**Hệ thống gợi ý bài hát dựa trên thói quen của người dùng bằng phương pháp K-means**

Le Dang Hung, Le Hong Quang, Tran Tuan Anh

**Tóm tắt:**

Hiện nay có rất nhiều ứng dụng nghe nhạc ra đời như Spotify, Apple Music,… do đó người dùng có nhiều lựa chọn hơn trong việc chọn ra cho mình danh sách những bài hát yêu thích từ những trang web đó. Bên cạnh đó, vẫn có nhiều người dùng có thói quen tải những bài nhạc về để nghe ngoại tuyến, việc này đã trở thành thói quen và đem lại sự thuận tiện cho họ. Tuy nhiên, hiếm có người dùng nào có thể tạo ra một danh sách gợi ý các bài hát cho những lần tải về tiếp theo của mình. Chính vì thế, chúng tôi mong muốn giới thiệu một hệ thống gợi ý âm nhạc dựa trên thuật toán K-means clustering. Hệ thống của chúng tôi sử dụng dữ liệu từ thói quen nghe nhạc của người dùng rồi chia họ thành từng nhóm người dùng khác nhau và gợi ý cho những nhóm đó về những bài hát có chủ đề, thể loại, nghệ sĩ, xuất xứ,…mà chúng có điểm chung. Mong muốn của chúng tôi với mô hình của mình đó chính là phù hợp với những người không quá quan tâm đến vấn đề kỹ thuật của mô hình mà chỉ cần kết quả của quy trình máy học. Chúng tôi mong rằng mô hình của mình sẽ là tiền đề giúp cho người nghe nhạc cảm thấy càng yêu thích sở thích của mình hơn.

**Từ khóa:** Gợi ý bài hát, Nghệ sĩ, Xuất xứ, Hành vi người dùng

1. **Giới thiệu**

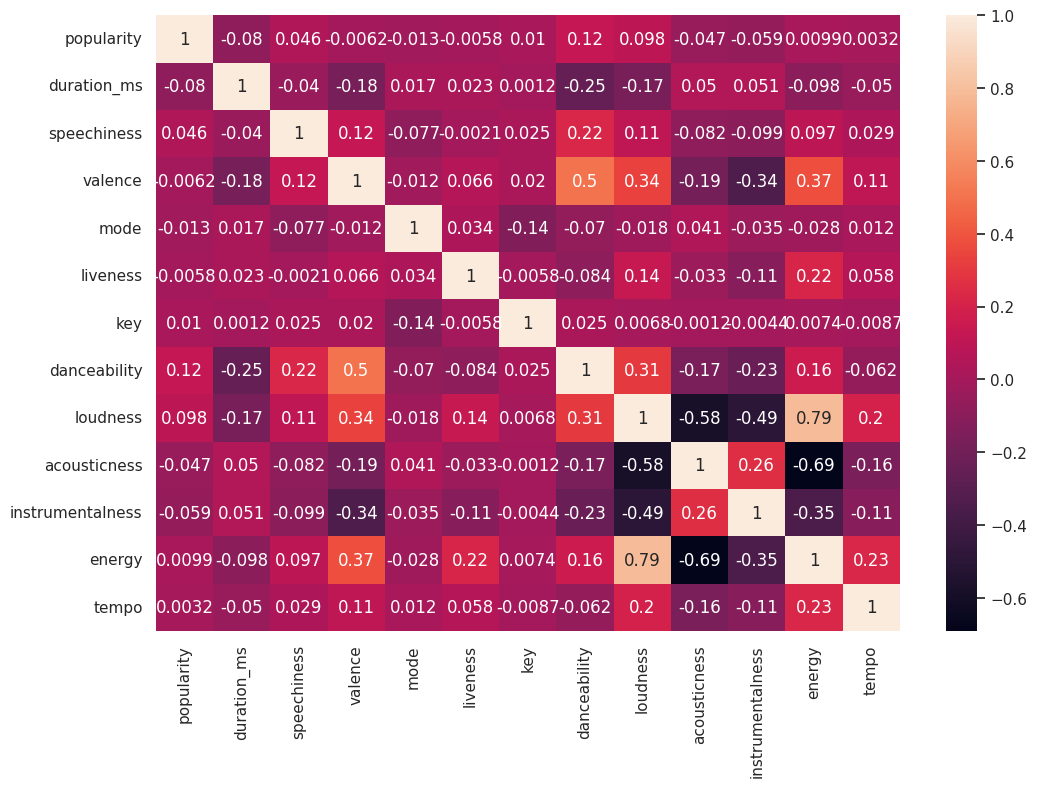
Một số lượng lớn người dùng có thể dễ dàng tải xuống các bài hát từ các trang web. Nhưng ít có người dùng nào có thể tạo ra riêng cho mình một hệ thống gợi ý các bài hát mà họ có thể yêu thích một cách cá nhân hóa. Nhìn chung, có hai cách để tiếp cận đối với một hệ thống gợi ý cơ bản đó là hệ thống gợi ý dựa trên nội dung (content-based) và hệ thống khuyến nghị hợp tác (collaborative). Hệ thống gợi ý dựa trên nội dung phân tích nội dung của các đối tượng mà người dùng ưa thích trước đây và đề xuất những nội dung có liên quan. Còn đối với khuyến nghị hợp tác, hệ thống đề xuất các đối tượng mà nhóm người dùng có sở thích tương tự cũng thích. Với hệ thống đề xuất dựa trên nội dung cần nhiều thời gian để ổn định và có thể đề xuất các bài hát không phù hợp cho người dùng vì nó không xem xét đến sở thích của từng người dùng. Bên cạnh đó, mô hình khuyến nghị hợp tác lại không thể đề xuất các bài hát có ít người từng nghe qua nó [2]. Vì lý do đó, chúng tôi sử dụng một cách tiếp cận quen thuộc hơn chính là sử dụng K-means clustering.

