Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Curitiba

TRABALHO DE MINERAÇÃO DE DADOS

Popularidade de Músicas do Spotify

Características que impulsionam o sucesso de uma faixa

Bruno Ikeda Silva

Curitiba
06/2023

SUMÁRIO

1. OBJETIVO

O objetivo é desenvolver um modelo preditivo utilizando os atributos musicais disponíveis para prever a popularidade das músicas. Além disso, pretende-se identificar quais atributos têm maior influência na popularidade, fornecendo insights sobre os fatores que impulsionam o sucesso das músicas na plataforma do Spotify.

2. DADOS E ABORDAGEM

A fonte de dados utilizada foi um dataset contendo atributos musicais de faixas do Spotify.

2.1. DADOS

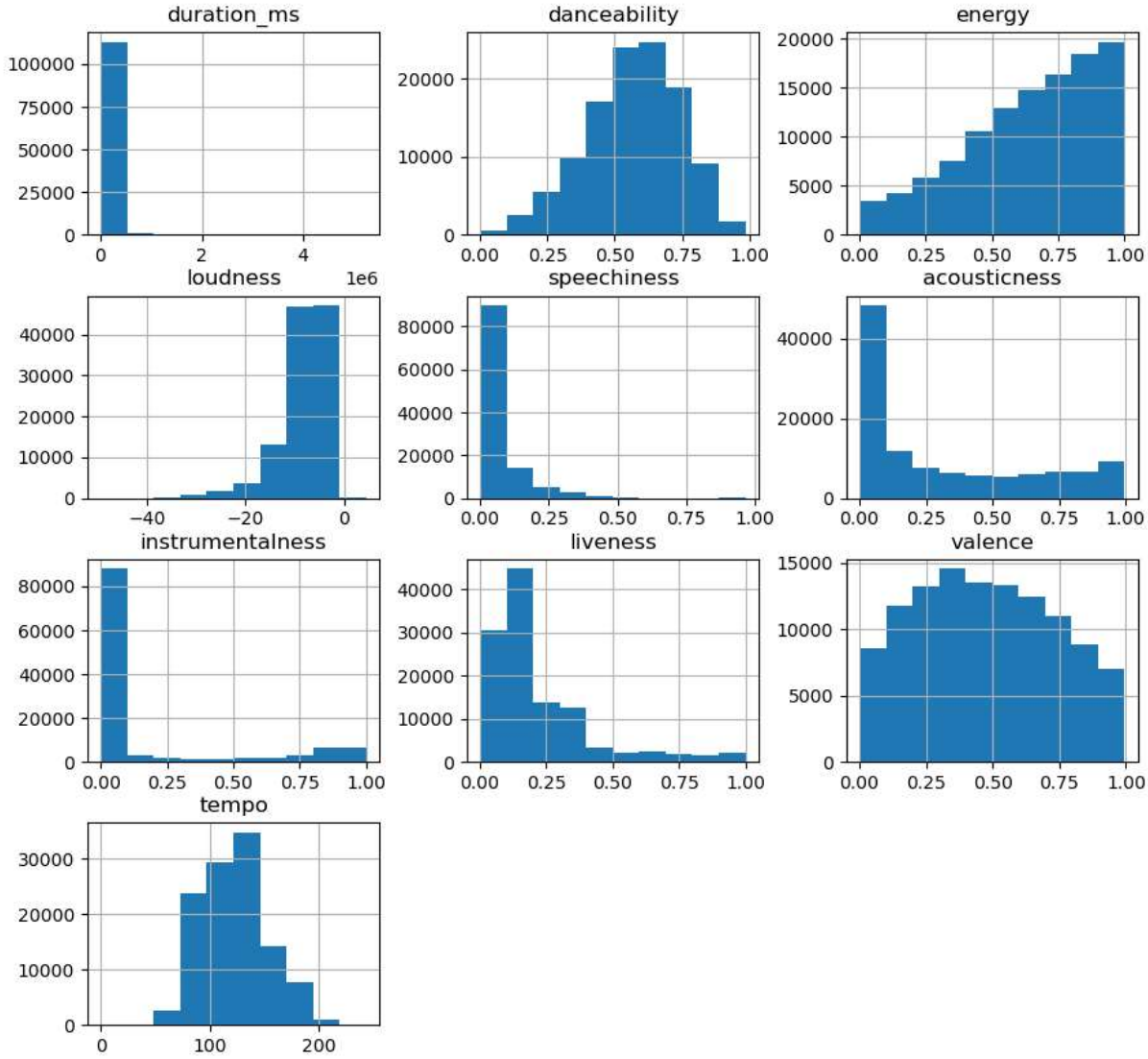
https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

