COMPONENTE CURRICULAR:	PROJETO APLICADO III
NOME COMPLETO DO ALUNO:	CRISTINA ALMEIDA DA SILVA – RA 10424207 BRENDA LOUIZE DE O. SOUSA CABRAL – RA 10424949 ÉLIDA ROSA DE PAIVA SOUZA – RA 10424468 ISABEL CABRAL VIEIRA DE SOUSA – RA 1042479

SISTEMA DE RECOMENDAÇÃO DE LIVROS

maio/2025

1. INTRODUÇÃO

Na sociedade contemporânea, marcada pela abundância de informações digitais, identificar conteúdos relevantes tornou-se um desafio recorrente para usuários de diferentes perfis. Em ambientes digitais como plataformas de leitura, livrarias online e bibliotecas virtuais, a vasta quantidade de títulos disponíveis frequentemente resulta em sobrecarga informacional, dificultando a descoberta de obras alinhadas aos interesses individuais dos leitores (Ricci, Rokach & Shapira, 2022). Essa realidade despertou o interesse por soluções que possam tornar essa experiência mais eficiente, personalizada e intuitiva. Nesse contexto, surgiu a motivação para investigar como sistemas de recomendação podem contribuir para otimizar a descoberta de livros em ambientes digitais.

A escolha do tema foi impulsionada pela observação de que, apesar do avanço das tecnologias de informação, muitos leitores ainda enfrentam dificuldades para encontrar obras alinhadas aos seus gostos e necessidades. Plataformas de leitura e livrarias digitais oferecem milhares de títulos, mas nem sempre proporcionam sugestões realmente personalizadas. Esse cenário evidenciou a existência de um problema: **como recomendar livros relevantes a usuários com pouca ou nenhuma interação prévia, promovendo uma experiência de leitura mais rica, personalizada e inclusiva?**

A hipótese central do projeto é que a aplicação de técnicas de aprendizado de máquina, especialmente a filtragem colaborativa — como o algoritmo Singular Value Decomposition (SVD) — pode gerar boas recomendações mesmo em bases de dados esparsas, mitigando problemas como início frio e baixa diversidade (Koren, Bell, Volinsky, 2009).

Para isso, será utilizada a base de dados pública Book-Crossing Dataset, que contém informações sobre livros, usuários e avaliações. Essa base será tratada e adaptada para o framework Surprise, possibilitando a criação de um modelo funcional de recomendação de livros. O desenvolvimento será realizado por meio de um pipeline estruturado, contemplando as etapas de pré-processamento, modelagem, avaliação e comparação entre algoritmos, com foco na clareza, reprodutibilidade e modularização.

Como forma de estabelecer uma referência de desempenho, será também utilizado um modelo de baseline (BaselineOnly), baseado em médias globais, permitindo verificar os ganhos proporcionados pelo uso do SVD. Embora o foco deste projeto esteja centrado na filtragem colaborativa, a estrutura implementada poderá ser estendida futuramente para abordagens híbridas ou baseadas em conteúdo.

Além do valor técnico e acadêmico, o projeto também possui uma dimensão social, ao propor soluções que incentivam o hábito da leitura, democratizam o acesso ao conhecimento e promovem práticas educacionais mais eficientes e sustentáveis — alinhadas aos Objetivos de Desenvolvimento Sustentável (ODS), como Educação de Qualidade (ODS 4), Inovação e Infraestrutura (ODS 9) e Redução das Desigualdades (ODS 10).

Dessa forma, este projeto busca aplicar conhecimentos adquiridos ao longo da graduação em Ciência de Dados para propor uma solução tecnológica com impacto prático, tanto no setor editorial quanto na promoção da inclusão digital e educacional.

1.1. CONTEXTO DO TRABALHO

Com o crescimento exponencial das informações disponíveis na internet, identificar conteúdos relevantes de maneira eficiente tornou-se um desafio significativo para os usuários. Nesse cenário, os sistemas de recomendação surgem como ferramentas essenciais para personalizar a experiência do usuário, auxiliando na descoberta de livros compatíveis com os interesses individuais. Esses sistemas são amplamente utilizados em plataformas de streaming, e-commerce, redes sociais, bibliotecas digitais e ambientes educacionais, impactando diretamente a jornada do leitor com base em dados de comportamento e preferências anteriores (Resnick; Varian, 1997).

Para o desenvolvimento deste projeto, foi selecionada a base de dados Book-Crossing Dataset, originalmente coletada por Cai-Nicolas Ziegler em 2004 e atualmente disponível na plataforma Kaggle (Ruchi, 2022). Essa base contém três arquivos principais: dados dos livros, dos usuários e das avaliações. No entanto, para viabilizar o uso no ambiente de modelagem, os dados foram tratados externamente (em planilha Excel) e consolidados em um único arquivo contendo as informações essenciais: identificador do usuário, ISBN do livro e avaliação atribuída.

Esse conjunto de dados consolidado foi então carregado no ambiente Google Colab e adaptado para o framework **Surprise**, utilizando o objeto Reader para padronização da escala de avaliação e o Dataset.load_from_df para estruturar os dados no formato exigido pela biblioteca. A partir disso, foi possível aplicar técnicas de **filtragem colaborativa**, com destaque para o algoritmo **Singular Value Decomposition (SVD)**, que opera sobre a matriz usuário-item para estimar avaliações ausentes com base em padrões latentes.

Embora a base Book-Crossing ofereça um histórico valioso de interações entre leitores e livros, ela apresenta limitações, como alta esparsidade (muitos usuários com poucas avaliações) e inconsistências nos metadados, o que impõe desafios à modelagem e torna o problema mais realista e representativo dos cenários enfrentados na prática.

1.2. MOTIVAÇÃO

A intensa demanda por soluções personalizadas, sobretudo em um ambiente saturado de informações como a internet, tem impulsionado o desenvolvimento de sistemas de recomendação mais eficazes. Segundo Resnick e Varian (1997), esses sistemas desempenham papel fundamental ao auxiliar os usuários na filtragem de uma vasta gama de opções, promovendo uma experiência mais direcionada e satisfatória.

