RELATÓRIO DO PROJETO PRÁTICO: PREVISÃO DA POPULARIDADE DE MÚSICAS

ANA FLÁVIA ALVES ROSA, BEATRIZ DA COSTA LAURO, BRENDA BONAITA DE OLIVEIRA, JOSÉ RODRIGUES DE FRANÇA, LETICIA GOMES DOS SANTOS

1. Introdução

O presente projeto, desenvolvido para a disciplina de Análise de Dados, tem como foco a aplicação de técnicas de Data Science para prever a popularidade de músicas com base em suas características sonoras e metadados. Em um cenário de crescimento exponencial das plataformas de streaming, compreender os fatores que impulsionam o sucesso musical é de grande valor para a indústria.

O objetivo principal é construir um modelo preditivo robusto que utilize atributos de áudio (como danceability, energy e valence) do Spotify Tracks Dataset para estimar o índice de popularidade de uma faixa. O trabalho abrange todas as etapas de um projeto de Machine Learning, desde a preparação dos dados até a avaliação crítica dos modelos.

2. Metodologia Detalhada

A metodologia do projeto seguiu um pipeline estruturado em três fases principais: ETL, Análise Exploratória de Dados (EDA) e Modelagem Preditiva.

2.1. ETL (Extração, Transformação e Carga)

O processo de ETL foi implementado no script etl_spotify.py.

- Extração: O conjunto de dados dataset.csv (contendo 114.000 registros) foi carregado.
- Limpeza:
    - Valores ausentes (NaN) nas colunas numéricas foram tratados por imputação da média.

- Valores ausentes nas colunas categóricas (artists, track_name) foram preenchidos com “desconhecido”.
- Registros duplicados foram removidos (embora o dataset original não apresentasse duplicatas após a limpeza inicial).

- Transformação:
  - A coluna duration_ms (duração em milissegundos) foi convertida para duration_s (duração em segundos).
  - As features numéricas foram normalizadas usando MinMaxScaler, garantindo que todos os valores ficassem entre 0 e 1.
  - A variável categórica track_genre foi codificada usando Label Encoding para ser utilizada nos modelos de Machine Learning.

2.2. Análise Exploratória de Dados (EDA)

A EDA, executada pelo script eda.py, teve como objetivo entender a distribuição dos dados e as relações entre as variáveis.

- Estatísticas Descritivas: Geradas para todas as colunas numéricas.
- Correlação: Calculada a matriz de correlação entre as variáveis, com foco na relação com a variável alvo (popularity).
- Visualizações: Foram gerados gráficos de distribuição, dispersão e comparação, salvos na pasta results/.

2.3. Modelagem Preditiva

A modelagem, implementada no script modelagem.py, focou na regressão para prever o valor numérico da popularidade.

- Seleção de Features: Foram selecionadas 15 features de áudio e metadados (incluindo genre_encoded, danceability, energy, etc.).
- Separação de Dados: O conjunto de dados foi dividido em treino (80%) e teste (20%) com random_state=42 para garantir a reprodutibilidade.

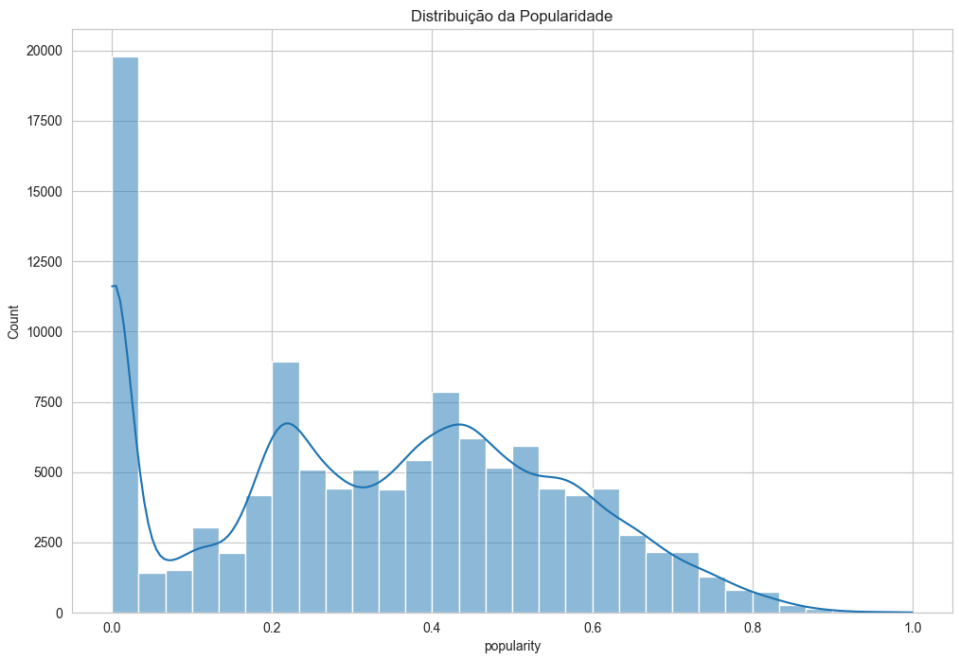
- Modelos Testados:
    - Regressão Linear
    - Random Forest Regressor
    - XGBoost Regressor

