**🎶 Progetto XAI: "Perché una canzone diventa un successo?"**

**1. Dataset consigliato**

**Spotify Tracks Genre** (Kaggle):  
Massimo ~114.000 brani, 20 colonne tra cui: popularity, danceability, energy, loudness, valence, tempo, genre, ecc. [medium.com+1ischool.berkeley.edu+1](https://medium.com/%40mansisu08/spotify-tracks-dataset-abf6d02cdc49?utm_source=chatgpt.com)

* **popularity** = punteggio da 0 a 100
* **audio\_features** = caratteristiche utili e facili da interpretare
* **genere** = categoria del brano

👉 Perfetto per classificazione o regressione (predire la popolarità o il genere)!

**2. Obiettivo del modello**

Puoi scegliere una di due strade:

**🎯 A) Regressione**

* **Target**: popularity
* Task: predire popolarità (contatore streaming, hit/non-hit)

**🎯 B) Classificazione Multiclasse**

* **Target**: genre (es. Pop, Jazz, Rock, Hip‑Hop...)
* Task: capire quali caratteristiche determinano il genere

**3. Modello consigliato**

* Per regressione: **Random Forest Regressor** o **XGBoost Regressor**
* Per classificazione: **Random Forest** o **XGBoost**
* Più gestibile: **Regressione**, dato che popularity è numerico

**4. Applicazione con SHAP**

**🔹 Summary Plot (Beeswarm)**

* Mostra le feature più influenti su popularity
* Es. danceability e loudness spingono in alto, tempo potrebbe abbassare

**🔹 Dependence Plot**

* Per scoprire trend su popularity, es. come cambia in base a valence

**🔹 Force / Waterfall Plot**

* Per una canzone specifica: es. “Questo brano ha pop=85 perché ha energia alta (+0.15) ma valence media (-0.05)”

**5. Risultati e insight attesi**

* Scoprire quali elementi musicali influenzano maggiormente la popolarità
* Identificare **soglie critiche** (es. danceability > 0.7 = +10 punti)
* Trovare **pattern non lineari** (es. valence molto basso o molto alto diminuisce la popolarità)
* Spiegare intuitivamente perché un brano sarà più o meno popolare

**6. Struttura del progetto**

1. **Import e preprocessamento**  
   Caricamento dati, gestione NaN, codifica delle variabili se serve
2. **Modellazione**  
   Split train/test, addestramento Random Forest
3. **XAI con SHAP**
   * Summary plot
   * Dependence plot per 1–2 feature
   * Waterfall/force plot per 3–5 brani rappresentativi
4. **Interpretazione**  
   Spiegazioni semplici da inserire in report/slide
5. **Conclusioni**
   * Quali feature contano davvero
   * Pattern interessanti
   * Possibili bias o limiti

DATASET: <https://www.kaggle.com/datasets/priyamchoksi/spotify-dataset-114k-songs>