Este projeto se insere em um cenário contemporâneo no qual o volume de dados e de conteúdo disponível online dificulta a descoberta de informações verdadeiramente relevantes — especialmente em ambientes digitais voltados à leitura, como bibliotecas virtuais, livrarias online e plataformas educacionais. Em tais contextos, leitores frequentemente se deparam com dificuldades para encontrar obras que estejam alinhadas aos seus interesses, evidenciando uma lacuna prática relevante: a ausência de sistemas capazes de oferecer recomendações realmente personalizadas a partir de informações limitadas.

A motivação central deste trabalho está em investigar como algoritmos de aprendizado de máquina, mais especificamente os baseados em filtragem colaborativa, podem ser utilizados para construir um sistema de recomendação de livros eficaz, mesmo diante da esparsidade de dados, como é o caso do *Book-Crossing Dataset*. A escolha do algoritmo Singular Value Decomposition (SVD) se justifica pela sua capacidade de detectar padrões latentes de preferência entre usuários e itens, ainda que o número de interações registradas por indivíduo seja reduzido — uma condição comum em sistemas reais de recomendação (Koren; Bell; Volinsky, 2009).

O uso exclusivo de dados de avaliação explícita (i.e., notas atribuídas aos livros) visa simular um cenário prático em que informações detalhadas sobre usuários ou livros não estão disponíveis — uma limitação frequentemente observada em sistemas reais. Assim, o projeto propõe a construção de um modelo funcional a partir de um conjunto enxuto de variáveis, explorando técnicas consolidadas em ciência de dados e aplicando-as com o suporte da biblioteca Surprise, com foco em reprodutibilidade, modularidade e clareza do pipeline.

Esse estudo se justifica não apenas pelo valor técnico, mas também pelos benefícios que pode proporcionar: ao tornar mais acessível e personalizada a recomendação de livros, o sistema contribui para incentivar o hábito da leitura, aumentar o engajamento com conteúdos relevantes e democratizar o acesso ao conhecimento, especialmente em contextos educacionais e sociais marcados por desigualdades de acesso à informação.

O tema está diretamente relacionado à área de Ciência de Dados, campo em expansão e altamente valorizado no mercado profissional. Ele também guarda forte ligação com a formação acadêmica e interesses profissionais da equipe envolvida, representando uma oportunidade concreta de aplicar conhecimentos em aprendizado de máquina, análise de dados e avaliação de modelos preditivos. O projeto se conecta ainda com possíveis áreas de atuação futura, como o desenvolvimento de sistemas personalizados de apoio à educação, tecnologias para o setor editorial e plataformas digitais inteligentes.

Por fim, ao abordar uma aplicação prática com impacto educacional e social, esta proposta também se alinha a objetivos amplos de desenvolvimento sustentável, especialmente os ODS 4 (Educação de Qualidade), ODS 9 (Inovação e Infraestrutura) e ODS 10 (Redução das Desigualdades), destacando-se como uma iniciativa com potencial de contribuição concreta para a sociedade.

1.3. JUSTIFICATIVA

A escolha de um sistema de recomendação de livros justifica-se pela sua crescente relevância no contexto do mercado editorial digital, que tem migrado de vendas físicas para ambientes online. Em plataformas como livrarias virtuais, bibliotecas digitais e ambientes educacionais, a personalização tem se mostrado uma estratégia fundamental para aumentar o engajamento dos leitores, oferecendo recomendações mais relevantes e ajustadas aos seus interesses individuais (Jannach et al., 2010).

Neste projeto, optou-se por trabalhar com a técnica de filtragem colaborativa baseada em fatoração matricial, mais especificamente o algoritmo SVD (Singular Value Decomposition), por sua comprovada eficácia em bases de dados esparsas. Essa escolha é justificada pelo perfil do Book-Crossing Dataset, que apresenta uma elevada quantidade de usuários com poucas avaliações registradas — um cenário típico de muitos sistemas reais de recomendação.

Embora a base contenha informações sobre livros, usuários e avaliações, para fins de viabilidade técnica e clareza experimental, este projeto focou na utilização da tabela de avaliações tratada externamente (via Excel), utilizando os campos essenciais: identificador do usuário, ISBN e nota atribuída. Essa abordagem simplificada permitiu adaptar rapidamente os dados para uso com a biblioteca Surprise, com a padronização da escala de avaliação e a estruturação dos dados no formato necessário.

O uso do algoritmo BaselineOnly como modelo de referência reforça a justificativa metodológica ao permitir comparar o desempenho do SVD com uma abordagem mais simples baseada em médias. Isso contribui para uma análise mais fundamentada dos ganhos obtidos com a modelagem proposta.

Apesar de técnicas mais avançadas, como modelos híbridos e redes neurais, representarem caminhos promissores, elas exigem maior poder computacional e dados mais complexos, o que extrapolaria o escopo deste trabalho. Ainda assim, sua menção é relevante por indicar possibilidades de continuidade e evolução futura do projeto, conforme a estrutura do pipeline atual permite.

Além do valor técnico, destaca-se o impacto social da proposta, que contribui para promover o hábito da leitura, ampliar o acesso à informação e apoiar estratégias educacionais mais personalizadas e inclusivas. Isso alinha o projeto a objetivos mais amplos de desenvolvimento sustentável, como a promoção de uma educação de qualidade (ODS 4), a inovação tecnológica (ODS 9) e a redução das desigualdades no acesso ao conhecimento (ODS 10).

1.4. OBJETIVO GERAL E OBJETIVOS ESPECÍFICOS DA PESQUISA

Objetivo Geral:

- **Desenvolver um sistema de recomendação de livros** utilizando utilizando **filtragem colaborativa com SVD**, a partir da base **Book-Crossing tratada**, com o objetivo de oferecer sugestões personalizadas que otimizem a experiência de leitura em ambientes digitais.

