Data Mining 作業二 Clustering

此次作業主要目的在讓同學學習運用 Clustering Algorithms 及其效果評估。

1. 資料集
   (1) 來源：https://www.kaggle.com/datasets/mrmorj/dataset-of-songs-in-spotify
   (2) 說明：此資料集來自於全球最大的音樂串流網站 Spotify。資料集中共有 35,877 首音樂。每首音樂有 22 個欄位，包括 12 種音訊特徵(audio features)、音樂類別(genre)、歌曲名稱、網址、歌曲長度(duration)等。其中音樂類別共有 Trap, Techno, Techhouse, Trance, Psytrance, Dark Trap, DnB (drums and bass), Hardstyle, Underground Rap, Trap Metal, Emo, Rap, RnB, Pop and Hiphop 共 15 種類別。12 種音樂特徵包括 Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Time Signature.

2. 使用工具：請運用 Python sklearn.cluster, sklearn.mixture.GaussianMixture 實驗 K-means, Hierarchical Clustering, DBSCAN, GMM 的效果。

3. 參考網頁：
   https://scikit-learn.org/stable/modules/clustering.html
   https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

4. 繳交方式：作業每人繳交一份報告，檔案類型以 pdf 為限。上傳檔名格式為 學號_HW1，EX: 110753XXX_HW1.pdf.

5. 繳交期限：2022/12/18 23:59

6. 題目
   (1) 請列出每個 Audio Feature 的值域及其意義，同時觀察是否有 missing value 或 noise.
   (2) 如何做分群前的資料前處理(Preprocessing, 包括 Data Clean, Feature Normalization)？
   (3) 請執行 K-means，並列出 K-means 最佳分群的結果。結果包括 Elbow Diagram、參數、群數、每群的數量、主要的音樂類別、Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient, Confusion Matrix.
   (4) 請執行 Hierarchical Clustering，並列出最佳分群的結果。結果包括 Elbow Diagram、參數、群數、每群的數量、主要的音樂類別、Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient, Confusion Matrix.

(5) 請執行 DBSCAN，並列出最佳分群的結果。結果包括 Elbow Diagram、參數、群數、每群的數量、主要的音樂類別、Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient, Confusion Matrix.

(6) 請執行 GMM，並列出最佳分群的結果。結果包括 Elbow Diagram、參數、群數、每群的數量、主要的音樂類別、Rand Index, Normalized Mutual Information, Adjusted Mutual Information, V-measure, Fowlkes-Mallows Scores, Silhouette Coefficient, Confusion Matrix.

(7) 針對以上的分群方法，哪個分群方法效率最佳？為什麼？

(8) 有哪些可能的方法，可以提升分群的效果？