UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

CIBELE OLIVEIRA FERREIRA
Orientador: Prof. Dr. Guilherme Tavares de Assis

PROPOSTA E DESENVOLVIMENTO DE UMA ESTRATÉGIA DE PREDIÇÃO DE SUCESSO MUSICAL BASEADA EM CARACTERÍSTICAS ACÚSTICAS E GÊNERO MUSICAL

Ouro Preto, MG
2024

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

CIBELE OLIVEIRA FERREIRA

PROPOSTA E DESENVOLVIMENTO DE UMA ESTRATÉGIA DE PREDIÇÃO DE SUCESSO MUSICAL BASEADA EM CARACTERÍSTICAS ACÚSTICAS E GÊNERO MUSICAL

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Guilherme Tavares de Assis

Ouro Preto, MG
2024

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Cibele Oliveira Ferreira

Proposta e desenvolvimento de uma estratégia de predição de sucesso musical baseada em características acústicas e gênero musical

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 9 de Fevereiro de 2024.

Membros da banca:

Guilherme Tavares de Assis (Orientador) - Doutor - Universidade Federal de Ouro Preto
Jadson Castro Gertrudes (Examinador) - Doutor - Universidade Federal de Ouro Preto
Marcelo Luiz Silva (Examinador) - Mestre - Universidade Federal de Ouro Preto

Guilherme Tavares de Assis, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 9/02/2024.

Documento assinado eletronicamente por **Guilherme Tavares de Assis**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 15/02/2024, às 17:10, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0665727** e o código CRC **06301626**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho a todos aqueles que encontram na música a chave que abre as portas de suas emoções mais profundas.

Agradecimentos

Agradeço à todos da minha família que me acompanharam e torceram por mim nessa trajetória.

Agradeço ao meu orientador, Guilherme Tavares de Assis, por sua paciência e sensibilidade terem me ajudado a transformar todo o processo em realidade.

Por fim, agradeço aos meus amigos por me apoiarem e serem pacientes comigo, obrigada por sempre estarem presentes.

"Se a vida fere com a sensação do brilho.

De repente a gente brilhará."

– Gilberto Gil

Resumo

O avanço da tecnologia tem mudado a forma como a música é consumida como, por exemplo, a mudança das mídias físicas para os serviços de *streaming*. Essa mudança gerou grandes conjuntos de dados que podem ser explorados por meio de técnicas de mineração de dados, que podem auxiliar a compreensão das tendências e padrões das características musicais associadas à popularidade e podem extrair informações de grandes conjuntos de dados de gravações musicais, incluindo recursos como tempo, ritmo, melodia e harmonia. A pesquisa nessa área é praticamente inexistente no Brasil, mas pode trazer benefícios para a indústria musical, considerando a pouca quantidade de trabalhos registrados nesse campo. Desta forma, o objetivo geral deste trabalho é propor, desenvolver e validar uma estratégia para prever a popularidade de músicas com base em características acústicas e gêneros musicais. Para alcançá-lo, uma arquitetura de funcionamento em quatro etapas foi definida, a saber: (a) coleta dos dados por meio do consumo da API do *Spotify*; (b) pré-processamento desses dados; (c) identificação de padrões e tendências nos dados pré-processados; e (d) utilização de modelos de predição para prever o percentual de popularidade de músicas no conjunto de teste. Experimentações práticas, envolvendo dois bancos de dados, geraram resultados satisfatórios quanto às análises realizadas; considerando dados de gêneros musicais puros brasileiros, os resultados destacam desafios na generalização para esses gêneros mais específicos. Os modelos *Ridge Regression* e *Huber Regressor* apresentaram melhor capacidade de generalização (R^2 de aproximadamente 68.55%) sendo posteriormente reavaliados com a seleção de características relevantes, contribuindo para a compreensão e previsão da popularidade musical, com potencial aplicação na indústria musical brasileira.

Palavras-chave: Recuperação de informação musical. Predição de popularidade musical. Características acústicas. Gênero musical. Mineração de dados.

Abstract

The advancement of technology has changed the way music is consumed, such as the shift from physical media to streaming services. This change has generated large datasets that can be explored through data mining techniques, aiding in understanding trends and patterns in musical features associated with popularity. These techniques can extract information from extensive datasets of music recordings, including features like tempo, rhythm, melody, and harmony. Research in this area is virtually nonexistent in Brazil, but it could bring benefits to the music industry given the scarcity of documented work in this field. Therefore, the overall objective of this study is to propose, develop, and validate a strategy for predicting the popularity of songs based on acoustic features and music genres. To achieve this, a four-stage operating architecture was defined, namely: (a) data collection through the consumption of the Spotify API; (b) preprocessing of this data; (c) identification of patterns and trends in the preprocessed data; and (d) use of prediction models to forecast the popularity percentage of songs in the test set. Practical experiments involving two databases yielded satisfactory results for the conducted analyses. Considering data from pure Brazilian music genres, the results highlight challenges in generalizing to these more specific genres. The Ridge Regression and Huber Regressor models demonstrated better generalization capabilities (R^2 of approximately 68.55%), subsequently reassessed with the selection of relevant features, contributing to the understanding and prediction of musical popularity. This has potential applications in the Brazilian music industry.

Keywords: Musical information retrieval. Prediction of musical popularity. Acoustic features. Musical genre. Data mining.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

UFOP	Universidade Federal de Ouro Preto
RIAA	Recording Industry Association of America
RIM	Recuperação da Informação Musical
KDD	Knowledge Discovery in Databases
API	Application Programming Interface
KNN	K-Nearest Neighbors

Sumário

1 Introdução

Com o avanço tecnológico, ocorreram mudanças na forma como as pessoas acessam e consomem músicas, filmes e outras mídias. De acordo com Panda et al. (2021), o consumo de música, que antes girava em torno de mídias físicas como fita cassete ou vinil, assumiu uma propriedade por meio de serviços de *streaming* que fornecem acesso a milhões de faixas musicais.

Segundo a *Recording Industry Association of America* (RIAA), em 2018, a indústria musical movimentou US$ 9,8 bilhões somente nos Estados Unidos. A RIAA também destaca que o *streaming* é a principal fonte de receita para a indústria musical nos Estados Unidos, o que mostra a importância da oferta de música por meio desse serviço. A música é uma parte integral da nossa vida: não se pode negar sua importância, dada sua onipresença, no nosso convívio social e ambiente físico; dito isto, os recursos de *streaming* são frequentemente procurados por pessoas que buscam ambientar-se como seres que sentem, pensam e agem. Ademais, existem poucas técnicas efetivas de organização e recuperação de músicas; portanto, se explorados, estes dados podem descobrir relações com vários fenômenos econômicos, sociais, tecnológicos e culturais.

Com o movimento crescente da indústria, realizar estudos focados em investigar a popularidade das músicas tornou-se cada vez mais comum, segundo Interiano et al. (2018). A mineração de dados pode ser uma grande aliada durante o processo de análise de dados, visto que há uma quantidade cada vez maior de informações acessíveis em decorrência do consumo de música digital.

A análise da popularidade musical com base nas características acústicas e no gênero musical pode fornecer informações valiosas sobre os fatores que contribuem para o sucesso de diferentes músicas. Técnicas de extração de dados podem ser usadas para obter informações de grandes conjuntos de dados de gravações musicais, incluindo recursos como tempo, ritmo, melodia e harmonia. Os resultados dessa extração podem ser comparados em diferentes gêneros musicais para identificar tendências e padrões nas características musicais associadas à popularidade.

Aplicada no contexto apresentado, uma metodologia de modelos de regressão pode ser adotada para antecipar a popularidade de uma música, considerando suas características acústicas e o gênero musical associado. Ela é motivada por uma variedade de fatores, permitindo a avaliação da relevância de diferentes recursos para diferentes gêneros musicais. Ao examinar possíveis resultados obtidos por essa metodologia, é possível obter uma compreensão mais precisa e detalhada das características que contribuem para a popularidade dentro de cada gênero musical. Em outras palavras, essa metodologia pode ser empregada para embasar decisões relacionadas à produção e promoção de músicas em diversos contextos musicais.

Vale ressaltar que pesquisas na área de recuperação da informação musical no Brasil são

praticamente inexistentes na literatura. Por um lado, o tema é relevante e inovador, conforme aponta Santini e Souza (2007) mas, por outro lado, os pesquisadores enfrentam grandes desafios durante o processo de pesquisa, apesar de ser uma parte importante da cultura brasileira.

Este capítulo encontra-se organizado como se segue. A Seção 1.1 apresenta a motivação para a realização desse trabalho. A Seção 1.2 descreve os objetivos geral e específicos. A Seção 1.3 aborda o método utilizado no desenvolvimento desse trabalho. Finalmente, a Seção 1.4 apresenta o delineamento do restante da monografia.

1.1 Justificativa

Atualmente, poucos trabalhos foram registrados na área de recuperação da informação musical no Brasil. É importante que haja mais engajamento em pesquisas nesta área para que se possa avançar e compreender como a tecnologia pode ser utilizada para melhorar a experiência musical e promover a indústria musical no Brasil. Este campo de pesquisa é uma área que necessita de novas técnicas de recuperação, dado o crescente volume de música digital disponível na *internet* brasileira.

Os conhecimentos adquiridos nesta área podem fornecer benefícios para todas as partes envolvidas no ciclo de vida de uma música. A capacidade de prever com precisão a popularidade de uma música pode ser útil em vários seguimentos como, por exemplo, se uma análise for realizada em uma música prestes a lançar; desta forma, é possível determinar o impacto antes do lançamento da obra.

Diante dessas ponderações, estratégias ou abordagens, que podem analisar características acústicas e prever a popularidade musical de uma determinada música, terão grande relevância para a área de recuperação de informação musical.