2.2. ABORDAGEM

1. Limpeza e tratamento dos dados
2. Análise Exploratória
3. Normalização
4. Label Encoding
5. Oversampling
6. Construção dos Modelos
7. Avaliação dos Modelos
8. Avaliação dos Atributos Influentes

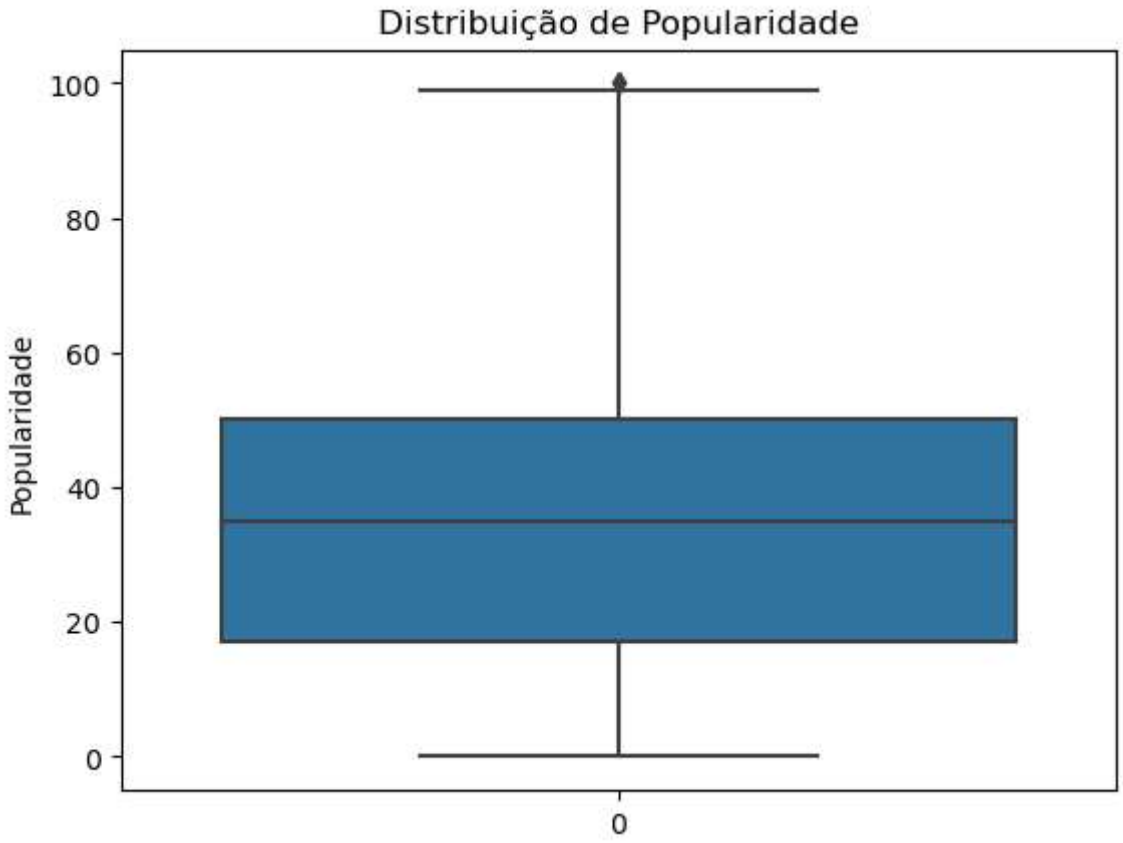
3. RESULTADOS E DISCUSSÃO

3.1. HISTOGRAMA DE VARIÁVEIS QUANTITATIVAS

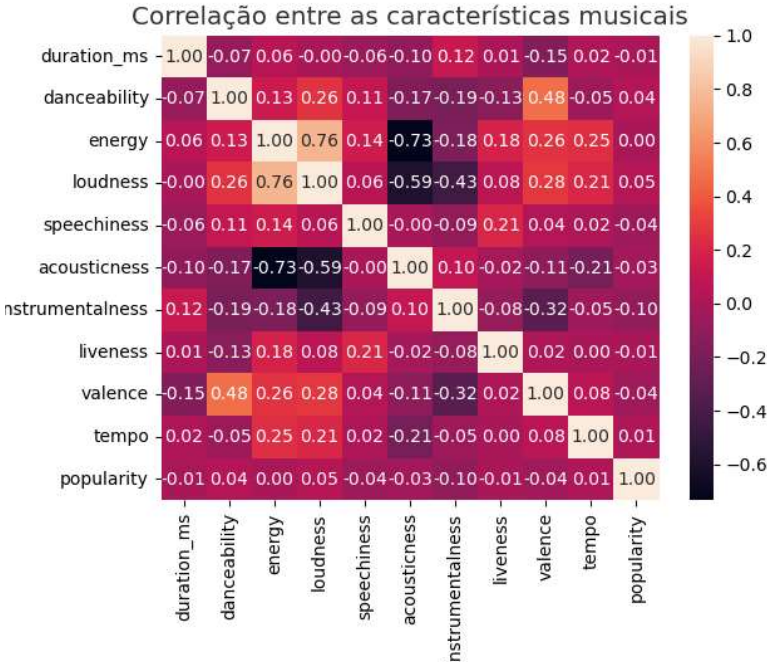


- energy segue uma distribuição crescente
