TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

KHOA TOÁN - TIN HỌC

Báo cáo cuối kì
Python cho khoa học dữ liệu

GV bộ môn: Hà Văn Thảo

Nhóm 10

Châu Gia Kiệt	22110095
Trương Hồng Kiệt	22110096
Trần Trọng Kiên	22110093
Trương Minh Quân	22110174

Hồ Chí Minh, Tuesday, 14 January 2025

Mục lục

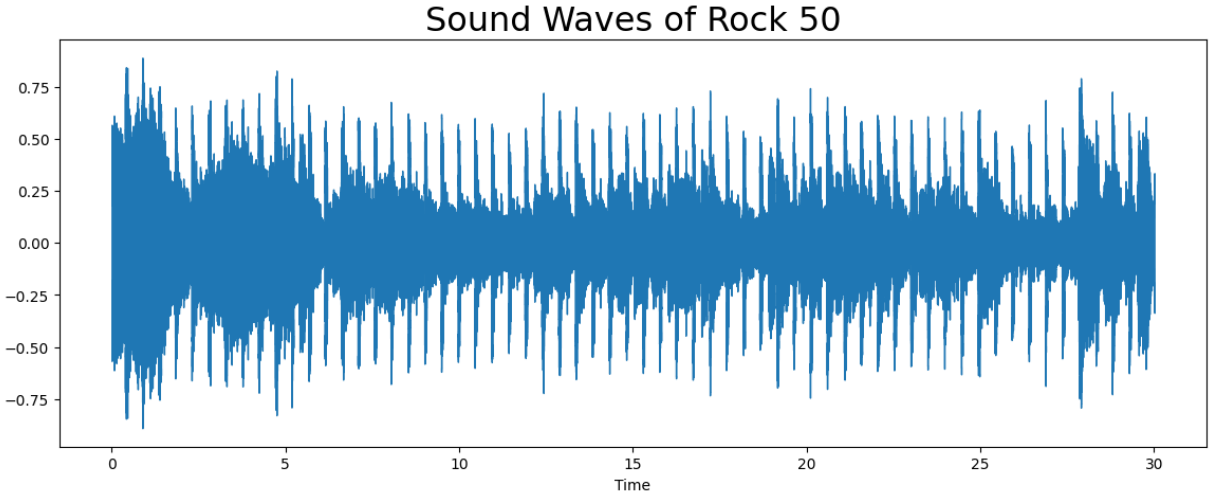
I. Giới thiệu đề tài

Âm nhạc là một phần không thể thiếu trong văn hóa và cuộc sống con người. Trong thời đại số hóa, lượng dữ liệu âm nhạc khổng lồ đã mở ra cơ hội để phân tích và ứng dụng công nghệ. Đề tài này tập trung vào việc khám phá các tập dữ liệu âm nhạc nhằm tìm ra những đặc điểm và xu hướng, như nhịp điệu, tâm trạng, và cấu trúc âm thanh. Bằng cách áp dụng các kỹ thuật học máy, dự án hướng đến việc phân loại chính xác các bài hát theo thể loại nhạc, giải quyết những thách thức như sự chồng chéo giữa các thể loại và sự đa dạng trong phong cách âm nhạc. Bên cạnh đó, đề tài còn triển khai một hệ thống gợi ý nhạc, sử dụng dữ liệu về sở thích người dùng và độ tương đồng giữa các bài hát để cá nhân hóa trải nghiệm. Đây là sự kết hợp giữa phân tích dữ liệu, học máy và triển khai thực tiễn, mang lại đóng góp quan trọng trong lĩnh vực công nghệ âm nhạc và nâng cao trải nghiệm người dùng.

II. Cơ sở lý thuyết

1. Tìm hiểu khái niệm về kiểu dữ liệu âm thanh.

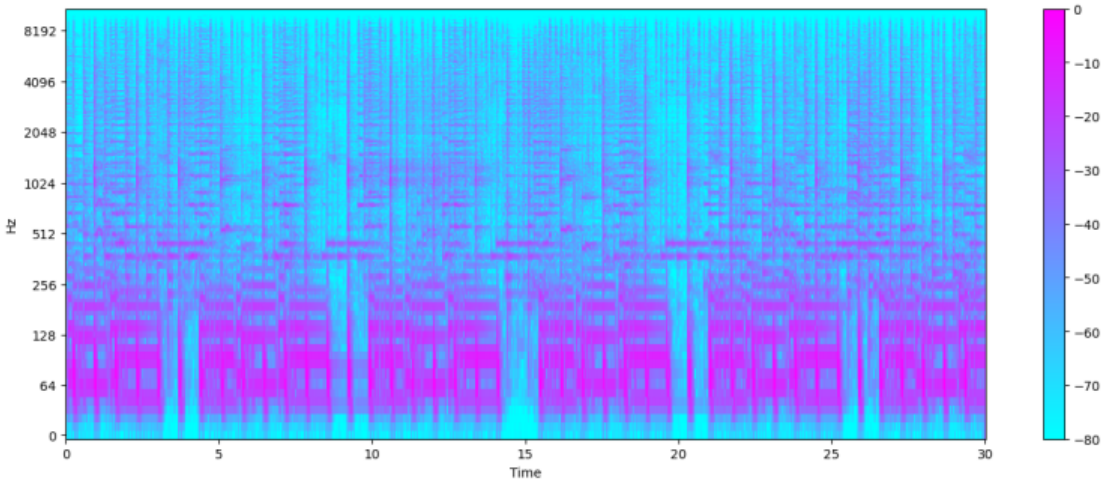
- **Âm thanh** là chuỗi dao động với cường độ áp suất thay đổi. Ví dụ sóng âm của một tệp âm thanh:



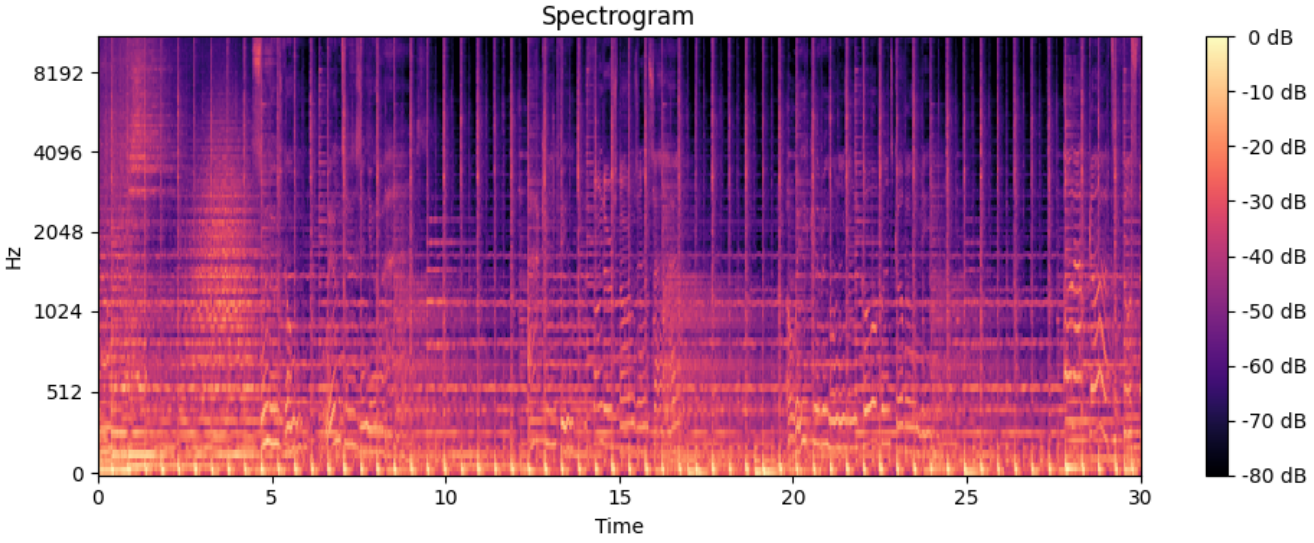
- **Tần số mẫu** (Sample Rate): Sample rate là số lượng mẫu âm thanh có thể thu thập từ tệp âm thanh trong mỗi giây. Mỗi mẫu (sample) có thể chứa các đặc trưng của âm thanh mà ta sẽ tìm hiểu thêm bên dưới. Tần số mẫu có thể được

đo bằng đơn vị Hz hoặc kHz.

- **Biểu đồ phổ**: Một biểu đồ phổ (spectrogram) là một cách biểu diễn trực quan của phổ tần số của một tín hiệu khi nó thay đổi theo thời gian. Nó giúp chúng ta quan sát cách năng lượng của các thành phần tần số khác nhau thay đổi theo thời gian như âm thanh. Khi áp dụng cho tín hiệu âm thanh, biểu đồ phổ đôi khi được gọi là sonograph, voiceprint, hoặc voicegram. Ví dụ:



- **Mel Spectrogram** là một dạng biểu đồ phổ mà trong đó tần số được ánh xạ sang thang Mel - một thang đo tần số dựa trên cách con người cảm nhận âm thanh. Thang Mel gần giống với cách tai người phân biệt âm thanh ở các tần số khác nhau. Ví dụ:

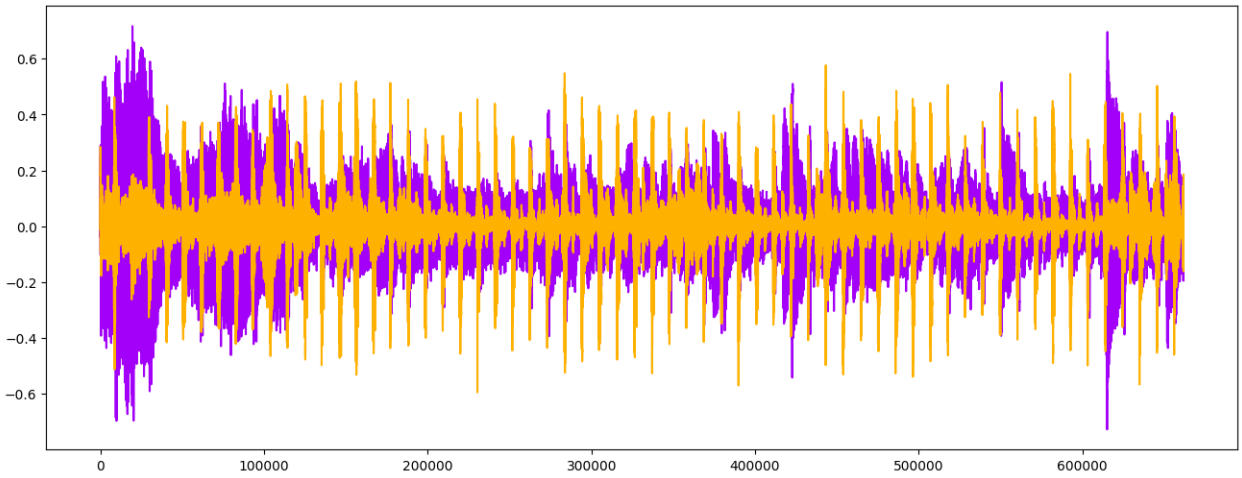


- **Zero Crossing Rate:** Đây là tỷ lệ mà tín hiệu thay đổi từ giá trị dương sang

giá trị âm hoặc ngược lại (thay đổi quanh giá trị 0) trong một khoảng thời gian nhất định. Đây là một đặc trưng phổ biến trong xử lý tín hiệu âm thanh, đặc biệt hữu ích để phân biệt giữa âm thanh tông cao (như giọng nói) và âm thanh ồn hoặc nhạc cụ.

