FUNDAÇÃO EDUCACIONAL DE FERNANDÓPOLIS
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

SISTEMAS DE RECOMENDAÇÕES UTILIZANDO INTELIGÊNCIA ARTIFICIAL

RECOMMENDATION SYSTEMS UTILIZING ARTIFICIAL INTELLIGENCE

José Pedro Bezerra Riva[1], Gustavo Andrey Marcelino Bordin[2], Jefferson Antonio Ribeiro Passerini[3]

[1]José Pedro Bezerra Riva, josepedeo14@gmail.com
[2] Gustavo Andrey Marcelino Bordin, andreygustavo32011@gmail.com
[3] Jefferson Antonio Ribeiro Passerini, jefferson.passerini@fef.edu.br

Área: Ciência de Dados.
Subárea: Ciência da Computação

RESUMO

Sistemas de recomendação são amplamente utilizados em plataformas digitais, como e-commerce, serviços de streaming e redes sociais, com a finalidade de personalizar a experiência do usuário, sugerindo produtos, conteúdos ou serviços com base em seu perfil e comportamento. Este Trabalho de Conclusão de Curso tem como objetivo o estudo e desenvolvimento de sistemas de recomendação utilizando técnicas de inteligência artificial (IA). A pesquisa aborda os principais tipos de sistemas de recomendação: baseados em conteúdo, filtragem colaborativa e modelos híbridos, além de explorar o uso de algoritmos de aprendizado de máquina, como redes neurais artificiais e técnicas de processamento de linguagem natural (PLN). Para a parte prática, foi desenvolvido um protótipo funcional utilizando a linguagem Python e bibliotecas como Scikit-learn e Pandas, permitindo testar modelos de recomendação com dados reais ou simulados. Os resultados obtidos demonstram a eficácia das abordagens baseadas em IA na personalização de recomendações, evidenciando a importância desses sistemas na melhoria da experiência do usuário e na geração de valor para empresas digitais.

Palavras-chave: Sistemas de recomendação. E-Commerce. Técnicas de Processamento de Linguagem Natural. Python.

ABSTRACT

Recommendation systems are widely used on digital platforms such as e-commerce, streaming services, and social networks to personalize user experiences by suggesting products, content, or services based on user profiles and behavior.
This Final Year Project aims to study and develop recommendation systems using artificial intelligence (AI) techniques. The research covers the main types of recommendation systems: content-based filtering, collaborative filtering, and hybrid models. It also explores the use of machine learning algorithms, including artificial neural networks and natural language processing (NLP) techniques. For the practical part, a functional prototype was developed using the Python programming language and libraries such as Scikit-learn and Pandas, allowing for the testing of recommendation models with real or simulated data. The results demonstrate the effectiveness of AI-based approaches in providing personalized recommendations, highlighting the importance of these systems in enhancing user experience and adding value to digital businesses.

Keywords: *Recommender Systems, E-Commerce, Natural Language Processing Techniques, Python.*

1 INTRODUÇÃO

Com o crescimento exponencial de dados disponíveis na internet e a popularização de plataformas digitais, tornou-se essencial oferecer experiências personalizadas aos usuários. Nesse contexto, os sistemas de recomendação desempenham um papel fundamental ao sugerir produtos, conteúdos ou serviços que estejam alinhados com os interesses individuais de cada usuário.

Presentes em diversas aplicações do cotidiano, como serviços de streaming, redes sociais e plataformas de comércio eletrônico, esses sistemas utilizam algoritmos inteligentes para processar grandes volumes de dados e prever preferências. A inteligência artificial (IA), especialmente por meio de técnicas de aprendizado de máquina, tem ampliado significativamente a eficácia dos sistemas de recomendação.

Algoritmos como filtragem colaborativa, sistemas baseados em conteúdo e modelos híbridos são cada vez mais aprimorados com o uso de redes neurais e outras abordagens modernas da IA. Isso permite não apenas uma melhor acurácia nas sugestões, mas também uma maior adaptabilidade ao comportamento dinâmico dos usuários.

Este trabalho tem como objetivo investigar as principais técnicas de recomendação impulsionadas por inteligência artificial, bem como desenvolver um protótipo funcional que demonstre, na prática, como essas tecnologias podem ser aplicadas para oferecer recomendações eficientes e personalizadas.

2 REFERENCIAL TEÓRICO

Esta seção tem como objetivo apresentar os fundamentos necessários para compreender a metodologia adotada, incluindo os principais conceitos de processamento de linguagem natural, além dos modelos e algoritmos utilizados ao longo deste estudo.