- tempo, danceability e valence aparentemente se aproximam de uma Gaussiana

3.2. POPULARITY



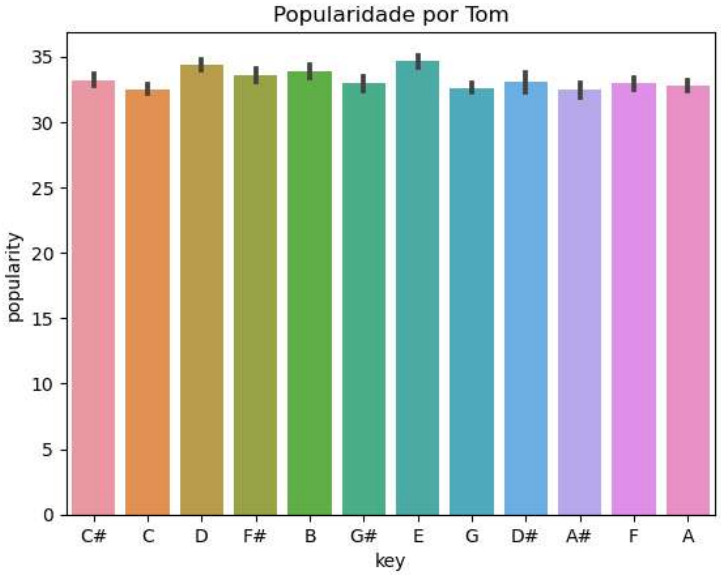
3.3. CORRELAÇÃO ENTRE VARIÁVEIS

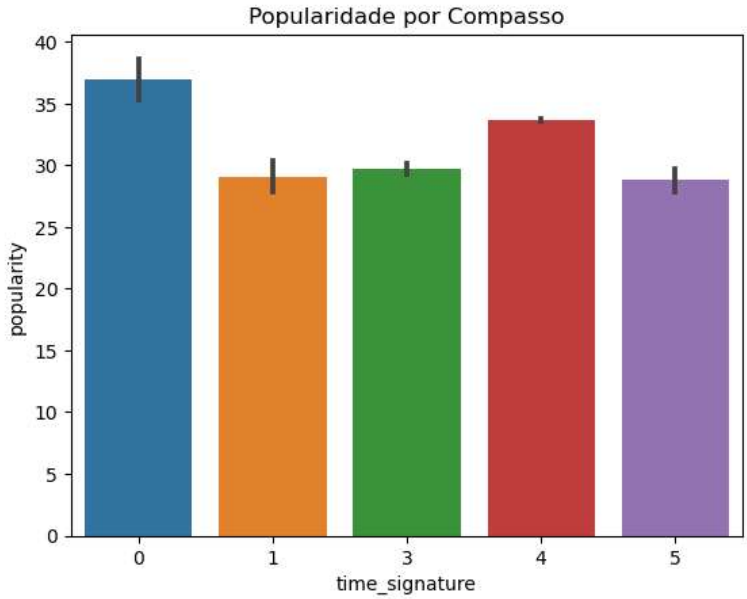


- Nenhuma das variáveis contínuas possui alta correlação com popularity (nossa variável alvo)

- energy e acousticness possuem alta correlação negativa, o que pode causar algum problema de multicolinearidade posteriormente, assim como loudness e energy.

