

TP N° 1

La fouille de données ou "data mining" est un processus qui permet de découvrir des connaissances inconnues au préalable, enfouies dans une masse volumineuse de données.

Dans le cadre du tp N°3 il est question de :

- lire un benchmark dans le but d'effectuer la tâche de classification de données : il faut évidemment procéder à la phase de preprocessing des données d'abord
- partager le benchmark en une Partie « apprentissage » (80%) et laisser 20% pour les tests, en tirant aléatoirement les instances et sans remise.
- appliquer l'algorithme KPPV ou K_NN pour l'ensemble de test , faire varier K de 1 à 10

Puis calculer les mesures de performances de cet algorithme pour le benchmark « Diabete » par exemple : TP, TN , FP, FN et les afficher , afficher la valeur de la precision, de rappel et Fmesure pour chaque valeur de K.

Dessiner la courbe de la precision du classificateur en fonction de k.

Determiner alors la meilleure valeur de K qui a donné la plus haute precision du classificateur.