**UNIVERSIDADE PRESBITERIANA MACKENZIE**

**Faculdade de Computação e Informática**

**Ciência De Dados**



**Sistema de Recomendação - Spotify**

**Gabriel Chaves Gonçalves**

**Italo Aparecido Lopes**

**São Paulo**

**2025**

**Resumo**

**Lista de Figuras**

Nenhuma entrada de índice de ilustrações foi encontrada.

**Lista de Equações**

Nenhuma entrada de índice de ilustrações foi encontrada.

**SUMÁRIO**

[1. INTRODUÇÃO 6](#_Toc209960997)

[1.1. MOTIVAÇÃO 7](#_Toc209960998)

[1.2. JUSTIFICATIVA 7](#_Toc209960999)

[1.3. OBJETIVOS 8](#_Toc209961000)

[2. REFERÊNCIAL TEÓRICO 8](#_Toc209961001)

[2.1. APRENDIZADO NÃO SUPERVISIONADO 8](#_Toc209961002)

[2.2. SISTEMAS DE RECOMENDAÇÃO 9](#_Toc209961003)

[2.2.1. FILTRAGEM POR CONTEÚDO 9](#_Toc209961004)

[2.2.1.1. SIMILARIDADE DE COSSENO 10](#_Toc209961005)

[2.2.2. K-MEANS 10](#_Toc209961006)

[2.2.3. KNN 11](#_Toc209961007)

[2.2.4. SILHOUETTE SCORE 11](#_Toc209961008)

[2.2.5. VETORIZAÇÃO DE TEXTO 12](#_Toc209961009)

[2.2.5.1. TF-IDF 12](#_Toc209961010)

[3. METODOLOGIA 13](#_Toc209961011)

[4. RESULTADOS E DISCUSSÃO 13](#_Toc209961012)

[5. CONCLUSÃO 13](#_Toc209961013)

[6. PRÓXIMAS ETAPAS 13](#_Toc209961014)

[7. LINK GITHUB 13](#_Toc209961015)

[8. REFERÊNCIAS BIBLIOGRÁFICAS 13](#_Toc209961016)

# **INTRODUÇÃO**

Os serviços de streaming musical oferecem suporte emocional relevante; estudos indicam que ouvir música em casa está positivamente correlacionado ao bem-estar subjetivo, reduzindo o estresse e promovendo vínculos sociais (BOER; ABUBAKAR, 2023).

Além disso, durante períodos de crise, como a pandemia da COVID-19, o consumo de músicas com conteúdo positivo aumentou, sugerindo uso compensatório para manter o equilíbrio emocional (MUÑIZ-TALAVERA et al., 2024).

Nos últimos anos, com o crescimento das plataformas de streaming, o consumo de música passou por uma grande transformação. Serviços como Spotify, Deezer, Apple Music e YouTube Music oferecem milhões de faixas em tempo real, tornando desafiador oferecer experiências personalizadas para cada usuário.

O setor musical tem se beneficiado economicamente com o streaming: em 2024, a receita global de música gravada chegou a US$ 29,6 bilhões, impulsionada por assinaturas que cresceram 9,5%, superando o crescimento médio do PIB global (IFPI, 2025).

Plataformas como o Spotify também democratizam o mercado musical, permitindo que artistas independentes obtenham receitas significativas; em 2024, músicos independentes geraram mais de US$ 5 bilhões em receitas apenas na plataforma (SPOTIFY, 2025).

Dada a relevância cultural, social e econômica dos serviços de streaming, torna-se essencial o desenvolvimento de ferramentas que ampliem a personalização e melhorem a experiência do usuário. Nesse sentido, a inteligência artificial desempenha papel estratégico, permitindo que grandes volumes de dados musicais sejam analisados e convertidos em recomendações que atendem tanto às preferências individuais quanto às tendências globais de consumo.

Nesse cenário, a aprendizagem não supervisionada tem papel fundamental, pois permite descobrir padrões ocultos em grandes bases de dados sem a necessidade de rótulos pré-definidos. Técnicas como clustering e regras de associação são utilizadas para agrupar usuários com preferências semelhantes e identificar padrões de consumo relevantes (AFOUDI; LAZAAR; AL ACHHAB, 2021).