1.2 Objetivos Geral e Específicos

O objetivo geral desse trabalho consiste em propor, desenvolver e validar uma estratégia para predição da popularidade musical, baseada em características acústicas e gêneros musicais das mesmas. Essa estratégia engloba, em sua proposta, a análise de como caracteristicas acústicas, sob a perspectiva de gênero musical, podem influenciar no sucesso de uma música.

De um modo geral, os principais objetivos específicos, alcançados neste trabalho, foram:

- identificação de padrões e tendências relacionados a popularidade musical por meio da análise de características acústicas e gêneros musicais;
- avaliação do desempenho de diferentes modelos de predição;
- validação da aplicabilidade da estratégia proposta no contexto da indústria musical.

1.3 Método de Trabalho

Visando o alcance do objetivo geral deste trabalho, foi definida e elaborada uma arquitetura que descreve o funcionamento de uma estratégia para a predição de popularidade musical. Essa estratégia inclui etapas como coleta e pré-processamento de dados, utilização de modelos de aprendizado de máquina, treinamento e avaliação dos mesmo. Baseada em tal arquitetura, a estratégia adotada envolveu a limpeza e organização dos dados coletados, eliminando valores ausentes e aberrantes, e transformando os dados em um formato adequado para análise. Em fases subsequentes, foram aplicadas técnicas de mineração de dados para extrair informações pertinentes dos dados selecionados.

Para cada fase experimental, que envolve a aplicação de modelos de aprendizado de máquina em duas bases de dados distintas, avaliou-se o desempenho por meio de métricas específicas. Inicialmente, quatro modelos foram submetidos a esses experimentos, sendo que os mais promissores foram selecionados para uma análise mais aprofundada. Durante essas avaliações, foram consideradas métricas pertinentes à regressão, como o coeficiente de determinação (R^2) e o erro médio absoluto (MAE), para mensurar a eficácia de cada modelo. Com base nos resultados obtidos, realizou-se uma seleção de características, visando otimizar o desempenho dos modelos escolhidos. Essa abordagem proporcionou uma compreensão mais profunda das características que impactam significativamente a precisão da predição de popularidade musical.

1.4 Organização do Trabalho

O restante desta monografia encontra-se organizado como se segue. O Capítulo 2 apresenta a revisão de literatura necessária para a realização deste trabalho, envolvendo fundamentação teórica e trabalhos diretamente relacionados. O Capítulo 3 descreve a estratégia proposta neste trabalho, envolvendo suas características e arquitetura de funcionamento. No Capítulo 4, são descritos os experimentos realizados, quanto à utilização da estratégia de predição de popularidade musical desenvolvida, e são apresentados e discutidos os resultados experimentais obtidos. Por fim, o Capítulo 5 apresenta conclusões deste trabalho e as perspectivas de trabalho futuro.

2 Revisão Bibliográfica

Este capítulo apresenta a revisão bibliográfica feita para a realização deste trabalho. Para tanto, encontra-se organizado da seguinte maneira: a Seção 2.1 aborda a fundamentação teórica necessária ao desenvolvimento deste trabalho e a Seção 2.2 apresenta trabalhos diretamente relacionados.

2.1 Fundamentação Teórica

Esta seção tem, como objetivo, apresentar conceitos relevantes para a fundamentação e a construção da proposta deste trabalho, bem como contribuir para um melhor entendimento da metodologia aplicada. Os assuntos abordados estão dispostos da seguinte forma: a Subseção 2.1.1 discorre sobre a recuperação da informação musical, a Subseção 2.1.2 aborda o serviço de *streaming Spotify*, a Subseção 2.1.3 fala sobre o processo de descoberta de conhecimento a partir de dados e a Subseção 2.1.4 envolve todos os modelos de regressão essenciais para a compreensão da estratégia.

2.1.1 Recuperação da Informação Musical

Métodos convencionais de recuperação de informação, que lidam com armazenamento automático e recuperação de documentos, a partir de linguagem natural ou controlada, geralmente estão associados a objetos de dados compostos por textos. Os documentos consistem em palavras como representação simbólica de determinado assunto e os mecanismos de busca para recuperá-los são construídos em uma estrutura de definições e das relações entre as palavras dos documentos.

Conforme mencionado em McLane (1996), a informação musical não possui em sua premissa um assunto intrínseco ou o significado dos termos utilizados; entretanto, por meio da análise de seu conteúdo, diferentes perspectivas podem ser captadas de um mesmo documento musical, exigindo a necessidade de representação de diferentes pontos da mesma obra. Segundo Cruz (2008), a estrutura da música incorpora elementos adicionais que permitem defini-la como um objeto informacional musical mais amplo, dotado de conteúdo (atributos internos e metadados descritivos) e de contexto (associações com outros objetos musicais e não musicais e com situações ou eventos em que este objeto musical está inserido).

A Recuperação da Informação Musical (RIM) é uma área multidisciplinar de pesquisa que, de forma geral, pretende desenvolver formas de gestão de coleções de obras musicais para preservação, busca, acesso e uso Santini e Souza (2007). O crescimento pela pesquisa sobre a RIM está vinculado à explosão do desenvolvimento de coleções em rede, com formatos de

compressão da informação musical e custos decrescentes do armazenamento digital e da conexão banda-larga Futrelle e Downie (2002).

Para a classificação de uma obra musical, segundo McLane (1996), é necessário considerar três perspectivas sobre a representação da música, a saber:

a) Visão Subjetiva: esta visão consiste no uso do esquema de notação para representar a obra musical e a informação bibliográfica. A mesma nota ou sequência de notas pode ser representada de diferentes maneiras na partitura, sendo essa representação uma inferência do profissional.

b) Visão Objetiva: esta visão pode ser considerada a mais completa representação da música, na medida em que inclui as seguintes características: tom, tempo, harmonia, editorial e timbre. O som musical é objetivo; uma vez gravado, a representação da música por meio da gravação é definitiva e não mais sujeita às variações editoriais e de performance.

c) Visão Interpretativa: esta visão é definida pela classificação e por esquemas analíticos que elucidam características (não tão claras) de uma obra musical; ou seja, revisões musicais, que são avaliadas de maneira rigorosa, fazem parte da visão interpretativa.

De acordo com McLane (1996), "qualquer representação da música irá consistir em uma ou mais destas três visões", dependendo das necessidades de informação do usuário. Portanto, as estratégias de desenvolvimento e gestão de coleções devem considerar a necessidade e a linguagem do público, mas sem desconsiderar o contexto de produção dos documentos.

2.1.2 *Spotify*

O *Spotify* surgiu, em 2006, como uma solução tecnológica para a distribuição de conteúdos baseados numa plataforma *peer-to-peer* Fleischer e Snickars (2017). Hoje, com 433 milhões de usuários, o *Spotify* é o serviço de subscrição de *streaming musical* mais popular do mundo Fleischer e Snickars (2017), impulsionando descoberta e engajamento e contribuindo com uma média de US$ 20 por usuário para a indústria da música (ao contrário dos menos de US$ 1 do *Youtube*), sendo considerado, por muitos, como o salvador da indústria musical Ellis-Petersen (2016).

O *Spotify* fornece dados crescentes sobre o comportamento de seus ouvintes e sobre o conteúdo musical de faixas, de forma fácil e legalmente aceita. Neste contexto, os dados registrados pelo *Spotify* apresentam vantagens sobre outros tipos de registros digitais relacionados ao comportamento humano. Primeiro, escutar música é um comportamento intrínseco e, portanto, tem o potencial de capturar informações sutis sobre preferências pessoais de seus ouvintes. Segundo, a música induz e comunica emoções e ativa regiões cerebrais ligadas à emoção e à criatividade Juslin e Laukka (2003). E, por último, escutar música abrange escalas de tempo, às

vezes estendendo-se por toda a atividade diária de uma pessoa, segundo North, Hargreaves e Krause (2009), capturando uma imagem mais completa das atividades e rotinas diárias de uma pessoa.

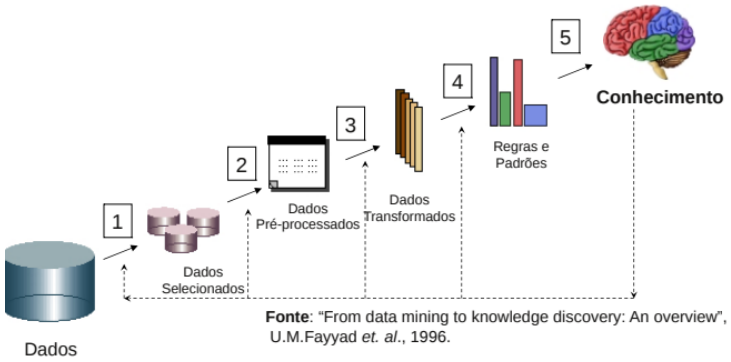
No intuito de recomendar conteúdo personalizado para cada usuário, tradicionalmente, o *Spotify* usa uma técnica para entender o gosto musical de um usuário específico com base no histórico de músicas escutadas Johnson (2014). Essas técnicas são independentes de conteúdo, contando apenas com os padrões de consumo dos usuários. Ademais, o *Spotify* inclui fontes adicionais de informações, impulsionado pela aquisição do *The Echo Nest* (um serviço de inteligência musical que fornece extração de dados de músicas por rastreamento na *web*), em 2014, sendo capaz de estimar a dançabilidade de uma música ou sua valência.

Neste trabalho, o *Spotify* é utilizado como base de dados da estratégia proposta. Ele fornece conjuntos de dados precisos e atualizados sobre as músicas e contribui com pesquisas em diversas áreas, desde recomendação até modelagem de estratégias voltadas à criação ou à descoberta musical.

2.1.3 Processo de Descoberta de Conhecimento

A evolução de recursos computacionais ocorrida nos últimos anos tem provocado a necessidade de técnicas e ferramentas capazes de lidar com os grandes volumes de dados gerados. O processo de descoberta de conhecimento em base de dados, também conhecido como *Knowledge Discovery in Databases* (KDD), segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), é um processo não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. A Figura 2.1 ilustra as etapas associadas ao processo de descoberta de conhecimento, considerando a definição apresentada.

