Université de Sherbrooke
Département d'informatique

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

Hiver 2020

Examen intratrimestriel

Professeure :
Aïda Ouangraoua

À remettre **avant le lundi 16 mars à 23h59**
sur **opus.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. Pour la soumission du travail, se connecter dans un fureteur au serveur http://opus.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.

Cet examen comporte 4 questions et 2 pages.

Question 1:	points
Question 2:	points
Question 3:	points
Question 4:	points

Total:	points

NOM : _____.

PRÉNOM : _____.

MATRICULE : _____.

SIGNATURE : _____.

Données :

On vous fournit des données contenant des informations sur des revues de publication « Open Access ». Ces données ont été utilisées pour créer la ressource http://flourishoa.org/

Récupérer les 3 tables du jeu de données sur GitHub : https://github.com/FlourishOA/Data

Question 1 : Exploration-Description (15 pts)

a) Présenter une description de chacun des attributs des 3 tables, avec des graphiques pour la visualisation des statistiques descriptives au besoin.

Question 2 : Prétraitement-Représentation (35 pts)

a) Effectuer un prétraitement des données pour supprimer les duplications et corriger les incohérences s'il y en a. (10 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)

c) Construire un modèle pour prédire les valeurs de catégorie de journaux manquantes de la façon la plus précise possible (cela inclut la sélection d'attributs informatifs, le choix et le paramétrage d'un modèle de classification, le calcul du score du modèle, l'application du modèle pour prédire les catégories manquantes). Justifier les choix effectués. (15 pts)

Question 3 : Régression-Clustering (50 points)

a) Supprimer tous les attributs ayant plus de 50% de données manquantes. (5pts)

b) Construire un modèle pour prédire le coût actuel de publication (attribut « price ») à partir des autres attributs (cela inclut la sélection d'attributs informatifs, le choix et le paramétrage d'un modèle de régression, le calcul du score du modèle, l'application du modèle pour prédire les coûts). Justifier les choix effectués.
Lister les 10 revues qui s'écartent le plus (en + ou -) de la valeur prédite. (15)

c) Construire un modèle pour grouper les revues suivant le coût actuel de publication (attribut « price ») et le score d'influence (attribut « proj_ai ») (cela inclut la détermination du nombre de clusters, le choix et le paramétrage d'un modèle de clustering, l'application du modèle pour trouver les clusters). Justifier les choix effectués. (15 pts)

d) Présenter des statistiques descriptives des clusters obtenus, et lister les revues du meilleur cluster en termes de rapport moyen : score d'influence / coût de publication. (15 pts)