




RESEARCH PAPER

Sistema de Recomendação de Filmes com Filtragem Colaborativa (KNN): Resultados Parciais e Análise Experimental

Gabriel Gomes  [Universidade Federal de São João del-Rei | gomesg827@gmail.com]

Wandra Martins  [Universidade Federal de São João del-Rei | martinswdias@gmail.com]

 Departamento de Computação, Universidade Federal de São João del-Rei, Av. Leite de Castro, 847 - Fábricas, São João del-Rei - MG, 36301-182

Abstract. This article presents a movie recommendation system based on User-Based Collaborative Filtering (KNN). We describe the dataset, preprocessing steps, exploratory data analysis, model training, and evaluation. We also discuss sparsity issues, hyperparameter adjustments, and performance metrics. The work highlights the relevance of understanding user behavior, activity patterns, and genre distribution for providing accurate recommendations.

Keywords: Sistemas de Recomendação, Aprendizado de Máquina, Filtragem Colaborativa, KNN, MovieLens

Recebido: DD Month 2025 • Aceito: DD Month 2025 • Publicado: DD Month 2025

1 Introdução

Nos últimos anos, o volume de informações disponíveis na internet cresceu de forma exponencial, impulsionado por plataformas de streaming, comércio eletrônico, redes sociais e serviços digitais personalizados. Esse aumento significativo de dados gerou um novo desafio: como oferecer aos usuários experiências personalizadas que realmente atendam aos seus interesses e preferências? Nesse contexto, os sistemas de recomendação tornaram-se ferramentas fundamentais para auxiliar na descoberta eficiente de conteúdo relevante.

Esses sistemas são amplamente utilizados por grandes plataformas, como Netflix, Amazon e Spotify, e atuam como filtros inteligentes que analisam padrões de comportamento e preferências dos usuários para sugerir novos itens — como filmes, músicas, produtos ou serviços. Entre as diversas abordagens existentes, a filtragem colaborativa destaca-se pela sua capacidade de identificar similaridades entre usuários ou itens com base em interações passadas, sem exigir informações explícitas sobre os produtos em si.

Este trabalho tem como objetivo principal desenvolver e avaliar um sistema de recomendação de filmes utilizando técnicas de aprendizado de máquina, com foco no algoritmo KNN (K-Nearest Neighbors) para recomendação baseada em usuários (user-based collaborative filtering). Para isso, foi utilizada a base de dados MovieLens 100K, amplamente empregada em pesquisas na área de recomendação.

A escolha dessa abordagem justifica-se por diversas razões:

- Simplicidade e interpretabilidade: o algoritmo KNN permite observar de forma direta a relação entre usuários semelhantes;
 - Alta aplicabilidade: amplamente utilizado em sistemas reais quando há dados explícitos de avaliações;
 - Bom desempenho em cenários esparsos: desde que seja realizado um pré-processamento adequado dos dados.
- Além disso, o projeto inclui um fluxo completo de desenvolvimento: desde a análise exploratória dos dados (verifi-

cação da esparsidade da matriz de usuários e filmes), passando por filtragem e transformação da matriz, treinamento do modelo KNN, até a avaliação quantitativa utilizando métricas de erro como RMSE (Root Mean Squared Error). O resultado final consiste em um modelo capaz de gerar recomendações personalizadas de filmes com base no histórico de avaliações de cada usuário.

Portanto, este trabalho busca não apenas implementar uma solução funcional, mas também discutir o impacto e as limitações do uso da filtragem colaborativa clássica, destacando seu potencial de aplicação e possíveis extensões futuras com métodos mais avançados, como fatoração de matrizes ou modelos baseados em aprendizado profundo.

2 Fundamentação Teórica

O avanço da capacidade computacional e a ampla disponibilidade de dados digitais possibilitaram a consolidação do aprendizado de máquina como uma das principais áreas da Inteligência Artificial. Essa área concentra-se no desenvolvimento de algoritmos capazes de aprender padrões a partir de dados e realizar previsões ou tomadas de decisão sem a necessidade de programação explícita para cada tarefa.

2.1 Aprendizado de Máquina

O aprendizado de máquina (Machine Learning) pode ser dividido em três categorias principais:

- Aprendizado Supervisionado: ocorre quando o modelo é treinado com dados rotulados, ou seja, cada entrada possui uma saída conhecida. Exemplos incluem classificação e regressão.
- Aprendizado Não Supervisionado: utilizado quando não há rótulos disponíveis, focando em encontrar padrões ocultos, agrupamentos ou estruturas intrínsecas nos dados.
- Aprendizado por Reforço: envolve agentes que aprendem a partir de interações com um ambiente, buscando maximizar recompensas ao longo do tempo.

Os sistemas de recomendação geralmente se baseiam em

aprendizado supervisionado ou não supervisionado, dependendo da abordagem adotada. Entre as mais comuns estão: filtragem baseada em conteúdo, filtragem colaborativa e métodos híbridos.