Figura 2.1 – Processo de descoberta de conhecimento em base de dados.



Fonte: *From data mining to knowledge discovery: An overview*, U.M.Fayyad *et. al.*, 1996.

De acordo com a Figura 2.1, o processo de descoberta de conhecimento possui as seguintes etapas: (a) seleção de dados, que seleciona os dados que possuem uma certa relevância para a análise de descoberta; (b) pré-processamento, que inclui a limpeza, correspondente à remoção

de ruídos e dados inconsistentes, e a integração de dados, que combina dados obtidos de diversas fontes; (c) transformação de dados, que transforma os dados em formatos apropriados para a mineração; (d) mineração dos dados, que consiste na aplicação de métodos capazes de extrair novas informações a partir dos dados pré-processados; (e) e geração de conhecimento, que consiste na identificação de padrões úteis, com base nos resultados obtidos, para a validação.

Relativo ao processo de descoberta de conhecimento, particularmente, a Subseção 2.1.3.1 descreve a extração de dados, que é um processo intrínseco da etapa de pré-processamento e transformação dos dados, e a Subseção 2.1.3.2 aborda a mineração de dados.

2.1.3.1 Extração de dados

A extração de dados é o processo de encontrar, em um grande volume de dados, informações com características específicas, conforme descrito por Alvarez (2007). Essa tarefa, em grandes bases de dados, pode ocasionar um aumento na chance de encontrar padrões pouco significativos e até mesmo espúrios. Além disso, esse grande volume de dados pode apresentar uma estruturação na organização dos dados, mas também pode ser totalmente desestruturado.

Para Silva, Barros e Prudêncio (2005), os textos estruturados seguem um formato inflexível e, por terem uma estrutura familiar e imutável, a extração pode ocorrer com facilidade a partir de regras baseadas em delimitadores da linguagem ou da ocorrência de termos. No caso dos textos não estruturados, esse padrão não é considerado em sua formação visto que as sentenças podem ser descritas em alguma linguagem natural, inviabilizando a extração com base apenas na formatação.

O objetivo das técnicas de extração de dados é a construção de sistemas que consigam encontrar, a partir da combinação de padrões selecionados, informações relevantes Cowie e Lehnert (1996). Os autores explicam que, pelo fato da extração de informação ser um processo de selecionar estruturas e combinar dados encontrados em textos, a sua produção final, uma vez estruturada, é um procedimento com o claro objetivo de criação de um repositório ou de um banco de dados.

Neste trabalho, a estratégia de extração de dados selecionada é o consumo da *Application Programming Interface* (API) do *Spotify*. Esta técnica envolve a coleta de dados relevantes por meio de chamadas automatizadas à API do serviço.

2.1.3.2 Mineração de dados

A Mineração de Dados ou *Data Mining* trata-se da exploração de grandes quantidades de dados em busca de padrões, regras associadas ou sequências temporais, para extrair um significado, sendo usado tanto para descrever características do passado como predizer tendências para o futuro Corrêa e Sferra (2003).

Neste contexto, o principal objetivo da mineração de dados, segundo Dias et al. (2001), é

fornecer subsídios para que, a partir de um histórico, sejam feitas previsões de tendências futuras e seja descoberta qual é a relação entre os dados. Os autores discorrem que os resultados da mineração de dados podem ser utilizados no gerenciamento de informação, tomada de decisão, controle de processos, dentre outras aplicações.

Para que a descoberta de conhecimento seja relevante, metas bem definidas devem ser estabelecidas, permitindo assim um entendimento claro dos dados que originaram o conhecimento produzido. As técnicas de mineração de dados podem ser aplicadas a tarefas; essas técnicas vão desde as tradicionais da estatística multivariada até modelos mais atuais de aprendizagem.

Algumas tarefas realizadas por técnicas de mineração de dados, para Dias et al. (2001), podem ser descritas como: (a) classificação, onde há a construção de um modelo em que possa ser aplicado a dados não classificados a fim de categorizá-los em classes; (b) estimativa, onde se define um valor para alguma variável contínua desconhecida; (C) associação, onde se determina relações entre campos de uma mesma transação;(d) agrupamento, onde há o processo de partição de uma população em vários subgrupos ou grupos baseados em medidas de similaridade ou modelos probabilísticos; e (e) sumarização, onde técnicas, para se encontrar uma descrição compacta para um subconjunto de dados, são definidas e aplicadas.

A estratégia de mineração de dados utilizada na predição de popularidade musical deste trabalho envolve a aplicação de vários modelos de análise de regressão, que buscam estimar a relação entre uma variável dependente (no caso, a popularidade da música) e uma ou mais variáveis independentes (como por exemplo, características acústicas). Nas próximas subseções, são apresentados os modelos de regressão que foram utilizados na abordagem de predição de popularidade musical.

2.1.4 Modelos de regressão

Esta Subseção apresenta diferentes modelos de regressão relevantes para a compreensão da estratégia proposta. Portanto, as Subseções 2.1.4.1, 2.1.4.2, 2.1.4.3, 2.1.4.4 descrevem, respectivamente, o funcionamento dos modelos de regressão *Huber Regressor*, *K-Nearest Neighbor*, *Random Forest* e *Ridge Regression*.

2.1.4.1 *Huber Regressor*

O *Huber Regressor*, introduzido por Huber (1992), é uma técnica robusta de regressão que aborda a presença de valores atípicos ou contaminados nos dados. Diferentemente do método de mínimos quadrados tradicional, o *Huber Regressor* busca minimizar uma função de perda que combina características do erro quadrático médio (utilizado nos mínimos quadrados) e erro absoluto (usado na mediana).

Ao estimar um parâmetro de localização em distribuições normais contaminadas, o *Huber Regressor* é projetado para ser menos sensível a desvios extremos nos dados, proporcionando

uma abordagem mais resistente a influências de valores atípicos. Em conjuntos de dados, onde a presença de *outliers* pode impactar significativamente a performance do modelo, o *Huber Regressor* oferece uma alternativa eficaz, especialmente quando a distribuição exata dos dados é desconhecida ou aproximada.

Por fim, segundo Franke, Hardle e Martin (2012), a capacidade do *Huber Regressor* de equilibrar a minimização de erros quadráticos e absolutos contribui para uma estimativa robusta dos parâmetros, evitando que influências de *outliers* distorçam significativamente as previsões. Isso é crucial ao lidar com a complexidade e a variedade de dados presentes na indústria musical. A presença de *outliers* nos dados acústicos utilizados para a predição de popularidade musical não é incomum, podendo resultar de variações abruptas nas características das músicas; as variações podem impactar negativamente a precisão de modelos tradicionais de regressão, que são mais sensíveis a essas influências extremas. Sua formulação matemática é dada pela função de perda definida como:

$$L(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{para } |y - \hat{y}| \leq \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta), & \text{caso contrário} \end{cases}$$

onde y é o valor real, $\hat{y}$ é o valor predito, e δ é um parâmetro que controla a sensibilidade aos *outliers*. Quando $|y - \hat{y}| \leq \delta$, a função de perda é quadrática, caso contrário, é linear.

2.1.4.2 *K-Nearest Neighbor*

Conforme em Taunk et al. (2019), o *K-Nearest Neighbor* (KNN) é uma técnica de aprendizado de máquina simples e eficaz, aplicável tanto à classificação quanto à regressão. Sua abordagem consiste em agrupar dados em *clusters* coesos ou subconjuntos e classificar novos dados com base em sua semelhança com dados previamente treinados. A entrada é atribuída à classe com a qual compartilha a maioria dos vizinhos mais próximos.

Apesar de ser uma escolha popular em diversas aplicações devido às suas simplicidade e escalabilidade, o KNN apresenta desafios, como a determinação do valor de k e a escolha da métrica de distância. Ainda em Taunk et al. (2019), é apresentado que, em grandes conjuntos de dados, a determinação pode ser custosa e a fase de predição pode ser lenta.

Entretanto, o KNN pode ser uma escolha viável, especialmente quando a relação entre as características e a popularidade é complexa e não linear. Sua capacidade de lidar com dados não lineares é valiosa quando não há conhecimento prévio dos dados, considerando a diversidade musical. Ele pode ser formalmente expresso da seguinte maneira:

$$\hat{y}(x) = \frac{1}{k} \sum_{i=1}^{k} y_i$$

onde $\hat{y}(x)$ é a predição para o novo dado x, k é o número de vizinhos mais próximos considerados e y_i é o valor da variável de saída para o i-ésimo vizinho mais próximo de x.

2.1.4.3 *Random Forest*

A *Random Forest* é um algoritmo avançado de aprendizado de máquina que utiliza múltiplas "árvores de decisão"para fazer previsões mais precisas. Árvores de decisão são estruturas que imitam o processo de tomada de decisões humanas, dividindo dados em diferentes caminhos com base em condições específicas.

Em Ali et al. (2012), o processo de construção de cada árvore da floresta é descrito como a escolha de um dos conjuntos diferentes de características aleatoriamente definidas para orientar as decisões; o número de características utilizadas afeta a performance do modelo, sendo que mais características resultam em menor erro de previsão para cada árvore, mas aumentam a correlação entre as árvores. Durante esse processo, cada árvore é desenvolvida até o máximo possível, sem restrições.

No contexto da predição de popularidade musical, a *Random Forest* oferece vantagens significativas. O algoritmo utiliza seleção aleatória de características, escolhendo um subconjunto de variáveis a partir do conjunto total disponível. Isso contribui para a diversidade das árvores na floresta, tornando o modelo mais robusto e capaz de lidar com diferentes nuances nos dados. A formulação matemática do modelo pode ser representada como segue:

Seja T o conjunto de árvores na floresta e X um vetor de características da música. Cada árvore t em T pode ser representada como uma função $h_t(X)$. A predição final é obtida por meio da média das predições de todas as árvores na floresta, dado por

$$\hat{Y}(X) = \frac{1}{|T|} \sum_{t \in T} h_t(X)$$

onde $\hat{Y}(X)$ é a predição final para o vetor de características X.