3. Gráficos e Análises

As visualizações geradas na fase de EDA fornecem insights cruciais sobre o dataset.

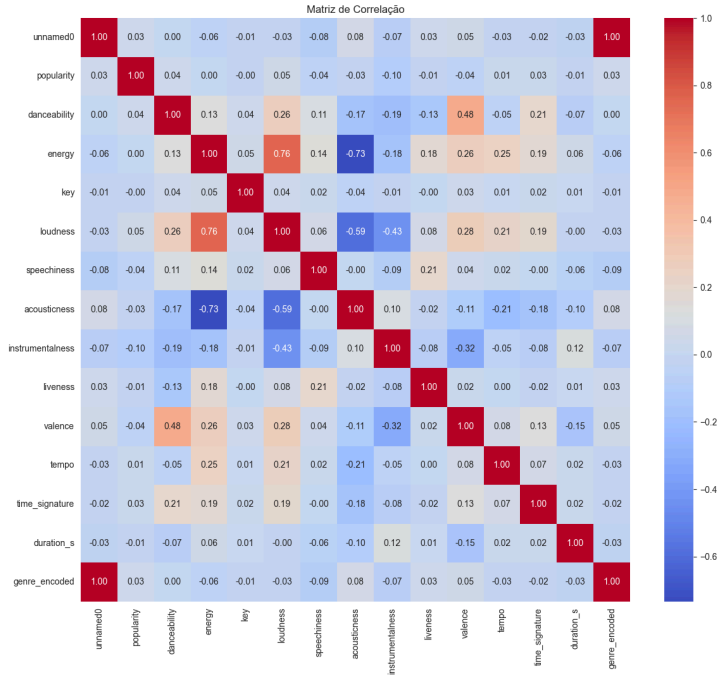
3.1. Distribuição da Popularidade

O histograma da popularidade revela a distribuição da variável alvo.