Para enfrentar esse desafio, empresas do setor utilizam sistemas de recomendação que analisam dados e padrões de comportamento para sugerir músicas personalizadas.

Esses sistemas se beneficiam da aprendizagem não supervisionada, pois conseguem identificar perfis de usuários e similaridades entre itens de forma autônoma. Nos últimos anos, os sistemas têm evoluído para oferecer experiências altamente personalizadas, reduzindo o esforço de busca do usuário e priorizando conteúdos de maior relevância (ZOU; ZHOU, 2025).

Nesse contexto, tais sistemas analisam padrões de consumo e atributos das músicas para indicar novas faixas, artistas ou listas de reprodução, contribuindo para maior engajamento e satisfação do usuário nas plataformas de streaming.

Este projeto propõe o desenvolvimento de um protótipo que, a partir de dados do Spotify Dataset 1921-2020, aplicará algoritmos de filtragem por conteúdo como a similaridade de cosseno e KMeans, para montar os sistemas de recomendação com a finalidade de indicar novas músicas aos usuários. Além disso, serão realizados testes com usuários finais, permitindo avaliar a efetividade das recomendações e a experiência proporcionada.

O trabalho também envolve etapas essenciais da ciência de dados, como análise exploratória, transformação e modelagem dos dados, integrando teoria e prática para construção de uma solução aplicada.

# **MOTIVAÇÃO**

A crescente disponibilidade de dados digitais e o avanço das técnicas de análise automatizada têm transformado a forma como empresas compreendem e interagem com seus clientes.

A possibilidade de explorar esses dados para oferecer experiências personalizadas é um dos principais motores da inovação em sistemas de recomendação.

A motivação deste trabalho surge da oportunidade de aplicar técnicas de aprendizagem não supervisionada para desenvolver um sistema capaz de construir perfis personalizado e sugerir músicas com características semelhantes às já apreciadas pelo usuário, sem a necessidade de rótulos explícitos, contribuindo para recomendações mais inteligentes e adaptativas.

# **JUSTIFICATIVA**

A filtragem baseada em conteúdo é uma abordagem eficaz para sistemas de recomendação, especialmente em cenários onde não há dados suficientes sobre outros usuários ou quando se deseja preservar a individualidade das sugestões.

Este trabalho se justifica pela relevância prática da aplicação dessas técnicas em um ambiente real, contribuindo para o avanço de soluções inteligentes na área de Ciência de Dados.

# **OBJETIVOS**

**Objetivo Geral**

Desenvolver um sistema de recomendação musical utilizando técnicas de filtragem baseada em conteúdo, com base nos atributos disponíveis na base de dados do Spotify, visando oferecer sugestões personalizadas aos usuários.

**Objetivos Específicos**

* Realizar a exploração e o pré-processamento da base de dados do Spotify, incluindo limpeza, normalização e seleção de atributos relevantes.
* Construir perfis de usuários simulados com base em faixas previamente escutadas ou avaliadas.
* Implementação dos algoritmos de similaridade de cosseno e KMeans para sistemas de recomendação.
* Avaliar a qualidade das recomendações por meio de testes com usuários finais e métricas como Silhouette Score e Davies-Bouldin Index.
* Discutir os resultados obtidos, destacando as vantagens e limitações da abordagem adotada.

**Objetivo extensionista**

Este trabalho busca contribuir para os Objetivos de Desenvolvimento Sustentável (ODS) da ONU, como Saúde e Bem-Estar (ODS 3) e Educação de Qualidade (ODS 4) ao propor um sistema de recomendação que possibilite ampliar o acesso democrático a conteúdos musicais, promovendo maior diversidade cultural e valorização de artistas independentes nas plataformas digitais.

# **REFERÊNCIAL TEÓRICO**

A fundamentação teórica é uma etapa essencial para a construção de um trabalho acadêmico, pois fornece embasamento conceitual e metodológico ao estudo.