2.1.4.4 *Ridge Regression*

O *Ridge Regression*, conforme descrito por McDonald (2009), destaca-se como um método de estimação de parâmetros utilizado para lidar com o problema de colinearidade frequentemente presente na regressão linear múltipla; a multicolinearidade surge quando duas ou mais variáveis independentes em um modelo de regressão são altamente correlacionadas, tornando difícil distinguir seus efeitos individuais sobre a variável dependente.

Para lidar eficazmente com cenários de multicolinearidade, O *Ridge Regression* introduz um fator de regularização que suaviza as estimativas dos coeficientes, evitando instabilidades na presença de variáveis independentes altamente correlacionadas.

Na predição de popularidade musical, em que diversas características acústicas podem estar interrelacionadas, a multicolinearidade pode ser uma preocupação. A abordagem do *Ridge Regression* oferece uma solução ao lidar de forma eficaz com essa interdependência, contribuindo para estimativas mais estáveis e confiáveis dos parâmetros do modelo. Sua formulação matemática é dada por:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

onde $\hat{\beta}^{ridge}$ representa os estimadores dos coeficientes sob o modelo de *Ridge Regression*, y_i é a variável dependente, X_i é a matriz de variáveis independentes, β é o vetor de coeficientes a ser estimado, λ é o parâmetro de regularização e p é o número de variáveis independentes.

2.2 Trabalhos Relacionados

Na literatura, existem trabalhos que relacionam a popularidade de músicas com as preferências, cada vez mais incidentes, de gênero musical e com as prosperidades corporativas do artista.

Neste contexto, Cosimato et al. (2019) propõem um modelo para estimar o *ranking* na *Billboard 200 Chart* de um álbum de músicas, utilizando abordagens de classificação como *Random Forest* e *Support Vector Machine*[1], no intuito de explorar e classificar dados provenientes de redes sociais. Em tal trabalho, um estudo de análise de sentimentos dos *tweets* de usuários do Twitter, em relação à música, foi conduzido para treinar um modelo de classificação que, por sua vez, prevê a classificação de álbuns de músicas no *ranking*. Para a experimentação prática proposta, os experimentos foram realizados com o auxílio de um sistema de verificação de precisão denominado B200P; foram selecionados 19 álbuns e, utilizando essa entrada, o sistema alcançou uma precisão[2] de 94%. Tal abordagem demonstrou-se promissora, mas a análise de sentimentos dos usuários do *Twitter* pode ser influenciada por fatores externos, como tendências temporárias, e não necessariamente refletir a qualidade musical de uma obra. Além disso, a abordagem proposta está restrita aos níveis de artista e música, podendo afetar a precisão do modelo e sua capacidade de generalização.

Outra abordagem para a predição de popularidade de músicas é apresentado em Oliveira, Lacerda e Moro (2022) que tem, como objetivo, analisar a colaboração do artista sob uma perspectiva de gênero para entender melhor como as conexões de gênero impactam o sucesso musical. Sua metodologia inclui a construção de uma rede de gêneros musicais baseada no sucesso para detectar perfis de colaboração e analisar sua evolução através do tempo. Diferentemente

[1] *Support Vector Machine* é um algoritmo de aprendizado de máquina que é usado para classificação de dados em duas ou mais classes e tem sido amplamente utilizado em problemas de aprendizado supervisionado.

[2] A precisão é uma métrica que representa a proporção de instâncias corretamente classificadas pelo modelo em relação ao total de instâncias classificadas.

de Cosimato et al. (2019), essa pesquisa evidencia o efeito significativo que as conexões entre artistas geram no sucesso musical; ademais, também evidencia a importância de se considerar a evolução do gênero musical ao longo do tempo e a influência de tendênciais musicais na predição de sucesso.

Já em Interiano et al. (2018), foram utilizadas as características acústicas das músicas para realizar predições. Para a análise da estratégia proposta, foram coletadas mais de 500.000 músicas a partir do *Top 100 de Singles* do Reino Unido e do site *MusicBrainz*; os autores extraíram as características acústicas das faixas, incluindo variáveis binárias que indicam a emoção transmitida pela música. Com essas informações, foi possível determinar se uma música seria ou não um sucesso, garantindo uma precisão de 70% com o classificador *Random Forest*. Também foram realizados experimentos nos quais os autores adicionavam uma propriedade que indicasse a contribuição de um *superstar* na composição da música, definido como um artista que possuísse pelo menos uma música na primeira posição do *ranking* nos cinco anos anteriores. O acréscimo desse campo possibilitou que o modelo atingisse uma precisão de 85%. Em contrapartida, a abordagem possui limitações quanto o uso de dados, estando limitada à uma fonte desatualizada; a ampliação da amostra de dados pode ser significativa na obtenção de uma visão mais ampla e diversificada da popularidade musical.

Relacionando os trabalhos citados, foi produzido um comparativo ilustrado na Tabela 2.1: avaliou-se os pontos em comum de cada trabalho, qual a relação entre eles e como eles contribuem para o presente trabalho. A Tabela 2.1 apresenta os seguintes critérios: (a) diversidade musical, que se refere à variedade de estilos, gêneros e formas musicais presentes em uma determinada cultura ou região; (b) dados atualizados, que indicam se a informação utilizada para análise ou tomada de decisão reflete a situação mais recente; e (c) características acústicas, que consideram a percepção emocional e subjetiva da música pelo ouvinte.

Tabela 2.1 – Comparativo dos trabalhos relacionados

Autores	Características		
	Diversidade musical	Dados atualizados	Características acústicas
COSIMATO et al. (2019)		x	
OLIVEIRA et al. (2022)	x	x	
INTERIANO et al. (2018)			x
Estratégia proposta	x	x	x

Fonte: Elaborado pela autora

De acordo com a Tabela 2.1, observa-se que o critério "dados atualizados" foi abordado pela maior parte dos trabalhos relacionados descritos: a utilização de uma fonte de dados atualizada é essencial na predição de popularidade musical, uma vez que as preferências musicais do público estão em constante evolução. O critério "diversidade musical" também foi abordado em dois dos trabalhos citados, levando em consideração a variedade de gostos musicais na sociedade: uma ampla gama de fatores que afetam as preferências musicais das pessoas são considerados pelo modelo.

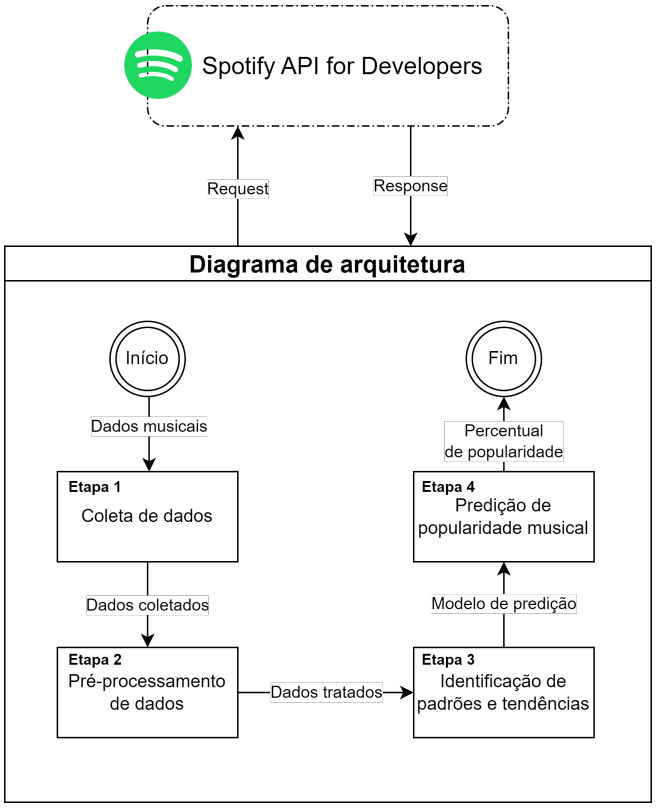
Outro ponto importante observado foi a consideração de características acústicas na análise de dados. Poucos trabalhos fazem uso dessas informações, apesar de estudos sugerirem que as características acústicas, como timbre, ritmo, harmonia, entre outras, podem influenciar a percepção da música pelo ouvinte, gerando sentimentos que podem estar vinculados à popularidade de uma determinada música.

Este trabalho visa realizar um estudo de predição de popularidade de músicas no mercado musical como os demais trabalhos mencionados. Pretende-se compreender alguns aspectos relevantes por trás da popularidade e, tal como Interiano et al. (2018), características acústicas para realizar predições; o uso de dados atualizados e representativos para o modelo de predição é fundamental para obter resultados precisos e confiáveis.

3 Desenvolvimento

Como mencionado, este trabalho possui, como objetivo geral, a proposta, o desenvolvimento e a validação de uma estratégia para predição da popularidade de músicas, baseada em características acústicas e gêneros musicais das mesmas. Para tanto, uma arquitetura de funcionamento foi definida (vide Figura 3.1), englobando desde a coleta e pré-processamento de dados relevantes até a aplicação de algoritmos de aprendizado de máquina. Desta forma, é possível criar um modelo de predição capaz de estimar a popularidade de uma nova música com uma precisão significativa.

Figura 3.1 – Arquitetura de funcionamento da estratégia de predição proposta



Fonte: Elaborado pela autora

De acordo com a Figura 3.1, o funcionamento da estratégia de predição proposta é composto por quatro etapas, a saber: (a) a primeira etapa, descrita na Seção 3.1, consiste em coletar dados por meio do consumo da API do *Spotify* gerando, para a próxima etapa, um conjunto de dados coletados; (b) a segunda etapa, apresentada na Seção 3.2, consiste em realizar o pré-processamento desses dados gerando, para a próxima etapa, um conjunto de dados pré-processados; (c) a terceira etapa, descrita na Seção 3.3, consiste em realizar a identificação de padrões e tendências nos dados pré-processados gerando, para etapa final, os modelos de predição a serem utilizados; (d) a quarta etapa, apresentada na Seção 3.4, consiste em utilizar dos modelos de predição criados na etapa anterior para prever o percentual de popularidade de músicas no conjunto de teste. Nas próximas seções, são apresentados com mais detalhes as etapas desenvolvidas para o funcionamento da estratégia.