3.4. Popularidade x Variáveis Qualitativas





3.5. Rápida primeira análise dos atributos

Dep. Variable:	popularity	**R-squared (uncentered):**	0.692
Model:	OLS	**Adj. R-squared (uncentered):**	0.692
Method:	Least Squares	**F-statistic:**	1.604e+04
Date:	Tue, 27 Jun 2023	**Prob (F-statistic):**	0.00
Time:	15:14:26	**Log-Likelihood:**	-5.1515e+05
No. Observations:	113999	**AIC:**	1.030e+06
Df Residuals:	113983	**BIC:**	1.030e+06
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
duration_ms	-0.1833	0.068	-2.715	0.007	-0.316	-0.051
danceability	1.0172	0.078	13.009	0.000	0.864	1.170
energy	-0.4366	0.074	-5.925	0.000	-0.581	-0.292
speechiness	-1.4699	0.072	-20.530	0.000	-1.610	-1.330
instrumentalness	-2.4370	0.071	-34.330	0.000	-2.576	-2.298
liveness	0.2808	0.070	4.025	0.000	0.144	0.418
valence	-2.5310	0.081	-31.236	0.000	-2.690	-2.372
tempo	0.4255	0.068	6.217	0.000	0.291	0.560
artists	6.447e-05	7.16e-06	9.008	0.000	5.04e-05	7.85e-05
album_name	9.146e-05	5.13e-06	17.841	0.000	8.14e-05	0.000
track_name	1.479e-05	3.27e-06	4.525	0.000	8.39e-06	2.12e-05
explicit	3.9371	0.251	15.704	0.000	3.446	4.428
key	0.1478	0.019	7.909	0.000	0.111	0.184
mode	0.2064	0.137	1.505	0.132	-0.062	0.475
time_signature	8.9030	0.079	112.001	0.000	8.747	9.059
track_genre	0.0371	0.002	18.756	0.000	0.033	0.041

Utilizando a biblioteca statsmodel para criar um rápido modelo de regressão linear, apesar de rústico, conseguimos compreender que as variáveis: artists, album_name e track_name, não serão relevantes para nosso futuro modelo pois seus coeficientes são muito próximos de zero.

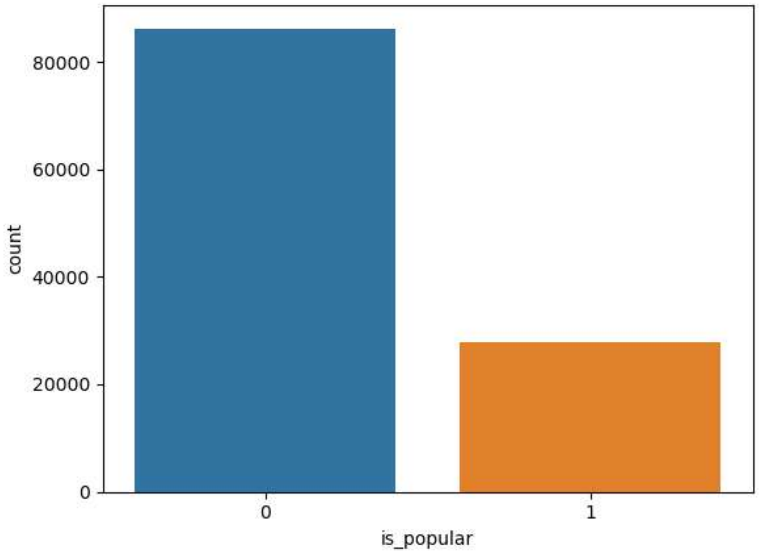
Embora o R2_score esteja relativamente alto, para dados de teste a regressão linear não performa com qualidade como veremos posteriormente.