Nessa seção, são apresentados os principais conceitos, modelos e teorias que sustentam a pesquisa, permitindo uma compreensão mais ampla do tema investigado. Portanto, neste capítulo, abordaremos os conceitos utilizados no desenvolvimento deste trabalho.

# **APRENDIZADO NÃO SUPERVISIONADO**

O aprendizado não supervisionado, busca extrair padrões e estruturas latentes a partir de dados não rotulados, sem necessidade de supervisão explícita. Através de técnicas como clustering, análise de componentes principais e detecção de anomalias, é possível explorar a distribuição dos dados e gerar representações úteis para tarefas subsequentes.

Em trabalho recente, Kauffmann et al. (2025) destacam que modelos de aprendizado não supervisionado podem, inadvertidamente, apresentar efeitos do tipo Clever Hans, ou seja, basear suas decisões em artefatos no dado que não generalizam bem, o que pode comprometer o desempenho em aplicações reais (KAUFFMANN et al., 2025).

# **SISTEMAS DE RECOMENDAÇÃO**

Sistemas de recomendação são algoritmos que sugerem itens relevantes aos usuários com base em padrões de comportamento, preferências ou atributos dos produtos. Eles são amplamente utilizados em plataformas digitais para personalizar experiências e reduzir a sobrecarga informacional.

As principais abordagens incluem a filtragem colaborativa, que utiliza interações entre usuários, e a recomendação baseada em conteúdo, que analisa características dos itens. A combinação dessas técnicas forma sistemas híbridos, capazes de lidar com limitações individuais de cada abordagem (RICCI; ROKACH; SHAPIRA, 2011).

Aguiar et al. (2020) realizaram um estudo comparativo sobre sistemas de recomendação baseados em personalidade, utilizando o IBM Watson Personality Insights. A pesquisa mostrou que características psicológicas podem melhorar a acurácia das recomendações em relação aos métodos tradicionais. Esse estudo reforça a importância de incorporar variáveis contextuais e subjetivas nos sistemas de recomendação. Ao considerar traços de personalidade, os algoritmos se tornam mais sensíveis ao perfil do usuário, aumentando a relevância das sugestões.

# **FILTRAGEM POR CONTEÚDO**

A filtragem por conteúdo é uma técnica de recomendação que sugere itens com base nas características dos produtos e no perfil individual do usuário. O sistema compara atributos dos itens com preferências previamente registradas, buscando similaridade entre eles.

Essa abordagem é eficaz em contextos com pouca interação entre usuários, pois depende apenas do histórico do próprio indivíduo. Sistemas baseados em conteúdo utilizam metadados e atributos estruturados para prever a relevância de novos itens, sendo amplamente aplicados em e-commerce, redes sociais e serviços de streaming (MUREL; KAVLAKOGLU, 2024).

Vieira et al. (2023) analisaram o uso da filtragem por conteúdo em sistemas de recomendação aplicados a bibliotecas. O estudo mostrou que essa técnica é amplamente adotada por sua simplicidade e capacidade de personalização, mesmo em ambientes com pouca interação entre usuários. Segundo os autores, a filtragem por conteúdo permite construir recomendações mais precisas ao considerar atributos específicos dos documentos. No entanto, o modelo pode apresentar limitações, como a dificuldade de sugerir itens fora do perfil conhecido do usuário.

# **SIMILARIDADE DE COSSENO**

A similaridade de cosseno é uma métrica que avalia o grau de semelhança entre dois vetores com base no ângulo entre eles. Ela é amplamente utilizada em sistemas de recomendação, pois permite comparar itens com base em suas características vetoriais, independentemente da magnitude.

Essa técnica foi originalmente proposta por Salton e McGill (1983) no contexto da recuperação de informação. Desde então, tornou-se uma ferramenta essencial para identificar padrões em dados multidimensionais, como perfis de usuários ou atributos de produtos.

Araújo (2024) aplicou a similaridade de cosseno em um sistema de recomendação de animes, utilizando avaliações de usuários como vetores. O estudo demonstrou que essa abordagem melhora a precisão das sugestões ao considerar a proximidade entre perfis de preferência.

