ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

KHAI PHÁ DỮ LIỆU (CO3093)

BÀI TẬP LỚN

PHÂN TÍCH & KHAI PHÁ TÍN HIỆU EEG PHỤC VỤ PHÂN LOẠI GIAI ĐOẠN GIẤC NGỦ

Giảng viên hướng dẫn: Đỗ Thanh Thái

Sinh viên thực hiện: Lương Ngọc Trung - 2313668
Nguyễn Quang Tùng - xxxxxxx
Nguyễn Quang Huy - xxxxxxx

Thành phố Hồ Chí Minh, Tháng 11 Năm 2025

Mục lục

1 Tổng quan về đề tài

1.1 Đặt vấn đề

Trong kỷ nguyên số, dữ liệu nghe nhạc trực tuyến được thu thập với quy mô lớn và đa dạng về dạng thức — từ thông tin cơ bản về bài hát (title, artist, album) đến các *audio features* định lượng (tempo, danceability, energy, valence, loudness, v.v.), cùng các chỉ số hành vi người dùng (lượt nghe, playlist, skip rate, listening session). Spotify — một trong những nền tảng streaming lớn nhất thế giới — cung cấp API mở (Spotify Web API) cho phép truy xuất rất nhiều thông tin hữu ích cho nghiên cứu dữ liệu và khai phá tri thức.

Việc nghiên cứu khai thác dữ liệu từ API Spotify mang lại nhiều giá trị thực tiễn và học thuật trong các lĩnh vực như marketing âm nhạc, dự báo xu hướng âm nhạc, hệ thống gợi ý (recommendation systems), phân khúc người dùng, phân tích hành vi tiêu dùng nội dung, và cả nghiên cứu văn hóa âm nhạc. Một số ứng dụng cụ thể bao gồm:

- Xây dựng hệ thống gợi ý cá nhân hoá (personalized recommendation) dựa trên *content-based* và *collaborative filtering*.
- Dự báo mức độ phổ biến (popularity) của bài hát/album/artists trong tương lai để tối ưu chiến dịch quảng bá.
- Phân khúc người nghe theo hành vi (nhóm "fan hâm mộ", nhóm nghe ngẫu nhiên, nhóm thích mood buồn/vui...) phục vụ marketing mục tiêu.
- Phân tích xu hướng âm nhạc theo thời gian, khu vực, thể loại — hỗ trợ nhà sản xuất/nhãn hiệu âm nhạc quyết định đầu tư.
- Tự động tạo playlist theo tâm trạng (mood), hoàn cảnh (workout, study), hoặc theo mục tiêu kinh doanh (tăng thời gian nghe, giảm skip rate).

So với bộ dữ liệu chuẩn dạng bảng như Adult Census Income, dữ liệu từ Spotify có tính đa nguồn (track metadata, audio features, artist metadata, user interactions), có chiều thời gian rõ rệt (lượt nghe theo thời gian), và thường kèm theo các dữ liệu phi cấu trúc (ví dụ mô tả bài hát, lời bài hát nếu có). Những đặc điểm này đòi hỏi quy trình tiền xử lý, feature engineering và thiết kế mô hình khác biệt, cũng như chú ý đến các vấn đề về bảo mật, quyền riêng tư và tuân thủ chính sách API.

1.2 Mục tiêu của đề tài

Đề tài hướng đến xây dựng một quy trình khai phá dữ liệu toàn diện và áp dụng các thuật toán học máy trên dữ liệu thu thập từ Spotify API, nhằm giải quyết một hoặc nhiều bài toán phân tích/học máy liên quan đến âm nhạc. Các mục tiêu cụ thể có thể bao gồm:

1. **Khảo sát và phân tích bộ dữ liệu:** mô tả chi tiết các nguồn dữ liệu có thể truy xuất từ Spotify API (track metadata, audio features, artist & album metadata, playlists, popularity, user listening events nếu có quyền truy cập), phân tích phân bố, kiểm tra dữ liệu thiếu, trùng lặp và nhiễu.
2. **Tiền xử lý và chuẩn hóa dữ liệu:** thiết kế quy trình thu thập (API rate limits, pagination), làm sạch dữ liệu, xử lý missing values, chuẩn hóa định dạng ngày giờ (release_date), mã hóa các thuộc tính phân loại (genre, key, mode), và scale các thuộc tính liên tục (tempo, loudness, duration_ms).

3. **Feature engineering:** tạo các đặc trưng mới có ý nghĩa (ví dụ: tempo buckets, normalized popularity per genre, recency features, user-level aggregates như avg_listen_duration), kết hợp metadata với audio features để gia tăng năng lực dự đoán.
4. **Xây dựng và đánh giá mô hình:** thử nghiệm nhiều kỹ thuật phù hợp cho các nhiệm vụ khác nhau: phân lớp (ví dụ phân loại bài hát có *hit* hay không), hồi quy (dự đoán popularity score), phân cụm (phân đoạn bài hát hoặc người nghe), và hệ thống gợi ý (collaborative, contentbased, hybrid). Đánh giá bằng các metric thích hợp (Accuracy, Precision, Recall, F1, AUC cho phân lớp; RMSE/MAE cho hồi quy; MAP/NDCG cho recommendation).
5. **So sánh, chọn lựa và đề xuất ứng dụng:** lựa chọn mô hình tối ưu dựa trên kết quả thực nghiệm và đưa ra đề xuất ứng dụng thực tế (playlist generation, targeted marketing, forecasting).
6. **Đề xuất hướng phát triển:** phân tích giới hạn nghiên cứu hiện tại và đề xuất các bước mở rộng (sử dụng dữ liệu user-level chi tiết, tích hợp lyrics/lyric features, áp dụng học sâu cho audio raw signal, hoặc triển khai mô hình real-time).

1.3 Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu

- Dữ liệu thu thập từ Spotify Web API: *track metadata* (id, name, artists, album, release_date, duration_ms, popularity), *audio features* (danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time_signature), *artist metadata* (genres, followers, popularity), và *playlist metadata* (tên playlist, tracks, followers).
- Các thuật toán khai phá và học máy phù hợp với dữ liệu dạng thời gian và bảng: phân lớp, hồi quy, phân cụm, và các kỹ thuật recommendation.

Phạm vi nghiên cứu

- Tập trung vào nghiên cứu phân tích dữ liệu và xây dựng mô hình trên dữ liệu public/thu thập được qua Spotify API (không thu thập dữ liệu cá nhân nhạy cảm nếu không được phép).
- Các bài toán nghiên cứu có thể bao gồm: dự đoán mức độ phổ biến (popularity) của track, phân loại track vào nhóm "hit / non-hit", phân cụm tracks theo đặc trưng âm thanh, phân khúc người nghe dựa trên hành vi (nếu có dữ liệu user-level), và thiết kế hệ thống gợi ý cơ bản.
- Hạn chế: không đi sâu vào xử lý tín hiệu audio thô (raw waveform) hoặc huấn luyện mô hình deep learning phức tạp trên audio trên quy mô lớn (trừ khi có tài nguyên tính toán và dữ liệu đủ lớn). Không triển khai hệ thống real-time production; nghiên cứu dừng ở mức proof-of-concept và thí nghiệm offline.
- Tuân thủ chính sách sử dụng API của Spotify (rate limits, terms of service), và các nguyên tắc đạo đức về dữ liệu (không lưu trữ hoặc phân tích dữ liệu cá nhân nhạy cảm mà không có consent).

1.4 Cấu trúc báo cáo (đề xuất cập nhật)

Báo cáo có thể được tổ chức lại để phù hợp với dataset mới như sau:

Chương 1 – Tổng quan về đề tài: lý do chọn dataset Spotify, tầm quan trọng của nghiên cứu, mục tiêu, phạm vi.

Chương 2 – Cơ sở lý thuyết: giới thiệu các khái niệm về audio features, recommendation systems, supervised learning, clustering, các metric đánh giá; tóm tắt cơ chế tính popularity của Spotify (nếu tìm được nguồn/căn cứ).

Chương 3 – Dữ liệu và phương pháp thu thập: mô tả chi tiết endpoints của Spotify API đã sử dụng, cách thiết lập OAuth (nếu cần), chiến lược sampling, lưu trữ dữ liệu thô.

Chương 4 – Tiền xử lý và feature engineering: trình bày pipeline tiền xử lý, xử lý missing, mã hóa, scaling, danh sách các feature mới tạo và lý do chọn.

Chương 5 – Thiết kế thực nghiệm và mô hình: mô tả các bài toán (classification/regression/recommendation), các thuật toán thử nghiệm, cách chia train/test (temporal split nếu cần), cross-validation, hyperparameter tuning.

Chương 6 – Kết quả và đánh giá: kết quả thực nghiệm, bảng so sánh metric, phân tích ý nghĩa các features quan trọng, trực quan hoá kết quả.

Chương 7 – Thảo luận, hạn chế và hướng phát triển: thảo luận về kết quả, các hạn chế của dữ liệu và mô hình, đề xuất nghiên cứu tiếp theo (ví dụ: tích hợp lyrics, raw audio modeling, real-time recommendation, A/B testing).

2 Cơ Sở Lý Thuyết

2.1 Tổng quan về Khai phá Dữ liệu (Data Mining)

2.1.1 Khái niệm và Tầm quan trọng

Khai phá dữ liệu (Data Mining - DM), còn được biết đến với tên gọi Phát hiện Tri thức trong Cơ sở dữ liệu (Knowledge Discovery in Databases - KDD), là một lĩnh vực liên ngành kết hợp giữa **Thống kê**, **Học máy (Machine Learning)**, **Trí tuệ nhân tạo (AI)** và **Quản trị cơ sở dữ liệu**. Về bản chất, Data Mining là quá trình tự động hoặc bán tự động nhằm **trích xuất** và **khám phá** các mẫu (patterns), mối quan hệ (relationships), quy tắc (rules) và tri thức (knowledge) có giá trị, tiềm ẩn, chưa được biết đến trước đó từ các tập dữ liệu lớn (Big Data). Các tri thức này phải có khả năng **diễn giải được**, **mang tính hành động** (actionable), và có thể áp dụng vào các mục tiêu kinh doanh hoặc nghiên cứu cụ thể.

