BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC KINH TẾ TP HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH



Logo, company name

Description automatically generated

**ĐỒ ÁN MÔN HỌC**

DỮ LIỆU LỚN VÀ ỨNG DỤNG

**Đề tài:**

***Ứng dụng Recommender System đề xuất bài hát từ bộ Spotify Data***

**Thành viên:** Lê Minh Hoàng

Trương Văn Kiệt

Nguyễn Văn Hoàng Dũng

**Giảng viên:** Đặng Nhân Cách

TP Hồ Chí Minh, tháng 3 năm 2022

***Mục Lục Nội Dung***

[**LỜI NÓI ĐẦU** 4](#_Toc131544549)

[**CHƯƠNG 1. GIỚI THIỆU** 5](#_Toc131544550)

[**1.** **Xác định đề tài** 5](#_Toc131544551)

[**2.** **Hướng tiếp cận** 5](#_Toc131544552)

[**3.** **Ngôn ngữ và thư viện sử dụng** 6](#_Toc131544553)

[**CHƯƠNG 2. CƠ SỞ LÍ THUYẾT VÀ ỨNG DỤNG LIÊN QUAN** 8](#_Toc131544554)

[**1.** **Cơ sở lý thuyết** 8](#_Toc131544555)

[**1.1.** **Hồi quy (*Regresison*)** 8](#_Toc131544556)

[**1.2.** **Phân cụm (*Clustering*)** 8](#_Toc131544557)

[**1.3.** **Các loại mô hình** 8](#_Toc131544558)

[**2.** **Ứng dụng liên quan** 9](#_Toc131544559)

[ **Hồi quy** 9](#_Toc131544560)

[ **Phân cụm** 9](#_Toc131544561)

[**CHƯƠNG 3: BỘ DỮ LIỆU VÀ ĐỀ XUẤT MÔ HÌNH** 10](#_Toc131544562)

[**1.** **Bộ dữ liệu** 10](#_Toc131544563)

[**2.** **Đề xuất mô hình** 11](#_Toc131544564)

[**2.1.** **Phương hướng tiếp cận 1** 11](#_Toc131544565)

[**2.2.** **Phương hướng tiếp cận 2** 12](#_Toc131544566)

[**CHƯƠNG 4: THỰC NGHIỆM VÀ PHÂN TÍCH** 13](#_Toc131544567)

[**1.** **Tiền xử lý dữ liệu** 13](#_Toc131544568)

[**1.1.** **Lọc các tên bài hát là chữ Latinh** 13](#_Toc131544569)

[**1.2.** **Xử lý dữ liệu bị thiếu** 14](#_Toc131544570)

[**1.3.** **Xử lý các dữ liệu “nhiễu”** 14](#_Toc131544571)

[**2.** **Phân tích khai phá dữ liệu (EDA)** 14](#_Toc131544572)

[**3.** **Chuẩn hóa dữ liệu** 17](#_Toc131544573)

[**4.** **Xây dựng mô hình** 18](#_Toc131544574)

[**4.1.** **Mô hình gợi ý bài hát dựa trên thang điểm đánh giá (1 – 10)** 18](#_Toc131544575)

[**4.2.** **Mô hình gợi ý bài hát dựa trên danh sách yêu thích** 21](#_Toc131544576)

[**CHƯƠNG 5: KẾT LUẬN** 25](#_Toc131544577)

***Mục lục ảnh***

[Hình 1. Bảng tóm tắt hướng tiếp cận 6](#_Toc131544374)

[Hình 2. Các thư viện được sử dụng 7](#_Toc131544375)

[Hình 3. Bảng thông tin các features 10](#_Toc131544376)

[Hình 4. Kiểu dữ liệu của các features 11](#_Toc131544377)

[Hình 5. Thông tin bộ dữ liệu sau khi lọc chữ Latinh 13](#_Toc131544378)

[Hình 6. Số dòng có dữ liệu bị thiếu của các cột 14](#_Toc131544379)

[Hình 7. Bảng chỉ số Thống kê của Spotify\_data.csv 14](#_Toc131544380)

[Hình 8. Biểu đồ countplot theo cột "release\_year" 15](#_Toc131544381)

[Hình 9. Biểu đồ thể hiện phân phối các giá trị của cột "popularity" 16](#_Toc131544382)

[Hình 10. Biểu đồ thể hiện mối quan hệ của "energy" và "tempo" 16](#_Toc131544383)

[Hình 11. Biểu đồ thể hiện mối quan hệ của "loudness" và "acousticness" 17](#_Toc131544384)

[Hình 12. Bảng kết quả chuẩn hóa dữ liệu 17](file:///C:\Users\baske\Downloads\BaiTapBigdata_Spotify%20(1).docx#_Toc131544385)

[Hình 13. Biểu đồ hệ số tương quan của các biến 18](#_Toc131544386)

[Hình 19. Khởi tạo mạng neural 20](#_Toc131544387)

# **LỜI NÓI ĐẦU**

Trong những năm vừa qua, thị trường **âm nhạc kỹ thuật số** đang phát triển với mức tăng trưởng chóng mặt và trở thành thị trường đầy tiềm năng. Sự đa dạng và tiện nghi là những đặc tính vượt trội của loại thị trường này. Những điều đó khiến cho thị trường này ngày càng thu hút được nhiều doanh nghiệp tham gia vào và chứng tỏ được chỗ đứng vững chãi.

