UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

JOÃO PAULO PRATA COSTA

INVESTIGANDO RECOMENDAÇÃO HÍBRIDA PONDERADA EM RELAÇÃO A FAIRNESS E SENSIBILIDADE AO RISCO

Ouro Preto, MG
2026

JOÃO PAULO PRATA COSTA

INVESTIGANDO RECOMENDAÇÃO HÍBRIDA PONDERADA EM RELAÇÃO A FAIRNESS E SENSIBILIDADE AO RISCO

Monografia II apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Reinaldo Silva Fortes

Ouro Preto, MG
2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

João Paulo Prata Costa

Investigando Recomendação Híbrida Ponderada em Relação a Fairness e Sensibilidade ao Risco

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 27 de Fevereiro de 2026

Membros da banca

Reinaldo Silva Fortes (Orientador) - Doutor - Universidade Federal de Ouro Preto
Anderson Almeida Ferreira (Examinador) - Doutor - Universidade Federal de Ouro Preto
Daniel José Chaves Ferreira (Examinador) - Bacharel - PPGCC / UFOP

Reinaldo Silva Fortes, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 27/02/2026

Documento assinado eletronicamente por **Reinaldo Silva Fortes**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 03/03/2026, às 09:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1062467** e o código CRC **2A121DAA**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Agradecimentos

Agradeço aos meus pais por apoiarem sempre minhas decisões, aos amigos de Divinópolis pela amizade e apoio nos momentos difíceis da jornada, a todos os professores do DECOM pelos ensinamentos, ao meu orientador Reinaldo que me guiou neste trabalho, e por fim, aos companheiros da República Território Xavante, por fazerem minha passagem pela UFOP e Ouro Preto inesquecível.

Resumo

Sistemas de recomendação são ferramentas que ajudam os usuários a encontrar itens relevantes dentre muitas opções. Eles são baseados em algoritmos que analisam os dados do usuário, como histórico de compras, avaliações e comportamentos de navegação, para sugerir produtos ou serviços relevantes e personalizados para cada pessoa. Esses sistemas são amplamente utilizados em *e-commerces*, serviços de *streaming* de vídeo e de música, redes sociais, entre outros. Um dos principais desafios enfrentados pelos sistemas de recomendação é atender a diversas métricas de avaliação, como precisão, diversidade e novidade das recomendações, pois essas métricas podem ser conflitantes entre si. Outro desafio enfrentado no desenvolvimento de sistemas de recomendação é lidar com a variabilidade dos resultados, isto é, garantir que sejam consistentes e precisos em diversos cenários. Neste trabalho, são abordados dois conceitos relacionados à variabilidade: **justiça** e **sensibilidade ao risco**. Justiça nas recomendações significa que os algoritmos devem ser projetados para evitar preconceitos e discriminações, levando em consideração fatores como gênero, raça, idade, entre outros, para garantir que as recomendações tenham uma qualidade equiparável para todos os usuários. Já a sensibilidade ao risco está relacionada à capacidade do sistema de reduzir a probabilidade de resultados insatisfatórios. Para lidar com a complexidade desses múltiplos objetivos, a hibridização de algoritmos, que consiste na combinação de diferentes modelos de recomendação clássicos para aliar seus pontos fortes e atenuar suas fraquezas individuais, surge como uma abordagem relevante. Ao integrar múltiplos vieses indutivos, levanta-se a hipótese teórica de que sistemas híbridos não apenas melhorem a precisão global, mas também possam suavizar disparidades entre diferentes perfis e conferir maior robustez contra falhas extremas. Sendo assim, neste trabalho, investiga-se empírica e estatisticamente como estratégias de hibridização impactam as avaliações desses critérios. O objetivo principal é realizar experimentos para avaliar se técnicas de hibridização contribuem para recomendações mais justas e menos sensíveis ao risco. Os resultados obtidos por meio da validação temporal demonstram que, paradoxalmente, a hibridização baseada em regressão atuou como um nivelador genérico, reduzindo a personalização, amplificando as disparidades de desempenho entre os grupos de usuários e aumentando a sensibilidade ao risco global do sistema. **Palavras-chave**: Sistemas de Recomendação. Justiça Algorítmica. Sensibilidade ao Risco. Hibridização. Filtragem Colaborativa.

Abstract

Recommender systems are tools that help users find relevant items among many options. They are based on algorithms that analyze user data, such as purchase history, ratings, and browsing behavior, to suggest relevant and personalized products or services for each individual. These systems are widely used in e-commerce, video and music streaming services, social networks, among others. One of the main challenges faced by recommender systems is meeting various evaluation metrics, such as accuracy, diversity, and novelty, since these may conflict with each other. Another challenge is dealing with variability in results, that is, ensuring that the results are consistent and accurate in various scenarios. In this work, two concepts related to variability are addressed: **fairness** and **risk-sensitiveness**. Fairness in recommendations means that algorithms should be designed to avoid biases and discrimination, taking into account factors such as gender, race, and age, to ensure that recommendations have comparable quality for all users. Risk-sensitiveness, on the other hand, is related to the system's ability to minimize the probability of unsatisfactory results. To deal with the complexity of these multiple objectives, algorithm hybridization — which consists of combining different classic recommendation models to ally their strengths and mitigate their individual weaknesses — emerges as a relevant approach. By integrating multiple inductive biases, the theoretical hypothesis is raised that hybrid systems not only improve global accuracy but may also smooth out disparities between different profiles and provide greater robustness against extreme failures. Therefore, this work empirically and statistically investigates how hybridization strategies impact the evaluations of these criteria. The main objective is to conduct experiments to answer whether hybridization techniques contribute to fairer and less risk-sensitive results in recommendations. The results obtained through temporal validation demonstrate that, paradoxically, regression-based hybridization acted as a generic leveler, reducing personalization, amplifying performance disparities between user groups, and increasing the overall risk sensitivity of the system.

