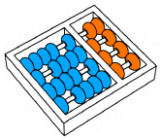

Avaliação de Justiça em Sistemas de Recomendação

G. S. Costa *M. F. Ferraz* *T. C. Loesch* *E. Martins*

Relatório Técnico - IC-PFG-25-51
Projeto Final de Graduação
2025 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Avaliação de Justiça em Sistemas de Recomendação

Gabriel da Silva Costa Miguel Figueira Ferraz Tomás Conti Loesch
Eliane Martins

Resumo

Sistemas de Recomendação tornaram-se onipresentes em plataformas digitais, moldando o consumo de informação e entretenimento. Tradicionalmente, a eficácia desses sistemas é medida por métricas de acurácia, como o erro quadrático médio. No entanto, a maximização da acurácia pode, inadvertidamente, amplificar vieses e perpetuar desigualdades presentes nos dados históricos. Este trabalho propõe uma avaliação conjunta de desempenho preditivo e justiça (*fairness*) em algoritmos de Filtragem Colaborativa. Utilizando a biblioteca *Surprise* para modelagem e a biblioteca *Fairlearn* para auditoria de viés, foram analisados modelos clássicos sobre quatro bases de dados de contextos distintos: MovieLens (100k e 1M), Amazon Music, Book Crossing e MyAnimeList. A análise investigou não apenas a justiça sob a ótica do usuário, considerando atributos como idade e gênero, mas também sob a ótica do item, avaliando categorias como gênero da obra e país de origem. Os resultados indicam que, embora os modelos mantenham níveis aceitáveis de paridade demográfica entre grupos de usuários, eles apresentam disparidades significativas na recomendação de itens, penalizando sistematicamente gêneros de nicho e produções fora do *mainstream*.

1 Introdução

Sistemas de Recomendação (SRs) desempenham um papel fundamental em diversas aplicações modernas, atuando como filtros essenciais em plataformas de *streaming*, comércio eletrônico e redes sociais. O objetivo primordial desses sistemas é mitigar a sobrecarga de informação, auxiliando usuários na descoberta de novos conteúdos ao predizer preferências com base em interações anteriores e padrões de comportamento.

Tradicionalmente, o estado da arte em SRs foca na otimização de métricas de acurácia, buscando minimizar a distância entre a predição do modelo e a avaliação real do usuário. Entretanto, métricas de erro, como RMSE ou MAE, são insuficientes para capturar o impacto social das recomendações. A acurácia, por si só, não garante que o sistema seja isento de preconceitos; pelo contrário, algoritmos de aprendizado de máquina podem reproduzir e até amplificar vieses históricos contidos nos dados de treinamento, afetando de maneira desigual diferentes grupos demográficos ou categorias de produtos.

Diante desse cenário, a comunidade científica tem voltado sua atenção para a "Justiça Algorítmica" (*Fairness*). A questão central deixa de ser quão assertivas são as recomendações e passa a olhar para como essas recomendações são distribuídas dentre os subgrupos dos *datasets*.

Este trabalho propõe uma análise comparativa de métricas tradicionais de avaliação de SRs em conjunto com métricas de justiça. O estudo utiliza algoritmos de Filtragem Colaborativa aplicados a múltiplos domínios (filmes, livros, músicas e animes) para compreender as limitações das abordagens convencionais. O objetivo é explorar como diferentes modelos balanceiam o equilíbrio (*trade-off*) entre precisão preditiva e equidade na distribuição de oportunidades, tanto para os usuários que consomem o conteúdo quanto para os itens que são recomendados.

O texto está organizado em sete seções que seguem uma progressão lógica da fundamentação teórica à análise empírica e conclusões. Inicialmente, a Seção 2 estabelece a base conceitual, apresentando os princípios dos Sistemas de Recomendação, com foco na Filtragem Colaborativa, e introduz as definições de Justiça Algorítmica (*Fairness*) que orientam toda a avaliação posterior.

A Seção 3 detalha os quatro *datasets* selecionados: MovieLens, Amazon Music, Book Crossing e MyAnimeList, descrevendo suas características, níveis de esparsidade e as etapas de pré-processamento realizadas para padronização. Em seguida, a Seção 4 expõe a metodologia experimental, apresentando a biblioteca *Surprise*, os algoritmos de recomendação avaliados (como k-NN e SVD) e as métricas de acurácia tradicionais utilizadas (RMSE, MAE e FCP).

Os resultados são apresentados e discutidos em duas etapas principais: a Seção 5 analisa o desempenho preditivo dos modelos em cada domínio, enquanto a Seção 6 conduz a avaliação de justiça, empregando a biblioteca *Fairlearn* para contrastar a equidade sob a perspectiva do usuário (atributos demográficos) e do item (categorias de conteúdo). Por fim, a Seção 7 sintetiza as descobertas do estudo, destacando as disparidades identificadas entre justiça do usuário e do item, e propõe direções para pesquisas futuras.

2 Fundamentação

Sistemas de Recomendação têm se tornado componentes centrais em aplicações modernas, como plataformas de *streaming*, comércio eletrônico e redes sociais, auxiliando os usuários a descobrirem novos conteúdos relevantes. Esses sistemas podem ser classificados em três grandes categorias:

- **Baseados em Conteúdo**: utilizam informações sobre os itens (como gênero, descrição ou características) e sobre o histórico do usuário para recomendar itens semelhantes aos que ele já demonstrou interesse.
- **Filtragem Colaborativa**: baseiam-se em padrões de interação entre usuários e itens, inferindo preferências futuras a partir de similaridades entre perfis de usuários ou itens avaliados. É a abordagem foco deste trabalho.
- **Híbridos**: combinam técnicas baseadas em conteúdo e colaborativas para aproveitar as vantagens de ambas.

Além da questão da acurácia, que tradicionalmente é a principal métrica de avaliação em SRs, a literatura recente tem destacando a importância de considerar aspectos de *justiça* (*fairness*). Nesse contexto, busca-se garantir que os algoritmos não apenas apresentem boas predições, mas também tratem diferentes grupos de usuários e itens de forma equilibrada, evitando a amplificação de desigualdades já existentes nos dados.

Assim, a fundamentação teórica deste trabalho abrange tanto os conceitos de filtragem colaborativa e suas variantes quanto a discussão sobre justiça em sistemas de recomendação, que servirá de base para a análise desenvolvida nas seções seguintes.

3 Conjuntos de Dados

A seleção de *datasets* para este estudo visa cobrir uma variedade de desafios e domínios comuns na pesquisa em Sistemas de Recomendação, garantindo a avaliação da acurácia e justiça além da generalização dos modelos. As fontes de dados incluem o repositório caserec/Datasets-for-Recommender-Systems [2] e a biblioteca de software Surprise [1].

3.1 Análise Exploratória dos Conjuntos de Dados

Os quatro *datasets* escolhidos, MovieLens, Book Crossing, Amazon Music e MyAnimeList, apresentam naturezas de interação e níveis de esparsidade distintos. As características essenciais de cada conjunto de dados, incluindo o número de avaliações, usuários, itens e o nível de esparsidade, são sumarizadas na Tabela 1.

Tabela 1: Análise Exploratória dos Conjuntos de Dados

Dataset	**Avaliações**	**Usuários**	**Itens**	**Esparsidade**
MovieLens 100k	100.000	943	1.682	93,70%
MovieLens 1M	1.000.209	6.040	3.705	95,53%
Amazon Music	64.706	5.541	3.568	99,68%
Book Crossing	62.656	1.295	14.684	99,67%
MyAnimeList	419.943	4.714	7.157	98,75%

3.1.1 MovieLens

Este *dataset* é o mais utilizado para benchmarking em SRs. Ele consiste em um grande volume de avaliações explícitas de filmes, além de dados demográficos detalhados dos usuário. O conjunto foi obtido diretamente por meio do código da biblioteca Surprise. Por ser uma fonte amplamente utilizada e já organizada para experimentos, é frequentemente o ponto de partida para a validação de novos modelos.

3.1.2 Book Crossing

Um *dataset* notório pela sua extrema esparsidade e pela natureza mista de suas avaliações. Uma característica notável desse conjunto de dados, é o fato dos itens (livros) terem poucas avaliações no geral, mas cada usuário tem muitos livros avaliados.

3.1.3 Amazon Music

Representando um ecossistema de e-commerce de grande volume, este dataset contém avaliações de música e possui uma rica dimensão temporal (carimbos de data/hora), sendo crucial para experimentos que abordam escalabilidade e a dinâmica temporal das preferências.

3.1.4 MyAnimeList

Um *dataset* de nicho que registra avaliações de usuários para animações japonesas (animes). Também possui dados de atividade dos usuários que fizeram acesso à página de um produto, mas não deixaram uma avaliação. Sua inclusão é importante para avaliar a performance de modelos em contextos especializados, onde a alta concordância ou polarização de gostos pode ser um fator.

3.2 Pré-Processamento

O pré-processamento dos dados visa padronizar os *datasets* utilizados, normalizando os dados e removendo ruído.

A principal técnica de pré-processamento empregada nos *datasets* é o Filtro de k-core, que remove usuários e itens com um número de interações abaixo de um limiar predefinido (k). Essa filtragem é essencial para mitigar o efeito de *cold start*, onde um item ou usuário novos, têm suas recomendações afetadas pela falta de dados históricos. [4]

Devido à natureza dos dados, que possuem diferentes faixas de avaliação, foi necessário definir como padrão notas de 1–5. Para isso aplicamos uma normalização nas avaliações dos *datasets* MyAnimeList e BookCrossing, que possuem notas de 1–10.

