UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

WEBER CORDEIRO GODOI

AVALIAÇÃO DA INFLUÊNCIA DE CARACTERÍSTICAS MUSICAIS SOBRE A POPULARIDADE DE CANÇÕES NAS CHARTS DO SPOTIFY BRASILEIRO EM 2021

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

DOIS VIZINHOS
2022

WEBER CORDEIRO GODOI

AVALIAÇÃO DA INFLUÊNCIA DE CARACTERÍSTICAS MUSICAIS SOBRE A POPULARIDADE DE CANÇÕES NAS CHARTS DO SPOTIFY BRASILEIRO EM 2021

EVALUATION OF THE INFLUENCE OF MUSICAL CHARACTERISTICS ON THE POPULARITY OF SONGS ON THE BRAZILIAN SPOTIFY CHARTS IN 2021

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Jefferson Tales Oliva

DOIS VIZINHOS
2022

4.0 Internacional

Esta licença permite remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es) e que licenciem as novas criações sob termos idênticos. Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

WEBER CORDEIRO GODOI

AVALIAÇÃO DA INFLUÊNCIA DE CARACTERÍSTICAS MUSICAIS SOBRE A POPULARIDADE DE CANÇÕES NAS CHARTS DO SPOTIFY BRASILEIRO EM 2021

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de aprovação: 18/novembro/2022

Jefferson Tales Oliva
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

Dalcimar Casanova
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

Marcelo Teixeira
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

DOIS VIZINHOS
2022

AGRADECIMENTOS

A minha mãe Silvia por todo amor e carinho dedicados ao meu desenvolvimento pessoal. Agradeço por todo apoio nas minhas decisões de carreira.

Aos meus irmãos Wyver e Wylli, por todo companheirismo, conhecimento e experiências compartilhadas.

RESUMO

Atualmente, o formato digital é o principal canal de distribuição de músicas e a plataforma do Spotify se destaca como a principal distribuidora. As receitas geradas pelas músicas são compartilhadas entre distribuidoras, artistas, compositores e gravadoras, sendo proporcionais ao número de reproduções. Nesse cenário, distribuidoras que adotam métodos capazes de prever o potencial de popularidade de canções antes de seu lançamento são mais competitivas à medida que podem otimizar a alocação de recursos financeiros em músicas mais propensas a atingir altos índices de popularidade. Dessa forma, esse trabalho tem por objetivo avaliar a influência de características musicais sobre o desempenho das canções nas charts brasileiras do Spotify em 2021, valendo-se da utilização de modelos de classificação de aprendizado de máquina floresta aleatória e da utilização técnicas de interpretabilidade de modelos caixa preta como Feature Importance, Permutation Importance e o SHAP Feature Importance. Também são aplicadas, nesse trabalho, técnicas não dependentes de modelos de aprendizado de máquina como a correlação de Pearson e o Information Value. Os dados referentes ao número de reproduções, características de artistas e características musicais de canções que atingiram o top 200 do Spotify brasileiro no ano de 2021 foram coletados e um índice de popularidade foi calculado para rotulação das canções entre músicas populares e comuns. Os resultados mostram que as características dos artistas como o número de seguidores, o índice de popularidade do artista e os gêneros musicais atribuídos ao artista são as três características mais importantes para determinação da popularidade de uma canção, seguidas pelas características musicais como a proporção de palavras faladas na canção, acusticidade, energia, tempo e o índice relativo à presença de audiência.

Palavras-chave: Características musicais; Spotify, Popularidade; Interpretabilidade; Modelos caixa preta.

ABSTRACT

Currently, the digital format is the main music's distribution channel and the spotify platform stands out as the main distributor. The revenues generated by the songs are shared between distributors, artists, composers and record studios, being proporcional to the number of streams. In this scenario, distributors that adopt methods capable of predicting song's popularity before their release are more competitive as they can optimize the allocation of financial resources to songs that are more likely to reach high popularity. This work have the main goal to the influence of musical characteristics on the performance of songs on Spotify's Brazilian charts in 2021, using random forest classification model and interpretability techniques of black box models such as features importance, permutation importance and SHAP feature Importance. Other techniques not dependent on machine learning models such as pearson's correlation and information value are also applied. Data regarding the number of streams, artist characteristics and musical characteristics of songs that reached the top 200 of brazilian Spotify charts in 2021 were collected and a popularity index was calculated for labeling the songs between popular and common songs. The results show that artist characteristics such as the number of followers, the artist's popularity index and the musical genres assigned to the artist are the three most important characteristics for determining the popularity of a song, followed by musical characteristics such as speechiness, acousticness, energy, tempo and liveness.

Keywords: Musical characteristics; Spotify; Popularity; Interpretability; Black box models.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CD	Compact Disc
MP3	MPEG-1/2 Audio Layer 3
IV	Information Value
WOE	Weight of Evidence
CSV	Comma-separated values
API	Application Programming Interface
ROC	Receiver operating characteristic curve
AUC	Area under the ROC Curve
FI	Feature Importance
PI	Permutation Importance

SUMÁRIO

1 INTRODUÇÃO

O meio mais antigo de reprodução de música que se tem conhecimento é a forma síncrona, em que músicos e ouvintes compartilham o mesmo ambiente. Contudo, com avanço da tecnologia, novas maneiras de reprodução, consumo e distribuição de músicas passaram a coexistir e competir com os métodos tradicionais (SILVA, 2015)

Atualmente, o formato digital é o principal canal de distribuição de música. Na história recente, os principais meios de distribuição evoluíram desde o Vinil, CD, MP3 aos serviços de streaming como o Spotify. Esse, no que lhe concerne, tem um papel de destaque no mercado de distribuição musical, sendo considerado a maior plataforma e serviço de música por streaming do mundo com 40 milhões de usuários e presença em 57 países (ARAÚJO; OLIVEIRA, 2014).

A receita gerada pelas músicas aumenta conforme o desempenho no número de reproduções alcançadas na plataforma. O modelo de negócio do Spotify permite o compartilhamento dessas verbas entre a distribuidora, artistas, compositores e gravadoras (KLEINA, 2018). Para as indústrias fonográficas, emplacar hits de sucesso, portanto, é uma das formas de gerar receitas milionárias e motiva a busca por meios que permitam analisar o potencial de popularidade das canções antes do lançamento e distribuição.

Na literatura, estudos foram conduzidos com a finalidade de identificar padrões nas preferências dos ouvintes brasileiros no Spotify por meio de análises exploratórias das características musicais (CUSTÓDIO et al., 2021). Num contexto semelhante, Middlebrook e Sheik (2019) propõe a aplicação de modelos de aprendizado de máquinas para classificação de músicas populares e comuns.

Ambas as abordagens partem da premissa do sucesso da canção provir da sua presença nas charts do Spotify e Billboard respectivamente. Araujo, Cristo e Giusti (2019), por outro lado, propõe uma metodologia semelhante ao aplicar modelos de aprendizado de máquina para prever músicas populares, considerando a entrada no top 50 nas charts do Spotify como fator determinante para o sucesso.

Esse estudo, contudo, propõe um cálculo de um indicador de popularidade a partir do número de streams e do tempo de permanência nas charts do Spotify e o decil superior desse indicador é usado como referência para rotulação de músicas populares e comuns.

Além disso, serão abordados, nesse trabalho, técnicas quantitativas e também modelos de aprendizado de máquina combinadas á técnicas de interpretabilidade para determinação da influência das características musicais sobre a popularidade das músicas. Os resultados de ambas as abordagens serão, posteriormente, analisadas e comparadas.

Esse trabalho se propõe á analisar a influência das características musicais sobre a popularidade de músicas, para isso, será utilizado um modelo aprendizado de máquina e aplicada técnicas de interpretabilidade além de outras técnicas não dependentes do modelo. Os resultados de ambas as abordagens serão, posteriormente, comparadas e analisadas.

