Faculdade de Computação e Informática

Título - Análise de Fatores que Influenciam a Popularidade das Músicas no Spotify: Uma Abordagem Baseada em Características de Áudio, Gêneros e Artistas

Isabella Abreu Comelli[1], Pedro Henrique Wege Barbosa[1]

[1]Programa de graduação em Sistemas de Informação – Faculdade de Computação e Informática (FCI) – Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

{isabella.comelli, pedrohenriquewege.barbosa}@mackenzista.com.br

Resumo. *Este projeto tem como objetivo identificar os fatores que influenciam a popularidade das músicas no Spotify, com foco nas características de áudio presentes no banco de dados da plataforma. Além disso, busca-se explorar a influência de gêneros musicais e artistas, complementando a análise sonora. Para o estudo, foram selecionadas as 100 músicas mais populares do Spotify, permitindo uma investigação detalhada sobre os elementos que contribuem para seu sucesso. Os módulos pandas e scikit-learn, disponíveis na linguagem de programação Python, serão utilizados para a manipulação dos dados e o treinamento de modelos de aprendizado de máquina, analisando atributos como energia, valência e tempo. A eficácia do modelo será avaliada com base na precisão das associações identificadas entre as características sonoras e a popularidade das faixas. A pesquisa busca contribuir para a compreensão dos padrões que tornam certas músicas mais populares.*

1. Introdução

A popularidade de uma música no Spotify é influenciada por uma série de fatores, desde aspectos subjetivos, como preferências do público e tendências culturais, até características mais técnicas, como sua estrutura sonora. Com o avanço das tecnologias de análise de dados, tornou-se possível identificar padrões que contribuem para o sucesso de determinadas faixas na plataforma. A compreensão desses padrões pode beneficiar tanto artistas e produtores, que buscam otimizar suas produções para alcançar um público maior, quanto o próprio Spotify, que pode aprimorar suas estratégias de recomendação e marketing.

O mercado musical está em constante transformação, e entender os fatores que impulsionam o sucesso de uma música pode ser um diferencial para artistas, gravadoras e plataformas de streaming. Identificar características acústicas comuns em músicas populares pode fornecer insights valiosos sobre tendências sonoras e preferências do público. Além disso, a análise da influência de gêneros musicais e da notoriedade dos artistas na popularidade das músicas pode auxiliar na formulação de estratégias de lançamento e promoção. Ao utilizar dados públicos do Spotify e técnicas de aprendizado de máquina, este estudo busca contribuir para uma melhor compreensão do que torna uma música popular.

Este projeto tem como objetivo principal identificar os fatores que influenciam a popularidade das músicas no Spotify, com ênfase nas características de áudio extraídas diretamente da plataforma. Além disso, busca-se explorar a relação entre gêneros musicais e a popularidade, bem como o impacto do reconhecimento prévio dos artistas no sucesso de suas faixas. Para isso, foram selecionadas as 100 músicas mais populares do Spotify, permitindo uma investigação detalhada dos elementos que contribuem para seu sucesso.

O estudo será conduzido por meio da coleta e análise de dados extraídos da API pública do Spotify, organizados em dois conjuntos principais: um contendo informações gerais das músicas, como nome, artista e popularidade, e outro focado nas características de áudio, incluindo dançabilidade, energia e tonalidade. As técnicas de aprendizado de máquina serão aplicadas para identificar padrões entre esses atributos e o sucesso das músicas. O projeto será implementado na linguagem Python, utilizando bibliotecas como pandas, numpy, matplotlib e scikit-learn para manipulação de dados, visualização e modelagem preditiva.

2. Descrição do Problema

A popularidade das músicas no Spotify é um fenômeno complexo, influenciado por diversos fatores, desde a exposição midiática até as características sonoras das faixas. Com o crescimento das plataformas de streaming, compreender o que torna uma música popular tornou-se uma questão estratégica para artistas, gravadoras e para o próprio Spotify. No entanto, a definição de popularidade muitas vezes é subjetiva e depende de uma combinação de fatores que vão além do simples número de reproduções.