3.6. Pré-processamento

Após realizar a normalização e o Label Encoding dos dados, foi realizada uma etapa de binarização da popularidade.

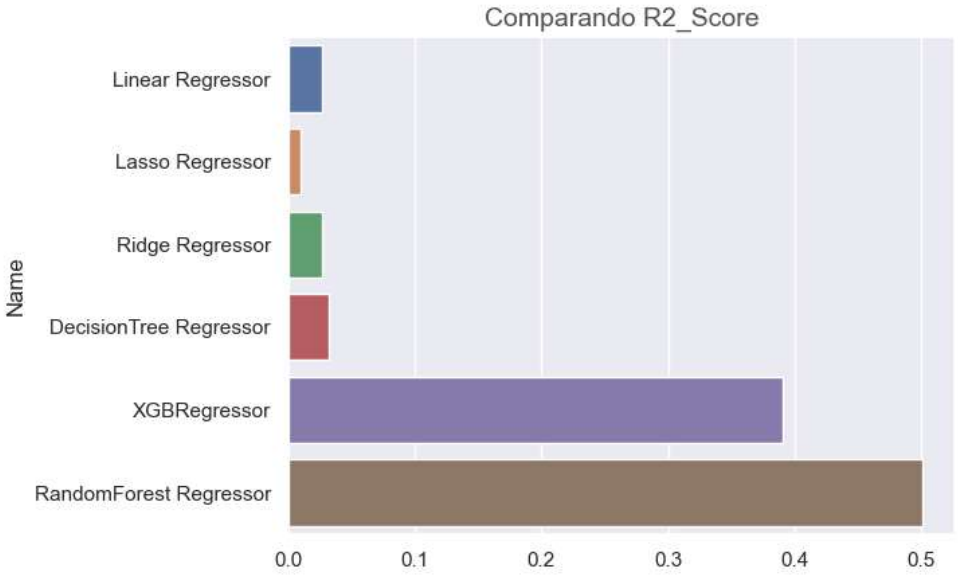
Cria-se um novo atributo chamado *is_popular* que recebe 1 caso a popularidade seja maior que 50 ou recebe 0 caso contrário.

Motivo: Utilizar modelos de classificação para prever apenas se a música é ou não é popular.



- Muito discrepante a diferença.
- Posteriormente foi realizada técnicas de oversampling.