Keywords: Recommender Systems. Algorithmic Fairness. Risk-Sensitiveness. Hybridization. Collaborative Filtering.

Lista de Tabelas

Lista de Abreviaturas e Siglas

SR	Sistema de Recomendação
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
CF	Filtragem Colaborativa - do inglês *Colaborative Filtering*
FH	Filtragem Hibrida
CB	Filtragem Baseada em Conteúdo - do inglês *Content Based Filtering*
DP	*Demographic Parity*
EO	*Equal Opportunity*

Sumário

1 Introdução

Atualmente, o mundo digital oferece abundantemente conteúdo, como músicas, filmes, séries, produtos físicos e vídeos, o que torna as possibilidades de escolha dos usuários muito abrangentes. Esses itens ficam disponíveis em plataformas, entre elas, serviços de *streaming*, redes sociais e *e-commerces*, lugares onde a experiência e satisfação do usuário são importantes para que ele continue utilizando os serviços. É nesse contexto que os Sistemas de Recomendação (SRs) ganham destaque, pois auxiliam os usuários na filtragem de conteúdo relevante (MA et al., 2011).

Os SRs traçam o perfil do usuário analisando um conjunto amplo de dados, que vai além do histórico de consumo de conteúdo. Eles englobam o comportamento de navegação, o histórico de compras, avaliações explícitas (como notas e reviews) e, frequentemente, características demográficas e contextuais. Por exemplo, em uma plataforma de *streaming* como a Netflix, a frequência com que uma pessoa assiste a filmes de ação pode ser usada para inferir o interesse dela em conteúdos desse gênero. Com o perfil do usuário então traçado, os SRs fazem recomendações de itens utilizando técnicas como a Filtragem Colaborativa, a Baseada em Conteúdo ou a Híbrida.

A Filtragem Colaborativa (FC), inicialmente proposta por Goldberg et al. (1992), parte do princípio de que usuários que consumiram o mesmo conteúdo tendem a ter preferências semelhantes. Para exemplificar, imagine um cenário simplificado onde um usuário A assista aos filmes *Forrest Gump* e *Star Wars*, e um usuário B assista aos filmes *Star Wars* e Rei Leão. Apesar de *Forrest Gump* e Rei Leão serem filmes de gêneros totalmente distintos, a filtragem colaborativa infere que o filme Rei Leão pode ser relevante para o usuário A, pois o comportamento do usuário B demonstrou que pessoas que assistem a *Star Wars* também podem gostar de assistir ao Rei Leão.

Já a Filtragem Baseada em Conteúdo (CB) recomenda itens que compartilham atributos semelhantes aos já consumidos pelo usuário. Utilizando o exemplo de filmes citado anteriormente, o usuário A, que assistiu a *Star Wars*, poderia receber uma lista de recomendações com obras como *Star Trek* e *Transformers*, pois os três filmes possuem gêneros semelhantes (ação, aventura e ficção científica). Note que, nesse caso, a recomendação não depende da avaliação de itens por outros usuários; em contrapartida, os atributos dos filmes são relevantes.

Entretanto, ambas as abordagens isoladas apresentam desafios. A FC sofre com o problema do *Cold Start* (dificuldade em lidar com novos usuários ou itens sem histórico) e a esparsidade dos dados. Por outro lado, a CB tende à superespecialização, recomendando sempre itens muito semelhantes aos já vistos, limitando a descoberta de novidades (serendipidade) e dependendo da qualidade dos metadados descritivos.

Por fim, a Filtragem Híbrida (FH) combina diferentes algoritmos de FC e CB, de modo

que cada um possa complementar as limitações dos outros. Dessa forma, a FH consegue produzir recomendações mais confiáveis, com redução na variabilidade dos resultados e mitigação dos problemas de partida a frio e de superespecialização.

Além dos aspectos técnicos de precisão, ao gerar uma recomendação, deve-se considerar a ética do algoritmo, desconsiderando atributos protegidos do usuário. Hardt, Price e Srebro (2016) abordam o risco de considerar características como cor, raça, sexo e idade, o que pode resultar em um algoritmo discriminatório e que se distancia do ideal de justiça. O conceito de justiça está relacionado à variabilidade dos resultados da recomendação e refere-se à capacidade que um SR tem de tratar entidades similares de forma não discriminatória (PITOURA; STEFANIDIS; KOUTRIKA, 2021).

Outro critério fundamental, relacionado à variabilidade dos resultados, é a Sensibilidade ao Risco. Em vez de otimizar apenas a média global de acertos, a sensibilidade ao risco foca na capacidade do sistema de minimizar a probabilidade de gerar resultados severamente insatisfatórios. Apresentado inicialmente na literatura por Wang, Bennett e Collins-Thompson (2012) no contexto da Recuperação de Informação, esse conceito também pode ser aplicado a SRs. Como demonstrado por Fortes (2022), a incorporação explícita de medidas sensíveis ao risco contribui diretamente para a elaboração de recomendações mais robustas. Nesse contexto, a robustez traduz a capacidade do algoritmo de operar de forma estável em diferentes situações, equilibrando múltiplas métricas e garantindo que o sistema não penalize grupos específicos de usuários.