# **K-MEANS**

O K-Means é um algoritmo de aprendizado de máquina não supervisionado amplamente utilizado para agrupamento de dados. Seu objetivo é particionar um conjunto de dados em k grupos (clusters), de acordo com a similaridade entre os pontos (KARIM KHAN et al., 2024).

O processo inicia com a escolha de k centróides, que representam os centros de cada grupo. Cada ponto é atribuído ao cluster com centróide mais próximo, geralmente utilizando a distância euclidiana como métrica. Em seguida, os centróides são recalculados e o processo se repete até que não haja mudanças significativas, indicando a convergência do modelo (BALIGODUGULA; AMSAAD, 2025).

Embora eficiente, o K-Means apresenta limitações, como a necessidade de definir previamente o número de clusters, a sensibilidade a outliers e dificuldades com dados de formatos não esféricos. Pesquisas recentes propõem adaptações para mitigar esses problemas, como métodos híbridos e métricas de validação, como o silhouette score (KARIM KHAN et al., 2024).

# **KNN**

O método K-Nearest Neighbors (KNN) é uma técnica de aprendizado de máquina não paramétrica amplamente utilizada para problemas de classificação, onde a variável resposta é categórica, podendo assumir dois ou mais valores. Segundo James et al. (2021), o método KNN classifica uma nova observação com base nas classes das k observações mais próximas presentes no conjunto de treinamento, utilizando uma métrica de distância como a Euclidiana.

De acordo com Bishop (2006), a escolha do valor k influencia diretamente no desempenho do modelo, e geralmente é determinada por meio de técnicas como validação cruzada. Além disso, recomenda-se a normalização prévia dos dados para evitar que variáveis com diferentes escalas tenham influência desproporcional no resultado.

O método KNN é amplamente utilizado em áreas como diagnóstico médico, sistemas de recomendação e classificação de imagens devido à sua simplicidade conceitual e fácil implementação. Entretanto, pode apresentar desafios em conjuntos grandes ou de alta dimensionalidade, exigindo métodos auxiliares para melhorar sua eficiência.

# **SILHOUETTE SCORE**

O Silhouette Score é uma métrica interna amplamente utilizada para avaliar a qualidade de agrupamentos em algoritmos de aprendizado não supervisionado.

Esse valor varia entre -1 e 1, onde valores próximos a 1 indicam que o ponto está bem atribuído ao seu cluster (boa coesão interna e separação entre clusters), valores próximos a 0 indicam sobreposição entre clusters, e valores negativos denotam possível atribuição incorreta (o ponto estaria mais próximo de outro cluster) (RAUTENSTRAUCH; OHLER, 2025)

Esse valor varia entre -1 e 1, onde valores próximos a 1 indicam que o ponto está bem atribuído ao seu cluster (boa coesão interna e separação entre clusters), valores próximos a 0 indicam sobreposição entre clusters, e valores negativos denotam possível atribuição incorreta (o ponto estaria mais próximo de outro cluster) (RAUTENSTRAUCH; OHLER, 2025).

Apesar de ser bastante usada, pesquisas recentes apontam limitações do Silhouette em contextos específicos como integração de dados de célula única (single-cell). Nesses casos, as suposições do Silhouette (clusters bem definidos e separáveis) são frequentemente violadas, comprometendo sua confiabilidade nessas aplicações (RAUTENSTRAUCH; OHLER, 2025).

Em trabalhos mais recentes, há propostas de extensões do Silhouette clássico. Por exemplo, Vardakas, Papakostas e Likas (2024) propõem o soft silhouette, uma formulação probabilística do coeficiente de silhueta incorporada a redes neurais para clustering profundo, que busca formar clusters compactos e bem separados de maneira diferenciável e otimizada (VARDAKAS; PAPAKOSTAS; LIKAS, 2024)

# **VETORIZAÇÃO DE TEXTO**

A vetorização de texto é um dos principais processos do pré-processamento de dados textuais, permitindo que textos sejam convertidos em representações numéricas para utilização em modelos de aprendizado de máquina.