3.1 Coleta de dados

Esta seção aborda o processo de coleta de dados, fornecendo uma visão aprofundada da execução da Etapa 1 da arquitetura de funcionamento da estratégia de predição proposta (vide Figura 3.1).

A fonte primária de dados é a API do *Spotify*, reconhecida pela sua abrangência e confiabilidade no cenário musical. Especificamente, foi utilizado o *Spotipy*, a biblioteca em *Python* para a API do *Spotify*, para obter dados sobre faixas e artistas. Cada entrada de dados possui informação de classificação, como posição no *ranking*, nome da faixa, nome do artista e informações detalhadas sobre as músicas, como características acústicas e gêneros musicais associados aos artistas. Dentro do conjunto de dados, existem 13 características acústicas, que abrangem elementos como acusticidade, dançabilidade, duração (em milissegundos), energia, instrumentalidade, chave, vivacidade, volume, modo, expressividade vocal, tempo, assinatura de tempo e valência. A descrição detalhada e o tipo de dado de cada uma dessas características acústicas são detalhados na Tabela 3.1. Essas características fornecem uma base para a compreensão das nuances musicais presentes no conjunto de dados.

Tabela 3.1 – Detalhes das características acústicas

Característica acústica	**Tipo de dado**	**Descrição do valor**
Acusticidade	float	Indica se a faixa é acústica, sendo uma medida de confiança, que varia de 0.0 a 1.0. Um valor de 1.0 representa alta confiança de que a faixa é acústica.
Dançabilidade	float	Indica o quão adequada uma faixa é para a dança. Um valor de 0.0 indica que é menos adequada para dançar, enquanto 1.0 indica que é altamente adequada para dançar.
Duração	int	Indica a duração da faixa em milissegundos.
Energia	float	Indica a medida perceptual de intensidade e atividade, que varia de 0.0 a 1.0. Características perceptuais que contribuem para esse atributo incluem amplitude dinâmica, percepção de volume, timbre, taxa de início e entropia geral.
Instrumentalidade	float	Indica se uma faixa não contém vocais. Valores acima de 0.5 são destinados a representar faixas instrumentais, mas a confiança aumenta à medida que o valor se aproxima de 1.0.
Chave	int	Indica a tonalidade da faixa. Números inteiros mapeiam para notas usando a notação padrão de classes de altura (*Pitch Class*). Por exemplo, 0 = C, 1 = $C\sharp/D\flat$, 2 = D, e assim por diante. Se nenhuma tonalidade foi detectada, o valor é -1.
Vivacidade	float	Indica a presença de uma plateia na gravação. Um valor acima de 0.8 indica uma forte probabilidade de que a faixa seja ao vivo.
Volume	float	Indica a sonoridade geral de uma faixa em decibéis (dB), que varia entre -60 e 0 dB. A sonoridade é a qualidade de um som que é o principal correlato psicológico da intensidade física (amplitude).
Modo	int	Indica a modalidade (maior ou menor) de uma faixa, ou seja, o tipo de escala a partir da qual seu conteúdo melódico é derivado. Maior é representado por 1 e menor por 0.
Expressividade vocal	float	Indica a presença de palavras faladas em uma faixa. Quanto mais exclusivamente parecida com discurso a gravação for, mais próxima de 1.0 será o valor do atributo.
Tempo	float	Indica a velocidade ou ritmo de uma faixa musical e deriva diretamente da duração média das batidas.
Assinatura de tempo	int	Indica quantas batidas há em cada compasso.
Valência	float	Indica a positividade musical transmitida por uma faixa, que varia de 0.0 a 1.0.

Fonte: Elaborada pela autora.

A busca por faixas musicais foi conduzida de maneira segmentada por gênero. Cada gênero foi tratado como uma categoria independente, otimizando a representatividade das amostras. A seleção abrangeu o cenário musical brasileiro, com o intuito de capturar a riqueza e diversidade da produção musical no contexto nacional. Para cada gênero, um processo de seleção aleatória foi conduzido, considerando até 50 faixas. Durante o processo, medidas foram implementadas para evitar duplicatas, garantindo que cada faixa fosse processada apenas uma vez. Em casos de exceções ou falta de características acústicas, mensagens de erro apropriadas foram exibidas.

Os dados recuperados desempenham um papel fundamental na construção de um conjunto abrangente de informações. Após a conclusão dessa etapa, os dados obtidos são transformados na Etapa 2 (vide Figura 3.1), sendo preparados para serem empregados por algoritmos na identificação de padrões e tendências.

3.2 Pré-processamento de dados

Esta seção trata do pré-processamento, relativo a Etapa 2 (vide Figura 3.1), necessário para preparar o conjunto de dados obtido na Etapa 1 para treinamento de modelos de aprendizado de máquina. A comparação entre as duas bases revela distinções notáveis: a base maior mostra uma ampla variação nos atributos e uma média de popularidade mais alta, indicando preferência por músicas populares. Por outro lado, a base menor exibe menor variação e média de popularidade mais baixa, sugerindo uma possível diversidade de gêneros menos explorados.

Para padronização e uniformidade dos dados, utilizou-se a técnica de normalização, representada pelo *StandardScaler*, que busca ajustar as variáveis numéricas para uma escala padrão, de modo a garantir que as diferentes escalas não prejudiquem a interpretação e a análise comparativa dos atributos. Destaca-se a decisão intencional de preservar *outliers*, considerando que os dados musicais, por natureza, exibem uma vasta gama de características, refletindo a diversidade intrínseca da produção musical; essa decisão é respaldada pela compreensão de que esses valores extremos podem conter informações valiosas sobre músicas únicas ou inovadoras que fogem das normas convencionais.

Em relação à diversidade de gêneros musicais, empregou-se a técnica de codificação binária conhecida como *one-hot encoding*; as informações categóricas dos gêneros musicais foram transformadas em variáveis binárias distintas, onde a presença ou ausência de um determinado gênero é representado por 1 ou 0, respectivamente. Sobre a preservação dos *outliers*, certos gêneros musicais menos convencionais ou emergentes podem ser considerados *outliers* em um conjunto de dados mais amplo.

Por fim, os dados tratados servem como entrada para a Etapa 3 (vide Figura 3.1), onde modelos de aprendizado de máquina são empregados para analisar padrões e comportamentos musicais.

3.3 Identificação de padrões e tendências

A Etapa 3 tem, como objetivo, analisar os dados provenientes da Etapa 2, a fim de discernir padrões subjacentes e identificar tendências significativas relacionadas à popularidade musical. Dentre os algoritmos empregados destacam-se o *Ridge Regression*, *Huber Regressor*, *Random Forest* e *K-Nearest Neighbor*, que foram treinados utilizando o conjunto de treinamento derivado da divisão das bases de dados em treino e teste que, correspondem, respectivamente, a 70% e 30%.

O *Ridge* busca modelar a relação entre as características acústicas e a popularidade, incorporando uma penalidade que contribui para a estabilidade do modelo, especialmente quando lidamos com conjuntos de dados complexos. A escolha deste modelo se fundamenta na sua eficácia em lidar com correlações entre características preditivas, promovendo uma interpretação mais estável, dada a natureza multifacetada do cenário musical. Para este propósito, o modelo foi configurado com os seguintes parâmetros: [’alpha’: 1, ’solver’: ’saga’].

O *Huber* é introduzido para mitigar o impacto de valores extremos no conjunto de dados, garantindo uma abordagem mais resiliente em face de potenciais outliers que podem influenciar de maneira desproporcional os resultados. Sua robustez torna-se viável no contexto musical, onde determinadas músicas podem se destacar devido a características únicas ou inovadoras. O modelo *Huber* é configurado com os seguintes parâmetros: [’alpha’: 0.001, ’epsilon’: 1.1, ’max_iter’: 200, ’tol’: 0.0001].

O *Random Forest* é particularmente relevante para a predição de popularidade musical devido às suas características distintas; ao contrário de uma única árvore de decisão, que pode ser suscetível a *overfitting* ou a considerar padrões específicos do conjunto de treinamento, ele busca mitigar esses problemas ao agregar várias árvores. Ao explorar a complexidade e não linearidade das relações entre as características acústicas e a popularidade, esse modelo pode identificar padrões mais abrangentes e representativos, proporcionando uma visão mais holística do conjunto de dados. O modelo é configurado com os seguintes parâmetros: [’max_depth’: None, ’min_samples_leaf’: 1, ’min_samples_split’: 5, ’n_estimators’: 250].

O KNN oferece uma perspectiva valiosa ao considerar a similaridade entre músicas em termos de características acústicas e gêneros musicais. A identificação dos vizinhos mais próximos possibilita inferências sobre a popularidade musical com base em padrões encontrados em casos semelhantes. O modelo KNN foi configurado com os seguintes parâmetros: [’algorithm’: ’auto’, ’leaf_size’: 5, ’metric’: ’manhattan’, ’n_neighbors’: 5, ’p’: 1, ’weights’: ’distance’].

Uma vez que os modelos de predição tenham sido treinados, eles podem ser utilizados para predizer a popularidade de músicas que não foram utilizadas no treinamento. Para isso, o conjunto de teste é fornecido como entrada aos modelos treinados na Etapa 4 (vide Figura 3.1).