Observa-se, portanto, um claro ponto de convergência teórica entre esses três conceitos: enquanto a hibridização é adotada com o intuito de estabilizar o sistema e reduzir as falhas dos algoritmos clássicos, a justiça e a sensibilidade ao risco atuam como métricas fundamentais para avaliar os impactos éticos e de robustez gerados por essa variabilidade. Diante da hipótese teórica de que a combinação de múltiplos algoritmos possa não apenas melhorar a precisão global, mas também atenuar disparidades e evitar falhas extremas, emergem os questionamentos centrais investigados neste trabalho:

- Técnicas de hibridização contribuem para resultados mais justos na recomendação?
- Técnicas de hibridização contribuem para resultados menos sensíveis ao risco?

As respostas a tais indagações poderão contribuir para o estudo de melhorias das recomendações, dando ênfase a *Fairness* e a *Risk-Sensitiveness*.

O restante deste capítulo é organizado da seguinte forma. A justificativa para explorar a justiça nas recomendações é apresentada na Seção 1.1. A Seção 1.2 apresenta os objetivos do trabalho e a Seção 1.3, a organização deste trabalho.

1.1 Justificativa

Os Sistemas de Recomendação (SRs) têm se mostrado uma importante ferramenta do mundo tecnológico, estando presentes nas mais diversas plataformas de *e-commerce*, redes sociais e serviços de *streaming*. Tradicionalmente, o sucesso desses sistemas tem sido medido pela precisão média global. No entanto, focar exclusivamente na média pode mascarar problemas graves decorrentes da variabilidade dos resultados. Nesse contexto, buscar melhorias em métricas como *Fairness* (justiça) e *Risk-Sensitiveness* (sensibilidade ao risco), que são o foco deste trabalho, é de suma importância, pois um algoritmo descalibrado pode acarretar sérias consequências éticas e comerciais.

A urgência de discutir a justiça algorítmica torna-se evidente ao observarmos os danos reais causados por sistemas enviesados em áreas críticas. Um exemplo emblemático é o algoritmo COMPAS (Perfil de Gerenciamento Corretivo de Infratores para Sanções Alternativas), projetado para estimar o risco de reincidência criminal. A classificação gerada pelo sistema julgava pessoas negras como fortes candidatas a voltarem a cometer crimes, enquanto indivíduos brancos com históricos semelhantes eram classificados como de baixo risco. Ao final do estudo, comprovou-se que indivíduos negros com alta classificação de risco frequentemente não cometiam novos crimes, enquanto brancos classificados como de baixo risco voltavam a delinquir (ANGWIN et al., 2016).

Embora o COMPAS atue na esfera penal, a essência desse problema de viés discriminatório e otimização enviesada manifesta-se diretamente nos SRs. No contexto da recomendação, um algoritmo injusto pode marginalizar grupos específicos de consumidores. Por exemplo, ao ser otimizado apenas para agradar o perfil majoritário do conjunto de dados, o modelo pode gerar recomendações de alta qualidade para usuários superativos ou pertencentes a um gênero predominante, enquanto entrega resultados genéricos e de baixa precisão para usuários menos ativos ou de grupos demográficos minoritários. Na prática, isso cria uma experiência excludente e perpetua desigualdades no acesso ao conteúdo.

De maneira complementar, a sensibilidade ao risco avalia a robustez do sistema diante de falhas severas. Um SR altamente sensível ao risco é aquele que, embora acerte na média, erra de forma drástica em uma parcela dos usuários. Recomendações severamente descalibradas podem quebrar a confiança do indivíduo, resultando em profunda insatisfação e no abandono da plataforma.

Desse modo, evidencia-se a importância de estudos empíricos como o aqui apresentado. Investigar se escolhas arquiteturais comuns na indústria, como a hibridização, atuam como mecanismos de calibração ou se amplificam os vieses e riscos inerentes aos dados é fundamental para a evolução técnica e ética dos SRs e para a contribuição de trabalhos futuros na área.

1.2 Objetivos

Neste trabalho, o objetivo principal consiste em investigar o impacto de técnicas de hibridização na qualidade dos Sistemas de Recomendação, avaliando-as especificamente sob as óticas de justiça (*Fairness*) e sensibilidade ao risco (*Risk Sensitiveness*). Adicionalmente, ao comparar os resultados obtidos nessas duas frentes, busca-se analisar se existe uma relação de dependência ou correlação entre os conceitos — isto é, investigar se a variação na sensibilidade ao risco de um algoritmo reflete, de alguma forma, em um impacto direto na justiça das recomendações e vice-versa.

Para o alcance deste propósito, os objetivos específicos são definidos como:

- ***Fairness***: explorar se a hibridização através de técnicas de regressão contribui para uma recomendação mais justa.
- ***Risk Sensitiveness***: explorar se a hibridização através de técnicas de regressão contribui para uma recomendação menos sensível ao risco.

1.3 Organização do Trabalho

Este documento está organizado da seguinte forma. O Capítulo 2 contém os fundamentos e trabalhos relacionados. O Capítulo 3 apresenta os recursos e a metodologia utilizada. O Capítulo 4 apresenta resultados. Por último, o Capítulo 5 contém a conclusão e discussões sobre trabalhos futuros.