Có ba thách thức dành cho hệ thống gợi ý bài hát được cá nhân hóa: Thứ nhất là làm thế nào để phân tích thuộc tính, thứ hai là làm thế nào để phân tích sở thích của người dùng và thứ ba là làm thế nào để chọn những bản nhạc gần với sở thích của người dùng. Vì vậy, chúng tôi đề xuất một giải pháp cho ba câu hỏi này. Giải pháp đầu tiên mà chúng tôi sử dụng chính là phân tích sóng âm thanh bằng cách đưa chúng về dưới các giá trị số, chúng tôi có thể trích xuất thuộc tính âm nhạc của từng bài hát khác nhau. Giải pháp tiếp theo là sử dụng danh sách nhạc mà người dùng đã tải xuống hoặc nghe và từ đó suy ra tùy chọn của người dùng từ danh sách đã khai thác được. Để có thể giải quyết bài toán này, chúng tôi đề xuất thuật toán K-means clustering để phân tích thói quen của người dùng. Số lượng cụm (K) sẽ cần phải được cập nhật liên tục để mô hình có thể hoạt động chính xác nhất. Nếu K được lựa chọn đúng, độ chính xác của mô hình sẽ cao hơn, người dùng có thể nhận được các gợi ý chính xác hơn dựa trên những gì họ đã nghe trước đó. Số lượng cụm (K) là vô cùng quan trọng, Phạm vi của một cụm cần đủ nhỏ để chứa các bài hát tương đương nhau vì nếu phạm vi cụm quá lớn, những bài hát không quá giống với cụm cũng sẽ được gom vào. Trung tâm của các cụm K cũng chính là sở thích của người dùng đã được thu thập thông qua danh sách. Để có được K tốt nhất, giá trị K trong thuật toán K-Means được cập nhật bằng cách sử dụng phương pháp Elbow. Đoạn mã chạy qua một loạt các giá trị K khác nhau (từ 2 đến 20) và tính toán Sum of Squared Distances (SSD) cho mỗi giá trị K. SSD được tính bằng cách cộng tổng bình phương khoảng cách từ mỗi điểm đến trung tâm của cụm tương ứng. Sau đó, đồ thị Elbow được vẽ để hiển thị biểu đồ giữa số lượng cụm (K) và SSD. Phương pháp Elbow giả sử rằng có một "khuỷu tay" trên đồ thị và giá trị K tại điểm khuỷu tay là giá trị tối ưu cho số lượng cụm. Trong trường hợp này, giá trị K là điểm nơi đường cong trở nên dẹt khi K tăng lên, như một "khuỷu tay" [1]. Sau khi xác định được giá trị tối ưu của K bằng phương pháp Elbow, thuật toán K-Means chạy lại với giá trị K này để phân chia dữ liệu thành các cụm tương ứng.