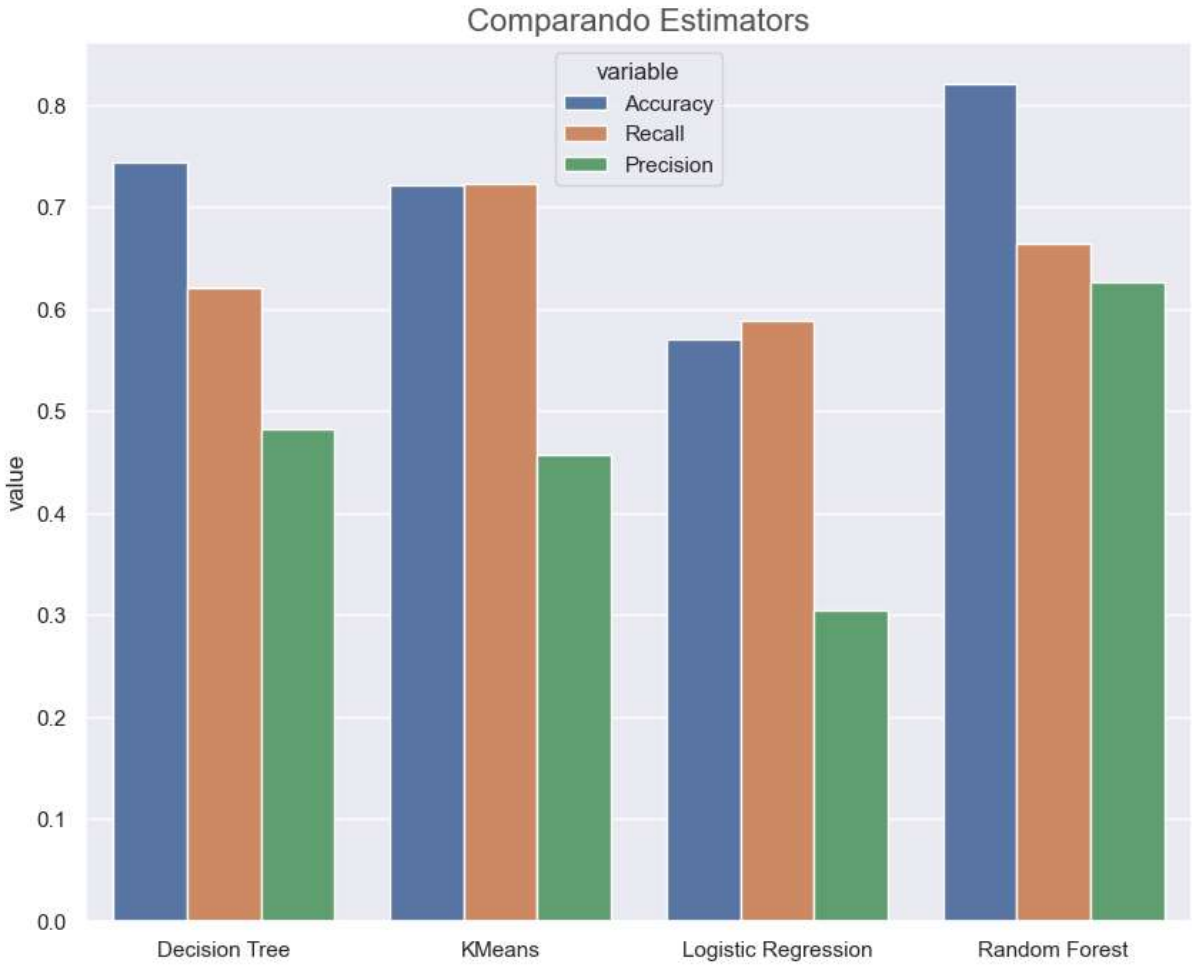
3.7. COMPARANDO MODELOS DE REGRESSÃO



Os modelos de regressão não tiveram uma boa performance com os dados de teste. Para a explicação de fatores influentes, podem ser boas ferramentas, entretanto, para serem utilizados como modelos preditivos ainda estão precários. (O melhor modelo tem 0.5 em seu R2)

3.8. COMPARANDO MODELOS DE CLASSIFICAÇÃO

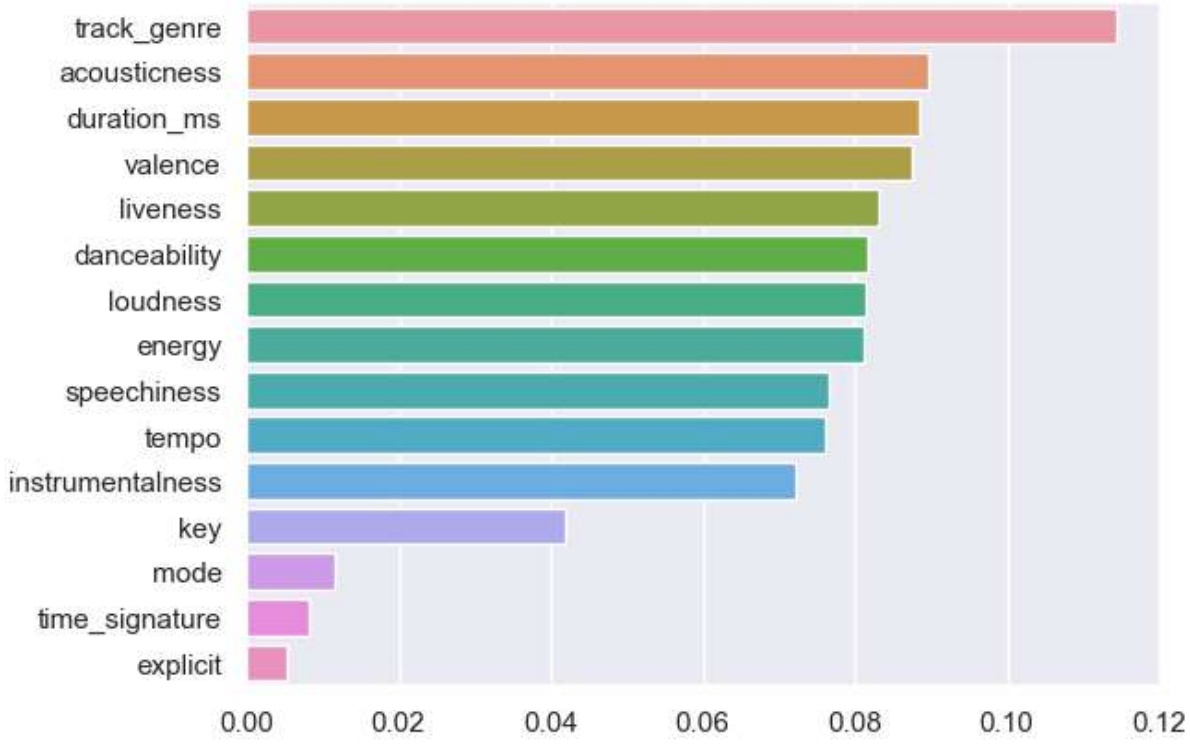
Após a binarização da variável *popularity,* precisamos manipular sua distribuição. Por esse motivo, foi utilizado a técnica de SMOTE para oversampling.