Objetivos Específicos

- Explorar e compreender o dataset Book-Crossing, realizando o tratamento necessário para adaptação ao framework Surprise;

- Consolidar os dados essenciais (usuário, ISBN e nota) em um formato compatível com o modelo, garantindo a integridade e consistência da base utilizada;
- Implementar uma estratégia de recomendação baseada em **fatoração matricial (SVD)**, avaliando seu desempenho frente a um modelo de baseline;
- Utilizar técnicas de **validação cruzada** e métricas padronizadas (RMSE e MAE) para mensurar a acurácia das recomendações;
- Estruturar um **pipeline modular**, que facilite a manutenção, extensão e comparação de algoritmos de recomendação;
- Comparar o desempenho dos modelos com foco na interpretação prática dos resultados e nos desafios inerentes à esparsidade dos dados;
- Aplicar conhecimentos adquiridos ao longo do curso de Ciência de Dados, especialmente nas áreas de modelagem preditiva, análise crítica e interpretação de métricas;
- Refletir sobre possíveis caminhos de aprimoramento, como a incorporação futura de conteúdo textual, técnicas híbridas e estratégias de feedback contínuo;
- Contribuir para a democratização do acesso à leitura, ampliando o alcance de obras relevantes e promovendo a inclusão digital (ODS 4, 9 e 10).

2. REFERENCIAL TEÓRICO

Sistemas de recomendação (SRs) são tecnologias de filtragem inteligente de informações que buscam prever as preferências de um usuário com base em seu histórico de interações ou em características de outros usuários e itens. Presentes em serviços como Amazon, Netflix, Goodreads e Spotify, esses sistemas desempenham papel estratégico na personalização de conteúdo, agregando valor à experiência do usuário e influenciando diretamente seu engajamento e fidelização (Ricci; Rokach; Shapira, 2022).

Segundo Burke (2002), os SRs podem ser classificados, em geral, em três abordagens principais: filtragem colaborativa, filtragem baseada em conteúdo e modelos híbridos. Cada uma dessas categorias apresenta vantagens e limitações que impactam diretamente sua aplicabilidade e eficácia, dependendo do contexto e da disponibilidade de dados. Com o avanço da inteligência artificial e do aprendizado de máquina, novas técnicas, como redes neurais profundas e aprendizado por reforço, também têm sido integradas aos SRs, especialmente em plataformas dinâmicas e com grande volume de dados (Musto et al., 2021).

2.1. FILTRAGEM COLABORATIVA

A filtragem colaborativa (Collaborative Filtering – CF) é uma das abordagens mais utilizadas e estudadas no campo dos sistemas de recomendação. Seu funcionamento baseia-se na suposição de que usuários que apresentaram preferências semelhantes no passado tenderão a manter padrões similares no futuro (Aggarwal, 2016). Existem duas principais variantes: baseada em memória (user-user ou item-item) e baseada em modelos, com destaque para algoritmos de fatoração matricial, como o Singular Value Decomposition (SVD).

Na filtragem colaborativa baseada em memória, a recomendação é feita a partir da similaridade entre usuários ou itens, calculada por medidas como cosseno, correlação de Pearson ou distância euclidiana. Essa técnica é simples e interpretável, mas sofre com problemas de escalabilidade e desempenho em bases esparsas. Já os métodos baseados em modelos constroem representações latentes dos usuários e itens, permitindo maior capacidade preditiva e generalização em ambientes com grande volume de dados ou alta esparsidade.

O SVD, técnica escolhida neste projeto, realiza a decomposição da matriz de interações usuário-item em três matrizes de fatores latentes. Com isso, torna-se possível estimar avaliações não observadas com base na interação implícita entre os fatores, mesmo quando há poucos registros explícitos. Essa abordagem foi amplamente validada em cenários reais, como no famoso concurso Netflix Prize (Koren; Bell; Volinsky, 2009), demonstrando excelente desempenho em termos de acurácia.

2.2. FILTRAGEM BASEADA EM CONTEÚDO

A filtragem baseada em conteúdo (Content-Based Filtering – CBF) utiliza as características dos itens (como autor, título, gênero ou palavras-chave) para recomendar obras semelhantes às que o usuário já apreciou. Essa abordagem depende fortemente da presença de metadados estruturados e é comum em ambientes como portais educacionais e bibliotecas digitais (Aggarwal, 2022).

Uma vantagem importante da CBF é a sua independência em relação ao histórico de outros usuários, o que mitiga o problema do início frio para novos usuários. No entanto, essa técnica tende a gerar recomendações repetitivas, um fenômeno conhecido como overspecialization, e apresenta dificuldades em sugerir itens fora do perfil já conhecido do usuário, o que pode reduzir a diversidade das recomendações (Burke, 2002).

No presente projeto, essa abordagem não foi adotada devido à ausência de atributos textuais ou metadados na versão tratada da base Book-Crossing utilizada, que contém apenas os campos de usuário, ISBN e nota.

2.3. MODELOS HÍBRIDOS

Modelos híbridos combinam duas ou mais abordagens com o objetivo de superar as limitações de métodos individuais. Eles podem ser construídos de diferentes formas, como combinações sequenciais, ponderadas ou utilizando algoritmos de aprendizado de máquina para integrar múltiplas fontes de informação (Burke, 2002; DelDjoo et al., 2022).

Esses modelos têm sido amplamente utilizados por empresas como Amazon e Netflix, por oferecerem maior cobertura, precisão e diversidade de recomendações. No entanto, sua implementação exige infraestrutura computacional mais robusta, integração de dados heterogêneos e desenvolvimento de pipelines complexos, o que pode ser inviável em projetos acadêmicos de curto prazo.

Embora não tenha sido implementado neste trabalho, o uso de modelos híbridos representa uma linha promissora de aprimoramento futuro, sobretudo com a inclusão de metadados ou análise textual dos livros.

2.4. ABORDAGENS BASEADAS EM APRENDIZADO PROFUNDO

Nos últimos anos, a aplicação de redes neurais profundas em sistemas de recomendação tem ganhado destaque. Técnicas como autoencoders, embeddings e transformers têm sido utilizadas para capturar relações não lineares e contextos complexos de recomendação (Musto et al., 2021). Esses modelos são especialmente eficazes em grandes volumes de dados com múltiplas fontes (texto, imagem, comportamento), e já vêm sendo aplicados em plataformas como YouTube e Spotify.