2.1 INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA

A Inteligência Artificial (IA) é uma área da ciência da computação que busca desenvolver sistemas capazes de simular comportamentos inteligentes, como percepção, raciocínio e tomada de decisão. Dentro desse campo, o Aprendizado de Máquina (ou *Machine Learning*) se destaca como uma abordagem que permite a criação de algoritmos que aprendem padrões a partir de dados históricos e realizam previsões com base nesses padrões (ALPAYDIN, 2014).

O aprendizado de máquina pode ser classificado em três categorias principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. Em sistemas de recomendação, os métodos supervisionados e não supervisionados são os mais utilizados, sendo a filtragem colaborativa e os sistemas baseados em conteúdo os principais paradigmas aplicados.

2.2 SISTEMAS DE RECOMENDAÇÃO

Sistemas de recomendação são ferramentas computacionais projetadas para sugerir itens relevantes aos usuários, com base em seu comportamento anterior ou em características dos itens. Essas ferramentas têm se tornado indispensáveis em aplicações de comércio eletrônico, plataformas de streaming e redes sociais, uma vez que ajudam a personalizar a experiência do usuário e aumentar a retenção e engajamento (RICCI et al., 2011).

Os tipos mais comuns de sistemas de recomendação incluem a Filtragem colaborativa que recomenda itens com base nas preferências de usuários semelhantes; os sistemas baseado em conteúdo, que recomendam itens com base nas características dos itens anteriormente avaliados pelo usuário; e os sistemas híbridos, que combinam elementos das duas abordagens anteriores para melhorar a precisão das recomendações.

Na aprendizagem não supervisionada, são identificados padrões de entrada, mas sem ter um feedback explícito. No caso, os dados não são rotulados. Sendo assim, a tarefa mais comum na aprendizagem não supervisionada é o agrupamento de grupos identificados como úteis. Por exemplo, pode ser desenvolvido gradualmente o conceito de um dia bom ou um dia ruim de tráfego para um taxista, sem essa informação ter sido rotulada por um humano (ALPAYDIN, 2014).

Na aprendizagem por reforço, o algoritmo aprende com recompensas e punições. Ainda no exemplo do taxista, a falta de gorjeta após uma corrida dá a indicação de que algo está errado. Bem como a vitória em um jogo de damas indica que o jogador fez a coisa certa. É conferido então ao agente decidir quais das ações anteriores ao reforço são responsáveis por estes resultados (ALPAYDIN, 2014).

Na aprendizagem supervisionada, o agente observa exemplos de pares de entrada e saída e adquire um entendimento de como transformar a entrada em saída, no primeiro componente mencionado, as entradas consistem em percepções, e a saída é dada por um instrutor, que indica "Freie!" ou "Vire à esquerda". No segundo componente, as entradas são imagens da câmera, e as saídas são fornecidas por um instrutor que as classifica como "isso é um ônibus" (ALPAYDIN, 2014).

No terceiro componente, a teoria da frenagem é uma função que leva em consideração os estados e as ações relacionadas à frenagem até a distância de parada. Nesse caso, o valor de saída é diretamente derivado da percepção do agente (posteriormente); o ambiente atua como o instrutor.

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN) é um campo da IA que visa possibilitar a interação entre computadores e linguagem humana. No contexto de sistemas de recomendação, técnicas de PLN são empregadas na análise de textos como descrições de produtos, resenhas, comentários e títulos, proporcionando ao sistema um entendimento mais profundo do conteúdo textual (JURAFSKY; MARTIN, 2021).

2.4 TF-IDF

Ao transformar textos em representações numéricas por meio de técnicas como *Bag-of-Words*, *Word Embeddings* e TF-IDF, os algoritmos de recomendação passam a ser capazes de identificar semelhanças entre documentos ou produtos com base em conteúdo textual. TF-IDF é uma sigla em inglês que representa "*Term Frequency - Inverse Document Frequency*", que em português pode ser traduzida como "Frequência do Termo - Frequência Inversa do Documento".

De acordo com a explicação fornecida por Moser (2022), o TF-IDF é um

mecanismo de Recuperação de Informação associado ao Processamento de Linguagem Natural.

O termo "TF" refere-se à contagem da frequência com que uma palavra ocorre em um conjunto de documentos, atribuindo pesos às palavras com base em sua frequência. Já o "IDF" analisa a importância das palavras, levando em consideração que uma palavra pode aparecer muitas vezes, mas isso não implica necessariamente que ela seja relevante para a análise.

