# 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**



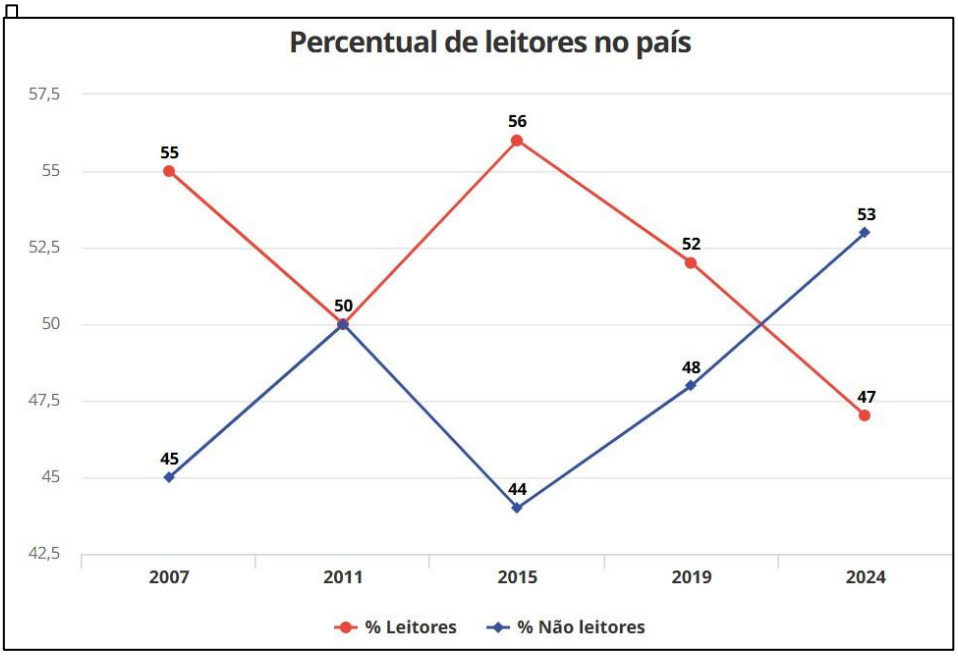
# 1 INTRODUÇÃO

## 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.

**Abaixo segue o link do Github deste projeto:**

https://github.com/AllanaSS/ProjetoAplicado3_Grupo25/tree/main

## 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 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.

## 3.2 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.

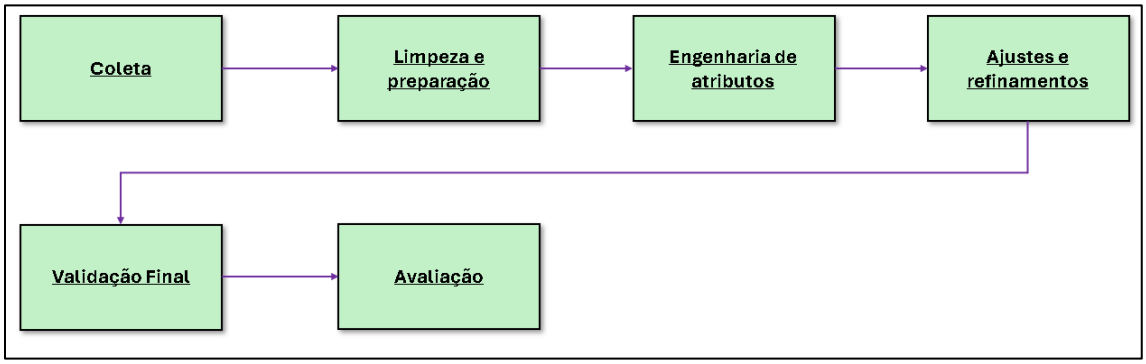
Validação e avaliação de desempenho

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

Avaliação quantitativa por métricas Precision@K, Recall@K e RMSE.

Visualização de desempenho por gráfico de comparação entre modelos.

## 3.2 DRIAGRAMA DO PIPELINE




## 4 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-menos-pesquisaaponta-que-pais-perdeu-quase-7-milhoes-de-leitores-em-4-anos-veja-raio-x.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.