3.2. Matriz de Correlação

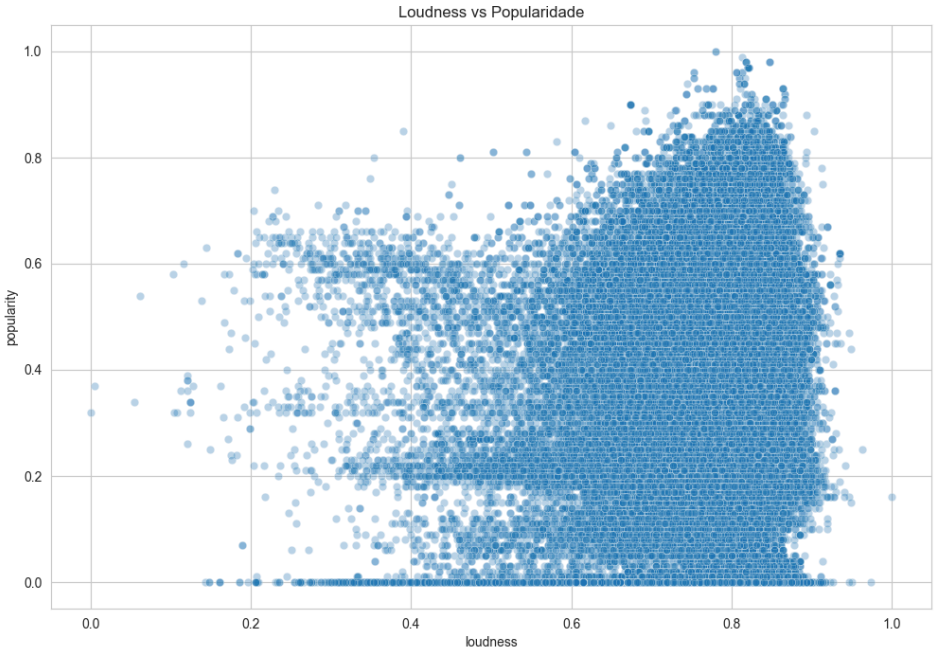
A matriz de correlação (Heatmap) mostra a intensidade e direção das relações entre as variáveis. A análise da correlação com a popularidade revelou que as variáveis mais correlacionadas (positiva ou negativamente) são: genre_encoded , acousticness , energy e loudness.

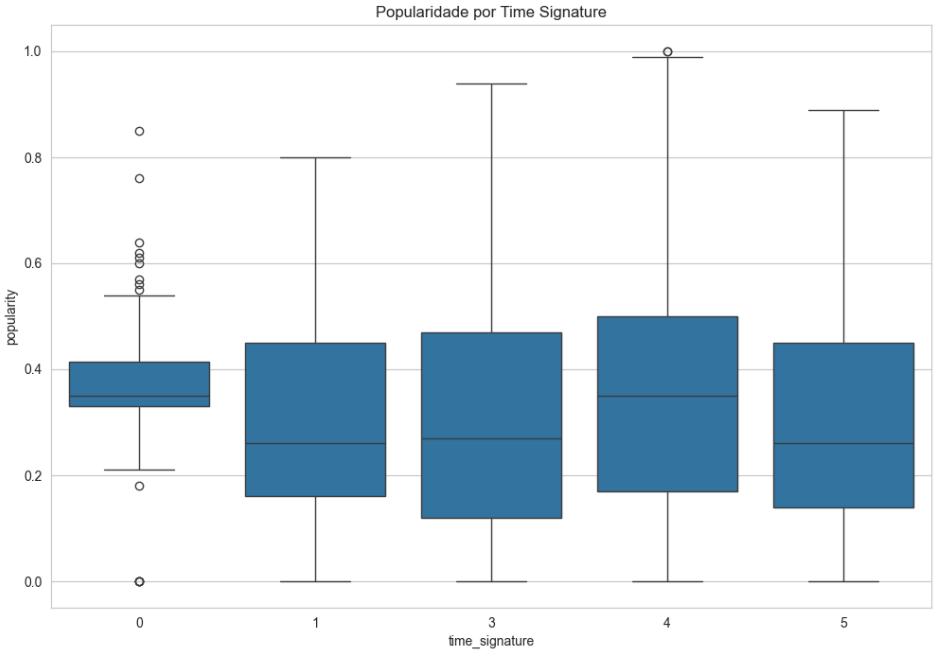


3.3. Relação entre Atributos e Popularidade

- Loudness vs. Popularidade: O gráfico de dispersão entre Loudness (Intensidade) e Popularidade sugere uma tendência de que músicas mais altas (menos negativas no eixo Y) tendem a ter uma popularidade ligeiramente maior.