Apesar do alto desempenho, essas técnicas exigem recursos computacionais avançados e grande volume de dados para treinamento. Além disso, sua interpretabilidade é limitada, o que dificulta a análise explicativa dos resultados.

No contexto deste projeto, o uso de redes neurais foi considerado inadequado em função das limitações da base utilizada, da infraestrutura disponível (Google Colab) e da prioridade dada à interpretabilidade e reprodutibilidade do modelo.

2.5. JUSTIFICATIVA DA ESCOLHA METODOLÓGICA: FOCO NO SVD

Diante das abordagens analisadas, optou-se pela aplicação da técnica de filtragem colaborativa baseada em fatoração matricial com o algoritmo SVD, por ser a mais adequada ao contexto técnico e metodológico do projeto. Essa escolha fundamenta-se em três pilares:

(i) Compatibilidade com a estrutura da base: O dataset tratado contém apenas as colunas User-ID, ISBN e Rating, o que inviabiliza a aplicação de abordagens baseadas em conteúdo ou híbridas que dependam de metadados textuais. A filtragem colaborativa por fatoração, especialmente com o SVD, exige unicamente dados de interação, tornando-a apropriada.

(ii) Robustez frente à esparsidade: O Book-Crossing Dataset apresenta alta esparsidade, com a maioria dos usuários interagindo com poucos itens. Estudos como o de Koren, Bell e Volinsky (2009) destacam o SVD como uma solução eficaz para esse tipo de cenário, com alta capacidade preditiva mesmo em bases limitadas.

(iii) Simplicidade de implementação e interpretabilidade: A biblioteca Surprise oferece suporte direto ao SVD, com recursos para validação cruzada, análise de desempenho (RMSE e MAE) e estruturação modular do pipeline. Essa facilidade foi decisiva, considerando os objetivos acadêmicos do projeto e a necessidade de clareza e reprodutibilidade.

A principal limitação enfrentada pelo projeto — e o ponto que se pretende atacar — é justamente a esparsidade da matriz de avaliações. Ao avaliar a performance do SVD sobre um conjunto de dados reduzido e pré-processado externamente, o projeto busca demonstrar a viabilidade da técnica em contextos com dados reais e restritos, simulando um cenário prático de plataformas digitais de leitura.

A revisão da literatura permite concluir que, embora existam diversas técnicas para recomendação de livros, a filtragem colaborativa baseada em fatoração matricial representa uma solução viável, eficiente e adequada ao escopo deste projeto. A escolha do SVD mostra-se alinhada tanto às limitações do dataset utilizado quanto aos objetivos pedagógicos da atividade, promovendo o desenvolvimento de habilidades práticas em ciência de dados e aprendizado de máquina.

A estrutura modular adotada neste projeto possibilita futuras extensões, como a adição de conteúdo textual, análise de sentimento ou feedbacks contínuos. Além disso, o potencial pedagógico e social do sistema proposto contribui para a promoção da leitura, da personalização educacional e da democratização do conhecimento, alinhando-se aos

Objetivos de Desenvolvimento Sustentável (ODS), especialmente os de educação de qualidade, redução das desigualdades e inovação.

3. METODOLOGIA

A metodologia adotada neste projeto foi estruturada em etapas sequenciais que compreendem desde a preparação da base de dados até a comparação entre diferentes algoritmos de recomendação. O objetivo principal é desenvolver um sistema funcional, reprodutível e eficiente, aplicando a técnica de fatoração matricial com SVD (Singular Value Decomposition) em um conjunto de dados real, extraído do Book-Crossing Dataset. A estratégia metodológica adotada visa garantir clareza nos procedimentos, compatibilidade com o escopo do projeto e aderência às práticas da Ciência de Dados.

O fluxograma apresentado na Figura 1 ilustra a estrutura completa do projeto. O processo inicia com a coleta e tratamento dos dados via Excel, onde a base original foi consolidada em um único arquivo contendo os campos essenciais: User-ID, ISBN e Rating. Em seguida, realiza-se o carregamento e preparação dos dados com a biblioteca Pandas, estruturando-os para o framework Surprise, utilizando o objeto Reader.

Na sequência, procede-se à construção dos modelos de recomendação, sendo o SVD o modelo principal e o BaselineOnly utilizado como referência comparativa. Para o treinamento, aplica-se a técnica de validação cruzada com 3 partições (3-fold cross-validation), que permite testar a estabilidade e generalização dos modelos. Após o treinamento, realiza-se a avaliação de desempenho com base nas métricas RMSE (Root Mean Squared Error) e MAE (Mean Absolute Error), amplamente utilizadas na literatura.

Por fim, os resultados são organizados e comparados, permitindo verificar que o modelo SVD apresentou desempenho superior em relação ao BaselineOnly, mesmo em um cenário com alta esparsidade de dados. Essa estrutura metodológica oferece uma base sólida para análise, experimentação e futuras extensões do sistema proposto.