Bài toán được đặt ra ở đây là đối với bộ dữ liệu được khai thác từ spotify - một doanh nghiệp có ảnh hưởng tương đối lớn trong thị trường âm nhạc kỹ thuật số thì làm sao để khai thác và sử dụng bộ dữ liệu đó một cách hiệu quả? Chính vì điều đó, nhóm em đã quyết định cào dữ liệu từ Spotify nhằm đưa ra những dự đoán về hành vi của các đối tượng khách hàng dùng Spotify.

Trong bài đồ án này, bộ dữ liệu thu thập được có tên là “**Spotify\_data.csv**” được cào về thông qua **Spotify API**. Bộ dữ liệu được sử dụng để phục vụ cho việc nghiên cứu mức độ hiệu quả của các thuật toán máy học, hình thành những phân tích, dự đoán liên quan đề hành vi của khách hàng nghe nhạc trên ứng dụng, website Spotify. Đây cũng là lý do nhóm em dành sự quan tâm và chọn đề tài “*Ứng dụng Máy Học vào dữ liệu được cào từ Spotify*” làm đồ án cuối kỳ học phần môn Dữ liệu lớn (Big Data) khoa Công nghệ thông tin (BIT) trường Đại học Kinh tế (UEH). Do điều kiện còn hạn chế về thời gian cũng như kiến thức bản thân có được nên sẽ khó tránh khỏi những sai sót. Kính mong giảng viên sẽ có những nhận xét và nhiệt tình góp ý để nhóm có thể tiếp thu kiến thức tốt hơn trong học phần này.

Cuối cùng, nhóm cũng xin phép cảm ơn thầy Đặng Nhân Cách đã hết sức tận tâm truyền đạt những kiến thức quý báu và hỗ trợ nhóm.

# **CHƯƠNG 1. GIỚI THIỆU**

## **Xác định đề tài**

Spotify là một dịch vụ cung cấp nhạc, podcast và các video kỹ thuật số cho phép người dùng truy cập hàng triệu bài hát và các nội dung khác của các nghệ sĩ trên khắp thế giới. Spotify luôn được người dùng biết đến với khả năng hiểu người dùng đến một cách kì lạ: Vào mỗi thứ 2 hàng tuần, hơn 100 triệu người dùng Spotify trên toàn thế giới sẽ có một playlist hoàn toàn mới đang đợi họ trên Spotify gọi là *Discovery Weekly*. Đó là một danh sách gồm 30 bài hát mà họ chưa bao giờ nghe trước đây, nhưng hầu như mọi người đều yêu thích nó, và điều đó thật kì diệu.

Vậy làm cách nào mà Spotify có thể chọn riêng 30 bài hát cho mỗi người mỗi tuần, và thậm chí hiện nay còn là mỗi ngày với *Daily Mix*? Hãy tìm hiểu xem cách mà Spotify mang âm nhạc đến với người dùng của họ.

Spotify không sử dụng duy nhất một mô hình mà sử dụng kết hợp một số dịch vụ để tạo ra công cụ đề xuất cực kỳ mạnh mẽ và đặc biệt. Để tạo ra *Discorvery Weekly* hay *Daily Mix*, họ sử dụng 3 mô hình chính:

* Mô hình lọc cộng tác (*Collaborative Filtering*): một mô hình phân tích hành vi của tất cả người dùng
* Mô hình xử lý ngôn ngữ tự nhiên (*Natural Language Processing – NLP*): phân tích ngôn từ
* Mô hình âm thanh: phân tích những audio track

Chúng ta sẽ tập trung sâu vào Mô hình lọc cộng tác hay *Collaborative Filtering* của Spotify ở đồ án này để có thể hiểu được cách thức hoạt động của mô hình thế nào mà lại hiệu quả tới mức thu hút được đông đảo người dùng như vậy và dựa vào đó để xây dựng một hệ thống đơn giản gợi ý bài hát cho người dùng (*Recommendation System*).