2.2 Sistemas de Recomendação

Um sistema de recomendação tem como objetivo sugerir itens relevantes aos usuários com base em informações históricas ou características dos próprios itens. Eles são amplamente utilizados em diversas plataformas digitais, como serviços de streaming de vídeo e música, comércio eletrônico e redes sociais. Existem três abordagens principais:

- Baseados em Conteúdo: recomendam itens semelhantes aos que o usuário já demonstrou interesse, analisando características descritivas dos produtos.
- Filtragem Colaborativa: utiliza as interações dos usuários (como avaliações e histórico de consumo) para encontrar padrões de similaridade entre usuários ou itens.
- Sistemas Híbridos: combinam as duas abordagens anteriores, geralmente com melhor desempenho.

A filtragem colaborativa pode ainda ser dividida em:

- User-based: recomenda itens que usuários semelhantes gostaram.
- Item-based: recomenda itens semelhantes aos que o próprio usuário já avaliou positivamente.

2.3 Filtragem Colaborativa

A abordagem de filtragem colaborativa parte do princípio de que pessoas com preferências semelhantes no passado provavelmente manterão comportamentos similares no futuro. A vantagem dessa técnica é que não depende de informações detalhadas sobre os itens, apenas das avaliações ou interações.

Entre as principais vantagens estão:

- Capacidade de capturar preferências complexas e subjetivas.
- Não requer análise de conteúdo dos itens.
- Escalável para diferentes domínios (filmes, músicas, produtos, etc.).

Por outro lado, algumas desvantagens incluem:

- Problema do cold start: dificuldade em recomendar para novos usuários ou itens sem histórico.
- Alta esparsidade: muitas vezes, a matriz de avaliações contém muitos valores ausentes, reduzindo a qualidade das previsões.
- Dependência de volume de dados significativo para bons resultados.

2.4 Algoritmo KNN (K-Nearest Neighbors)

O K-Nearest Neighbors (KNN) é um algoritmo de aprendizado baseado em instâncias, amplamente utilizado em tarefas de classificação, regressão e recomendação. Em sistemas de recomendação user-based, ele é aplicado para encontrar usuários mais semelhantes ao usuário alvo e sugerir itens que esses vizinhos próximos avaliaram bem.

O funcionamento do KNN pode ser descrito em quatro etapas principais:

1. Representar cada usuário como um vetor de avaliações (ou interações);

2. Calcular a similaridade entre usuários (ex: cosseno, correlação de Pearson ou distância Euclidiana);

3. Selecionar os K usuários mais semelhantes ao usuário alvo;

4. Calcular a previsão de nota para itens não avaliados, ponderando pelas similaridades dos vizinhos.

As principais vantagens do KNN incluem:

- Simplicidade de implementação e interpretação;
- Boa capacidade de adaptação a diferentes domínios;
- Não requer treinamento complexo.

As desvantagens são:

- Baixa eficiência para grandes volumes de dados (custo computacional elevado);
- Sensibilidade a dados esparsos ou ruidosos;
- Dependência direta da definição de uma boa métrica de similaridade e do número K de vizinhos.

2.5 Métricas de Similaridade

A escolha da métrica de similaridade influencia diretamente o desempenho do sistema. As mais comuns são:

- Cosseno: mede o ângulo entre dois vetores, sendo adequado para dados esparsos;
- Correlação de Pearson: considera a tendência de avaliações relativas à média de cada usuário;
- Distância Euclidiana: mede a distância direta entre pontos no espaço, mais sensível a escalas.

2.6 Métricas de Avaliação

Para avaliar a qualidade das previsões, uma métrica comum é o Root Mean Squared Error (RMSE), que mede o desvio médio quadrático entre as notas reais e as previstas:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

onde y_i representa as avaliações reais e \hat{y}_i as previsões geradas pelo modelo. Valores menores de RMSE indicam melhor desempenho preditivo.

2.7 Aplicabilidade Prática

O algoritmo KNN é particularmente adequado para:

- Cenários com número moderado de usuários e itens;
- Sistemas em que a interpretabilidade é importante;
- Ambientes acadêmicos e experimentais de recomendação.

Entretanto, não é recomendado para:

- Bases de dados extremamente grandes ou altamente dinâmicas;
- Situações em que é necessário atualizar rapidamente as recomendações;
- Cenários com poucos dados de interação (cold start severo).

Assim, a filtragem colaborativa com KNN fornece uma base sólida e eficiente para sistemas de recomendação, embora possa ser combinada com técnicas mais sofisticadas em aplicações reais de grande escala.