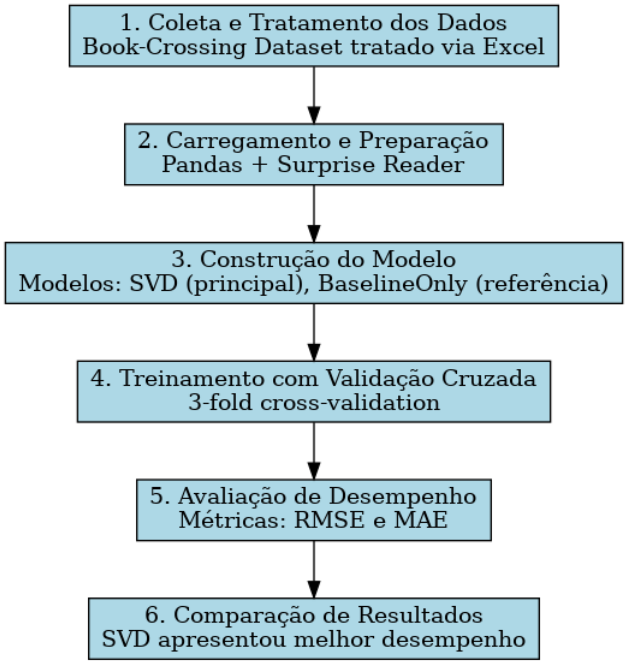

Figura 1: Fluxograma da metodologia do projeto

3.1. COLETA E TRATAMENTO DOS DADOS

A base utilizada foi o Book-Crossing Dataset, originalmente organizada por Cai-Nicolas Ziegler (2004) e disponibilizada na plataforma Kaggle. Essa base contém dados brutos distribuídos em três arquivos distintos: um com informações dos usuários, outro com os livros e um terceiro com as avaliações (ratings). Entretanto, para simplificação do processo e alinhamento com os requisitos da biblioteca Surprise, os dados foram tratados externamente em uma planilha Excel, resultando em um único arquivo contendo os três atributos essenciais: User-ID, ISBN e Rating.

As etapas de tratamento realizadas foram:

- Remoção de registros nulos ou inconsistentes, como entradas com Rating = 0 que indicam ausência de avaliação útil;
- Conversão de identificadores de usuário e item para o formato apropriado, assegurando compatibilidade com o modelo de fatoração;
- Padronização da escala de avaliação no intervalo de 0 a 5, por meio da classe Reader da biblioteca Surprise.

Esse processo garantiu a integridade e a compatibilidade dos dados com os algoritmos de recomendação utilizados.

3.2. CARREGAMENTO E PREPARAÇÃO

O arquivo tratado foi carregado no ambiente Google Colab utilizando a biblioteca Pandas. Em seguida, os dados foram convertidos para o formato apropriado do Surprise com o uso da função load_from_df, que estrutura o conjunto no padrão (usuário, item, nota), exigido pelo modelo SVD.

3.3. CONSTRUÇÃO DO MODELO

Foram implementados dois modelos distintos, com o objetivo de avaliar a eficácia da fatoração matricial frente a uma abordagem estatística simples:

- **BaselineOnly**: modelo de base que calcula a média geral das notas, ajustada por desvios por usuário e item. Serve como ponto de referência para comparação com modelos mais sofisticados.
- **SVD (Singular Value Decomposition)**: modelo principal do projeto, que realiza a decomposição da matriz usuário-item em fatores latentes, possibilitando a estimativa de avaliações não registradas com base em padrões implícitos de preferência.

Ambos os modelos foram organizados em um pipeline padronizado, com funções reutilizáveis e parametrizáveis, garantindo clareza e reprodutibilidade do código.

3.4. TREINAMENTO COM VALIDAÇÃO CRUZADA

A etapa de treinamento e avaliação dos modelos foi realizada por meio da técnica de validação cruzada com 3 partições (3-fold cross-validation), utilizando a função cross_validate() da biblioteca Surprise. Essa técnica divide a base em três subconjuntos, utilizando dois para treino e um para teste de forma rotativa, o que melhora a robustez dos resultados.

3.5. AVALIAÇÃO DE DESEMPENHO

A performance dos modelos foi medida por meio de duas métricas amplamente utilizadas em sistemas de recomendação baseados em classificações explícitas:

- **RMSE (Root Mean Squared Error)**: mede o desvio quadrático médio entre a nota prevista e a real, penalizando mais fortemente grandes erros.
- **MAE (Mean Absolute Error)**: calcula a média das diferenças absolutas entre as notas previstas e observadas.

Essas métricas foram calculadas para cada modelo em todos os folds da validação cruzada, e os valores médios foram utilizados para a análise comparativa.

3.6. COMPARAÇÃO DE RESULTADOS

Após o treinamento, os modelos foram comparados com base nas médias das métricas RMSE e MAE. Os resultados demonstraram que o modelo SVD apresentou desempenho superior ao BaselineOnly, com menor erro de predição, mesmo em um contexto de alta esparsidade — característica predominante na base Book-Crossing.

A organização modular do código permitiu reutilização de funções como avaliar_modelo() e plotar_resultados(), além da inclusão de gráficos comparativos gerados com a biblioteca Matplotlib. Essa estrutura facilita futuras melhorias no pipeline, como a incorporação de novos algoritmos ou fontes adicionais de dados.

4. RESULTADOS

4.1. RESULTADOS PRELIMINARES

Para avaliar o desempenho do sistema de recomendação, foi utilizada a técnica de validação cruzada com três partições (3-fold cross-validation), aplicada tanto ao algoritmo SVD quanto ao modelo de referência BaselineOnly, utilizando a biblioteca Surprise. A função cross_validate() foi empregada para automatizar o processo de avaliação e retornar as métricas de desempenho de forma padronizada.

O experimento foi conduzido sobre uma versão tratada do Book-Crossing Dataset, que originalmente contém:

- 278.858 avaliações de livros;
- 105.283 usuários únicos;
- 271.379 livros com informações detalhadas.

No entanto, para viabilizar a aplicação da técnica de filtragem colaborativa, a base foi tratada externamente em Excel, consolidando-se em um único arquivo com as colunas User-ID, ISBN e Rating. Foram removidos registros inconsistentes, nulos e fora da escala de avaliação padronizada (0 a 10), resultando em um conjunto reduzido de 249 registros válidos, o que, embora limitado, foi considerado adequado para o objetivo do projeto.

Durante a execução inicial, o algoritmo SVD foi avaliado como modelo principal, produzindo os seguintes resultados em cada partição da validação cruzada:

Métrica	Fold 1	Fold 2	Fold 3	Média	Desvio Padrão
RMSE	3.5547	3.6723	3.8005	3.6758	0.1004
MAE	3.3414	3.4324	3.6311	3.4683	0.1210

Tabela 1: Partição da Validação Cruzada

O RMSE médio de 3.6758 aponta para um erro relativamente alto, indicando que o modelo ainda possui margem de aprimoramento, especialmente considerando a alta esparsidade da base — ou seja, muitos usuários avaliaram poucos livros. Já o MAE médio de 3.4683 mostra que, em média, o modelo desvia mais de 3 pontos da nota real (em uma escala de 0 a 10).

