Exercice 3
Pré-traitement des données

Discrétisation des données

Il peut être intéressant de discrétiser les attributs continus d'un dataset pour traiter les valeurs aberrantes (outliers) et lisser les données. La discrétisation consiste en la division des N valeurs possibles (généralement N>9) d'un attribut en un nombre fini de catégories.

Méthode 1 : Discrétisation en classes d'effectifs égaux (Quantiles, Equal-frequency)

- Diviser les N valeurs possibles en Q quantiles (Q à définir).
- La position du I^{ème} Quantile est égale à Position=N*i/Q.
- Toutes les valeurs appartenant à l'intervalle [Quantile Q_i , Q_{i+1}[sont représentées par une même catégorie $0 \leq i < Q$.

Méthode 2 : Discrétisation en classes d'amplitudes égales (Intervalles égaux, Equal-width)

- Définir ou calculer le nombre d'intervalle k à utiliser.
- La largeur de chaque intervalle de valeurs est égale à : $(Valeur_{max} - Valeur_{Min})/k$.
- Toutes les valeurs appartenant à un même intervalle sont représentées par une même catégorie.

Normalisation des données

La normalisation des données permet de modifier les N valeurs possibles de chaque attribut afin d'utiliser une échelle commune.

Méthode 1 : Normalisation avec Min-Max

$$Valeur_{(i,\ new)} = \frac{Valeur_{(i,old)} - Valeur_{(min,old)}}{Valeur_{(max,old)} - Valeur_{(min,old)}} (Valeur_{(max,\ new)} - Valeur_{(min,\ new)}) + Valeur_{(min,\ new)}$$

Méthode 2 : Normalisation avec z-score

$$Valeur_{(i,new)} = \frac{Valeur_{(i,old)} - Valeur_{(mean,old)}}{S} \quad ,S = \frac{1}{N}\sum_{i=1}^{N} | Valeur(i, old) - Valeur(mean, old) |^2$$

Questions :

1- Écrire une fonction Python permettant de discrétiser les valeurs du dataset "Dataset-Exos.txt" avec la Méthode 2. (Utiliser la formule de Huntsberger ⇒ K=1+(3/10)*log(n,base=10)).
2- Remplacer les valeurs discrétisées par la moyenne de l'intervalle correspondant.
3- Écrire une fonction Python permettant de normaliser les valeurs du dataset "Dataset-Exos.txt" avec la Méthode 1. (Tester avec $Valeur_{min,\ new}=0$, $Valeur_{max,\ new}=1$).

Belkadi Widad Hassina (G1-G2): belkadi.wh.usthb@gmail.com
Khelfa Celia (G3-G4): khelfacelia99@gmail.com