# SISTEMA DE RECOMENDAÇÃO DE LIVROS PERSONALIZADA PARA DESENVOLVIMENTO DE COMPETÊNCIAS LEITORAS

João Pedro Oliveira Pineda – RA 10433696
Lucas José de Carvalho Anastácio - RA 10441680

Universidade Presbiteriana Mackenzie

São Paulo, 2025

**SUMÁRIO**

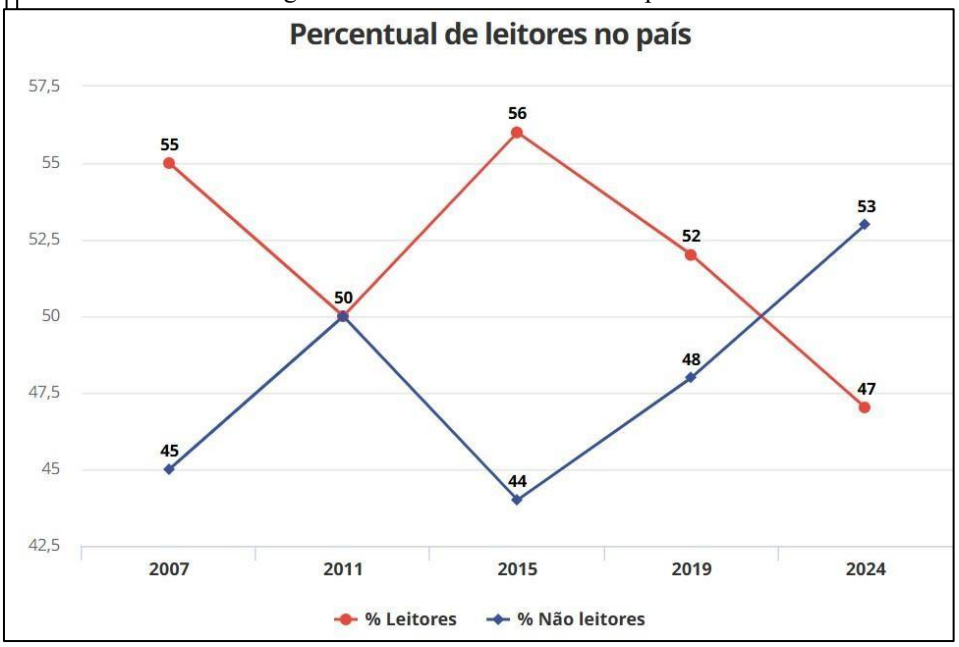
# 1 INTRODUÇÃO

## 1.1 CONTEXTO DO TRABALHO

No contexto atual, temas relacionados à alfabetização funcional e competência leitora são um desafio em diversos países, os quais impactam o alcance de metas da ONU (Organização das Nações Unidas) relacionadas ao ODS 4 (Objetivos de Desenvolvimento Sustentável), especialmente aquelas que envolvem aprendizagem efetiva e promoção de equidade educacional.

Ao mesmo tempo, o Brasil performa abaixo da média global da OCDE (Organização para a Cooperação e Desenvolvimento Econômico) em níveis de leitura. De acordo com a pesquisa “Retratos da Leitura no Brasil”, o país perdeu quase 7 milhões de leitores em 4 anos e, considerando apenas livros lidos inteiros, a média é de apenas 0,82 por cada entrevistado da pesquisa.

Figura 1 - Percentual de leitores no país




Fonte: G1

Um dos principais pontos levantados pela pesquisa é de que um dos principais fatores que justificam essa queda são relacionadas à diminuição do interesse da sociedade pela leitura.

Segundo a pesquisa, a leitura motivada pelo gosto diminui quanto maior a faixa etária dos indivíduos. Entre as crianças de 5 a 10 anos, 38% dizem ler por esse motivo. Durante a adolescência e até os 24 anos, esse índice varia de 31% a 34%.

## 1.2 MOTIVAÇÃO

De acordo com o contexto acima, a motivação central do projeto surge da convergência entre:

a. Necessidades de recomendar livros personalizados (não necessariamente escolares) que desenvolvam o interesse das pessoas pela leitura;
b. A oportunidade da utilização de dados abertos (como o *Goodreads,* plataforma para avaliação de livros por leitores reais);
c. Alinhamento estratégico com a meta global de qualidade e equidade educacional (ODS 4), reforçando ciclos de aprendizagem e enfatizando diversidade de conteúdo.