Por exemplo, no inglês, o artigo "*the*", que equivale a "o" ou "a" em português, costuma ocorrer frequentemente, mas não tem grande importância no contexto geral da análise. Segundo Moser (2022), a fórmula para calcular TF-IDF é $TF\_IDF = TF * IDF$. Sendo que *Term Frequency* (TF) representa a frequência de uma palavra "*p*" em um documento "*d*". É fundamental computar a contagem de todas as palavras; neste caso a fórmula apresentada é:

$$TF(p, d) = (count(p, d)).$$

É importante destacar que uma palavra pode aparecer com frequência elevada em um documento, mas isso não necessariamente a torna mais importante. Portanto, é comum aplicar uma redução na frequência bruta, utilizando o logaritmo na base 10 da frequência. Além disso, costuma-se adicionar 1 à contagem, uma vez que não é possível calcular o logaritmo de 0. Portanto, a fórmula fica como segue abaixo:

$$TF(p, d) = (count(p, d) + 1).$$

Conforme mencionado anteriormente, o IDF refere-se à Frequência Inversa de Documento e tem como objetivo equilibrar o TF, ou seja, o IDF atribui menor importância às palavras que são frequentes em todos os documentos. O cálculo da equação resulta em um vetor de pesos, um para cada palavra do documento, onde "D" representa o conjunto de documentos e "df" é a função que conta a frequência de um termo "t" em cada documento:

$$IDF(t) = \left(\frac{count(D)}{df(t)}\right)$$

O objetivo desta abordagem é gerar um vetor com os dados preparados para que se possa aplicar um classificador de forma a determinar a predição final do texto.

2.5 A LINGUAGEM PYTHON COM FERRAMENTA PARA SISTEMAS DE RECOMENDAÇÃO

A linguagem de programação Python consolidou-se como uma das mais utilizadas na área de ciência de dados, aprendizado de máquina e desenvolvimento de sistemas de recomendação.

Sua popularidade se deve à sintaxe clara, à extensa documentação e à disponibilidade de bibliotecas voltadas para o processamento de dados, aprendizado de máquina e visualização.

Entre as bibliotecas mais utilizadas estão: pandas, para manipulação de dados; scikit-learn, para criação de modelos de aprendizado de máquina; NLTK e spaCy, para processamento de linguagem natural; matplotlib e seaborn, para visualização gráfica; e surprise, para sistemas de recomendação.

Projetos como o abordado no curso "*Machine Learning e Sistema de Recomendação com Python*", ministrado por Jones Granatyr (UDEMY, 2024), utilizam tais ferramentas para desenvolver sistemas práticos de recomendação, com aplicação real e acessível para estudantes e profissionais da área.

2.6 MÉTRICAS DE DESEMPENHO

A avaliação de sistemas de recomendação exige o uso de métricas específicas que considerem não apenas a exatidão das previsões, mas também a relevância, ordenação e utilidade prática das recomendações fornecidas aos usuários. Cada tipo de sistema — como filtragem colaborativa, baseado em conteúdo (por exemplo, com TF-IDF) ou híbrido — apresenta características particulares que tornam certas métricas mais apropriadas do que outras (HERLOCKER et al., 2004; RICCI et al., 2011).

2.6.1 MÉTRICAS PARA SISTEMAS BASEADO EM CONTEÚDO

Modelos baseados em conteúdo, como aqueles que utilizam TF-IDF para comparar similaridade textual entre descrições de itens, são frequentemente avaliados por métricas que analisam a relevância e a ordenação dos itens recomendados. Precisão e revocação avaliam a proporção de itens relevantes entre os recomendados e vice-versa.

São adequadas para cenários em que o sistema recomenda itens similares com

base em suas características textuais. F1-Score combina precisão e revocação em uma única medida harmônica, é útil quando se busca equilibrar ambos os aspectos.

NDCG (*Normalized Discounted Cumulative Gain*) é uma métrica muito utilizada para verificar se os itens mais relevantes aparecem no topo da lista de recomendações, o que é essencial em aplicações baseadas em conteúdo.

Essas métricas são particularmente eficazes quando o sistema utiliza representações vetoriais, como o modelo TF-IDF, que prioriza termos informativos e reduz o peso de palavras genéricas.

2.6.2 MÉTRICAS DE FILTRAGEM COLABORATIVA

Sistemas de recomendação baseados em filtragem colaborativa, principalmente os que utilizam abordagens baseadas em vizinhos mais próximos (KNN) ou fatoração de matrizes, lidam com predição de notas ou classificações fornecidas por usuários.

As métricas RMSE e MAE são amplamente utilizadas para avaliar o erro entre as notas previstas pelo sistema e as notas efetivamente atribuídas pelos usuários, sendo que valores menores indicam melhor desempenho do modelo (ALPAYDIN, 2020). Métricas de ranqueamento, como Precision@k e Recall@k, medem a proporção de itens relevantes dentro de um conjunto limitado de recomendações e são essenciais para avaliar a qualidade de listas recomendadas em sistemas top-N (RICCI et al., 2022).

Além disso, a métrica Coverage é utilizada para medir a proporção do catálogo que é efetivamente recomendada pelo sistema, permitindo verificar se o algoritmo consegue explorar diferentes regiões do conjunto de itens e evitar recomendações excessivamente concentradas (JANNACH et al., 2011; BURKOV, 2019).

