Nome do projeto: Análise e Segmentação de Dataset Musical (1950 - 2019)

Nomes dos integrantes do grupo: André de Medeiros, Cristovão Cronje, Murilo da Costa, Pablo Philipe França

Data de entrega: 16/07/2025

Resumo

O presente projeto tem como objetivo explorar e segmentar um acervo de músicas com base em suas características musicais e temas líricos, extraídos de um banco de dados que abrange o período de 1950 a 2019. Por meio de técnicas de agrupamento não supervisionado, a análise buscou identificar padrões ocultos e criar perfis musicais distintos. Os principais métodos utilizados incluíram a Análise de Componentes Principais (PCA) para redução de dimensionalidade , clusterização com o algoritmo K-Means (testando k=3 e k=8) , e Clusterização Hierárquica para análise da estrutura dos grupos. Para visualização e validação dos agrupamentos, foram empregadas técnicas como t-SNE e UMAP. Os resultados revelaram segmentações claras, onde uma divisão em 3 clusters agrupa as músicas em categorias amplas (ex: Energéticas, Acústicas, Focadas em Letra), enquanto uma divisão em 8 clusters oferece uma granularidade maior, distinguindo subgrupos baseados em temas líricos e nuances sonoras específicas.

Introdução

Contextualização do Problema

Este projeto aborda o desafio de organizar e compreender grandes volumes de dados musicais. A análise visa explorar um conjunto de músicas, segmentando-as com base em uma variedade de características, como dançabilidade, energia e temas líricos. O objetivo é ir além da catalogação tradicional por gênero, utilizando aprendizado não supervisionado para descobrir "perfis musicais" inerentes aos próprios dados.

Justificativa

A escolha do problema justifica-se pela crescente necessidade de sistemas de recomendação mais inteligentes e organização de bibliotecas musicais de forma personalizada. A identificação de padrões e perfis musicais permite criar playlists customizadas e entender as preferências dos ouvintes de maneira mais profunda. A base de dados foi selecionada por sua riqueza em atributos quantitativos, tanto musicais quanto temáticos, permitindo uma análise multifacetada.

Objetivos Específicos

O projeto foi guiado pelas seguintes questões principais:

- Quantos clusters representam adequadamente a diversidade musical presente na base de dados?
- Quais são as características médias e distintivas de cada grupo musical identificado?
- É possível interpretar cada cluster em termos de estilos musicais, temas líricos ou tipos de música?
- Como as segmentações com 3 e 8 clusters se comparam em termos de detalhamento e relevância prática?

Descrição da Base de Dados

Origem e Características Gerais

O dataset utilizado foi importado de um arquivo CSV através da biblioteca pandas. Ele contém uma coleção de músicas com diversas características numéricas que representam tanto atributos musicais (ex: danceability, energy, acousticness) quanto temas líricos (ex: romantic, violence). Embora o número exato de registros e variáveis não seja especificado no extrato, a análise focou exclusivamente nas colunas numéricas relevantes para a clusterização.

Qualidade dos Dados

Durante a fase de preparação, foram identificados valores ausentes ou inválidos nas colunas selecionadas. Além disso, a análise exploratória apontou para a existência de outliers que poderiam distorcer a formação dos clusters, especialmente em algoritmos sensíveis à distância como o K-Means.

Discussão sobre os Atributos Mais Relevantes

Os atributos selecionados para a análise foram as variáveis numéricas que descrevem características musicais e temas líricos quantitativos. As análises subsequentes mostraram que variáveis como

energy, valence, danceability e acousticness são cruciais para diferenciar os clusters. Da mesma forma, temas líricos como

obscene e family/gospel provaram ser determinantes para a formação de grupos específicos, independentemente das características de áudio.