3.4 Predição de popularidade musical

A Etapa 4 assume o papel de realizar a predição da popularidade musical no conjunto de teste. Com base nos modelos de aprendizado de máquina treinados na etapa anterior, a avaliação do desempenho é conduzida comparando as previsões geradas com a verdadeira popularidade das músicas no conjunto de teste. Essa análise não apenas avalia a capacidade preditiva dos modelos, mas também proporciona fatores determinantes da popularidade musical. A conclusão desta etapa representa não apenas o resultado preditivo, mas também uma contribuição significativa para a compreensão dos padrões e tendências que influenciam a aceitação do público em relação às músicas.

4 Experimentação Prática

Neste capítulo, são apresentados e analisados os experimentos de validação da estratégia de predição de popularidade musical, seguindo a arquitetura proposta no Capítulo 3. A Seção 4.1 descreve os experimentos realizados e a Seção 4.2 apresenta e avalia os resultados obtidos por meio dos experimentos.

4.1 Descrição experimental

Para realizar os experimentos de validação da estratégia de predição de popularidade musical, considerando a utilização dos modelos de aprendizado de máquina (vide Seção 3.3), foram gerados, por meio do consumo da API do *Spotify*, dois bancos de dados, sendo que: (a) o primeiro é composto por dados musicais de gêneros brasileiros restritos[1] e (b) o segundo representa gêneros musicais brasileiros puros[2], conforme apresentado na Tabela 4.1. A Figura 4.1 apresenta a estrutura dos bancos de dados gerados, destacando as características acústicas das músicas e os gêneros musicais associados.

Tabela 4.1 – Gêneros musicais brasileiros considerados

Categoria	**Gêneros musicais**
Restritos	Brazilian Funk, Sertanejo, Samba, Forró, Pagode, Brazil, Jovem Guarda, Bossa Nova, Brazilian Classical, Brazilian Jazz, Latin Classical, MPB, Funk Pop, Pagode Baiano, Sertanejo Universitário, Piseiro, Afrobeat Brasileiro, Boom Bap Brasileiro, Brazilian Hip Hop, Brazilian Reggae, Rap Nacional, Rock Nacional, Afrofuturismo Brasileiro, Manguebeat
Puros	Sertanejo Pop, Sertanejo Universitário, Samba, Forró, Pagode, Jovem Guarda, MPB, Funk Carioca, Rock Nacional

Fonte: Elaborada pela autora.

Cada um dos banco de dados foi utilizado para treinar os modelos de aprendizado de máquina. Os dados foram divididos em conjuntos de treinamento e teste. A popularidade da música[3] foi definida como a variável dependente nesses bancos de dados. As variáveis independentes são as características associadas a cada faixa; essas características incluem tanto informações relacionadas às características acústicas quanto informações de gênero musical.

[1] Os gêneros musicais brasileiros restritos tendem a ser mais específicos, incluindo subgêneros e influências contemporâneas do Brasil

[2] Os gêneros musicais brasileiros puros representam as principais correntes musicais originadas no Brasil

[3] Campo do tipo inteiro, fornecido pela API do *Spotify*, que é atribuido a uma faixa musical, representando uma medida que varia de 0 a 100, sendo 100 a pontuação máxima de popularidade

Todas as características acústicas disponíveis foram selecionadas, exceto "Assinatura de tempo" e "Modo", pois são propriedades consideradas irrelevantes para o poder explicativo. As fórmulas para calcular os valores das características são mantidas em segredo pelo *Spotify* (provavelmente por razões comerciais).

Como os modelos de aprendizado de máquina lidam melhor com entradas númericas, cada gênero musical é representado em uma coluna de uma tabela. Cada coluna indica a presença do gênero para uma música específica, atribuindo o valor 1 se a faixa musical pertence ao gênero e 0 caso contrário; esse método, conhecido como *codificação one-hot*, permite uma representação mais flexível dos gêneros musicais, por associar diferentes pesos aos diferentes gêneros.

Para otimizar o desempenho dos modelos, foi conduzida uma exploração sistemática de diferentes configurações por meio da técnica *GridSearchCV* (*Grid Search Cross-Validation*). Essa abordagem permitiu examinar um conjunto diversificado de hiperparâmetros para cada modelo de aprendizado de máquina, buscando identificar as combinações mais eficazes para a tarefa de predição da popularidade musical. Durante esse processo, o conjunto de treinamento é dividido em várias dobras, e o modelo é treinado e validado em diferentes subconjuntos.

Os cenários de experimentação foram delineados em duas etapas distintas. Na primeira etapa, todos os modelos foram considerados utilizando os dois bancos de dados disponíveis. Na segunda etapa, os melhores modelos identificados na etapa anterior, foram treinados novamente, mas desta vez utilizando apenas as *features* selecionadas pelo *SelectFromModel*[4]

Após o treinamento com as *features* selecionadas, os modelos foram avaliados quanto ao seu desempenho utilizando métricas relevantes. O Coeficiente de Determinação (R^2) foi utilizado para medir a capacidade dos modelos em explicar a variabilidade dos dados de popularidade musical. Ele fornece uma medida da qualidade das previsões em relação à variação dos valores observados, variando de 0 a 1, onde 1 indica uma correspondência perfeita. O Erro Médio Absoluto (MAE) foi empregado como uma métrica adicional para avaliar a precisão das previsões, representando a média das diferenças absolutas entre as previsões do modelo e os valores reais de popularidade.

4.2 Análise dos Resultados

Nesta seção, são apresentados e analisados os resultados obtidos por meio da experimentação prática descrita na Seção 4.1.

A Tabela 4.2 apresenta os resultados obtidos da primeira etapa de experimentação com os modelos de aprendizado de máquina aplicados aos bancos de dados musicais criados.

[4] Técnica de seleção de *features* utilizada para escolher as características mais relevantes para modelos durante a fase de treinamento.

Modelo	Treino R^2	Teste R^2	Treino MAE	Teste MAE
Random Forest	0.4728	0.1408	0.5360	0.6420
Ridge Regression	0.3429	0.0555	0.6121	0.6558
Huber Regressor	0.3884	-0.1157	0.4287	0.6680
K-Nearest Neighbor	0.9999	0.1056	0.0005	0.6311

(a) Desempenho dos modelos com gêneros puros brasileiros

Modelo	Treino R^2	Teste R^2	Treino MAE	Teste MAE
Random Forest	0.9083	0.6200	0.1973	0.4462
Ridge Regression	0.7762	0.6855	0.3192	0.4169
Huber Regressor	0.7577	0.6380	0.2392	0.3644
K-Nearest Neighbor	0.9777	0.5586	0.0166	0.4416

(b) Desempenho dos modelos com gêneros restritos brasileiros

Tabela 4.2 – Desempenho dos modelos na primeira etapa

Os resultados obtidos para o banco de dados de gêneros puros brasileiros revelam desafios adicionais no ajuste dos modelos. À medida que se restringe o escopo dos dados para gêneros musicais mais específicos, observa-se a dificuldade dos modelos em generalizar para novos bancos de dados não vistos durante o treinamento. Todos os modelos enfrentaram uma queda significativa no R^2 e um aumento no MAE quando aplicados a dados mais puros. Essa tendência sugere uma sensibilidade desses modelos à diversidade dos gêneros musicais mais específicos.

Em relação ao desempenho dos modelos com gêneros restritos brasileiros, nota-se que os modelos *Random Forest* e *K-Nearest Neighbor* demonstraram um ajuste expressivo aos dados de treinamento, evidenciado pelo elevado R^2, mas o desempenho em dados de teste sugere que eles possam estar enfrentando dificuldades em generalizar seus conhecimentos para novos bancos de dados que não foram previamente vistos durante o treinamento. Essa situação pode ser interpretada como um sinal de sobreajuste, indicando que os modelos podem ter se adaptado demais aos detalhes específicos dos dados de treinamento, perdendo a capacidade de generalização para novos dados.

Ao analisar os resultados do *Ridge Regression* e do *Huber Regressor*, no contexto dos gêneros restritos brasileiros, observa-se um comportamento distinto do *Random Forest* e do *K-Nearest Neighbor* em relação ao desempenho nos conjuntos de treinamento e teste. Ambos os modelos demonstraram ser opções válidas, pois a disparidade entre as métricas é menor, indicando uma melhor capacidade de generalização para novos dados. O *Ridge Regression*, conhecido por lidar eficientemente com a multicolinearidade ao introduzir penalidades na função de custo, apresenta coeficientes menos propensos a flutuações extremas. O *Huber Regressor*, por outro lado, é projetado para ser menos sensível a valores discrepantes; no entanto, seus resultados podem variar dependendo da natureza específica dos dados.

Modelo	Treino R^2	Teste R^2	Treino MAE	Teste MAE
Ridge Regression	0.7664	0.6858	0.3270	0.4108
Huber Regressor	0.7602	0.6235	0.2441	0.3624

Tabela 4.3 – Desempenho dos modelos na segunda etapa

Na segunda etapa de experimentação prática, a validação foi realizada nos dois melhores modelos da primeira etapa, o *Ridge Regression* e o *Huber Regressor*, considerando a seleção de características relevantes para a predição do popularidade musical. Os resultados consolidados na Tabela 4.3 evidenciam que ambos os modelos mantiveram um bom desempenho, com métricas relativamente consistentes nos conjuntos de treinamento e teste. Uma vez que o *Ridge Regression* manteve seu padrão de comportamento, sugerindo que as características removidas durante o processo de seleção não exerceram uma influência significativa no desempenho geral do modelo, não foi realizada uma análise específica das características mais importantes para esse modelo; a estabilidade inerente do *Ridge Regression*, decorrente de sua regularização, limita a variação dos coeficientes, tornando essa análise menos informativa. Por outro lado, o *Huber Regressor* apresentou um desempenho ainda mais robusto após a seleção de características, indicando que a escolha criteriosa dessas variáveis contribuiu para uma melhoria na capacidade preditiva do modelo.

A Tabela 4.4 mostra as características mais importantes para o *Huber Regressor* após a seleção.