1.1 Objetivos

Os objetivos do trabalho são apresentados a seguir.

1.1.1 Objetivo Geral

O objetivo desse trabalho é avaliar a influência das características musicais sobre o desempenho nas charts do Spotify brasileiro no ano de 2021. Para isso, foram definidos os seguintes objetivos específicos:

1.1.2 Objetivos Específicos

- Utilização de métodos de aprendizado de máquina para levantamento da importância das características musicais;
- Aplicação de técnicas de interpretabilidade de modelos de aprendizado de máquina;
- Aplicação de outras técnicas para levantamento da importância das características;

1.2 Hipótese

É possível mensurar a influência das características musicais, por meio de técnicas de interpretabilidade de modelos de aprendizado de máquina e de outras técnicas não dependentes do modelo.

1.3 Justificativa

Prever a popularidades de músicas é de grande valia para indústria fonográfica, uma vez que se abre uma oportunidade de otimizar a alocação de recursos financeiros, priorizando investimentos em músicas com maior propensão a atingir altos índices de popularidade (MIDDLEBROOK; SHEIK, 2019).

Ao analisar e quantificar as características que favorecem a popularidade de canções, esse trabalho contribui para orientar escolhas de repertório de álbuns ou ajustes em faixas musicais com objetivo de otimizar o desempenho nas charts do Spotify.

1.4 Organização do Trabalho

O trabalho está organizado da seguinte forma:

- Capítulo 2: Fundamentação Teórica
- Capítulo 3: Metodologia
- Capítulo 4: Resultados e Discussões
- Capítulo 5: Considerações finais

2 FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo será apresentado conceitos teóricos que fundamentam as técnicas utilizadas nesse trabalho para extração das características musicais e de artistas mais importantes para determinação da popularidade de canções.

2.1 Modelos de Aprendizado de Máquina

2.1.1 Modelo de floresta aleatória (Random Forest)

Random forest é um algorítimo baseado em ensembles de árvores de decisões em que Cada árvore contribui para classificação de uma amostra por meio de voto (PETKOVIC et al., 2021). Ao invés de utilizar a melhor característica para divisão de um nó, o algorítimo busca a melhor opção de um subconjunto de características gerada aleatóriamente, isso introduz uma componente aleatória ao crescimento das árvores de decisões geradas que contribui para diminuição da variância e aumento do viés, e que, de modo geral, produz um modelo melhor (GÉRON, 2019). Além disso, cada árvore é construída a partir de um subconjunto dos dados de treinamento que são gerados aleatoriamente do dataset de treinamento.

Além disso, cada árvore é construída a partir de um subconjunto de mesmo tamanho dos dataset de treinamento por meio de amostragem com reposição, e que também são, portanto, um importante fator de aleatoriedade do modelo.

Devido a complexidade do funcionamento e alta variedade de árvores geradas, o modelo de floresta aleatória é considerado um modelo caixa preta cujas previsões são difíceis de explicar em termos simples (GÉRON, 2019).

2.2 Tecnicas interpretabilidade de modelos caixa preta

Um dos fatores limitantes das técnicas de interpretabilidade de modelos de aprendizado de máquina é que o erro está diretamente ligado ao erro do modelo utilizado. nessa seção será discutido 3 técnicas distintas: Feature Importance, Permutation Importance e SHAP Feature Importance.

2.2.1 Feature Importance

O algorítimo de floresta aleatória seleciona as melhores características de um subconjunto de características para construção das árvores de decisão. Por conta disso, é possível mensurar a importância das características por meio do calculo da pureza dos nós das árvores que compõe a floresta (ROGERS; GUNN, 2005). Uma das maneiras de calcular essa impureza é avaliando o ganho de informação ou entropia por meio da equação (GÉRON, 2019):

$$H_1 = \sum_{k=1|P_{i,k}\neq 0}^{n} P_{i,k} \log_2(P_{i,k}) \tag{1}$$

A importância de cada característica pode ser computada de acordo com algorítimo presente no anexo **??** desse trabalho detalhado por Bischl et al. (2022).

2.2.2 Permutation Feature Importance

O Permutation Feature Importance é uma das técnicas capazes de mensurar a importância das características em predições de modelos de aprendizado de máquina. Para isso, a importância é calculada por meio da avaliação do incremento no erro de classificação das predições ao permutar os valores das características individuais. Incrementos elevados configuram uma característica importante, em contrapartida, incrementos nulos, pequenos ou muito próximos a 0 configuram características insignificantes ou de menor importância para as predições (MOLNAR, 2022).

A principal desvantagem da aplicação dessa técnica é que seu valor é dependente do erro do modelo. Além disso, a presença de características correlacionadas entre si pode diminuir a importância individual calculada de cada característica (MOLNAR, 2022).

2.2.3 SHAP Feature Importance

Shapley Additive exPlanations ou SHAP é uma técnica capaz de mensurar a contribuição de cada característica em previsões individuais gerada por modelos de aprendizado de máquina. Esse método se baseia na teoria dos jogos em que os valores atribuídos a cada características se comportam como jogadores e a previsão como a recompensa (MOLNAR, 2022).

As combinações entre os valores de duas características devem ser computadas para o calculo do valor de Shapley. De acordo com Molnar (2022) o calculo pode se tornar complexo e de alto custo computacional a medida que o número de características aumentam. Para lidar com esse problema, o valor de Shapley pode ser calculado por meio da aproximação da amostragem de Monte-Carlo detalhado em Štrumbelj e Kononenko (2014). A aproximação, contudo não será aplicada nesse trabalho.

A contribuição global de cada característica ou SHAP Feature Importance, pode ser calculada computando a média dos valores absolutos de Shapley de todas as previsões individuais realizadas pelo modelo de Machine Learning de acordo com a equação detalhado em (MOLNAR, 2022).

$$SHAP_{FI} = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(i)} \right| \tag{2}$$

Em que ϕ corresponde ao valor de Shapley.

2.3 Outras técnicas

As técnicas abordadas nessa seção não são dependentes de modelos de aprendizado de máquina.

2.3.1 Correlação de Pearson

De acordo com o Wikipedia, a correlação de pearson é uma medida de correlação linear entre dois conjuntos de dados. Essa medida é calculada por meio da razão entre a covariância de duas variáveis pelo produto de seus desvios padrões individuais segundo a equação abaixo:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

O valor da correlação de Pearson pode variar de -1 á 1. Valores próximos de 1 caracterizam variáveis fortemente correlacionadas positivamente, valores próximos de -1 remetem á variáveis fortemente correlacionadas negativamente e valores próximos de 0 denotam variáveis com baixa ou nenhuma correlação linear.

É importante salientar que a correlação de Pearson é uma representação da relação linear de duas variáveis e, portanto, não é capaz de capturar relações não lineares (GÉRON, 2019).

2.3.2 Information Value (IV)

Essa técnica é aplicada tradicionalmente em análise de risco de crédito (EGAN, 2021). A aplicação dessa técnica requer, como pré-requisito, o cálculo do Weight of Evidence (WOE).

O WOE mensura o poder preditivo de uma variável dependente em relação a variável independente (BHALLA, 2016). Cada característica contínua é subdividida em categorias ou intervalos, essa divisão, no entanto, é reaproveitada caso a características seja do tipo categórica. Para cada intervalo ou categoria o WOE pode ser calculado de acordo a formula:

$$WoE = ln\left(\frac{\%_{Eventos}}{\%_{NaoEventos}}\right) \tag{4}$$

Segundo Bhalla (2016) o número de categorias assim como seus limites superiores e inferiores são definidos de forma a respeitar os seguintes critérios:

- Cada categoria criada deve conter ao menos 5% de eventos
- As categorias não podem conter eventos e não eventos nulos;
- O WOE deve ser distinto para cada categoria;
- O WOE deve ser monotônico, crescente ou decrescente dentro de um mesmo agrupamento de categorias;
- Os valores faltantes são categorizados separadamente;
- Índice de liveness pequeno;
- Tempo intermediário á elevado.