Além disso, alguns conjuntos de dados do repositório *caserec* [2] também receberam uma melhoria nos identificadores dos usuários e itens, mapeando dados mais complexos para identificadores inteiros simples, visando uma melhoria na performance dos modelos.

O tratamento específico aplicado a cada base de dados é descrito a seguir:

- **MovieLens:** A biblioteca Surprise já lida com o download, limpeza básica e formatação em um formato padronizado. Porém, aplicamos um filtro k-core, limitando o número mínimo de avaliações por filme e número mínimo de avaliações por usuário.
- **Book Crossing:** Aplicação de filtro k-core, normalização das avaliações (de 1–10 para 1–5) e remoção de avaliações implícitas (0) para isolar o sinal de rating explícito.
- **Amazon Music:** Aplicação de filtro k-core e extração da tupla de interação (ID, ID, Rating, Timestamp), conversão de identificadores alfanuméricos para números inteiros.
- **MyAnimeList:** Aplicação de filtro k-core e normalização das avaliações (de 1–10 para 1–5).

4 Modelos de Filtragem Colaborativa

4.1 Biblioteca Surprise

A biblioteca Surprise (Simple Python Recommendation System Engine)[1], é um scikit projetado para a construção e análise de sistemas de recomendação, com foco em algoritmos de Filtragem Colaborativa. A Surprise foi desenvolvida para fornecer um conjunto de ferramentas acessíveis e simples de usar. A biblioteca possui documentação clara que explica o funcionamento dos algoritmos implementados.

A Surprise foi projetada para lidar com dados de avaliação explícita, por exemplo os dados em que os usuários fornecem uma nota numérica direta para os itens, como em uma escala de 1 a 5 estrelas. A biblioteca não oferece suporte para feedback implícito, como cliques, visualizações e compras, e nem para informações baseadas em conteúdo, como gênero de filme e descrição do produto.

4.2 Modelos

A biblioteca Surprise oferece diversos algoritmos para treinamento de sistemas de recomendação já implementados. Esses modelos incluem algoritmos de baseline, modelos de vizinhança e baseados em fatoração de matrizes.

- **Baseline:** Prediz avaliações usando apenas a média global e os vieses de usuário e item. Modelo rápido mas não muito preciso.

- **k-NN Basic:** Usa as avaliações dos k-vizinhos mais próximos para prever uma avaliação. É um algoritmo de similaridade, utiliza os dados dos k-vizinhos mais próximos para prever a nota do usuário.
- **k-NN Baseline:** Combina a abordagem KNN com estimativas de linha de base para maior precisão. Procura descontar vieses e achar os k-vizinhos de maneira mais precisa. É melhor para tratar de problemas de esparsidade.
- **SVD:** Decompõe a matriz de avaliações em fatores latentes de usuário e item. Como o modelo generaliza com base nos padrões, consegue tratar o cold start, utilizando poucos dados para estimar fatores latentes.
- **SVD++:** Estende o SVD para incorporar informações de feedback implícito. O modelo leva em consideração o fato de que o usuário escolheu avaliar um item, não leva em consideração somente as notas.

4.3 Métricas Quantitativas de Performance

A biblioteca *Surprise* oferece diversas métricas para avaliar modelos de predição contínua, que são baseados na diferença entre o valor previsto e o valor real. Essas métricas representam variações na forma de calcular o erro. A seguir, apresentamos as principais utilizadas neste trabalho:

- **RMSE (Root Mean Squared Error)**: Métrica que calcula o desvio médio entre as predições do modelo e os valores reais das avaliações. Equivalente ao desvio padrão dos erros de predição, o RMSE mede a magnitude média do erro, penalizando mais fortemente erros maiores devido à sua natureza quadrática. Assim, um erro de 2,0 pontos contribui quatro vezes mais para o erro total do que um erro de 1,0 ponto. Calculado conforme a Equação (1).

$$RMSE = \sqrt{\frac{1}{|\hat{R}|} \sum_{(u,i)\in\hat{R}} (r_{ui} - \hat{r}_{ui})^2} \tag{1}$$

  r_{ui} é a avaliação verdadeira, $\hat{r}_{ui}$ é a avaliação prevista e $|\hat{R}|$ é o número total de avaliações no conjunto de teste.

- **MAE (Mean Absolute Error)**: Erro absoluto médio, uma alternativa ao RMSE que oferece uma interpretação diferente. O MAE mede a média da magnitude absoluta dos erros. O MAE trata os erros de forma linear, diferente do RMSE. Um erro de 2.0 pontos tem o dobro do peso que um erro de 1.0 ponto, tornando a métrica mais intuitiva e robusta a outliers. Definido pela Equação (2).

$$MAE = \frac{1}{|\hat{R}|} \sum_{(u,i)\in\hat{R}} |r_{ui} - \hat{r}_{ui}| \tag{2}$$

  r_{ui} é a avaliação verdadeira, $\hat{r}_{ui}$ é a avaliação prevista e $|\hat{R}|$ é o número total de avaliações no conjunto de teste.

- **FCP (Fraction of Concordant Pairs)**: Este modelo é uma métrica de ranking, ele busca saber se o usuário prefere o item A ao item B. O FCP compara pares de itens e analisa como par concordante, quando o modelo acerta a predição, e par discordante, quando o modelo erra a predição.

$$FCP = \frac{N_{concordantes}}{N_{concordantes} + N_{discordantes}} \tag{3}$$

$N_{concordantes}$ é o número de pares concordantes e $N_{discordantes}$ é o número de pares discordantes.

4.4 Experimentos com a Biblioteca Surprise

Foram conduzidos experimentos com algoritmos de filtragem colaborativa com o objetivo de obter uma avaliação da acurácia dos modelos fornecidos pela biblioteca Surprise em diferentes domínios, dado a diversidade de *datasets*.

Para essa avaliação, realizou-se uma validação cruzada dos modelos utilizando a técnica de **k-Fold**, na qual o conjunto de dados é dividido em k subconjuntos (ou *folds*) de tamanho semelhante. Em cada iteração, um dos subconjuntos é utilizado como conjunto de teste, enquanto os k-1 restantes servem como conjunto de treino, repetindo-se o processo até que todos os subconjuntos tenham sido usados como teste. [3] Para os experimentos realizados utilizamos uma validação cruzada com cinco *folds* ($k = 5$), os resultados à seguir representam a média das métricas coletadas para cada *fold*.

Foram coletadas métricas de acurácia disponíveis pela biblioteca (RMSE, MAE e FCP), além do tempo de execução de cada treinamento. Essa métricas indicam o erro encontrado à partir das predições realizadas, por cada modelo treinado, em cima do subconjunto de teste.

5 Avaliação de acurácia

5.1 Resultados MovieLens 100k

Modelo	RMSE	MAE	FCP	Tempo (s)
Baseline	0.9406	0.7457	0.6987	4.05
k-NN	1.0059	0.7964	0.7098	25.52
k-NN Baseline	0.9157	0.7183	0.7190	30.09
SVD	0.9307	0.7334	0.7043	5.93
SVD++	0.9241	0.7295	0.7079	44.16

Tabela 2: Resultados de acurácia do dataset MovieLens 100k por modelo.

Neste *dataset*, o modelo k-NN Baseline apresentou o melhor desempenho em termos de acurácia, atingindo os menores valores de erro, RMSE e MAE. A abordagem de vizinhança que considera os vieses foi a mais eficaz. O FCP indica que k-NN Baseline também foi o melhor modelo na predição de ranking do usuário.

5.2 Resultados MovieLens 1M

Modelo	RMSE	MAE	FCP	Tempo (s)
Baseline	0.9083	0.7191	0.7240	78.46
k-NN	0.9881	0.7809	0.7197	1155.02
k-NN Baseline	0.8726	0.6824	0.7487	1227.09
SVD	0.8965	0.7092	0.7316	88.82
SVD++	0.8891	0.7017	0.7331	713.51

Tabela 3: Resultados de acurácia do dataset MovieLens 1M por modelo.

No *dataset* MovieLens 1M, observa-se um comportamento semelhante ao encontrado no conjunto menor. O modelo k-NN Baseline novamente apresentou o melhor desempenho geral, alcançando os menores valores de RMSE e MAE entre todos os modelos testados. Entretanto, o custo computacional aumentou drasticamente: tanto k-NN quanto k-NN Baseline apresentaram tempos de execução muito elevados.

Os métodos baseados em fatoração, SVD e SVD++, apresentaram desempenho competitivo e bem menor em termos de tempo de execução, especialmente quando comparados ao k-NN.

No geral, os resultados indicam que, embora k-NN Baseline ofereça a melhor acurácia, sua escalabilidade é limitada. Já os modelos fatorados mantêm um bom equilíbrio entre desempenho e eficiência computacional em *datasets* maiores como o MovieLens 1M.

5.3 Resultados Amazon Music

Modelo	RMSE	MAE	FCP	Tempo (s)
Baseline	0.9270	0.7045	0.6123	1.63
k-NN	1.0687	0.7851	0.6370	7.77
k-NN Baseline	0.9919	0.7174	0.5927	7.64
SVD	0.9149	0.6814	0.5959	3.76
SVD++	0.9209	0.6915	0.5989	7.90

Tabela 4: Resultados de acurácia para Amazon Music por modelo.

Neste *dataset*, o modelo com melhor acurácia foi o SVD, tendo o segundo melhor tempo de execução também. O Modelo k-NN teve um erro consideravelmente alto neste experimento. O FCP indica que mesmo o modelo SVD tendo melhores métricas de erro de predição, o modelo k-NN é melhor em acertar a ordem de preferência do usuário.