2.6.3 MÉTRICAS DE SISTEMAS HÍBRIDOS

Sistemas híbridos combinam estratégias de conteúdo e colaborativas, e, por isso, podem ser avaliados com uma combinação das métricas anteriores. No geral, os modelos híbridos visam maximizar precisão, diversidade e personalização.

Serendipity e Diversidade avaliam o quanto as recomendações surpreendem positivamente o usuário ou o expõem a itens fora de sua zona de interesse usual.

São especialmente valorizadas em sistemas híbridos que buscam evitar a recomendação de itens redundantes. Cobertura e Novidade são importantes para verificar se o sistema está recomendando novos itens e alcançando uma maior parte do acervo disponível.

2.7 APLICAÇÃO DAS MÉTRICAS NO DESENVOLVIMENTO COM PYTHON

Durante a implementação prática de sistemas de recomendação com Python, bibliotecas como *scikit-learn, surprise*, pandas e *numpy* oferecem suporte para o cálculo de praticamente todas as métricas mencionadas. Por exemplo, no curso de *Machine Learning* e Sistemas de Recomendação com Python, são utilizados modelos de filtragem colaborativa com *KNNBasic* (da biblioteca *surprise*) avaliados por RMSE e MAE, e também modelos baseados em conteúdo, avaliados com precisão e similaridade de cosseno entre vetores TF-IDF (UDEMY, 2024).

A escolha das métricas, portanto, deve estar sempre alinhada ao modelo adotado e aos objetivos da aplicação, como priorizar recomendações precisas, diversificadas ou exploratórias.

2.8 MEDIDAS DE SIMILARIDADE E ALGORITMOS DE DISTÂNCIA

Para que um sistema de recomendação identifique itens semelhantes ou usuários com gostos parecidos, é necessário quantificar a "proximidade" entre eles. Matematicamente, isso é feito por meio de cálculos de distância em um espaço vetorial, onde cada produto ou usuário é representado como um ponto em um espaço multidimensional.

2.8.1 Distância de Minkowski

A Distância de Minkowski é uma métrica generalizada utilizada em espaços vetoriais normados. Ela serve como uma fórmula base que pode assumir diferentes comportamentos dependendo do valor atribuído ao parâmetro p. De acordo com Han et al. (2012), a distância de Minkowski entre dois pontos X e Y (ou dois itens com n atributos) é definida pela equação:

$$D(x,y) = \left(\sum_{\{i=1\}}^{\{n\}} \left| x_i - y_i \right|^p \right)^{\left\{\frac{1}{p}\right\}}$$

Onde:

- n é o número de dimensões (atributos do produto);
- xi e yi são os valores dos atributos para os itens X e Y, respectivamente;
- p é o parâmetro de ordem que define a métrica de distância.

A vantagem da utilização de Minkowski é a sua flexibilidade. Ao ajustar o valor de p, é possível alterar a sensibilidade do algoritmo a diferenças grandes ou pequenas entre os atributos dos produtos, permitindo uma calibragem mais fina do sistema de recomendação.

2.8.2 Distância Euclidiana

A Distância Euclidiana é o caso mais comum e intuitivo derivado da métrica de Minkowski, obtida quando se define o parâmetro p = 2. Geometricamente, ela representa o comprimento do segmento de reta que conecta dois pontos no espaço, sendo a métrica escolhida para o desenvolvimento deste projeto, sua fórmula é:

$$d(x,y) = \sqrt{\left\{ \sum_{\{i=1\}}^{n} \left(y_i - x_i \right)^2 \right\}}$$

Em sistemas de recomendação, a distância Euclidiana é amplamente utilizada para calcular a similaridade entre itens que possuem atributos numéricos contínuos (como preço, peso ou avaliações). Quanto menor o valor da distância Euclidiana entre dois vetores de produtos, maior é a similaridade entre eles. É fundamental, no entanto, que os dados estejam normalizados, pois atributos com escalas muito grandes podem dominar o cálculo da distância, distorcendo as recomendações (AGGARWAL, 2016).

2.8.3 K-Nearest Neighbors (KNN)

O algoritmo *K-Nearest Neighbors* (K-Vizinhos Mais Próximos), ou KNN, é um método de aprendizado supervisionado (para classificação) ou não supervisionado (para busca de similaridade) que utiliza métricas de distância, como a Euclidiana,

para tomar decisões. No contexto de sistemas de recomendação, o KNN opera sob a premissa de que itens similares tendem a estar próximos uns dos outros no espaço vetorial.

O funcionamento básico do algoritmo consiste em calcular a distância entre o item-alvo (o produto que o usuário está visualizando) e todos os demais itens da base de dados, selecionar os K itens com as menores distâncias, ou seja, os vizinhos mais próximos, e, por fim, recomendar esses itens ao usuário.