Trong bối cảnh của dự án này, dữ liệu được thu thập từ Spotify API cung cấp một lượng lớn các **đặc trưng âm học (audio features)** và **siêu dữ liệu (metadata)** cho mỗi bài hát. Các đặc trưng này, như `danceability`, `energy`, `valence` (cảm xúc), `acousticness`, `tempo`, v.v., là một biểu diễn định lượng về bản chất âm nhạc. Bằng các kỹ thuật Data Mining, chúng ta có thể:

- **Phân cụm (Clustering)**: Nhóm các bài hát có đặc tính âm học tương tự lại với nhau mà không cần biết trước nhãn thể loại, từ đó phát hiện ra các "việt-thể loại"(micro-genres) hoặc tạo các playlist có tâm trạng, không khí đồng nhất.
- **Phân loại (Classification)**: Dự đoán thể loại hoặc mức độ phổ biến của một bài hát dựa trên các đặc trưng của nó.
- **Luật kết hợp (Association Rule Learning)**: Tìm ra các quy tắc kiểu "những người thích bài hát A cũng thường thích bài hát B", là nền tảng cho hệ thống gợi ý (Recommendation System).
- **Giảm chiều dữ liệu (Dimensionality Reduction)**: Trực quan hóa toàn bộ thư viện âm nhạc trong không gian 2D hoặc 3D để dễ dàng quan sát và phân tích.

2.1.2 Quy trình Phát hiện Tri thức trong Cơ sở dữ liệu (KDD Process)

Quy trình KDD là một quy trình chuẩn hóa, lặp đi lặp lại, bao gồm các bước sau để đảm bảo kết quả khai phá có ý nghĩa và đáng tin cậy:

1. **Lựa chọn Dữ liệu (Data Selection)**: Xác định mục tiêu phân tích và tập trung vào một tập hợp con dữ liệu liên quan từ các nguồn lớn hơn. Trong dự án này, bước này tương ứng với việc xác định các đặc trưng âm học và thể loại âm nhạc cần lấy từ Spotify API.
2. **Tiền xử lý Dữ liệu (Data Preprocessing)**: Đây là bước quan trọng và tốn nhiều công sức nhất, nhằm nâng cao chất lượng dữ liệu để các thuật toán hoạt động hiệu quả. Các tác vụ chính bao gồm:
    - **Làm sạch dữ liệu (Data Cleaning)**: Xử lý dữ liệu thiếu (missing values), nhiễu (noise), và không nhất quán (inconsistencies). Ví dụ: một số bài hát có thể thiếu giá trị cho đặc trưng `loudness`.
    - **Tích hợp dữ liệu (Data Integration)**: Kết hợp dữ liệu từ nhiều nguồn khác nhau (nếu có).

- **Chuyển đổi Dữ liệu (Data Transformation)**: Bao gồm **Chuẩn hóa (Normalization)** hoặc **Scaling** để đưa các đặc trưng có phạm vi giá trị khác nhau (ví dụ: `tempo` từ 50-200 và `acousticness` từ 0.0-1.0) về cùng một thang đo, thường là [0,1] hoặc phân phối chuẩn. Điều này rất quan trọng đối với các thuật toán dựa trên khoảng cách như K-Means. Các kỹ thuật phổ biến là Min-Max Scaling và Z-score Standardization.

3. **Biến đổi Dữ liệu (Data Transformation)**: Bước này tập trung vào việc giảm độ phức tạp của dữ liệu. Các kỹ thuật bao gồm:
   - **Giảm chiều dữ liệu (Dimensionality Reduction)**: Như Principal Component Analysis (PCA) hoặc t-SNE, giúp giảm số lượng đặc trưng trong khi vẫn cố gắng bảo toàn cấu trúc của dữ liệu, hỗ trợ trực quan hóa và tăng hiệu suất thuật toán.
   - **Rút trích đặc trưng (Feature Extraction)**: Tạo ra các đặc trưng mới, có ý nghĩa hơn từ các đặc trưng gốc.
4. **Khai phá Dữ liệu (Data Mining)**: Lõi của quy trình, nơi các thuật toán được áp dụng để trích xuất các mẫu từ dữ liệu. Tùy thuộc vào mục tiêu, các nhiệm vụ chính bao gồm:
   - **Học có giám sát (Supervised Learning)**: Phân loại (Classification), Hồi quy (Regression).
   - **Học không giám sát (Unsupervised Learning)**: Phân cụm (Clustering), Luật kết hợp (Association Rules).
   - **Phát hiện bất thường (Anomaly Detection)**.

   Trong dự án này, chúng ta sẽ tập trung vào **phân cụm** bằng thuật toán K-Means.
5. **Diễn giải và Đánh giá (Interpretation/Evaluation)**: Phân tích và đánh giá các mẫu đã được phát hiện. Các mẫu này cần được đánh giá về tính hữu ích, tính mới lạ, tính dễ hiểu và tính hợp lệ (thông qua các chỉ số đánh giá như Silhouette Score, v.v.). Tri thức thu được sau đó được trình bày cho các bên liên quan dưới dạng báo cáo, biểu đồ, hoặc tích hợp vào các hệ thống ra quyết định.

2.2 Thuật toán phân cụm K-Means

2.2.1 Khái niệm phân cụm

Phân cụm (Clustering) là nhiệm vụ học không giám sát nhằm phân chia một tập hợp các đối tượng dữ liệu thành các nhóm (gọi là các cụm - clusters) sao cho:

- Các đối tượng trong cùng một cụm có độ **tương đồng cao** (high intra-cluster similarity).
- Các đối tượng thuộc các cụm khác nhau có độ **tương đồng thấp** (low inter-cluster similarity).

Trong ngữ cảnh âm nhạc, phân cụm giúp chúng ta khám phá cấu trúc nội tại của dữ liệu mà không cần biết trước thể loại, từ đó phát hiện ra các nhóm bài hát có cùng tâm trạng, năng lượng, hoặc phong cách biểu diễn.

2.2.2 Thuật toán K-Means

K-Means là một trong những thuật toán phân cụm dựa trên tâm cụm (centroid-based) phổ biến nhất do đơn giản, hiệu quả và dễ hiện thực.

2.2.2.1 Ý tưởng Thuật toán hoạt động bằng cách xác định **K** điểm trung tâm (centroids), mỗi điểm đại diện cho một cụm, sau đó gán từng điểm dữ liệu vào cụm có centroid gần nó nhất. Quá trình này được lặp lại cho đến khi các centroid ổn định.

2.2.2.2 Các bước thực hiện

1. **Khởi tạo (Initialization)**: Chọn ngẫu nhiên K điểm dữ liệu từ tập dữ liệu làm centroid ban đầu. (Các phương pháp cải tiến như K-Means++ được sử dụng để chọn các centroid ban đầu xa nhau, giúp thuật toán hội tụ tốt hơn).
2. **Gán cụm (Assignment)**: Với mỗi điểm dữ liệu x_i trong tập dữ liệu, tính toán khoảng cách (thường là khoảng cách Euclid) đến tất cả K centroid. Gán x_i vào cụm có centroid gần nó nhất.
$$\text{Cụm}(x_i) = \arg\min_{j=1}^{K} ||x_i - \mu_j||^2$$
3. **Cập nhật centroid (Update)**: Với mỗi cụm C_j vừa được gán, tính toán lại centroid mới μ_j của cụm đó bằng cách lấy trung bình cộng của tất cả các điểm dữ liệu đã được gán vào cụm C_j.
$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$
4. **Lặp lại (Iteration)**: Lặp lại Bước 2 và Bước 3 cho đến khi thỏa mãn điều kiện dừng (ví dụ: các centroid không thay đổi đáng kể, hoặc đạt số vòng lặp tối đa).

2.2.2.3 Hàm mục tiêu K-Means tìm cách cực tiểu hóa hàm **Tổng bình phương khoảng cách trong cụm (Within-Cluster Sum of Squares - WCSS)**:

$$J = \sum_{j=1}^{K} \sum_{x_i \in C_j} ||x_i - \mu_j||^2$$

Trong đó:

- K: Số cụm.
- C_j: Tập các điểm thuộc cụm thứ j.
- μ_j: Centroid của cụm j.
- x_i: Một điểm dữ liệu.

2.2.3 Ưu điểm và Hạn chế của K-Means

2.2.3.1 Ưu điểm

- **Đơn giản và dễ hiện thực**: Thuật toán có ý tưởng rõ ràng và dễ dàng cài đặt.
- **Hiệu quả về thời gian tính toán**: Độ phức tạp tính toán là $O(n \times K \times I \times d)$, trong đó n là số điểm dữ liệu, K là số cụm, I là số vòng lặp, và d là số chiều dữ liệu. Điều này làm nó phù hợp với các tập dữ liệu lớn.

- **Hội tụ nhanh**: Trong thực tế, thuật toán thường hội tụ sau một số ít vòng lặp.
- **Dễ diễn giải**: Kết quả phân cụm có thể được giải thích thông qua các centroid, đại diện cho "điểm trung bình"của cụm.

2.2.3.2 Hạn chế

- **Cần xác định trước số cụm K**: Việc chọn K không phù hợp có thể dẫn đến kết quả phân cụm kém chất lượng. Phương pháp Elbow Method thường được dùng để ước lượng K tối ưu bằng cách quan sát điểm "khuỷu tay"trên đồ thị WCSS theo K.
- **Nhạy cảm với việc khởi tạo ban đầu**: Việc chọn ngẫu nhiên centroid ban đầu có thể dẫn đến kết quả cục bộ (local optimum) khác nhau. K-Means++ là một cải tiến để giảm thiểu vấn đề này.
- **Nhạy cảm với nhiễu và ngoại lai (outliers)**: Các điểm dữ liệu nằm xa các cụm chính có thể kéo centroid lệch khỏi vị trí thực sự của cụm.
- **Giả định cụm có dạng hình cầu (spherical)**: K-Means hoạt động tốt khi các cụm có dạng hình cầu và kích thước tương đương. Nó hoạt động kém hiệu quả với các cụm có hình dạng phức tạp, không lồi (non-convex).
- **Chỉ hoạt động với dữ liệu số**: Thuật toán dựa trên khoảng cách nên chỉ áp dụng được cho dữ liệu định lượng.