5.4 Resultados Book Crossing

Modelo	RMSE	MAE	FCP	Tempo (s)
Baseline	0.6602	0.5061	0.5629	2.20
k-NN	0.8487	0.6504	0.5302	3.41
k-NN Baseline	0.7212	0.5434	0.5574	3.81
SVD	0.6595	0.5009	0.5554	3.24
SVD++	0.6588	0.5012	0.5540	16.11

Tabela 5: Resultados de acurácia do dataset Book Crossing por modelo.

Como este é um *dataset* com grande esparsidade, os modelos que lidam melhor com isso, SVD e SVD++, têm melhores métricas de acurácia. Devido a esparsidade, todos modelos tiveram métricas de FCP ruins, indicando uma dificuldade em acertar a ordem de preferência do usuário.

5.5 Resultados MyAnimeList

Modelo	RMSE	MAE	FCP	Tempo (s)
Baseline	0.5401	0.4121	0.6873	28.00
k-NN	0.6547	0.5011	0.6857	291.03
k-NN Baseline	0.5198	0.3920	0.7117	309.73
SVD	0.5369	0.4085	0.6879	38.55
SVD++	0.5380	0.4096	0.6868	284.20

Tabela 6: Resultados de acurácia do dataset Anime por modelo.

O modelo k-NN Baseline obteve as melhores métricas de acurácia porém o maior tempo de execução. Os modelos SVD e SVD++ tiveram métricas muito parecidas entre si porém o SVD++ demandou um tempo de execução muito maior que o SVD. O modelo k-NN Baseline também obteve a melhor métrica de FCP, mostrando ser mais eficiente na predição de ordem de preferência do usuário.

6 Avaliação de Justiça

A justiça em aprendizado de máquina é um conceito multidimensional que visa assegurar que os algoritmos operem sem discriminação ou viés indevido. No contexto de Sistemas de Recomendação, a justiça pode ser definida como a ausência de preconceito sistemático nas sugestões fornecidas aos usuários.

A literatura divide a justiça em duas categorias principais:

- **Justiça Individual:** Exige que indivíduos semelhantes sejam tratados de maneira semelhante.
- **Justiça de Grupo:** Exige que grupos definidos por atributos protegidos ou sensíveis (como gênero, raça, idade ou categoria de produto) recebam tratamento estatisticamente equivalente por parte do modelo.

Devido à utilização da biblioteca *Fairlearn* como ferramenta de auditoria neste trabalho, nossa avaliação foca na Justiça de Grupo. A *Fairlearn* opera avaliando discrepâncias de métricas entre subgrupos pré-definidos nos dados. Portanto, para realizar a avaliação, é necessário discretizar os dados em categorias específicas para verificar se o modelo privilegia um grupo em detrimento de outro.

Nas fórmulas apresentadas a seguir, utilizadas para calcular as métricas de justiça, como Paridade Demográfica e Igualdade de Oportunidade, as variáveis representam:

- A: O atributo sensível ou o grupo que está sendo avaliado (ex: Gênero = Feminino).
- Y: O rótulo real (*ground truth*), ou seja, se o item é realmente relevante para o usuário (ex: o usuário de fato avaliou positivamente).
- $\hat{Y}$: A predição do modelo, ou seja, se o sistema classificou o item como relevante.

Assim, buscamos medir se a relação entre a predição $\hat{Y}$ e o valor real Y se mantém constante independentemente do grupo A analisado.

6.1 Métricas de Justiça

A avaliação de justiça em sistemas de recomendação busca identificar e mitigar vieses sistemáticos que possam favorecer ou desfavorecer determinados grupos de usuários. Esses vieses geralmente se originam dos próprios dados de treino, que refletem desigualdades históricas ou padrões de uso enviesados, e podem ser amplificados pelos modelos de recomendação.

Para esta análise, foram consideradas três métricas clássicas da literatura de aprendizado justo, todas disponíveis na biblioteca *Fairlearn*. Cada uma delas mede aspectos distintos de equidade no desempenho do sistema:

- **Demographic Parity (Paridade Demográfica)**: Essa métrica avalia se a probabilidade de o modelo fazer uma predição positiva (ou de alta recomendação) é independente do grupo sensível (como gênero ou idade). Formalmente, exige que:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

onde A representa o atributo sensível. Em sistemas de recomendação, a paridade demográfica indica que homens e mulheres (ou jovens e idosos, por exemplo) têm chances semelhantes de receber recomendações de alta pontuação. Valores próximos de zero indicam menor disparidade entre grupos.

- **Equal Opportunity (Igualdade de Oportunidade)**: Mede se todos os grupos têm a mesma taxa de verdadeiros positivos, isto é, se os itens realmente relevantes têm igual chance de serem corretamente recomendados para diferentes grupos. Formalmente:

$$P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1)$$

  Essa métrica é útil quando o foco é garantir que usuários de todos os grupos tenham as mesmas oportunidades de receber boas recomendações.

- **Equalized Odds (Oportunidades Equalizadas)**: É uma generalização da métrica anterior. Ela exige que tanto as taxas de verdadeiros positivos quanto as taxas de falsos positivos sejam equivalentes entre os grupos:

$$P(\hat{Y} = 1|Y = y, A = 0) = P(\hat{Y} = 1|Y = y, A = 1) \quad \forall y \in \{0, 1\}$$

  Essa métrica é mais restritiva, pois visa garantir que o modelo não apenas recomende igualmente bem, mas também cometa erros de forma balanceada entre os grupos.

6.2 Avaliação de Justiça com o dataset Amazon Music

O *dataset* de *Amazon Music* não possui dados demográficos dos usuários que avaliaram os álbuns, com isso, foi realizada uma avaliação de justiça para os itens do conjunto de dados.

Os álbuns contidos no *dataset* vieram previamente classificados dentre 52 gêneros musicais, coletados as métricas de acurácia dentro de cada um desses grupos.

Justiça por Gênero do Álbum

Os resultados de acurácia por gênero musical no *Amazon Music* (Figuras 1 e 2) revela padrões distintos entre os modelos. O SVD demonstra desempenho superior consistente, alcançando os menores valores de RMSE e MAE na maioria dos gêneros. Os modelos k-NN apresentaram alta variabilidade, com desempenho superior ao SVD para alguns gêneros, mas com erro superior à 1.0 para outros gêneros.

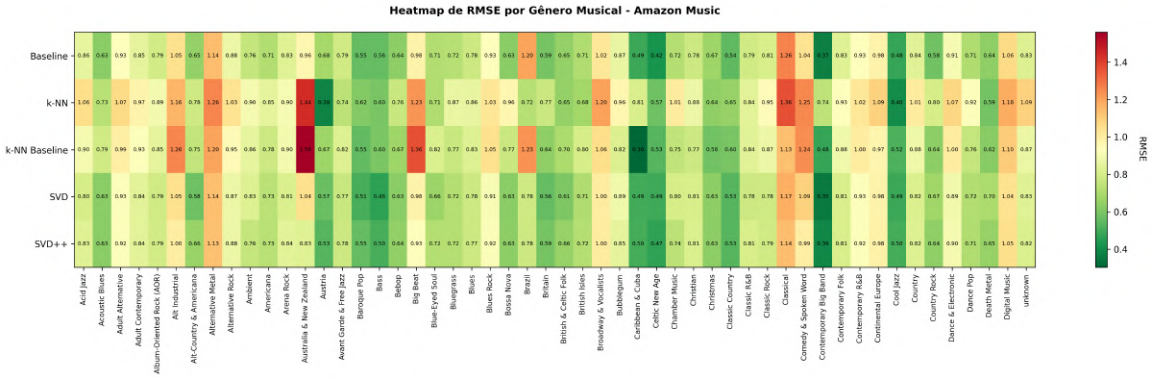

Figura 1: Acurácia RMSE por gênero do álbum.

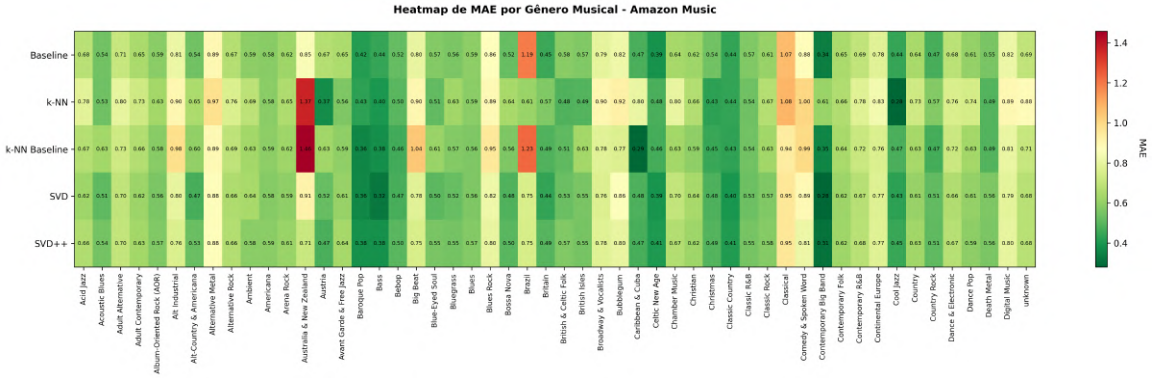

Figura 2: Acurácia MAE por gênero do álbum.

