UNIVERSITÉ DE SHERBROOKE
DÉPARTEMENT D'INFORMATIQUE

IFT 870 / BIN 710 - Forage de donnée

TP#2 : Prétraitement et représentation de données

Hiver 2025

Le but de ce devoir est de pratiquer le prétraitement et la représentation de données : auscultation, nettoyage, intégration, réduction.

Ce devoir est à faire en équipe de deux ou de trois. Il devra être complété avant le vendredi 14 Février 2025 à 23h59. Vous devez remettre, sur `turnin.dinf.usherbrooke.ca`, un fichier Ipython notebook (nommé `tp2.ipynb`) contenant votre rapport et vos scripts Python pour ce devoir.

Description des tâches à réaliser : Base de données des codes de médicaments au É-U

On vous fournit un jeu de données composé de deux tables au format `csv` : `product2.csv` et `package2.csv`. Vous pouvez trouver la description des attributs de ces tables aux adresses `https://www.fda.gov/drugs/drug-approvals-and-databases/ndc-product-file-definitions` et `https://www.fda.gov/drugs/drug-approvals-and-databases/ndc-package-file-definitions`.

1. Auscultez les données et présentez un résumé de votre auscultation (nombre d'attributs pour chaque table, types d'attributs, valeurs manquantes, incohérences intra-attribut, incohérences inter-attribut entre attributs reliés, vraissemblance et interprétabilité des attributs) ;
2. Listez les relations/règles observées entre les attributs (informations communes, chaînes de caractéres communes, attribut inclus dans un autre, ordre des valeurs) ;
3. Détectez et corrigez les incohérences entre des valeurs d'attributs dans les deux tables ; pour chaque règle identifiée à la question précédente, détectez et corrigez les cas où la règle n'est pas respectée ;
4. Proposez et appliquez une méthode pour compléter les données manquantes dans les deux tables ;
5. Détectez et retirez les objets dupliqués dans les deux tables ;
6. Intégrez les deux tables et nettoyez le résultat (données dupliquées, incomplètes, incohérentes, erronées) ;

7. Proposez un nouvel ensemble d'attributs (représentation) qui élimine la redondance des informations dans les valeurs des attributs, et qui permet de transformer l'attribut PHARM_CLASSES en un ensemble d'attributs distincts correspondant à ses différents champs EPC, CS, MOA, PE etc. ;
8. À partir de la nouvelle représentation, proposez un ensemble d'attributs à utiliser pour prédire le plus précisément possible toutes les classes pharmacologiques établies d'un médicament (champ EPC dans l'attribut PHARM_CLASSES) ;
9. En se basant sur la réduction de dimension obtenue à la question précédente, appliquez un modèle de classification pour prédire les classes pharmacologiques établies des médicaments pour lesquels l'information est manquante ;

Remise du travail

Pour soumettre votre travail, connectez-vous, dans un fureteur, au serveur `http://turnin.dinf.usherbrooke.ca` en utilisant votre CIP, puis choisissez le cours `IFT870 (BIN710)` et le projet `TP2`. Chargez votre fichier `tp2.ipynb` et soumettez-le. Le nom de votre fichier de remise doit être exactement `tp2.ipynb`. Indiquez bien les noms des deux, ou potentiellement trois membres de l'équipe dans le fichier. Ne faites qu'une seule soumission par équipe. Ne remettez pas d'autre fichier.