## **Hướng tiếp cận**

Mô hình này được dùng đầu tiên bởi Neflix và được sử dụng dựa trên những đánh giá về bộ phim của người dùng để đề xuất bộ phim đó cho người dùng tương tự khác. Sau sự thành công của Neflix, mô hình ngày càng trở nên phổ biến và là điểm bắt đầu cho bất cứ một ai muốn làm một mô hình đề xuất cho người dùng. Cũng giống như Neflix nhưng thay vì sử dụng mô hình dựa trên những đánh giá thì Spotify lại sử dụng dữ liệu phản hồi ngầm (*implicit feedback*), ví dụ như việc người dùng đó lưu bài hát đó vào playlist của mình hay là họ ghé thăm trang của ca sĩ sau khi nghe một bài hát của họ.

Dựa trên hệ thống gợi ý bài hát cho người dùng trên Spotify, nhóm chúng em sẽ sử dụng *Machine Learning* để có có thể có một mô hình dự đoán được bài hát nào sẽ được khách hàng ưa thích và từ đó gợi ý bài hát đó cho khách hàng. Nhằm tìm ra phương pháp phù hợp và hiệu quả, chúng em sẽ tiếp cận bài toán theo 2 hướng khác nhau (*gồm* *Hồi quy* và *Phân cụm)*:

* Với hướng tiếp cận đầu tiên, chúng em sẽ dựa vào đánh giá của người nghe cho từng bài nhạc trong một list được chuẩn bị sẵn (mức điểm dự kiến sẽ nằm trong khoảng từ 0 - 10), từ đó áp dụng hình hồi quy nhằm đánh giá và gán điểm cho từng bài hát và trả lại cho user gợi ý những bài hát có số điểm cao nhất từ kết quả của mô hình. Hạn chế của hướng tiếp cận này ở thực tế rằng mọi người khi nhận xét về về một bài hát chỉ có khả năng là thích hoặc không, họ sẽ không có 1 thang điểm để đánh giá cho từng bài.
* Với hướng tiếp cận thứ 2, chúng ta sẽ đánh giá dựa trên danh sách các bài hát mà người dùng thích, cách tiếp cận này có thể dễ dàng để sử dụng trong thực tế khi mỗi lần nghe xong một bài nhạc, user chỉ cần đánh giá là có thích nó hoặc không và hệ thống sẽ khi nhận lại và gợi ý các bài nhạc tương tự với các bài mà người nghe đánh giá là thích. Điểm khác biệt trong áp dụng mô hình học máy giữa 2 hướng tiếp cận này không chỉ ở việc chúng ta sẽ áp dụng mô hình hồi quy ở cách tiếp cận 1 và áp dụng mô hình phân lớp ở cách tiếp cận 2 mà còn là về tập mẫu để huấn luyện mô hình. Với cách tiếp cận 1 chúng ta chỉ cần huấn luyện mô hình với danh sách các bài hát đã được user đánh giá (từ 0-10) và áp dụng mô hình đã huấn luyện vào kho dữ liệu các bài nhạc để tìm ra những bài hát có số điểm cao nhất rồi gợi ý cho user. Ngược lại với cách tiếp cận 2, những bài nhạc mà người dùng đánh giá là thích sẽ được đánh dấu là 1, những bài nhạc còn lại sẽ tự động đánh dấu là 0, từ đó ta sẽ sử dụng toàn bộ kho dữ liệu để huấn luyện mô hình phân lớp và tìm ra xem bài nhạc nào có khả năng sẽ là bài nhạc được người dùng thích từ đó đưa ra gợi ý cho người dùng.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Ý tưởng** | **Mô hình học máy** | **Dữ liệu huấn luyện** |
| **Cách tiếp cận 1** | Sử dụng thang điểm đánh giá từ 0 - 10 cho mỗi bài nhạc user đã nghe | Hồi quy | Các bài nhạc đã được đánh giá |
| **Cách tiếp cận 2** | User chỉ cần đánh giá có thích một bài nhạc hay không. Với 1 là thích, 0 là không | Phân lớp | Toàn bộ dữ liệu |

Hình 1. Bảng tóm tắt hướng tiếp cận

## **Ngôn ngữ và thư viện sử dụng**

Trong bài này, chúng em chọn sử dụng ngôn ngữ Python - phổ biến với nhiều thư viện hỗ trợ các quy trình trong toàn chuỗi.

Các thư viện cụ thể sẽ được sử dụng trong bài gồm có *pandas* với mục đích đọc dữ liệu đầu vào, hỗ trợ dataframe, *seaborn* cho mục đích vẽ các biểu đồ

trực quan và *sklearn* dùng để tạo các mô hình học máy cũng như đánh giá mô hình. Bên cạnh đó, thư viện Spotipy cũng được sử dụng trong việc cào dữ liệu nhưng sẽ không được đề cập trong file code đính kèm.