Segundo Jurafsky e Martin (2021), essa transformação é essencial, pois os algoritmos computacionais operam sobre números, tornando necessário mapear palavras ou documentos para um espaço vetorial. Métodos tradicionais incluem o Bag of Words (BoW) e o Term Frequency-Inverse Document Frequency (TF-IDF), que representam os textos com base na frequência das palavras, capturando padrões estatísticos do vocabulário.

# **TF-IDF**

O método Term Frequency-Inverse Document Frequency (TF-IDF) é uma técnica amplamente utilizada na vetorização de texto, permitindo a transformação de documentos em representações numéricas para análise e modelagem computacional.

Segundo Manning, Raghavan e Schütze (2008), o TF-IDF combina duas métricas principais: a Frequência do Termo (TF), que mede a recorrência de uma palavra em um documento, e a Frequência Inversa do Documento (IDF), que avalia a importância do termo considerando sua presença em vários documentos. Esse equilíbrio garante que palavras comuns, como artigos e preposições, tenham menor peso, enquanto termos mais relevantes recebam maior ênfase na representação vetorial.

Segundo Manning, Raghavan e Schütze (2008), o TF-IDF combina duas métricas principais: a Frequência do Termo (TF), que mede a recorrência de uma palavra em um documento, e a Frequência Inversa do Documento (IDF), que avalia a importância do termo considerando sua presença em vários documentos. Esse equilíbrio garante que palavras comuns, como artigos e preposições, tenham menor peso, enquanto termos mais relevantes recebam maior ênfase na representação vetorial.

# **METODOLOGIA**

# **RESULTADOS E DISCUSSÃO**

# **CONCLUSÃO**

# **PRÓXIMAS ETAPAS**

# **LINK GITHUB**

<https://github.com/lopesita/Projeto-III-System_recomendation>

# **REFERÊNCIAS BIBLIOGRÁFICAS**

AFOUDI, Yassine; LAZAAR, Mohamed; AL ACHHAB, Mohammed. Intelligent recommender system based on unsupervised machine learning and demographic attributes. Simulation Modelling Practice and Theory, v. 107, p. 1–16, fev. 2021. DOI: 10.1016/j.simpat.2020.102198. Acesso em 10 set 2025.

AGUIAR, Janderson Jason B.; ARAÚJO, Joseana M. F. R. de; COSTA, Evandro de B. Estudo comparativo de abordagens para sistemas de recomendação baseados em personalidade com uso do serviço IBM Watson Personality Insights. Revista Ibérica de Sistemas e Tecnologias de Informação, n. 40, p. 73–88, dez. 2020. DOI: 10.17013/risti.40.73-88. Acesso em: 21 set. 2025.

ARAUJO, Arthur Frade de. Desenvolvimento de sistema de recomendação de animes: uma aplicação da similaridade por cossenos. Recife: Universidade Federal de Pernambuco, Centro de Informática, 2024. Trabalho de Graduação (Bacharelado em Ciência da Computação).

BALIGODUGULA, Vishnu Vardhan; AMSAAD, Fathi. Unsupervised Learning: Comparative Analysis of Clustering Techniques on High-Dimensional Data. [preprint], mar. 2025. Acesso em: 28 set. 2025.

BISHOP, Christopher M. Pattern Recognition and Machine Learning. New York: Springer, 2006. Disponível em: https://link.springer.com/book/10.1007/978-0-387-45528-0. Acesso em: 28 set. 2025.

IBRAHIM, Osman Ali Sadek; et al. Revisiting recommender systems: an investigative survey. Neural Computing and Applications, v. 37, p. 2145–2173, jan. 2025.DOI: 10.1007/s00521-024-10828-5. Acesso em 10 set 2025.

KAUFFMANN, J.; et al. Explainable AI reveals Clever Hans effects in unsupervised learning. Nature Machine Intelligence, v. (a definir), p. (a definir), dez. 2024. DOI: 10.1038/s42256-024-01000-2. Acesso em: 28 set. 2025.