- Aparentemente RandomForest foi o melhor modelo.
- Embora não tenha uma precisão tão alta, KMeans possuiu o maior valor para o Recall, ou seja, de todos os casos positivos no dataset, ele foi o que mais acertou.
- Mesmo realizando técnicas de oversampling, ainda tivemos grandes erros ao prever a classe positiva em todos os modelos.
- Mesmo realizando técnicas de oversampling, ainda tivemos grandes erros ao prever a classe positiva em todos os modelos. Talvez uma solução para isso fosse realizar um undersampling ao invés.

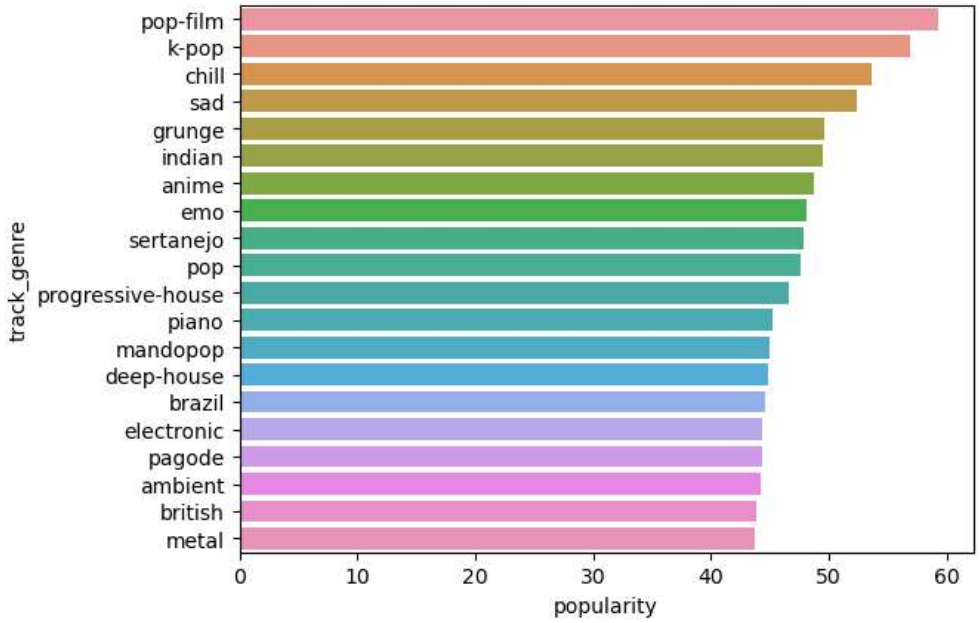
3.9. AVALIANDO OS ATRIBUTOS MAIS INFLUENTES

Para realizar essa análise, o modelo utilizado foi o RandomForestClassifier, pois apresentou o melhor resultado.