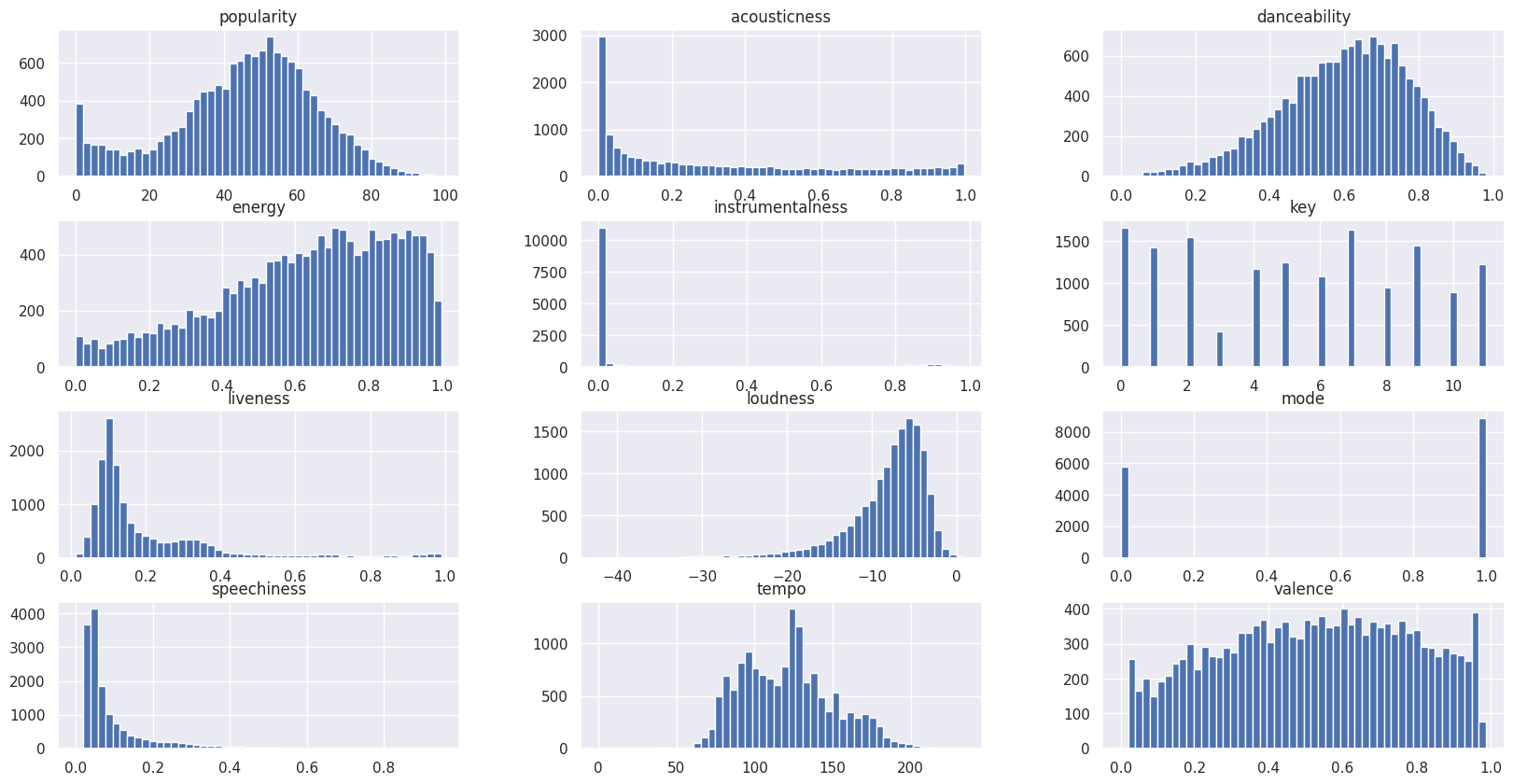
1. **Chuẩn bị dữ liệu**

Trong bài viết này, chúng tôi sử dụng bộ dữ liệu từ Spotify API truy cập thông tin về bài hát, playlist, và người nghe từ Spotify để tạo dự đoán dựa trên lịch sử và thói quen nghe nhạc của người dùng [3]. Tập dữ liệu đầu vào dùng để phân cụm bao gồm 14675 hàng và 19 cột cùng với tập dữ liệu lịch sử nghe nhạc của người dùng có 85 bài hát.

Đầu tiên chúng ta sẽ đến với bước tiền xử lý dữ liệu. Chúng tôi chỉ chọn những cột có thuộc tính ảnh hưởng đến quá trình phân cụm để đưa vào tập dữ liệu huấn luyện và loại bỏ đi những hàng bị trống dữ liệu. Sau đó, chúng tôi vẽ biểu đồ thể hiện mối tương quan giữa các thuộc tính (Hình\_1) và phân tích sự phân bố của các thuộc tính đó qua biểu đồ histogram (Hình 2). Các thuộc tính ‘key’, ‘mode’ và ‘popularity’ sẽ không hữu ích cho phần phân cụm, vì vậy chúng tôi sẽ loại bỏ đi những cột có chứa thuộc tính này. Như vậy 9 thuộc tính còn lại sẽ được chọn và đưa vào danh sách để tiến hành phân cụm.