2 Revisão Bibliográfica

Neste capítulo, serão apresentadas técnicas de filtragem utilizadas nos SR, além de métricas de avaliação de predições. Também são apresentados os conceitos de justiça e sensibilidade ao risco. A Seção 2.1, introduz o leitor no contexto SR, apresentando técnicas de filtragem e métricas de avaliação. A Seção 2.2 apresenta o conceito de justiça, e a Seção 2.3, o de Sensibilidade ao Risco. Por último, a Seção 2.4 apresenta os trabalhos relacionados.

Sistemas de recomendação são algoritmos que analisam dados de usuários e itens para gerar recomendações personalizadas, com o objetivo de aumentar a satisfação e o engajamento do usuário. Eles podem ser utilizados em diversas áreas, como comércio eletrônico, entretenimento, redes sociais, entre outras. Nas subseções seguintes, serão apresentadas diferentes estratégias de recomendação além de métricas usadas para avaliar a qualidade das predições.

2.1 Sistemas de Recomendação

Diante da vasta quantidade de informações disponíveis digitalmente, os Sistemas de Recomendação (SRs) surgem como ferramentas essenciais para contornar a sobrecarga de informação. Seu objetivo principal é filtrar grandes volumes de dados e sugerir itens relevantes e personalizados para cada usuário (como filmes, músicas ou produtos), baseando-se na previsão de interesse a partir de históricos de interações e padrões de consumo.

Para alcançar esse objetivo, a literatura estabelece diferentes abordagens algorítmicas. As subseções a seguir detalham as técnicas clássicas — com foco na filtragem colaborativa (Subseção 2.1.1), na filtragem baseada em conteúdo (Subseção 2.1.2) e na hibridização de métodos (Subseção 2.1.3) — e apresentam as métricas matemáticas utilizadas para avaliar o desempenho desses sistemas (Subseção 2.1.5).

2.1.1 Filtragem colaborativa

SRs baseados em filtragem colaborativa fazem a recomendação a partir da similaridade entre itens e usuários. Essa filtragem parte do principio de que usuários que já manifestaram gostos semelhantes por itens tendem a a manter essa similaridade de preferências. Para ilustrar essa ideia, considere um SR que fornece sugestões de séries aos seus usuários. Suponha que existam quatro usuários e quatro séries no sistema, e que cada usuário tenha avaliado alguma dessas séries, como exibido na matriz de ratings a seguir:

A partir dessas informações, é possível utilizar alguma medida para calcular a similaridade entre usuários, a fim de encontrar os vizinhos mais próximos de cada um, como Correlação de

Tabela 2.1 – Tabela de avaliações de séries

	Série 1	**Série 2**	**Série 3**	**Série 4**
Usuário 1	5	4	-	-
Usuário 2	-	5	4	3
Usuário 3	4	-	3	5
Usuário 4	3	-	-	4

Pearson, Índice de Dice e Similaridade de Cosseno. Para esse exemplo, considere a similaridade de cosseno, que é dada por:

$$sim(A,B) = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}|| \cdot ||\vec{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.1}$$

Onde, $\vec{A}$ e $\vec{B}$ são os vetores de *ratings* para cada usuário no cálculo de similaridade, n é a quantidade de termos nos vetores e A_i e B_i são as coordenadas dos vetores. A aplicação da fórmula resulta em um número entre -1 e 1, de forma que um resultado igual a 1 indicaria que os vetores são exatamente iguais, e igual a -1 que são totalmente opostos. Para este cálculo, é necessário que os usuários comparados possuam avaliações em comum para pelo menos dois itens. Considerando os usuários 3 e 4, e as séries 1 e 4, resulta-se em:

$$\vec{A}.\vec{B} = (4 * 3) + (5 * 4) = 12 + 20 = 32 \tag{2.2}$$

$$||\vec{A}|| = \sqrt{4^2 + 5^2} = \sqrt{16 + 25} = \sqrt{41} \tag{2.3}$$

$$||\vec{B}|| = \sqrt{3^2 + 4^2} = \sqrt{9 + 16} = \sqrt{25} \tag{2.4}$$

$$sim(A,B) = 32/(\sqrt{41} * \sqrt{25}) = 32/\sqrt{1025} = 0.9949 \tag{2.5}$$

A partir da similaridade é possível aplicar algoritmos , como o *K-Nearest Neighbors*, que utiliza a média dos *ratings* fornecidos pelos K vizinhos mais próximos de um usuário U, sobre alguns itens, para inferir o valor dos *ratings* para o mesmo usuário e os mesmos itens. Note que a FC não depende de informações detalhadas sobre os itens, já que, para fazer a recomendação, ela utiliza a associação do comportamento entre usuários (no exemplo citado, os *ratings*). Perceba também, que ela promove a descoberta de novos gêneros, pois o item recomendado para o usuário A com base nas preferências de um usuário B, podem ser diferentes do que o usuário A está acostumado a consumir. Apesar de não depender dos atributos detalhados dos itens recomendados, a FC enfrenta dificuldades ao lidar com usuários novos no sistema, pois, nesse caso, não é possível identificar características suficientes para calcular a similaridade entre usuários.