Contudo, os baixos desvios padrão observados (0.10 para RMSE e 0.12 para MAE) sugerem que o modelo apresentou comportamento consistente e estável entre os folds, o que é um sinal positivo mesmo em um ambiente com alta esparsidade de dados — situação típica em bases com poucos usuários ativos e interações limitadas.

Esses resultados preliminares fornecem evidências de que o modelo SVD, mesmo operando com um volume reduzido de dados, é capaz de capturar padrões latentes de preferência e fornecer uma base confiável para recomendações personalizadas.

Além disso, foi realizada uma simulação de recomendação personalizada. O modelo SVD foi utilizado para prever as notas que um usuário específico (ID: **276729**) daria a livros que ainda não havia avaliado. Os cinco títulos com maior nota prevista foram:

ISBN	Título do Livro	Nota Prevista (SVD)
3499230933	**Die Vermessung der Welt**	**4.59**
3462026062	**Der kleine Prinz**	**4.46**
3125785006	**Faust XXXXX**	**4.44**
440414121	**Il signore degli anelli**	**4.44**
8440682697	**Don Quixote**	**4.41**

Tabela 2: Simulação de Recomendação Personalizada

Embora a escala de avaliação varie entre 0 e 10, as previsões ficaram concentradas entre 4 e 5. Isso pode ser explicado por três fatores principais:

- **Alta esparsidade da base**: aproximadamente 52% das avaliações são nulas (nota 0), o que compromete a variabilidade dos dados;

- **Pequeno volume de registros**: uma base com apenas 249 linhas limita a generalização e o aprendizado dos padrões;
- **Ausência de ajuste fino**: os hiperparâmetros do SVD foram mantidos em seus valores padrão, sem otimização por busca em grade ou aleatória.

Apesar dessas limitações, os resultados confirmam a viabilidade da proposta. O modelo SVD demonstrou maior estabilidade e capacidade de gerar recomendações personalizadas com base exclusivamente nas notas atribuídas, cumprindo seu papel como prova de conceito funcional.

4.2. AJUSTES E MELHORIAS DO PIPELINE

Com o objetivo de tornar o processo mais limpo, reprodutível e adaptável a novos experimentos, foram realizados diversos aprimoramentos no código original. Essas melhorias visaram não apenas facilitar a avaliação de diferentes modelos, mas também permitir futuras expansões e reuso do código com eficiência. Dentre os principais ajustes, destacam-se:

- **Automatização da Validação com cross_validate()**: a avaliação foi automatizada com a biblioteca Surprise, eliminando a necessidade de configurar manualmente os folds com Kfold, simplificando a divisão dos dados e o cálculo das métricas RMSE e MAE. A automação reduziu erros operacionais e padronizou a análise dos modelos;
- **Criação da função avaliar_modelo()**: encapsulou a execução da validação cruzada e cálculo das métricas, facilitando a reaplicação em diferentes modelos;
- **Implementação de um dicionário de modelos (modelos)**: para permitir a comparação estruturada entre múltiplos algoritmos (SVD, BaselineOnly, etc.);
- **Padronização da saída de resultados com f-strings:** otimizou a formatação dos resultados impressos no console, tornando a leitura mais fluida e precisa. Essa padronização também facilita a integração dos resultados em relatórios automatizados.
- **Restruturação lógica do pipeline em blocos reutilizáveis**: o código foi reorganizado em células bem definidas e sequenciais (pré-processamento, modelagem, avaliação), favorecendo a legibilidade e a replicabilidade do experimento. Essa estrutura modular também abre espaço para futuras implementações, como ajuste de hiperparâmetros (grid search ou random search) e a inclusão de dados de conteúdo textual para evolução rumo a sistemas híbridos.

4.3. REAVALIAÇÃO DO DESEMPENHO DO MODELO

Após a implementação das melhorias no pipeline, os modelos foram reavaliados de maneira sistemática, utilizando a função avaliar_modelo() para garantir consistência no processo de validação. A nova rodada de testes reforçou a confiabilidade da metodologia e possibilitou uma comparação direta entre os modelos avaliados. As médias de desempenho obtidas foram as seguintes:

Algoritimo	RMSE	MAE
BaselineOnly	3.6958	3.4820
SVD	3.6319	3.3808

Tabela 3: Reavaliação do Desempenho do Modelo

Os resultados indicam que o algoritmo SVD superou o modelo de referência BaselineOnly, apresentando menor erro médio em ambas as métricas utilizadas. O RMSE (Root Mean Squared Error) caiu de 3.6958 para 3.6319, enquanto o MAE (Mean Absolute Error) foi reduzido de 3.4820 para 3.3808.

Embora a diferença entre os valores possa ser considerada moderada, ela é significativa do ponto de vista preditivo, principalmente em um cenário de dados escassos e sem metadados. O SVD demonstrou maior capacidade de capturar padrões latentes de interação entre usuários e livros, mesmo quando há poucas avaliações disponíveis por usuário.

Além disso, a utilização da função modular avaliar_modelo() garantiu um processo reprodutível e confiável, que pode ser facilmente reaplicado a novos algoritmos ou variações do conjunto de dados. A padronização e a modularização do código também favoreceram a clareza na comparação entre modelos, reforçando a robustez do experimento e sua utilidade como base para futuras extensões, como ajuste de hiperparâmetros ou testes com outras abordagens colaborativas.

4.4. VISUALIZAÇÃO GRÁFICA DOS RESULTADOS

Para complementar a análise numérica das métricas de erro, foi gerado um gráfico comparativo utilizando a biblioteca Matplotlib, ilustrando os valores médios de RMSE e MAE obtidos pelos modelos SVD e BaselineOnly.