2.2.4 Ứng dụng K-Means trong phân tích dữ liệu Spotify

Trong dự án này, K-Means được áp dụng để phân cụm các bài hát dựa trên các đặc trưng âm học từ Spotify API. Quy trình cụ thể như sau:

- **Đặc trưng đầu vào**: Các đặc trưng như `danceability`, `energy`, `valence`, `acousticness`, `instrumentalness`, `liveness`, `speechiness`, `tempo` (sau khi đã chuẩn hóa).
- **Xác định số cụm**: Sử dụng Elbow Method hoặc Silhouette Analysis để xác định số cụm K tối ưu cho tập dữ liệu.
- **Phân cụm**: Áp dụng K-Means để gom nhóm các bài hát thành K cụm.
- **Diễn giải kết quả**: Phân tích các centroid của từng cụm để hiểu được đặc tính âm nhạc của từng nhóm. Ví dụ:
  - Một cụm có centroid với `energy` cao, `danceability` cao, `valence` cao có thể được gán nhãn là "Nhạc hạnh phúc, sôi động".
  - Một cụm khác có `acousticness` cao, `energy` thấp, `tempo` chậm có thể là "Nhạc acoustic, tĩnh lặng".
- **Ứng dụng**: Kết quả phân cụm có thể được sử dụng để tạo playlist tự động, đề xuất bài hát mới dựa trên sự tương đồng về cụm, hoặc phân tích xu hướng âm nhạc.

2.2.5 Các chỉ số đánh giá chất lượng phân cụm

Để đánh giá và so sánh chất lượng của các kết quả phân cụm khác nhau, một số chỉ số được sử dụng:

- **Silhouette Score**: Đo lường mức độ "chặt chẽ"và "tách biệt"của các cụm. Giá trị nằm trong khoảng [-1, 1]. Giá trị càng gần 1 cho thấy cụm càng chặt chẽ và tách biệt tốt.
- **Inertia (WCSS)**: Tổng bình phương khoảng cách của các điểm tới centroid của cụm chứa nó. Inertia càng nhỏ càng tốt, nhưng có xu hướng giảm khi K tăng.
- **Calinski-Harabasz Index**: Tỷ lệ giữa độ phân tán giữa các cụm và độ phân tán trong từng cụm. Giá trị càng cao càng tốt.

3 Tổng quan về dữ liệu

3.1 Giới thiệu nguồn dữ liệu

Trong khuôn khổ của đề tài, bộ dữ liệu được sử dụng được thu thập trực tiếp từ Spotify Web API, một trong những nguồn dữ liệu âm nhạc phong phú và chi tiết nhất hiện nay. Mục tiêu của việc thu thập này là xây dựng một bộ dữ liệu (dataset) đủ lớn và đa dạng, chứa các đặc trưng âm thanh (audio features) – yếu tố then chốt để thực hiện bài toán phân cụm (clustering) và xây dựng hệ thống gợi ý nhạc.
Bộ dữ liệu cốt lõi của đề tài được thu thập trực tiếp từ Spotify, một trong những nền tảng lưu trữ và phân tích âm nhạc số lớn nhất thế giới. Thay vì sử dụng một bộ dữ liệu có sẵn, chúng ta đã chủ động xây dựng bộ dữ liệu riêng để đảm bảo tính đa dạng và phù hợp với mục tiêu.

Quá trình này được thực hiện bằng cách sử dụng các công cụ lập trình để tự động truy vấn (query) hệ thống dữ liệu của Spotify. Chúng ta đã áp dụng chiến lược thu thập ngẫu nhiên, tìm kiếm các bài hát thuộc nhiều thể loại, nghệ sĩ, và mức độ phổ biến khác nhau. Mục tiêu là để tạo ra một "bức tranh"tổng thể về âm nhạc, thay vì chỉ tập trung vào một nhóm cụ thể. Kết quả là một bộ dữ liệu lớn, phong phú, sẵn sàng cho việc khám phá.

3.2 Cấu trúc bộ dữ liệu

Bộ dữ liệu cuối cùng (spotify.csv) là một bảng dữ liệu lớn, chứa thông tin của 114.000 bài hát với 21 cột thông tin. Để phục vụ cho các bài toán Data Mining, chúng ta có thể chia 21 cột này thành 3 nhóm chức năng chính như sau:

- **Thông tin định danh (Identifiers)** Đây là các cột dùng để nhận diện, mô tả bài hát. Chúng không được dùng trong việc huấn luyện mô hình (vì là dạng chữ hoặc ID) nhưng rất quan trọng để diễn giải, kiểm tra kết quả sau khi phân tích.

  1. `index`
  2. `track_id` — (Mã định danh duy nhất)
  3. `artists` — (Nghệ sĩ)
  4. `album_name` — (Tên album)
  5. `track_name` — (Tên bài hát)

- **Đặc trưng âm thanh (Audio Features)** Đây là "trái tim"của bộ dữ liệu cho đề tài phân cụm (clustering). Chúng là các cột dữ liệu số, mô tả định lượng đặc tính âm nhạc của bài hát và sẽ là đầu vào (features) chính cho mô hình K-Means.

  Danh sách biến (mô tả ngắn):

  1. `danceability` — Khả năng bắt tai của bài hát
  2. `energy` — Năng lượng bài nhạc mang lại
  3. `key` — Khóa/âm giai
  4. `loudness` — Âm lượng
  5. `mode` — integer số năm học tương ứng
  6. `speechiness` — Tính "nhiều từ"trong bài
  7. `acousticness` — Tính 'acoustic'/mộc

8. `instrumentalness` — Tính nhạc cụ trong bài
9. `liveness` — Tính trực tiếp
10. `valence` — Độ tích cực/vui vẻ
11. `tempo` — Nhịp độ - BPM
12. `duration_ms` — Thời lượng
13. `explicit` — Nội dung nhạy cảm, sẽ được chuyển thành 0/1

- **Thông tin bối cảnh & Tham khảo (Contextual)** Các cột này cung cấp thêm bối cảnh và rất hữu ích để làm biến mục tiêu cho các bài toán khác (như Phân loại, Hồi quy), hoặc dùng để đối chiếu và kiểm chứng kết quả sau khi phân cụm.

1. `popularity` — Độ phổ biến (0-100)
2. `time_signature` — Chỉ số nhịp
3. `track_genre` — Thể loại nhạc

3.3 Đặc điểm thống kê sơ bộ và thách thức

Để dữ liệu có thể sử dụng cho phân tích và huấn luyện mô hình, cần đảm bảo các tiêu chí sau:

- Thu thập được ổn định (không bị chặn, không mất nguồn bất ngờ).
- Dữ liệu đủ sạch, đủ nhiều.
- Có thể xử lý bằng công cụ mà nhóm làm chủ (Python, Microsoft SQL server . . .).
- Đủ chất lượng và tính liên quan để phân tích bên trong.

Nhận xét tính chất của dữ liệu thu thập:

a. Tính đầy đủ (Completeness)

Dữ liệu từ hai nguồn chính:

- **Spotify Audio Features:** hơn 113.000 mẫu bài hát (có nhiều mức phổ biến khác nhau) có thông tin chi tiết về đặc trưng âm nhạc (tempo, energy, danceability, valence, ...), tuy nhiên do thiếu hụt thuộc tính thời gian ra bài hát nên ta sẽ dùng **API Spotify** để cập nhật.
- **Billboard Dataset:** một tập dữ liệu chứa khoảng 2000 bài hát được lưu dưới dạng tệp .csv về những bài nhạc xu hướng trong giai đoạn từ đầu thế kỉ 21 đến năm 2020.
- Trường dữ liệu **Popularity** bằng 0 trong vài bản ghi nhận, khả năng là do điều kiện khách quan dẫn đến thuộc tính này có giá trị 0, nên những bài hát có giá trị bằng 0 ta cũng sẽ cân nhắc không xét tới.

Đánh giá: Đạt yêu cầu, thiếu sót không ảnh hưởng đến phân tích hoặc có giải pháp để bổ sung thông tin. Tuy nhiên vẫn cần loại ra những dữ liệu bất thường.

b. Tính nhất quán (Consistency)

Dữ liệu đã được chuẩn hoá:

- Các cột tên bài hát và nghệ sĩ được thống nhất về định dạng chữ thường.
- Thời gian (năm phát hành) được chuyển về kiểu datetime hoặc int để dễ nhóm theo giai đoạn.
- Các giá trị số (**danceability**, **liveness**, **instrustmentalness**, **energy**, ...) nằm trong cùng một thang đo (0–1).
- Các thuộc tính như tempo hay loudness vẫn tuân thủ đơn vị chuẩn.
- **Mức độ phổ biến (popularity)** trong phạm vi (0-100) - đây là một trong những trường dữ liệu quan trọng để ta đánh giá độ nổi tiếng, tỉ lệ thành công của một bài hát.

Đánh giá: dữ liệu tương đối nhất quán, có thể xử lý trực tiếp bằng Python (pandas, seaborn).

c. Tính liên kết (Linkability)

Hai nguồn sở hữu những cặp khóa có thể liên kết qua cặp, một khóa hợp lệ có thể là ***(artist, track)***.

Do cách ghi tên có thể hơi khác nhau giữa các nguồn (chữ hoa/thường, ký tự đặc biệt), cần xử lý chuẩn hóa trước để đồng bộ.

Đánh giá: khả năng liên kết cao, đủ để kết hợp phân tích xu hướng giữa Billboard và Spotify dataset đã thu thập.

d. Độ tin cậy (Reliability)

Spotify Audio Features và **BillBoard Dataset** đều được thu thập từ nền tảng **Kaggle**, là nguồn dữ liệu có độ tin cậy, uy tín cao và cập nhật thường xuyên dù bên cạnh đó sẽ có những sai sót nhỏ ở những dữ liệu cập nhật theo ngày.

Đánh giá: đáng tin cậy cho phân tích thống kê và xu hướng âm nhạc.

e. Tính đại diện và khối lượng (Representativeness & Size)

Tổng cộng gần **115000 mẫu**, đủ lớn để phân tích hành vi âm nhạc qua nhiều giai đoạn.