2. Những đặc trưng của âm thanh

- **Harmonics and Perceptual** (Hài âm và đặc trưng cảm nhận):
  - Hài âm (Harmonics) là các tần số bội số của tần số cơ bản, tạo ra màu sắc âm thanh (timbre) mà tai người không thể phân biệt riêng lẻ, giúp phân biệt giữa các nhạc cụ hoặc giọng nói.
  - Đặc trưng cảm nhận (Perceptual features) liên quan đến cảm nhận của con người về âm thanh, như nhịp điệu và cảm xúc, ví dụ âm thanh mạnh mẽ hoặc thay đổi nhanh có thể gợi cảm giác phấn khích hoặc căng thẳng.



- **Tempo** (BPM - Beats Per Minute) là tốc độ của một bản nhạc, được đo bằng số nhịp (beats) trong một phút (BPM).
- **Spectral Centroid và Spectral Spread:** Đây là hai phương pháp đơn giản để trích xuất thông tin về vị trí và hình dạng phổ. Spectral Centroid chỉ ra vị trí "trung tâm khối lượng"của âm thanh, phản ánh nơi có nhiều năng lượng tần số nhất trong phổ âm thanh. Nó được tính toán như trung bình có trọng số của các tần số có mặt trong âm thanh, với trọng số là biên độ hoặc năng lượng của mỗi

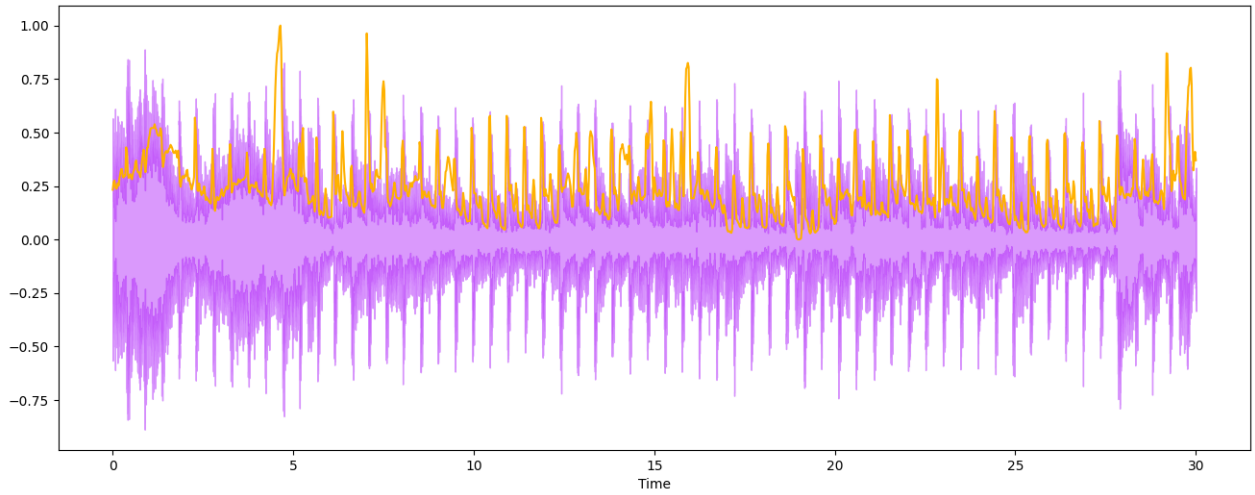
tần số. Trung tâm phổ được tính bằng công thức:

$$\text{Spectral Centroid} = \frac{\sum_{i=1}^{N} f_i \cdot A_i}{\sum_{i=1}^{N} A_i}$$

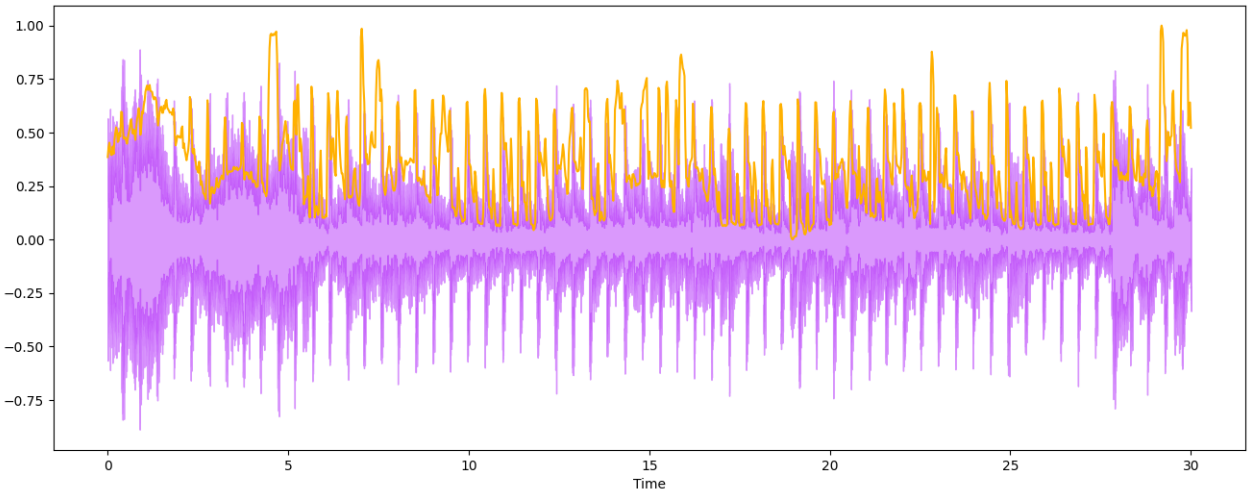
Trong đó:

- f_i: Tần số thứ i.
- A_i: Biên độ (hoặc năng lượng) tại tần số f_i.
- N: Tổng số tần số trong phổ.

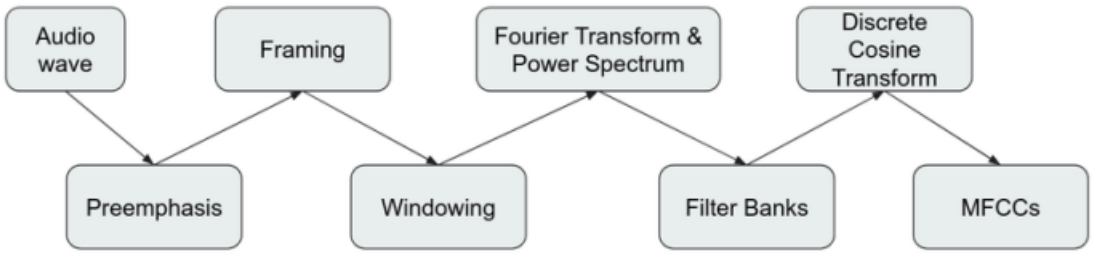
- **Spectral spread** đo sự phân bố của phổ xung quanh tâm của nó. Đây là một ví dụ về Spectral centroid:



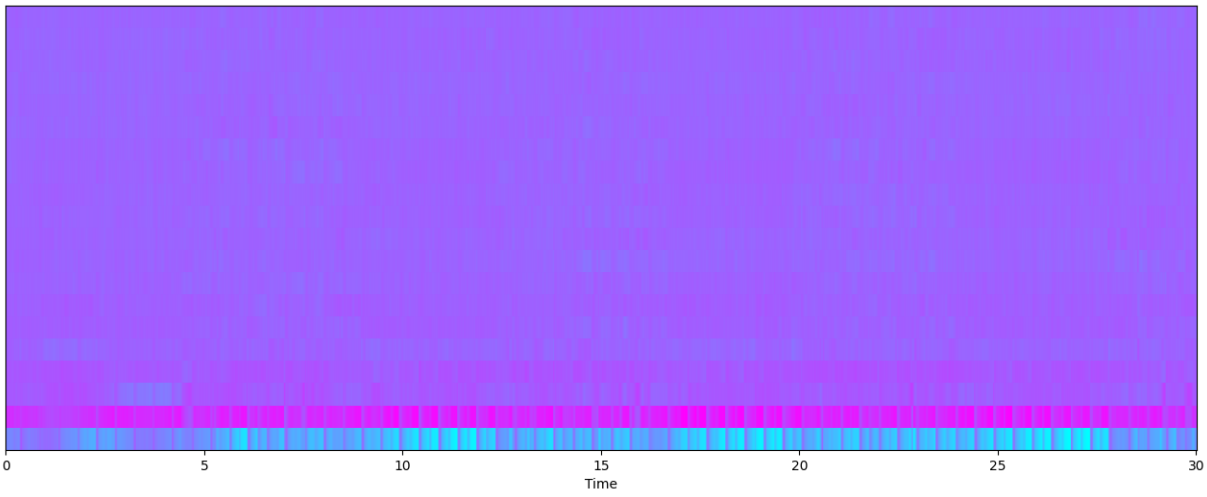
- **Spectral rolloff:** Đặc trưng này được định nghĩa là tần số mà dưới mức đó một phần trăm nhất định (thường là khoảng 0.9), phân bố cường độ của phổ được tập trung. Nó giúp mô tả hình dạng của tín hiệu âm thanh và phân biệt giữa các loại âm thanh có đặc trưng phổ khác nhau.