INTERNATIONAL FEDERATION OF THE PHONOGRAPHIC INDUSTRY (IFPI). Amidst highly competitive market, global recorded music revenues grew 4.8% in 2024. 2025. Disponível em: [https://www.ifpi.org/ifpi-amidst-highly-competitive-market-global-recorded-music-revenues-grew-4-8-in-2024/](https://www.ifpi.org/ifpi-amidst-highly-competitive-market-global-recorded-music-revenues-grew-4-8-in-2024/%20).Acesso em: 10 set 2025.

JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3. ed. London: Pearson, 2021. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 28 set. 2025.

KARIM KHAN, Iliyas; DAUD, Hanita Binti; ZAINUDDIN, Nooraini Binti; SOKKALINGAM, Rajalingam; MUSEEB, Abdul; INAYAT, Agha. Addressing limitations of the K-means clustering algorithm: outliers, non-spherical data, and optimal cluster selection. AIMS Mathematics, v. 9, n. 9, p. 25070-25097, 2024. DOI: 10.3934/math.20241222. Acesso em: 28 set. 2025.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. Disponível em: https://nlp.stanford.edu/IR-book/. Acesso em: 28 set. 2025.

MUÑIZ-TALAVERA, Miguel; et al. The Soundtrack of a Crisis: More Positive Music Preferences During Economic and Social Adversity. Journal of Happiness Studies, v. 25, art. 44, 2024. DOI: 10.1007/s10902-024-00757-4. Acesso em 10 set 2025.

MUREL, Jacob; KAVLAKOGLU, Eda. O que é filtragem baseada em conteúdo? IBM Think, 21 mar. 2024. Disponível em: <https://www.ibm.com/br-pt/think/topics/content-based-filtering>. Acesso em: 21 set. 2025.

NAÇÕES UNIDAS BRASIL. Objetivos de Desenvolvimento Sustentável. Nações Unidas no Brasil, 2025. Disponível em: <https://brasil.un.org/pt-br/sdgs>. Acesso em: 10 set. 2025.

RAUTENSTRAUCH, Pia; OHLER, Uwe. Shortcomings of silhouette in single-cell integration benchmarking. Nature Biotechnology, 30 jul. 2025. DOI: 10.1038/s41587-025-02743-4. Acesso em: 28 set. 2025.

RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha. Introduction to recommender systems. In: RICCI, Francesco et al. (Ed.). Recommender Systems Handbook. Boston: Springer, 2011. p. 1–35.

SHAN, Xin; ZHANG, Yan; DENG, Jie; MA, Haixia; HU, Xiaoxi. The Association between Music Listening at Home and Subjective Well-Being. Behavioral Sciences, v. 14, n. 9, art. 767, 2024. DOI: 10.3390/bs14090767. Acesso em 10 set 2025.

SALTON, Gerard; MCGILL, Michael J. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

SPOTIFY. Beyond profits: how the music industry’s cultural and financial impact define its success in 2025. Spotify Newsroom, 12 mar. 2025. Disponível em: [https://newsroom.spotify.com/2025-03-12/beyond-profits-how-the-music-industrys-cultural-and-financial-impact-define-its-success-in-2025/.](https://newsroom.spotify.com/2025-03-12/beyond-profits-how-the-music-industrys-cultural-and-financial-impact-define-its-success-in-2025/.%20) Acesso em: 10 set. 2025.

VARDAKAS, Georgios; PAPAKOSTAS, Ioannis; LIKAS, Aristidis. Deep Clustering Using the Soft Silhouette Score: Towards Compact and Well-Separated Clusters. [preprint], fev. 2024. Disponível em: https://arxiv.org/abs/2402.00608. Acesso em: 28 set. 2025.

VIEIRA, Bruna Beatriz de Moura; PASSOS, Ketry Gorete Farias dos; SALM, Vanessa Marie. Sistemas de recomendação em bibliotecas: iniciativas e proposta de um modelo teórico híbrido. BiblioCanto, v. 9, n. 1, p. 1–15, 2023. DOI: 10.21680/2447-7842.2023v9n1ID32504. Acesso em: 21 set. 2025.