Dữ liệu bao phủ đa dạng thể loại (Pop, Hip-Hop, Rock, EDM, Latin, v.v) và nhiều ngôn ngữ.

Đánh giá: tập dữ liệu có tính đại diện tốt, phản ánh xu hướng toàn cầu.

f. Tính liên quan (Relevance)

Các thuộc tính như **energy, danceability, tempo, acousticness, popularity** phản ánh trực tiếp đặc trưng âm nhạc.

Các biến này rất hữu ích cho việc phân tích xu hướng thể loại hoặc dự đoán đặc điểm bài hát lọt bảng xếp hạng.

Đánh giá: dữ liệu rất phù hợp với mục tiêu nghiên cứu.

4 Phương pháp luận và tiền xử lý dữ liệu

4.1 Tiền xử lý dữ liệu

4.2 Phân tích dữ liệu

Những dữ liệu từ những dataset này là cơ sở để ta có thể dự đoán xu hướng âm nhạc qua từng giai đoạn từ quá khứ đến hiện tại. Từ đó có thể xây dựng thêm mô hình dự đoán ở tương lai về những yếu tố liên quan.

4.2.1 Phân tích tổng quan - Tập dữ liệu spotify_clean_with_date.csv

Code thực thi dựa trên ý tưởng khảo sát mức độ tương quan dữ liệu đặc trưng âm nhạc, thời lượng, số bài hát với **mức độ phổ biến (popularity)** qua từng giai đoạn **(từng giai đoạn là 5 năm)**:

Phân tích biểu đồ tổng quan:

Hình 1: *Ma trận tương quan của toàn bộ dữ liệu*

Nhìn chung, **độ phổ biến (popularity)** của các bài hát trên Spotify có mối tương quan dương nhẹ với các yếu tố như **mức độ dễ nhảy (danceability)** và **độ lớn âm thanh (loudness)**.

Điều này cho thấy người nghe hiện nay có xu hướng ưa chuộng các bài hát sôi động, tiết tấu rõ ràng và mang năng lượng tích cực, dù mức độ tương quan không quá mạnh (khoảng 0.08 - 0.09). Đây là xu hướng dễ hiểu, vì các nền tảng nghe nhạc hiện đại thường ưu tiên các ca khúc có tiết tấu bắt tai, dễ dùng trong video ngắn hoặc playlist phổ biến (ví dụ: workout, party, trending hits).

Ngược lại, **instrumentalness** có tương quan âm rõ rệt (-0.2) với độ phổ biến. Điều này phản ánh thực tế rằng các bài hát thuần nhạc (ít hoặc không có giọng hát) thường ít được nghe rộng rãi hơn, chủ yếu phục vụ mục đích thư giãn, học tập, hoặc nền nhạc. Trong khi đó, các ca khúc có lời, dễ hát theo, dễ viral trên mạng xã hội lại được chia sẻ nhiều hơn, giúp tăng độ phổ biến.

Tóm lại, người nghe toàn cầu có xu hướng ưu tiên nhạc có nhịp điệu rõ, sôi động và giàu

năng lượng, trong khi các thể loại thiên về instrumental hoặc quá nhẹ nhàng có phạm vi khán giả hẹp hơn. Xu hướng này phản ánh sự phát triển của văn hóa nghe nhạc trực tuyến, nơi yếu tố cảm xúc tức thời và tính giải trí ngắn hạn đang chiếm ưu thế.

Phân tích từng giai đoạn cụ thể:

Hình 2: *Ma trận tương quan giai đoạn 2000 - 2004*

Vào giai đoạn năm 2000-2004, ta có thể thấy rõ **độ phổ biến (popularity)** tỉ lệ thuận với **độ lớn âm thanh (loudness)** (0.18) và tỉ lệ nghịch khá rõ với **instrumentalness** (-0.28) bên cạnh đó dữ liệu cũng cho thấy **speechiness** (-0.09) cũng ảnh hưởng tiêu cực đến độ phổ biến. Qua mức tương quan này, có thể nhận thấy rằng xu hướng nghe nhạc toàn cầu trong giai đoạn này thiên về những ca khúc mạnh mẽ, có tiết tấu rõ và giàu năng lượng.

Đây là thời kỳ mà các thể loại pop, rock, R&B và hip-hop phát triển mạnh mẽ, được quảng bá rộng rãi qua đài phát thanh, MTV, và các bảng xếp hạng như Billboard. Những bài hát có giọng hát nổi bật, nhịp điệu rõ ràng, âm lượng cao thường tạo được ấn tượng mạnh và dễ lan

tỏa hơn.

Ngược lại, các bản nhạc thuần instrumental hay mang tính thư giãn, không lời chưa phổ biến rộng rãi do thói quen nghe nhạc lúc bấy giờ vẫn gắn với ca sĩ và giai điệu có lời. Xu hướng này phản ánh thời kỳ đầu của kỷ nguyên nhạc đại chúng hiện đại, khi người nghe chú trọng cảm xúc mạnh mẽ và yếu tố giải trí trong âm nhạc hơn là tính nghệ thuật thuần túy.

img/PhÕãn tÕắch tòệEng quan/pttq_6_bs.png

Hình 3: *Ma trận tương quan giai đoạn 2005 - 2009*

Vào giai đoạn 2005-2009, ta thấy độ phổ biến (popularity) vẫn tương quan dương với **độ lớn âm thanh** = +0.15, đồng thời tương quan âm với **speechiness** (-0.14) và **instrumentalness** (-0.25).
Điều này cho thấy xu hướng âm nhạc trong giai đoạn này tiếp tục nghiêng về những ca khúc có âm thanh mạnh mẽ, sản xuất chuyên nghiệp và có giai điệu bắt tai hơn là lời thoại hoặc nhạc không lời.

Thời kỳ này cũng là giai đoạn âm nhạc đại chúng chuyển mình mạnh mẽ, với sự phát triển

của nhạc số, iTunes, và nền tảng chia sẻ video như YouTube (ra mắt năm 2005). Các ca khúc nổi tiếng thường có độ nén âm cao hơn, tạo cảm giác "to và rõ"khi phát trên thiết bị di động hoặc trực tuyến.

Mặt khác, hệ số tương quan âm với **speechiness** cho thấy người nghe ít ưa chuộng các bài mang yếu tố "nói"nhiều như rap thuần hoặc bài hát có đoạn nói dài; thay vào đó, melody (giai điệu) và hook dễ nhớ vẫn là yếu tố then chốt quyết định độ phổ biến.

Còn với **instrumentalness**, giá trị âm tiếp tục khẳng định rằng các bản nhạc không lời vẫn chưa chiếm được vị trí cao trong xu hướng nghe nhạc đại chúng, khi thị trường vẫn ưu tiên những ca khúc có phần thể hiện khả năng hát của ca sĩ.

Hình 4: *Ma trận tương quan giai đoạn 2010 - 2014*

Vào giai đoạn 2010-2014, các hệ số tương quan giữa **độ phổ biến (popularity)** và các thuộc tính âm nhạc nhìn chung dao động quanh mức $\pm$**0.1**, cho thấy mối liên hệ giữa đặc trưng âm thanh và mức độ phổ biến đã trở nên yếu và đa dạng hơn so với những thập niên trước.

Cụ thể, **loudness** giảm xuống còn 0.08 , thể hiện xu hướng người nghe ưa chuộng các bản thu âm cân bằng, dễ nghe hơn là âm thanh quá mạnh hoặc nén cao. Trong khi đó, **speechiness** (-0.17) và **instrumentalness** (-0.21) vẫn duy trì tương quan âm rõ rệt, phản ánh việc những bài có yếu tố "nói"nhiều hoặc nhạc thuần instrumental vẫn chưa được đại chúng đón nhận mạnh mẽ. Ngoài ra, **energy** và **liveness** có giá trị tương quan âm nhẹ (-0.07), cho thấy những ca khúc quá sôi động hay mang tính biểu diễn trực tiếp phải là xu hướng chủ đạo trong giai đoạn này.

Đây là thời kỳ chuyển giao mạnh mẽ của âm nhạc toàn cầu, khi Spotify, YouTube và các nền tảng streaming bắt đầu định hình lại cách người nghe tiếp cận âm nhạc. Thay vì tập trung vào các bản hit sôi động như trước, khán giả bắt đầu tìm kiếm sự đa dạng và tính cá nhân hóa cao hơn - từ indie pop, alternative cho đến electronic nhẹ nhàng.

Xu hướng này cho thấy thị hiếu âm nhạc toàn cầu đang dịch chuyển dần từ tính đại chúng thuần túy sang sự cân bằng giữa cảm xúc, nội dung và trải nghiệm nghe linh hoạt, phù hợp với sự phát triển của công nghệ và lối sống hiện đại.

img/PhÕãn tÕách tòệEng quan/pttq_8_bs.png

Hình 5: *Ma trận tương quan giai đoạn 2015 - 2019*

Trong giai đoạn 2015-2019, ta nhận thấy **thời lượng bài hát (duration_ms)** có tương quan âm nhẹ với **độ phổ biến** (-0.11), cho thấy những ca khúc ngắn hơn lại có xu hướng được nghe nhiều hơn. Điều này phản ánh sự thay đổi rõ rệt trong thói quen nghe nhạc của người dùng thời đại streaming - ưu tiên các bài hát súc tích, dễ nghe, dễ chia sẻ trên nền tảng như Spotify, TikTok hay YouTube. Giai đoạn này số lượng bài hát bùng nổ nhất (tăng 100% so với 5 năm trước).

Bên cạnh đó, **danceability** tăng nhẹ lên +0.09, cho thấy âm nhạc có nhịp điệu rõ và dễ nhảy tiếp tục được ưa chuộng, đặc biệt trong thời kỳ bùng nổ của EDM, tropical house và pop điện tử. Các yếu tố như **speechiness** (-0.10) và **instrumentalness** (-0.22) vẫn duy trì mối tương quan âm, thể hiện việc các bài có lời vẫn chiếm ưu thế rõ rệt so với nhạc không lời trong dòng chảy nhạc đại chúng.