Text

Description automatically generated

Hình 2. Các thư viện được sử dụng

# **CHƯƠNG 2. CƠ SỞ LÍ THUYẾT VÀ ỨNG DỤNG LIÊN QUAN**

## **Cơ sở lý thuyết**

### **Hồi quy (*Regresison*)**

Hồi quy là một phương pháp dùng để tính toán một mối quan hệ số giữa một biến đầu vào với một biến dự đoán đầu ra để dự đoán giá trị của biến đầu ra khi giá trị của biến đầu vào đã biết trước. Các mô hình hồi quy được sử dụng để dự đoán tương lai dựa trên cơ sở các biến đầu vào đã quan sát được.

### **Phân cụm (*Clustering*)**

Phân cụm là một kĩ thuật được sử dụng để phân nhóm các đối tượng có đặc điểm tương tự nhau. Kỹ thuật phân cụm hoạt động bằng các tìm cách phân chia các đối tượng có trong tập dữ liệu thành nhiều nhóm khác nhau dựa trên sự giống nhau của chúng.

Khác với phân lớp (Classifying), phân cụm là một phương pháp học Không giám sát (*Unsupervised learning*) vì nó không có thông tin giám sát từ phía người dùng. Thay vào đó, phân cụm tìm cách khám phá sự tương đồng tự nhiên trong tập dữ liệu và cố gắng xác định các nhóm đối tượng có đặc điểm giống nhau lại với nhau.

### **Các loại mô hình**

#### Linear Regression

Hồi quy tuyến tính là một kỹ thuật phân tích dữ liệu thống kê trong Machine Learning. Mô hình tuyến tính là một mô hình toán học trong đó với một tập các giá trị x của biến độc lập, mô hình sẽ tìm ra một hàm số tuyến tính để dự đoán biến phụ thuộc y.

Trong học máy, phương pháp hồi quy tuyến cũng được thực hiện như định nghĩa toán học, kỹ thuật này sẽ hoạt động tìm cách khớp dữ liệu với các đường tuyến tính (*linear lines*) để dự đoán giá trị của biến phụ thuộc trong tương lai.

#### Neural Network

Mạng *neural* được xây dựng dựa trên cơ sở các nơ ron nhân tạo (*artificial neurons*) – các đơn vị xử lý thông tin cơ bản của mạng. Các nơ ron nhân tạo được kết nối với nhau để tạo thành các lớp và mạng *neural*. Mỗi nơ ron nhân tạo có thể nhận đầu vào từ nhiều nơ ron nhân tạo khác, xử lý dữ liệu và sản xuất đầu ra. Các nơ ron nhân tạo được kết nối với nhau qua các trọng số (*weights*) để tạo thành các kết nối trọng số trong mạng.

Trong mạng *neural*, quá trình của việc học dựa trên việc điều chỉnh các giá trị trọng số để tìm ra một mô hình dự đoán tốt nhất cho bài toán được đưa ra. Các thuật toán như Gradient Descent và Backprogation được sử dụng để điều chỉnh các trọng số trong mạng *neural*.

#### Logistic Regression

Logistic Regression sử dụng để dự đoán xác suất của một biến phụ thuộc nhị phân (*Binary dependent variable*) dựa trên các biến đầu vào để sau đó phân loại chúng. Hàm sigmoid được sử dụng để ước tính xác suất các nhãn; hàm này đi qua một giá trị số thực đầu vào và trả về giá trị số thực trong khoảng [0,1], đại diện cho xác suất lớn nhất của biến phụ thuộc trong cả 2 trường hợp.

#### Random Forest

*Random Forest* là một thuật toán trích xuất tính toán và phân loại của một tập dữ liệu lớn bằng cách tạo ra một loạt các *Decision Tree.* Mỗi *Decision Tree* được tạo ra bằng cách chọn ngẫu nhiên một tập dữ liệu con từ tất cả dữ liệu và một tập chỉ mục con cho phép các *node* được phân tách.

Mỗi cây quyết định được sử dụng để phân loại dữ liệu, sau đó kết quả được tổng hợp từ tất cả các cây quyết định để tạo ra một kết quả phân loại chung. Quá trình này được thực hiện bằng cách sử dụng phương pháp Voting (bầu cử): quyết định cuối cùng sẽ được lấy bằng cách đếm số lương phiếu ủng hộ mỗi loại.

#### Gradient Boosting

*Gradient Boosting* xây dựng các mô hình dự đoán đáng tin cậy dựa trên việc kết hợp các mô hình đơn giản như *Decision Tree*. Điểm nổi bật của mô hình này là mỗi cây sẽ tập trung vào việc cải thiện sai sót của cây trước đó.