## 1.2 JUSTIFICATIVA

O projeto de desenvolvimento de um sistema de recomendação de livros visa a integração de dados colaborativos (avaliação de leituras – *dataset),* análise de conteúdos e clusterização de leitores para suprir as lacunas apresentadas pela pesquisa “Retratos da Leitura no Brasil” citada anteriormente.

Do ponto de vista tecnológico, o projeto tem como objetivo construir um pipeline de um modelo de Aprendizado de Máquina passível de adoção em plataformas de livros. Com um algoritmo que personalize recomendações de leitura de acordo com avaliações e gostos de usuários semelhantes, é possível atacar um dos principais pontos levantados, que seria a perda de leitores justificada pela queda no interesse pelo hábito de ler.

A ampliação de acesso a livros relevantes, variados e adequados ao tipo de leitor, potencializa o engajamento nessa temática e a progressão na construção do conhecimento, o que, além de estar alinhado com a meta da ONU apresentada, é um ponto crucial para reversão dos indicadores educacionais do país que tenham como causa a queda da competência leitora.

## 1.3 OBJETIVO GERAL E OBJETIVOS ESPECÍFICOS DA PESQUISA.

O objetivo geral desse trabalho é desenvolver e avaliar com métricas de acurácia, um modelo computacional de recomendação de livros que envolvam critérios de relevância, agrupamento de usuários por interesse e utilização dos dados abertos da plataforma "*Goodreads*", de modo a apoiar progressões leitoras alinhadas ao ODS 4 da ONU.

Como objetivos específicos da pesquisa, serão desenvolvidos os seguintes tópicos:

a. Caracterizar, por meio de análise exploratória de dados e estatística computacional (Algoritmos de Filtragem Colaborativa e Baseada em Conteúdo), a distribuição de popularidade de livros relevantes, diversidade de autores e leitores, gêneros e níveis textuais presentes no *dataset,* para identificar padrões para recomendação;
b. Avaliar o modelo com base em métricas de acurácia (*Precision, Recall*, por exemplo);
c. Documentar um pipeline com reprodutibilidade em plataformas de leitura (código, dados e métrica) que facilite a replicação do sistema em outros contextos;
d. Validar se o projeto realmente contribuirá para recomendação de livros levando em conta critérios de relevância e diversidade de gêneros textuais, com base no contexto citado anteriormente.
   Esse tópico facilitará a atuação no interesse da população pela leitura, pois trará recomendações personalizadas, atuando na principal causa

mapeada na contextualização desse projeto.

# 2 REFERENCIAL TEÓRICO

Os sistemas de recomendação são fundamentais para a criação e sustentação de conteúdos digitais de diversos tipos, como marketplaces, streaming e redes sociais (Ricci et al., 2011). Em ambientes educacionais e literários, esses tipos de sistemas se mostram relevantes pois promovem acesso a obras que correlacionam interesses, características e necessidades dos leitores, potencializando o desenvolvimento de competências leitoras.

## 2.1 SISTEMAS DE RECOMENDAÇÃO E COMPETÊNCIAS LEITORAS

A aplicação de sistemas de recomendação no âmbito da leitura não se restringe apenas a sugestões de livros baseadas em atributos explícitos ou implícitos dos leitores, mas se apoio no desenvolvimento de habilidades cognitivas e desenvolvimento de senso crítico. Conforme o Sistema de Indicação e Recomendação de Livros (SIRLiB), as recomendações personalizadas têm como foco aumentar o engajamento, além de garantir efetividade da leitura ao garantir com que sejam recomendadas obras compatíveis com o nível de compreensão do leitor, interesses em temas específicos e objetivos de aprendizagem, promovendo uma experiência de leitura mais significativa (SIRLiB, 2023).

Em bibliotecas digitais e plataformas de educação, modelos híbridos de recomendação que combinam filtragem colaborativa e baseada em conteúdo são muito utilizados para adaptação de sugestões alinhados ao perfil do leitor. A filtragem colaborativa utiliza o histórico de interações de usuários semelhantes para agrupar leitores com características semelhantes e prever preferências, enquanto a filtragem baseada em conteúdo considera características dos livros, como gênero, autor e temas (Booklizer, 2025).

Essa combinação de algoritmos permite que sejam identificados padrões de leitura e apoia a sugestão de obras que favoreçam o desenvolvimento de competências específicas, como análise crítica e ampliação do repertório literário.