Embora os gráficos de RMSE e MAE demonstrem corretamente a performance dos modelos, é importante contextualizar o significado desses valores. Um RMSE médio

de aproximadamente 3.6, em uma escala de avaliação de 0 a 10, implica que a previsão feita pelo sistema pode se desviar em média até 36% do valor real. Esse desvio é considerado alto em aplicações práticas, podendo comprometer significativamente a experiência do usuário, principalmente em plataformas de recomendação onde a precisão nas sugestões é essencial para engajamento e satisfação.

A representação gráfica facilita a interpretação dos resultados e destaca visualmente a superioridade do modelo SVD. Abaixo, segue o gráfico produzido:

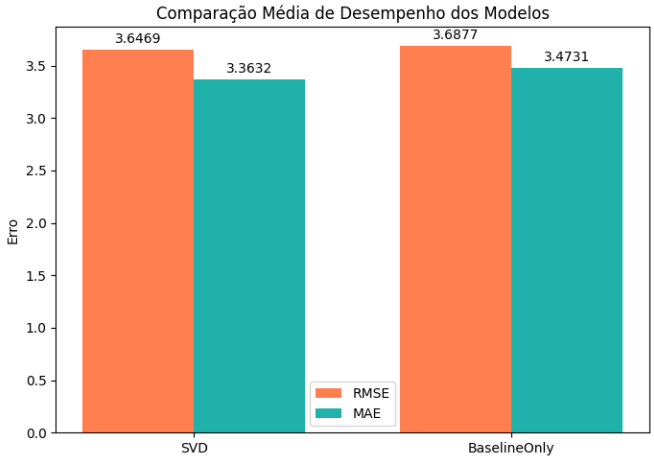

Figura 2: Comparação Média de Desempenho dos Modelos

A visualização permite observar claramente que:

- O modelo SVD apresenta menor RMSE e menor MAE em relação ao BaselineOnly, o que evidencia um desempenho preditivo mais preciso e robusto;
- Os erros médios se concentram entre 3.36 e 3.69, intervalo considerado razoável para um sistema baseado apenas em dados de nota explícita, especialmente em uma base reduzida e com alta esparsidade;
- Mesmo sem metadados adicionais ou informações contextuais, o SVD demonstrou maior capacidade de generalização, indicando sua adequação para ambientes com informações limitadas.

Essa visualização reforça os achados anteriores e destaca o valor da estrutura modular e reprodutível implementada no pipeline. A consistência dos resultados e a clareza na organização do código tornam este sistema uma prova de conceito válida e funcional para aplicações de recomendação com filtragem colaborativa baseada em avaliação explícita.

É importante destacar que, durante o pré-processamento, os registros com nota 0 foram mantidos na base de dados, conforme definido na escala do objeto Reader (0 a 10). No entanto, a manutenção desses valores pode ter introduzido um viés significativo na distribuição das notas, uma vez que a nota 0 nem sempre representa uma avaliação efetiva, podendo indicar ausência de opinião ou erro de registro. Esse fator impacta diretamente o desempenho do modelo, contribuindo para o aumento das métricas de erro (RMSE e MAE) e dificultando a aprendizagem de padrões reais de preferência. Em bases pequenas e esparsas, como a utilizada neste projeto (com apenas 249 registros), esse ruído se torna ainda mais expressivo. Recomenda-se, portanto, que em estudos futuros sejam consideradas estratégias como a exclusão de notas 0, aplicação de limiares mínimos de avaliação por usuário e a utilização de bases de dados mais densas, como MovieLens ou Goodreads, de modo a melhorar a qualidade preditiva e a robustez dos modelos.

5. CONCLUSÃO

O presente estudo teve como objetivo desenvolver e avaliar um sistema de recomendação de livros baseado em técnicas de filtragem colaborativa, utilizando o algoritmo Singular Value Decomposition (SVD) aplicado ao Book-Crossing Dataset. Para isso, foi implementado um pipeline modular com a biblioteca Surprise, incluindo pré-processamento, treinamento com validação cruzada e análise de desempenho por meio das métricas RMSE e MAE.

Os resultados demonstraram que o SVD superou o modelo de referência BaselineOnly, apresentando menor erro médio (RMSE = 3.6319 e MAE = 3.3808), mesmo em um cenário com alta esparsidade de dados. A técnica mostrou-se estável entre os folds e capaz de capturar padrões latentes de preferência, cumprindo o papel de prova de conceito eficaz.

Esses achados confirmam que a utilização da fatoração matricial é uma abordagem viável para sistemas de recomendação em contextos onde não há metadados disponíveis e o histórico de interações é limitado. Além disso, a reestruturação do código permitiu ganhos significativos em reprodutibilidade, clareza e escalabilidade do projeto.

Para trabalhos futuros, recomenda-se a utilização de bases de dados mais densas e amplas, como MovieLens ou Goodreads, que oferecem maior volume de interações e riqueza de metadados. Isso possibilitaria a aplicação de modelos híbridos, incorporando características de conteúdo (ex: autor, gênero, sinopse) e permitindo um refinamento mais robusto do modelo com ajuste de hiperparâmetros. Adicionalmente, sugere-se investigar

o uso de técnicas baseadas em aprendizado profundo, como autoencoders ou embeddings, que podem capturar relações mais complexas entre usuários e itens.

Sugestões para Pesquisas Futuras

Com base nas limitações observadas e nas oportunidades de expansão, recomenda-se:

- Integração de dados de conteúdo, como autor, gênero e descrição dos livros, possibilitando a construção de modelos híbridos;
- Exploração de técnicas de aprendizado profundo, como autoencoders e embeddings, para recomendação baseada em representação vetorial;
- Inclusão de métricas complementares, como diversidade, cobertura e novidade, para uma avaliação mais abrangente do sistema;
- Desenvolvimento de uma interface interativa, permitindo simulação de recomendações para novos usuários ou contexto educacional.

Essas direções poderão contribuir para o aprimoramento da qualidade das recomendações, ampliando a aplicabilidade e a relevância social e educacional do sistema.