Característica	Importância
Valência	0.2681
Acusticidade	-0.0936
Vivacidade	-0.1965
Instrumentalidade	-0.2680
Energia	-0.3724

Tabela 4.4 – Características Mais Importantes para *Huber Regressor*

As características selecionadas foram aquelas que tiveram um impacto mais significativo na capacidade do modelo *Huber Regressor* de prever a popularidade da música. O *Huber Regressor* atribui uma importância positiva significativa à "Valência", indicando que a positividade emocional da música tem uma influência forte na popularidade. Porém, os coeficientes negativos para "Acusticidade", "Vivacidade", "Instrumentalidade" e "Energia" indicam que faixas menos acústicas, menos prováveis de serem gravadas ao vivo, menos instrumentais e com menor energia têm uma maior probabilidade de alcançar popularidade.

De modo geral, os resultados obtidos foram satisfatórios. A abordagem adotada na seleção de características contribuiu para um aprimoramento significativo no desempenho do *Huber Regressor*. Esses resultados têm implicações práticas para a compreensão e previsão da

popularidade musical, oferecendo uma base sólida para futuras investigações e aplicações na indústria musical.

Figura 4.1 – Estrutura dos bancos criados

genres	duration_ms	acousticness	danceability	instrumentalness	valence	energy	liveness	loudness	mode	speechiness	tempo	popularity
['brazilian hip hop', 'funk carioca']	149091	0.447	0.955	0.000489	0.497	0.886	0.0825	-4.235	1	0.195	132.08	0
['bossa nova', 'brazilian boogie', 'brazilian soul', 'instrumental funk']	219173	0.151	0.632	0.825	0.938	0.791	0.127	-5.509	1	0.0507	101.675	42
['brazilian rock', 'mpb', 'pop rock brasileiro', 'rock mineiro', 'rock nacional brasileiro']	238547	0.0671	0.715	9.2e-05	0.786	0.741	0.413	-7.786	1	0.0301	96.987	54
['bossa nova', 'brazilian boogie', 'brazilian soul', 'instrumental funk']	219173	0.151	0.629	0.825	0.936	0.791	0.127	-5.509	1	0.0542	101.398	0
['brazilian bass', 'brazilian edm', 'complextro', 'electro house', 'melbourne bounce']	185238	0.00421	0.786	0.151	0.651	0.974	0.289	-5.919	1	0.0853	125.964	42
['brazilian percussion', 'canadian latin']	264933	0.0982	0.805	0.0015	0.743	0.914	0.0803	-5.134	1	0.102	108.253	16
['brazilian edm', 'brazilian house']	191136	0.00784	0.733	0.804	0.414	0.952	0.351	-5.383	1	0.0522	123.984	29
['brazilian hip hop']	422857	0.0548	0.685	8.16e-06	0.674	0.567	0.179	-4.073	0	0.124	104.22	51
['brazilian rock', 'mpb', 'pop rock brasileiro', 'rock mineiro', 'rock nacional brasileiro']	253173	0.0208	0.775	0.00112	0.923	0.863	0.0479	-5.529	0	0.0419	117.002	44
['bossa nova jazz', 'brazilian jazz', 'jazz funk']	389093	0.422	0.795	0.759	0.933	0.568	0.0575	-11.691	1	0.075	97.498	19
['boom bap brasileiro', 'brazilian hip hop', 'brazilian reggae', 'rap nacional antigo']	243760	0.406	0.727	1.37e-06	0.941	0.854	0.0754	-5.147	0	0.138	90.627	45
['brazilian hip hop']	282000	0.408	0.64	0.0	0.5	0.69	0.0696	-8.221	1	0.271	81.812	47
['brazilian dnb', 'deep liquid bass', 'jump up', 'jungle']	267225	0.00417	0.706	0.764	0.0863	0.573	0.155	-3.907	1	0.141	174.037	25
['brazilian punk', 'brazilian punk rock', 'brazilian reggae', 'brazilian rock', 'hard rock brasileiro', 'rock brasiliense', 'rock nacional brasileiro']	240107	0.00174	0.58	1.35e-05	0.377	0.973	0.0696	-5.391	1	0.101	91.044	38
['brazilian emo', 'brazilian rock', 'novo rock gaucho', 'pop rock brasileiro']	194500	0.674	0.551	9.63e-05	0.366	0.415	0.126	-1.742	1	0.0792	162.378	17
['brazilian edm']	125366	0.016	0.957	0.00172	0.707	0.673	0.103	-6.186	1	0.103	122.958	27
['brazilian rock', 'mpb', 'pop rock brasileiro', 'rock mineiro', 'rock nacional brasileiro']	160827	0.0544	0.765	0.0	0.477	0.787	0.099	-6.114	1	0.0792	127.933	41
['brazilian edm']	210270	0.0138	0.772	0.00102	0.619	0.791	0.28	-4.979	0	0.0547	122.016	29
['brazilian hip hop']	192052	0.153	0.578	0.0	0.38	0.781	0.135	-6.278	0	0.267	79.939	16
['brazilian hip hop', 'brazilian indie']	211173	0.0176	0.659	5.4e-05	0.919	0.781	0.0171	-8.088	1	0.139	188.078	36
['bossa nova', 'brazilian boogie', 'brazilian soul', 'instrumental funk']	230040	0.065	0.796	0.0517	0.944	0.568	0.102	-9.973	0	0.0478	117.025	17
['bossa nova', 'brazilian boogie']	313493	0.538	0.879	0.000333	0.802	0.495	0.0827	-10.584	0	0.0765	102.652	16
['brazilian hip hop', 'manguebeat']	240726	0.537	0.78	0.0	0.324	0.527	0.0918	-8.863	1	0.29	102.088	33
['brazilian reggae', 'rock gaucho']	169053	0.225	0.703	4.84e-06	0.707	0.911	0.329	-4.268	0	0.0575	129.964	14
['brazilian percussion']	202507	0.106	0.533	0.914	0.928	0.971	0.409	-9.504	1	0.0614	106.634	20
['brazilian modern jazz']	196266	0.0108	0.647	0.782	0.635	0.691	0.0829	-9.869	1	0.0498	132.801	10
['brazilian boogie', 'brazilian soul', 'mpb']	193040	0.131	0.419	3.99e-05	0.522	0.686	0.187	-7.263	0	0.0788	119.697	26
['brazilian hip hop']	162397	0.136	0.761	0.0	0.366	0.306	0.115	-15.75	0	0.302	108.928	37
['brazilian boogie']	406067	0.254	0.784	0.000334	0.865	0.484	0.0628	-9.884	1	0.14	80.17	12
['brazilian reggae', 'rock gaucho']	240653	0.097	0.605	0.0	0.617	0.949	0.885	-4.073	0	0.212	103.76	12
['brazilian hip hop']	162377	0.3	0.779	0.0	0.789	0.958	0.299	-4.988	0	0.262	63.505	13
['brazilian hip hop']	662791	0.106	0.579	3.8e-05	0.357	0.629	0.197	-10.828	1	0.211	172.077	14
['bossa nova', 'brazilian boogie', 'brazilian soul', 'instrumental funk']	267160	0.82	0.414	0.00292	0.702	0.675	0.161	-10.48	1	0.317	91.785	10
['brazilian hip hop']	212187	0.0109	0.636	0.0	0.787	0.676	0.628	-11.664	1	0.307	91.916	11
['brazilian hip hop']	373099	0.237	0.704	0.0	0.635	0.615	0.184	-7.705	0	0.441	87.816	36
['brazilian hip hop']	190877	0.663	0.575	1.19e-06	0.78	0.763	0.335	-6.297	1	0.185	90.027	12
['brazilian hip hop']	239778	0.199	0.549	0.0	0.718	0.708	0.0868	-9.428	1	0.378	178.06	11
['brazilian rock', 'mpb', 'musica maranhense']	222111	0.117	0.771	0.000183	0.691	0.862	0.863	-7.079	0	0.0936	133.05	11
['brazilian dnb', 'deep liquid bass', 'jump up', 'jungle']	254788	0.00527	0.692	0.246	0.455	0.952	0.359	-2.537	1	0.303	174.019	16
['brazilian boogie']	232627	0.128	0.863	0.00743	0.595	0.409	0.0594	-15.126	0	0.0787	100.277	16
['brazilian jazz']	190133	0.523	0.578	2.82e-06	0.722	0.731	0.189	-8.882	0	0.0536	106.797	13
['brazilian tech house']	163810	0.00937	0.836	0.0859	0.891	0.955	0.0825	-4.963	0	0.051	126.003	27
['arrocha', 'forro', 'sertanejo', 'sertanejo universitario']	168217	0.18	0.636	0.0	0.761	0.933	0.544	-3.707	1	0.0615	160.109	77
['arrocha', 'sertanejo', 'sertanejo universitario']	190294	0.115	0.347	0.0	0.577	0.802	0.418	-5.435	1	0.0447	74.689	75
['agronejo', 'arrocha', 'forro', 'piseiro', 'sertanejo universitario']	135688	0.1	0.571	0.0	0.968	0.929	0.354	-2.455	0	0.0398	158.008	74
['brazilian gospel', 'sertanejo gospel']	213421	0.759	0.541	0.0	0.314	0.346	0.127	-9.926	0	0.0332	113.369	76
['arrocha', 'forro', 'sertanejo', 'sertanejo universitario']	150864	0.215	0.656	0.0	0.745	0.913	0.0827	-3.316	1	0.0522	149.969	75
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	173968	0.112	0.477	0.0	0.521	0.922	0.371	-3.394	1	0.0606	101.571	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	173968	0.112	0.477	0.0	0.521	0.922	0.371	-3.394	1	0.0606	101.571	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	173968	0.112	0.477	0.0	0.521	0.922	0.371	-3.394	1	0.0606	101.571	1
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	173968	0.112	0.477	0.0	0.521	0.922	0.371	-3.394	1	0.0606	101.571	0
['agronejo', 'sertanejo', 'sertanejo pop']	175726	0.463	0.781	0.0	0.744	0.894	0.174	-2.722	0	0.0566	131.963	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	173968	0.112	0.477	0.0	0.521	0.922	0.371	-3.394	1	0.0606	101.571	0
['agronejo', 'sertanejo', 'sertanejo pop']	199346	0.18	0.634	0.0	0.68	0.806	0.415	-4.301	1	0.224	180.09	0
['agronejo', 'sertanejo', 'sertanejo pop']	175726	0.463	0.781	0.0	0.744	0.894	0.174	-2.722	0	0.0566	131.963	0
['agronejo', 'sertanejo', 'sertanejo pop']	175726	0.463	0.781	0.0	0.744	0.894	0.174	-2.722	0	0.0566	131.963	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	178251	0.39	0.359	0.0	0.496	0.76	0.247	-4.632	1	0.0513	154.136	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	184307	0.249	0.568	0.0	0.442	0.938	0.979	-3.752	1	0.225	119.869	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	183308	0.588	0.47	0.0	0.56	0.694	0.692	-6.476	0	0.0473	171.615	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	180607	0.548	0.545	0.0	0.52	0.726	0.366	-4.584	1	0.0358	96.224	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	181458	0.308	0.644	0.0	0.507	0.872	0.536	-4.187	1	0.0563	142.02	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	172005	0.13	0.483	0.0	0.576	0.91	0.395	-3.474	1	0.0845	166.057	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	172005	0.13	0.483	0.0	0.576	0.91	0.395	-3.474	1	0.0845	166.057	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	174118	0.316	0.556	4.55e-06	0.48	0.874	0.753	-2.651	1	0.0497	117.849	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	183431	0.668	0.729	0.0	0.638	0.752	0.219	-4.1	0	0.0339	113.865	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	153410	0.617	0.486	0.0	0.663	0.871	0.97	-3.642	0	0.258	151.175	1
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	180607	0.548	0.545	0.0	0.52	0.726	0.366	-4.584	1	0.0358	96.224	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	183308	0.588	0.47	0.0	0.56	0.694	0.692	-6.476	0	0.0473	171.615	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	180607	0.548	0.545	0.0	0.52	0.726	0.366	-4.584	1	0.0358	96.224	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	151093	0.194	0.635	0.0	0.666	0.789	0.966	-5.037	1	0.0365	106.12	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	172005	0.13	0.483	0.0	0.576	0.91	0.395	-3.474	1	0.0845	166.057	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	183308	0.588	0.47	0.0	0.56	0.694	0.692	-6.476	0	0.0473	171.615	1
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	178251	0.39	0.359	0.0	0.496	0.76	0.247	-4.632	1	0.0513	154.136	1
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	163914	0.668	0.499	0.0	0.442	0.838	0.349	-3.035	1	0.0636	117.715	2
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	181458	0.308	0.644	0.0	0.507	0.872	0.536	-4.187	1	0.0563	142.02	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	178251	0.39	0.359	0.0	0.496	0.76	0.247	-4.632	1	0.0513	154.136	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	183431	0.668	0.729	0.0	0.638	0.752	0.219	-4.1	0	0.0339	113.865	1
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	172005	0.13	0.483	0.0	0.576	0.91	0.395	-3.474	1	0.0845	166.057	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	183308	0.588	0.47	0.0	0.56	0.694	0.692	-6.476	0	0.0473	171.615	0
['agronejo', 'arrocha', 'sertanejo', 'sertanejo universitario']	153410	0.617	0.486	0.0	0.663	0.871	0.97	-3.642	0	0.258	151.175	0