Các thuộc tính khác như **energy**, **loudness** hay **valence** nhìn chung giữ ổn định hoặc tăng nhẹ, phản ánh sự cân bằng giữa năng lượng và cảm xúc tích cực trong âm nhạc. Tổng thể, xu hướng âm nhạc toàn cầu giai đoạn này hướng đến tính giải trí cao, giai điệu dễ tiếp nhận, và thời lượng ngắn gọn, phù hợp với môi trường nghe liên tục, đa nhiệm của người dùng hiện đại.

Hình 6: *Ma trận tương quan giai đoạn 2020 - 2024*

Trong giai đoạn năm 2020-2024, âm nhạc bắt đầu có sự biến đổi theo chiều hướng hiện đại và "tinh gọn"hơn. Cụ thể, **thời lượng bài hát (duration_ms)** tiếp tục có tương quan âm nhẹ với độ phổ biến (-0.09), cho thấy người nghe vẫn ưa chuộng những ca khúc ngắn gọn, dễ tiếp cận và dễ viral - phù hợp với xu hướng nghe nhạc trên nền tảng video ngắn như TikTok hay Reels.

Đáng chú ý, **danceability** tăng trở lại lên +0.10 , phản ánh sự quay lại của các ca khúc có nhịp điệu bắt tai, dễ nhảy, đặc biệt trong các thể loại như pop dance, EDM, latin pop hoặc các bản phối mang năng lượng tích cực sau đại dịch. Điều này đặc biệt diễn ra xuyên suốt trên những nền tảng TikTok (loại hình video ngắn có nhạc và nhảy rất thu hút người xem).
Speechiness dần ổn định quanh mức -0.04, cho thấy người nghe đã chấp nhận sự pha trộn giữa hát và nói, đặc trưng của rap melodic, R&B hiện đại và pop lai hip-hop.

Trong khi đó, **instrumentalness** (-0.17) vẫn duy trì tương quan âm, phản ánh việc nhạc có lời vẫn chiếm ưu thế trong thị trường đại chúng, nhưng không còn chênh lệch mạnh như giai đoạn trước.

Các yếu tố khác như **energy**, **valence**, **loudness**, **liveness** đều có tương quan dương nhẹ (0.01-0.05), cho thấy thị hiếu nghe nhạc toàn cầu đã trở nên đa dạng và cân bằng hơn: người nghe không chỉ tìm đến nhạc sôi động hay ballad nhẹ, mà còn hướng tới âm nhạc mang lại cảm xúc tích cực, thoải mái và phù hợp với nhiều bối cảnh.

Giai đoạn này đánh dấu sự trưởng thành của thị trường âm nhạc số, khi khán giả toàn cầu không còn chạy theo xu hướng đơn lẻ, mà hướng đến sự cá nhân hóa, cảm xúc và tính ứng dụng cao của âm nhạc trong đời sống hàng ngày.

Cái nhìn tổng quan:

Nhìn sơ lược qua các giai đoạn này, ta có 1 nhận xét trước tiên là Số lượng bài hát qua từng giai đoạn ngày càng tăng dần, điều này có thể giải thích bởi 2 lí do chính: thứ nhất là điều kiện ngày càng tốt, con người dễ tiếp cận với âm nhạc toàn cầu thông qua các nền tảng trực tuyến, dẫn đến nhiều bài nhạc được ra đời hơn, nhiều nhạc sĩ, ca sĩ không cần qua các kênh chính thống như trong quá khứ. Thứ hai là do chất liệu âm nhạc ngày càng phong phú, sự phát triển tư duy âm nhạc của nhiều khu vực đã giúp cho sự ra đời của các bài hát trở nên bùng nổ.

Như trên biểu đồ có thể thấy từ 4022 bài hát (2000 - 2004), tăng lần lượt lên 6253 bài (2005 - 2009), 10139 bài (2010 - 2014), 19762 bài (2015 - 2019), 23666 bài (2020 - 2024) bài hát mỗi 5 năm.

Trong suốt hơn hai thập kỷ từ 2000 đến 2024, xu hướng âm nhạc toàn cầu cho thấy sự dịch chuyển rõ rệt từ tính đại chúng thuần túy sang sự đa dạng và cá nhân hóa sâu sắc.

Ở giai đoạn đầu (2000-2009), người nghe ưa chuộng các ca khúc có âm thanh mạnh mẽ, **độ ồn của nhạc (loudness)** cao và giàu năng lượng, điển hình cho thời kỳ nhạc pop, rock và R&B thống trị. Các bài hát có lời, dễ hát theo và giàu cảm xúc chiếm ưu thế tuyệt đối, trong khi nhạc instrumental hoặc rap thuần chưa được đón nhận rộng rãi.

Bước sang thập kỷ 2010, cùng với sự phát triển mạnh của streaming (Spotify, YouTube, Apple Music), thị hiếu nghe nhạc trở nên phân tán hơn. Các yếu tố như **loudness**, **energy** hay **valence** dần ổn định quanh mức trung bình, phản ánh việc người nghe có xu hướng tìm kiếm âm nhạc dễ chịu, cân bằng và phù hợp với ngữ cảnh thay vì chỉ chú trọng vào độ sôi động. Sự xuất hiện của playlist theo tâm trạng hoặc hoạt động cũng làm thay đổi cách tiếp cận âm nhạc - người dùng nghe theo cảm xúc hơn là thể loại.

Từ năm 2015 trở đi, thời lượng bài hát có xu hướng ngắn dần, **danceability** tăng trở lại, và **speechiness** giảm nhẹ, phù hợp với môi trường tiêu thụ âm nhạc nhanh, ngắn và liên tục. Giai đoạn này chứng kiến sự lên ngôi của EDM, pop điện tử, latin pop và hip-hop hiện đại, song song với sự lan tỏa mạnh mẽ của các nền tảng xã hội và video ngắn.

Đến giai đoạn 2020-2024, âm nhạc toàn cầu bước vào thời kỳ ổn định và tinh gọn, nơi người nghe hướng đến sự hài hòa giữa nhịp điệu, cảm xúc và khả năng "thích ứng với mọi khoảnh khắc". Các yếu tố âm thanh không còn quyết định mạnh đến độ phổ biến, mà thay vào đó, nội dung, cảm xúc và tính lan tỏa xã hội đóng vai trò quan trọng hơn.

Qua đó có thể thấy, xu hướng âm nhạc qua hơn 20 năm cho thấy sự chuyển dịch từ **đại chúng → cá nhân hóa → cân bằng cảm xúc**, phản ánh cách công nghệ, văn hóa và xã hội hiện đại định hình lại thói quen thưởng thức âm nhạc toàn cầu.

4.2.2 Phân tích liên quan những bài lọt top bảng xếp hạng - Tập dữ liệu songs_normalize.csv

Nếu như phần ở trên là phân tích tổng quan, thì phần này ta sẽ tập trung phân tích những bài hát lọt vào bảng xếp hạng âm nhạc theo thời gian, qua đó ta có thể nhận diện được sự khác nhau giữa những bài hát nói chung và những bài hát được đám đông hưởng ứng nói riêng.

Với tệp dữ liệu **songs_normalize.csv**, với khoảng 2000 bài hát lọt bảng xếp hạng âm nhạc Spotify qua từng năm (giai đoạn 2000-2020), ta sẽ xem xét sự thay đổi trong các đặc trưng âm thanh (audio features) cũng như sự phân bố thể loại và đặc điểm cấu trúc bài hát.

Mục tiêu của phần này là nhận diện các xu hướng biến đổi trong phong cách âm nhạc đại chúng theo thời gian – từ đó rút ra mối liên hệ giữa đặc trưng kỹ thuật của bài hát và sự phổ biến của chúng trên nền tảng nghe nhạc trực tuyến.

Trước tiên ta tiến hành nhóm các bài hát theo từng năm và thêm vào một cột chu kì:

img/PhÕãn tÕách xu hÔặòệ̣Ẹng/code.png

Hình 7: *Mã nguồn một số khởi tạo cần thiết*

Để nhận thấy sự biến đổi cảu các đặc trưng âm nhạc rõ rệt nhất, ta dùng biểu đồ đường để biểu diễn qua các năm:

```
img/PhÕãn tÕách xu hÔặòệ̣Ẹng/code_2.png
```

Hình 8: *Mã nguồn để vẽ biểu đồ đường cho các đặc trưng âm nhạc*

Hình 9: *Biểu đồ đường biểu diễn giá trị của các đặc trưng âm nhạc trong giai đoạn 2000 - 2020*

Nhận xét:

- **Danceability :** Xu hướng này cho thấy chỉ số dễ nhảy của các bài hát lọt top có sự tăng trưởng ổn định qua các năm, đặc biệt từ sau năm 2010. Điều này khẳng định rằng yếu tố nhịp điệu và khả năng kích thích chuyển động đang trở thành yếu tố cốt lõi để một bài hát trở nên phổ biến trong kỷ nguyên nhạc Pop/Hip-hop hiện đại.
- **Energy :** Chỉ số energy của Top Songs thường duy trì ở mức cao xuyên suốt hai thập kỷ. Mức năng lượng cao (thể hiện qua cường độ và tốc độ nhận thức) cho thấy thị trường âm nhạc ưa chuộng các bản phối mạnh mẽ, dồn dập, phù hợp với môi trường tiệc tùng, tập luyện, hoặc các ứng dụng giải trí cần cảm giác bùng nổ.
- **Acousticness :** Đặc trưng này có xu hướng giảm rõ rệt theo thời gian. Sự sụt giảm cho thấy các bài hát lọt top ngày càng ít sử dụng các yếu tố âm thanh thuần mộc, không được

xử lý bằng điện tử. Đây là minh chứng cho sự chuyển đổi của ngành công nghiệp sang các bản phối sử dụng công nghệ sản xuất âm thanh hiện đại và tổng hợp.