Segundo *Scikit-Learn Developers* (2024), o KNN é um algoritmo "preguiçoso" (*lazy learner*), pois não constrói um modelo interno complexo, mas sim armazena os dados de treinamento e realiza o cálculo de similaridade em tempo real durante a execução. Isso o torna simples de implementar e muito eficaz para filtragem baseada em conteúdo e filtragem colaborativa.

3. METODOLOGIA

3.1 A BASE DE DADOS: AMAZON SALES DATASET

Para o desenvolvimento e validação do sistema de recomendação, foi utilizada a base de dados *Amazon Sales Dataset*, que contém registros de vendas e avaliações de produtos eletrônicos e acessórios comercializados na plataforma da Amazon. O conjunto de dados, em sua versão processada para este projeto, conta com aproximadamente 1.351 registros de produtos únicos, estruturados em 16 atributos distintos. A escolha desta base justifica-se pela riqueza de informações textuais e numéricas, que permitem a aplicação de abordagens baseadas em conteúdo.

3.1.1 Estrutura e Dicionário de Dados

Os dados estão organizados em formato tabular (CSV), onde cada linha representa um produto específico. As principais colunas utilizadas no processamento do sistema de recomendação são detalhadas a seguir:

- Informações de Identificação e Categorização:
    - `product_id`: Identificador único do produto na plataforma.

    - `product_name`: Título comercial do produto. Fundamental para a identificação visual e busca textual.
    - `category`: Categoria hierárquica do item (ex: *Computers & Accessories* | *Cables* | *USB Cables*). Este atributo permite filtrar recomendações dentro de nichos específicos.
- Atributos Numéricos (Métricas de Distância):
    - `rating`: Nota média atribuída pelos usuários ao produto, variando tipicamente em uma escala de 0 a 5. Este valor serve como indicador qualitativo.
    - `rating_count`: Número total de avaliações que o produto recebeu. Este atributo atua como um indicador de popularidade e confiabilidade da nota média. No pré-processamento, este campo (originalmente textual devido à pontuação) é convertido para valor numérico para permitir o cálculo de Distância Euclidiana.
- Atributos Textuais (Processamento de Linguagem Natural):
    - `about_product`: Descrição detalhada das características e funcionalidades do item. É o principal atributo utilizado na vetorização por TF-IDF, pois contém as palavras-chave que definem o conteúdo semântico do produto.
    - `review_content`: Conteúdo textual das avaliações dos usuários, que pode ser utilizado para enriquecer a análise de similaridade ou para análise de sentimento.

3.1.2 Pré-processamento e Adequação

Antes da aplicação dos algoritmos de Inteligência Artificial, a base de dados passa por uma etapa de limpeza e transformação (conforme descrito na seção de tratamento de dados). Isso inclui a remoção de caracteres especiais das colunas de preço e contagem de avaliações, a normalização dos textos (remoção de *stopwords* e padronização para letras minúsculas) e o tratamento de dados faltantes.

Essa estrutura híbrida — contendo tanto texto rico para o modelo de TF-IDF quanto métricas numéricas para o cálculo de distância (Minkowski/Euclidiana) — torna o *dataset* ideal para demonstrar a eficácia de sistemas de recomendação que operam além da simples filtragem colaborativa.

3.1.3 Tipo de Pesquisa

A presente pesquisa caracteriza-se como aplicada, com abordagem quantitativa e experimental, pois tem como objetivo o desenvolvimento e a validação de um sistema de recomendação de produtos por meio de técnicas de Inteligência Artificial. O trabalho visa compreender e aplicar modelos matemáticos e estatísticos para sugerir itens com base em dados reais, simulando o ambiente de um E-Commerce.

3.2 Ferramentas e Tecnologias Utilizadas

O desenvolvimento foi realizado em linguagem Python, utilizando um conjunto de bibliotecas voltadas ao processamento, análise e visualização de dados, bem como à implementação de algoritmos de aprendizado de máquina. As principais ferramentas empregadas neste trabalho foram as seguintes: Pandas, utilizada para manipulação, tratamento e organização dos dados em estruturas tabulares (*DataFrames*); NumPy, responsável pelo processamento numérico e pelas operações vetoriais e matriciais necessárias aos cálculos de similaridade entre produtos; Matplotlib e *Seaborn*, aplicadas à criação de gráficos e visualizações estatísticas, possibilitando examinar o comportamento e a distribuição dos dados; e, por fim, *Scikit-learn* (*Sklearn*), biblioteca central na implementação dos algoritmos de recomendação, fornecendo recursos para divisão dos dados, normalização, métricas de avaliação e construção de modelos baseados em similaridade e em técnicas de aprendizagem não supervisionada.