As características de áudio, como energia, dançabilidade e valência, podem desempenhar um papel significativo na recepção do público. Entretanto, a relação entre esses atributos e o sucesso de uma música ainda não é totalmente compreendida. Da mesma forma, a influência dos gêneros musicais e da notoriedade dos artistas na popularidade das faixas é um aspecto relevante que carece de uma análise aprofundada.

Este estudo busca responder a questões fundamentais: Quais características sonoras estão mais associadas à popularidade das músicas? Existe um padrão comum entre as faixas que alcançam maior sucesso? A notoriedade do artista impacta diretamente a popularidade da música, independentemente de suas características sonoras? Para responder a essas perguntas, será utilizada a API do Spotify, permitindo a extração e análise de dados de músicas populares na plataforma.

A utilização de APIs como ferramenta de extração de dados tem se tornado uma prática comum em diversas áreas, incluindo a análise musical. Segundo documentos da Google Developers (2024) e Camelo (2023), as APIs possibilitam acesso a grandes volumes de informações, permitindo automação e processamento eficiente dos dados coletados. No contexto deste projeto, a API do Spotify será fundamental para coletar dados sobre as músicas e suas características sonoras, viabilizando uma abordagem analítica baseada em evidências.

Dessa forma, este estudo visa explorar a relação entre atributos sonoros e popularidade no Spotify, utilizando aprendizado de máquina e análise estatística para identificar padrões e tendências que possam contribuir para uma melhor compreensão do sucesso musical na era do streaming.

3. Ética e Responsabilidade

O uso de inteligência artificial para análise de dados e tomada de decisões tem se tornado cada vez mais comum, trazendo benefícios significativos para diversas áreas, incluindo o setor musical. No entanto, a aplicação dessas tecnologias levanta questões éticas importantes, especialmente no que diz respeito à transparência, viés algorítmico e impacto na indústria da música.

Um dos principais desafios éticos na implementação de soluções baseadas em IA é garantir que os modelos utilizados sejam justos e imparciais. Algoritmos treinados com dados históricos podem perpetuar padrões existentes, favorecendo determinados gêneros musicais, artistas ou características sonoras em detrimento de outros. Isso pode reforçar desigualdades no mercado, prejudicando músicos independentes ou gêneros menos populares. Para mitigar esse risco, é fundamental que os dados utilizados no estudo sejam representativos e que haja um monitoramento contínuo dos resultados gerados pela IA.

Além disso, a transparência no desenvolvimento da solução é um aspecto essencial. Usuários e profissionais da indústria musical devem compreender os critérios utilizados pelos modelos para determinar padrões de popularidade. Soluções de IA que influenciam a recomendação de músicas, por exemplo, podem impactar diretamente a visibilidade de artistas, tornando necessário um compromisso ético na construção dos algoritmos. O uso de APIs e IA deve ser pautado por boas práticas, garantindo que as

decisões automatizadas sejam explicáveis e auditáveis (GOOGLE DEVELOPERS, 2024; CAMELO, 2023).

Outro ponto relevante é a responsabilidade no uso dos dados. Como o projeto envolve a coleta de informações por meio da API do Spotify, é essencial garantir a conformidade com políticas de privacidade e uso responsável dos dados. A anonimização e o respeito às diretrizes da plataforma são medidas necessárias para evitar o uso indevido das informações.

Por fim, a implementação da IA na análise de popularidade musical deve ser feita com o compromisso de agregar valor à indústria da música, sem substituir a criatividade humana. O objetivo não é criar um modelo que dite o que deve ser produzido, mas sim fornecer insights que possam auxiliar artistas, produtores e plataformas de streaming na compreensão das tendências musicais de forma ética e responsável.