Fonte: Elaborado pela autora

5 Considerações Finais

Neste capítulo, são apresentados aspectos conclusivos sobre o trabalho desenvolvido. A Seção 5.1 apresenta as considerações finais do trabalho e a Seção 5.2 apresenta as perspectivas de trabalho futuro.

5.1 Conclusão

Este trabalho tem, como objetivo, investigar a popularidade das músicas por meio da análise de suas características acústicas e do gênero musical. Para tanto, uma estratégia de predição foi definida e experimentos de validação da mesma foram realizados

Por meio dos experimentos, quanto ao foco associado ao desempenho dos modelos nos bancos de dados, pode-se concluir que os resultados obtidos foram satisfatórios, mas há espaço para aprimoramento. As análises realizadas podem auxiliar a indústria musical na compreensão dos fatores que contribuem para a popularidade de uma música, destacando a importância das características acústicas e do gênero musical nesse processo.

Com base nas análises dos modelos, destaca-se a importância dos algoritmos *Ridge Regression* e *Huber Regressor*, que demonstraram boa capacidade de generalização nos bancos de dados testados. A seleção de características relevantes durante a segunda etapa de experimentação contribuiu para aprimorar ainda mais o desempenho do *Huber Regressor*, indicando a eficácia dessa abordagem na melhoria da capacidade preditiva do modelo.

Apesar da pesquisa na área de recuperação da informação musical ser pouco desenvolvida no Brasil, este trabalho contribui para o seu avanço, especialmente na aplicação de estratégias em grandes bancos de dados de gravações musicais.

5.2 Trabalhos Futuros

Como perspectivas de trabalho futuro, pretende-se: (1) realizar uma exploração de novos fatores que podem influenciar a popularidade musical, incluindo análises regionais para compreender como características específicas de determinadas regiões, como o contexto cultural e preferências musicais, podem impactar a popularidade das músicas; (2) utilizar bancos de dados maiores e mais representativos, permitindo uma avaliação mais abrangente do desempenho dos modelos em cenários diversos; e (3) explorar diferentes modelos de predição, a fim de comparar e identificar aqueles que melhor se adequam ao contexto musical.

Referências

PANDA, R.; REDINHO, H.; GONÇALVES, C.; MALHEIRO, R.; PAIVA, R. P. How does the spotify api compare to the music emotion recognition state-of-the-art? **Proceedings of the 18th Sound and Music Computing Conference (SMC 2021)**, 2021.

INTERIANO, M.; KAZEMI, K.; WANG, L.; YANG, J.; YU, Z.; KOMAROVA, N. Musical trends and predictability of success in contemporary songs in and out of the top charts. **IEEE Access**, 2018.

SANTINI, R. M.; SOUZA, R. F. Recuperação da informação de música e a ciencia da informação: tendências e desafios de pesquisa. **VIII ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação**, 2007.

MCLANE, A. Music as information. **Annual Review of Information Science and Technology (ARIST)**, v. 31, p. 225, 1996. ISSN 0066-4200.

CRUZ, F. W. Necessidades de informação musical de usuários não especializados. **Tese (Doutorado em Ciência da Informação)**, Universidade de Brasília, 2008.

FUTRELLE, J.; DOWNIE, J. S. Interdisciplinary communities and research issues in music information retrieval. In: **ISMIR**. [S.l.: s.n.], 2002. v. 2, p. 215–221.

FLEISCHER, R.; SNICKARS, P. Discovering spotify-a thematic introduction. **Culture Unbound**, v. 9, n. 2, p. 130–145, 2017.

ELLIS-PETERSEN, H. Music streaming hailed as industry's saviour as labels enjoy profit surge. **The Guardian**, v. 29, 2016.

JUSLIN, P. N.; LAUKKA, P. Communication of emotions in vocal expression and music performance: Different channels, same code? **Psychological bulletin**, American Psychological Association, v. 129, n. 5, p. 770, 2003.

NORTH, A. C.; HARGREAVES, D. J.; KRAUSE, A. E. Music and consumer behaviour. **Oxford handbook of music psychology**, Oxford University Press Oxford, UK, p. 481–490, 2009.

JOHNSON, C. Algorithmic music recommendations at spotify. 2014.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, 1996.

ALVAREZ, A. C. Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem. **Dissertação (Mestrado em Ciências de Computação e Matemática Computacional)**, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2007.

SILVA, E. F.; BARROS, F. A.; PRUDÊNCIO, R. B. Uma abordagem de aprendizagem híbrida para extração de informação em textos semi-estruturados. v. 1, p. 504–13, 2005.

COWIE, J.; LEHNERT, W. Information extraction. **Communications of the ACM**, ACM New York, NY, USA, v. 39, n. 1, p. 80–91, 1996.

CORRÊA, Â.; SFERRA, H. H. Conceitos e aplicações de data mining. **Revista de ciência & tecnologia**, v. 11, n. 19-34, p. 20, 2003.

DIAS, M. M. et al. Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados. Florianópolis, SC, 2001.

HUBER, P. J. Robust estimation of a location parameter. In: **Breakthroughs in statistics: Methodology and distribution**. [S.l.]: Springer, 1992.

FRANKE, J.; HARDLE, W.; MARTIN, D. **Robust and nonlinear time series analysis**. [S.l.]: Springer Science Business Media, 2012. v. 26.

TAUNK, K.; DE, S.; VERMA, S.; SWETAPADMA, A. A brief review of nearest neighbor algorithm for learning and classification. In: **2019 International Conference on Intelligent Computing and Control Systems (ICCS)**. [S.l.: s.n.], 2019.

ALI, J.; KHAN, R.; AHMAD, N.; MAQSOOD, I. Random forests and decision trees. **International Journal of Computer Science Issues (IJCSI)**, International Journal of Computer Science Issues (IJCSI), v. 9, n. 5, p. 272, 2012.

MCDONALD, G. C. Ridge regression. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley Online Library, 2009.

COSIMATO, A.; PRISCO, R. D.; GUARINO, A.; MALANDRINO, D.; LETTIERI, N.; SORRENTINO, G.; ZACCAGNINO, R. The conundrum of success in music: Playing it or talking about it? **IEEE Access**, v. 7, p. 123289–123298, 2019.

OLIVEIRA, G.; LACERDA, A.; MORO, M. Analyses of musical success based on time, genre and collaboration. **Anais do XXXV Concurso de Teses e Dissertações**, SBC, p. 81–90, 2022.