- **Mel-Frequency Cepstral Coefficients (MFCC):** Đây là tập hợp các đặc trưng (thường từ 10–20) mô tả hình dạng tổng thể của phổ âm thanh. MFCCs mô phỏng đặc điểm của giọng người, giúp phân tích và nhận dạng giọng nói hoặc âm thanh. Đây là sơ lược cách trích xuất MFCCs của âm thanh:

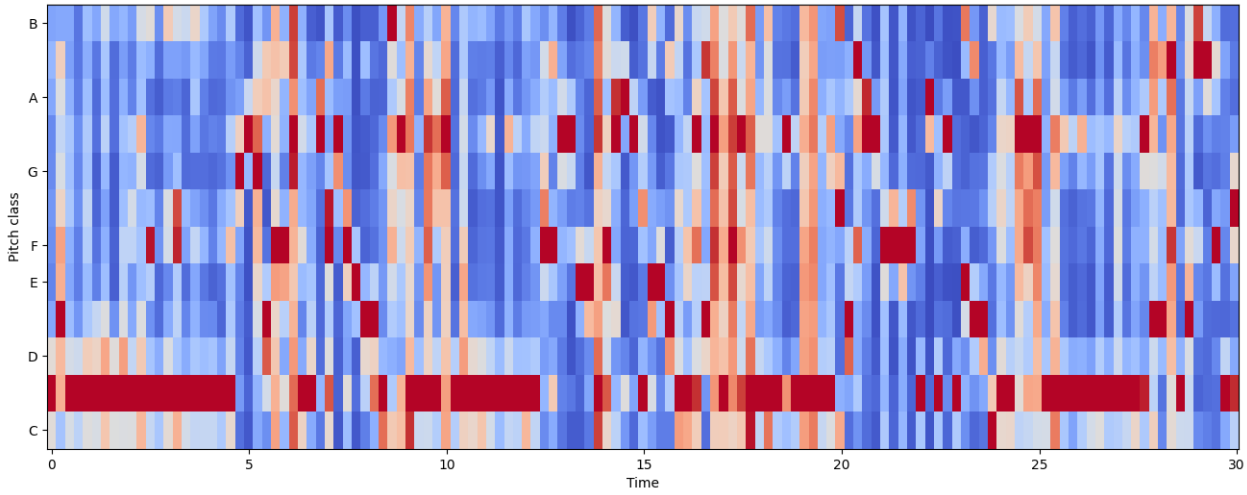


Đây là MFCSs của một tệp âm thanh bất kỳ.



- **Chroma Frequencies** là đặc trưng âm thanh đại diện cho âm nhạc, trong đó toàn bộ phổ âm thanh được chiếu lên 12 bin, tương ứng với 12 nửa cung trong

một quãng tám nhạc. Điều này giúp phân tích các yếu tố hòa âm trong âm nhạc.



III. Xử lý dữ liệu âm thanh sử dụng mô hình học máy

1. XGBoost (Extreme Gradient Boosting):

- **Khái quát về Gradient Boosting:**
  Gradient Boosting là một phương pháp học máy mạnh mẽ, thường được sử dụng trong các bài toán hồi quy và phân loại. Đặc điểm của Gradient Boosting là xây dựng mô hình thông qua việc kết hợp nhiều mô hình yếu (weak learners), thường là cây quyết định, theo phương pháp nâng cao dần (boosting). Quá trình này bắt đầu bằng việc huấn luyện một mô hình đơn giản, sau đó tính toán sai số (residuals) của mô hình đó. Mỗi mô hình tiếp theo sẽ được huấn luyện để sửa chữa sai số còn lại, bằng cách tập trung vào các điểm dữ liệu mà mô hình trước đó dự đoán sai. Mỗi mô hình bổ sung được kết hợp vào mô hình tổng thể, giúp cải thiện dần độ chính xác của dự đoán.
  Một trong những thuật toán nổi bật sử dụng phương pháp Gradient Boosting là XGBoost, LightGBM và CatBoost, đều là những công cụ phổ biến trong việc giải quyết các bài toán thực tế, đặc biệt là trong các cuộc thi máy học như Kaggle. Phương pháp này có thể hoạt động rất hiệu quả khi dữ liệu có nhiều biến phức tạp.
- **Khái quát về XGBoost:** XGBoost (Extreme Gradient Boosting) là một giải

thuật được base trên gradient boosting, tuy nhiên kèm theo đó là những cải tiến to lớn về mặt tối ưu thuật toán, về sự kết hợp hoàn hảo giữa sức mạnh phần mềm và phần cứng, giúp đạt được những kết quả vượt trội cả về thời gian training cũng như bộ nhớ sử dụng.
Mã nguồn mở với xấp xỉ 350 contributors và xấp xỉ 3,600 commits trên Gihub, XGBoost cho thấy những khả năng ứng dụng đáng kinh ngạc của mình như :

- XGBoost có thể được sử dụng để giải quyết được tất cả các vấn đề từ hồi quy (regression), phân loại (classification), ranking và giải quyết các vấn đề do người dùng tự định nghĩa.
- XGBoost hỗ trợ trên Windows, Linux và OS X.
- Hỗ trợ tất cả các ngôn ngữ lập trình chính bao gồm C ++, Python, R, Java, Scala và Julia.
- Hỗ trợ các cụm AWS, Azure và Yarn và hoạt động tốt với Flink, Spark và các hệ sinh thái khác.

XGBoost cung cấp hiệu suất và độ chính xác cao, xử lý các tập dữ liệu lớn một cách hiệu quả, ngăn ngừa quá khớp thông qua chính quy hóa, hỗ trợ xử lý song song để đào tạo nhanh hơn, quản lý các giá trị bị thiếu một cách hiệu quả và cung cấp đánh giá tầm quan trọng của tính năng, khiến nó trở thành lựa chọn được ưa chuộng cho mô hình dự đoán

2. Convolutional Neural Network:

1. Giới Thiệu

Mạng Neural Tích chập (CNN) là một kiến trúc mạng neural sâu (deep learning) được thiết kế đặc biệt để xử lý dữ liệu có cấu trúc không gian như hình ảnh, video hoặc âm thanh. CNN trở nên nổi tiếng nhờ khả năng tự động trích xuất đặc trưng (feature extraction) từ dữ liệu đầu vào mà không cần can thiệp thủ công, giúp chúng vượt trội trong các bài toán thị giác máy tính (computer vision) như phân loại ảnh, nhận diện vật thể, và phân vùng ảnh (segmentation).

2. Kiến Trúc Cơ Bản của CNN

- Lớp Tích chập (Convolutional Layer)
  - Chức năng: Áp dụng các bộ lọc (filters/kernels) để trích xuất đặc trưng từ dữ liệu đầu vào.
  - Cơ chế:
    - Mỗi bộ lọc trượt qua ảnh (theo stride - bước nhảy) và tính tích chập (dot product) giữa các trọng số của bộ lọc và vùng ảnh tương ứng.
    - Padding: Thêm các pixel giá trị 0 xung quanh ảnh để giữ nguyên kích thước đầu ra (same padding) hoặc không thêm (valid padding).
    - Kích thước đầu ra: Phụ thuộc vào kích thước ảnh đầu vào, bộ lọc, stride, và padding.
- Lớp Kích hoạt (Activation Layer): Hàm ReLU (Rectified Linear Unit): Thường được sử dụng để thêm tính phi tuyến (non-linearity), giúp mạng học các mẫu phức tạp. Công thức: $f(x) = max(0, x)$
- Lớp Gộp (Pooling Layer)
  - Chức năng: Giảm kích thước không gian của dữ liệu, giảm tính toán và tránh overfitting.
  - Max Pooling: Chọn giá trị lớn nhất trong vùng cửa sổ.
  - Average Pooling: Lấy giá trị trung bình của vùng cửa sổ.
- Lớp Kết nối Đầy đủ (Fully Connected Layer): Đặt ở cuối mạng, biến đổi dữ liệu thành vector 1D để phân loại hoặc hồi quy.

- Huấn Luyện CNN
  - Hàm mất mát (Loss Function): Ví dụ: Cross-entropy cho bài toán phân loại.
  - Tối ưu hóa: Sử dụng các thuật toán như SGD, Adam để cập nhật trọng số.
  - Kỹ thuật Regularization:
    * Dropout: Tắt ngẫu nhiên một số neuron trong quá trình huấn luyện.
    * Batch Normalization: Chuẩn hóa đầu vào của các lớp để tăng tốc độ hội tụ.
    * Data Augmentation: Tăng cường dữ liệu bằng cách xoay, lật, phóng to/thu nhỏ ảnh.

  CNN là công cụ mạnh mẽ trong deep learning, đặc biệt với dữ liệu có cấu trúc không gian. Sự phát triển của các kiến trúc như ResNet hay ứng dụng transfer learning tiếp tục mở rộng khả năng của CNN trong cả nghiên cứu và thực tiễn. Tuy nhiên, việc hiểu rõ cơ chế hoạt động và tối ưu hóa mạng vẫn là thách thức quan trọng.

IV. Tổng quan về bộ dữ liệu

1. Giới thiệu về bộ dữ liệu

Bộ dữ liệu về âm thanh: GTZAN

- **Genres Original:** Đây là tập dữ liệu GTZAN nổi tiếng, được ví như "MNIST của âm thanh". Bộ dữ liệu bao gồm 10 thể loại nhạc, mỗi thể loại chứa 100 tệp âm thanh với độ dài 30 giây.
- **Images Original:** Một biểu diễn trực quan cho mỗi tệp âm thanh được tạo ra. Để phân loại dữ liệu, một cách tiếp cận là sử dụng mạng nơ-ron nhân tạo (Neural Networks - NN). Vì mạng NN (đặc biệt là CNN, loại được sử dụng trong nghiên cứu này) thường tiếp nhận dữ liệu đầu vào dưới dạng hình ảnh, nên các tệp âm thanh đã được chuyển đổi thành Mel Spectrograms để phù hợp. Phần này sẽ được giải thích chi tiết hơn trong nghiên cứu.