4. Dataset

Os dados utilizados neste estudo foram coletados por meio da API pública do Spotify, sendo organizados em dois dataframes principais: df_tracks e df_audio_features. O primeiro contém informações gerais sobre as músicas, como título, popularidade, gênero musical, duração, nome do artista e do álbum. Já o segundo abrange as características de áudio de cada faixa, incluindo atributos como dançabilidade, energia, tom e instrumentalidade.

A base de dados extraída da API apresentou uma qualidade satisfatória, com informações completas para a maioria das músicas. Foram identificadas poucas ocorrências de valores ausentes, principalmente em colunas relacionadas a álbuns, e algumas duplicatas que precisaram ser removidas. Como o foco da análise é a relação entre características de áudio e popularidade, os outliers foram considerados músicas populares que não apresentavam padrões acústicos similares às demais faixas de destaque.

Durante a etapa de pré-processamento, algumas transformações foram aplicadas para garantir a padronização e a qualidade dos dados. No dataframe df_tracks, a coluna track_explicit foi convertida de um formato booleano para inteiro, e os subgêneros musicais foram agrupados dentro de gêneros principais, como no caso de synth-pop, que foi classificado simplesmente como pop. Além disso, todas as linhas duplicadas foram removidas.

Após a junção dos dois dataframes, algumas colunas numéricas que continham valores ausentes foram preenchidas com zero, garantindo a consistência do dataset. Foi realizada a conversão de valores em notação científica na coluna

track_instrumentalness, transformando-os em valores de ponto flutuante (float). Também foi criada a coluna track_main_genre_id, que classifica as músicas de acordo com gêneros musicais padronizados. Algumas colunas foram descartadas por não agregarem valor à análise, como informações sobre álbuns e subgêneros, além de atributos de áudio como track_mode, track_energy, track_acousticness, track_liveness, track_valence, track_loudness e track_time_signature.

Para facilitar a análise dos dados, algumas colunas foram convertidas em variáveis categóricas. Atributos como track_danceability, track_speechiness e track_instrumentalness foram segmentados em duas categorias, sendo classificados com valores binários (0 ou 1) com base em seus percentis medianos. Além disso, a coluna track_popularity foi transformada em categorias de popularidade, agrupando os valores por faixas. Por exemplo, uma música com popularidade 85 foi classificada na "casa dos 80" e recebeu o valor 8, enquanto uma faixa com popularidade 74 foi categorizada na "casa dos 70" e recebeu o valor 7.

A preparação dos dados foi conduzida utilizando Python e as bibliotecas pandas, numpy e matplotlib, permitindo a manipulação eficiente das tabelas, a aplicação de transformações e a visualização inicial dos padrões nos dados. A análise exploratória incluiu estatísticas descritivas para identificar distribuições e possíveis correlações entre os atributos de áudio e a popularidade das músicas. Esse processo foi fundamental para garantir a qualidade dos dados utilizados no estudo e a confiabilidade dos resultados obtidos.

5. Metodologia

A metodologia deste trabalho foi estruturada com base em uma abordagem quantitativa, focando na análise estatística e na aplicação de algoritmos de aprendizado de máquina supervisionado. O objetivo foi investigar a influência de atributos sonoros e metadados de faixas musicais na variável de popularidade no Spotify, plataforma dominante no mercado global de streaming (IFPI, 2024).

5.1 Coleta de Dados

Os dados foram coletados por meio da API pública do Spotify, utilizando a biblioteca Spotipy na linguagem Python. A API permite o acesso a dados de músicas disponíveis na plataforma, incluindo métricas como popularidade, dançabilidade, energia, instrumentalidade, tempo, valência, entre outros (SPOTIFY FOR DEVELOPERS, 2024). Foram selecionadas as 100 faixas mais populares do momento, conforme ranking da própria plataforma, permitindo capturar as tendências sonoras e estruturais presentes nas músicas de maior sucesso.