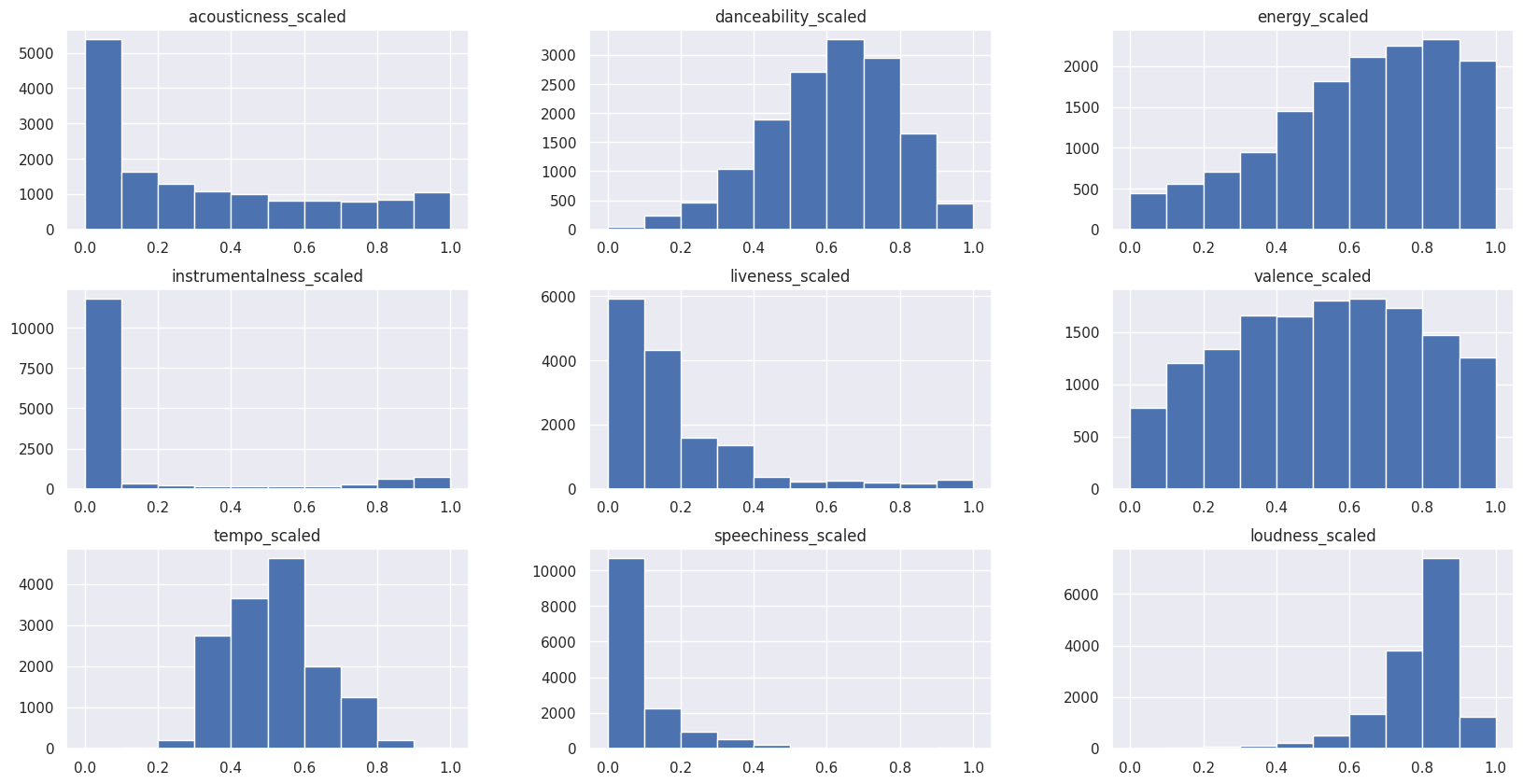


Hình 1: Biểu đồ mối tương quan



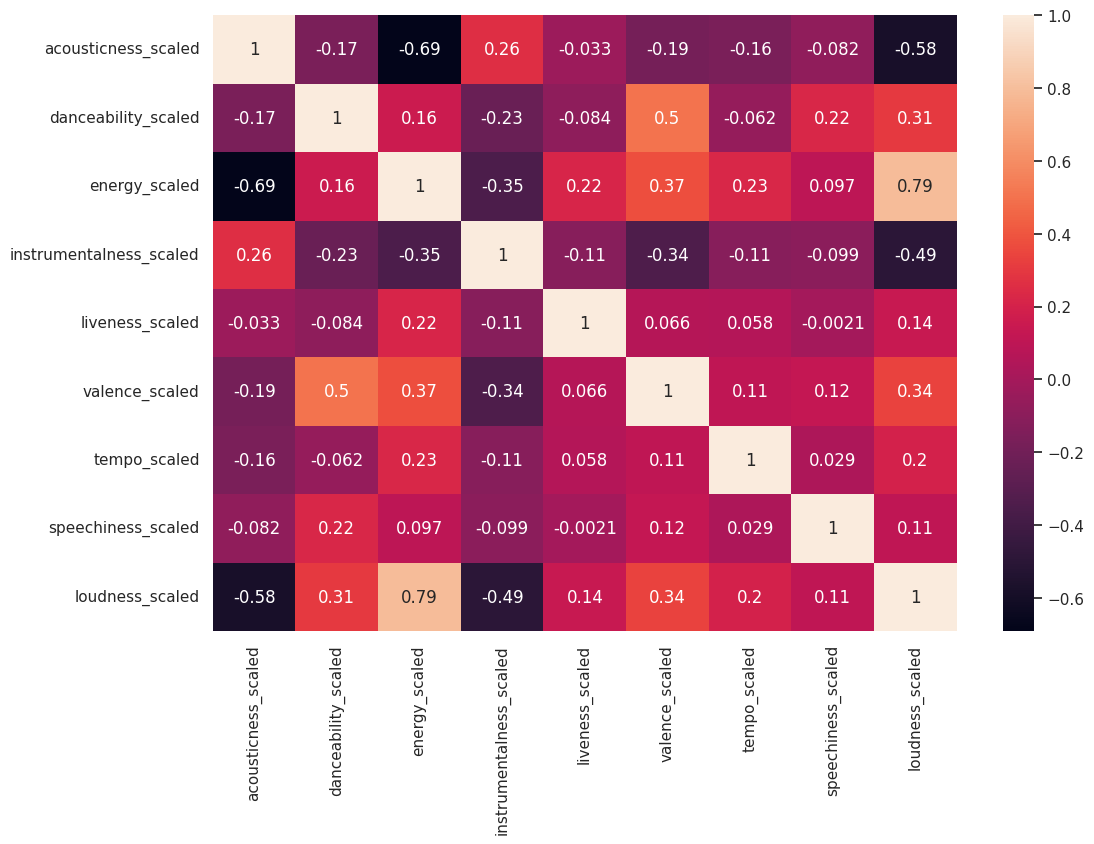
Hình 2: Biểu đồ phân bố tập dữ liệu

Tiếp theo, chúng tôi sẽ thay đổi tỉ lệ của tập dữ liệu để có thể thực hiện việc phân cụm ở phía sau. Về thang đo của các thuộc tính, để sử dụng được thuật toán K-means, tất cả các thuộc tính phải có cùng giá trị trung bình và phương sai. Chúng tôi sẽ chuyển tất cả chúng sang phạm vi từ 0,0 đến 1,0, có nghĩa là chuẩn hóa chúng. Có 2 phương pháp có thể sử dụng ở đây để có thể chuẩn hóa dữ liệu là StandardScaler và MinMaxScaler. Chúng tôi hiện đang sử dụng MinMaxScaler do giá trị inertia nhỏ trong K-means. Sau khi thay đổi tỉ lệ có thể thấy được sự phân bố của các thuộc tính ổn hơn (Hình 3).



Hình 3: Biểu đồ chuẩn hóa dữ liệu

Để có một kết quả tốt hơn, chúng ta nên tìm các thuộc tính có mối tương quan cao và nếu có, chúng ta nên loại bỏ chúng. Dựa vào biểu đồ thể hiện các giá trị tương quan giữa các thuộc tính (Hình 4), chúng ta thấy không có mối tương quan nào là không đáng chú ý. Vì vậy, chúng tôi sẽ giữ lại tất cả và bây giờ dữ liệu đã sẵn sàng để thực hiện phân cụm.



Hình 4: Biểu đồ mối tương quan dữ liệu được chuẩn hóa

1. **Phương pháp**

Tóm tắt phương pháp  
Sau khi dữ liệu đầu vào đã được xử lý, chúng tôi sẽ tiến hành huấn luyện mô hình K-Means để phân loại bài hát trong thư viện âm nhạc theo số phân cụm tối ưu nhất. Tiếp đó, Principal Component Analysis (PCA) sẽ được sử dụng để trích xuất các đặc điểm chính của dữ liệu giúp cho gia tăng độ chính xác của thuật toán phân cụm K-means. Cuối cùng, dựa vào dữ liệu nghe của người dùng, từ đó sẽ gợi ý cho người nghe những bài hát mà họ có thể hợp với sở thích nghe của bản thân.