Por fim, o Information Value (IV) é calculado considerando todas as categorias de acordo com a formula:

$$IV = \sum(\%_{Eventos} - \%_{NaoEventos}) * WoE \tag{5}$$

A implementação do numero ótimo de categorias, calculo de WOE e IV é facilitada por meio da biblioteca XVerse.

3 MATERIAIS E MÉTODOS

Esse capítulo tem como objetivo apresentar os materiais e apontar as metodologias utilizadas.

O Python foi utilizado como linguagem de programação para processamento dos dados.

3.1 Coleta dos dados

O URL identificador das músicas e o número de streams diárias de músicas presentes no top 200 das charts brasileiras do Spotify entre 01/01/2021 á 31/03/2022 foram coletados do site do Spotify e armazenados em disco no formato CSV (comma-separated values).

A coleta das características musicais das músicas e dos artistas foi realizada com auxílio da biblioteca SpotiPy e a API do Spotify. Ao todo, características de 1381 músicas e 388 artistas distintos foram extraídas sendo cada uma delas descrita nas subseções abaixo segundo o Spotify (2022):

3.1.1 Características musicais

1. **acousticness:** Medida que varia entre 0 e 1 que indica se uma música é acústica.
2. **danceability:** Medida que varia entre 0 e 1 que descreve o quão adequada uma faixa é para dançar. Um valor de 0,0 é o menos dançável e 1,0 é o mais dançável.
3. **duration_ms:** Duração da música em milissegundos.
4. **energy:** Medida que varia entre 0 e 1 que representa uma percepção de intensidade e atividade. Normalmente, as faixas energéticas parecem rápidas, altas e barulhentas
5. **instrumentalness:** Medida que varia entre 0 e 1 que indica se uma faixa não contém vocais. Os sons "Ooh"e "aah"são tratados como instrumentais neste contexto. Faixas de rap ou de palavras faladas são consideradas "vocais". Quanto mais próximo o valor de instrumentalness estiver de 1,0, maior a probabilidade de a faixa não conter conteúdo vocal. Valores acima de 0,5 destinam-se a representar faixas instrumentais, mas a confiança é maior à medida que o valor se aproxima de 1,0.
6. **key:** Numero inteiro que indica a nota que a faixa está. Os números inteiros indicam a notação padrão da classe de notas. Por exemplo. 0 = C, 1 = C/D, 2 = D, e assim por diante. Se nenhuma chave foi detectada, o valor é -1.
7. **liveness:** Medida que varia entre 0 e 1 que indica a presença de uma audiência na gravação. Valores mais altos de liveness representam uma probabilidade maior de que a faixa tenha sido executada ao vivo. Um valor acima de 0,8 fornece uma forte probabilidade de que a faixa esteja ao vivo.

8. **loudness:** Medida que varia de -60 á 0 que representa o volume geral de uma faixa em decibéis (dB).
9. **mode:** Medida binária que varia de 0 ou 1 que indica a modalidade (maior ou menor) de uma faixa, o tipo de escala da qual seu conteúdo melódico é derivado. Maior é representado por 1 e menor é 0.
10. **speechiness:** Medida que varia entre 0 e 1 que indica a presença de palavras faladas em uma faixa. Quanto mais exclusivamente falada a gravação (por exemplo, talk show, audiolivro, poesia), mais próximo de 1 o valor do atributo. Valores acima de 0,66 descrevem faixas que são provavelmente feitas inteiramente de palavras faladas. Valores entre 0,33 e 0,66 descrevem faixas que podem conter música e fala, seja em seções ou em camadas, incluindo casos como música rap. Os valores abaixo de 0,33 provavelmente representam músicas e outras faixas que não são do tipo que contém falas.
11. **tempo:** Número decimal que representa as batidas por minuto (BPM) geral estimado de uma faixa. Na terminologia musical, tempo é a velocidade ou ritmo de uma determinada peça e deriva diretamente da duração média da batida.
12. **time_signature:** Numero inteiro que representa a assinatura de tempo estimada. Por exemplo, a fórmula de compasso (medidor) é uma convenção de notação para especificar quantas batidas existem em cada compasso (ou medida). A assinatura de tempo varia de 3 a 7 indicando assinaturas de tempo de "3/4", a "7/4", por exemplo.
13. **valence:** Medida que varia entre 0 e 1 que descreve a positividade musical transmitida por uma faixa. Faixas com alta valência soam mais positivas (por exemplo, feliz, alegre, eufórica), enquanto faixas com baixa valência soam mais negativas (por exemplo, triste, deprimida, irritada).

3.1.2 Características dos artistas

1. **popularity:** Medida que varia entre 0 e 100 que representa a popularidade do artista. Essa popularidade é calculada a partir da popularidade de todas as faixas do artista.
2. **followers:** Numero inteiro que descreve o número total de seguidores do artista na plataforma do Spotify.
3. **genres:** Texto que apresenta uma lista dos gêneros separados por vírgula aos quais o artista está associado.

3.2 Pré-Processamento dos dados

A etapa de pré-processamento foi subdividida em 3 subseções, sendo elas:

3.2.1 Popularidade de músicas

Visando criar um índice de popularidade para cada dia único, o número de streams nas charts do top 200 do Spotify foram redimensionadas num índice que varia entre 0 e 1.

A fim de considerar a relevância da música ao longo do tempo, para cada música distinta, efetuou-se o somatório desse índice entre a data da primeira ocorrência nos charts e após de 60 dias corridos. O resultado foi dividido pela janela de tempo de 60 dias.

Devido à dificuldade em coletar o número de streams de canções que não compõe o top 200, dias em que as músicas não entraram nas charts entre a data da primeira ocorrência e 60 dias após foram penalizados, atribuindo-se o número de streams igual a zero.

Por fim, foram eliminadas do dataset as músicas cuja data da primeira entrada nos charts do spotify não ocorreu no ano de 2021.

3.2.2 Definição variável alvo

O rótulo de música popular foi atribuída ao decil superior da popularidade de músicas, totalizando 138 músicas populares e 1243 músicas comuns que compõe o dataset final.

3.2.3 Genero musical provável

A fim de relacionar apenas um gênero por canção, foi atribuído a cada música distinta todos os gêneros musicais ao qual o artista pertencia. Realizou-se uma análise de frequências de gêneros musicais dos artistas e arbitrou-se o gênero musical provável aquela a qual a música estivesse relacionada e que possuísse maior a frequência no dataset.

3.3 Amostragem

A partir dos dados coletados, foi realizado uma separação do dataset em treino, validação e teste aplicando-se amostragem estratificada. O dataset de treino foi composto por 883 registros sendo 88 musicas populares, o dataset de validação composto por 199 registros sendo 22 musicas populares e o dataset de teste composto por 249 registros sendo 28 musicas populares.

3.4 Interpretabilidade de modelos de aprendizado de máquina para avaliação de características musicais

3.4.1 Modelo Random Forest

O modelo de aprendizado de máquina floresta aleatória foi escolhido arbitrariamente, pois habilita a utilização do método de Feature Importance nativa do algorítimo.

O modelo treinado foi utilizado para aplicação de técnicas de interpretabilidade globais Features Importance, Permutation Importance e SHAP Feature Importance, e também da técnica de interpretabilidade local de Shapley.