Ngoài ra còn hai tệp dữ liệu kiểu CSV:

- Tệp đầu tiên chứa các đặc trưng (features) của mỗi bài hát (độ dài 30 giây). Cụ thể, mỗi bài hát được tính toán giá trị trung bình và phương sai trên nhiều đặc trưng có thể trích xuất từ tệp âm thanh.
- Tệp thứ hai cũng có cấu trúc tương tự, nhưng các bài hát được chia nhỏ thành các đoạn âm thanh dài 3 giây. Điều này làm tăng lượng dữ liệu lên gấp 10 lần, từ đó cải thiện hiệu quả của các mô hình phân loại. Trong xử lý dữ liệu, nhiều dữ liệu hơn thường mang lại kết quả tốt hơn.

2. Chi tiết về bộ dữ liệu

Trong phần này ta sẽ đi sâu vào chi tiết của 2 tệp CSV
features_30_seconds.csv: đây là tệp chứa các đặc trưng được trích xuất từ độ dài đầy đủ trên một tệp âm thanh

- filename: Tên tệp nhạc, dùng để định danh đoạn nhạc dài 30 giây.
- tempo: Tốc độ của bài hát, đo bằng số nhịp mỗi phút (BPM).
- beats: Số lượng nhịp đập được phát hiện trong đoạn nhạc
- chroma_stft: Trung bình giá trị phổ sắc điệu, biểu diễn cường độ năng lượng tại từng nốt nhạc.
- spectral_centroid: Tần số trung bình trọng số của đoạn nhạc, thể hiện độ sáng của âm thanh
- spectral_bandwidth: Độ rộng phổ tần số, biểu thị mức độ biến thiên tần số của đoạn nhạc.
- rolloff: Tần số mà 85
- zero_crossing_rate: Tần suất mà tín hiệu vượt qua trục không
- mfcc1, mfcc2, ..., mfcc20: Các hệ số Mel-frequency cepstral coefficients (MFCCs), đại diện cho đặc trưng tần số của âm thanh.
- genre: Thể loại nhạc của đoạn nhạc

features_3_seconds.csv: tệp này chứa các đặc trưng như sau:

- filename: Tên tệp nhạc, dùng để định danh đoạn nhạc dài 3 giây.
- tempo: Tốc độ của đoạn nhạc, đo bằng số nhịp mỗi phút (BPM).

- beats: Số lượng nhịp đập được phát hiện trong đoạn nhạc.
- chroma_stft, spectral_centroid, spectral_bandwidth, rolloff, zero_crossing_rate, mfcc1, ..., mfcc20: Các đặc trưng giống hoàn toàn như trong tệp features_30_sec.csv, nhưng được tính toán trên các đoạn nhạc ngắn hơn (3 giây).
- genre: Thể loại nhạc của đoạn nhạc

V. Xử lí dữ liệu

1. Khai phá dữ liệu

Trong phần này, chúng ta sẽ thực hiện khai phá dữ liệu từ tệp `features_30_sec.csv`. Tệp này chứa các đặc trưng âm thanh đã được trích xuất từ các tệp âm thanh trong bộ dữ liệu. Mỗi tệp âm thanh có độ dài 30 giây, và trong đó, mỗi thể loại có 100 bài hát, tổng cộng là 1000 bài hát, tương ứng với 10 thể loại.
Tệp csv có 1000 dòng (mỗi dòng tương ứng với một bài hát), và 60 cột gồm tên của tệp âm thanh, thể loại của bài hát và 58 đặc trưng (bao gồm giá trị trung bình và phương sai).
Dưới đây là cách ta đọc tệp CSV và kiểm tra dữ liệu:

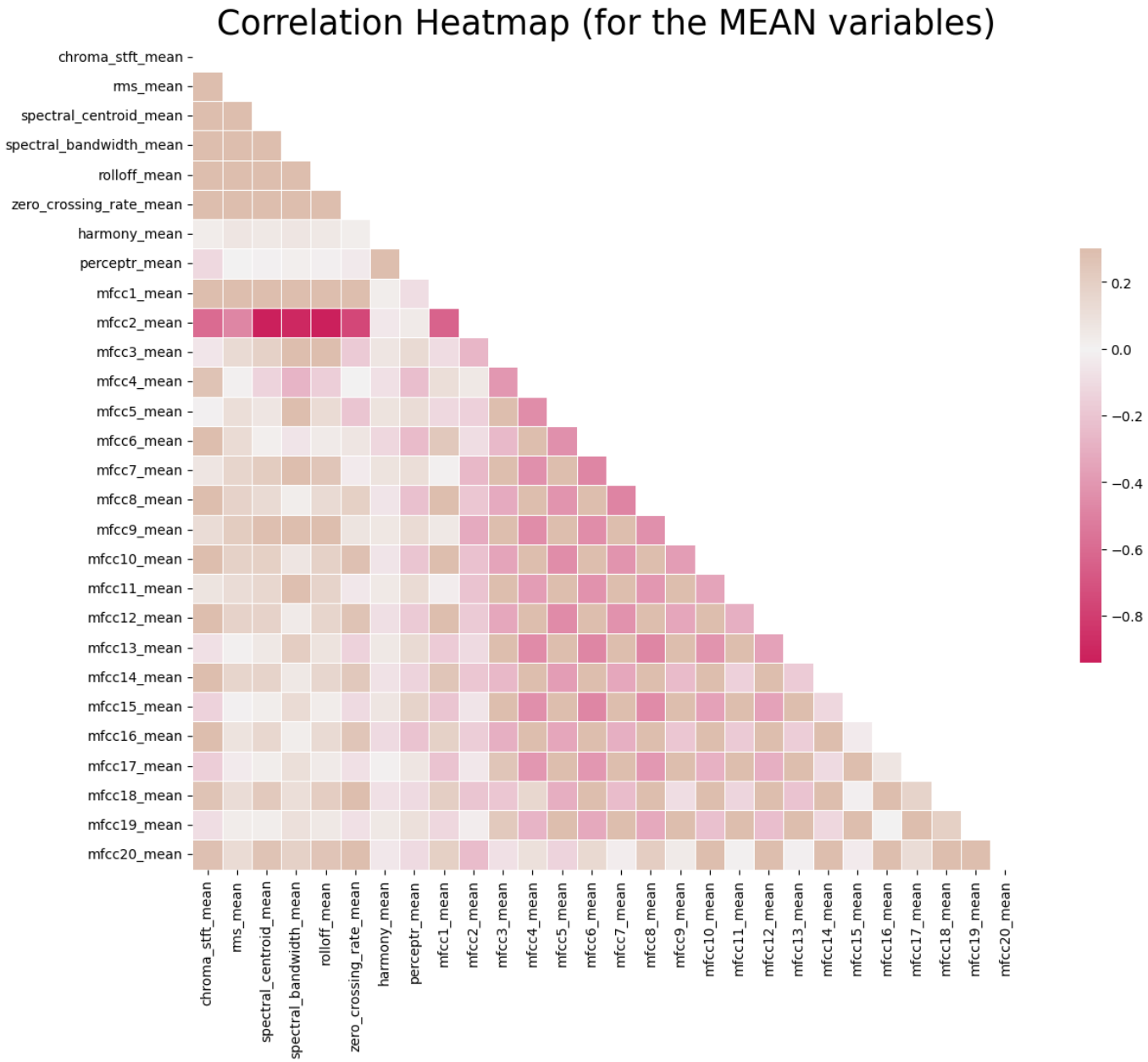
	filename	length	chroma_stft_mean	chroma_stft_var	rms_mean	rms_var	spectral_centroid_mean	spectral_centroid_var	spectral_bandwidth_mean
0	blues.00000.wav	661794	0.350088	0.088757	0.130228	0.002827	1784.165850	129774.064525	2002.449060
1	blues.00001.wav	661794	0.340914	0.094980	0.095948	0.002373	1530.176679	375850.073649	2039.036516
2	blues.00002.wav	661794	0.363637	0.085275	0.175570	0.002746	1552.811865	156467.643368	1747.702312
3	blues.00003.wav	661794	0.404785	0.093999	0.141093	0.006346	1070.106615	184355.942417	1596.412872
4	blues.00004.wav	661794	0.308526	0.087841	0.091529	0.002303	1835.004266	343399.939274	1748.172116

spectral_bandwidth_var	...	mfcc16_var	mfcc17_mean	mfcc17_var	mfcc18_mean	mfcc18_var	mfcc19_mean	mfcc19_var	mfcc20_mean	mfcc20_var	label
85882.761315	...	52.420910	-1.690215	36.524071	-0.408979	41.597103	-2.303523	55.062923	1.221291	46.936035	blues
213843.755497	...	55.356403	-0.731125	60.314529	0.295073	48.120598	-0.283518	51.106190	0.531217	45.786282	blues
76254.192257	...	40.598766	-7.729093	47.639427	-1.816407	52.382141	-3.439720	46.639660	-2.231258	30.573025	blues
166441.494769	...	44.427753	-3.319597	50.206673	0.636965	37.319130	-0.619121	37.259739	-3.407448	31.949339	blues
88445.209036	...	86.099236	-5.454034	75.269707	-0.916874	53.613918	-4.404827	62.910812	-11.703234	55.195160	blues

