Business Analytics – 2024

By Farzad Minooei, PhD, MBA

Homework #6: Clustering

Q1: To answer the assignment, use hw_6.csv.

The file includes gene expression data that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

(a) Apply hierarchical clustering to the samples using ward, complete, average, and single linkage to separate the samples into the two groups.
(b) Apply k-means++ clustering to separate the samples into the two groups.
(c) Compare all models based on TWSS and Average Silhouette method.
(d) Can clustering model identify healthy group from diseased group? Do your results depend on the type of your model?