As informações extraídas foram organizadas em dois conjuntos principais: o dataframe df_tracks, contendo dados gerais como nome da música, artista, popularidade e gênero; e o dataframe df_audio_features, com as características técnicas das faixas extraídas diretamente da API.

5.2 Preparação dos Dados

A etapa de preparação dos dados consistiu em diversas transformações e limpezas necessárias para garantir a qualidade da base. Inicialmente, foram removidos registros duplicados e tratados valores ausentes em colunas específicas, especialmente aquelas relacionadas a álbuns. A coluna track_explicit foi convertida de valor booleano para inteiro, enquanto gêneros similares ou subgêneros foram agrupados em categorias mais amplas (e.g., “synth-pop” foi classificado como “pop”).

Adicionalmente, variáveis contínuas como track_danceability, track_speechiness e track_instrumentalness foram transformadas em variáveis binárias com base na mediana dos dados, conforme metodologia adotada também por Camelo (2023). A variável track_popularity, por sua vez, foi discretizada em faixas de 10 pontos, o que permitiu categorizar músicas em diferentes níveis de sucesso e viabilizou o uso de classificadores discretos.

5.3 Modelagem Preditiva

Para a modelagem, foram empregados dois algoritmos supervisionados de classificação: K-Nearest Neighbors (KNN) e Decision Tree. Os modelos foram implementados utilizando a biblioteca scikit-learn (SCIKIT-LEARN, 2024), que fornece suporte robusto para tarefas de machine learning com fácil integração ao ecossistema de dados do Python.

O KNN foi escolhido por sua simplicidade e capacidade de prever a popularidade de uma música com base nas similaridades dos atributos sonoros com outras faixas. Já o modelo de Árvore de Decisão foi utilizado por sua interpretabilidade, permitindo visualizar os critérios decisórios que levam uma música a ser classificada como popular ou não. A acurácia dos modelos foi calculada por meio de validação cruzada, com os dados embaralhados em diversas execuções para testar a robustez das previsões.

5.4 Ferramentas Utilizadas

As seguintes ferramentas e bibliotecas foram utilizadas no desenvolvimento do projeto:

- Python: Linguagem principal de desenvolvimento, escolhida por sua sintaxe acessível e vasta gama de bibliotecas para ciência de dados e IA (GOOGLE

DEVELOPERS, 2024).

- Pandas: Usado para manipulação e limpeza dos dados, possibilitando a criação e transformação de dataframes para análise tabular eficiente (PANDAS, 2024).

- NumPy: Aplicado no suporte a operações matemáticas, manipulação de arrays e vetores numéricos de alta performance, funcionando como base para estruturas matriciais e cálculos estatísticos (HARRIS et al., 2020).

- Matplotlib: Empregado para geração de gráficos e visualizações dos dados e resultados dos modelos, como distribuição de atributos e acurácias por categoria (MATPLOTLIB, 2024).

- Scikit-learn: Biblioteca central para aplicação de algoritmos de aprendizado de máquina e divisão de dados em treino e teste, além de fornecer ferramentas para avaliação do desempenho dos modelos (SCIKIT-LEARN, 2024).

- Spotipy: Utilizada para autenticação e comunicação com a API do Spotify, permitindo a extração automatizada de dados musicais diretamente da plataforma (SPOTIFY FOR DEVELOPERS, 2024).

6. Resultados

Após a aplicação dos modelos preditivos, os resultados indicaram que variáveis como energy, danceability e valence (energia, dançabilidade e positividade, respectivamente) apresentaram forte correlação com altos níveis de popularidade. A predominância do gênero pop nas músicas populares também foi observada, corroborando o papel do apelo comercial e do gosto massivo do público.

Os modelos de aprendizado de máquina tiveram desempenhos satisfatórios, com a Árvore de Decisão apresentando maior interpretabilidade e o KNN alcançando melhor acurácia média. Os resultados mostraram taxas de acurácia entre 68% e 73%, com estabilidade mantida em torno de 70% mesmo em execuções aleatórias, evidenciando a consistência do conjunto de dados e a robustez dos algoritmos utilizados.