Os modelos baseados em fatoração matricial superam consistentemente os métodos baseados em vizinhança. Gêneros nichados como Classical e Comedy & Spoken Word apre-

sentam erros elevados em todos os modelos, sugerindo maior dificuldade de predição. O Baseline mantém performance intermediária estável, enquanto k-NN Baseline sofre com *overfitting* em gêneros específicos. A disparidade entre gêneros indica necessidade de aplicar tratamento nos dados para categorias com menor volume de avaliações.

A análise por gênero musical (Tabela 7) revelou disparidades significativas em todos os modelos. Valores elevados de Paridade Demográfica (0,74-0,83) indicam forte viés na distribuição de recomendações entre gêneros, enquanto o Equalized Odds máximo (1,000) demonstra que as taxas de erro são completamente desbalanceadas. Estes resultados sugerem que os algoritmos amplificam vieses existentes na base, favorecendo certos gêneros musicais de forma sistemática.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.7733	1.0000
k-NN	0.7447	1.0000
k-NN Baseline	0.8300	1.0000
SVD	0.7805	1.0000
SVD++	0.7564	1.0000

Tabela 7: Justiça por gênero do álbum.

6.3 Avaliação de Justiça com o dataset MyAnimeList

Visto que a biblioteca *Fairlearn* avalia a justiça baseada em grupos, foi necessário definir explicitamente quais segmentos dos dados seriam submetidos à análise de disparidade.

Como o *dataset* MyAnimeList não fornece dados demográficos dos usuários, a estratégia adotada foi avaliar a justiça sob a perspectiva do **Item**. Para isso, os animes foram segmentados em grupos com base em suas características intrínsecas. Escolhemos dividir os grupos pelo número de episódios (para verificar se obras curtas são prejudicadas em relação a obras longas) e pelo gênero da obra. Essa abordagem permite verificar se o algoritmo oferece "oportunidades iguais" de recomendação para diferentes tipos de produções artísticas.

Os diferentes animes analisados foram então classificados por duas informações: número de episódios e gênero.

Para a primeira classificação, dividimos as animações em 4 grupos: episódios únicos, obras curtas (até 12 episódios), obras de tamanho padrão (entre 13 e 25 episódios) e obras longas (com mais de 25 episódios).

As diferentes obras presentes no *dataset* vieram classificadas dentre 40 diferentes gêneros , utilizamos essa divisão para calcular as métricas de justiça dos itens.

Justiça por Número de Episódios

A análise por duração revela que animes com número padrão de episódios, recebem recomendações mais precisas em todos os modelos, enquanto episódios únicos e séries longas têm performance inferior. O modelo k-NN Baseline demonstrou a melhor acurácia geral dentre todas as categorias de duração.

Modelo	Episódio Único		Curto (<=12)		Padrão (<=25)		Longo (>=26)	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Baseline	0.5615	0.4273	0.5416	0.4122	0.5255	0.4014	0.5367	0.4158
k-NN	0.6681	0.5117	0.6691	0.5111	0.6357	0.4862	0.6409	0.4978
k-NN Baseline	0.5356	0.4019	0.5241	0.3948	0.5061	0.3821	0.5190	0.3964
SVD	0.5574	0.4230	0.5372	0.4081	0.5232	0.3985	0.5371	0.4143
SVD++	0.5583	0.4241	0.5376	0.4086	0.5235	0.3988	0.5366	0.4141

Tabela 8: Acurácia por número de episódios.

A análise de justiça por categoria de episódios revela vieses moderados, com valores de Paridade Demográfica entre 0,0087 (k-NN) e 0,0262 (k-NN Baseline), e Equalized Odds entre 0,0445 (k-NN) e 0,1350 (k-NN Baseline). Animes de episódio único apresentam maior RMSE, enquanto séries padrão obtêm melhor acurácia (0,51-0,64), provavelmente efeito do volume de dados. O k-NN Baseline, apesar de mais acurado globalmente, exibe o maior viés entre categorias.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.0140	0.0779
k-NN	0.0087	0.0445
k-NN Baseline	0.0262	0.1350
SVD	0.0185	0.0979
SVD++	0.0180	0.1027

Tabela 9: Justiça por número de episódios.

Justiça por Gênero do Anime

A análise de acurácia por gênero primário revela variações significativas no desempenho dos modelos de recomendação. Os valores de RMSE (Figura 3) variam de 0,09 (Policial) a 1,29 (Sobrenatural), evidenciando heterogeneidade nos padrões de avaliação. Gêneros como Policial, Vampiro e Samurai apresentam melhor desempenho preditivo, enquanto Sobrenatural, Yaoi e Hentai demonstram maiores erros((Figura 4). O k-NN Baseline obteve o melhor ranking médio (2,15), contrastando com o k-NN tradicional (4,70).

A superioridade do k-NN Baseline neste dataset se deve à sua capacidade de capturar padrões locais fortes enquanto normaliza vieses individuais, sendo mais robusto à gêneros nichados, com baixo volume de dados.

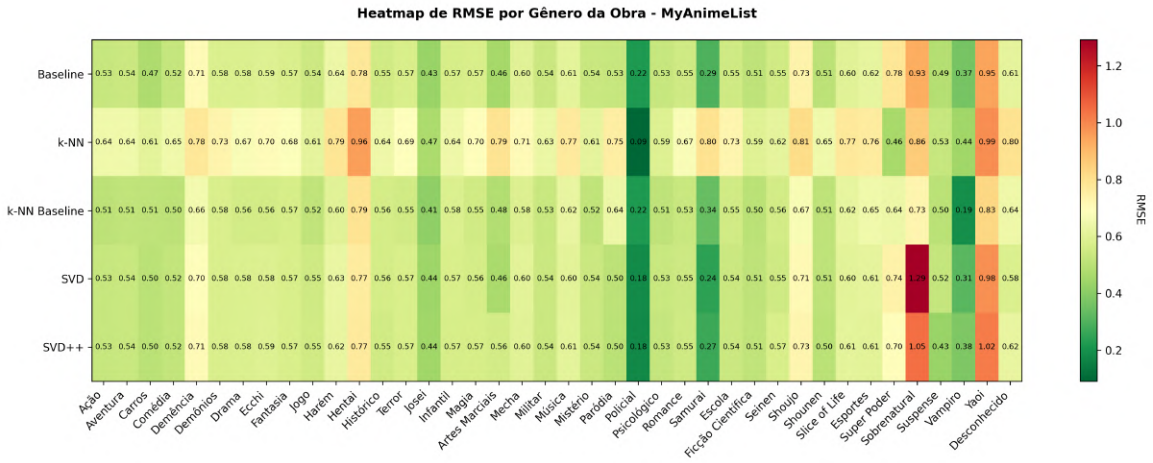

Figura 3: Acurácia RMSE por gênero.

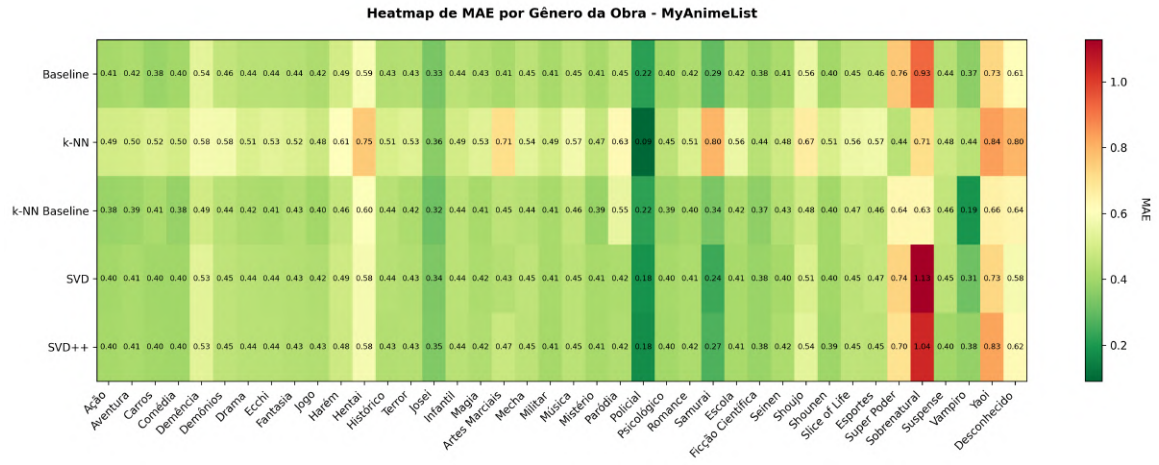

Figura 4: Acurácia MAE por gênero.

Foram identificadas (Tabela 10) disparidades significativas entre gêneros de anime, com valores elevados de Paridade Demográfica (0,18-0,65) e Equalized Odds máximo em todos os modelos. Os algoritmos baseados em vizinhança (k-NN) mostraram os maiores vieses, enquanto o Baseline foi relativamente menos enviesado, porém ainda problemático.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.1834	1.0000
k-NN	0.6500	1.0000
k-NN Baseline	0.6564	1.0000
SVD	0.2733	1.0000
SVD++	0.2426	1.0000

Tabela 10: Justiça por gênero.

6.4 Avaliação de Justiça com o dataset Book Crossing

O *dataset* de avaliações de livros do *Book Crossing* também possui dados do usuário, idade a país de origem, além da classificação dos livros.

Os usuários foram agrupados em faixas de idade, jovem (até 25 anos), adulto (25 a 60 anos) e idoso (acima de 60 anos), além de serem classificados por país de origem.

Também fizemos a análise da justiça utilizando os dados de classificação dos itens, por época de publicação e número de livros publicados pela editora.