Các bước hoạt động của mô hình:

- Xây dựng một cây dự đoán đơn giản hoạt động trên toàn bộ các đặc trưng

- Tính toán sai số giữa giá trị dự đoán và giá trị thực tế

- Trọng số các trường hợp một cách thông minh để tập trung vào tập dữ liệu mà cây phân loại sai

- Xây dựng cây mới

## **Ứng dụng liên quan**

Một số ứng dụng của mô hình hồi quy và mô hình phân cụm trong học máy được áp dụng vào thực tế là:

### **Hồi quy**

- Dự đoán giá cổ phiếu: dự đoán giá cổ phiếu dựa trên các yếu tố như kinh tế, kỹ thuật, xã hội

- Dự đoán doanh thu: dự đoán doanh thu dựa trên các yếu tố kinh tế, marketing, sản phẩm

- Phân tích dữ liệu: xác định quan hệ giữa các biến trong dữ liệu

- Dự đoán các nguy cơ tai nạn giao thông: dự đoán dựa trên các yếu tố như mật độ giao thông, điều kiện đường bộ, thời tiết

- Dự báo kết quả bầu cử: dựa trên các yếu tố như điều kiện kinh tế, tầm ảnh hưởng để dự báo sự ủng hộ của các cử tri

### **Phân cụm**

- Phân tích khách hàng: phân tích các thông tin khách hàng như lịch sử mua hàng, độ tuổi, giới tính, khu vực sống và sở thích để tạo ra các nhóm khách hàng

- Dự đoán khách hàng: dự đoán xác suất khách hàng mua hàng và các mặt hàng mà họ có thể quan tâm, từ đó tối ưu chiến lược bán hàng

- Giám sát sản xuất: theo dõi quá trình sản xuất và phân loại các quá trình gây ra lỗi, thất thoát

- Phân tích dữ liệu y khoa: phân loại các bệnh nhân dựa trên các yếu tố như độ tuổi, giới tính, triệu chúng và lịch sử bệnh

- Phân loại tín hiệu thông minh: phân tích hành vi của các tín hiệu thông minh như máy móc tự động và tự lái

- Phân loại ảnh: phân loại ảnh dựa trên các đặc trưng như màu sắc, hình dạng và kích thước

# **CHƯƠNG 3: BỘ DỮ LIỆU VÀ ĐỀ XUẤT MÔ HÌNH**

## **Bộ dữ liệu**

Bộ dữ liệu được nhóm chúng em cào về từ kho dữ liệu của Spotify thông qua API. Hạn chế của Spotify API là giới hạn số lượng dữ liệu được tải về nên chúng em quyết định chọn cách sử dụng dữ liệu 10,000 mẫu được cào sẵn để tiếp cận thay vì sử dụng dữ liệu cào live từ Spotify.

Graphical user interface, text, application

Description automatically generated

Hình 3. Bảng thông tin các features

Trong bộ dữ liệu mẫu cào được có 17 features, phục vụ mục đích nhận biết, đo lường khác nhau.

Với feature *like* được tụi em tự chuẩn bị nhằm phục vụ cho mục đích nghiên cứu hướng tiếp cận thứ 2, khi mà người nghe có 1 danh sách các bài nhạc họ thích.

Table

Description automatically generated

Hình 4. Kiểu dữ liệu của các features

Nhìn qua các đánh giá sơ bộ của dữ liệu nhằm đưa ra các phương hướng xử lý dữ liệu phù hợp. Dễ thấy được, bộ dữ liệu cào về được chỉ tồn tại khoảng 300 dữ liệu rỗng → dễ dàng loại bỏ khỏi dữ liệu rỗng chỉ chiếm 3% của toàn bộ dữ liệu. Trong bộ dữ liệu có 11 dữ liệu định lượng → phù hợp với các mô hình hồi quy, phân lớp mà không cần bước vector hóa.

## **Đề xuất mô hình**

Để có thể chạy mô hình gợi ý nhạc, ta cần giả lập một người nghe có gu âm nhạc riêng. Ở bài này, người nghe sẽ thích những bài nhạc thuộc thể loại hip hop, rap hơn nên sẽ đánh giá cao những bài trong thể loại cũng như những bài này sẽ nằm trong list yêu thích giả lập này. Các dữ liệu của người nghe giả lập này sẽ được nhắc cụ thể hơn trong từng phương thức tiếp cận cũng như với mô hình máy học áp dụng tương ứng ở phần **Xây dựng mô hình**.

### **Phương hướng tiếp cận 1**

Như đã được đề cập, ý tưởng của mô hình theo phương hướng này hướng tới việc một người nghe sẽ đánh giá các bài hát từ 0 - 10 dựa vào trải nghiệm của họ. Sau khi có một danh sách các bài nhạc đã được người nghe gán điểm, chúng ta có thể dễ dàng huấn luyện mô hình hồi từ danh sách đó và sử dụng mô hình vào trong kho dữ liệu các bài nhạc để đưa ra gợi ý phụ hợp cho người nghe. Mỗi người nghe sẽ có những điểm đánh giá khác cho từng bài nhạc khác, từ đó sẽ tạo nên những mô hình được huấn luyện khác nhau dẫn đến đưa ra những gợi ý khác nhau phù hợp với từng cá nhân.

