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

Lista de Equações

SUMÁRIO

1. 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.

1.1. 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.

1.2. 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.

1.3. 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.

2. 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.

2.1. 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).

2.2. 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.

2.2.1. 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.

2.2.1.1. 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.

2.2.2. K-means

O K-Means é um método de aprendizado não supervisionado utilizado para particionar dados em k grupos por similaridade geométrica, minimizando a distância entre cada ponto e o centróide do seu cluster. Segundo Géron (2022), o algoritmo alterna iterativamente duas etapas: (i) atribuir cada instância ao centróide mais próximo — geralmente pela distância euclidiana — e (ii) recalcular os centróides como média dos pontos atribuídos, repetindo até convergência. A escolha do valor de k influencia diretamente a qualidade da partição, sendo comumente guiada por métricas como Elbow e Silhouette.

Embora simples e eficiente para grandes volumes, o K-Means pressupõe clusters aproximadamente esféricos e de tamanho similar, sendo sensível a outliers, à escala das variáveis (requer normalização prévia) e à alta dimensionalidade, podendo demandar variantes como Mini-Batch K-Means em cenários de grande escala (Géron, 2022).

2.2.3. 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.

2.2.3.1. TF-IDF

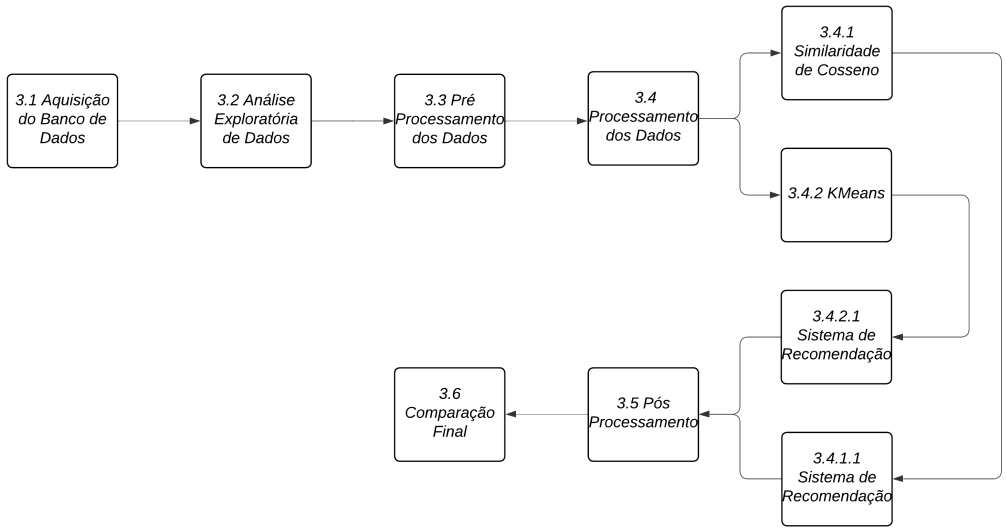
O método Term Frequency-Inverse Document Frequency (TF-IDF) é uma técnica amplamente utilizada na vetorização de texto, permitindo a transformação de documentos em representações numéricas para análise e modelagem computacional.

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.

3. METODOLOGIA

Este capítulo detalha a metodologia utilizada para o desenvolvimento deste trabalho. Serão apresentados o delineamento da pesquisa e o fluxo das atividades, que foi estruturado em sete etapas principais, cada uma contendo atividades específicas, assim como ilustra a Figura 1.

Figura 1 - Fluxograma do Projeto



3.1. Aquisição do Banco de Dados

O banco de dados utilizado neste projeto foi obtido no repositório público Kaggle, no dataset “Spotify Tracks Dataset”, disponibilizado em formato CSV. Essa base reúne metadados musicais e atributos numéricos extraídos pela API oficial do Spotify, incluindo danceability, energy, instrumentalness, valence, entre outros, amplamente utilizados em pesquisa aplicada sobre recomendação musical.

A base foi escolhida porque apresenta estrutura tabular com variáveis numéricas diretamente compatíveis com técnicas de normalização, redução de dimensionalidade e clusterização, eliminando a necessidade de pré-processamento de arquivos de áudio bruto. Isso permite concentrar o esforço metodológico na modelagem e avaliação do sistema.

3.2. Análise Exploratória de Dados

A Análise Exploratória de Dados (AED), foi crucial para a compreensão da estrutura e da qualidade do conjunto de dados brutos. Para otimizar e sistematizar esta fase, utilizou-se a biblioteca ydata-profiling do Python, uma ferramenta robusta que automatiza a geração de um relatório descritivo completo. Este relatório, exportado em

formato HTML, forneceu estatísticas descritivas, distribuições de variáveis e matrizes de correlação de maneira concisa e visual.