Definimos alguns intervalos para as épocas de publicação: 1920-1949, 1950-1969, 1970-1989, 1990-1999 e 2000-2010. Também classificamos as editoras pelo número de livros publicados, criando algumas categorias de tamanho: até 5 publicações, de 6 a 20, de 21 a 50, de 51 a 99, de 100 a 249 e acima de 250 publicações.

6.4.1 Justiça de Usuário

Justiça por Idade

A análise por faixa etária revela que usuários idosos recebem recomendações consideravelmente mais precisas, enquanto jovens têm a pior performance em todos os modelos. O SVD++ destacou-se como o melhor algoritmo no geral, com o k-NN apresentando resultados inferiores dentre todas as faixas etárias.

Modelo	Jovem		Adulto		Idoso	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Baseline	0.6779	0.5193	0.6571	0.5044	0.6201	0.4835
k-NN	0.8922	0.6935	0.8389	0.6414	0.8406	0.6471
k-NN Baseline	0.7282	0.5423	0.7205	0.5441	0.6742	0.5215
SVD	0.6789	0.5126	0.6559	0.4984	0.6132	0.4734
SVD++	0.6775	0.5125	0.6561	0.4995	0.6088	0.4720

Tabela 11: Acurácia por faixa etária do usuário.

Os modelos apresentam comportamentos distintos em termos de justiça etária. O k-NN demonstrou excelente equidade com valores muito baixos, enquanto os demais algoritmos mostraram disparidades moderadas, especialmente no Equalized Odds (0,31-0,33), indicando variações nas taxas de erro entre diferentes faixas etárias.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.0476	0.3276
k-NN	0.0086	0.0314
k-NN Baseline	0.0468	0.3136
SVD	0.0445	0.3093
SVD++	0.0467	0.3331

Tabela 12: Justiça por faixa etária do usuário.

Justiça por País

Os modelos SVD++ e SVD apresentaram melhor desempenho geral, seguidos pelo Baseline, enquanto o k-NN teve pior performance. Trinidad e Tobago apresentou maior erro, enquanto Nova Zelândia tiveram menor erro, sendo ambos com um baixo volume de dados. Países com mais dados, como EUA e Alemanha, mostraram erros moderados(Figuras 5 e 6).

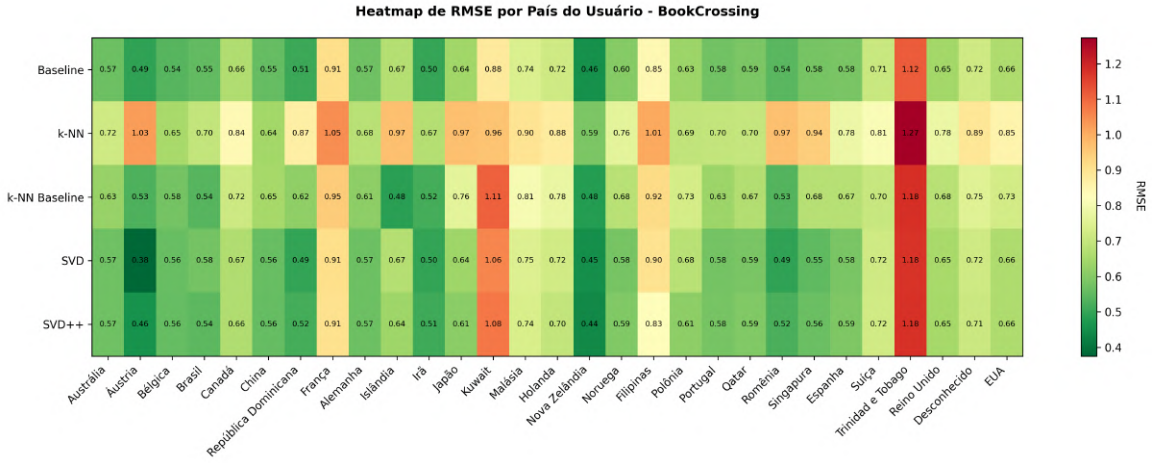

Figura 5: Acurácia RMSE por país do usuário.

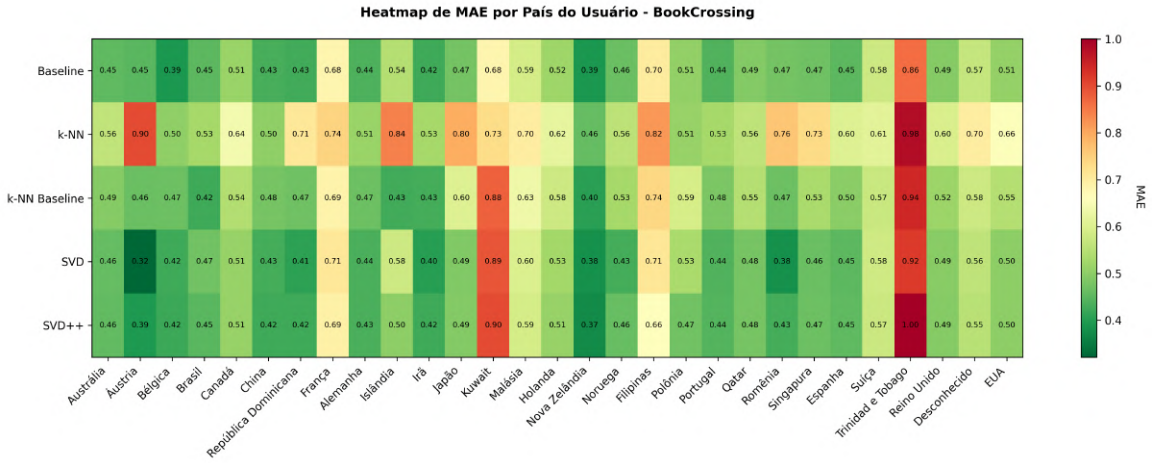

Figura 6: Acurácia MAE por país do usuário.

Foram identificadas (Tabela 13) disparidades extremas entre nacionalidades, com Equalized Odds máximo (1,000) em todos os modelos, indicando que as taxas de erro variam completamente entre países. O Baseline foi o menos enviesado, porém todos os algoritmos apresentam problemas graves de justiça geográfica.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.0788	1.0000
k-NN	0.1556	1.0000
k-NN Baseline	0.1486	1.0000
SVD	0.0911	1.0000
SVD++	0.0962	1.0000

Tabela 13: Justiça por país do usuário.

6.4.2 Justiça de Item

Justiça por Época de Publicação

Livros mais antigos (1920-1949) recebem recomendações mais precisas, especialmente com o modelo SVD++, enquanto obras dos anos 70-90 apresentam performance inferior. O SVD++ destacou-se como o melhor algoritmo dentre todas as faixas de tempo, com o k-NN mostrando resultados consistentemente piores.

Modelo	1920–1949		1950–1969		1970–1989		1990–1999		2000–2010	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Baseline	0.6026	0.5021	0.6517	0.5312	0.6588	0.5070	0.6646	0.5097	0.6538	0.5010
k-NN	0.8727	0.6738	0.8277	0.6411	0.8569	0.6584	0.8567	0.6559	0.8282	0.6342
k-NN Baseline	0.6730	0.5603	0.7061	0.5604	0.7154	0.5428	0.7294	0.5488	0.7108	0.5363
SVD	0.6252	0.5192	0.6465	0.5211	0.6554	0.4985	0.6631	0.5033	0.6547	0.4967
SVD++	0.5755	0.4756	0.6490	0.5213	0.6551	0.4998	0.6629	0.5041	0.6538	0.4966

Tabela 14: Acurácia por época de publicação do item.

Os modelos apresentam boa paridade demográfica entre épocas de publicação, porém o Equalized Odds moderado a alto (0,26-0,50) revela que as taxas de erro variam significativamente entre diferentes décadas. O Baseline foi o mais equilibrado, enquanto o SVD++ mostrou as maiores disparidades.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.0123	0.2587
k-NN	0.0346	0.4360
k-NN Baseline	0.0319	0.2977
SVD	0.0185	0.3001
SVD++	0.0266	0.4950

Tabela 15: Justiça por época de publicação de item.

Justiça por Número de Publicações da Editora

Os modelos SVD++, SVD e Baseline apresentaram melhor desempenho, consistente em todas as categorias. Diferente da análise por país, a performance foi notavelmente uniforme entre categorias de editoras, com variação mínima de erro. Editoras com de 51 a 99 e mais

de 250 livros apresentaram ligeiramente menor erro. A homogeneidade dos resultados sugere que o tamanho do catálogo da editora tem impacto limitado na acurácia das predições.

Modelo	<=5		6–20		21–50		51–99		100–249		>=250	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Baseline	0.6648	0.5097	0.6697	0.5155	0.6728	0.5183	0.6490	0.4979	0.6589	0.5063	0.6511	0.4974
k-NN	0.8448	0.6449	0.8471	0.6489	0.8514	0.6512	0.8396	0.6416	0.8525	0.6531	0.8467	0.6529
k-NN Baseline	0.7223	0.5437	0.7318	0.5541	0.7392	0.5579	0.7048	0.5297	0.7191	0.5439	0.7156	0.5384
SVD	0.6662	0.5052	0.6668	0.5087	0.6691	0.5108	0.6470	0.4902	0.6582	0.5009	0.6512	0.4917
SVD++	0.6653	0.5061	0.6683	0.5106	0.6709	0.5132	0.6473	0.4921	0.6563	0.5000	0.6503	0.4917

Tabela 16: Acurácia por número de publicações da editora.