### **Phương hướng tiếp cận 2**

Khác với mô hình 1, ở mô hình này, chúng em đã chuẩn bị sẵn một feature *like* nhằm đánh giá xem người nghe có thích bài nhạc đó hay không với 1 là thích, 0 là không thích. Với mô hình này ta sẽ sử dụng các thuật toán phân lớp để huấn luyện toàn bộ dữ liệu. Nhưng để đánh giá được thuật toán, chúng ta sẽ chia bộ dữ liệu thành 2 phần, 80% để huấn luyện và 20% để test. Sau khi đã có bộ dữ liệu huấn luyện chúng ta bắt đầu thử các thuật toán phân lớp, bắt đầu với hồi quy logistic, thuật toán phổ biến nhất.

# **CHƯƠNG 4: THỰC NGHIỆM VÀ PHÂN TÍCH**

## **Tiền xử lý dữ liệu**

### **Lọc các tên bài hát là chữ Latinh**

Nhằm đảm bảo tính thống nhất của bộ dữ liệu, ta sẽ tiến hành lọc và chỉ chọn các bài hát có tên là chữ Latinh cho máy học.

Ta xác định *Biểu thức chính quy (Regular Expression)* cho các chữ latinh là  *“^[a-zA-Z\s]+$ “* và thực hiện lọc trên toàn bộ cột *track\_name* bằng phương thức *Match()* cho các kiểu dữ liệu chuỗi (*String*).

Table

Description automatically generated

Hình 5. Thông tin bộ dữ liệu sau khi lọc chữ Latinh

Kết quả cho ta thấy được bộ dữ liệu đã thỏa mãn được yêu cầu trên, không có bài hát nào có tên không phải là chữ latinh.

### **Xử lý dữ liệu bị thiếu**

Ta kiểm tra số lượng dòng bị thiếu của từng cột:

A picture containing table

Description automatically generated

Hình 6. Số dòng có dữ liệu bị thiếu của các cột

Bộ dữ liệu có 206 dòng bị thiếu dữ liệu tương ứng với các cột *tempo*, *speechiness*, *loudness*, *liveness*, *instrumentalness*, *energy*, *acousticness*, *valance*. Lượng dữ liệu bị thiếu ứng với khoảng 2% so với toàn bộ dữ liệu (10000 dòng), ít hơn so với chuẩn 5% nên ta thực hiện loại bộ đi các dữ liệu bị thiếu.

### **Xử lý các dữ liệu “nhiễu”**

Việc tìm các dữ liệu bị nhiễu sẽ dựa vào điểm z (giá trị z sẽ được tính bằng cách lấy hiệu của giá trị đang xét và trung bình của cột ứng với giá trị chia cho độ lệch chuẩn của cột): các dòng có điểm z > 3 hoặc điểm z < 3 sẽ ứng với các dòng có dữ liệu bị nhiễu, việc tìm sẽ được áp dụng trên phạm vi các cột có dữ liệu số.

Sau khi thực hiện nhóm em thu được kết quả bộ dữ liệu không có dữ liệu “nhiễu”.

## **Phân tích khai phá dữ liệu (EDA)**

A picture containing table

Description automatically generated

Hình 7. Bảng chỉ số Thống kê của Spotify\_data.csv

Ta có thể thấy được sự tăng trưởng trưởng số lượng bài hát tương đối ổn định của thị trường Âm nhạc kỹ thuật số từ năm 2000 đến 2015 và bắt đầu có sự giảm nhẹ ở 3 năm tiếp theo cho tới khi giảm đột ngột ở năm 2019 (Hình 8). Một lý do khác giải thích cho việc này là ngày càng nhiều các bài nhạc tiếng nước ngoài (các ngôn ngữ không phải chữ latinh) được phát hành trên Spotify.

Chart, bar chart

Description automatically generated

Hình 8. Biểu đồ countplot theo cột "release\_year"

Sự đa dạng của các thể loại nhạc được thể hiện rõ nét qua việc đã có 1387 thể loại nhạc xuất hiện trên Spotify.

