**Sistema de recomendação de filme baseado no perfil do usuário**

**Enrico Najjar Galdeano, Gabriel Shihao Chen Yin, Samuel Lopes Pereira**

Faculdade de Computação e Informática (FCI)  
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

{10402924, 10408981, [10403767}@mackenzista.com.br](mailto:10403767}@mackenzista.com.br)

Link Github: https://github.com/Enrico104029/Projeto-IA-.git

***Resumo.*** O objetivo do projeto é o desenvolvimento de um sistema de recomendações de filmes personalizado que usa como base o estudo de K-Nearest Neighbors (KNN) (GEEKSFORGEEKS, 2017). O processo consiste em analisar o perfil do usuário com base em suas preferências de gêneros e autores, na intenção de facilitar a escolha dos conteúdos disponíveis nas principais plataformas de streaming. A proposta se justifica dada a dificuldade que os usuários encontram ao escolher um título disponível, mesmo sabendo quais são suas preferências. Serão usadas para o desenvolvimento do sistema técnicas de aprendizado de máquina, que serão responsáveis pelo algoritmo de indicação assertivo aos conteúdos relevantes por perfil.

**1. Introdução**

**a. Contextualização**

Com o crescimento acelerado das plataformas de *streaming*, o consumo de filmes e séries aumentou significativamente nos últimos anos (SOUZA; BARTH, 2024). Contudo, a grande quantidade de títulos disponíveis pode gerar indecisão e frustração nos usuários, que gastam tempo excessivo na busca por conteúdos de seu interesse. Nesse cenário, sistemas de recomendação inteligentes surgem como ferramentas essenciais para personalizar a experiência do usuário.

**b. Justificativa**

A justificativa para o desenvolvimento deste projeto está relacionada à necessidade de soluções que otimizem o processo de escolha de filmes e ampliem a satisfação do consumidor. Ao empregar técnicas de inteligência artificial e redes neurais em grafos, é possível capturar de maneira mais precisa as relações entre usuários e filmes, oferecendo recomendações mais adequadas em comparação com métodos tradicionais. Além disso, o crescimento do mercado de streaming reforça a importância de mecanismos inovadores e escaláveis para melhorar a experiência do usuário.

**c. Objetivo**

O objetivo principal do projeto é desenvolver um sistema de recomendação de filmes baseado em K-Nearest Neighbors (KNN), capaz de analisar as preferências do usuário e indicar conteúdos similares em termos de gênero e autor. Como objetivo secundário, pretende-se avaliar a eficiência do modelo no conjunto de dados TMDB 5000, verificando a acurácia e a relevância das recomendações.

**d. Opção do projeto**

O projeto se enquadra na área de Machine Learning, mais especificamente na subárea de filtragem baseada em conteúdo, por utilizar o algoritmo K-Nearest Neighbors (KNN) para encontrar a similaridade entre os próprios filmes. A escolha dessa abordagem se deve ao fato de que ela é um método simples e eficaz para encontrar filmes com características (gênero e autor) semelhantes. O KNN permite explorar de forma eficiente essas relações, possibilitando a geração de recomendações personalizadas com alta precisão, encontrando os filmes mais "próximos" às preferências do usuário. Dessa forma, a opção pelo uso do KNN se mostra adequada para o desenvolvimento de um sistema de recomendação eficiente.

**2. Fundamentação Teórica**

Os sistemas de recomendação podem ser classificados em três tipos principais: baseados em conteúdo, colaborativos e híbridos (ISINKAYEVA; ISINKAYEV, 2021). O sistema proposto utiliza uma abordagem baseada em conteúdo, que se concentra em recomendar itens que são semelhantes àqueles que o usuário demonstrou interesse no passado. Neste projeto, o algoritmo K-Nearest Neighbors (KNN) será a técnica principal. O KNN é um método não-paramétrico usado para classificação e regressão. Ele funciona encontrando os "k" vizinhos mais próximos de um ponto de dados e fazendo uma previsão com base neles. No nosso caso, o KNN será usado para encontrar filmes que sejam os mais "próximos" às preferências de gênero e autor do usuário. A biblioteca scikit-learn será utilizada para a implementação do KNN, enquanto o dataset TMDB 5000 fornecerá a base de dados para o desenvolvimento e a avaliação.

**3. Descrição do Problema**

O problema central identificado é a dificuldade que usuários enfrentam para encontrar conteúdos que atendam plenamente aos seus interesses em meio à grande oferta de títulos em plataformas de *streaming*. Esse excesso de opções pode gerar indecisão, perda de tempo e frustração, um fenômeno conhecido como o Paradoxo da Escolha (SCHWARTZ, 2004). Métodos tradicionais de recomendação, como filtros simples ou listas genéricas, não são suficientes para capturar a complexidade das preferências individuais. Dessa forma, enxerga-se necessária uma abordagem mais sofisticada, que considere as características dos filmes, como gênero e autor, para oferecer recomendações mais precisas.