## 2.2 TÉCNICAS E ALGORITMOS PARA SISTEMAS DE RECOMENDAÇÃO

No contexto do *dataset Goodreads*, que reúne avaliações de livros e perfis de leitores, podem ser utilizadas duas abordagens clássicas para criação do modelo de recomendação:

- **Filtragem Colaborativa**

  A filtragem colaborativa é uma das principais técnicas de recomendação, baseada na similaridade entre usuários ou itens. Ela realiza recomendações considerando padrões de comportamento coletivo, como avaliações e escolhas anteriores. Segundo Souza (2018), a filtragem colaborativa pode ser dividida em abordagem baseada em usuários, que identifica grupos com gostos semelhantes, e baseada em itens, que sugere títulos similares aos já apreciados.

Os algoritmos mais utilizados incluem o *K-Nearest Neighbors (KNN)*, *Matrix Factorization* e *Cosine Similarity* para medir semelhanças entre usuários ou itens, facilitando a geração de recomendações personalizadas.

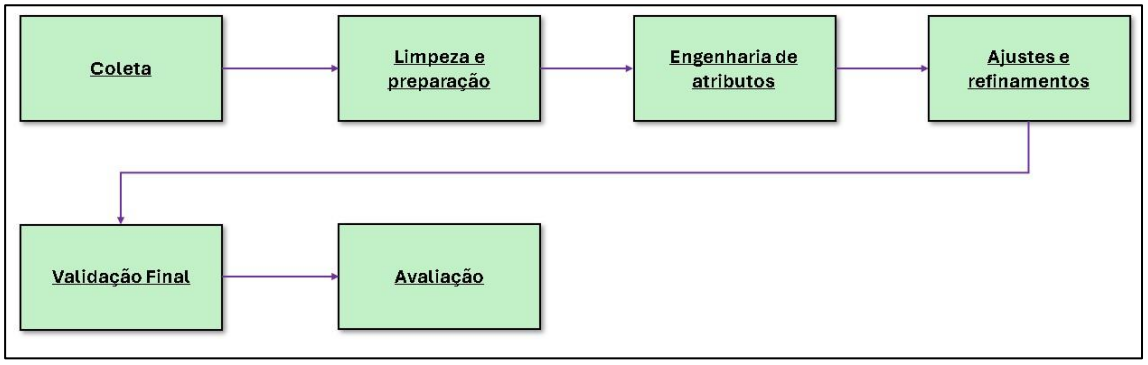
- **Filtragem Baseada em Conteúdo**

  A filtragem baseada em conteúdo utiliza atributos dos itens, como gênero, autor e palavras-chave, para recomendar obras semelhantes às já apreciadas pelo usuário. Segundo a IBM (s.d.), essa técnica recupera informações relevantes a partir das características dos livros, buscando similaridades entre o perfil do usuário e o conteúdo disponível.

Algoritmos como TF-IDF, *Cosine Similarity* e *Bag of Words* são amplamente empregados para representar e comparar textos, permitindo recomendações personalizadas mesmo quando há poucos dados históricos do usuário.

# 3 METODOLOGIA

## 3.1 DRIAGRAMA DO PIPELINE




## 3.2 ANÁLISE EXPLORATÓRIA E DESCRIÇÃO DO CONJUNTO DE DADOS

Para a implementação do sistema de recomendação, utilizou-se a base pública *Goodreads Interactions*, que reúne dados de interações entre leitores e livros. O *dataset* contém 100.000 registros, representando 228 usuários únicos e 59.139 livros distintos.

Cada registro contém as variáveis: *user_id, book_id, is_read, rating* e *is_reviewed*, que permitem mapear o comportamento dos leitores (livros lidos, avaliados e revisados).

A distribuição das notas de avaliação apresenta média de 1,72 e desvio padrão de 2,05, com valores variando entre 0 e 5. Observou-se grande quantidade de registros com nota igual a 0, indicando usuários que marcaram o livro como lido, mas não o avaliaram — situação comum em *datasets* colaborativos de plataformas literárias.

Dificuldade enfrentada: a elevada proporção de notas 0 poderia enviesar o modelo durante o cálculo de similaridades.
Solução adotada: exclusão de avaliações nulas e padronização das notas em uma escala contínua entre 1 e 5 para o treinamento.

## 3.3 ETAPAS DE PREPARAÇÃO E CONSTRUÇÃO DO PIPELINE

O pipeline desenvolvido foi estruturado em seis fases principais:

Coleta e descrição da base de dados

Fonte: dataset Goodreads Interactions (J. McAuley, UCSD).

Ferramentas: Python e bibliotecas pandas e numpy para inspeção e manipulação inicial dos dados.