Todos os modelos demonstraram excelente equidade em relação ao tamanho das editoras, com valores muito baixos em ambas as métricas. O k-NN destacou-se como o mais justo, indicando que os algoritmos não favorecem sistematicamente editoras grandes ou pequenas.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.0111	0.0735
k-NN	0.0114	0.0370
k-NN Baseline	0.0174	0.0866
SVD	0.0132	0.0812
SVD++	0.0129	0.0861

Tabela 17: Justiça por número de publicações da editora.

6.5 Avaliação de Justiça com o dataset MovieLens 100k

O *dataset* MovieLens 100k possui dados demográficos dos usuários, incluindo idade, gênero e ocupação, além da classificação dos filmes por gênero.

Os usuários foram agrupados por faixas etárias: jovem (até 25 anos), adulto (25 a 45 anos), meia-idade (45 a 60 anos) e idoso (acima de 60 anos), além de serem classificados por gênero e ocupação. As ocupações incluem 21 categorias diferentes, como administrador, artista, educador, engenheiro, estudante, médico, programador e aposentado.

Também fizemos a análise da justiça utilizando os dados de classificação dos itens, agrupando os filmes por gênero primário. O *dataset* possui 17 gêneros cinematográficos: ação, aventura, animação, infantil, comédia, crime, documentário, drama, fantasia, filme noir, terror, musical, mistério, romance, ficção científica, suspense e desconhecido.

6.5.1 Justiça de Usuário

Justiça por Idade

Usuários adultos recebem recomendações significativamente mais precisas na maioria dos modelos, enquanto jovens apresentam a pior performance. Os algoritmos k-NN Baseline e SVD++ destacaram-se para a faixa adulta, com o k-NN convencional mostrando resultados inferiores em todas as idades.

Modelo	Jovem		Adulto		Idoso	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Baseline	0.9718	0.7725	0.9266	0.7341	0.9325	0.7288
k-NN	1.0440	0.8323	0.9894	0.7816	0.9607	0.7499
k-NN Baseline	0.9406	0.7387	0.9056	0.7098	0.9324	0.7213
SVD	0.9599	0.7567	0.9173	0.7234	0.9348	0.7348
SVD++	0.9498	0.7491	0.9112	0.7190	0.9367	0.7291

Tabela 18: Acurácia por idade do usuário.

Os modelos demonstraram excelente equidade entre faixas etárias, com valores muito baixos em ambas as métricas. O SVD++ destacou-se como o mais justo, seguido pelo SVD, enquanto todos os algoritmos mantiveram alta imparcialidade na distribuição de recomendações por idade.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.0530	0.0749
k-NN	0.0862	0.1090
k-NN Baseline	0.0501	0.0785
SVD	0.0348	0.0741
SVD++	0.0242	0.0801

Tabela 19: Justiça por idade do usuário.

Justiça por Gênero

Foram identificadas disparidades significativas entre gêneros, com usuários masculinos recebendo recomendações consistentemente mais precisas em todos os modelos. O k-NN Baseline apresentou a melhor performance para homens, porém a diferença de acurácia entre gêneros permanece uma preocupação em todos os algoritmos.

Modelo	Feminino		Masculino	
	RMSE	MAE	RMSE	MAE
Baseline	0.9942	0.7912	0.9217	0.7301
k-NN	1.0765	0.8563	0.9807	0.7762
k-NN Baseline	0.9724	0.7656	0.8968	0.7027
SVD	0.9825	0.7773	0.9122	0.7187
SVD++	0.9762	0.7735	0.9047	0.7128

Tabela 20: Acurácia por gênero do usuário.

Os resultados demonstram excelente equidade de gênero, com valores muito baixos em ambas as métricas. O k-NN Baseline destacou-se como o mais justo, seguido pelo SVD++, indicando que os algoritmos distribuem recomendações de forma balanceada entre usuários masculinos e femininos.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.0139	0.0209
k-NN	0.0212	0.0469
k-NN Baseline	0.0061	0.0166
SVD	0.0145	0.0194
SVD++	0.0078	0.0198

Tabela 21: Justiça por gênero do usuário.

Justiça por Ocupação

Analisando os resultados de acurácia por ocupação profissional (Figuras 7 e 8) notamos disparidades significativas entre grupos no dataset MovieLens 100k. O grupo de médicos (doctor) apresenta consistentemente os melhores resultados, com RMSE variando entre 0.8087 (SVD++) e 0.9174 (k-NN), enquanto profissionais de saúde (healthcare) registram os piores desempenhos, com RMSE atingindo 1.3063 no k-NN. Aposentados (retired) e engenheiros (engineer) também demonstram acurácia superior. O modelo SVD++ oferece o melhor equilíbrio geral, minimizando disparidades entre grupos. Notavelmente, ocupações com menor volume de dados, como healthcare e homemaker, sofrem maior degradação de performance.

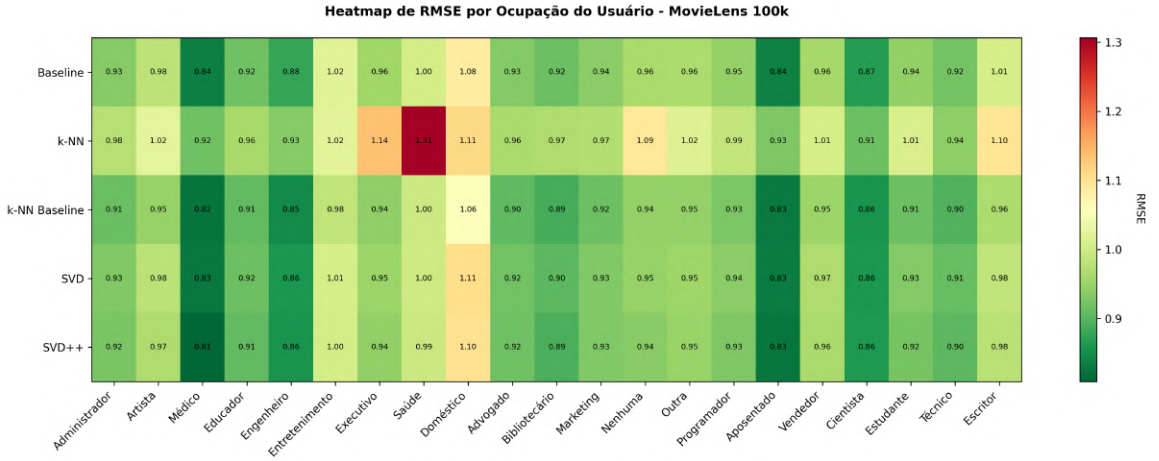

Figura 7: Acurácia RMSE por ocupação profissional do usuário.

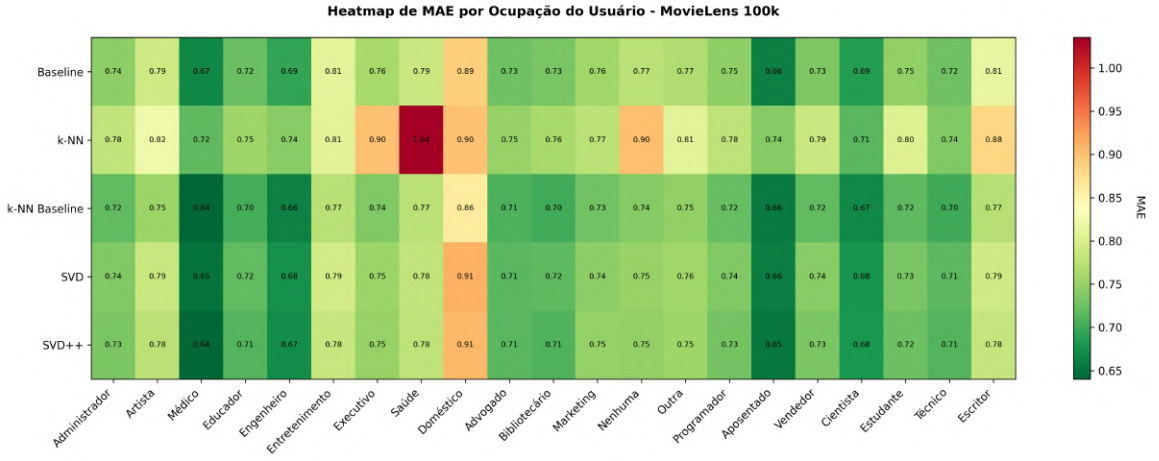

Figura 8: Acurácia MAE por ocupação profissional do usuário.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.3284	0.4052
k-NN	0.2437	0.3535
k-NN Baseline	0.3531	0.4336
SVD	0.3538	0.4543
SVD++	0.3553	0.4202

Tabela 22: Justiça por ocupação profissional do usuário.

6.5.2 Justiça de Item

Justiça por Gênero do Filme

Os resultados de acurácia por gênero de filmes no *MovieLens 100k* revelam disparidades consideráveis entre categorias cinematográficas. O gênero Suspense apresenta os piores resultados em todos os modelos, enquanto os filmes de Fantasia demonstram acurácia superior, com RMSE (Figura 9) consistentemente abaixo de 0.90. O SVD++ novamente oferece o melhor desempenho geral.

Nota-se que gêneros com mais dados obtêm métricas mais favoráveis, enquanto categorias de nicho, como Noir e Terror, sofrem degradação de performance.

