Exercise 3
Data Preprocessing

Data Discretization

Discretizing continuous attributes in a dataset can be useful for handling outliers and smoothing the data. Discretization involves dividing the N possible values (typically N > 9) of an attribute into a finite number of categories.

Method 1: Discretization into Equal-Frequency Intervals (Quantiles, Equal-Frequency)

- This method divides the N possible values into Q quantiles (where Q is defined).
- The position of the i-th quantile is calculated as: Position = N * i / Q.
- All values that fall within the interval [Quantile Q_i, Q_{i+1}) are represented by the same category, where $0 \leq i < Q$.

Method 2: Discretization into Equal-Width Intervals

- Define or calculate the number of intervals, k, to use.
- The width of each interval is equal to: (MaxValue - MinValue) / k.
- All values falling within the same interval are represented by the same category.

Data Normalization

Data normalization is the process of adjusting the N possible values of each attribute to use a common scale.

Method 1: Min-Max Normalization

The formula for Min-Max normalization is:

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

Method 2: Z-Score Normalization

The formula for Z-score normalization is:

$$Value_{(i, new)} = \frac{Value_{(i, old)} - Value_{(mean, old)}}{S} \text{ with } S = \frac{1}{N} \sum_{i=1}^{N} \left| Value_{(i, old)} - Value_{(mean, old)} \right|$$

Questions :

1- Write a function to discretize the values of the "DatasetExos.csv" using Method 2 (Equal-width intervals). Use Huntsberger's formula to calculate the number of intervals $K=1+(10/3)*\log_{10}(n)$.

2- Replace the discretized values with the average of the corresponding interval.

3- Write a function to normalize the values of the "DatasetExos.csv" using Method 1. Use $Value_{min, new}=0$ and $Value_{max, new}=1$.

Dataset: DatasetExos.csv is available on this link.