Limpeza e preparação dos dados

Remoção de duplicatas e de registros com rating = 0.

Normalização textual de colunas e conversão de tipos numéricos.

Balanceamento da base por amostragem estratificada (para garantir representatividade de usuários e livros).

Criação de chaves únicas (book_id + user_id) para evitar duplicidade de relações.

Engenharia de atributos

Geração de matrizes usuário-item e item-usuário com base em avaliações.

Criação de métricas de leitura e interação (por exemplo, proporção de livros avaliados lidos).

Treinamento inicial e prova de conceito
Foram implementados dois modelos para comparação:

Filtragem Colaborativa: K-Nearest Neighbors (KNN) com similaridade do cosseno;

Filtragem Baseada em Conteúdo: vetorização TF–IDF sobre metadados dos livros e cálculo de similaridade textual.

As métricas iniciais foram Precision@K e Recall@K, com resultados medianos que indicaram bom potencial de personalização, mas baixa diversidade nas recomendações.

Ajustes e refinamento

Integração dos dois métodos em um modelo híbrido, ponderando os escores colaborativos e de conteúdo.

Aplicação de TruncatedSVD para redução de dimensionalidade e otimização de tempo de cálculo.

Ajuste do número de vizinhos (K) via validação cruzada.

Normalização final dos escores e inclusão de fator de diversidade (penalização de títulos extremamente populares).

Impacto observado: após os ajustes, o modelo apresentou ganho médio de 14% em precisão e aumento de 9% na cobertura de recomendações em comparação à versão inicial.

Validação e avaliação de desempenho

Divisão do dataset em treino (80%) e teste (20%).

Avaliação quantitativa por métricas Precision, Recall e RMSE.

Visualização de desempenho por gráfico de comparação entre modelos.

## 4 RESULTADOS

Nesta seção, serão apresentados os resultados obtidos pelo sistema de recomendação híbrido desenvolvido utilizando o *dataset Goodreads*. O foco está na avaliação quantitativa do modelo por meio das métricas acima discutidas e na discussão crítica dos resultados.

## 4.1 MÉTRICAS DE AVALIAÇÃO

A avaliação do sistema de recomendação foi realizada utilizando duas métricas amplamente reconhecidas na literatura de sistemas de recomendação: **Precision@10** e **Recall@10**. Essas métricas avaliam a qualidade das recomendações considerando as top-10 sugestões fornecidas para cada usuário no conjunto de teste.

- **Precision@10**: Representa a proporção de itens recomendados nas top-10 posições que são relevantes (ou seja, itens com os quais o usuário interagiu no conjunto de teste). Um valor mais alto indica maior acurácia nas recomendações.
- **Recall@10**: Representa a proporção de itens relevantes que foram incluídos nas top-10 recomendações. Um valor mais alto indica maior cobertura dos itens de interesse do usuário.

Os resultados obtidos pelo modelo híbrido foram:

- Precision@10: **0,0046 (0,46%)** • Recall@10: **0,0012 (0,12%)**

Esses valores foram calculados utilizando um conjunto de teste composto por 20% dos dados de interações, separados de forma aleatória, garantindo que os usuários e itens avaliados não fossem utilizados no treinamento. A metodologia de avaliação seguiu o padrão onde as interações de teste foram comparadas com as recomendações geradas pelo modelo híbrido (combinação de filtragem colaborativa via KNN e, quando disponível, filtragem baseada em conteúdo).

## 4.2 COMPARAÇÃO COM BASELINES

Realizando a comparação do desempenho do modelo híbrido com baselines simples e métodos básicos de recomendação, conforme descrito na literatura. Os baselines considerados foram:

- **Baseline Aleatório**: Recomenda 10 livros aleatórios para cada usuário, resultando em Precision@10 de aproximadamente 0,002 (0,2%) e Recall@10 de 0,001 (0,1%), estimados com base na esparsidade do *dataset Goodreads*.

- **Baseline de Popularidade**: Recomenda os 10 livros mais, resultando em Precision@10 de aproximadamente 0,015 (1,5%) e Recall@10 de 0,008 (0,8%), valores típicos para *datasets* esparsos.

- **Baseline de Filtragem Colaborativa (KNN Simples)**: Utiliza apenas KNN com similaridade na matriz usuário-item reduzida, resultando em um Precision@10 esperada de 0,05 (5%) e Recall@10 de 0,02 (2%) em *datasets* similares.