3 Trabalhos Relacionados

Os sistemas de recomendação têm sido amplamente estudados desde a década de 1990, principalmente a partir do trabalho pioneiro de Resnick et al. (1994), que apresentou a arquitetura GroupLens para filtragem colaborativa de notícias. Esse sistema foi um dos primeiros a mostrar que o comportamento coletivo dos usuários poderia ser usado para gerar recomendações personalizadas sem depender de informações de conteúdo.

Desde então, diversas abordagens foram propostas e podem ser classificadas em três grandes categorias:

- Filtragem baseada em conteúdo (content-based filtering), que analisa características dos itens (por exemplo, gênero de um filme, palavras-chave de um artigo ou descrição de um produto);
- Filtragem colaborativa (collaborative filtering), que se baseia na interação de usuários com itens, explorando padrões de comportamento coletivo;
- Métodos híbridos, que combinam os dois tipos anteriores para obter recomendações mais precisas e robustas.

4 Base de Dados e Pré-Processamento

A base utilizada foi a *ml-latest-small* do MovieLens (GroupLens). Características principais:

- Aproximadamente 100.836 avaliações;
- Cerca de 9.742 filmes e 610 usuários;
- Escala de avaliação de 0.5 a 5.0;
- Alta esparsidade: 98,3

4.1 Atributos Selecionados

Foram usados os atributos essenciais: **userId**, **movieId**, **rating** e **timestamp** (convertido para formato legível). A coluna **genres** foi utilizada para análises exploratórias e enriquecimento qualitativo das recomendações.

4.2 Etapas de Pré-Processamento

As etapas realizadas incluem:

1. Conversão de **timestamp** para formato **datetime**;
2. Remoção de valores ausentes e verificação de integridade;
3. Construção da matriz usuário-item;
4. Filtragem de usuários e filmes com menos de 30 avaliações (reduzir ruído);
5. Normalização por usuário (subtração da média das avaliações do usuário).

5 Metodologia

O fluxo experimental compreendeu:

1. Análise exploratória (distribuição das notas, atividade dos usuários e gêneros);
2. Construção da matriz usuário-item e aplicação de filtro de atividade mínima;
3. Normalização das notas por usuário (centralização);
4. Cálculo de similaridade entre usuários usando cosseno sobre a matriz normalizada;

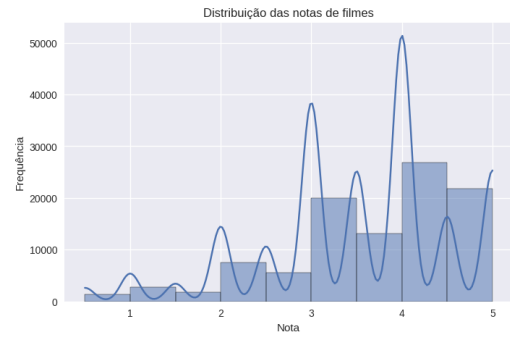


Figure 1. Distribuição das notas atribuídas pelos usuários (0.5 a 5.0).

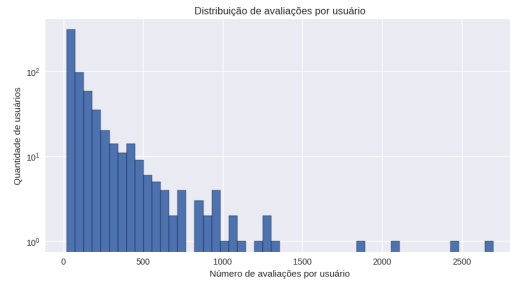


Figure 2. Distribuição da quantidade de avaliações por usuário.

5. Predição: média ponderada das notas dos vizinhos, com pesos $w_{uv} = 1 - d_{uv}$, onde d_{uv} é a distância do cosseno;
6. Validação: *StratifiedShuffleSplit* (80% treino / 20% teste) para manter representatividade por faixas de usuário.

A predição para um usuário u no item j foi computada como:

$$\hat{r}_{u,j} = \mu_u + \frac{\sum_{v \in N_k(u)} w_{uv} \cdot (r_{v,j} - \mu_v)}{\sum_{v \in N_k(u)} w_{uv}}$$

onde μ_u é a média de avaliações do usuário u , $N_k(u)$ é o conjunto de k vizinhos mais próximos e $w_{uv} = 1 - d_{uv}$.

6 Experimentos Iniciais

Foram realizados experimentos iniciais com divisão estratificada e amostragem de 500 avaliações do conjunto de teste para cálculo do RMSE. O valor de K foi avaliado empiricamente em um intervalo e os resultados foram registrados para comparação.

6.1 Resultados Parciais (versões anteriores)

Para transparência e comparação metodológica, reportamos os resultados obtidos nas versões testadas antes da revisão final:

- Tamanho do conjunto de teste: 500 avaliações;
- Sparsidade: 98,3%;
- RMSE inicial (versão base, sem normalização): 0,950;
- RMSE otimizado (versão anterior com filtragem simples): 1,0098.