- **Speechiness :** Chỉ số này cho thấy sự tăng trưởng nhẹ và liên tục sau năm 2005. Xu hướng tăng này phản ánh sự trỗi dậy mạnh mẽ của thể loại Hip-hop và Rap trong bảng xếp hạng, nơi lời nói và nhịp điệu đóng vai trò chủ đạo thay vì giai điệu thuần túy.
- **Valence :** Đặc trưng cảm xúc này thường biến động nhưng ổn định hoặc có thể giảm nhẹ trong các năm gần đây. Nhận xét này cho thấy sự phổ biến của một bài hát không hoàn toàn phụ thuộc vào việc nó có vui vẻ hay hạnh phúc (high valence) hay không, mà các sắc thái cảm xúc phức tạp (ví dụ: nhạc buồn, nhạc thư giãn) vẫn có thể đạt được thành công lớn.

img/PhÕãn tÕách xu hÔặòệ̣ng/code_3.png

Hình 10: *Mã nguồn để vẽ biểu đồ đường cho thuộc tính nhịp độ* ***(tempo)***

Hình 11: *Biểu đồ đường thể hiện nhịp độ các bài hát xu hướng trong giai đoạn 2000 - 2020*

Nhận xét:

Quan sát biểu đồ cho thấy nhịp độ trung bình của các bài hát có xu hướng dao động mạnh nhưng nhìn chung **có xu hướng tăng** về cuối giai đoạn.

- Từ 2007–2011, xuất hiện các đỉnh nhịp độ rõ rệt (đặc biệt năm 2008 và 2011), trùng với thời kỳ bùng nổ của các bản Pop/Dance sôi động, chịu ảnh hưởng từ trào lưu EDM đầu thập niên 2010.
- Sau năm 2012, nhịp độ có dấu hiệu ổn định quanh 120 BPM, rồi tăng mạnh trở lại vào năm 2019, cho thấy sự quay lại của xu hướng nhạc nhanh và giàu năng lượng trong các bản hit hiện đại.

Nhìn tổng thể, sự biến thiên này phản ánh sự linh hoạt của thị hiếu nghe nhạc toàn cầu, khi tốc độ bài hát thay đổi theo chu kỳ trào lưu, nhưng **về lâu dài vẫn là tăng**: giai đoạn đầu

2000s ưu tiên Pop nhẹ, đến 2010s nổi bật với Dance/Electronic nhanh, và cuối kỳ quay lại các giai điệu sôi động hơn nhằm phục vụ nhu cầu giải trí ngắn trên nền tảng streaming.

Tiếp theo, ta sẽ tiến hành xem xét các thể loại nhạc hay góp mặt ở trong những danh sách bài hát hay, nổi tiếng nhất:

```
img/PhÕãn tÕắch xu hÔặòệ̣ng/code_4.png
```

Hình 12: *Mã nguồn để vẽ biểu đồ cột cho các thể loại nhạc theo chu kì 5 năm*

Hình 13: *Biểu đồ cột thể hiện các thể loại nhạc lên xu hướng nhiều nhất (trong giai đoạn 2000-2020)*

Từ biểu đồ cột, ta có thể thấy những thể loại **(genre)** hay lọt bảng xếp hạng qua các chu kì cũng không hề thay đổi, chỉ khác ở số lượng bài hát và cơ cấu từng chu kì:

- Sự Thống trị của Pop/Hip-hop: Trong hầu hết các giai đoạn, thể loại Pop và Hip-hop/Rap (hoặc Rap) chiếm thị phần lớn nhất trong các bài hát lọt top. Điều này xác nhận rằng dữ liệu đã được thu thập có tính đại diện cao cho xu hướng tiêu dùng âm nhạc đại chúng toàn cầu.
- Tăng trưởng của Dance/Electronic và Latin: Nhận thấy sự tăng trưởng đáng kể về số lượng bài hát thuộc thể loại Dance/Electronic và Latin (nếu có) trong các giai đoạn gần đây (ví dụ: 2015–2019). Sự tăng trưởng này liên kết trực tiếp với xu hướng tăng của danceability và energy trên Biểu đồ Đường, chỉ ra một sự dịch chuyển gu nghe nhạc hướng tới sự sôi động và đa dạng văn hóa hơn.

- Thị phần Giảm của Rock/Country: Các thể loại truyền thống như Rock hoặc Country (dựa vào kết quả thực tế của biểu đồ) có thể cho thấy sự duy trì hoặc suy giảm nhẹ thị phần trong Top 100/Top Chart. Điều này gợi ý rằng các thể loại này đang chuyển dịch khỏi vị trí thống trị trên các bảng xếp hạng thương mại.

Một yếu tố không thể bỏ qua là **thời lượng của bài hát (duration_ms)**:

```
img/PhÕãn tÕắch xu hÔặòệẸng/code_5.png
```

Hình 14: *Mã nguồn vẽ biểu đồ đường cho thời lượng các bài hát xu hướng*

Hình 15: *Biểu đồ đường thể hiện độ dài trung bình của các bài hát xu hướng trong giai đoạn 2000 - 2020 (đơn vị: phút)*

Quan sát biểu đồ cho thấy độ dài trung bình của các bài hát lọt top Spotify **giảm dần một cách rõ rệt** trong suốt hai thập kỷ qua.

- Cụ thể, giai đoạn 2000 – 2005, phần lớn các ca khúc phổ biến có độ dài trung bình khoảng 4.0 – 4.2 phút, đặc trưng cho cấu trúc truyền thống của nhạc Pop và Ballad, nơi bài hát thường có phần mở đầu, cao trào và kết thúc rõ ràng.
- Bước sang 2006 – 2010, thời lượng trung bình bắt đầu giảm xuống còn 3.8 – 3.9 phút, cho thấy xu hướng rút gọn dần cấu trúc bài hát để đáp ứng nhu cầu nghe nhanh và tiện lợi hơn.
- Đặc biệt từ 2015 trở đi, đồ thị thể hiện xu hướng giảm mạnh còn khoảng 3.4 – 3.6 phút, trùng khớp với giai đoạn bùng nổ của các nền tảng phát trực tuyến như Spotify, Apple Music hay TikTok.

Sự thay đổi này không chỉ là hiện tượng kỹ thuật mà còn mang ý nghĩa sâu sắc về sự thích ứng của ngành công nghiệp âm nhạc với hành vi người dùng hiện đại.

Trong thời đại kinh tế đi liền với phát trực tuyến, thời lượng ngắn giúp bài hát được phát lại nhiều lần hơn, tăng tổng lượt nghe và doanh thu bản quyền. Đồng thời, khoảng chú ý của người nghe trên nền tảng số ngày càng giảm, buộc các nghệ sĩ và nhà sản xuất phải tối ưu cấu trúc bài hát sao cho phần điệp khúc hoặc hook xuất hiện sớm, dễ gây nghiện và lan tỏa nhanh trên mạng xã hội.

Việc phân tích thời lượng bài hát qua thời gian giúp hiểu rõ cách mà thị trường âm nhạc toàn cầu đang tiến hóa về mặt cấu trúc sáng tác và tiêu thụ nội dung. Đây là chỉ báo quan trọng cho thấy ngành công nghiệp đang chuyển dịch mạnh sang mô hình ngắn – nhanh – lặp lại, phù hợp với xu hướng tiêu thụ giải trí hiện đại.

4.2.3 Ứng dụng của dữ liệu được phân tích

Việc phân tích dữ liệu âm nhạc Spotify giai đoạn 2000–2024 mang lại nhiều ứng dụng thực tiễn, đặc biệt trong lĩnh vực khoa học dữ liệu và kỹ thuật phần mềm.

Trước hết, các mô hình tương quan giữa các đặc trưng âm nhạc và độ phổ biến có thể được sử dụng làm nền tảng để xây dựng hệ thống gợi ý bài hát (music recommendation system) — giúp cá nhân hóa trải nghiệm người dùng dựa trên sở thích và xu hướng nghe nhạc theo thời gian.

Bên cạnh đó, ta có thể huấn luyện mô hình học máy (machine learning) nhằm dự đoán mức độ phổ biến của một bài hát mới dựa trên các thuộc tính như **danceability**, **energy** hay **instrumentalness**, từ đó hiểu rõ hơn các yếu tố ảnh hưởng đến thị hiếu công chúng.

Dữ liệu cũng có thể được ứng dụng trong phân tích cảm xúc (sentiment analysis), phân cụm thể loại (clustering by genre), hoặc khai thác cho các bài toán thị trường âm nhạc – như xác định giai đoạn bùng nổ của thể loại cụ thể (EDM, hip-hop, ballad).

Ngoài ra, kết quả phân tích còn giúp sinh viên rèn luyện kỹ năng xử lý dữ liệu thực tế, từ thu thập, làm sạch, trực quan hóa đến suy luận xu hướng, qua đó hình thành tư duy phân tích ứng dụng trong các dự án thực tế về dữ liệu số, giải trí hoặc AI.

Nhìn chung, bộ dữ liệu âm nhạc không chỉ phản ánh sự thay đổi trong văn hóa nghe nhạc toàn cầu, mà còn là nguồn học liệu quý giá để thực hành, nghiên cứu và phát triển ứng dụng thông minh trong thời đại dữ liệu lớn (Big Data).

4.3 Phân tích Dữ liệu Khám phá (EDA)

5 Kỹ thuật Phân cụm K-Means

5.1 Mục tiêu và thiết lập thực nghiệm

Mục tiêu:

Mục tiêu của thực nghiệm này là áp dụng thuật toán K-Means, một thuật toán học không giám sát, để phân chia 114,000 bài hát trong bộ dữ liệu thành các nhóm (cụm) riêng biệt. Mục tiêu cốt lõi là khám phá các cấu trúc tiềm ẩn trong dữ liệu.

Việc phân cụm này dựa trên 14 đặc trưng âm thanh và thuộc tính (bao gồm: `popularity`, `duration_ms`, `danceability`, `energy`, `key`, `loudness`, `mode`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valence`, `tempo`, và `explicit_int`) nhằm xác định các "bản sắc"hoặc "tâm trạng"âm nhạc (musical profiles) một cách tự động. Kết quả này không dựa trên thông tin thể loại (`track_genre`) cho trước, mà hoàn toàn dựa trên các đặc tính âm thanh của bài hát. Các cụm này sẽ là nền tảng để xây dựng các logic gợi ý nhạc (ví dụ: "Gợi ý các bài hát tương tự"hoặc "Xây dựng hồ sơ người dùng").