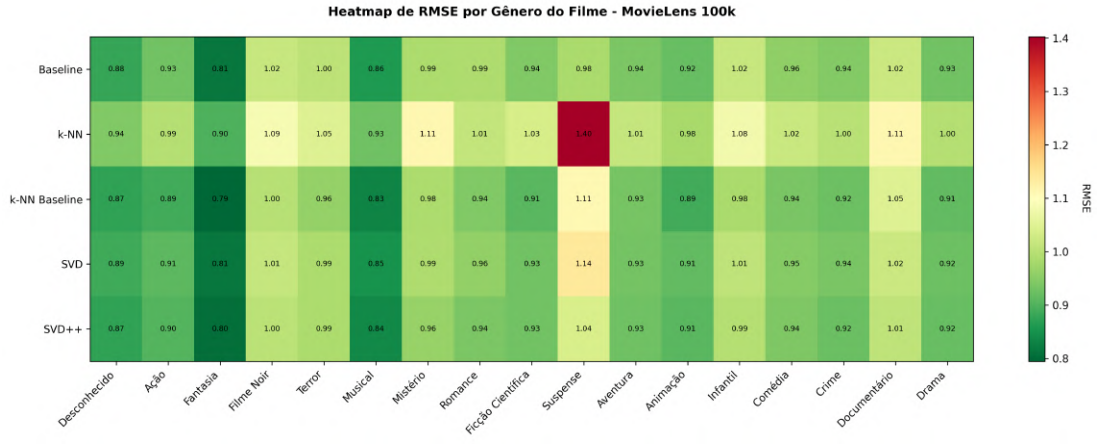

Figura 9: Acurácia RMSE por gênero do filme.

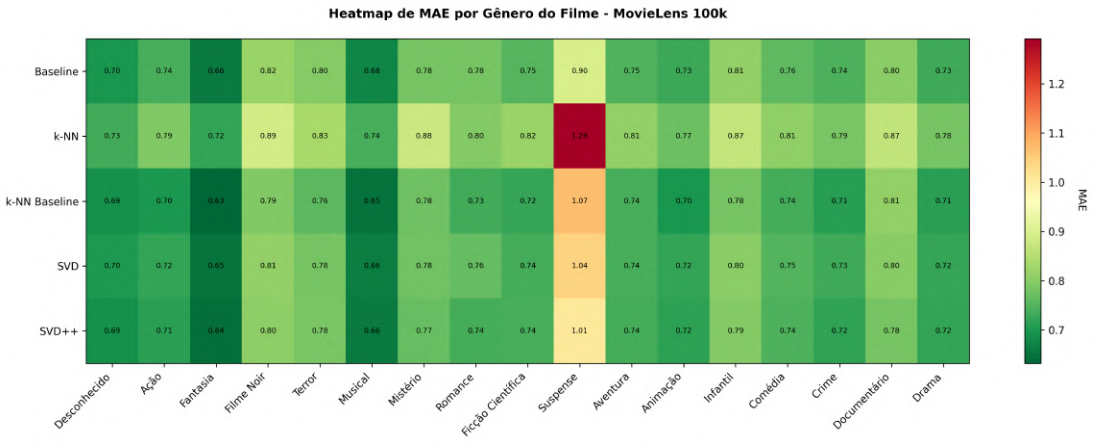

Figura 10: Acurácia MAE por gênero do filme.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.5756	0.6460
k-NN	0.8351	0.8502
k-NN Baseline	0.6077	0.7887
SVD	0.5669	0.6246
SVD++	0.5854	0.7048

Tabela 23: Justiça por gênero do filme.

6.6 Avaliação de Justiça com o dataset MovieLens 1M

O *dataset* MovieLens 1M constitui uma versão ampliada do MovieLens 100k, contendo aproximadamente 1 milhão de avaliações de 6.040 usuários para 3.900 filmes. Mantendo a mesma estrutura do conjunto menor, este *dataset* inclui informações demográficas detalhadas dos usuários, idade, gênero e ocupação, e metadados dos filmes, com classificação por múltiplos gêneros. A escala expandida oferece uma base mais robusta para análise de padrões de comportamento e vieses sistêmicos.

Para a análise de justiça do usuário, aplicamos os mesmos critérios de agrupamento do MovieLens 100k: as faixas etárias foram definidas como jovem (até 25 anos), adulto (25 a 45 anos), meia-idade (45 a 60 anos) e idoso (acima de 60 anos). A classificação por gênero e ocupação também foi mantida, abrangendo as mesmas 21 categorias ocupacionais—entre elas administrador, artista, educador, engenheiro, estudante, médico, programador e aposentado, o que permite uma comparação direta entre os dois conjuntos de dados.

Na análise de justiça do item, consideramos a classificação por gênero primário dos filmes, totalizando os mesmos 17 gêneros cinematográficos presentes no *dataset* menor: ação, aventura, animação, infantil, comédia, crime, documentário, drama, fantasia, filme noir, terror, musical, mistério, romance, ficção científica, suspense e desconhecido. Esta consistência na categorização permite avaliar se os vieses identificados no conjunto menor se mantêm, se intensificam ou se atenuam com o aumento significativo do volume de dados.

6.6.1 Justiça de Usuário

Justiça por Idade

Padrão consistente com o MovieLens 100k: usuários idosos recebem recomendações mais precisas, seguidos por adultos, enquanto jovens têm a pior performance. Os modelos k-NN Baseline e SVD++ mantiveram-se como os melhores, com o k-NN convencional apresentando resultados inferiores.

Modelo	Jovem		Adulto		Idoso	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Baseline	0.9574	0.7618	0.8971	0.7104	0.8797	0.6916
k-NN	1.0373	0.8233	0.9769	0.7715	0.9595	0.7569
k-NN Baseline	0.9125	0.7153	0.8626	0.6753	0.8564	0.6652
SVD	0.9451	0.7510	0.8865	0.7016	0.8727	0.6860
SVD++	0.9327	0.7392	0.8774	0.6918	0.8657	0.6782

Tabela 24: Acurácia por idade do usuário.

Os modelos apresentam boa equidade entre faixas etárias, com valores moderados em ambas as métricas. O k-NN Baseline destacou-se como o mais justo, seguido pelo SVD++, indicando que os algoritmos mantêm imparcialidade razoável na distribuição por idade.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.1160	0.1165
k-NN	0.1338	0.1187
k-NN Baseline	0.0998	0.0791
SVD	0.1098	0.1069
SVD++	0.1066	0.0994

Tabela 25: Justiça por idade do usuário.

Justiça por Gênero

Persiste a disparidade observada no MovieLens 100k, com usuários masculinos recebendo recomendações mais precisas em todos os modelos. O k-NN Baseline manteve-se como o melhor algoritmo para homens, porém a diferença de performance entre gêneros continua sendo uma questão relevante.

Modelo	Feminino		Masculino	
	RMSE	MAE	RMSE	MAE
Baseline	0.9369	0.7407	0.8987	0.7121
k-NN	1.0088	0.7959	0.9812	0.7758
k-NN Baseline	0.9071	0.7103	0.8611	0.6735
SVD	0.9275	0.7331	0.8877	0.7028
SVD++	0.9177	0.7243	0.8781	0.6925

Tabela 26: Acurácia por gênero do usuário.

Os resultados confirmam excelente equidade de gênero, com valores muito baixos em ambas as métricas. O SVD++ destacou-se como o mais justo, seguido pelo k-NN, demonstrando que os algoritmos distribuem recomendações de forma balanceada entre usuários de diferentes gêneros.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.0128	0.0232
k-NN	0.0059	0.0231
k-NN Baseline	0.0072	0.0229
SVD	0.0110	0.0213
SVD++	0.0030	0.0197

Tabela 27: Justiça por gênero do usuário.

Justiça por Ocupação

A avaliação da acurácia por ocupação profissional (Figuras 11 e 12) oferece uma visão detalhada de como os modelos performam para diferentes perfis de usuário. A análise dos *heatmaps* revela que, embora as diferenças não sejam extremas, existem variações consistentes na acurácia das recomendações conforme a ocupação. Por exemplo, ocupações como *Executivo* e *Autônomo* tendem a apresentar erros menores, indicando que os modelos conseguem captar melhor as preferências desses grupos. Em contrapartida, ocupações como *Desempregado* e *Estudante* mostram erros maiores. Além disso, observa-se que os modelos *k-NN Baseline* e *SVD++* mantêm desempenho mais estável e com menores erros na maioria das ocupações, enquanto o *k-NN* apresenta os valores mais altos, especialmente em certas profissões. Essas disparidades sugerem que o perfil ocupacional do usuário influencia a qualidade das recomendações recebidas.

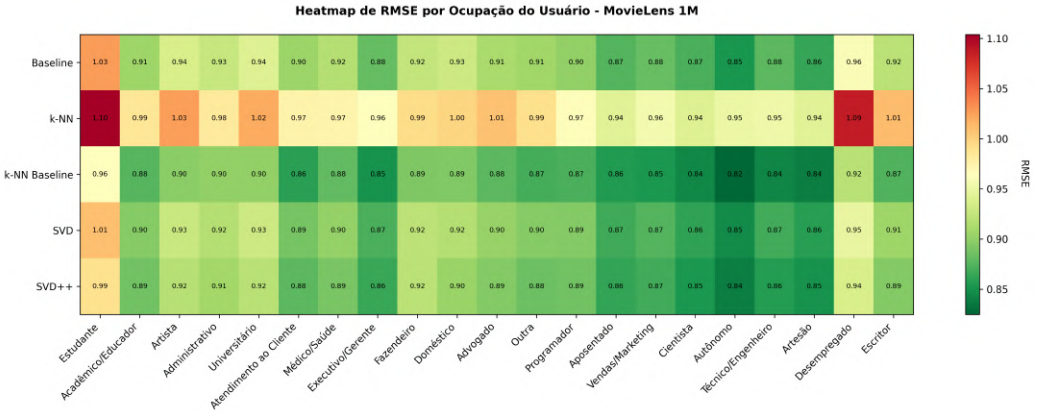

Figura 11: Acurácia RMSE por ocupação profissional do usuário.