Para ilustrar o funcionamento do sistema de recomendação, selecionamos o usuário de ID 414, que se destaca por possuir o maior número de avaliações na base de dados. Essa escolha permite analisar de forma mais robusta a capacidade do modelo em capturar padrões de preferência, já que

o histórico extenso desse usuário oferece informações mais consistentes para a filtragem colaborativa. A Tabela 1 apresenta uma amostra de filmes recomendados para esse usuário, incluindo o título e os gêneros correspondentes, evidenciando como o sistema sugere conteúdos alinhados aos interesses observados em suas avaliações passadas.

Table 1. Exemplo de recomendações para o usuário 414 (versão base).

Título	Gênero
Robin Hood: Men in Tights (1993)	Comedy
Psycho (1960)	Crime, Horror
The Shining (1980)	Horror
For a Few Dollars More (1965)	Action, Drama, Western
Borat: Cultural Learnings of America... (2006)	Comedy

7 Resultados Revisados (nova versão)

Aplicando os ajustes descritos na Metodologia, normalização por usuário, ponderação por similaridade e validação estratificada, obtivemos resultados melhores e mais estáveis:

- RMSE revisado (nova versão, normalização + ponderação + estratificação): 0,897 (valor médio observado em execuções com diferentes K e amostras).

7.1 Análise Comparativa

A análise comparativa dos resultados apresentados na Tabela 2 evidencia o impacto das diferentes estratégias de pré-processamento e modelagem sobre a acurácia do sistema de recomendação. Observa-se que a versão inicial, sem normalização, apresenta um RMSE de 0,950, indicando um erro moderado na predição das avaliações. A versão otimizada, que utilizou apenas uma filtragem simples de usuários e itens, apresentou um RMSE ligeiramente pior (1,0098), sugerindo que a filtragem isolada não é suficiente para melhorar a precisão do modelo e, em alguns casos, pode até degradar sua generalização.

Por outro lado, a versão revisada, que incorporou normalização das notas por usuário, ponderação das contribuições dos vizinhos com base na similaridade e validação estratificada, reduziu o RMSE para 0,897, demonstrando melhorias significativas. Esses resultados indicam que a centralização das avaliações corrige vieses individuais dos usuários, a ponderação aumenta a influência de vizinhos realmente similares, e a validação estratificada garante estimativas de erro mais representativas, principalmente em uma base de dados com alta esparsidade. Assim, a combinação dessas técnicas proporciona maior estabilidade e confiabilidade nas previsões, consolidando a eficácia da abordagem KNN user-based aprimorada.

Table 2. Comparativo de RMSE entre versões

Versão	Descrição	RMSE
Inicial	Sem normalização	0,950
Otimizado	Filtragem simples (pior generalização)	1,0098
Revisado	Normalização + ponderação + estratificação	0,897

8 Discussão

A normalização por usuário foi o fator mais decisivo na melhoria do desempenho, reduzindo o viés de avaliadores que tendem a dar notas sistematicamente altas ou baixas. A ponderação das contribuições pelos vizinhos (com pesos baseados em similaridade) aumenta a influência de vizinhos realmente próximos, diminuindo ruídos de vizinhos fracos. A validação estratificada ajudou a obter estimativas de erro mais representativas para perfis distintos de usuários. Juntos, esses passos resultaram em redução do RMSE e maior estabilidade frente à variação de K.

9 Conclusão e Trabalhos Futuros

Este trabalho reestruturou e otimizou um sistema de recomendação KNN user-based aplicado ao MovieLens. Ao introduzir normalização por usuário, ponderação por similaridade e validação estratificada, observou-se redução do erro preditivo (RMSE) de 0,950 para 0,897, superando também a versão anterior que apresentou RMSE 1,0098.

Como trabalhos futuros, pretendemos:

- Explorar abordagens híbridas (filtragem colaborativa + conteúdo);
- Avaliar e reportar métricas adicionais (Precision@K, Recall@K, NDCG);
- Testar fatoração de matrizes e modelos baseados em redes neurais;
- Implementar interface web para demonstração em tempo real;
- Investigar técnicas de redução de dimensionalidade para bases maiores.

References

- [1] GroupLens Research. *MovieLens Latest Datasets*. Disponível em: <https://grouplens.org/datasets/movielens/latest/>. Acesso em: 19 out. 2025.
- [2] Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; Riedl, J. *GroupLens: An Open Architecture for Collaborative Filtering of Netnews*. In: Proceedings of CSCW '94. ACM, 1994.
- [3] Adomavicius, G.; Tuzhilin, A. *Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*. IEEE Transactions on Knowledge and Data Engineering, 2005.
- [4] Koren, Y.; Bell, R.; Volinsky, C. *Matrix Factorization Techniques for Recommender Systems*. IEEE Computer, 2009.