Por outro lado, possui vantagens, uma delas, segundo Herlocker et al. (2001), é o seu modelo conceitual de baixa complexidade, que não se preocupa com o conteúdo dos itens, e sim com a avaliação dos usuários. Em contrapartida, o modelo também apresenta algumas desvantagens, como o problema de *Cold Start*, que ocorre quando não há dados suficientes sobre novos itens ou usuários para realizar recomendações. Tal problema pode ser tratado com a

utilização de *tags*, selecionadas pelos usuários, que são utilizadas para identificar similaridade com outros consumidores do sistema (ZHANG et al., 2010).

2.1.2 Filtragem baseada em conteúdo

A filtragem baseada em conteúdo é uma abordagem em SR que se concentra nas características dos itens. Neste método, o perfil do usuário é construído com base nos atributos dos itens que ele já consumiu, permitindo que novos itens semelhantes sejam recomendados. Por exemplo, se um usuário assiste frequentemente a filmes de ação e aventura, o SR pode recomendar conteúdos de gêneros similares.

Sistemas baseados em conteúdo apresentam algumas vantagens em relação aos baseados em filtragem colaborativa, pois as recomendações de um indivíduo não dependem da associação de seu perfil com o de outros usuários. Além disso, itens recentemente adicionados podem ser recomendados de imediato com base nos atributos já conhecidos sobre eles.

Apesar dessas vantagens, a filtragem por conteúdo pode acabar recomendando apenas itens muito semelhantes aos já consumidos pelo usuário, o que vai na contramão de um importante critério de recomendação: a diversidade. Além desse excesso de especialização, outra limitação significativa afeta os usuários recém-cadastrados (desafio amplamente conhecido na literatura como *Cold Start* de usuário). Como o algoritmo depende exclusivamente das interações passadas para traçar as preferências, usuários recentes que ainda não consumiram ou avaliaram itens não possuem um perfil delineado, o que inviabiliza a geração de recomendações personalizadas em um primeiro momento.

2.1.3 Filtragem hibrida

A filtragem híbrida é uma abordagem em sistemas de recomendação que combina duas ou mais técnicas de filtragem, como a filtragem colaborativa e a filtragem baseada em conteúdo, para melhorar a precisão das recomendações. Ela é usada para superar as limitações de cada técnica individual e aproveitar os pontos fortes de cada uma (BURKE, 2007). Burke (2002) descreve diferentes tipos de filtragem hibrida, entre elas:

- Ponderada: Nesse tipo, são utilizados diferentes métodos para gerar uma pontuação referente a relevância do item a ser recomendado, posteriormente, esses pesos passam por uma combinação linear ponderada para obter um resultado otimizado de recomendação.
- Misto: No método misto, as listas de recomendação geradas por diferentes técnicas são apresentadas conjuntamente. Isso pode envolver, por exemplo, a exibição de recomendações baseadas em conteúdo e colaborativas lado a lado, proporcionando maior diversidade aos usuários.

- Cascata: A abordagem de cascata estabelece uma hierarquia explícita entre os recomendadores, de modo que o recomendador de menor prioridade faça ajustes na lista do de maior prioridade. O recomendador secundário atua, principalmente, desempatando os scores gerados na lista do recomendador primário. Dessa forma, a filtragem híbrida em cascata permite aprimorar a qualidade das recomendações, aproveitando as vantagens de diferentes métodos.

A filtragem híbrida, portanto, oferece uma solução mais abrangente e personalizada para sistemas de recomendação, permitindo que os usuários tenham acesso a sugestões mais precisas e diversificadas.

2.1.4 Algoritmos de Recomendação

Nesta seção, apresentam-se as bases teóricas dos algoritmos de recomendação utilizados neste estudo. A relevância econômica destas tecnologias é evidenciada por projeções que estimam que o mercado global de e-commerce ultrapassará 8,1 trilhões de dólares até 2026 (TOMBUS, 2025). A seleção abrange métodos clássicos de vizinhança e fatoração de matrizes, além de modelos de regressão empregados como meta-modelos.

2.1.4.1 Algoritmos baseados em Vizinhança (KNN)

Os algoritmos de K-Vizinhos Próximos (*K-Nearest Neighbors*, KNN) fundamentam-se na premissa de que usuários ou itens com comportamentos similares no passado tendem a manter essa similaridade no futuro (STIJGER, 2025).

- **User-based KNN:** Identifica usuários com perfis de avaliação semelhantes ao usuário alvo, realizando predições baseadas na média ponderada das avaliações desses vizinhos (KOREN; BELL; VOLINSKY, 2009).
- **Item-based KNN:** Foca na similaridade entre os itens. A recomendação ocorre quando um item j possui um histórico de avaliações correlato ao item i. Pesquisas indicam que esta abordagem apresenta maior estabilidade em ambientes industriais (STIJGER, 2025).
- **StochasticItemKNN:** Utilizou-se a versão estocástica do *itemKNN*, que introduz um mecanismo de amostragem probabilística na seleção de vizinhos para gerar variabilidade e permitir cálculos de intervalos de confiança (STIJGER, 2025).

2.1.4.2 Fatoração de Matrizes

Esta classe de algoritmos decompõe a matriz de interações R em matrizes de menor dimensão, representando fatores latentes (KOREN; BELL; VOLINSKY, 2009).

- **SVD e BiasedSVD:** A variant com viés (*Biased*) incorpora parâmetros para capturar o desvio individual de usuários e popularidade de itens. A regra de predição é definida por $\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$ (KOREN; BELL; VOLINSKY, 2009).