Bản đồ nhiệt tương quan (Correlation Heatmap) cho trung bình của các đặc trưng

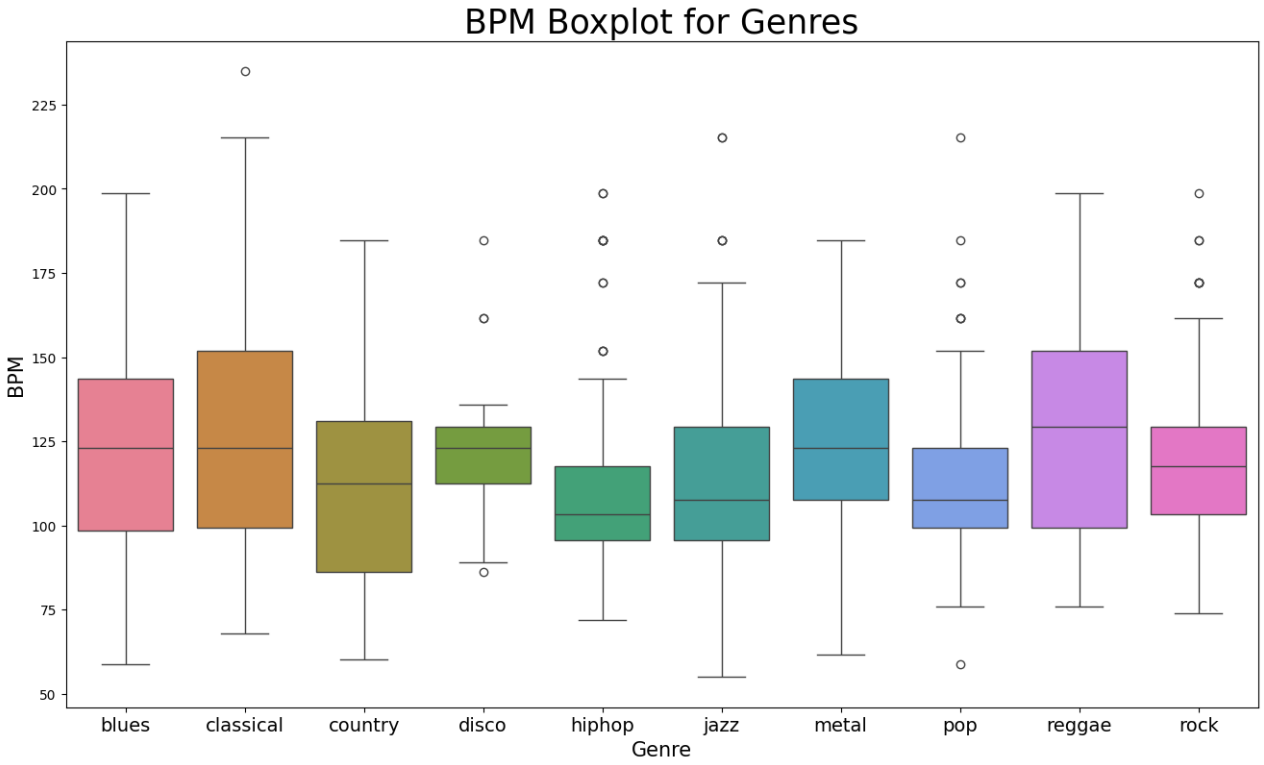
Bản đồ nhiệt tương quan giúp chúng ta quan sát mối quan hệ giữa các đặc trưng. Các đặc trưng có giá trị trung bình (mean) sẽ được tính toán tương quan và vẽ dưới dạng bản đồ nhiệt.

- Các đặc trưng MFCC có mức tương quan cao với nhau, đặc biệt là các cặp gần nhau như mfcc2_mean và mfcc3_mean. Điều này có thể gây ra đa cộng tuyến (multicollinearity) khi sử dụng các mô hình tuyến tính.
- Các đặc trưng như chroma_stft_mean, rms_mean, hoặc spectral_centroid_mean có tương quan thấp hơn với các đặc trưng MFCC. Điều này chứng tỏ chúng có thể cung cấp thêm thông tin độc lập.



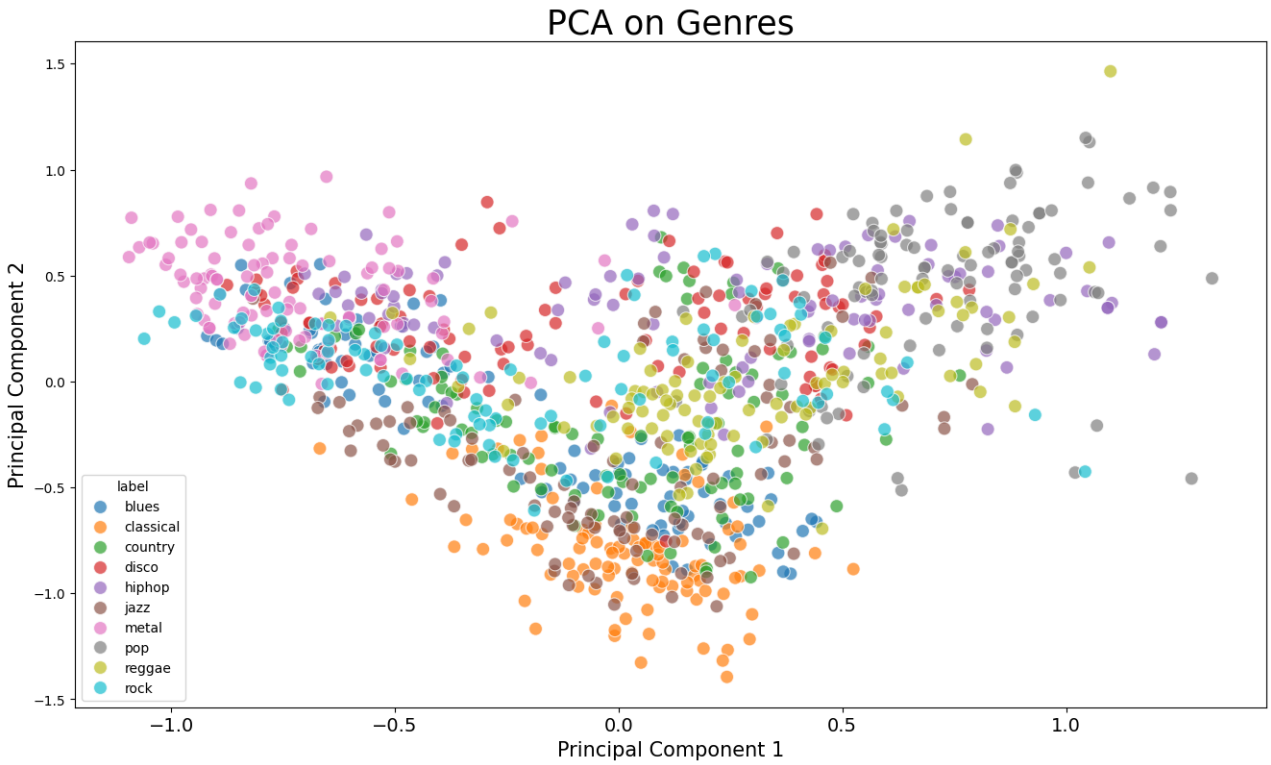
Biểu đồ hộp (Box plot) biểu diễn sự phân bố của các thể loại theo BPM.
Boxplot:

- Một số đặc trưng có sự khác biệt rõ ràng giữa các nhãn, nghĩa là chúng có khả năng phân biệt cao. Ví dụ: các đặc trưng như MFCC hoặc spectral có sự khác biệt lớn ở trung vị hoặc biên độ.
- Một số đặc trưng có outliers khá rõ ràng nên có thể ảnh hưởng đến các mô hình thống kê hoặc máy học. Ta có thể cân nhắc xử lý các outliers này (ví dụ như chuẩn hoá) trước khi sử dụng dữ liệu huấn luyện mô hình.



Phân tích thành phần chính (PCA) - Để trực quan hóa các nhóm thể loại nhạc

Phân tích thành phần chính (PCA) giúp chúng ta giảm chiều dữ liệu và trực quan hóa các nhóm thể loại nhạc. Để thực hiện PCA, dữ liệu cần được chuẩn hóa trước, sau đó chúng ta áp dụng PCA với 2 thành phần chính để trực quan hóa trên biểu đồ phân tán (scatter plot).



2. Phân loại thể loại âm nhạc (Music genres classification)

Các lớp thể loại: Trước khi đến với phần huấn luyện mô hình phân loại thể loại nhạc bằng máy học, ta sẽ chia 10 thể loại trong tập dữ liệu thành 10 lớp sử dụng trong bài toán này (bài toán phân loại đa lớp):

- Lớp 0: blues
- Lớp 1: classical
- Lớp 2: country
- Lớp 3: disco
- Lớp 4: hiphop
- Lớp 5: jazz
- Lớp 6: metal
- Lớp 7: pop
- Lớp 8: reggae
- Lớp 9: rock

3. Mô hình

Trong dự án này, ta sẽ thử nghiệm trên hai mô hình **XGBoost(Extreme Gradient Boosting)** và **Mạng nơ-ron tích chập (Convolutional Neural Network - CNN)**

Mô hình XGBoost (Extreme Gradient Boosting)

Mô hình XGBoost, một thuật toán mạnh mẽ dựa trên phương pháp boosting, để phân loại các thể loại âm nhạc dựa trên các đặc trưng của đoạn âm thanh.
Dữ liệu đầu vào của file feature_3_sec.csv chứa các đặc trưng của các đoạn âm thanh được chia thành các đoạn 3 giây. Chia thành hai phần: tập huấn luyện (70%) và tập kiểm tra (30%), giúp đánh giá hiệu năng của mô hình trên dữ liệu chưa từng thấy

Mô hình mạng nơ-ron tích chập (Convolutional Neural Network - CNN)

Mô hình CNN được xây dựng để phân loại thể loại âm nhạc trực tiếp từ các đặc trưng biểu đồ phổ trên thang đo Mel (Mel Spectrogram)
Dữ liệu âm thanh được chuyển đổi thành biểu đồ Mel Spectrogram, sau đó được chuẩn hóa và lưu trữ dưới dạng các tensor để làm đầu vào cho mô hình CNN. Mô hình CNN được xây dựng với hai khối tích chập (Convolutional Blocks) và các lớp kết nối đầy đủ (Fully Connected Layers).

Mô hình đề xuất đơn giản

Hệ thống đề xuất nhạc là một ứng dụng phổ biến trong các dịch vụ phát nhạc trực tuyến, giúp người dùng khám phá các bài hát tương tự dựa trên sở thích cá nhân hoặc đặc điểm âm thanh. Trong nghiên cứu này, chúng tôi sử dụng độ tương đồng cosine (cosine similarity) để xây dựng một hệ thống đề xuất nhạc đơn giản. Phương pháp này đo lường mức độ tương đồng giữa các bài hát dựa trên các đặc trưng của âm thanh.
Tổng quan về độ tương đồng cosine Độ tương đồng cosine là một phương pháp phổ biến để đo lường mức độ tương tự giữa hai vector trong không gian đa chiều

$$\text{Cosine Similarity} = \frac{A.B}{||A||.||B||}$$

Trong đó:

- $A.B$ là tích vô hướng (dot product) của hai vector.
- $||A||$ và $||B||$ là độ dài (norm) của vector A và B.