3.4.2 Pré-processamento e seleção de hiperparâmetros

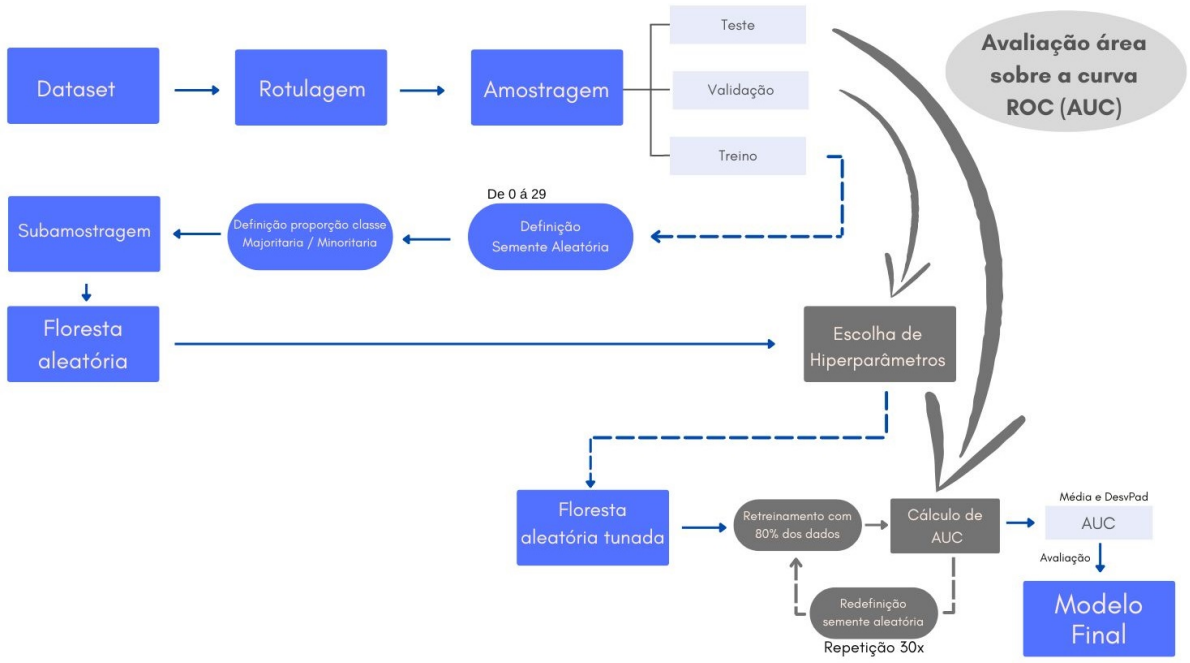
O número total de árvores, o número mínimo de amostras para dividir um nó interno foram escolhidos arbitrariamente como hiperparâmetros.

A fim de lidar o problema de classificação em bases desbalanceadas utilizou-se a técnica de subamostragem com auxílio da biblioteca Imblearn alterando-se o número de amostras para treinamento, modificando a relação do número de amostras da classe minoritária e majoritária no conjunto de treinamento. Complementarmente, foi escolhido o hiperparâmetro class weight a fim de lidar com o problema.

Os candidatos á melhores conjuntos de hiperparâmetros foram elencados avaliando-se a área sobre a curva ROC (AUC) em dados de validação após o treinamento do modelo sobre os dados provindos da subamostragem do conjunto de treino. O experimento foi repetido 30 vezes, variando-se a semente aleatória de iniciação do modelo de floresta aleatória.

Para cada conjunto distinto de hiperparâmetros candidatos, foi conduzido uma segunda batelada de experimentos com a finalidade de simular uma validação cruzada, repetindo-se por mais 30 vezes o treinamento do modelo com 80% dos dados provindos da subamostragem do conjunto de treino e avaliando a área sobre a curva ROC (AUC) em dados de testes. Novas sementes aleatórias foram geradas a cada repetição do experimento. O esquema completo está apresentado na figura na figura 1.

Figura 1 – Esquema de amostragem, subamostragem, tunagem e escolha do modelo final



Fonte: Autoria própria.

O conjunto de hiperparâmetros que apresentou maior média da área sobre a curva ROC (AUC) foi escolhido e a as sementes aleatórias geradas foram selecionadas para eleição

do modelo de floresta aleatória final conforme o esquema apresentado na figura 1.

3.4.3 Feature Importance

A importância de cada característica musical e de artistas pela técnica de Feature Importance nativa do modelo de floresta aleatória foi calculada sobre o modelo final com auxílio da biblioteca Scikit-learn.

3.4.4 Permutation Importance

Após o treinamento do modelo em dados de treino, com auxílio da biblioteca Eli5 instanciou-se o cálculo do Permutation Importance. Para essa técnica, apenas os dados de testes foram utilizados para o cálculo da importância das características musicais e de artistas.

3.4.5 SHAP Feature Importance

Calculou-se o valor de Shapley para cada característica musical e de artistas sobre os dados de teste com auxílio da biblioteca SHAP.

Extraiu-se o valor global SHAP Feature Importance aplicando-se a média dos valores absolutos dos valores de Shapley.

3.5 Avaliação características musicais com métodos quantitativos

3.5.1 Correlação de Pearson

Para todos os registros coletados calculou-se a correlação de Pearson de todas as características musicais e dos artistas, em comparação ao índice de popularidade calculado.

3.5.2 Information Value

Para o mesmo conjunto de dados utilizados no treinamento do modelo de aprendizado de máquina calculou-se, com auxílio da biblioteca Xverse, o "Weight of Evidence"de todas as características musicais e de artistas contínuas em 3 intervalos. O número de intervalos das variáveis categóricas permaneceu inalterado. Em seguida, calculou-se o Information Value para cada característica.

3.6 Analise de dados

Foi realizado uma análise dos resultados da importância das características musicais e de artista derivada dos 5 métodos aplicados.

Para cada técnica extraíram-se os valores absolutos das importâncias das características que, em seguida, foram redimensionadas para uma mesma escala de forma que a soma das

importâncias fosse equivalente a 1. Os resultados, por fim, foram comparados num gráfico do tipo mapa de calor.

Com o proposito de ranquear as características em ordem de importância, um sistema de votos foi proposto em que cada técnica contribui com 5 votos correspondentes as 5 características mais importantes. As características que assegurarem maior número de votos são consideradas as mais importantes para determinação da popularidade.

4 RESULTADOS

Nesse capitulo são apresentados os resultados das aplicações das técnicas de mensuração de importâncias de características musicais e de artistas para popularidade de canções. Os resultados de cada técnica adotada é discutida individualmente em seções. A última seção desse capítulo é reservado para comparação dos resultados.

4.1 Modelo de Machine Learning

Foram testados diversas composições de hiperparâmetros e do tamanho da subamostragem do conjunto de treinamento alimentadas ao modelo de floresta aleatória, variando-se a razão entre as classes pertencentes á classe minoritária e majoritária conforme apresentado no esquema da figura 1.

A tabela 1 apresenta os hiperparâmetros e as escolhas de pré-processamento que apresentaram maior média de área sobre a curva ROC (AUC) em dados de testes avaliados em 30 repetições de experimentos com modelo treinado com 80% dos dados de treino após a subamostragem.

Tabela 1 – Hiperparâmetros - Random Forest.

Hiperparâmetro	Valores
Class Weight 0	0.01
Class Weight 1	0.99
Criterion	Entropy
Max Depth	N/A
Min Samples Leaf	1
Min Samples Split	5
N Estimators	150

Fonte: Autoria própria.

A tabela 2 apresenta a média e desvio padrão das métricas de performance em dados de teste do modelo selecionado. A tabela 3 apresenta as métricas do modelo final.

Tabela 2 – Métricas do modelo Random Forest em %.

Acurácia	AUC	Sensibilidade	Especificidade	Precisão
52,57 $\pm 2{,}78$	60,57 $\pm 3{,}31$	69,52 $\pm 6{,}13$	50,67 $\pm 3{,}08$	13,70 $\pm 1{,}25$

Fonte: Autoria própria.

