Data Mining 作業三 Classification

目的: 學習運用 Classification 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 supervised learning 實驗 Random Forest, Support Vector Machine 音樂類別分類的效果。

3. 參考網頁：
   https://scikit-learn.org/stable/supervised_learning.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_HW3，EX: 110753XXX_HW3.pdf.

5. 繳交期限：2023/01/21(除夕) 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) 如何做分類前的資料前處理(Preprocessing, 包括 Data Clean, Feature Normalization) ?
   (3) 請執行 Random Forest，並列出最佳分類的結果。結果包括 Imbalance 處理(Over-Sampling、Under-Sampling) 、Cross-Validation、Random Forest 參數、Accuracy、 Confusion Matrix、哪些類別的音樂彼此之間比較不易分別、Feature Importance、運用哪些方法提升分類準確率。（執行 Output Accuracy 的畫面，請截圖）
   (4) 請執行 Support Vector Machine，並列出最佳分類的結果。結果包括 Imbalance 處理(Over-Sampling、Under-Sampling)、Cross-Validation、 SVM 參數、Accuracy、 Confusion Matrix、哪些類別的音樂彼此之間比較不易分別、Feature Importance、運用哪些方法提升分類準確率。（執行 Output Accuracy 的畫面，請截圖）
   (5) 請根據 Linear SVM 的 Feature Importance，選出 Top-N 重要的 Features，並運用這些 Features 重新執行作業二的 Clustering，觀察效果是否有提升。（執行 Output 效果的畫面，請截圖）