Chart, histogram

Description automatically generated

Hình 9. Biểu đồ thể hiện phân phối các giá trị của cột "popularity"

Chart, scatter chart

Description automatically generated

Hình 10. Biểu đồ thể hiện mối quan hệ của "energy" và "tempo"

Chart, scatter chart

Description automatically generated

Hình 11. Biểu đồ thể hiện mối quan hệ của "loudness" và "acousticness"

## **Chuẩn hóa dữ liệu**

A picture containing diagram

Description automatically generatedTa có thể dễ dàng nhận biết dữ liệu định tính đang có các thước đo khác nhau. Để có thể huấn luyện mô hình đạt hiệu quả hơn, chúng em quyết định chuẩn hóa dữ liệu, đưa các giá trị này về một tỉ lệ, loại bỏ các thước đo riêng để thuận tiện hơn trong việc so sánh cũng như chạy mô hình. Trong bài này, chúng em sẽ dùng thuật toán standardscaler để thay đổi tỉ lệ. thuật toán được tính toán như sau: z = (x-u) / s với z là giá trị mới, x là giá trị cũ, u là giá trị trung bình và s là độ lệch chuẩn. Sau khi chuẩn hóa dữ liệu ta có kết quả như sau:

Hình 12. Bảng kết quả chuẩn hóa dữ liệu

Sau khi đã chuẩn hóa dữ liệu, ta có thể tiến hành so sánh mức độ tương quan giữa các biến, từ đó loại bỏ các biến có độ tương quan cao để có thể xây dựng mô hình hồi quy tốt hơn. Trong bày này, chúng em chọn threshold tương quan cao giữa các biến là 0.8, nếu vượt quá sẽ bị loại. Tuy nhiên trong bộ dữ liệu này không có cặp biến nào có tương quan cao nên không phải loại biến nào cả.

Chart

Description automatically generated with low confidence

Hình 13. Biểu đồ hệ số tương quan của các biến

## **Xây dựng mô hình**

### **Mô hình gợi ý bài hát dựa trên thang điểm đánh giá (1 – 10)**

Với mục đích phục cho huấn luyện mô hình hồi quy, nhóm chọn ngẫu nhiên 20 bài hát và gán điểm cho chúng (đánh giá điểm cao với những bài thuộc thể loại hip hop, rap).

**Table

Description automatically generated with medium confidence**

Hình 14. 20 bài hát được chọn ngẫu nhiên

Table

Description automatically generated

Hình 15. Các chỉ số ứng với 20 bài hát được chọn

Sau khi đã có giả lập của người nghe, chúng ta sẽ chọn và thử xem mô hình hồi quy nào sẽ phù hợp với bộ dữ liệu này.

Đầu tiên chúng ta sẽ sử dụng mô hình hồi quy phổ biển cũng như cơ bản nhất - hồi quy tuyến tính (linear regression). Với hồi quy tuyến tính, tốc độ huấn luyện mô hình khá nhanh, nhưng vì là mô hình đơn giản nên đôi khi kết quả đầu ra sẽ không tốt. Sau khi áp dụng mô hình hồi quy tuyến tính đã được huấn luyện với bộ dữ liệu người nghe giả lập, ta sẽ xem top 10 bài hát được đánh giá cao nhất:

Graphical user interface

Description automatically generated with medium confidence

Hình 16. Chỉ số của các bài hát được đề xuất bởi mô hình hồi quy

Độ phân hóa này quá cao khi bài cao điểm nhất đạt 89 điểm trong khi bài cao điểm thứ 10 đạt 17 điểm. Bên cạnh đó, 4 bài nhạc trong danh sách trên là thuộc thể loại hip hop, khá thấp so với nhóm em kỳ vọng.

Table

Description automatically generated

Hình 17. Tên các bài hát được đề xuất bởi mô hình quy

Ta cùng thử một mô hình hồi quy khác, sử dụng Neural Network để huấn luyện dữ liệu. Trong mô hình này, ta sử chọn sử dụng hàm kích hoạt relu, hàm mất mát tính dựa trên mean\_absolute\_error và hàm tối ưu là adam (phương pháp kết hợp giữa Momentum và RMSprop). Với mô hình sử dụng Neural Network, tốc độ huấn luyện dữ liệu sẽ lâu hơn và sẽ ngày càng scale lên khi dữ liệu lớn dần. Tuy nhiên, mô hình này sẽ cho kết quả tốt hơn so với hàm hồi quy tuyến tính thông thường.

Graphical user interface, text

Description automatically generated

Hình 19. Khởi tạo mạng neural

Ta có thể thấy kết quả trả về từ mô hình sử dụng neural network tốt hơn so với hồi quy tuyến tính thông thường khi 7/10 bài nhạc được gợi ý có thể loại hip hop.