Através do relatório HTML gerado pelo ydata-profiling, foi possível identificar as características essenciais do dataset, como a natureza das variáveis e a presença de dados faltantes ou outliers críticos para o estudo. As informações coletadas nesta fase de diagnóstico subsidiaram diretamente a etapa subsequente de Pré-Processamento.

3.3. Pré Processamento de Dados

A etapa de Pré Processamento teve como objetivo preparar o conjunto de dados, garantindo que as features estivessem na escala e no formato exigidos por cada um dos modelos de análise subsequentes.

Nesta etapa também fizemos a seleção dos algoritmos de aprendizado de máquina que seriam utilizados (Similaridade de Cosseno e KMeans).

Dado que os algoritmos de aprendizado de máquina podem possuir sensibilidades distintas aos tipos de variáveis, a preparação dos dados foi abordada de forma individualizada para cada técnica. Os procedimentos específicos, incluindo a seleção e o tratamento das variáveis, são detalhados nas subseções a seguir.

3.3.1. Similaridade de Cosseno

O Pré Processamento para a Similaridade de Cosseno iniciou-se com a seleção de variáveis, eliminando metadados e identificadores irrelevantes. As colunas id, artists, name, year e release_date foram descartadas do dataset original. Esta exclusão focou em manter apenas atributos musicais quantitativos para a métrica de similaridade.

O conjunto de features restante para o algoritmo de similaridade de cosseno foi submetido à padronização, que é essencial para que todas as variáveis possuam o mesmo peso no cálculo vetorial da Similaridade de Cosseno. Para tal, utilizou-se a técnica StandardScaler (padronização por Z-Score), a qual transforma os dados para que o resultado tenha média zero (μ=0) e desvio-padrão unitário (σ=1). A transformação de cada ponto de dado (x) para seu valor padronizado (z) é definida pela Equação 1:

Equação 1- Fórmula para Padronização (StandardScaler/Z-Score)

$$z = \left(\frac{x - \mu}{\sigma}\right)$$

Onde:

- z: é o valor padronizado (o resultado da transformação).

- x: é o valor original da *feature*.
- μ: é a média da *feature* no conjunto de dados.
- σ: é o desvio-padrão da *feature* no conjunto de dados.

3.3.2. K-means

O pré-processamento dos dados para K-means iniciou-se com a representação textual dos nomes dos artistas por meio da técnica TF-IDF, de modo a converter informações simbólicas em uma matriz esparsa de pesos numéricos. Em paralelo, as variáveis quantitativas selecionadas foram padronizadas por meio do algoritmo StandardScaler (Equação 1), garantindo que diferenças de escala entre atributos não influenciassem de forma desproporcional o processo de agrupamento. Em seguida, as duas matrizes textuais transformadas por TF-IDF e numérica padronizada, foram combinadas para compor a estrutura final de entrada.

Sobre essa matriz combinada aplicou-se o PCA, representada através da Equação 2, com o objetivo de reduzir a dimensionalidade, concentrando a maior variabilidade do conjunto em um número reduzido de componentes principais, além de mitigar redundâncias e ruído. Por fim, os dados reduzidos foram submetidos ao algoritmo K-Means, tendo-se avaliado a qualidade dos agrupamentos por meio do índice de Silhouette Score, procedimento que possibilitou a definição do número adequado de clusters para o conjunto analisado.

Equação 2 – PCA

$$Z = X_C . V_k$$

Onde:

- Z matriz transformada (dados projetados nos novos componentes principais).
- X_C matriz de dados centralizada, obtida por $X_c = X - \bar{X}$ onde $\bar{X}$ é o vetor de médias das variáveis originais.
- V_k matriz composta pelos k autovetores principais (colunas), resultantes da decomposição da matriz de covariância.

3.4. Processamento dos Dados

Após a conclusão do pré-processamento, iniciou-se a etapa central de Processamento dos Dados. Esta fase corresponde à aplicação efetiva dos modelos de aprendizado não supervisionado e seu objetivo foi extrair os padrões da matriz de features já tratada.

O processamento foi dividido em duas abordagens analíticas paralelas. A primeira foi a aplicação da Similaridade de Cosseno. A segunda consistiu na execução do algoritmo de clusterização K-Means.

Embora ambas sejam métodos não supervisionados, suas abordagens no trabalho são distintas. A Similaridade de Cosseno foi utilizada para quantificar a semelhança entre itens. O K-Means foi usado para descobrir agrupamentos naturais de músicas.

Os resultados gerados por cada um desses algoritmos são a base das análises subsequentes. Eles permitiram tanto a criação de um sistema de recomendação quanto a segmentação do catálogo de músicas.