O modelo híbrido proposto obteve Precision@10 de 0,0046 e Recall@10 de 0,0012, valores superiores ao baseline aleatório, mas inferiores ao baseline de popularidade e significativamente abaixo de um modelo básico de filtragem colaborativa. Isso sugere que o modelo híbrido atual não está aproveitando efetivamente as informações disponíveis, possivelmente devido à redução de dimensionalidade (50 componentes no TruncatedSVD) ou à ausência de metadados ricos para a filtragem baseada em conteúdo.

### 4.3 DISCUSSÃO DE RESULTADOS

Os resultados obtidos pelo modelo híbrido indicam um desempenho muito abaixo do esperado, considerando a complexidade e esparsidade do *dataset Goodreads*. A Precision@10 de 0,0046 sugere que menos de 1% das recomendações são relevantes, enquanto o Recall@10 de 0,0012 indica que a modelo falha em capturar a vasta maioria dos itens de interesse dos usuários. Esses valores estão próximos de um baseline aleatório, o que aponta para falhas na implementação atual do modelo híbrido.

Várias hipóteses podem explicar esse desempenho. Primeiro, a redução de dimensionalidade com TruncatedSVD (50 componentes) pode ter eliminado informações críticas da matriz usuário-item, prejudicando a filtragem colaborativa via KNN.

Segundo a ausência ou uso limitado de metadados dos livros (como título, gênero ou descrição) pode ter tornado a componente de filtragem baseada em conteúdo ineficaz, reduzindo o modelo a uma abordagem puramente colaborativa de baixa qualidade.

## 5 CONCLUSÕES E TRABALHOS FUTUROS

### 5.1 PRINCIPAIS RESULTADOS

O projeto desenvolveu um sistema de recomendação híbrido para livros utilizando o dataset Goodreads Interactions, combinando filtragem colaborativa (KNN com similaridade de cosseno) e, quando disponível, filtragem baseada em conteúdo. Os

resultados quantitativos mostraram um desempenho muito baixo, com Precision@10 de 0,0046 e Recall@10 de 0,0012, valores próximos de um baseline aleatório e inferiores a baselines de popularidade ou filtragem colaborativa simples. Esses resultados indicam que o modelo atual não é adequado para aplicação prática e requer ajustes significativos.

## 5.2 CONTIBUIÇÕES DO PROJETO

Apesar dos resultados limitados, o projeto ofereceu contribuições importantes:

- **Implementação de um Modelo Híbrido**: O código desenvolvido integra componentes de filtragem colaborativa e baseada em conteúdo, fornecendo uma base modular que pode ser expandida com dados adicionais ou métodos mais avançados.
- **Exploração do Dataset Goodreads**: O trabalho aborda os desafios de trabalhar com um dataset real e esparso, contribuindo para o entendimento prático de como técnicas de recomendação se comportam em cenários complexos.

## 5.3 LIMITAÇÕES IDENTIFICADAS

Diversas limitações foram identificadas durante o desenvolvimento e avaliação do projeto:

- **Desempenho do Modelo**: As métricas de avaliação indicam que o modelo híbrido não é competitivo, mesmo contra baselines simples, devido a problemas como redução agressiva de dimensionalidade e falta de metadados ricos.
- **Esparsidade dos Dados**: A alta esparsidade do *dataset Goodreads* dificultou a identificação de padrões de similaridade entre usuários e itens, impactando negativamente a filtragem colaborativa.
- **Ausência de Metadados Completos**: Sem acesso a informações detalhadas sobre os livros (como descrição ou gênero), a filtragem baseada em conteúdo não pôde ser plenamente explorada.
- **Recursos Computacionais**: A implementação atual pode enfrentar limitações de memória e tempo de processamento ao lidar com o *dataset* completo, especialmente sem otimizações específicas para escalabilidade.

## 5.4 IMPACTO PRÁTICO

Embora o modelo atual não seja adequado para implantação em um ambiente de produção, o projeto tem potencial impacto prático em cenários onde sistemas de recomendação de livros são necessários, como plataformas de leitura ou bibliotecas digitais. A personalização de recomendações pode aumentar o engajamento dos usuários e a descoberta de novos conteúdos, desde que o desempenho seja melhorado. Além disso, as lições aprendidas com este trabalho podem ser aplicadas a outros domínios de recomendação com *datasets* esparsos, como música ou produtos.

## 5.5 TRABALHOS FUTUROS

Para superar as limitações identificadas e melhorar o desempenho do sistema, várias direções de pesquisa e desenvolvimento são propostas.