Graphical user interface, text, application

Description automatically generated

Hình 18. Tên các bài hát được đề xuất bởi mô hình Neural Network

Nhìn chung, với hướng tiếp cận đầu tiên, sử dụng thang điểm để đánh giá các bài nhạc nên được gợi ý cho người nghe áp dụng mô hình hồi quy, ta có thể dùng mạng neural network để đưa ra gợi ý phù hợp hơn. Tuy nhiên đối với tập dữ liệu thực tế, khi số người nghe không chỉ có 1 và số lượng bài không bị giới hạn bởi 20 thì chúng ta cần có những mô hình phù hợp hơn, đảm thời gian đưa ra gợi ý nhanh hơn nhưng vẫn đảm bảo được gợi ý đúng bài người nghe muốn.

## **Mô hình gợi ý bài hát dựa trên danh sách yêu thích**

Graphical user interface, application

Description automatically generated

Sau khi train và test mô hình hồi quy logistic, ta có thể thấy sai số rơi vào khoảng 0.78 - 0.8.  Áp dụng mô hình này vào việc gợi ý các bài nhạc phù hợp cho người nghe, ta có thể thấy từ mô hình sử dụng hồi quy tuyến tính có được 5/10 bài thuộc thể loại hip hop (hình 20).

Graphical user interface

Description automatically generated with medium confidence

Hình 19. Kết quả mô hình Logistic Regression dựa trên danh sách yêu thích

Tiếp theo ta sẽ thử mô hình với thuật toán Random Forest cũng phổ biến không kém. Với thuật toán này chúng ta có range sai số khi cross check rơi vào khoảng 0.76 - 0.83 cùng với đó là kết quả trả về khả quan hơn khi có 7/10 bài thuộc thể loại hip hop.

Text

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

Hình 20. Kết quả mô hình RandomForest dựa trên danh sách yêu thích

Cuối cùng ta sẽ thử áp dụng thuật toán Gradient Boosting vào mô hình, đây là một trong những thuật toán đánh giá khá mạnh của học máy. Với thuật toán được áp dụng, chúng ta có thể thấy sai số thấp hơn hẳn khi chỉ rơi vào 0.24 - 0.34, chính vì vậy mà kết quả trả về cũng đang thể hiện cực kì tốt khi 10 / 10 bài nhạc gợi ý đều thuộc thể loại hip hop.

Text

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

Hình 21. Kết quả mô hình GradientBoosting dựa trên danh sách yêu thích

1. **Đánh giá**

Trong 2 phương hướng trên, ta thấy được phương hướng tiếp cận dựa trên danh sách thích mang tính thực tế hơn khi bộ dữ liệu được huấn luyện dựa trên danh sách thích (không giới hạn bài hát) trong khi phương hướng tiếp cận còn lại thì bị phụ thuộc vào liệu bài hát có được tính điểm hay không, đồng thời cũng hiệu quả hơn nhờ vào các chỉ số thu được.

Xét chi tiết hơn về phương diện thực tế thì việc đánh giá yêu thích bài hát hay không là tiện ích hơn cho người dung so với việc đánh giá điểm. Đồng thời việc đánh giá yêu thích hay không cũng thêm bài hát vào danh sách yêu thích, vừa giúp người dùng có thể tìm kiếm lại một cách dễ dàng hơn và giúp chúng ta có nhiều dữ liệu hơn để huấn luyện mô hình.

Ở phương hướng tiếp cận dựa trên danh sách bài hát yêu thích thì mô hình *GradientBoosting* trả kết quả tốt hơn khi trả về sai số thấp (10/10 bài hát cùng thể loại) nhưng lại có thời gian huấn luyện, thực thi lâu hơn tương đối so với các mô hình còn lại.

# **CHƯƠNG 5: KẾT LUẬN**

Vì sự giới hạn dữ liệu cào về của **Spotify API** nên bộ dữ liệu thu thập được chỉ dừng lại ở mức 10000 ngàn dòng nhưng thu thập được đa dạng và đầy đủ các loại feature cần thiết.

Mô hình có kết quả thực nghiệm tương đối ổn (sai số ở mức chấp nhận được rơi vào khoảng 0.24 – 0.34) đối với bộ dữ liệu có hơn 9000 dòng và quá trình thực hiện huấn luyện cho mô hình cũng không quá mất nhiều thời gian. Nếu bộ dữ liệu lớn hơn thì khoảng thời gian chạy sẽ lâu hơn và kết quả xong cũng xong khó như được kì vọng ban đầu tuy nhiên vì đây là những mô hình sử dụng thuật toán tương đối đơn giản nên đây là việc có thể dự đoán trước.

Nói tóm gọn, mặc dù kết quả nhận được ở đây chỉ là một phần nhỏ khiếm khuyết so với hệ thống **Gợi ý bài hát** (Spotify áp dụng tích hợp nhiều phương thức bổ sung thêm như NLP và Audio Track)thực tế của Spotify nhưng kết quả nhận được có thể gọi là ở mức thành công và áp dụng được vào thực tiễn.