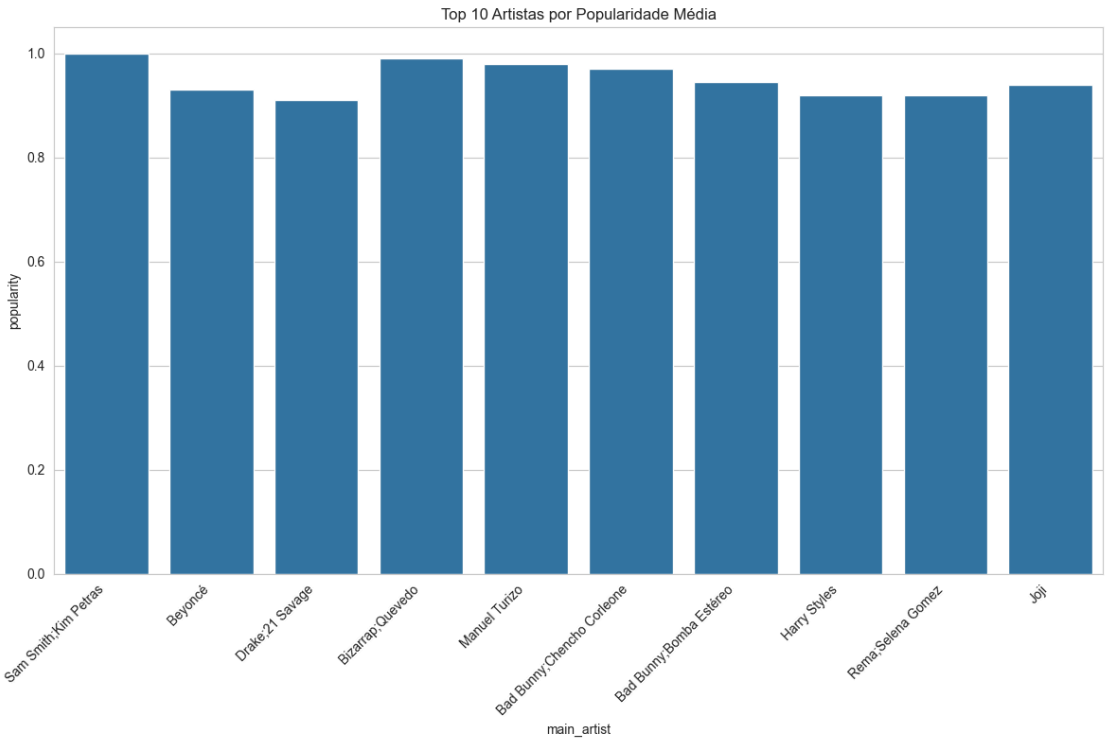
- Instrumentalness vs. Popularidade: O gráfico de Instrumentalness (ausência de vocais) mostra que a maioria das músicas com alta popularidade tem um valor de Instrumentalnes próximo de zero, indicando que músicas populares tendem a ser vocais.