3.1 Huấn luyện mô hình bằng thuật toán K-Means

K-means là một trong những thuật toán phân cụm phổ biến trong lĩnh vực học máy và khai phá dữ liệu. Nó được sử dụng để phân chia tập dữ liệu thành các cụm dựa trên sự tương đồng giữa các điểm dữ liệu.

Cách thức hoạt động của K-means như sau:

Input:

Số cụm mong muốn, k và cơ sở dữ liệu D={d1, d2,…dn} chứa n đối tượng dữ liệu.

Output:

Một tập hợp k cụm

Các bước:

1) Chọn ngẫu nhiên k đối tượng dữ liệu từ tập dữ liệu D làm trung tâm cụm ban đầu.

2) Lặp lại quá trình.

3) Tính khoảng cách giữa mỗi đối tượng dữ liệu d(i) (1 <= i <= n) và tất cả k tâm cụm c(j) (1 <= j <= k) và gán đối tượng dữ liệu di vào cụm gần nhất.

4) Với mỗi cụm j (1 <= j <= k), tính lại tâm cụm.

5) Cho đến khi không còn sự thay đổi ở trung tâm cụm.

Để có thể tìm ra số cụm tối ưu, chúng tôi sẽ sử dụng hai phương pháp đánh giá:

3.1.1. Giá trị Silhouette  
Giá trị Silhouette là một công cụ quan trọng trong K-Means clustering và các thuật toán phân cụm khác để đánh giá chất lượng của việc phân cụm dữ liệu. Giá trị này cho chúng ta biết được độ tách biệt giữa các cụm và xác định xem cách chia cụm hiện tại có tốt hay không. Để có thể tính được ra giá trị Silhouette cho một cụm, chúng ta sẽ tính trung bình của giá trị Silhouette của mỗi điểm theo công thức:

s(i) = (1)

Với b(i) là khoảng cách trung bình của điểm i tới tất cả các điểm thuộc cụm khác gần nhất, a(i) là khoảng cách trung bình của điểm i tới các điểm cùng cụm với nó và max{a(i), b(i)} là điểm lớn nhất giữa a(i) và b(i).

3.1.2. Phương pháp Elbow

Phương pháp Elbow là một kỹ thuật phổ biến được sử dụng trong phân cụm dữ liệu để xác định số lượng cụm tối ưu cho một thuật toán phân cụm như K-Means và được tính dựa trên tổng khoảng cách bình phương theo công thức:

SSD = (2)

Với K là số lượng cụm , Ci là tập hợp các điểm dữ liệu thuộc cụm thứ i, x là điểm dữ liệu trong Ci, μi là trung tâm của cụm Ci, và || x - μi || là khoảng cách Euclidean giữa điểm x và trung tâm μi.

Giá trị Silhouette và phương pháp Elbow là những phương pháp hữu ích trong việc tìm số lượng cụm tối ưu bởi tính dễ dùng, dễ áp dụng, thế nhưng cả hai thuật toán này đều có những nhược điểm khác nhau. Với giá trị Silhouette, phương pháp này phụ thuộc vào khoảng cách Euclidean, điều này sẽ tốn rất nhiều tài nguyên để tính toán từng điểm dữ liệu một. Trong khi đó, phương pháp Elbow trong một số trường hợp nhất định sẽ không tạo ra điểm uốn rõ ràng và cần thử nhiều giá trị K để tìm số cụm tối ưu dẫn đến tốn tài nguyên. Nhưng nhìn chung thì hai phương pháp trên vẫn là những phương pháp tốt nhất được ứng dụng trong việc tìm kiếm số lượng cụm cần phân chia.