Giá trị của độ tương đồng cosine nằm trong khoảng từ 0 đến 1, với giá trị càng lớn biểu thị mức độ tương đồng càng cao.
Khi áp dụng cosine similarity trong bài toán phân loại thể loại bài hát, ta có thể tính toán độ tương đồng giữa tất cả các cặp bài hát trong tập dữ liệu. Điều này dẫn đến một ma trận tương tự có kích thước 1000 x 1000, trong đó mỗi giá trị trong ma trận đại diện cho độ tương đồng giữa hai bài hát.

Ví dụ: Ta có 1000 bài hát, ta sẽ tạo ra một ma trận vuông 1000 x 1000, với mỗi phần tử tại vị trí (i, j) là giá trị cosine similarity giữa bài hát i và bài hát j. Ma trận này sẽ có tính đối xứng, vì cosine similarity giữa bài hát A và bài hát B luôn bằng cosine similarity giữa bài hát B và bài hát A.

Kết quả

Hệ thống hiển thị danh sách các bài hát tương tự dựa trên độ tương đồng cosine. Ví dụ, bài hát pop.00019.wav (Britney Spears - Hit Me Baby One More Time) có các bài hát tương tự như sau:

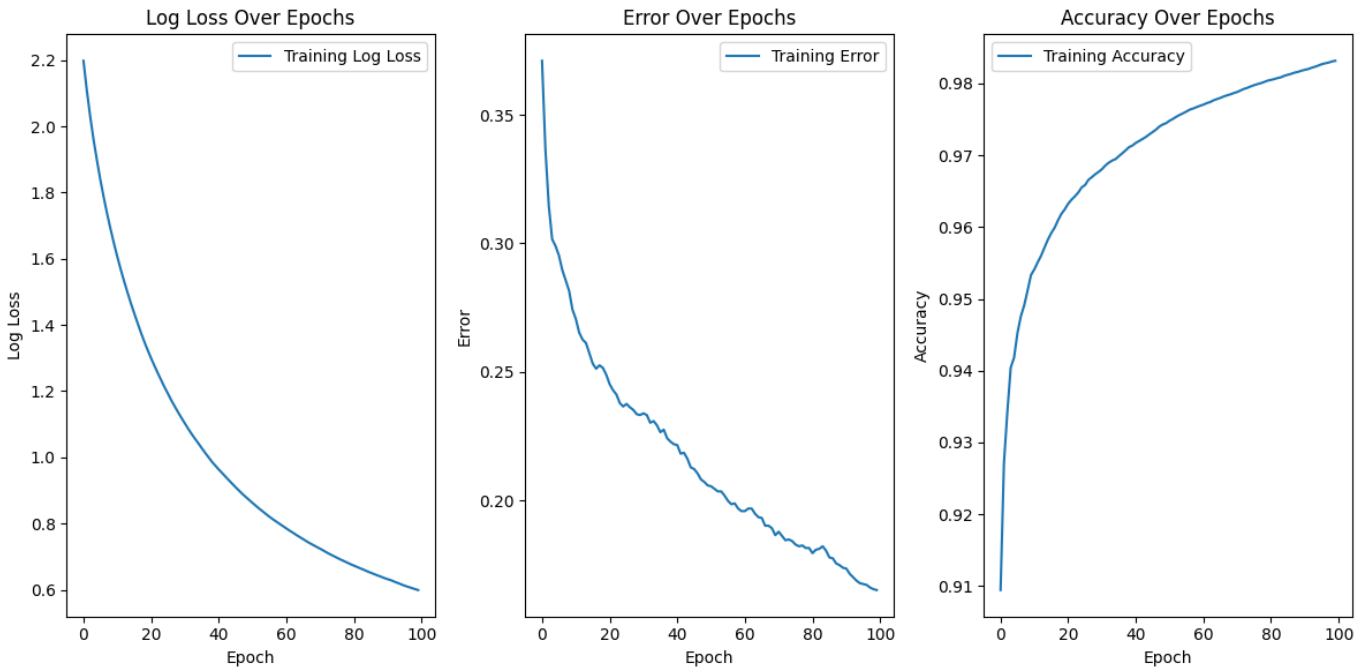
```
*******
Similar songs to  pop.00019.wav
filename
pop.00023.wav    0.862836
pop.00034.wav    0.860499
pop.00078.wav    0.829135
pop.00088.wav    0.824456
pop.00091.wav    0.802269
Name: pop.00019.wav, dtype: float64
```

Hệ thống đề xuất nhạc dựa trên độ tương đồng cosine là một phương pháp đơn giản nhưng hiệu quả trong việc phân tích và tìm kiếm bài hát tương tự. Hướng phát triển tiếp theo có thể bao gồm việc tích hợp các đặc trưng ngữ cảnh hoặc thông tin người dùng để nâng cao hiệu suất đề xuất.

VI. Tổng Kết

1. Mô hình XGBoost (Extreme Gradient Boosting):

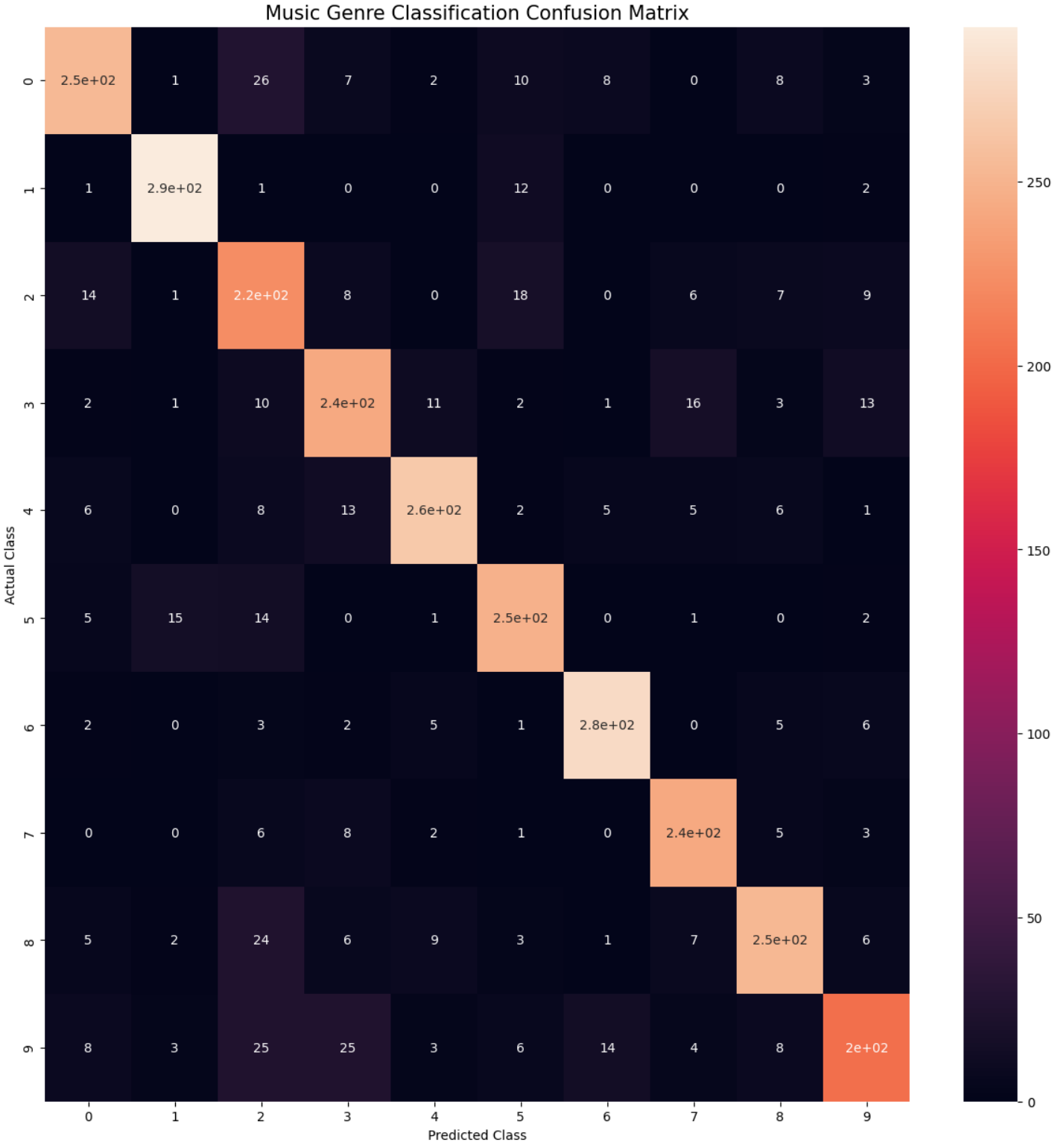
Mô hình XGBoost được huấn luyện với các tham số cơ bản như learning_rate và các chỉ số đánh giá mlogloss, merror, và auc.
Validation Accuracy: 0.8348348348348348
Biểu đồ dưới đây hiển thị các chỉ số qua các vòng lặp huấn luyện, giúp ta đánh giá hiệu năng của mô hình qua từng giai đoạn:



Dựa trên ba chỉ số đánh giá mlogloss, merror, và auc, ta có thể nhận xét như sau:

- Log Loss: giảm đều và không có sự dao động, chứng tỏ mô hình đang học tốt hơn theo thời gian và không có dấu hiệu overfitting hoặc underfitting.
- Error: giảm với xu hướng ổn định, chứng tỏ mô hình đang hội tụ và đang cải thiện dần độ chính xác.
- Accuracy: tăng đều đặn cho thấy mô hình đang học từ dữ liệu một cách hiệu quả. Tốc độ cải thiện giảm dần về cuối, chứng tỏ mô hình đã gần đạt trạng thái hội tụ.

Confusion Matrix



- Hiệu suất tổng thể:
  - Các giá trị trên đường chéo (từ 2.5e+02 đến 2.0e+02) cho thấy số lượng phân loại đúng tương đối cao.

  - Mỗi lớp đều có giá trị trên đường chéo mạnh, thể hiện hiệu suất chung tốt.