Thiết lập thực nghiệm:

- **Dữ liệu đầu vào (Input Data):** Dữ liệu thô là tệp `spotify.csv`. Sau Bước 1 (Lựa chọn và Tiền xử lý), chúng ta thu được một ma trận đặc trưng gồm 114,000 mẫu (bài hát) và 14 đặc trưng (cột).

  Chuẩn hóa dữ liệu (Standardization): Đây là bước bắt buộc. Thuật toán K-Means hoạt động dựa trên khoảng cách Euclidean (đo khoảng cách vật lý giữa các điểm dữ liệu). Các đặc trưng có thang đo lớn (như `duration_ms`, giá trị hàng trăm ngàn) sẽ hoàn toàn "lấn át"các đặc trưng thang đo nhỏ (như `danceability`, giá trị 0-1) trong phép tính khoảng cách này.

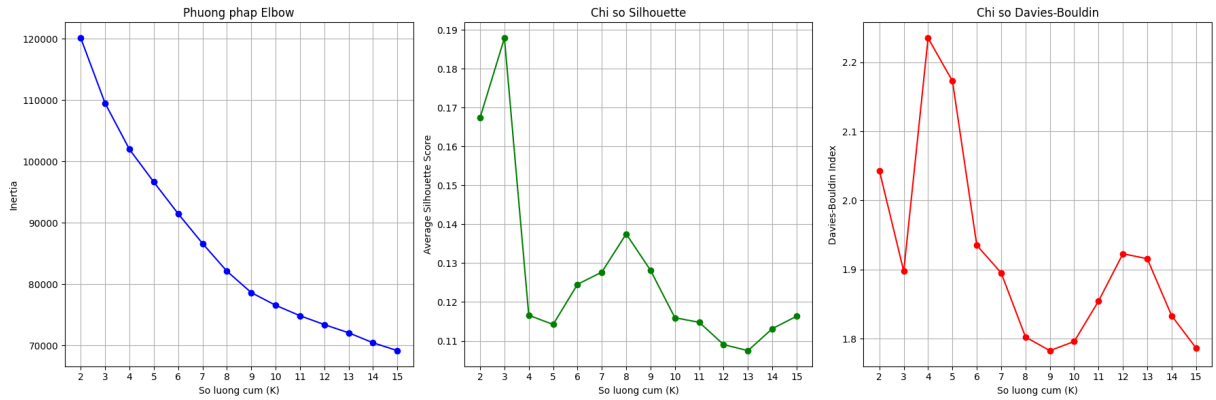
  Để giải quyết vấn đề này, chúng tôi đã sử dụng `StandardScaler` từ Scikit-learn. Kỹ thuật này biến đổi cả 14 đặc trưng để chúng có trung bình (mean) bằng 0 và độ lệch chuẩn (standard deviation) bằng 1. Điều này đảm bảo mọi đặc trưng đều có trọng số ảnh hưởng ngang nhau trong mô hình.

- **Lựa chọn Siêu tham số K (Số cụm):** Đây là quyết định quan trọng nhất trong mô hình K-Means. Để tìm giá trị `K` tối ưu một cách khách quan, chúng tôi đã chạy 3 phương pháp đánh giá khác nhau trên một mẫu ngẫu nhiên 10,000 bài hát. (Việc lấy mẫu là cần thiết vì tính toán chỉ số Silhouette có độ phức tạp $O(n^2)$, rất tốn kém trên toàn bộ dữ liệu).

  - **Phương pháp Elbow (Inertia):** Quán tính (Inertia) là tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm cụm (centroid) gần nhất của nó. Chúng ta tìm điểm "khuỷu tay"(elbow) - nơi mà việc thêm một cụm mới không còn giúp giảm quán tính một cách đáng kể.
  - **Chỉ số Silhouette (Silhouette Score):** Đo lường mức độ "đậm đặc"và "tách biệt"của các cụm. Chỉ số này nằm trong khoảng [-1, 1]. Giá trị gần 1 là tốt nhất (các điểm nằm gần các điểm khác trong cùng cụm, và xa các cụm khác). Chúng ta tìm `K` có điểm Silhouette cao nhất.
  - **Chỉ số Davies-Bouldin (Davies-Bouldin Index):** Đo lường tỷ lệ trung bình giữa độ "đậm đặc"trong cụm và sự tách biệt giữa các cụm. Giá trị gần 0 là tốt nhất. Chúng ta tìm `K` có điểm Davies-Bouldin thấp nhất.

- **Quyết định (Decision):** Dựa trên biểu đồ `k_selection_metrics.png` (Hình 5.1), kết quả phân tích rất rõ ràng:
  - Chỉ số Silhouette đạt đỉnh (peak) rõ rệt tại `K=9`, cho thấy 9 cụm là cấu hình "tách biệt"nhất.
  - Chỉ số Davies-Bouldin đạt đáy (valley) rõ rệt tại `K=10`, cho thấy 10 cụm có tỷ lệ trong-cụm/ngoài-cụm tốt nhất.
  - Phương pháp Elbow không cho thấy điểm "khuỷu tay"rõ ràng, nhưng bắt đầu thoải dần quanh mốc `K=9-10`.

  Việc các chỉ số gợi ý các giá trị `K` gần nhau (9 và 10) là rất phổ biến. Cả hai đều là lựa chọn tốt. Chúng tôi quyết định chọn `K=10` để huấn luyện mô hình cuối cùng. Lựa chọn này ưu tiên chỉ số Davies-Bouldin và đồng thời cung cấp một mức độ chi tiết (granularity) cao hơn một chút so với `K=9`, điều này được xem là có lợi cho việc xây dựng hệ thống gợi ý đa dạng.



Hình 16: *Biểu đồ 3 phương pháp đánh giá K*

- **Huấn luyện (Training):** Mô hình K-Means cuối cùng được huấn luyện trên toàn bộ 114,000 bài hát đã được chuẩn hóa. Chúng tôi sử dụng `KMeans` từ thư viện `scikit-learn` với các tham số sau:
  - `n_clusters=10`: Số cụm đã chọn.
  - `random_state=42`: Con số này đảm bảo rằng các vị trí khởi tạo tâm cụm ban đầu là giống nhau mỗi lần chạy. Điều này giúp kết quả của chúng ta có thể tái lập (reproducible).
  - `n_init=10`: (Đây là giá trị mặc định). Thuật toán sẽ chạy 10 lần với các điểm khởi tạo tâm cụm ngẫu nhiên khác nhau và tự động chọn kết quả tốt nhất (lần chạy có quán tính/inertia thấp nhất). Việc này giúp mô hình tránh bị "kẹt"ở một kết quả "tối ưu cục bộ"(local minimum) không tốt.

5.2 Kết quả phân cụm và trực quan hoá

Bảng Phân tích Đặc điểm Cụm (Centroids):

Sau khi huấn luyện mô hình (Bước 3) và gộp dữ liệu (Bước 4), chúng tôi thu được bảng đặc điểm trung tâm (centroid) của 10 cụm. Bảng này (Bảng 1) thể hiện "bản sắc"trung bình của

mỗi cụm, là cơ sở để chúng ta diễn giải ý nghĩa của chúng. Mỗi dòng đại diện cho một cụm, và mỗi cột là giá trị trung bình của 14 đặc trưng mà các bài hát trong cụm đó sở hữu.

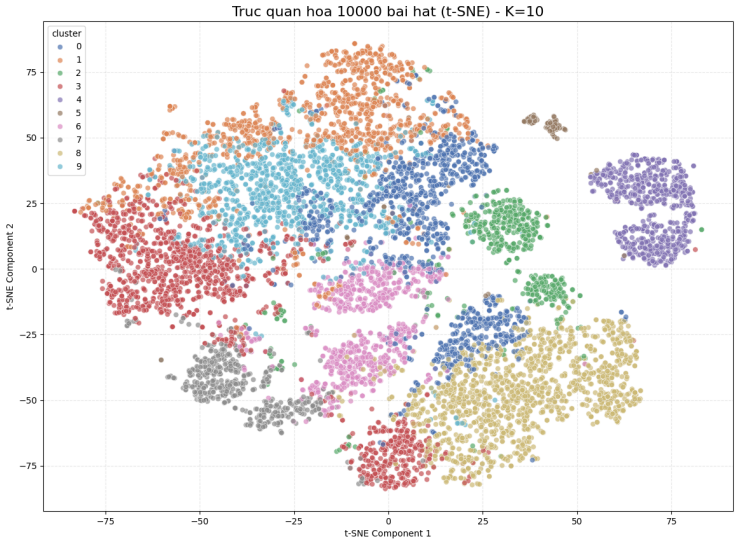
Bảng 1: *Phân tích đặc điểm trung tâm (centroid) của 10 cụm K-Means*

Cụm	Pop.	Dur. (ms)	Dance.	Energy	Key	Loud.	Mode	Speech.	Acoustic.	Instrum.	Live.	Valence	Tempo	Explicit
0	31.9	240k	0.41	0.84	5.2	-5.2	0.70	0.09	0.07	0.07	0.21	0.34	152.7	0.00
1	17.6	202k	0.67	0.73	5.1	-6.5	1.00	0.08	0.24	0.03	0.17	0.74	122.8	0.00
2	36.0	242k	0.51	0.75	5.3	-7.2	0.69	0.08	0.30	0.07	0.77	0.50	122.9	0.01
3	30.7	210k	0.51	0.33	5.0	-11.4	0.77	0.05	0.75	0.04	0.16	0.37	113.5	0.00
4	37.3	201k	0.64	0.72	5.4	-6.2	0.57	0.15	0.17	0.05	0.20	0.48	123.5	1.00
5	24.3	218k	0.57	0.68	5.1	-11.2	0.69	0.86	0.75	0.01	0.68	0.44	100.4	0.57
6	27.6	319k	0.60	0.73	5.6	-8.7	0.51	0.07	0.11	0.79	0.17	0.33	125.9	0.00
7	28.8	214k	0.35	0.17	4.8	-21.3	0.65	0.05	0.86	0.81	0.16	0.19	103.2	0.00
8	35.2	220k	0.66	0.73	6.2	-6.4	0.00	0.08	0.19	0.03	0.17	0.59	118.9	0.00
9	55.6	227k	0.61	0.68	4.9	-6.7	0.99	0.06	0.21	0.02	0.16	0.50	113.9	0.00

Trực quan hóa Cụm (t-SNE):

Do dữ liệu có 14 chiều (14 đặc trưng), chúng ta không thể trực quan hóa trực tiếp. Để "nhìn"được sự phân bố của 10 cụm, chúng tôi sử dụng kỹ thuật giảm chiều dữ liệu `t-SNE` (Bước 5) trên một mẫu 5,000 bài hát. `t-SNE` "nén"14 đặc trưng xuống còn 2 chiều (trục X và Y) để vẽ biểu đồ, cố gắng giữ nguyên cấu trúc lân cận của các điểm (các bài hát giống nhau ở 14 chiều sẽ có xu hướng nằm gần nhau ở 2 chiều).