3.2. Trực quan hóa bằng PCA

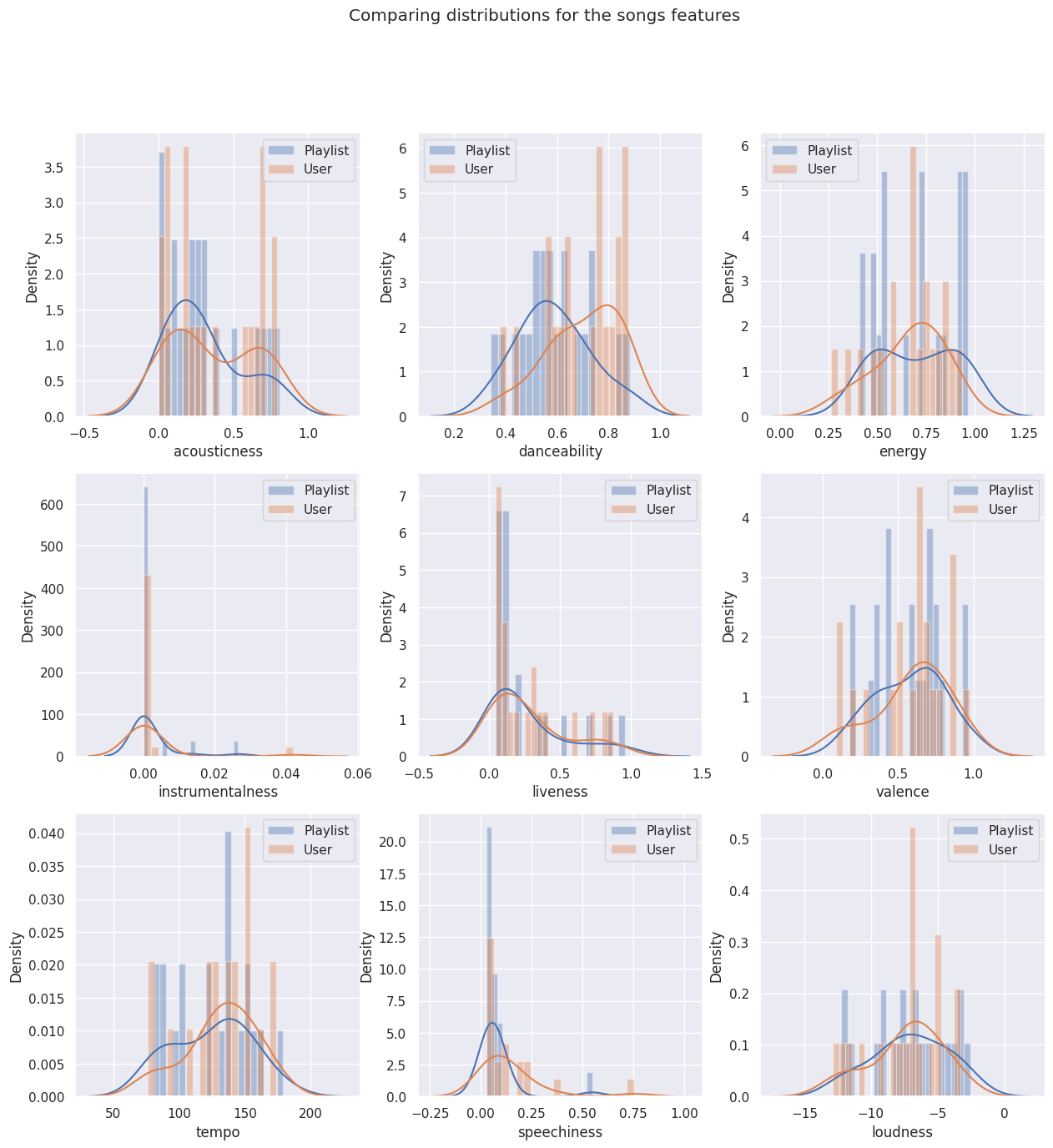
Cuối cùng, dù thuật toán đã tìm ra số cụm tối ưu, nhưng để có thể đánh giá độ chuẩn xác của thuật toán đã phân cụm, chúng tôi đã tích hợp thêm thuật toán phân tích thành phần chính. PCA là một thuật toán được sử dụng trong xử lý và phân tích dữ liệu để giảm chiều dữ liệu và trích xuất thông tin quan trọng, do đó thay vì chúng ta sẽ hiển thị ra dữ liệu trên những biểu đồ có nhiều chiều trở lên, phương pháp này sẽ giúp ta có thể biểu diễn dưới dạng dưới dạng 2D hoặc 3D và từ đó có thể dễ dàng thấy sự phân chia và tương tác giữa các cụm. [4] So với các phương pháp khác như t-SNE và UMAP, thì PCA có nhiều điểm như tốc độ xử lý, đơn giản hóa dữ liệu, trong khi đó các thuật toán còn lại sẽ làm mất nhiều thời gian, tìm tham số phù hợp và bị ảnh hưởng bởi những thành phần không quan trọng. Do đó chúng tôi đã chọn phương pháp này để đánh giá mô hình một cách nhanh nhất và thuận tiện nhất. Dù vậy, chúng ta không thể không bỏ qua những nhược điểm của PCA. PCA mất thông tin về cụm ban đầu và có thể dẫn đến mất mát thông tin về sự phân bố ban đầu của dữ liệu trong các cụm, đặc biệt nếu các thành phần chính mới không biểu thị rõ ràng các cụm hoặc các cụm trùng lặp.

Do đó, khi sử dụng PCA,chúng tôi cần cân nhắc và xem xét mức độ quan trọng của việc bảo toàn thông tin cụm ban đầu trong quyết định đánh giá mô hình.