- Các mẫu gây nhầm lẫn đáng chú ý:
  - Lớp 2 có sự nhầm lẫn đáng kể với các lớp khác, đặc biệt là 26 lần nhầm lẫn với Lớp 0.
  - Lớp 3 và Lớp 4 có mức độ nhầm lẫn vừa phải với nhau (13 và 11 lần).
  - Lớp 9 có sự nhầm lẫn đáng chú ý với Lớp 2 và Lớp 3 (25 lần mỗi lớp).
  - Lớp 8 có một số nhầm lẫn với Lớp 2 (24 lần).
- Các lớp có hiệu suất tốt nhất:
  - Lớp 6 có rất ít trường hợp bị nhầm lẫn với các lớp khác.
  - Lớp 1 cho thấy hiệu suất mạnh với mức nhầm lẫn tối thiểu.
  - Lớp 7 cũng có khả năng phân biệt tốt với các lớp khác.
- Các lớp gặp nhiều thách thức nhất:
  - Lớp 2 dường như gặp vấn đề lớn nhất, với sự nhầm lẫn đáng kể với nhiều lớp khác.
  - Lớp 9 có các mẫu nhầm lẫn lớn, cho thấy lớp này có thể chia sẻ nhiều đặc điểm với các lớp khác.
  - Lớp 0 có mức độ nhầm lẫn vừa phải lan rộng ra nhiều lớp.
- Khuyến nghị:
  - Tập trung cải thiện khả năng phân biệt giữa Lớp 2 và Lớp 0.
  - Nghiên cứu các đặc điểm gây nhầm lẫn giữa Lớp 3 và Lớp 4.
  - Xem xét lý do tại sao Lớp 9 thường xuyên bị nhầm lẫn với Lớp 2 và Lớp 3.

2. Mạng nơ-ron tích chập (Convolutional Neural Network)

Huấn luyện và đánh giá mô hình

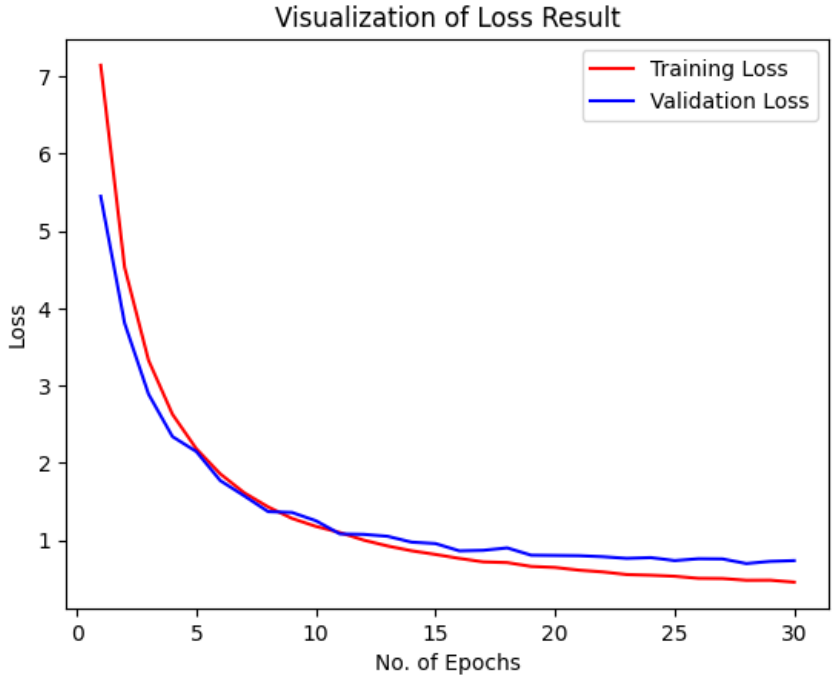
Mô hình được huấn luyện trên tập huấn luyện và đánh giá hiệu năng trên tập kiểm tra. (val_accuracy: 0.8828 - val_loss: 0.6790)

Trực quan hóa kết quả huấn luyện

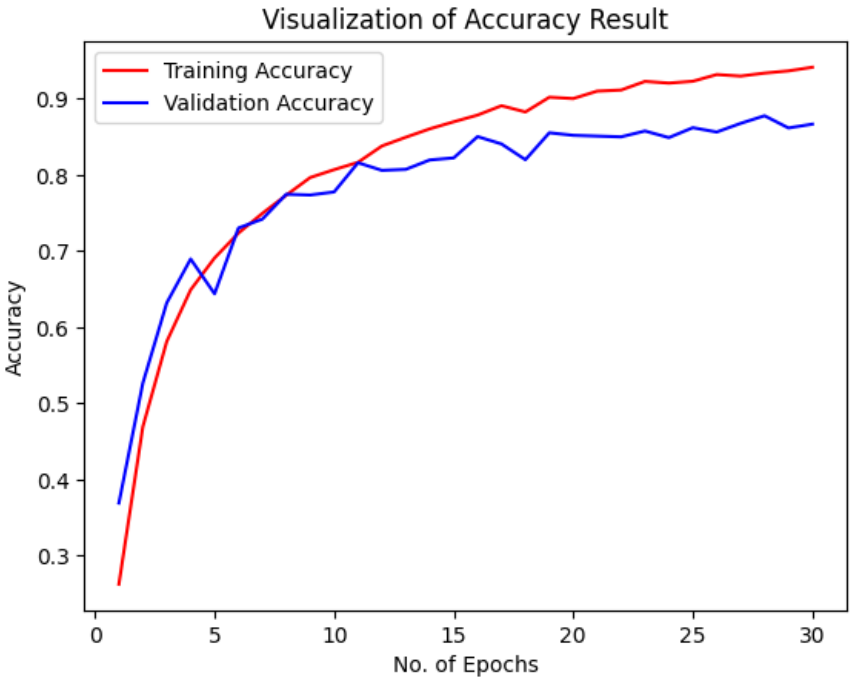
Kết quả huấn luyện được minh họa thông qua các biểu đồ, bao gồm độ mất mát

(Loss) và độ chính xác (Accuracy).

Loss



Accuracy



Nhận Xét:

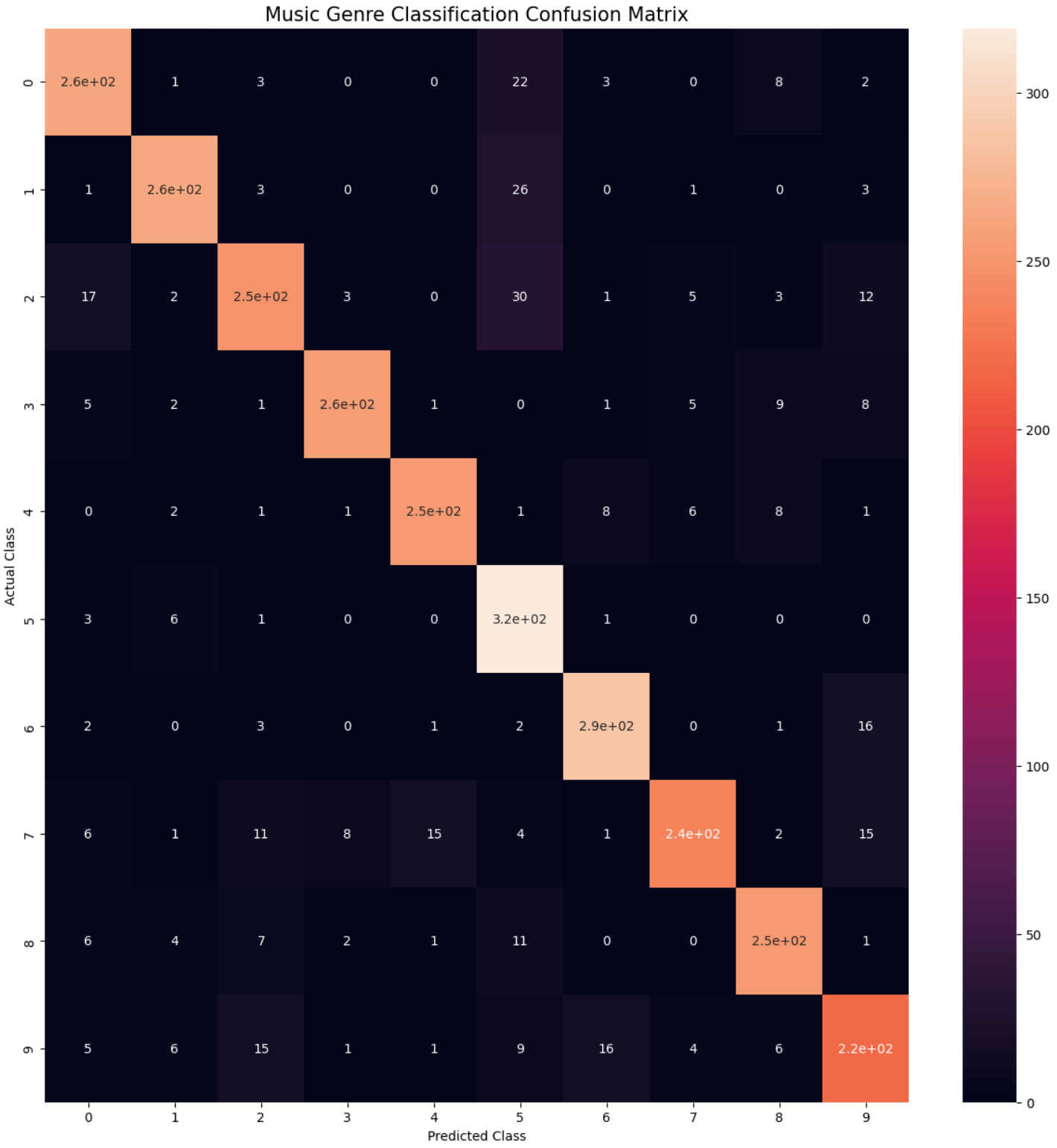
Loss:

- Loss giảm đều trên cả hai tập, cho thấy mô hình học tốt và đang hội tụ.
- Sự chênh lệch giữa training loss và validation loss không quá lớn, chứng tỏ mô hình không bị overfitting nghiêm trọng.
- Sau khoảng 15-20 epoch, validation loss dao động nhẹ trong khi training loss tiếp tục giảm, điều này có thể là dấu hiệu sớm cho overfitting.

