M2 LID – UGE
Projet Introduction à la Data Science

A partir de données nettoyées, transformées de la partie ETL du projet, vous allez maintenant expérimenter une approche d'apprentissage automatique avec la bibliothèque ml d'Apache Spark.

Vous allez utiliser l'approche supervisée dénommée arbre décisionnel (decision tree).
L'objectif est d'inférer l'idClass des médicaments, c'est_à-dire l'identifiant de la classe thérapeutique d'un médicament. Une classe thérapeutique, identifié par une valeur d'idClass, dépend de sa composition.
Pour les questions suivantes, vous fournirez systématiquement: une représentation de l'arbre décisionnel, une étude de la correction de la prédiction sur le jeu d'entrainement, la précision obtenue sur le jeu de données test et une explication en cas de problème observé. Pour chaque question, vous considérez des ensembles de classes thérapeutiques (idclass) et molécules différentes

Q6: idClass: 117, 372 et molécules: 94037 (nicotine), 2202 (paracétamol)
Q7 : idClass: 117, 372 et molécules: 94037, 2202, 2092 (ibuprofène)
Q8 : idClass: 117, 372, 394 et molécules: 94037, 2202, 2092, 1014
Q9 : idClass: 117, 372, 394, 204 et molécules: 94037, 2202, 2092, 1014, 24245

Q10: En fonction des problèmes rencontrés dans cette expérimentation, expliquez pourquoi la précision de la prédiction s'est dégradée.

Q11: Proposer 3 paires molécule, idClass qui peuvent être ajoutées à l'expérimentation et qui maintiendront une haute qualité de la prédiction.

Qx: paracétamol (2202), nicotine (94037), idclass (372,117)

Learned classification tree model:
DecisionTreeClassificationModel: uid=dtc_5154a4f06f4a, depth=1, numNodes=3, numClasses=2, numFeatures=1
If (feature 0 in {1.0})
Predict: 0.0
Else (feature 0 not in {1.0})
Predict: 1.0

Accuracy 100%

Qx': Qx + ibuprofène (2092)

Accuracy 100%