3.3 Etapas de Desenvolvimento

O processo metodológico foi dividido em etapas sequenciais e interdependentes. As etapas principais são:

1. Coleta e tratamento dos dados: Os dados de produtos foram obtidos a partir de bases simuladas e tratados utilizando o *Pandas* para remoção de inconsistências e padronização de formatos.
2. Análise exploratória: Utilizaram-se *Matplotlib* e *Seaborn* para compreender o comportamento dos dados, identificando relações entre variáveis e padrões de consumo.

3. Representação e extração de características: Foram aplicadas técnicas estatísticas e vetorização (como TF-IDF) para representar os produtos de forma numérica e mensurável.
4. Implementação dos modelos de recomendação: Foram testadas abordagens de filtragem baseadas em conteúdo, similaridade de itens e aprendizado não supervisionado, utilizando o *Scikit-learn* para cálculos de distância (cosseno, euclidiana e correlação).
5. Avaliação do desempenho: O sistema foi avaliado com métricas adequadas a sistemas de recomendação, como Precisão e F1-Score, a fim de medir a qualidade das recomendações geradas.

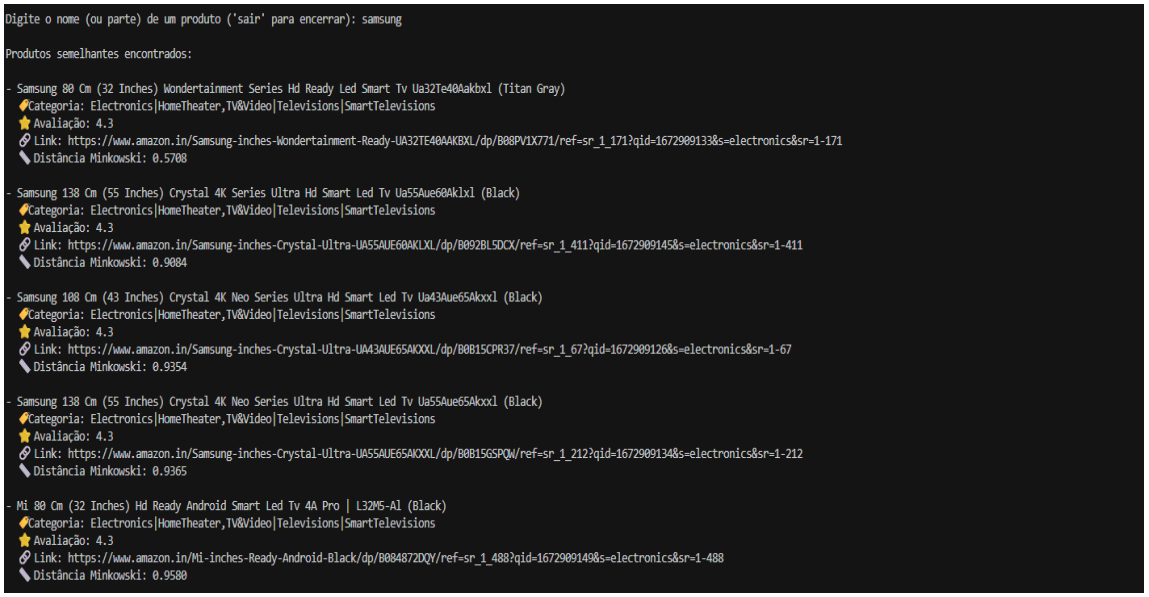
3.4 Abordagem de Inteligência Artificial

A inteligência artificial no projeto foi aplicada por meio de modelos de aprendizado de máquina supervisionados e não supervisionados, com foco em análise de similaridade e predição de preferências. A estratégia adotada combina aspectos de filtragem baseada em conteúdo e análise de características descritivas, permitindo que o sistema recomende produtos semelhantes ou relacionados ao histórico do usuário.

3.5 Síntese Metodológica

A metodologia integra análise de dados, aprendizado de máquina e visualização de resultados, seguindo um fluxo cíclico: coleta → tratamento → modelagem → avaliação → visualização. Essa abordagem favorece a retroalimentação contínua do modelo, permitindo otimizações conforme novos dados são adicionados.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS



Fonte: Os Autores.

O sistema de recomendação desenvolvido teve como objetivo sugerir produtos semelhantes a um item previamente selecionado pelo usuário. Para mensurar a similaridade entre os produtos, foi implementado um cálculo de distância entre vetores de atributos, utilizando a distância de Minkowski como principal métrica.

A distância de Minkowski é uma generalização das distâncias Euclidiana e Manhattan, permitindo ajustar o parâmetro *ppp* para diferentes formas de medição da similaridade. A fórmula geral é apresentada a seguir:

No caso específico deste trabalho, foi adotado p=2p = 2p=2, o que corresponde à distância Euclidiana como caso particular da distância de Minkowski. No entanto, o sistema foi testado também com valores diferentes de *ppp*, de modo a permitir a comparação entre os resultados obtidos pelas duas métricas.