Metodologia

Etapas da Análise

O projeto seguiu as seguintes etapas:

1. **Preparação dos Dados:** Importação do dataset, seleção de variáveis numéricas relevantes, tratamento de dados ausentes e padronização das escalas.
2. **Análise Exploratória de Dados (EDA):** Visualização da distribuição das variáveis com histogramas e boxplots para identificar assimetrias e outliers. Análise de correlações com um mapa de calor para entender as relações lineares entre as variáveis.
3. **Redução de Dimensionalidade:** Aplicação de PCA para reduzir a complexidade e visualizar as principais fontes de variância. Uso de t-SNE e UMAP para visualizações não lineares que preservam a estrutura local dos dados.
4. **Seleção do Número de Clusters:** Utilização do Método do Cotovelo (Elbow Method) e do Coeficiente de Silhueta (Silhouette Score) para determinar o número ideal de clusters para o K-Means. Análise de um dendrograma de Clusterização Hierárquica para identificar pontos de corte naturais.
5. **Clusterização e Análise:** Aplicação do algoritmo K-Means com k=3 e k=8, com e sem o uso de PCA, para comparar os resultados. Interpretação e caracterização de cada cluster formado.

Tratamento de Dados

- **Dados Ausentes:** Valores ausentes foram substituídos pela média da respectiva coluna, garantindo que os algoritmos de clusterização pudessem ser executados.
- **Padronização:** As variáveis foram padronizadas com StandardScaler para terem média zero e desvio padrão um, garantindo que todas as características tivessem peso equilibrado na análise.
- **Outliers:** Os outliers foram identificados visualmente com boxplots e quantificados pelo método do Intervalo Interquartil (IQR).

Ferramentas e Bibliotecas

As seguintes ferramentas foram utilizadas:

- **pandas:** Para importação e manipulação dos dados.
- **matplotlib:** Para a criação de visualizações estáticas 2D e 3D, como gráficos de PCA, hexbins e boxplots.
- **scikit-learn:** Para padronização (StandardScaler), PCA, K-Means, t-SNE e UMAP.
- **Plotly:** Para a criação de gráficos 3D interativos.

Resultados da EDA

Análise Univariada

Após a padronização, foram gerados histogramas para visualizar a distribuição de cada variável, permitindo detectar assimetrias e avaliar a variância. Adicionalmente, boxplots foram utilizados para identificar visualmente os outliers em cada variável , e o método IQR foi aplicado para quantificá-los, destacando as variáveis que mereciam atenção especial.

Análise Bivariada e PCA

Um mapa de calor da matriz de correlação foi construído para explorar as relações lineares entre as variáveis. Essa análise ajudou a identificar redundâncias e grupos de variáveis correlacionadas, o que é útil para a aplicação do PCA. A Análise de Componentes Principais (PCA) revelou que os primeiros 15 componentes explicam aproximadamente 90% da variância dos dados, e os primeiros 17 componentes explicam cerca de 95%. Isso indicou que um número entre 15 e 17 componentes seria uma escolha eficiente para reduzir a dimensionalidade, mantendo a maior parte da informação original.

Discussão dos Insights

Interpretação dos Resultados

Os resultados demonstram que o conjunto de dados pode ser segmentado de forma significativa em diferentes perfis musicais. A análise comparativa entre k=3 e k=8 foi particularmente reveladora.

- **Com k=3**, a segmentação é ampla, resultando em três grupos principais: **Cluster 0** (músicas de alta energia, dançantes e positivas), **Cluster 1** (músicas calmas, acústicas e introspectivas) e **Cluster 2** (músicas definidas por temas líricos específicos, como gospel ou conteúdo explícito).
- **Com k=8**, a segmentação torna-se mais granular. Os clusters mais amplos de k=3 são subdivididos em perfis mais refinados, como "Românticas e Melancólicas", "Conteúdo Explícito/Violento", "Acústicas e Instrumentais" e "Temas de Família/Espiritualidade". Isso mostra que, ao permitir mais grupos, os temas líricos se tornam um fator de separação mais forte.