1. **Kết quả**

Phần đầu tiên của dự án, mô hình đã được đào tạo và thử nghiệm bằng cách sử dụng K-means. Việc sử dụng phương pháp truyền thống đã cho chúng tôi ý tưởng về cách người dùng cần kiểm tra sự không nhất quán trong dữ liệu bằng cách sử dụng biểu đồ hoặc hàm toán học. Việc lựa chọn tính năng và lựa chọn mô hình được thực hiện bởi chúng tôi. Số lượng mô hình mà chúng tôi chọn mô hình tốt nhất dựa trên điểm chính xác là rất hạn chế do kiến thức về mô hình của chúng tôi còn hạn chế. Chúng tôi đã chọn ngẫu nhiên siêu tham số và không điều chỉnh nó vì chúng tôi nhận được kết quả như mong đợi. Tuy nhiên, nếu không đạt được kết quả như mong đợi, chúng tôi sẽ phải điều chỉnh các siêu tham số của mình, cung cấp cho nó các giá trị ngẫu nhiên mỗi khi chúng tôi muốn nó hoạt động tốt hơn.

Qua quá trình điều chỉnh số phân cụm cũng như đối chiếu dữ liệu từ 85 bài hát từ dữ liệu người dùng so với 20 bài hát mà hệ thống đưa ra, chúng tôi có thể đưa ra được những biểu đồ minh họa cho hiệu suất của mô hình (Hình 5).



Hình 5: Biểu đồ sự tương quan của danh sách bài hát gợi ý so với danh sách bài hát của người dùng

Dựa vào các biểu đồ, ta có thể thấy rằng các đặc trưng của các bài hát trong playlist được dự đoán bởi model có phân phối tương tự như các bài hát trong playlist của người dùng. Điều này cho thấy model đã hoạt động tốt trong việc học và dự đoán sở thích âm nhạc của người dùng. Tuy nhiên, để đánh giá chính xác hơn, bạn có thể cần phải thu thập phản hồi từ người dùng về chất lượng của playlist được dự đoán. Nếu người dùng thích các bài hát trong playlist mới, điều đó có nghĩa là model đã hoạt động tốt. Nếu không, bạn có thể cần phải điều chỉnh hoặc tinh chỉnh mô hình của mình.

1. **Kết luận và thảo luận**

Mô hình của của chúng tôi thực hiện khá tốt nhiệm vụ áp dụng K-means để phân cụm nhằm gợi ý bài hát tới người dùng dựa trên danh sách bài hát sẵn có của họ. Mô hình có thể được đưa vào thực tế trong quy mô nhỏ và đơn giản. Tuy nhiên, mô hình cũng vướng phải nhiều sự khó khăn do sự lựa chọn các bài hát thường không chỉ phụ thuộc vào thói quan mà còn dựa vào cảm xúc của người dùng.

Mô hình này của chúng tôi là chỉ là nỗ lực để xem làm cách nào có thể kết hợp một phương pháp đơn giản cho vấn đề gợi ý. Có rất nhiều các phương pháp khác có thể đem lại hiệu suất tốt hơn. Tuy nhiên việc sử dụng K-means đã giảm bớt gánh nặng của việc chọn thuật toán và điều chỉnh tham số. Nó trở nên hữu ích đặc biệt trong trường hợp huấn luyện lại, nơi người dùng có thể không luôn có thể chọn đúng loại thuật toán cho tập dữ liệu.

**Tài liệu tham khảo:**

[1] Robert L. Thorndike (1953). Who belongs in the family?. , 18(4), 267–276. doi:10.1007/bf02289263

[2] Kim, Dongmoon; Kim, Kun-su; Park, Kyo-Hyun; Lee, Jee-Hyong; Lee, Keon Myung (2007). [IEEE Sixth International Conference on Machine Learning and Applications (ICMLA 2007) - Cincinnati, OH, USA (2007.12.13-2007.12.15)] Sixth International Conference on Machine Learning and Applications (ICMLA 2007) - A music recommendation system with a dynamic k-means clustering algorithm. , (), 399–403. doi:10.1109/icmla.2007.97

[3] K. Allawadi and C. Vij, "A Smart Spotify Assistance and Recommendation System," 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2023, pp. 286-291, doi: 10.1109/InCACCT57535.2023.10141810.

[4] Z. Liumei, J. Fanzhi, L. Jiao, M. Gang and L. Tianshi, "K-means clustering analysis of Chinese traditional folk music based on midi music textualization," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2021, pp. 1062-1066, doi: 10.1109/ICSP51882.2021.9408762.