- **NMF (*Non-negative Matrix Factorization*):** Diferencia-se por impor que $W, H \geq 0$ na decomposição $V \approx WH$, o que facilita a interpretação aditiva das características (LEE; SEUNG, 1999; YANG; EMURA, 2025).

- **BiasedMF:** Foca na decomposição incorporando termos de viés para capturar regularidades globais, sendo robusto mesmo com reduções significativas no conjunto de treinamento (KOREN; BELL; VOLINSKY, 2009).

2.1.4.3 Algoritmos de Regressão para Hibridização

Na estratégia de hibridização por *stacking*, algoritmos de regressão atuam como meta-modelos. Esta arquitetura pode alcançar coeficientes de determinação R^2 superiores a 0,95 ao combinar predições complementares (CHEN; AL., 2025).

- **Modelos Lineares (Ridge e Bayesian Ridge):** O **Ridge** utiliza regularização L_2 para lidar com a multicolinearidade. O **Bayesian Ridge** introduz uma abordagem probabilística, sendo eficaz para estimar incertezas e evitar o sobreajuste (CHEN; AL., 2025; YANG; EMURA, 2025).

- **Modelos baseados em Árvores (*Random Forest* e *Gradient Boosting*):** O ***Random Forest*** reduz a variância através da média de múltiplas árvores. O ***Gradient Boosting*** constrói árvores sequenciais para corrigir erros residuais anteriores (CHEN; AL., 2025).

- **Modelos Especializados (*Tweedie* e *LinearSVR*):** O ***Tweedie Regressor*** é ideal para dados com inflação de zeros e assimetria positiva, comuns em interações esparsas (YANG; EMURA, 2025). O ***LinearSVR*** busca um hiperplano com margem de erro ϵ aceitável, garantindo robustez a *outliers* (CHEN; AL., 2025).

2.1.5 Avaliação de SRs

No contexto de SRs, algumas questões devem ser levantadas para avaliar o sucesso das recomendações. É importante observar se os usuários realmente consomem as recomendações e se ficam satisfeitos caso a resposta seja positiva. Em outras palavras, é preciso mensurar a acurácia das predições, tarefa que pode ser realizada através de métricas como as seguintes, onde $|E|$ representa o tamanho do conjunto de teste, r_{uj} representa o valor real de avaliação do usuário u, sobre o item j e $\hat{r}_{uj}$ representa o valor predito pelo sistema. Por fim, E representa o conjunto de entradas usadas para avaliação.

- MSE (Mean Squared Error): Tem como objetivo medir a diferença entre o conjunto de recomendações feitas, e o conjunto real. Pode ser calculado através da equação:

$$MSE = \frac{1}{|E|} \sum_{(u,j) \in E} (r_{uj} - \hat{r}_{uj})^2 \tag{2.6}$$

- A RMSE (Root Mean Squared Error), é uma forma mais interpretável de medir a diferença entre o conjunto de predições e o conjunto real, pois basicamente ela é a raiz do MSE, o que resulta em um valor no mesmo formato da variável dependente.

$$RMSE = \sqrt{\frac{1}{|E|} \sum_{(u,j) \in E} (r_{uj} - \hat{r}_{uj})^2} \tag{2.7}$$

- A MAE (Mean Absolute Error), assim como as métricas anteriores, é utilizada para medir o erro no resultado da predição, mas por não elevar o valor dos erros ao quadrado, ela não penaliza tanto os erros maiores.

$$MAE = \frac{1}{|E|} \sum_{(u,j) \in E} |r_{uj} - \hat{r}_{uj}| \tag{2.8}$$

- A NDCG (Normalized Discounted Cumulative Gain) é usada para medir a qualidade da ordenação dos resultados, isto é, se os itens que aparecem no topo da lista de recomendação são realmente os mais relevantes. NDCG pode ser obtida por:

$$NDCG = \frac{DCG}{IDCG} \tag{2.9}$$

  Sendo que, DCG (Discounted Cumulative Gain), mensura a relevância dos itens em cada posição da lista predita, e pode ser calculada por.

$$DCG = \sum_{i=1}^{k} \frac{e_{ui}}{\log_2(i+1)} \tag{2.10}$$

  De modo que k é a quantidade total de itens recomendados e e_{ui} é a relevância do item na posição i para o usuário u. A IDCG (Ideal Discounted Cumulative Gain) seria a aplicação da fórmula de DCG a uma lista ideal das recomendações. A relevância e_{ij} geralmente é obtida através da função exponencial $2^{r_{ij}} - 1$, onde r representa a relevância real do item i para o usuário j obtida através de heurísticas baseadas nos ratings do usuário. Note que NDCG resultará em um número entre 0 e 1, de forma que quanto mais próximo de 1, mais "ideal" será a ordenação da recomendação.