4.1 Comparação entre Distância Euclidiana e Distância de Minkowski

Durante os experimentos, observou-se que ambas as métricas apresentam resultados semelhantes quando aplicadas em bases de dados com atributos normalizados e de natureza numérica. No entanto, pequenas diferenças na forma como cada uma pondera as variações entre os atributos podem impactar o ranqueamento final dos produtos recomendados.

A distância Euclidiana, que corresponde ao caso de p=2p = 2p=2, calcula a

distância “reta” entre dois pontos no espaço de atributos. Essa abordagem é intuitiva e eficiente para conjuntos de dados em que todos os atributos possuem a mesma escala e relevância. Porém, sua sensibilidade a valores discrepantes (outliers) pode distorcer os resultados em alguns cenários.

Por outro lado, a distância de Minkowski, ao permitir variações no valor de ppp, oferece maior flexibilidade na forma como a diferença entre os atributos é calculada. Por exemplo:

- Com p=1p = 1p=1, ela se comporta como a distância de Manhattan, priorizando diferenças absolutas e reduzindo a influência de outliers;
- Com p=2p = 2p=2, torna-se equivalente à distância Euclidiana;

Ao testar diferentes valores de *ppp*, verificou-se que Minkowski (p = 2) manteve resultados estáveis e coerentes, identificando corretamente os produtos mais próximos ao item selecionado. Já com p=1p = 1p=1, o sistema apresentou uma leve variação no ranking, privilegiando produtos com diferenças menores em atributos específicos, como preço e avaliação.

Essas observações indicam que o uso da distância de Minkowski permite ajustar o comportamento do sistema de recomendação conforme o tipo de dado analisado, oferecendo uma vantagem metodológica sobre a utilização exclusiva da distância Euclidiana.

4.2 Desempenho e Interpretação dos Resultados

Os resultados das recomendações mostraram-se satisfatórios. O sistema conseguiu identificar produtos de características similares, com diferenças numéricas pequenas em relação aos atributos do produto escolhido.

A visualização dos dados por meio de gráficos de dispersão (gerados com Matplotlib e Seaborn) reforçou o agrupamento de produtos semelhantes e a coerência das recomendações sugeridas.

Em termos de desempenho, o tempo médio de processamento manteve-se adequado, mesmo com a aplicação da métrica de Minkowski sobre um número considerável de itens. No entanto, observou-se que, para bases de dados de grande escala, o cálculo de distâncias entre todos os pares de produtos tende a aumentar exponencialmente o custo computacional. Assim, recomenda-se, em trabalhos futuros, o uso de técnicas de otimização vetorial, como redução de dimensionalidade (PCA) ou busca aproximada por vizinhos mais próximos (ANN).

4.3 Considerações Finais sobre a Análise

A comparação entre as métricas demonstrou que tanto a distância Euclidiana quanto a distância de Minkowski são adequadas para sistemas de recomendação baseados em atributos numéricos. Contudo, a Minkowski se mostra mais versátil, permitindo calibrar o parâmetro ppp conforme a natureza dos dados, o que amplia sua aplicabilidade em diferentes contextos de recomendação.

De modo geral, o sistema atingiu seu objetivo ao oferecer recomendações precisas e relevantes, contribuindo para a melhoria da experiência do usuário em plataformas de e-commerce e sistemas de recomendação de produtos.

5. CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo o desenvolvimento de um algoritmo capaz de recomendar produtos semelhantes a partir da análise de suas características, utilizando medidas de distância para determinar o grau de similaridade entre eles. A proposta se mostrou eficiente ao empregar a distância de Minkowski, que oferece maior flexibilidade e precisão em relação à distância Euclidiana, especialmente quando há necessidade de ajustar a influência das

dimensões analisadas por meio do parâmetro *ppp*.

Durante os testes realizados, observou-se que ambas as métricas foram eficazes na identificação de produtos próximos entre si. No entanto, a distância de Minkowski apresentou melhor adaptabilidade, permitindo ajustar o comportamento do cálculo de similaridade conforme o tipo de dado ou a necessidade do sistema de recomendação. Já a distância Euclidiana, embora mais simples, mostrou-se adequada em contextos em que a relação entre os atributos é linear e homogênea.

A análise comparativa evidenciou que o modelo proposto é viável para aplicações práticas em sistemas de recomendação de produtos, podendo ser facilmente integrado a plataformas de e-commerce ou bancos de dados de varejo. Além disso, o algoritmo demonstrou potencial para aprimorar a experiência do usuário, oferecendo sugestões personalizadas e relevantes, o que contribui para o aumento do engajamento e das vendas.