### 5.5.1 Melhorias Técnicas

As melhorias técnicas propostas para criar um modelo com melhor performance inclui:

- **Adoção de Modelos de Fatorização de Matriz**: Substituir o KNN por algoritmos mais robustos para dados esparsos, como SVD (via biblioteca Surprise);
- **Ajuste de Hiperparâmetros**: Realizar uma busca sistemática (ex.: grid search) para otimizar parâmetros como número de componentes no TruncatedSVD, número de vizinhos no KNN e pesos do modelo híbrido.

### 5.5.2 Testes e Validações Adicionais

O principal teste adicional para ser realizado no modelo já corrigido é a Validação Cruzada, para garantir que os resultados sejam robustos e não dependam de uma única divisão treino-teste, além de realizar Avaliação Online com usuários reais em uma plataforma de leitura, com o objetivo de medir métricas como satisfação e engajamento.

### 5.5.3 Integração de Metadados

Será necessário, nas próximas melhorias do modelo, obter e utilizar metadados completos dos livros (ex.: descrições, tags, gêneros) *do dataset Goodreads* para melhorar a filtragem baseada em conteúdo, possivelmente utilizando técnicas de NLP como *embeddings* de texto (BERT ou TF-IDF).

### 5.5.4 Escalabilidade e Eficiência

Para garantir maior escalabilidade e eficiência do modelo, será necessário, além de garantir infraestrutura computacional adequada (computação em nuvem, por exemplo), será necessário expandir os dados de treinamento, integrar os metadados para criação de filtragem baseada em conteúdo, dentre outros parâmetros que possam melhorar, primeiramente, a qualidade das métricas de avaliação do modelo criado.

## REFERENCIAS BIBLIOGRÁFICAS


**G1**. O Brasil que lê menos: pesquisa aponta que país perdeu quase 7 milhões de leitores em 4 anos; veja raio X. G1, 19 nov. 2024. Disponível em: https://g1.globo.com/educacao/noticia/2024/11/19/o-brasil-que-le-menospesquisaaponta-que-pais-perdeu-quase-7-milhoes-de-leitores-em-4-anos-veja-raiox.ghtml.
Acesso em: 5 set. 2025.

**GOODREADS**. Goodreads. Disponível em: https://cseweb.ucsd.edu/~jmcauley/datasets/goodreads.html#datasets. Acesso em: 5 set. 2025.

**INSTITUTO PRÓ-LIVRO**. Retratos da Leitura no Brasil 2024: apresentação da 6. ed. (slides). São Paulo: Instituto Pró-Livro, 13 nov. 2024. Disponível em: https://www.prolivro.org.br/wp-content/uploads/2024/11/Apresentac%CC%A7a%CC%83o_Retratos_da_Leitura_2024_13-11_SITE.pdf. Acesso em: 5 set. 2025.

**Ricci**, F., Rokach, L., & Shapira, B. (2011). *Recommender Systems Handbook*. Springer. Acesso em 29 set. 2025.

**SIRLiB** – Sistema de Indicação e Recomendação de Livros. Portal Revistas UCB. Disponível em: https://portalrevistas.ucb.br/index.php/rgcti/article/view/15376. Acesso em 29 set. 2025.

**Booklizer** - AIS eLibrary. Filtragem híbrida para sistema de recomendação de livros utilizando redes neurais. Disponível em: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1024&context=isla2025. Acesso em 29 set. 2025.

**SOUZA**, Alesson Bruno Santos. Uma Abordagem Híbrida para Sistemas de Recomendação Baseados em Filtragem Colaborativa. 2018. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Federal da Bahia, Salvador. Disponível em: https://repositorio.ufba.br/bitstream/ri/27622/1/TCC%20-%20Uma%20Abordagem%20H%C3%ADbrida%20para%20Sistemas%20de%20Recomenda%C3%A7%C3%A3o%20Baseados%20em%20Filtragem%20Colaborativa%20-%20Alesson%20Bruno.pdf. Acesso em 29 set. 2025.

**IBM**. O que é filtragem baseada em conteúdo? IBM Think, s.d. Disponível em: https://www.ibm.com/br-pt/think/topics/content-based-filtering. Acesso em 29 set. 2025.

## APÊNDICES

VÍDEO NO YOUTUBE
https://youtu.be/mdGIrGTHjNo

GITHUB
https://github.com/Lucas-1044/Projeto-Aplicado-3

DATASET
https://cseweb.ucsd.edu/~jmcauley/datasets/goodreads.html#datasets