Tabela 3 – Métricas do modelo final.

Acurácia	AUC	Sensibilidade	Especificidade	Precisão
59,93%	66,66%	75,00%	58,23%	16,80%

Fonte: Autoria própria.

4.2 Feature Importance

A tabela 4 abaixo apresenta os resultados do calculo da importância de características.

Tabela 4 – Valores de Feature Importance.

Característica	Valores
seguidores cantor	0,1184
popularidade cantor	0,1169
tempo	0,11
acousticness	0,1002
danceability	0,0849
energy	0,0776
speechiness	0,0763
key	0,0749
liveness	0,0567
loudness	0,0522
duration ms	0,0512
valence	0,051
instrumentalness	0,0143
genero primario provavel	0,0082
mode	0,0051
time signature	0,002

Fonte: Autoria própria.

Duas das três características referentes aos artistas demonstraram maior poder de separabilidade das classes nos nós das árvores de decisões construídas que compõe a floresta aleatória. O número de seguidores do cantor e o índice de popularidade do cantor medido pela plataforma compõe as duas principais característica para determinação da popularidade da canção. O tempo, acousticness, danceabilidade e energy completam a lista das 6 principais características.

Por esse método, não é possível avaliar se ocorrem efeitos positivos ou negativos na popularidade de canções para cada característica.

4.3 Permutation Importance

A tabela 5 abaixo apresenta os resultados dos valores de Permutation Importance do modelo de floresta aleatória.

Tabela 5 – Valores de Permutation Importance.

Característica	Valores
seguidores cantor	0,0274
liveness	0,0209
popularidade cantor	0,0101
genero primario provavel	0,0094
energy	0,0087
speechiness	0,0072
duration ms	0,0072
key	0,0072
acousticness	0,0065
loudness	0,0051
tempo	0,0051
mode	0,0051
danceability	0,0036
instrumentalness	0,0014
valence	0,0014
time signature	0,0007

Fonte: Autoria própria.

O número de seguidores e a popularidade dos cantores compõem a primeira e terceira característica mais importante, respectivamente, sendo o número de seguidores a característica de maior importância, uma vez que, apresentou o maior incremento no erro na AUC de cerca de 2,74% após a permutação em dados de testes.

As características liveness, gênero primário provável, energy e speechiness compõe a lista das 6 principais características.

Analogamente á técnica de Feature Importance, por esse método, também não é possível avaliar a direção do efeito das características na popularidade de canções.

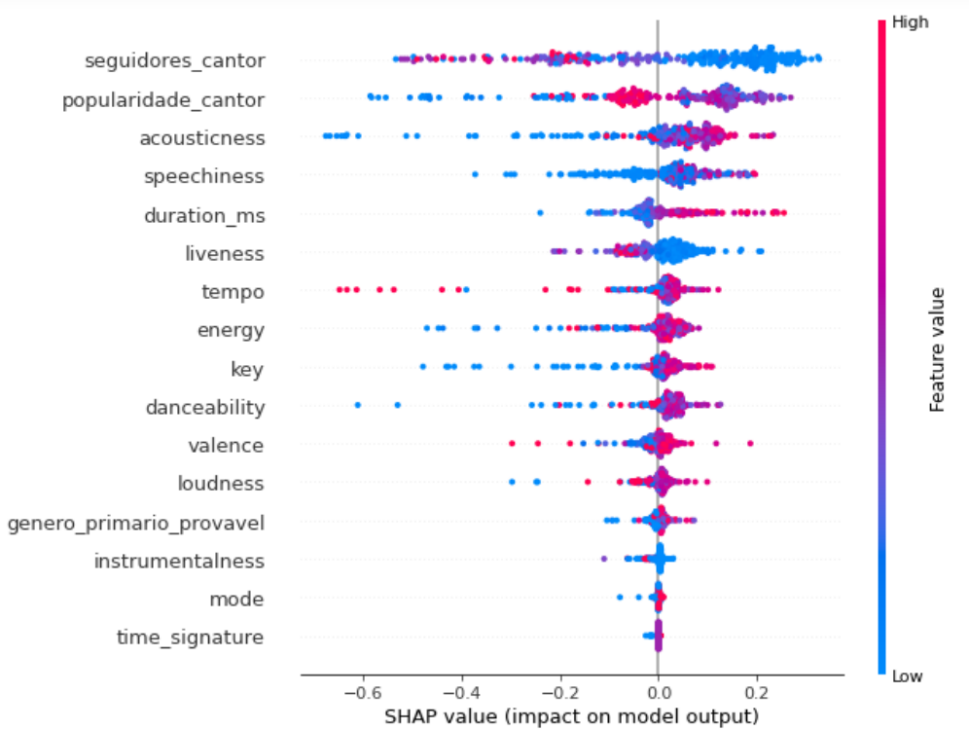
4.4 SHAP Feature Importance

Ao contrário das técnicas de Feature Importance e Permutation Importante, os valores de Shapley admitem números positivos e negativos, e, por meio da visualização dos dados, é possível extrair as seguintes informações:

1. Avaliar quais características surtem maior efeito global na popularidade de canções mensurando a distância dos pontos em relação á linha central do gráfico ou o valor absoluto do valor de Shapley;
2. Avaliar efeitos negativos ou positivos da popularidade de canções por meio da localização dos pontos em relação à linha central do gráfico, ou avaliando o sinal do valor de Shapley;
3. Avaliar a sensibilidade da variação das características sobre o efeito na popularidade das canções por meio da análise do efeito global em virtude da transição de cores referentes á grandeza das características.

A figura 2 abaixo apresenta os valores de Shapley para cada amostra dos dados de teste.

Figura 2 – Valores de Shapley vs Características



Fonte: Autoria própria.

O alto índice de popularidade de canções, portanto, pode ser explicado majoritariamente por:

- Número de seguidores do cantor na plataforma baixo;
- Índice de popularidade do cantor intermediário;
- Índice de acousticness elevado;
- Índice de speechiness elevado;
- Duração elevada;
- Índice de liveness baixo;
- Tempo entre intermediário á elevado.

Analogamente, baixos índices de popularidades de canções podem ser explicados majoritariamente por:

- Número de seguidores do cantor na plataforma entre intermediário e elevado;
- Índice de popularidade do cantor baixo ou alto;
- Índice de acousticness baixo;
- Índice de speechiness baixo;
- Duração baixo;
- Índice de liveness entre intermediário e elevado;
- Tempo elevado.

Tabela 6 – Estatística descritiva das variáveis

Caracteristica	Contagem	Média	DesvPad	Min	Max
key	277	5,5	3,5	0,0	11,0
mode	277	0,5	0,5	0,0	1,0
time_signature	277	4,0	0,3	1,0	5,0
genero_primario_provavel	277	16,0	13,7	0,0	41,0
danceability	277	0,7	0,1	0,0	1,0
energy	277	0,7	0,2	0,0	1,0
loudness	277	-5,5	2,5	-15,2	1,9
speechiness	277	0,1	0,1	0,0	0,8
acousticness	277	0,3	0,2	0,0	1,0
instrumentalness	277	0,0	0,0	0,0	0,6
liveness	277	0,2	0,2	0,0	1,0
valence	277	0,6	0,2	0,0	1,0
tempo	277	127,0	30,7	53,0	201,8
duration_ms	277	187998,5	48973,7	52062,0	421463,0
popularidade_cantor	277	70,2	12,7	31,0	95,0
seguidores_cantor	277	11509890	17008560	1455	99612040

Fonte: Autoria própria.

Para fins de consulta, a tabela 6 abaixo apresenta as estatísticas descritivas das características musicais e de artistas em dados de teste e expressa os seus graus de grandeza.