**4. Ética na Inteligência Artificial**

O uso de inteligência artificial em sistemas de recomendação envolve questões éticas importantes. Um dos principais desafios é evitar a criação de bolhas de filtro, em que o usuário é exposto apenas a conteúdos semelhantes aos já consumidos, reduzindo a diversidade cultural. Outro aspecto relevante é a transparência: o usuário deve compreender que as recomendações são geradas por algoritmos e que podem conter vieses presentes nos dados de treinamento. Além disso, é necessário garantir a privacidade e a segurança dos dados pessoais dos usuários, respeitando normas legais de proteção de dados. No desenvolvimento desta solução, a responsabilidade recai sobre os criadores em garantir que o sistema não discrimine, respeite a diversidade de interesses e atue como um facilitador da experiência de entretenimento.

**5. Dataset**

A origem do dataset utilizado neste projeto é a TMDB (The Movie Database). Serão utilizados dois arquivos de dados públicos, disponíveis para todos os usuários da plataforma: tmdb\_5000\_movies.csv e tmdb\_5000\_credits.csv. Juntos, eles reúnem informações sobre os 5.000 filmes mais votados e seus respectivos elencos e equipes de produção.

As colunas presentes nos datasets, incluem:

**tmdb\_5000\_movies.csv:**

ID: Identificação única de cada filme.

Title: Título do filme.

Genres: Gêneros do filme.

Original\_language: Língua original em que o filme foi produzido.

Overview: Resumo ou sinopse do filme.

Popularity: Métrica que indica o quão popular o filme é.

Release\_date: Data de estreia do filme.

Vote\_average: Média das avaliações recebidas.

Vote\_count: Quantidade total de votos recebidos.

**tmdb\_5000\_credits.csv:**

Movie\_id: Identificador que se conecta ao ID do arquivo de filmes.

Title: Título do filme.

Cast: Lista de atores principais.

Crew: Lista de pessoas que trabalharam no filme.

**6. Metodologia**

Na primeira fase do projeto, os dados passarão por um processo de pré-processamento, no qual será verificada a relevância de cada coluna, realizado o tratamento de valores nulos ou ausentes, e, dependendo da relevância, algumas colunas poderão ser combinadas. Também será feita uma análise exploratória para compreender melhor o conjunto de dados.

Na segunda fase, os dados serão transformados para o uso com o algoritmo K-Nearest Neighbors (KNN). Isso envolverá a extração dos dados de gênero e autores de cada filme e sua conversão em um formato vetorial (Vaz), permitindo a medição de similaridade entre os filmes.

Na terceira etapa, o algoritmo KNN (GEEKSFORGEEKS, 2017) será utilizado para gerar as recomendações. Com base nas preferências de gênero e autor do usuário, o KNN irá identificar os "k" filmes mais próximos ou similares no dataset, utilizando uma métrica de similaridade. A lista de filmes mais próximos às preferências do usuário será então apresentada como a recomendação final.

**7. Resultados Esperados**

Ao final do projeto, espera-se que o sistema seja capaz de recomendar um filme com base em uma entrada que represente o perfil do usuário, contendo informações como gêneros e atores favoritos.

**8. Referências bibliográficas**

SOUZA, D.; BARTH, M. Impactos no consumo de conteúdo audiovisual em plataformas de streaming durante a pandemia de Covid-19. **Revista UNINTER de Comunicação**, Curitiba, v. 10, n. 17, p. 69–80, 2024. Disponível em: <https://www.revistasuninter.com/revistacomunicacao/index.php/revista/article/view/886>. Acesso em: 06/09/2025.

ISINKAYEVA, G.; ISINKAYEV, M. The fundamentals of recommendation systems: collaborative filtering, content-based filtering, and hybrid methods. In: **International Conference on Information Technology and Applications (ICITA)**, 2021.

SCHWARTZ, Barry. O paradoxo da escolha: Por que mais é menos. Tradução de Bárbara Bentes. São Paulo: A Girafa, 2004.

GEEKSFORGEEKS. **KNearest Neighbor(KNN) Algorithm**. Disponível em: <https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/>.

VAZ, A. L. One-hot-encoding, o que é? Disponível em: <https://arthurlambletvaz.medium.com/one-hot-encoding-o-que-%C3%A9-cd2e8d302ae0>.

‌