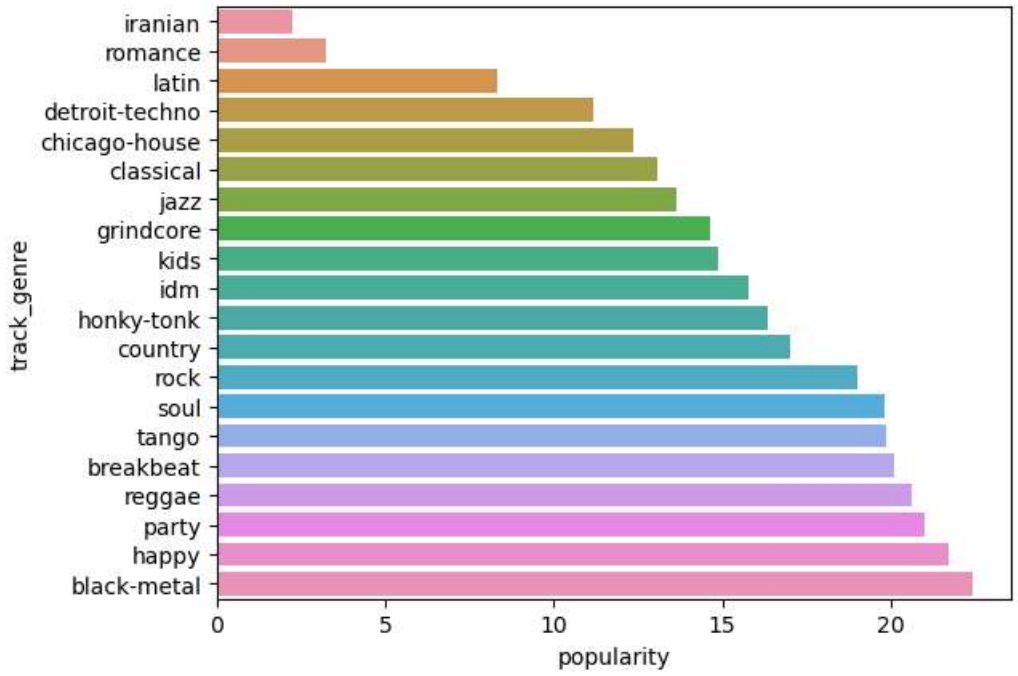
- É bastante perceptível que track_genre é o atributo mais relevante para o nosso modelo para sua previsão.
- explicit, time_signature (compasso) e mode (maior ou menor) não contribuem tanto para a previsão da popularidade da música.

3.9.1. GÊNEROS MUSICAIS MAIS POPULARES



- Alguns estilos musicais brasileiros se destacam no meio dos 20 gêneros mais populares, como o sertanejo, brazil, pagode.

3.9.2. GÊNEROS MUSICAIS MENOS POPULARES



- Alguns gêneros mais nichados são menos populares, como o jazz, black-metal, classical e até o rock.

4. CONSIDERAÇÕES FINAIS

O trabalho de criar um modelo preditivo para previsão da popularidade de uma música do Spotify é um tópico interessante e passível de um estudo mais aprofundado.

Pelo que conseguimos observar, o gênero musical possui a maior influência para prever a popularidade da música enquanto os outros atributos musicais tem menos importância. Isso pode ocorrer devido à existirem gêneros mais focados em atingir grande público enquanto outros podem ser considerados mais nichados.

Outro ponto a se analisar é que outros atributos não presentes no dataset poderiam ajudar o modelo a ser mais eficaz. Atributos como: ano, gravadora, país, idioma, etc.

A escolha de binarizar a popularidade nos trouxe vantagens e desvantagens. A grande vantagem é que dessa forma, conseguimos ter uma noção mais clara e resumida da popularidade da música (é ou não é popular), por isso modelos de classificação acabam sendo mais claros ao classificar uma música. A grande desvantagem é que perdemos a intensidade da popularidade da música, então não sabemos o quanto a música é popular.

Utilizar a regressão é um trabalho um pouco mais complexo, dado ao fato que muitos algoritmos trabalham com uma linearidade dos dados, o que é um problema ao se tratar de músicas e suas características, uma vez que sua popularidade não segue uma progressão linear dos seus atributos.

5. MATERIAIS UTILIZADOS

- Ambiente: Jupyter-notebook
- Kernel: Anaconda Python
- Bibliotecas:
    - numpy
    - pandas
    - matplotlib
    - seaborn
    - scikit-learn
    - imblearn
    - xgboost