Foi observado que uma escala intermediária á elevada no número de seguidores do cantor na plataforma geram efeitos negativos na popularidade das canções, já número de seguidores baixo geram efeitos positivos.

Por outro lado, apesar de elevados índices de popularidade dos cantores também causarem efeitos negativos, índices intermediários causam efeitos positivos na popularidade das canções.

A fim de avaliar o efeito global das características musicais, o SHAP Feature Importance foi calculado a partir da média dos valores absolutos de Shapley e os resultados são apresentados na tabela 7.

Tabela 7 – Valores de SHAP Absolutos.

Característica	Valores
seguidores cantor	0,1894
popularidade cantor	0,14
acousticness	0,0991
speechiness	0,069
duration ms	0,0467
liveness	0,0447
tempo	0,043
energy	0,0415
key	0,0387
danceability	0,0385
valence	0,0217
loudness	0,0184
genero primario provavel	0,011
instrumentalness	0,0078
mode	0,0025
time signature	0,0009

Fonte: Autoria própria.

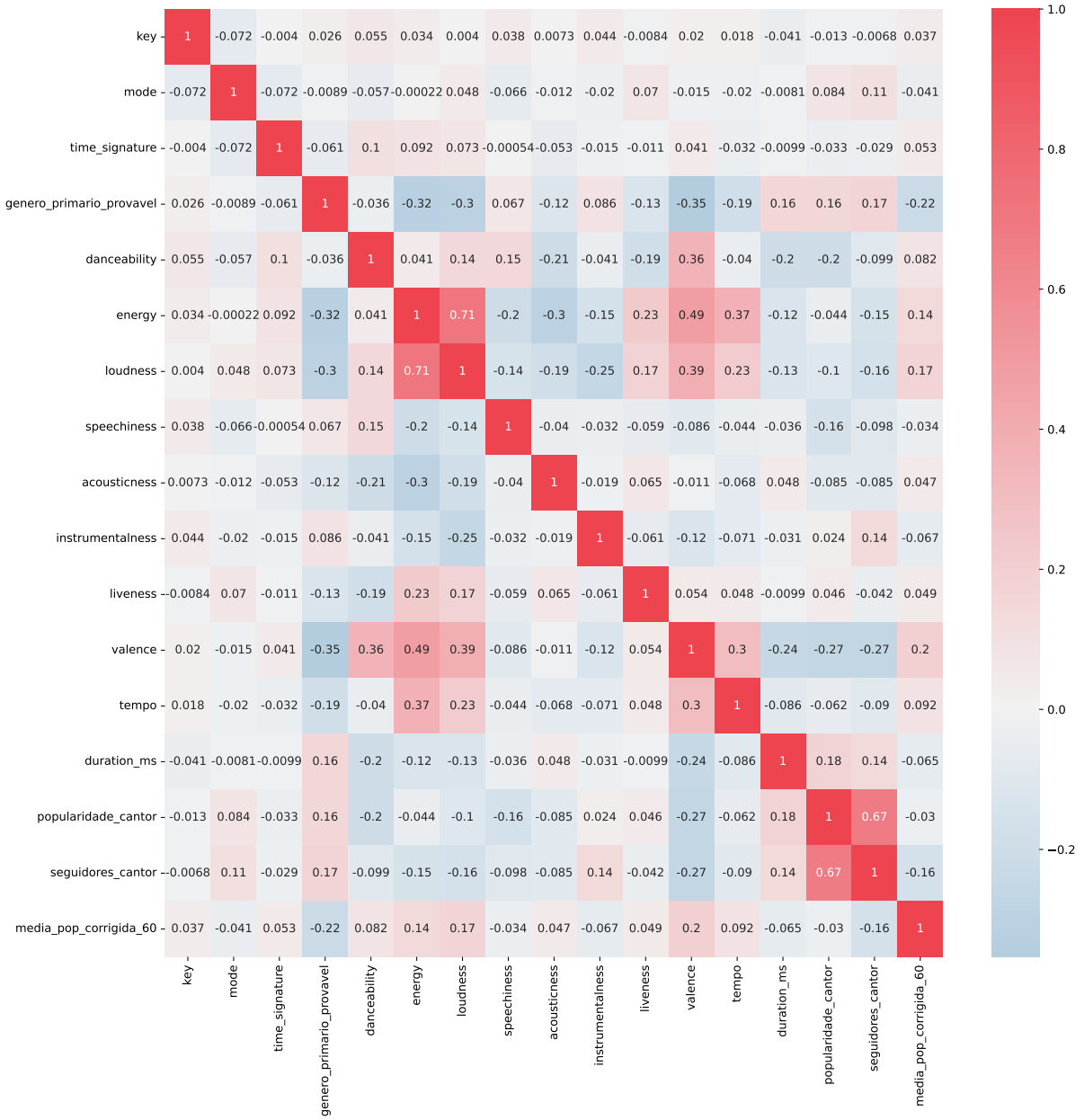
Por meio da interpretação do SHAP Feature Importance é possível mensurar o tamanho dos efeitos sobre a popularidade de canções, perdendo-se, contudo, a informação do sinal dos efeitos e a sensibilidade às características.

Dessa forma, compõem as 6 características mais importantes o número de seguidores, o índice de popularidade do cantor, acousticness, speechiness, duration e liveness.

4.5 Correlação de Pearson

A figura 3 abaixo apresenta os índices de correlação de Pearson para todos os dados coletados entre todas as variáveis, incluso a variável resposta do índice de popularidade calculado de canções identificada como media_pop_corrigida_60 no mapa de calor. É discutido na Seção 4.7 algumas implicações a respeito do impacto que características correlacionadas entre si podem gerar sobre a importância das características mensuradas por outras técnicas.

Figura 3 – Índice de Correlação de Pearson



Fonte: Autoria própria.

A tabela 8 abaixo apresenta os resultados da correlação das características com o índice de popularidade calculado de canções apenas.

Tabela 8 – Valores do índice de correlação de Pearson.

Característica	Valores
genero primario provavel	-0,2207
valence	0,2016
loudness	0,1676
seguidores cantor	-0,1556
energy	0,1415
tempo	0,0919
danceability	0,0818
instrumentalness	-0,0673
duration ms	-0,0652
time signature	0,0532
liveness	0,0488
acousticness	0,0474
mode	-0,0408
key	0,0368
speechiness	-0,0337
popularidade cantor	-0,0302

Fonte: Autoria própria.

A força da correlação das pode ser classificada conforme a tabela abaixo proposta por Schober, Boer e Schwarte (2018).

Tabela 9 – Tabela referencia de correlação de Pearson

Intervalo de correlação (p)	Força da correlação
0.00-0.10	Irrelevante
0.10–0.39	Fraca
0.40–0.69	Moderada
0.70–0.89	Forte
0.90–1.00	Muito forte

Fonte: Autoria própria.

Assim sendo, as características de gênero primário provável, valence, loudness, número de seguidores do cantor e energy compõe as 5 principais características e possuem correlação fraca com a variável alvo. Dessas variáveis, valence, energy e loudness possuem correlação positiva com a popularidade das canções e as variáveis genero primario provável e número de seguidores do cantor possuem correlação negativa. As demais características não possuem correlação relevante.

4.6 Information Value

A tabela 10 abaixo apresenta os valores de Information Value das características musicais e de artistas em dados de teste.

Tabela 10 – Valores de Information Value.

Característica	Valores
popularidade cantor	0,3167
key	0,2126
seguidores cantor	0,1501
speechiness	0,1137
genero primario provavel	0,1017
valence	0,0914
acousticness	0,0879
liveness	0,0851
loudness	0,0589
danceability	0,0438
duration ms	0,0348
energy	0,0201
mode	0,0012
time signature	0,0011
tempo	0,0009
instrumentalness	0

Fonte: Autoria própria.