- Precisão é o percentual de itens recomendados que realmente são relevantes para o usuário. A precisão de um conjunto recomendado R em relação a um conjunto de itens relevantes I para um usuário u pode ser dada por:

$$Prec(R_u, I_u) = 100 * \frac{|R_u \bigcap I_u|}{|R_u|} \tag{2.11}$$

- Revocação é o percentual dos itens relevantes que são recomendados. A revocação de um conjunto recomendado R em relação a um conjunto de itens relevantes I para um usuário u pode ser dada por:

$$Rec(R_u, I_u) = 100 * \frac{|R_u \bigcap I_u|}{|I_u|} \tag{2.12}$$

- F_1-meansure, aparece como uma forma de medir precisão e revocação em uma única métrica e pode ser obtida com a seguinte fórmula:

$$F_1(R_u, I_u) = \frac{2Prec(R_u, I_u)Rec(R_u, I_u)}{Prec(R_u, I_u) + Rec(R_u, I_u)} \tag{2.13}$$

2.2 Justiça nas Recomendações

Com a crescente popularidade dos SRs, cada vez mais escolhas baseadas em recomendações influenciam a vida das pessoas. No entanto, estudos recentes, como o de Burke (BURKE; SONBOLI; ORDONEZ-GAUGER, 2018), mostram que os SRs podem gerar recomendações injustas para diferentes indivíduos, caso não sejam adequadamente calibrados. Esse contexto destaca a importância de estabelecer métodos para medir a justiça nas recomendações, uma vez que tais sistemas podem perpetuar e ampliar as desigualdades existentes na sociedade. Além disso, garantir que as recomendações sejam justas não só é ético, mas também pode levar a um aumento da satisfação e da confiança dos usuários, bem como a uma maior adoção e fidelidade ao sistema de recomendação. Por essas razões, a pesquisa na área de justiça nas recomendações tem se tornado cada vez mais relevante e necessária.

As subseções a seguir apresentam alguns conceitos relativos à avaliação de justiça nas recomendações.

2.2.1 Justiça Individual

O conceito de justiça individual em SRs refere-se à igualdade de tratamento entre usuários individuais, ou seja, se a qualidade das recomendações é equiparável para indivíduos com características semelhantes. No contexto da recomendação, essa semelhança é fundamentada no histórico de interações do usuário no sistema (como o volume e o padrão de itens consumidos ou avaliados) ou em suas preferências declaradas. Dessa forma, a justiça individual pressupõe que dois usuários com comportamentos de consumo historicamente semelhantes devem receber níveis equivalentes de precisão, diversidade e utilidade em suas recomendações, evitando que o algoritmo favoreça arbitrariamente um indivíduo em detrimento de outro com perfil análogo.

A literatura apresenta diversas abordagens para mensurar a justiça individual, entre as quais podemos destacar:

- Índice de Gini (FU et al., 2020):

$$Fairness(q) = 1 - \frac{1}{2m \sum_{k=1}^{m} q(u_k)} \sum_{x=1}^{m} \sum_{y=1}^{m} |q(u_x) - q(u_y)| \tag{2.14}$$

Tal que $q(u_k)$ representa a medida de utilidade, isto é, o resultado da aplicação de qualquer métrica de avaliação de SRs, para o usuário u_k, m é o número total de usuários no conjunto de dados e $|q(u_x) - q(u_y)|$ representa o valor absoluto da diferença entre as medidas de utilidade para os usuários. Nessa equação, quanto mais o resultado se aproxima de 1, mais justas são as recomendações.

- Entropia:

  No contexto de sistemas de recomendação, a entropia é usada para avaliar a uniformidade da distribuição dos itens recomendados. Quanto maior a entropia, maior é a diversidade das recomendações, indicando uma distribuição mais equilibrada entre os itens. Por outro lado, uma entropia menor indica que a distribuição é mais concentrada em poucos itens, o que resulta em recomendações menos diversificadas. A entropia pode ser calculada por:

$$Entropy(L) = -\sum_{i \in I} p(i|L) \log p(i|L) \tag{2.15}$$

  De modo que $p(i|L)$ é o valor de probabilidade observado do item i nas listas de recomendação L.

Wang e Wang (2020) apresentam uma abordagem para a justiça individual do lado dos itens em um SR, baseada em *Deep Learning*, e buscam garantir que itens similares recebam uma cobertura semelhante nas recomendações. Ainda que existam estudos explorando justiça individual em diversas áreas, como aprendizado de máquina e recuperação da informação, o foco em SR está na justiça de grupo, apresentada na subseção seguinte.

2.2.2 Justiça de Grupo

A justiça de grupo está relacionada a grupos de usuários, que podem ser diferenciados por atributos como idade, sexo, cor ou condição econômica. Note que, apesar desses atributos raramente serem diretamente expostos pelos usuários, eles podem refletir no conjunto de dados usado para gerar a recomendação. Um exemplo de tal ocasião pode ser observado no *dataset* do *e-commerce Amazon*, que possui menos registros para usuários de condição econômica vulnerável, uma vez que esses possuem o poder de compra limitado e são menos ativos na plataforma. Nesse contexto, a atividade dos usuários se torna um critério de divisão entre grupos favorecidos (mais ativos) e vulneráveis (menos ativos), já que os algoritmos tendem a performar melhor para os usuários sobre os quais se tem mais informaçõesFu et al. (2020) (LI et al., 2021). Também é possível diferenciar grupos a partir de outras características do dataset, como demonstrado em Abdollahpouri et al. (2021) que divide os grupos a partir do interesse de usuários por itens

populares. Já Hardt, Price e Srebro (2016) apresentam métricas para justiça em grupo baseadas nos conceitos de *Demographic Parity* (DP) e *Equal Opportunity* (EO), isso em um contexto de modelos de classificação.