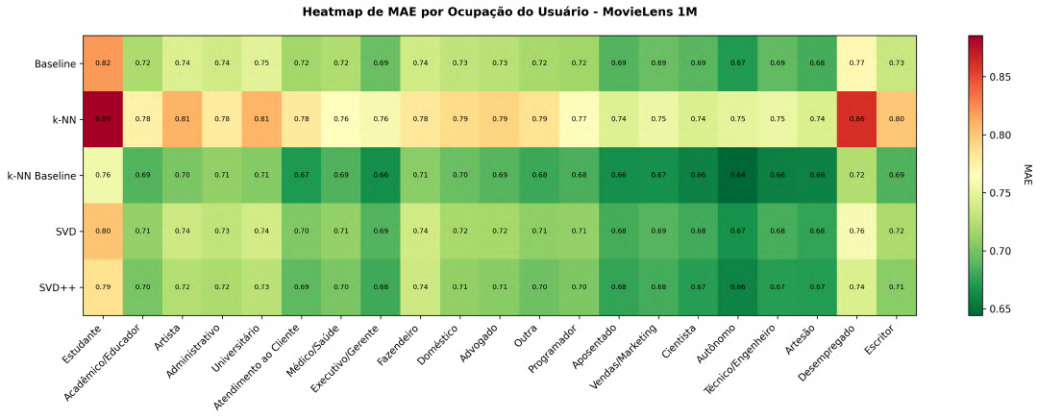

Figura 12: Acurácia MAE por ocupação profissional do usuário.

Foram observadas disparidades moderadas entre profissões, porém significativamente menores que no MovieLens 100k. O k-NN foi o menos enviesado, enquanto o SVD++ apresentou as maiores disparidades, indicando uma melhoria geral na justiça ocupacional com o conjunto de dados expandido.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.1971	0.2294
k-NN	0.1778	0.1900
k-NN Baseline	0.2049	0.2199
SVD	0.2007	0.2410
SVD++	0.2136	0.2580

Tabela 28: Justiça por ocupação profissional do usuário.

6.6.2 Justiça de Item

Justiça por Gênero do Filme

A análise da acurácia por gênero cinematográfico revela disparidades significativas na qualidade das predições. Como mostram os mapas de calor (Figuras 13 e 14), o gênero *Fantasia* apresenta os maiores erros, com MAE entre 0,88–0,98 e RMSE de 1,08–1,18, indicando que os modelos têm dificuldade em captar as preferências por esse gênero. Em contraste, gêneros como *Filme Noir* e *Mistério* têm os menores erros, sugerindo que as preferências nesses domínios são mais consistentes e previsíveis. O modelo *k-NN Baseline* destaca-se como o mais preciso na maioria dos gêneros, enquanto o *k-NN* tem desempenho inferior. Essas diferenças indicam que a natureza do conteúdo influencia diretamente a assertividade do sistema, o que pode levar a uma sub-recomendação sistemática de certos gêneros de nicho.

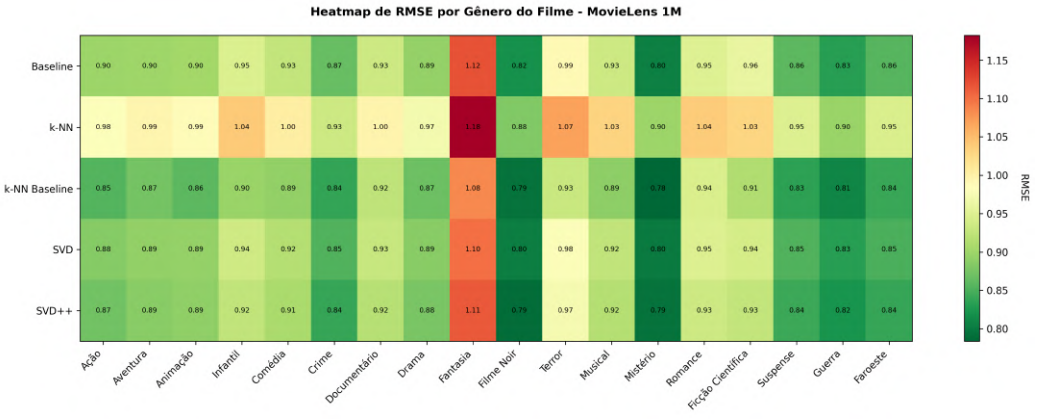

Figura 13: Acurácia RMSE por gênero do filme.

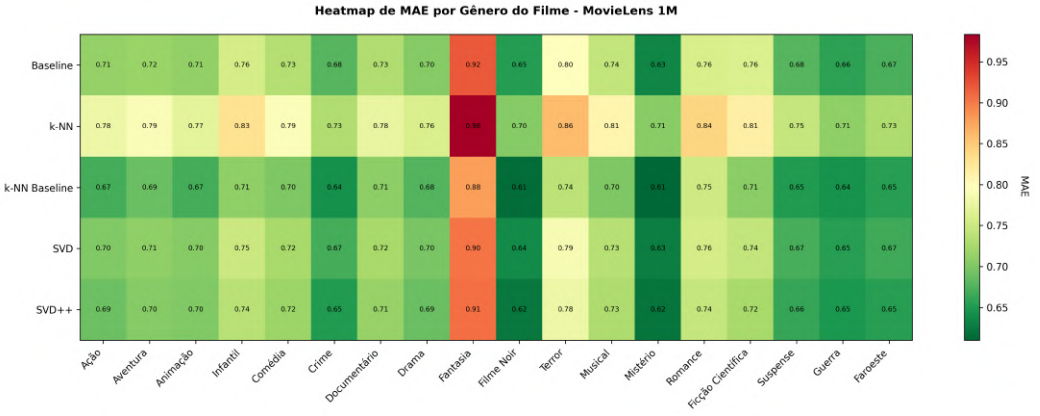

Figura 14: Acurácia MAE por gênero do filme.

Persistem fortes disparidades entre gêneros cinematográficos (Tabela 29), com valores elevados similares aos do MovieLens 100k. O k-NN Baseline foi o menos enviesado, enquanto o k-NN convencional manteve os maiores vieses, indicando que o problema de justiça entre gêneros é consistente dentre diferentes tamanhos de conjuntos de dados.

Modelo	Demographic Parity	Equalized Odds
Baseline	0.6568	0.7137
k-NN	0.7845	0.8279
k-NN Baseline	0.6209	0.6314
SVD	0.6311	0.6864
SVD++	0.6044	0.6454

Tabela 29: Justiça por gênero do filme.

7 Conclusões

Este trabalho realizou uma avaliação extensiva de algoritmos de Filtragem Colaborativa (Baseline, k-NN e SVD), utilizando a biblioteca *Surprise*, aplicados a quatro conjuntos de dados de domínios distintos: cinema, literatura, música e animação japonesa. O estudo contrastou métricas clássicas de acurácia (RMSE, MAE) com métricas de justiça de grupo (Paridade Demográfica e *Equalized Odds*) fornecidas pela biblioteca *Fairlearn*.

A análise dos resultados revelou um padrão consistente e importante: **há uma discrepância notável entre a justiça para o usuário e a justiça para o item.**

Observou-se que os algoritmos tendem a ser relativamente justos ao tratar grupos de usuário. As disparidades baseadas em idade ou gênero do usuário foram, em sua maioria, baixas. Isso sugere que os modelos conseguem generalizar bem as preferências através de diferentes demografias.

Por outro lado, a avaliação revelou graves problemas de justiça na recomendação de itens ou produtos. Grupos de itens definidos por gênero (ex: filmes de terror, músicas clássicas) ou origem (ex: livros de países fora do eixo comercial dominante) sofreram com taxas de erro desproporcionais e menor exposição.

A implicação prática é clara: a avaliação tradicional baseada apenas em métricas de acurácia é insuficiente para garantir sistemas de recomendação éticos e socialmente responsáveis. A persistência destes vieses dentre diferentes domínios e tamanhos de conjuntos de dados indica que intervenções proativas são necessárias, seja através de pré-processamento dos dados, incorporação de restrições focadas em justiça diretamente nos algoritmos, ou estratégias de pós-processamento. O desafio que se coloca é desenvolver sistemas que não apenas prevejam preferências com precisão, mas que também promovam ecossistemas de conteúdo diversos e inclusivos.

Referências

[1] Hug, N. (2024). *Surprise Python Scikit for recommender systems.* Disponível em: `https://surpriselib.com/`. Acessado em: 05 de setembro de 2025.

[2] Costa, A. F. (2023). *Datasets for Recommender Systems* [Repositório do GitHub]. Disponível em: `https://github.com/caserec/Datasets-for-Recommender-Systems`. Acessado em: 05 de setembro de 2025.

[3] Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* International Joint Conference on Artificial Intelligence (IJCAI), 14(2), 1137-1145.

[4] Anelli, V. W., Bellogin, A., Ferrara, A., Malitesta, D., Merra, F. A., Pomo, C., Donini, F. M., & Di Noia, T. (2021). *Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation.* 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), pp. 2405–2414. Disponível em: `http://dx.doi.org/10.1145/3404835.3463245`.