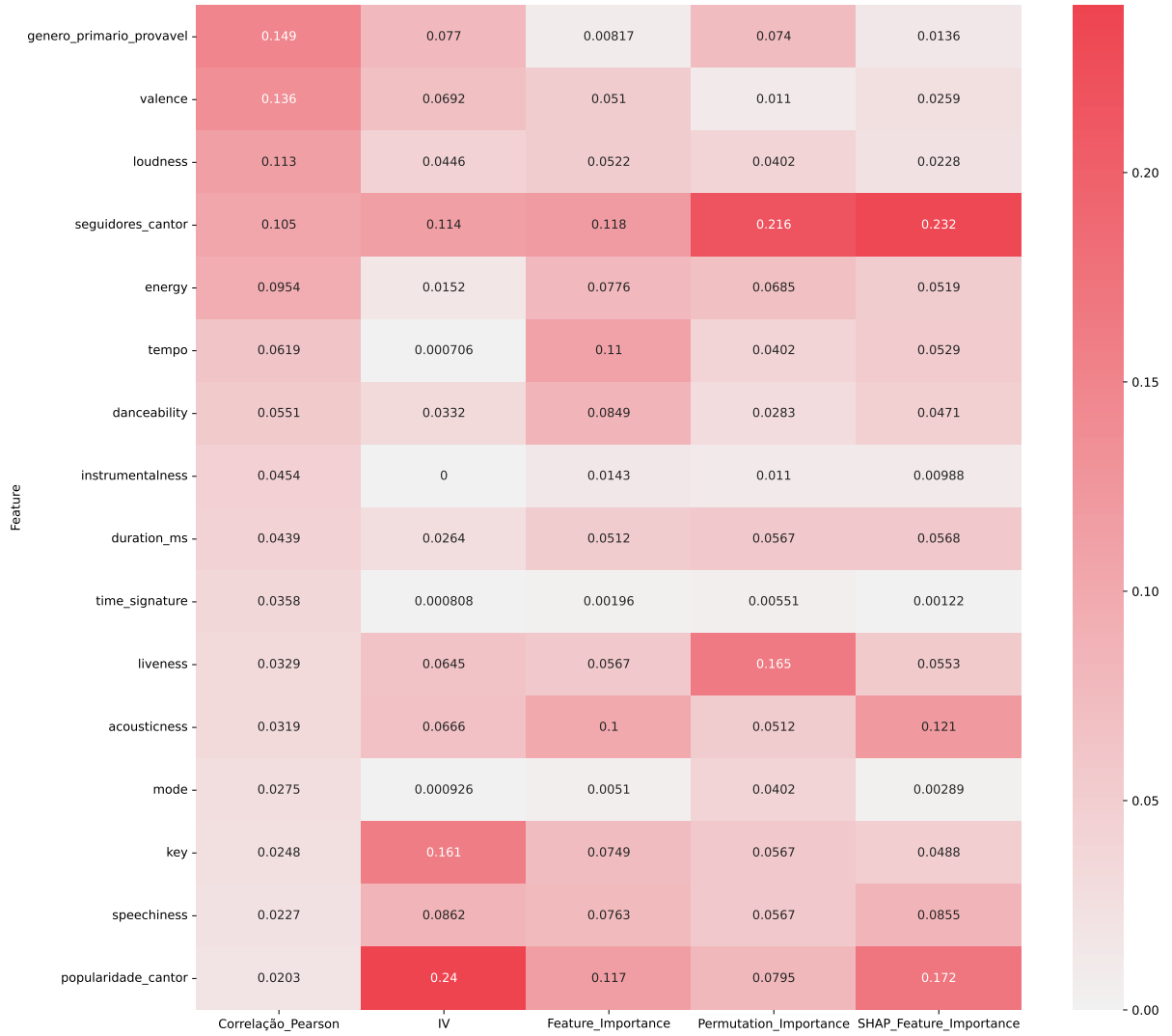
A popularidade do cantor, key, número de seguidores do cantor, speechiness, o gênero primário provável e valence são as características mais importantes e causam um maior efeito na variável resposta para essa técnica.

4.7 Análise comparativa

A fim de comparar a importância global das características sobre a popularidade das canções advindas das 5 técnicas, para cada técnica foi considerado o valor absoluto de cada importância calculando-se o módulo, e, em sequência, todas as importâncias foram colocadas numa mesma escala cuja soma resultasse em 1.

Os resultados foram plotados em um mapa de calor e podem ser observados na figura 4 abaixo:

Figura 4 – Análise comparativa de técnicas



Fonte: Autoria própria.

Por meio dessa visualização é possível verificar que as características referentes aos cantores são as mais importantes para determinação da popularidade de músicas. O número de seguidores do cantor, por exemplo, é a característica mais importante para as 3 técnicas de interpretabilidade do modelo de floresta aleatória, a terceira mais importante para técnica de Information Value e a quarta mais correlacionada linearmente com a popularidade das canções.

Já o índice de popularidade do cantor, é a segunda mais importante para as técnicas de interpretabilidade, com exceção da técnica Permutation Importance, em que ocupa a terceira posição. Além disso, é a característica mais importante para técnica de Information Value e, por fim, por outro lado, é a menos importante para o método de correlação de Pearson.

Visando ranquear as características, um sistema de votos foi proposto de forma que cada técnica foi responsável pela distribuição de 5 votos que correspondem às 5 características mais importantes em valores absolutos. Os resultados podem ser visualizados na tabela 11.

Tabela 11 – Top 5 Características musicais e de artistas mais importantes por técnicas

Característica	Pearson	IV	FI	PI	SHAP	Total Votos
seguidores_cantor	1	1	1	1	1	5
popularidade_cantor	0	1	1	1	1	4
genero_primario_provavel	1	1	0	1	0	3
acousticness	0	0	1	0	1	2
speechiness	0	1	0	0	1	2
energy	1	0	0	1	0	2
duration_ms	0	0	0	0	1	1
liveness	0	0	0	1	0	1
tempo	0	0	1	0	0	1
key	0	1	0	0	0	1
danceability	0	0	1	0	0	1
valence	1	0	0	0	0	1
loudness	1	0	0	0	0	1
instrumentalness	0	0	0	0	0	0
mode	0	0	0	0	0	0
time_signature	0	0	0	0	0	0

Fonte: Autoria própria.

Tabela 12 – Top 5 Características musicais mais importantes por técnicas

Característica	Pearson	IV	FI	PI	SHAP	Total Votos
speechiness	0	1	1	1	1	4
acousticness	0	1	1	0	1	3
liveness	0	1	0	1	1	3
tempo	1	0	1	0	1	3
energy	1	0	1	1	0	3
duration_ms	0	0	0	1	1	2
key	0	1	0	1	0	2
danceability	1	0	1	0	0	2
valence	1	1	0	0	0	2
loudness	1	0	0	0	0	1
instrumentalness	0	0	0	0	0	0
mode	0	0	0	0	0	0
time_signature	0	0	0	0	0	0

Fonte: Autoria própria.

As principais características, conforme discutido anteriormente, estão ligadas diretamente aos artistas e são em ordem de ranqueamento: o número de seguidores do cantor com 5 votos, a popularidade do cantor com 4 votos e o gênero primário provável da canção com 3 votos.

A tabela 12 apresenta os resultados ao reprocessar o sistema de votações, excluindo-se as características referentes aos artistas. As cinco características musicais mais importantes para determinação da popularidade da canção é em ordem de ranqueamento: speechiness com

4 votos, seguidas empatadas em segundo lugar por acousticness, liveness, tempo e energy com 3 votos cada.

A interpretação do gráfico dos valores de Shapley apresentado na figura 2, permite concluir que, considerando os intervalos das características nos dados de testes apresentados na tabela 6, valores elevados de speechiness, acousticness, tempo, energy e valores baixos de liveness favorecem canções mais populares.