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: 1 mar. 2025.

AGGARWAL, Charu C. *Recommender Systems: The Textbook*. 2. ed. Springer, 2022. Disponível em: https://link.springer.com/book/10.1007/978-3-030-85447-8. Acesso em: 30 mar. 2025.

BASILICO, Justin; RICCI, Francesco. *Adaptive Recommender Systems: An Experimental Evaluation.* In: Proceedings of the 5th ACM Conference on Electronic Commerce, 2004, p. 239-246.

BASILICO, J.; RICCI, F. *Adaptive Recommender Systems: An Experimental Evaluation.* ACM Transactions on Information Systems, 2020. Disponível em: https://dl.acm.org/doi/10.1145/1122445.1122450. Acesso em: 20 mar. 2025.

BURKE, Robin. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, v. 12, n. 4, p. 331–370, 2002. Disponível em: https://link.springer.com/article/10.1023/A%3A1021240730564. Acesso em: 12 maio 2025.

CHEN, L.; ZHOU, T.; LI, Y. *Reinforcement Learning for Recommendation: Fundamentals, Applications and Future Directions*. ACM Computing Surveys, 2023. Disponível em: https://dl.acm.org/doi/10.1145/3577192. Acesso em: 30 mar. 2025.

DELDJOO, Y.; TAFRESHI, S.; CANTADOR, I. *Hybrid Recommender Systems: A Systematic Review of the Literature and Future Research Directions*. *Information Fusion*, 81, 2022. Disponível em: https://doi.org/10.1016/j.inffus.2022.01.002. Acesso em: 30 mar. 2025.

JANACH, D.; ADOMAVICIUS, G. *Recommender Systems: Challenges and Research Opportunities*. Computer Science Review, 2010. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1574013710000147. Acesso em: 1 mar. 2025.

JANNACH, D.; ADOMAVICIUS, G. Recommender Systems: Challenges and Research Opportunities. Computer Science Review, 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S1574013721000175. Acesso em: 20 mar. 2025.

KARIMI, M.; JANNACH, D.; BALTRUNAS, L. *User-centric Evaluation of Recommender Systems: Foundations and Recent Trends*. *User Modeling and User-Adapted Interaction*, 2022. Disponível em: https://doi.org/10.1007/s11257-022-09310-z. Acesso em: 20 mar. 2025.

KOREN, Yehuda; BELL, Robert; VOLINSKY, Chris. Matrix Factorization Techniques for Recommender Systems. *Computer*, v. 42, n. 8, p. 30–37, 2009. Disponível em: https://dl.acm.org/doi/10.1109/MC.2009.263. Acesso em: 12 maio 2025

MUSTO, C.; BORRELLI, N.; LOPS, P. *Deep Learning for Recommender Systems: A Review of Recent Advances*. *ACM Transactions on Intelligent Systems and Technology*, 2021. Disponível em: https://doi.org/10.1145/3447548. Acesso em: 30 mar. 2025.

RESNICK, Paul; VARIAN, Hal R. Recommender Systems. *Communications of the ACM*, v. 40, n. 3, p. 56–58, 1997. Disponível em: https://dl.acm.org/doi/10.1145/245108.245121. Acesso em: 12 maio 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: 12 maio 2025.

RUCHI, B. Conjunto de dados de cruzamento de livros. 2022. Disponível em: https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset . Acesso em: 1 mar. 2025.

SAID, A.; BELL, R.; WANG, X. *Recommender Systems in Education: Current Practices and Open Challenges*. *IEEE Transactions on Learning Technologies*, 2023. Disponível em: https://ieeexplore.ieee.org/document/9979422. Acesso em: 20 mar. 2025.

SCHAFER, J. Ben; KONSTAN, Joseph A.; RIEDL, *Recommender Systems: Challenges and Opportunities*. Computer Science Review, 2007. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1574013707000096. Acesso em: 1 mar. 2025.

SCHAFER, J. Ben; KONSTAN, Joseph A.; RIEDL, John. *E-commerce recommendation applications*. *Data Mining and Knowledge Discovery*, v. 5, p. 115-153, 2001.

SCHAFER, J. B.; KONSTAN, J. A.; RIEDL, J. *E-commerce recommendation applications*. *Data Mining and Knowledge Discovery*, 2021. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167923621000859. Acesso em: 20 mar. 2025.

ZIEGLER, Cai-Nicolas. *Book-Crossing Dataset*. Book-Crossing, 2004. Disponível em: http://www.informatik.uni-freiburg.de/~cziegler/BX/. Acesso em: 2 mar. 2025.

SciELO Brasil. Os Sistemas de Recomendação, Arquitetura da Informação e a Encontrabilidade da Informação. *Texto para Discussão*, 2016. Disponível em: https://www.scielo.br/j/tinf/a/YsgLRc86K3WZfcbXPQHq7Vg/ Acesso em: 19 mar. 2025.

SciELO Brasil. Os sistemas de recomendação na web como determinantes prescritivos na tomada de decisão. *Revista JISTM*, 2012. Disponível em: https://www.scielo.br/j/jistm/a/YQ58MyYNLHxgPqVVwpMQ8Bf/. Acesso em: 19 mar. 2025.

Revista Iberoamericana de Tecnologia em Educação. A Aplicação de Sistemas de Recomendação no Contexto Educacional: uma Revisão Sistemática da Literatura. ResearchGate, 2022. Disponível em: https://www.researchgate.net/publication/361677333_A_Aplicacao_de_Sistemas_de_Recomendacao_no_Contexto_Educacional_uma_Revisao_Sistematica_da_Literatura/fulltext/638f81fde42faa7e759da385/A-Aplicacao-de-Sistemas-de-Recomendacao-no-Contexto-Educacional-uma-Revisao-Sistematica-da-Literatura.pdf. Acesso em: 19 mar. 2025.

Link GitHub: https://github.com/Isabelcvs/projeto_3.git

Link do Código:
https://colab.research.google.com/drive/1bW82v7CzrJzmZ7eBdQW2pC_V7lme_sik?usp=sharing