As representações gráficas produzidas com Matplotlib permitiram visualizar correlações entre atributos e popularidade, além de demonstrar a distribuição dos gêneros musicais no conjunto. Também foi possível identificar outliers (músicas populares com características acústicas destoantes) indicando que fatores não sonoros, como marketing ou reputação do artista, exercem influência importante.

7. Conclusão

Este projeto conseguiu alcançar seu principal objetivo: analisar e identificar os fatores sonoros que influenciam a popularidade das músicas no Spotify. A metodologia baseada em dados extraídos da API pública da plataforma, aliada a técnicas de aprendizado de máquina, permitiu a construção de modelos que preveem com razoável precisão os níveis de sucesso das faixas analisadas.

Os resultados confirmam a relevância de atributos como energia, dançabilidade e positividade, além da forte presença do gênero pop entre as faixas mais populares. Por outro lado, os modelos também mostraram limitações, especialmente no que diz respeito à previsão de sucessos baseados apenas em dados acústicos. Elementos externos, como tendências culturais, campanhas de divulgação e renome dos artistas, exercem papel decisivo e, muitas vezes, imprevisível.

Embora a inteligência artificial permita avanços significativos na análise musical, sua eficácia plena depende da integração com fatores sociais e contextuais. Como perspectiva futura, a análise poderia incluir variáveis textuais (letras das músicas), redes sociais e histórico de reprodução, além da aplicação de modelos mais sofisticados, como redes neurais ou análise de sentimentos. Tais aprimoramentos podem elevar a capacidade preditiva e apoiar de forma mais abrangente artistas, produtores e plataformas de streaming na tomada de decisões estratégicas.

8. Disponibilização e Apresentação do Projeto

A estrutura do projeto foi disponibilizada de forma completa no Github através do endereço https://github.com/pedrohwbarbosa/projeto_IA_mack, enquanto a apresentação está disponível no Youtube pelo endereço https://youtu.be/v0oH3Eu7iQ4.

9. Referências Bibliográficas

CAMELO, L. API do ChatGPT: como usar todo o poder da inteligência artificial? Pluga Blog, 2023. Disponível em: https://pluga.co/blog/api-chatgpt/. Acesso em: 10 fev. 2025.

GOOGLE DEVELOPERS. *Best practices for API design.* 2024. Disponível em: https://cloud.google.com/apis/design. Acesso em: 23 mai. 2025.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, v. 585, p. 357–362, 2020. DOI: 10.1038/s41586-020-2649-2. Disponível em: https://doi.org/10.1038/s41586-020-2649-2. Acesso em: 20 mai. 2025.

INTERNATIONAL FEDERATION OF THE PHONOGRAPHIC INDUSTRY – IFPI. Global Music Report 2024: State of the Industry. 2024. Disponível em: https://globalmusicreport.ifpi.org/. Acesso em: 20 mai. 2025.

MATPLOTLIB. Matplotlib 3.9.2 documentation. 2024. Disponível em: https://matplotlib.org/stable/index.html. Acesso em: 20 mai. 2025.

NUMPY. NumPy documentation. Disponível em: https://numpy.org/doc/stable/. Acesso em: 23 mai. 2025.

PANDAS. Pandas documentation. 2024. Disponível em: https://pandas.pydata.org/docs/index.html. Acesso em: 20 mai. 2025.

SCIKIT-LEARN. API Reference – Scikit-learn. 2024. Disponível em: https://scikit-learn.org/stable/api/index.html. Acesso em: 20 mai. 2025.

SPOTIFY FOR DEVELOPERS. Web API – Retrieve metadata from Spotify content, control playback or get recommendations. Sweden: Spotify, 2024. Disponível em: https://developer.spotify.com/documentation/web-api. Acesso em: 20 mai. 2025.