Para ambos os sistemas de recomendação, o input do usuário é o nome de um artista. Se este artista possui múltiplas músicas na base, o sistema calcula um "perfil agregado" através de um vetor médio de suas features. Este vetor representa o estilo geral do artista e serve como ponto de partida para a busca. Uma consequência desta abordagem é o foco na descoberta inter-artista: o sistema filtra as músicas do artista de entrada e recomenda apenas faixas de artistas diferentes que sejam vetorialmente similares ao perfil.

3.4.1. Similaridade de Cosseno

A Similaridade de Cosseno é uma métrica fundamental em sistemas de recomendação e recuperação de informação, sendo a técnica central de processamento de dados. Essa medida avalia a similaridade entre dois vetores (A e B) em um espaço multidimensional, calculando o cosseno do ângulo entre eles, independentemente da magnitude dos vetores.

O resultado varia no intervalo [-1, 1], onde 1 indica vetores na mesma direção (similaridade máxima) e 0 indica ortogonalidade (ausência de similaridade). A fórmula para o cálculo da Similaridade de Cosseno entre dois vetores, A e B, é dada pela Equação 3, que utiliza o produto escalar dos vetores dividido pelo produto de suas respectivas magnitudes (norma euclidiana):

Equação 3 - Similaridade de Cosseno

$$Similcosseno(A,B) = \frac{A\,.\,B}{||A||.\,||B||} = \frac{\sum_{i=1}^{n} AiBi}{\sqrt{\sum_{i=1}^{n} Ai^2}\sqrt{\sum_{i=1}^{n} Bi^2}}$$

Onde:

- Similcosseno(A,B): é a pontuação de similaridade de cosseno entre os vetores A e B.

- A e B: são os vetores de features (músicas) que estão sendo comparados.
- A e B: é o produto escalar dos vetores A e B (soma do produto de suas componentes).
- ||A|| e ||B||: são as magnitudes (normas) dos vetores A e B.
- n: é o número total de features (dimensões) nos vetores, após o pré-processamento.
- Ai e Bi: São os valores padronizados da i-ésima feature nos vetores A e B, respectivamente.

3.4.1.1. Sistema de Recomendação por Similaridade de Cosseno

Após a definição da métrica de similaridade de cosseno para quantificar a relação entre os vetores de features das músicas, foi desenvolvido um algoritmo de recomendação. Este algoritmo, encapsulado na função “recommend”, implementa um sistema de filtragem baseada em conteúdo (Content-Based Filtering). O seu funcionamento segue um fluxo processual de cinco etapas:

- Identificação da Entrada: o processo é iniciado quando o usuário fornece uma entrada de texto.
- Cálculo da Similaridade: uma vez que os índices dos itens de entrada são identificados, a partir da Equação 2 pode-se calcular a similaridade entre o(s) item(ns) de entrada e todos os outros itens do dataset.
- Agregação de Perfil: caso a busca inicial retorne múltiplos itens, o sistema o sistema não se baseia em apenas uma música. Em vez disso, ele calcula a média dos vetores de similaridade. Isso cria um "perfil de similaridade" agregado, que representa de forma mais robusta as características dos itens de entrada.
- Rankeamento e Filtragem: o vetor de similaridade resultante é ordenado em ordem decrescente, criando um ranking de todos os itens do dataset do mais similar ao menos similar.
- Apresentação dos Resultados: o sistema seleciona 5 recomendações mais similares e apresenta seus nomes e artistas ao usuário.

3.4.2. K-Means

O algoritmo K-Means é uma técnica clássica de partição de dados utilizada para agrupar observações em k grupos mutuamente exclusivos, com base na minimização da distância entre cada ponto e o centróide do seu respectivo cluster. Esse método

assume que os dados estão em um espaço métrico e busca descobrir estruturas latentes por proximidade geométrica.

O funcionamento baseia-se em duas etapas iterativas: (i) atribuir cada instância ao centróide mais próximo e (ii) recalcular os centróides como a média dos pontos atribuídos. O processo se repete até convergência, ou seja, até que as atualizações deixem de alterar as posições dos centróides ou atinjam um limite máximo de iterações.

A qualidade da partição é influenciada diretamente pelo valor de k, que precisa ser definido previamente. Valores baixos podem gerar grupos muito amplos e pouco discriminativos, enquanto valores altos podem produzir sobreparticionamento e perda de generalização. Por essa razão, a seleção de k costuma ser apoiada por métricas como o Silhouette Score, que avalia a coesão interna e a separação entre clusters.

O objetivo de otimização do K-Means consiste em minimizar a soma das distâncias quadráticas entre os pontos e seus centróides, conforme representada pela Equação 4 abaixo:

Equação 4 - K-means

$$J = \sum_{i=1}^{k} \sum_{x \in Si} || x - ci||^2$$

Onde:

- J: é a função de custo total, o valor que o K-Means tenta minimizar.
- $\sum_{i=1}^{k}$: é o somatório sobre todos os clusters (do cluster 1 até o cluster k).
- $\sum_{x \in S_i}$: é o somatório sobre todos os pontos x que pertencem àquele cluster específico Si.
- $||x - c_i||^2$: distância Euclidiana ao quadrado entre um ponto de dados (x) e o centroide (ci) do cluster ao quadrado ao qual ele pertence.