As diferentes técnicas de visualização (PCA, t-SNE, UMAP) foram cruciais para validar e interpretar os agrupamentos, cada uma oferecendo uma perspectiva complementar sobre a estrutura dos dados.

Aplicação em Contexto Real

Esses insights podem ser aplicados diretamente na criação de sistemas de recomendação musical e na organização de grandes catálogos. Por exemplo, um serviço de streaming poderia usar os perfis de cluster para gerar playlists personalizadas ("Músicas acústicas para relaxar", "Pop energético para treinar") ou para sugerir novas músicas a um usuário com base nas características do cluster que ele mais ouve.

Limitações da Análise

A análise possui algumas limitações. Primeiramente, ela é exploratória e seus resultados dependem das variáveis disponíveis no dataset. A inclusão de outros dados, como informações sobre o artista, o ano de lançamento ou o país de origem, poderia levar a segmentações diferentes e potencialmente mais ricas. Além disso, a determinação do número ideal de clusters (

K) sempre envolve um grau de subjetividade, embora métodos quantitativos como o coeficiente de silhueta tenham sido usados para guiar a decisão.

Conclusão

Resumo das Descobertas

Este trabalho demonstrou com sucesso a aplicação de técnicas de clusterização não supervisionada para identificar perfis musicais significativos a partir de características de áudio e temas líricos. A combinação de métodos como K-Means, Clusterização Hierárquica e visualizações avançadas (PCA, t-SNE, UMAP) permitiu uma análise robusta e confiável dos agrupamentos. A principal descoberta é a capacidade de segmentar o acervo musical em diferentes níveis de granularidade (k=3 vs. k=8), revelando desde categorias musicais amplas até nichos definidos por temas líricos específicos.

Com K=3 Clusters:

Cluster 0: Músicas de alta energia, dançantes e com sonoridade positiva (Pop, Eletrônica, Dance).

Cluster 1: Músicas mais calmas, acústicas e introspectivas (Baladas, Folk, Instrumental).

Cluster 2: Músicas definidas principalmente por temas líricos específicos (Ex: Gospel, Obscenidade), independentemente das características de áudio gerais.

Com K=8 Clusters:

Cluster 0: Energéticas e Dançantes (Pop, Eletrônica).

Cluster 1: Românticas e Melancólicas (Baladas, Soft Pop).

Cluster 2: Foco em Performance/Audiência (Hinos, Músicas ao Vivo).

Cluster 3: Conteúdo Explícito/Violento (Hip Hop, Rap, Metal).

Cluster 4: Foco em Comunicação/Interação (R&B, Pop lírico).

Cluster 5: Acústicas e Instrumentais (Clássica, Jazz, Ambiente).

Cluster 6: Temas de Família/Espiritualidade (Gospel, Religiosa).

Cluster 7: Foco em "Like/Girls" (Pop, R&B com temas de atração).

A escolha entre K=3 e K=8 (ou outro K) dependerá do nível de granularidade desejado para a análise ou aplicação subsequente.

Propostas de Estudos Futuros

Para trabalhos futuros, sugere-se:

1. **Aprofundar a Validação de K:** Empregar outras métricas quantitativas, como a estatística *gap* (gap statistic), para fornecer uma validação adicional na escolha do número ótimo de clusters.
2. **Enriquecimento do Dataset:** Incorporar novas *features* à análise, como metadados dos artistas (gênero, popularidade), ano de lançamento das músicas e informações contextuais, o que poderia revelar dimensões de similaridade ainda não exploradas.
3. **Modelagem Preditiva:** Utilizar os clusters gerados como variável alvo em um modelo de classificação para prever o perfil de novas músicas com base em suas características.

Referências

- **Fonte de Dados:** dataset: Music Dataset : 1950 to 2019
- **Bibliotecas e Ferramentas:**
    - Pandas
    - Matplotlib
    - Scikit-learn (utilizada para StandardScaler, PCA, KMeans, t-SNE, UMAP)
    - Plotly