Projet : Partie 2
Prétraitement des Données

Des données de mauvaise qualité conduiront à une extraction d'information de mauvaise qualité. C'est pourquoi la deuxième étape du Data Mining, et d'ailleurs la plus importante, est l'étape de prétraitement des données.

Le prétraitement de données inclut: Le **nettoyage des données** peut être appliqué pour éliminer le bruit et corriger les incohérences et valeurs aberrantes. L'**intégration des données** fusionne les données de plusieurs sources dans un seul dataset cohérent. La **réduction des données** permet de réduire la taille des données, via une agrégation, en éliminant les redondances ou en les discrétisant. Des **transformations de données** telles que la normalisation peuvent être appliquées, où les données sont mises à l'échelle. Cela peut améliorer la précision et l'efficacité des algorithmes de data mining impliquant des mesures de distance. Enfin, il est à noter que ces techniques ne s'excluent pas mutuellement; elles peuvent être appliquées ensemble.

Ainsi, dans cette deuxième partie du projet, il vous est demandé de procéder au prétraitement du dataset. Il est impératif d'utiliser l'analyse effectuée dans la partie 1 afin de vous guider dans cette 2ème partie. Le dataset obtenu après l'application de tous les traitements nécessaires doit être fonctionnel à 100% et le plus optimal possible pour les prochaines étapes. Le travail inclut:

A. Traitement des valeurs manquantes et aberrantes:
    a. Choix de la méthode de remplacement des valeurs manquantes.
    b. Choix de la méthode de traitement des valeurs aberrantes.

B. Réduction des données via la discrétisation des données continues:
    a. En classes d'effectifs égaux.
    b. En classes d'amplitudes égales.

C. Réduction des données (élimination des redondances) horizontales / verticales.

D. Normalisation des données:
    a. Méthode Min-Max.
    b. Méthode z-score.

Notes:

- Le rapport devra contenir le prétraitement du **Dataset 1** analysé en partie 1.
- Chaque binôme est tenu d'envoyer la version électronique du rapport au plus tard: **Samedi 29 Octobre 2022 à 23h59**.
- Chaque binôme devra présenter son interface et code source le **Dimanche 30 Octobre 2022** durant la séance de TP.

Bon courage !

HOUACINE Naila Aziza : n.a.houacine@gmail.com