3.4.2.1. Sistema de Recomendação por K-Means

Após a etapa de pré-processamento dos dados e da redução de dimensionalidade por PCA, procedeu-se ao desenvolvimento do módulo de recomendação. Nesta parte, optou-se por um sistema de filtragem baseada em conteúdo apoiado em agrupamento (Content-Based Filtering via Clustering), no qual a noção de similaridade entre músicas não é definida diretamente por uma métrica entre pares, mas mediada pela estrutura latente descoberta pelo algoritmo K-Means previamente treinado.

O fluxo de funcionamento do algoritmo de recomendação, encapsulado na função recomendar_por_artista, segue as seguintes etapas:

- Entrada do Usuário: o processo inicia-se quando o usuário fornece o nome de um artista como entrada textual.
- Localização das Faixas: o sistema identifica todas as faixas associadas ao artista no conjunto de dados, garantindo coerência entre o campo textual (artists) e seus vetores correspondentes no espaço PCA.
- Construção de Vetor Representativo: os vetores PCA das faixas do artista são agregados via média, produzindo uma assinatura numérica que sintetiza o perfil sonoro do artista.
- Projeção sobre o Modelo K-Means: esse vetor médio é então submetido ao modelo K-Means já treinado, permitindo identificar a qual cluster latente esse artista está associado dentro da estrutura aprendida.
- Seleção e Ordenação das Recomendações: recuperam-se as faixas pertencentes ao mesmo cluster, excluindo o próprio artista consultado e ranqueando por proximidade ao vetor médio do artista, retornando ao usuário aquelas mais similares segundo a distância euclidiana no espaço PCA.

Este mecanismo permite a geração de recomendações consistentes sem depender de rótulos pré-existentes ou de histórico de uso, explorando exclusivamente a organização emergente dos dados no espaço latente do clustering.

3.5. Pós Processamento dos Dados

O Pós Processamento foi a fase dedicada à interpretação e estruturação dos resultados brutos gerados pelos algoritmos. Esta etapa crucial converte as saídas matemáticas do KMeans e da Similaridade de Cosseno em um formato tratável. O objetivo final é preparar as informações para a análise de desempenho.

3.6. Comparação Final dos Sistemas de Recomendação

Após a implementação dos dois sistemas de recomendação, sendo um baseado em Similaridade do Cosseno e outro baseado em agrupamento via K-Means, foi realizada uma etapa comparativa com o objetivo de selecionar o melhor Sistema de Recomendação para o projeto.

As duas abordagens de recomendação foram avaliadas a partir de um conjunto de dez entradas simulando usuários, cada um informando artistas de sua preferência para consulta no sistema.

Como não existe no conjunto de dados um rótulo que indique qual seria a recomendação “correta” para cada usuário, não é possível calcular acurácia automaticamente no código. Dessa forma, foi conduzido um teste de validação.

O objetivo foi mensurar a eficácia perceptiva do sistema de recomendação. A validação envolveu 10 usuários de teste. Cada usuário forneceu o nome de um artista como dado de entrada. O sistema então gerou um total de 5 recomendações para cada entrada.

Os usuários classificaram cada sugestão individualmente. As etiquetas de classificação foram "Recomendação Boa" ou "Recomendação Ruim", como ilustra a figura 2.

Figura 2 - Tabela de Validação - Sistema de Recomendação

Usuário	Artista	Nº de Recomendações Boas	Nº de Recomendações Ruins
1	A	0	0
2	B	0	0
3	C	0	0
4	D	0	0
5	E	0	0
6	F	0	0
7	G	0	0
8	H	0	0
9	I	0	0
10	J	0	0
Total		0	0
Acurácia		0%	

Para quantificar o desempenho agregado, foi definida a métrica "Acurácia". Neste contexto, a acurácia representa a taxa de sucesso das sugestões, ela foi é calculada pela razão entre o número total de "Recomendações Boas" e o número total de recomendações realizadas (50), conforme elucida a Equação 5.

Equação 5 - Acurácia

$$Acurácia = \frac{\text{Número de Previsões Corretas}}{\text{Número Total de Previsões}}$$

Onde:

- Número de Previsões Corretas: corresponde ao "Total de Recomendações Boas".
- Número Total de Previsões: corresponde à soma de todas as recomendações avaliadas, ou seja, "Total de Recomendações Boas" + "Total de Recomendações Ruins".

4. RESULTADOS E DISCUSSÃO

5. CONCLUSÃO

6. PRÓXIMAS ETAPAS

7. LINK GITHUB

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

8. 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/.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/. 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.