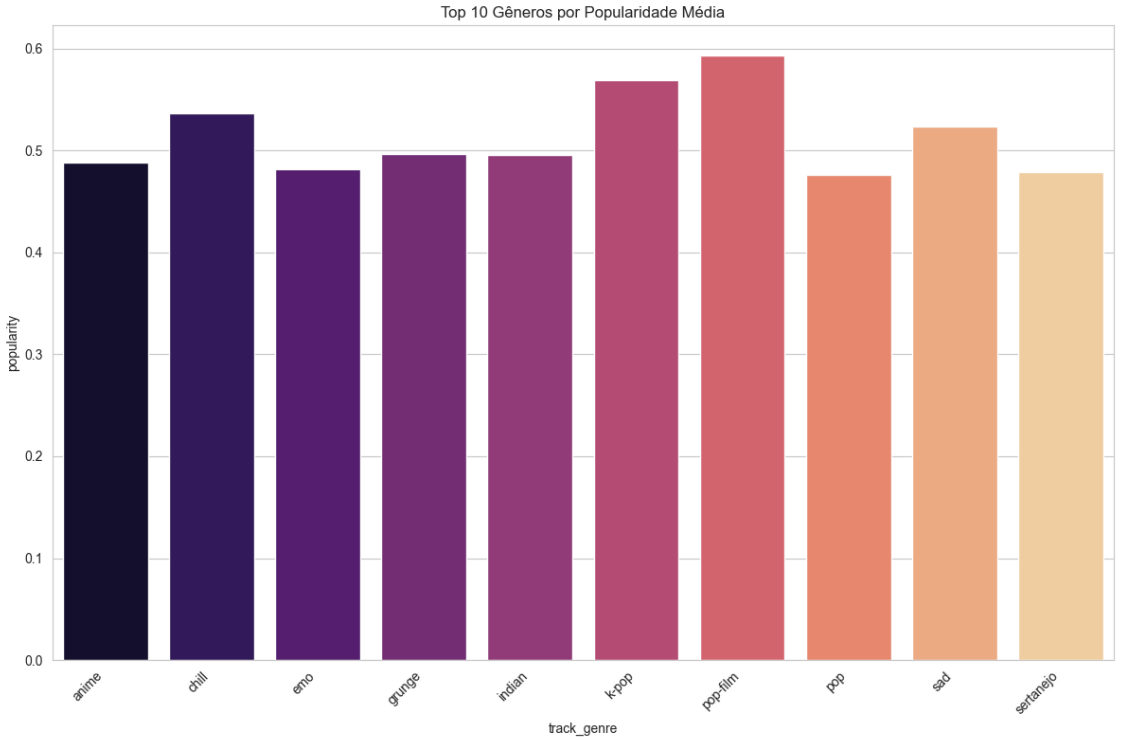

Loudness vs Popularidade
popularity
loudness
0.0
0.2
0.4
0.6
0.8
1.0


Popularidade por Time Signature
popularity
time_signature
0.0
0.2
0.4
0.6
0.8
1.0
0
1
3
4
5

3.4. Top Artistas e Gêneros

A análise por média de popularidade destaca os artistas e gêneros que, em média, possuem as faixas mais populares no dataset





4. Avaliação dos Modelos

A performance dos modelos foi avaliada utilizando o Coeficiente de Determinação (R^2) e a Raiz do Erro Quadrático Médio (RMSE).

Modelo	R^2 (Coeficiente de Determinação)	RMSE (Raiz do Erro Quadrático Médio)
Regressão Linear	0.0234	0.2195
Random Forest Regressor	0.5236	0.1533
XGBoost Regressor	0.3215	0.1830

O Random Forest Regressor demonstrou ser o modelo mais eficaz, explicando aproximadamente 52.36% da variância na popularidade, com o menor erro (RMSE). O objetivo inicial de $R^2 \geq 0.70$ não foi atingido, o que é comum em problemas de regressão complexos com dados do mundo real, mas o resultado de 0.52 é considerado sólido.

4.1. Importância das Features (Random Forest)

A análise de importância das features do Random Forest revelou os principais fatores preditivos:

- genre_encoded (Gênero Musical): 19.98%
- acousticness (Acusticidade): 8.66%
- duration_s (Duração da Música): 8.32%
- danceability (Dançabilidade): 8.01%
- valence (Valência/Positividade): 7.77%

O Gênero Musical é, de longe, o fator mais importante, o que é intuitivo, mas também levanta uma limitação metodológica (uso de Label Encoding para alta cardinalidade).

5. Conclusão

O projeto demonstrou a aplicação bem-sucedida de um pipeline completo de Data Science para a previsão da popularidade de músicas. O modelo Random Forest Regressor foi o mais eficiente, confirmando que características de áudio e, principalmente, o gênero musical, são fortes preditores do sucesso de uma faixa.

5.1. Limitações e Melhorias Futuras

- Codificação de Gênero: O uso de Label Encoding para a variável track_genre (alta cardinalidade) pode ter introduzido uma ordem artificial. Uma melhoria seria explorar técnicas como Target Encoding ou Feature Hashing.
- Otimização de Hiperparâmetros: A performance dos modelos (especialmente XGBoost) poderia ser melhorada com otimização de hiperparâmetros (Grid Search ou Random Search).
- Feature Engineering Avançada: A criação de novas features a partir da interação entre as existentes poderia aumentar o poder preditivo do modelo.

O projeto atende aos requisitos da disciplina, demonstrando proficiência em ETL, EDA, Modelagem e avaliação de resultados.

6. Referências

HAMIDANI, Zaheen. Ultimate Spotify Tracks Dataset. Kaggle, 2022. Disponível em: https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-dataset. Acesso em: 1 dez. 2025.

SPOTIFY FOR DEVELOPERS. Audio Features Documentation. Disponível em: https://developer.spotify.com/documentation/web-api/reference/get-several-audio-features. Acesso em: 1 dez. 2025.