Kết quả được thể hiện ở Hình 17. Mỗi điểm là một bài hát, được tô màu theo cụm (từ 0 đến 9) mà K-Means đã gán.



Hình 17: *Trực quan hóa 10 cụm K-Means sử dụng t-SNE trên 5,000 bài hát.*

5.3 Phân tích và đối chiếu

Phần này diễn giải ý nghĩa của 10 cụm dựa trên Bảng 1 và đối chiếu với Hình 17.

- **Cụm 7 (Nhạc Cổ điển/Ambient):** Nổi bật rõ ràng với `instrumentalness` (0.813) và `acousticness` (0.864) cao nhất. Đồng thời, đây là cụm "yên tĩnh"nhất với `energy` (0.170) và `loudness` (-21.3) cực kỳ thấp.
- **Cụm 4 (Rap/Hip-Hop):** Dễ dàng nhận diện là cụm Rap/Hip-Hop khi có `explicit_int` (1.000) – 100% bài hát là explicit – và chỉ số `speechiness` (0.148) cao.
- **Cụm 5 (Podcast/Nội dung nói):** Đây là cụm "không phải nhạc"duy nhất, với `speechiness` (0.860) cực cao, cho thấy nội dung chủ yếu là lời nói.
- **Cụm 3 (Acoustic/Ballad):** Tương tự Cụm 7 nhưng "có nhạc"hơn. Có `acousticness` (0.753) cao, nhưng `energy` (0.330) và `loudness` (-11.4) thấp, cho thấy đây là nhạc thư giãn, nhẹ nhàng.
- **Cụm 9 (Pop Thịnh hành):** Đây là cụm "mainstream"nhất, có `popularity` (55.57) trung bình cao nhất. Các chỉ số khác (`danceability`, `energy`, `valence`) đều ở mức trung bình "an toàn", dễ nghe.
- **Cụm 6 (Nhạc điện tử/Không lời dài):** Có `instrumentalness` (0.792) cao (nhưng không acoustic như Cụm 7) và `duration_ms` (319k) dài nhất. Đây có thể là nhạc Techno, Trance, hoặc Progressive Rock.
- **Cụm 2 (Nhạc Live/Trực tiếp):** Đặc trưng bởi chỉ số `liveness` (0.770) cao vượt trội, cho thấy đây là các bản thu âm trực tiếp (live recordings).

Đối chiếu với Trực quan hóa t-SNE (Hình 17):
Biểu đồ `t-SNE` đã củng cố mạnh mẽ kết quả phân tích.

- **Sự tách biệt rõ ràng:** Các cụm có "bản sắc"rất độc đáo như Cụm 7 (Cổ điển/Ambient) và Cụm 5 (Podcast) tạo thành các "đảo"riêng biệt, nằm tách xa khỏi phần còn lại, chứng tỏ K-Means đã xác định chúng rất chính xác.
- **Sự chồng chéo có ý nghĩa:** Phần trung tâm của biểu đồ là nơi tập trung các cụm "Pop"và "Dance"(như Cụm 1, 8, 9). Việc chúng nằm gần nhau và chồng chéo là hợp lý, vì chúng chia sẻ nhiều đặc tính âm thanh (như `danceability`, `energy`). Sự khác biệt của chúng có thể nằm ở các yếu tố nhỏ hơn như `mode` (trưởng/thứ) hoặc `popularity`, vốn là những khác biệt mà `t-SNE` có thể không ưu tiên khi chiếu xuống 2D.

Kết luận: Sự kết hợp giữa phân tích centroid (Bảng 1) và trực quan hóa t-SNE (Hình 17) cho thấy mô hình K-Means (K=10) đã thành công trong việc tạo ra 10 nhóm âm nhạc có ý nghĩa và tách biệt về mặt thống kê.

5.4 Ràng buộc, Hạn chế và Hướng phát triển

5.4.1 Ràng buộc và Hạn chế

Dù mô hình K-Means đã đạt được những kết quả khả quan, nghiên cứu này vẫn tồn tại một số hạn chế cần được thừa nhận:

- **Chất lượng dữ liệu từ Spotify API**: Các đặc trưng âm học được tính toán tự động bởi Spotify và có thể chứa nhiễu hoặc sai số nhất định. Độ chính xác của các chỉ số như `instrumentalness` hay `speechiness` có thể thay đổi giữa các bài hát.
- **Tính chủ quan trong diễn giải cụm**: Việc gán nhãn ý nghĩa cho các cụm (ví dụ: "Nhạc Pop thịnh hành", "Rap/Hip-Hop") dựa trên phân tích centroid mang tính chủ quan của nhóm nghiên cứu. Các diễn giải khác nhau có thể xuất hiện tùy thuộc vào góc nhìn của người phân tích.
- **Hạn chế của K-Means**:
  - Thuật toán giả định các cụm có dạng hình cầu (spherical) và kích thước tương đồng, điều này có thể không đúng với bản chất thực tế của dữ liệu âm nhạc vốn đa dạng và phức tạp.
  - K-Means nhạy cảm với việc khởi tạo centroid ban đầu, dù đã sử dụng `n_init=10` để giảm thiểu vấn đề này.
  - Việc lựa chọn K=10 dựa trên các chỉ số nhưng vẫn mang tính chủ quan. Các giá trị K khác (9, 11) có thể cho những cách phân chia có ý nghĩa khác.
- **Vấn đề cân bằng cụm**: Một số cụm có thể chứa rất ít bài hát trong khi các cụm khác lại rất lớn. Sự mất cân bằng này có thể ảnh hưởng đến tính hữu ích trong ứng dụng thực tế.
- **Hạn chế về tài nguyên tính toán**: Việc tính toán Silhouette Score cho toàn bộ dataset 114,000 điểm dữ liệu là bất khả thi do độ phức tạp $O(n^2)$, buộc phải sử dụng mẫu con 10,000 điểm để đánh giá.

5.4.2 Hướng phát triển trong tương lai

Dựa trên những hạn chế đã nêu, một số hướng phát triển có thể được đề xuất:

- **Thử nghiệm các thuật toán phân cụm khác**:
  - **DBSCAN**: Có thể phát hiện các cụm với hình dạng bất kỳ và tự động xác định số cụm, đồng thời xử lý tốt nhiễu và ngoại lai.
  - **Gaussian Mixture Models (GMM)**: Phù hợp hơn khi các cụm có phân phối hình elip và cho xác suất thuộc về cụm thay vì gán nhãn cứng.
  - **Agglomerative Hierarchical Clustering**: Cho phép khám phá cấu trúc phân cấp của dữ liệu âm nhạc.
- **Tích hợp thêm dữ liệu**:
  - Bổ sung lời bài hát (lyrics) để phân tích cảm xúc và chủ đề.
  - Thêm thông tin về ngữ cảnh người nghe (thời gian trong ngày, địa điểm, hoạt động).
  - Sử dụng dữ liệu lịch sử nghe của người dùng để cá nhân hóa phân cụm.

- **Phát triển hệ thống gợi ý lai (Hybrid Recommendation)**: Kết hợp kết quả phân cụm (content-based) với phương pháp collaborative filtering để tăng độ chính xác và đa dạng của gợi ý.
- **Xây dựng giao diện tương tác**: Phát triển dashboard cho phép người dùng khám phá các cụm âm nhạc, điều chỉnh tham số, và tùy chỉnh playlist dựa trên kết quả phân cụm.
- **Đánh giá chất lượng với chuyên gia âm nhạc**: Mời các chuyên gia hoặc nhạc sĩ đánh giá và gán nhãn cho các cụm để có diễn giải chính xác hơn.
- **Ứng dụng trong phân tích thị trường**: Sử dụng kết quả phân cụm để phân tích xu hướng âm nhạc theo thời gian, khu vực địa lý, hoặc nhân khẩu học người nghe.

5.4.3 Kết luận cho phương pháp K-Means

Mô hình K-Means với K=10 đã chứng minh hiệu quả trong việc khám phá cấu trúc tiềm ẩn của dữ liệu âm nhạc từ Spotify. Các cụm được phát hiện có ý nghĩa rõ ràng và tách biệt tốt, tạo nền tảng vững chắc cho việc phát triển hệ thống gợi ý nhạc dựa trên nội dung. Dù tồn tại một số hạn chế, phương pháp này cung cấp cái nhìn sâu sắc giá trị về cách các đặc trưng âm học có thể được sử dụng để phân loại âm nhạc một cách khách quan và tự động.

Những kết quả này không chỉ có giá trị học thuật mà còn có tiềm năng ứng dụng thực tế cao trong ngành công nghiệp âm nhạc và phát triển ứng dụng nghe nhạc.

6 Tài liệu tham khảo

Tài liệu

[1] Mô tả bài tập lớn.

[2] Kurose, James Ross, Keith. *Computer Networking: A Top-Down Approach*, 8th Edition.

[3] Lee, H., Li, B., Huang, Y., Chi, Y. & Lin, S. (2021). *"NCH Sleep DataBank: A Large Collection of Real-world Pediatric Sleep Studies with Longitudinal Clinical Data"* (version 3.1.0). PhysioNet. https://doi.org/10.13026/p2rp-sg37