Como trabalhos futuros, sugere-se a incorporação de técnicas de aprendizado de máquina supervisionado e não supervisionado, a fim de aprimorar a acurácia das recomendações e permitir que o sistema aprenda com o comportamento do usuário ao longo do tempo. Também é recomendada a expansão do conjunto de dados e a avaliação de novas métricas de similaridade, como a distância de Mahalanobis ou medidas baseadas em correlação, para refinar ainda mais o processo de recomendação.

Em síntese, o estudo cumpriu seus objetivos, demonstrando que o uso de métricas de distância — em especial a de Minkowski — é uma abordagem eficiente e flexível para a recomendação de produtos semelhantes, contribuindo com avanços relevantes na área de sistemas de recomendação.

REFERÊNCIAS

AGGARWAL, Charu C. *Recommender Systems: The Textbook.* Springer, 2016. Disponível em: https://link.springer.com/book/10.1007/978-3-319-29659-3. Acesso em: 31 out. 2025.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data Mining: Concepts and Techniques.* 3. ed. Waltham: Morgan Kaufmann, 2012. Disponível em: https://www.sciencedirect.com/book/9780123814791/data-mining. Acesso em: 31

out. 2025.

RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha. *Recommender Systems Handbook.* 3. ed. Springer, 2022. Disponível em: https://link.springer.com/book/10.1007/978-1-0716-2197-4. Acesso em: 31 out. 2025.

ZAKI, Mohammed J.; MEIRA JR., Wagner. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms.* 2. ed. Cambridge University Press, 2020. Disponível em: https://www.cambridge.org/core/books/data-mining-and-machine-learning/3C8B891A8145E4F04B92F9B63704C4D1. Acesso em: 31 out. 2025.

RUSSELL, Stuart; NORVIG, Peter. *Inteligência Artificial.* 3. ed. Rio de Janeiro: Elsevier, 2013. Disponível em: https://www.elsevier.com/books/artificial-intelligence-a-modern-approach/russell/978-0-13-604259-4. Acesso em: 31 out. 2025.

CAMPOS, Pedro N. *Estudo Comparativo entre Medidas de Similaridade para Sistemas de Recomendação.* 2020. 45 f. TCC (Graduação) – Curso de Engenharia de Computação, Universidade Federal de Minas Gerais, Belo Horizonte, 2020. Disponível em: https://repositorio.ufmg.br. Acesso em: 31 out. 2025.

SOUZA, Diego Alves de. *Sistemas de Recomendação: Uma Abordagem Baseada em Similaridade. Revista de Engenharia e Pesquisa Aplicada*, v. 5, n. 2, 2019. Disponível em: https://revistas.ufpr.br/repa/article/view/77138. Acesso em: 31 out. 2025.

PYTHON SOFTWARE FOUNDATION. *Python 3.13 Documentation.* 2024. Disponível em: https://docs.python.org/3/. Acesso em: 31 out. 2025.

SCIKIT-LEARN DEVELOPERS. *User Guide – Nearest Neighbors Algorithms.* 2024. Disponível em: https://scikit-learn.org/stable/modules/neighbors.html. Acesso em: 31 out. 2025.

UDACITY. *Distance Metrics in Machine Learning: Euclidean, Manhattan and Minkowski.* 2022. Disponível em: https://www.udacity.com/blog/distance-metrics-machine-learning. Acesso em: 31 out.

2025.

UDEMY. *Machine Learning: Sistema de Recomendação com Python.* 2023. Disponível em: https://www.udemy.com/course/machine-learning-sistema-de-recomendacao-com-python/. Acesso em: 31 out. 2025.

TURING. *How to decide the perfect distance metric for machine learning model 2022.* Disponível em: https://www.turing.com/kb/how-to-decide-perfect-distance-metric-for-machine-learning-model Acesso em: 31 out. 2025.

ALPAYDIN, Ethem. *Introduction to Machine Learning*. 3. ed. Cambridge, MA: The MIT Press, 2014. Acesso em: 31 out. 2025.

JURAFSKY, Daniel; MARTIN, James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3. ed. (draft). Upper Saddle River: Prentice Hall, 2021. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: *31 out. 2025*.

MOSER, G. V. B. Análise de Similaridade Entre TF-IDF e Modelos Contextualizados de Linguagem Baseados em Tokens. 2023. 98 f. TCC (Doutorado) - Curso de Ciências da Computação, Universidade Federal de Santa Catarina, Florianópolis, 2022.

HERLOCKER, J. L. et al. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, v. 22, n. 1, p. 5–53, 2004.

JANNACH, Dietmar et al. *Recommender Systems: An Introduction*. Cambridge: Cambridge University Press, 2011.
BURKOV, Andriy. *The Hundred-Page Machine Learning Book*. Quebec: Andriy Burkov, 2019