Accuracy:

- Mô hình học tốt vì cả hai độ chính xác đều tăng ổn định qua thời gian.
- Sự chênh lệch giữa training accuracy và validation accuracy không quá lớn cho thấy mô hình tổng quát tốt trên tập dữ liệu chưa được nhìn thấy.
- Validation accuracy có dao động ở các epoch sau, trong khi training accuracy tiếp tục tăng. Điều này có thể là dấu hiệu mô hình bị overfitting nhẹ.

Confusion matrix



- Hiệu suất tổng thể:
  - Các giá trị đường chéo (từ 2.3e+02 đến 3.0e+02) thể hiện số lượng phân loại đúng rất cao.

  - Tất cả các lớp đều có số lượng phân loại đúng chiếm ưu thế, phản ánh hiệu suất tổng thể tốt.

- Các mẫu gây nhầm lẫn đáng chú ý:
  - Lớp 2 có nhầm lẫn với Lớp 8 (36 lần) và Lớp 0 (12 lần).
  - Lớp 4 có sự nhầm lẫn với Lớp 5 (30 lần).
  - Lớp 9 có nhầm lẫn đáng kể với Lớp 8 (17 lần) và Lớp 2 (5 lần).
- Các lớp có hiệu suất tốt nhất:
  - Lớp 1 có rất ít nhầm lẫn với các lớp khác.
  - Lớp 6 cũng cho thấy khả năng phân loại tốt, ít nhầm lẫn.
- Các lớp gặp nhiều thách thức nhất:
  - Lớp 2 có sự nhầm lẫn lan rộng với nhiều lớp, đặc biệt là Lớp 8 và Lớp 0.
  - Lớp 4 gặp khó khăn khi phân biệt với Lớp 5.
  - Lớp 9 có một số nhầm lẫn với Lớp 8 và Lớp 2.
- Khuyến nghị:
  - Cần cải thiện khả năng phân biệt giữa Lớp 2 với Lớp 8 và Lớp 0.
  - Phân tích nguyên nhân dẫn đến nhầm lẫn giữa Lớp 4 và Lớp 5 để tăng cường độ chính xác.
  - Nghiên cứu kỹ hơn về đặc điểm của Lớp 9 để giảm nhầm lẫn với các lớp khác.

3. Kết luận

Hoàn thiện quy trình xử lý dữ liệu: Tiền xử lý dữ liệu: Dữ liệu âm thanh từ bộ dữ liệu GTZAN được xử lý kỹ lưỡng, bao gồm chuẩn hóa, chuyển đổi tín hiệu âm thanh thành hình ảnh phổ (spectrogram), cũng như chuẩn hoá các dạng đặc trưng khác dành cho mô hình XGBoost và chia tập dữ liệu thành các phần huấn luyện, kiểm tra, và đánh giá.

Phân tích khám phá dữ liệu (EDA):

- Thực hiện thống kê mô tả trên các đặc trưng âm thanh, như phổ tần số, nhịp điệu và mức độ biến thiên trong tín hiệu, để tìm hiểu sự khác biệt đặc trưng giữa

các thể loại nhạc.

- Phân tích mối tương quan giữa các đặc trưng như độ cao, năng lượng, và nhịp điệu, xác định các đặc trưng quan trọng trong phân loại thể loại âm nhạc.

Xây dựng và đánh giá mô hình:

- Triển khai và so sánh hai phương pháp chính: mô hình CNN làm việc trên hình ảnh phổ và XGBoost sử dụng các đặc trưng trích xuất từ dữ liệu âm thanh.
- CNN đạt độ chính xác 88.3thể loại nhạc có cấu trúc phổ đặc trưng như Classical và Rock.
- XGBoost cho hiệu suất khá tốt với độ chính xác 83.5trong việc khai thác mối quan hệ giữa các đặc trưng âm thanh cụ thể như năng lượng, nhịp điệu, và tần số.

Phát hiện đặc trưng quan trọng: Các đặc trưng như năng lượng trung bình, độ biến thiên nhịp điệu, và số lượng nhịp trong mỗi đoạn nhạc được xác định là những yếu tố quan trọng giúp cải thiện hiệu quả phân loại.

Dự án đã thành công trong việc phát triển một quy trình toàn diện, từ tiền xử lý dữ liệu, phân tích đặc trưng đến xây dựng và đánh giá mô hình phân loại thể loại âm nhạc. Mô hình CNN đã chứng minh hiệu quả vượt trội với độ chính xác 88.3cho thấy tiềm năng lớn khi xử lý dữ liệu dạng phổ.
Kết quả nghiên cứu không chỉ đóng góp vào lĩnh vực phân tích âm nhạc mà còn mở ra tiềm năng ứng dụng trong các hệ thống gợi ý nhạc, nhận dạng nhạc cụ, và phân tích cảm xúc qua âm nhạc.

VII. Thuận lợi và khó khăn

1. Khó khăn

Việc xây dựng mô hình phân loại âm thanh đối mặt với nhiều khó khăn và thách thức. Một trong những vấn đề lớn là sự đa dạng và phức tạp của dữ liệu âm thanh. Các âm thanh trong thực tế có thể bị ảnh hưởng bởi nhiều yếu tố như nhiễu nền, sự thay đổi về môi trường, độ lớn âm thanh, và chất lượng thu âm, điều này khiến cho việc phân biệt các loại âm thanh trở nên khó khăn hơn. Thêm vào đó, âm thanh

có tính phi tuyến và liên tục, không như hình ảnh hay văn bản, khiến cho việc trích xuất đặc trưng từ âm thanh đòi hỏi các kỹ thuật phức tạp hơn như biến đổi Fourier, Mel-frequency cepstral coefficients (MFCC), hoặc deep learning.

Một thách thức đặc biệt khi sử dụng bộ dữ liệu GTZAN – một trong những bộ dữ liệu phổ biến trong nghiên cứu phân loại âm thanh – là sự hạn chế về tính đa dạng và sự phân phối không đồng đều của các lớp âm thanh trong tập dữ liệu này. GTZAN chỉ bao gồm 1.000 đoạn âm thanh, mỗi đoạn dài 30 giây, chia thành 10 thể loại nhạc khác nhau. Tuy nhiên, các thể loại này có thể không đại diện đầy đủ cho toàn bộ sự đa dạng của âm nhạc thực tế, và độ dài ngắn của từng đoạn âm thanh có thể không phản ánh chính xác đặc trưng của các thể loại âm nhạc phức tạp hơn. Ngoài ra, các đoạn âm thanh trong GTZAN cũng có thể gặp phải vấn đề về chất lượng, chẳng hạn như sự trùng lặp trong dữ liệu hoặc nhiễu âm thanh, làm giảm độ chính xác của mô hình khi áp dụng cho các tác vụ thực tế. Những hạn chế này yêu cầu các kỹ thuật tiền xử lý và cải thiện mô hình tinh vi hơn để đạt được kết quả tốt hơn.

Đặc biệt, một vấn đề nổi bật khi sử dụng bộ dữ liệu GTZAN là sự xuất hiện của lỗi trong một số file âm thanh, chẳng hạn như file jazz.00054.wav. File này bị lỗi và không thể phát lại bình thường, điều này gây khó khăn trong quá trình xử lý dữ liệu và huấn luyện mô hình. Lỗi này có thể ảnh hưởng đến việc trích xuất đặc trưng âm thanh, gây gián đoạn trong quá trình phân loại, hoặc thậm chí làm giảm chất lượng mô hình cuối cùng nếu không được xử lý đúng cách. Khi gặp phải những vấn đề như vậy, người dùng cần phải xử lý các file bị lỗi, hoặc loại bỏ chúng khỏi bộ dữ liệu, điều này có thể làm giảm số lượng mẫu huấn luyện và ảnh hưởng đến tính đại diện của dữ liệu.

2. Thuận lợi

Mặc dù tồn tại một số hạn chế, việc sử dụng bộ dữ liệu GTZAN cũng mang lại nhiều thuận lợi quan trọng trong nghiên cứu phân loại âm thanh.

Thứ nhất, GTZAN là một trong những bộ dữ liệu phổ biến nhất trong lĩnh vực nghiên cứu phân loại âm thanh và nhạc, do đó, nó đã được sử dụng rộng rãi trong

nhiều nghiên cứu trước đây. Điều này mang lại lợi thế lớn cho các nhà nghiên cứu khi so sánh hiệu quả của mô hình mới với các mô hình đã có thông qua các kết quả benchmark. Việc sử dụng một bộ dữ liệu chuẩn cũng giúp tăng tính minh bạch và khả năng tái hiện của các nghiên cứu.

Thứ hai, GTZAN có cấu trúc dữ liệu rõ ràng với 1.000 đoạn âm thanh được chia đều thành 10 thể loại nhạc khác nhau. Điều này tạo điều kiện thuận lợi cho việc xây dựng và kiểm thử các mô hình phân loại âm thanh, đặc biệt là trong giai đoạn phát triển ban đầu khi nhà nghiên cứu cần một tập dữ liệu có quy mô vừa phải để nhanh chóng triển khai và đánh giá các thuật toán.

Thứ ba, các đoạn âm thanh trong GTZAN đều có độ dài cố định (30 giây), giúp giảm bớt công đoạn tiền xử lý dữ liệu, chẳng hạn như việc chuẩn hóa độ dài tín hiệu âm thanh. Điều này đặc biệt hữu ích khi áp dụng các phương pháp trích xuất đặc trưng như MFCC hoặc spectrogram, vốn đòi hỏi đầu vào có định dạng đồng nhất.

Cuối cùng, GTZAN cung cấp một môi trường thử nghiệm lý tưởng để kiểm chứng hiệu quả của các phương pháp học máy và deep learning. Dữ liệu có tính đa dạng vừa phải, bao gồm nhiều thể loại nhạc từ cổ điển đến hiện đại, cho phép các nhà nghiên cứu đánh giá khả năng tổng quát hóa của mô hình trên các loại dữ liệu khác nhau. Ngoài ra, do sự phổ biến của GTZAN, cộng đồng nghiên cứu đã phát triển nhiều công cụ và kỹ thuật xử lý tối ưu cho bộ dữ liệu này, giúp tiết kiệm thời gian và công sức trong việc xây dựng các mô hình.

Nhờ những thuận lợi trên, GTZAN vẫn là một lựa chọn hấp dẫn cho các nghiên cứu về phân loại âm thanh, đặc biệt trong các bài toán thí nghiệm và phát triển thuật toán mới.