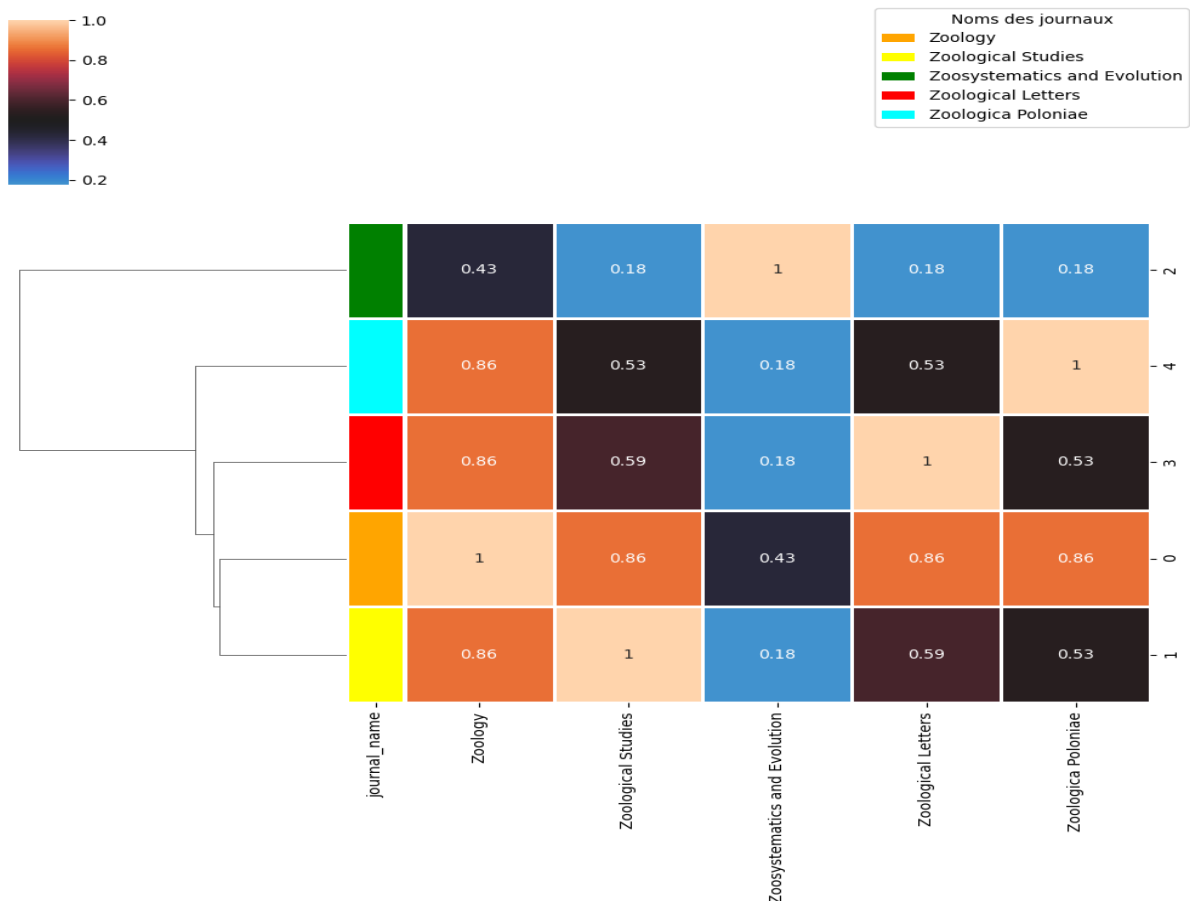
- b) Supprimer tous les attributs ayant plus de 50% de données manquantes. (5pts)
- c) Construire un modèle pour prédire le coût actuel de publication (attribut « price ») à partir des autres attributs. Cela inclut :
- i. La sélection d'attributs informatifs;
 - ii. Le choix et le paramétrage d'un modèle de régression

- Le but est de déterminer les meilleurs paramètres des deux (2) modèles suivants :
 - Le paramètre « n_neighbors » sur une plage de [1, 10] et le paramètre « metric » pour le modèle « KNeighborsRegressor » tel qu'implémenté par scikit-learn. Les distances à considérer pour en déduire la meilleure sont celles de Minkowski, de Manhattan et Euclidienne.
 - Les paramètres « max_depth » sur une plage de [5, 20] et « max_leaf_nodes » sur une plage de [10, 30] pour le modèle « DecisionTreeRegressor » tel qu'implémenté par scikit-learn.
- iii. Le calcul du score du modèle (le modèle avec les meilleurs paramètres obtenus précédemment). Lister les 10 journaux qui s'écartent le plus (en + ou -) de la valeur prédite. Que pouvez-vous conclure?
- iv. L'application du modèle pour prédire les coûts manquants.

Justifier les choix effectués. (20 pts)

Question 4 : Représentation-Visualisation (10 pts)

- a) On vous présente cette figure issue du prétraitement de ce jeu de données. Que représente cette illustration? Quelles conclusions pouvez-vous en tirer? (5pts)



- b) Proposer une démarche méthodologique et un script Python pour réaliser cette figure. (5pts)

Annexe ($dString$)

Soit a et b , deux chaînes de mots. $dString(a, b)$ représente la distance entre a et b et se définit comme suit :

$$dString(a, b) = 1 - \frac{common_words(a,b)}{\min(|words(a)|, |words(b)|)} \text{ avec}$$

$$common_words(a, b) = |words(a) \cap words(b)|$$

et

$\forall x \in \{a, b\}$, $words(x)$ représente l'ensemble des mots de x .

Exemple

Noter l'exclusion des mots : and, &... qui sont juste des délimiteurs des mots et ne doivent pas être considérés dans votre calcul de distance.

$a_1 = \text{Physics and Chemistry}$

$b_1 = \text{Physics}$

$c_1 = \text{Languages and Linguistics}$

$$dString(a_1, b_1) = 1 - \frac{|\{Physics\}|}{\min(|\{Physics, Chemistry\}|, |\{Physics\}|)} = 1 - \frac{1}{1} = 0$$

$$dString(a_1, c_1) = 1 - \frac{|\emptyset|}{\min(|\{Physics, Chemistry\}|, |\{Languages, Linguistics\}|)} = 1 - \frac{0}{2} = 1$$

-Fin de l'examen-