5 CONSIDERAÇÕES FINAIS

Foi levantado como justificativa para esse trabalho o potencial em orientar escolhas de repertório ou ajustes musicais que aumentem o potencial de popularidade de canções. Avaliar os efeitos que as características musicais e de artistas causam na popularidade, demonstrou ser importante e aderente a esse propósito.

As características musicais e de artistas mais importantes para determinação da popularidade de canções nos charts do Spotify brasileiro no ano de 2021 foram avaliados, atingindo-se, portanto, o objetivo geral desse trabalho.

Para isso, recorreu-se ao modelo de aprendizado de máquina de floresta aleatória e métodos nativos da biblioteca Scikit-Learn para determinação da importância das características com Feature Importance. Também foram aplicadas técnicas de interpretabilidade como o Permutation Importance e o SHAP Feature Importance, além de outros métodos não dependentes de modelos de aprendizado de máquina como a correlação de Pearson e Information Value.

Os resultados convergem e apontam maior importância às características referentes aos artistas para a popularidade das canções, sendo o número de seguidores do cantor, o índice de popularidade do cantor e o gênero primário provável as três mais importantes, seguidas pelas características referentes às músicas como a fração de palavras faladas na canção (speechiness), acusticidade, energia, tempo e o índice relativo de música ao vivo (liveness).

Dessa forma, a hipótese levantada é confirmada, pois foi possível mensurar a importância das características por meio da avaliação dos resultados obtidos através da aplicação das técnicas de interpretabilidade de modelo de aprendizado de máquina, e também, de outras técnicas não dependentes do modelo.

5.1 Limitações

Foram identificadas limitações referentes tanto à abordagem do problema que foi arbitrada pelo próprio autor quanto às limitações inerentes às técnicas utilizadas.

É sabido que, os erros relacionados a importância de características advindas das técnicas de interpretabilidade é indissociável ao erro do modelo. Além disso, o modelo de floresta aleatória foi arbitrariamente escolhido, pois, permitia a utilização das características selecionadas pelas árvores de decisões para o levantamento da importância de características por meio da aplicação da técnica de Feature Importance. Assim sendo, não foram utilizados, nesse trabalho, outros modelos de aprendizado de máquina que, eventualmente, poderiam apresentar melhores resultados.

Além disso, há limitações referentes às técnicas utilizadas. Por exemplo, a correlação de Pearson é indicada apenas para capturar correlações lineares, não sendo recomendada para mensurar correlações não lineares. Com relação à técnica Permutation Importance, não foi

avaliado eventuais ocorrências de divisão das importâncias entre duas características correlações entre si.

O algorítimo gerador das características musicais que foram analisadas nesse trabalho é de propriedade do Spotify. Para ter acesso a essas características é necessário que a música esteja, portanto, disponibilizada e distribuída na plataforma. Assim sendo, há limitações quanto ao uso indiscriminado do algoritmo para orientar as escolhas de repertório musical, sendo necessário, portanto, solicitar permissão á plataforma para o seu uso.

Outra limitação referente a esse trabalho foi na dificuldade de capturar o número de reproduções em dias que as canções não atingiram o top 200 nas charts do Spotify. Esse número pode, eventualmente, compor o cálculo do índice de popularidade de canções caso ocorra entre a data da primeira aparição nas charts e 60 dias após e isso pode afetar a rotulação de músicas populares e comuns, alterando o dataset final. Para contornar esse problema, foi atribuído o número de reproduções igual á zero nesse cenário.

5.2 Trabalhos Futuros

Sugere-se para aprimoramento desse trabalho a exploração dos seguintes tópicos:

- Avaliação de outras técnicas de aprendizado de máquina;
- Utilização de bibliotecas especializadas para processamento de áudio independentes da plataforma do Spotify para geração das características musicais;
- Exploração de outras definições de popularidade de canções que facilitam o aprendizado pelos modelos de aprendizado de máquina;
- Avaliação de características segmentada por gêneros musicais e por popularidade de cantores.

Referências

ARAUJO, C.; CRISTO, M.; GIUSTI, R. Will i remain popular? a study case on spotify. In: SBC. **Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2019. p. 599–610. Citado na página 11.

ARAÚJO, L. T.; OLIVEIRA, C. N. Música em fluxo: experiências de consumo musical em serviços de streaming. **Revista Temática**, v. 10, n. 10, 2014. Citado na página 11.

BHALLA, D. Weight of evidence (woe) and information value explained. **URL: https://www. listendata. com/2015/03/weight-of-evidence-woe-and-information. html (: 15.08. 2018)**, 2016. Citado na página 15.

BISCHL, B. et al. Introduction to machine learning (i2ml). **URL: https://slds-lmu.github.io/i2ml/chapters/07_forests/07-04-featureimportance/**, 2022. Citado na página 14.

CUSTÓDIO, K. C. d. L. et al. Padrões na preferência musical dos (as) brasileiros (as) sob a ótica do spotify. Universidade Federal de Campina Grande, 2021. Citado na página 11.

EGAN, C. **Improving Credit Default Prediction Using Explainable AI**. Tese (Doutorado) — Dublin, National College of Ireland, 2021. Citado na página 15.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**. [S.l.]: "O'Reilly Media, Inc.", 2019. Citado 2 vezes nas páginas 13 e 15.

KLEINA, N. **A história do Spotify e a revolução do streaming na música**. [S.l.], 2018. Disponível em: <https://www.tecmundo.com.br/mercado/131633-historia-spotify-revolucao-do-streaming-musica-video.htm>. Acesso em: 09 de novembro de 2022. Citado na página 11.

MIDDLEBROOK, K.; SHEIK, K. Song hit prediction: Predicting billboard hits using spotify data. **arXiv preprint arXiv:1908.08609**, 2019. Citado 2 vezes nas páginas 11 e 12.

MOLNAR, C. A guide for making black box models explainable. **URL: https://christophm. github. io/interpretable-ml-book**, 2022. Citado na página 14.

PETKOVIC, D. et al. Random forest model and sample explainer for non-experts in machine learning–two case studies. In: SPRINGER. **International Conference on Pattern Recognition**. [S.l.], 2021. p. 62–75. Citado na página 13.

ROGERS, J.; GUNN, S. Identifying feature relevance using a random forest. In: SPRINGER. **International Statistical and Optimization Perspectives Workshop"Subspace, Latent Structure and Feature Selection"**. [S.l.], 2005. p. 173–184. Citado na página 13.

SCHOBER, P.; BOER, C.; SCHWARTE, L. A. Correlation coefficients: appropriate use and interpretation. **Anesthesia & Analgesia**, Wolters Kluwer, v. 126, n. 5, p. 1763–1768, 2018. Citado na página 30.

SILVA, R. A. da. From gramophone to live streaming: the evolution of the modes of listening to music-some implications. **Revista da Tulha**, v. 1, n. 1, p. 251–263, 2015. Citado na página 11.

SPOTIFY. Spotify web api reference. **URL: https://developer.spotify.com/documentation/web-api/**, 2022. Citado na página 17.

ŠTRUMBELJ, E.; KONONENKO, I. Explaining prediction models and individual predictions with feature contributions. **Knowledge and information systems**, Springer, v. 41, n. 3, p. 647–665, 2014. Citado na página 14.

APÊNDICE A - Algorítimo Feature Importance para florestas aleatórias

Algoritmo 1: Random Forest Feature Importance

Input: Test Dataset
Output: Feature Importance
for *features* x_j, $j = 1$ *to* p **do**
    for *tree base learners* $b^{[m]}(x)$, $m = 1$ *to* M **do**
        Find all nodes N in $b^{[m]}(x)$ *that use* x_j.
        Compute improvement in splitting criterion achieved by them.
        Add up these improvements.
    end
    Add up improvements over all trees to get feature importance of x_j
end