Demographic Parity (DP) verifica se a proporção de usuários de diferentes grupos recebendo avaliações positivas é igual, *Equal Opportunity* (EO) compara a proporção de verdadeiros positivos para cada grupo de usuários. Lima et al. (2022) propõem uma adaptação dessas métricas para o contexto de SRs, de modo que DP se torna uma comparação entre as médias de predições de cada grupo de usuários, e EO resulta em uma comparação entre os erros médios de diferentes grupos de usuários. Os autores apresentam as seguintes equações:

- *Equal Opportunity*:

$$R_e(g) = (\frac{1}{|X_g|} \left|\left|\mathbf{X}g - \hat{\mathbf{X}}g\right|\right|_F^2 - \frac{1}{|X_{\neg g}|} \left|\left|\mathbf{X}\neg g - \hat{\mathbf{X}}\neg g\right|\right|_F^2)^2 \tag{2.16}$$

  Sendo que X_g e $X_{\neg}g$ são matrizes de avaliação reais considerando usuários no grupo g e fora do grupo g, respectivamente, $\hat{X}_g$ e $\hat{X}\neg g$ representam as avaliações preditas pelo sistema para os usuários que estão em g e não estão em g, respectivamente. O termo $||x||_F$ representa a norma de Frobenius da matriz x.

- *Demographic Parity*:

$$R_d(g) = \left| \frac{1}{|\hat{X}_g|} \sum_{\hat{r} \in \hat{X}_g} \hat{r} - \frac{1}{|\hat{X}_{\neg g}|} \sum_{\hat{r} \in \hat{X}_{\neg g}} \hat{r} \right| \tag{2.17}$$

Perceba que essa métrica compara a média das avaliações entre diferentes grupos, de modo que o resultado ideal da equação seria zero.

Outra maneira de calcular a justiça de grupo é através da diferença absoluta (AD). Quanto menor o valor de AD, mais justas são consideradas as recomendações, indicando que a diferença entre as utilidades dos grupos vulneráveis e favorecidos é menor. AD pode ser calculada pela equação:

$$AD = |f(G_0) - f(G_1)| \tag{2.18}$$

de modo que $f(G)$ é obtida através de alguma métrica de avaliação de SRs. Frequentemente são utilizadas as métricas NDCG (2.9) e F-1 (2.13) Fu et al. (2020) Li et al. (2021).

Outra métrica para avaliar se o sistema está tratando diferentes grupos de usuários de forma equitativa é a injustiça absoluta. A injustiça absoluta evidencia as diferenças na magnitude dos erros de estimativa entre os grupos de usuários. Quando a injustiça absoluta é baixa, significa que os erros de predição são equilibrados entre os grupos ou que ambos os grupos têm a mesma direção e magnitude de erro. No entanto, quando a injustiça absoluta é alta, isso indica que as

predições para um grupo são consistentemente superestimadas, enquanto as predições para o outro grupo são consistentemente subestimadas. A injustiça absoluta pode ser calculada por:

$$U_{\text{abs}} = \frac{1}{N}\sum_{j=1}^{N} \|(ED[y_{ui}]_j - ED[r]_j)| - |(EA[\hat{y}_{ui}]_j - EA[r]_j)\| \tag{2.19}$$

tal que, para N itens, $ED[\hat{y}_{ui}]_j$ é a pontuação média prevista para o j-ésimo item pelos usuários desfavorecidos, $EA[\hat{y}_{ui}]_j$ é a pontuação média prevista pelos usuários favorecidos, e $ED[r]_j$ e $EA[r]_j$ são as pontuações médias dos usuários desfavorecidos e favorecidos, respectivamente.

2.3 Sensibilidade ao Risco

Wang, Bennett e Collins-Thompson (2012) relatam o problema de "otimizar pela média" no contexto de Recuperação da Informação. Os autores apresentam um cenário problemático em que os algoritmos concentram suas técnicas apenas na melhoria da eficácia média global, sem considerar a robustez dos modelos. Nesse contexto, a robustez é entendida como a capacidade do sistema de performar de maneira consistente em diferentes situações, como, por exemplo, ao processar consultas mais complexas e ambíguas na busca, ou ao lidar com usuários que possuem comportamentos atípicos e históricos de consumo que fogem do padrão majoritário.

No cenário descrito pelos autores, apesar de os sistemas apresentarem bons resultados na maior parte das consultas, uma pequena parcela dos usuários recebia resultados severamente insatisfatórios. No artigo, apresentou-se um modelo que otimizou tanto a eficácia quanto a robustez, descrevendo *Risk-Sensitiveness* como a capacidade que os sistemas possuem de minimizar a probabilidade de entregar esses resultados insatisfatórios extremos.

Dinçer, Ounis e Macdonald (2014) apresentam equações para avaliar a sensibilidade ao risco, baseadas em diversas linhas de base de risco. Os autores afirmam que utilizar mais de uma base diminui a parcialidade ao mensurar *Risk-Sensitiveness*. Eles propuseram a função Z_{RISK}, que utiliza o teste qui-quadrado para calcular a variabilidade de desempenho em relação a diversos sistemas de recuperação de informação.

$$Z_{RISK}(i) = \left[\sum_{q \in Q^+} z_{iq} + (1+\alpha)\sum_{q \in Q^-} z_{iq}\right] \tag{2.20}$$

Sendo que $z_{iq} = x_{iq} - e_{iq}/\sqrt{e_{iq}}$; $e_{iq} = S_i \times Q_q/N$; e x_{iq} representa o desempenho da consulta q obtida com o sistema i. O elemento i representa um sistema no conjunto avaliado, e o elemento q representa uma consulta no sistema avaliado, de forma que $i \in 1, 2, ..., r$ e $q \in 1, 2, ..., c$, onde r e c representam respectivamente a quantidade de sistemas e consultas avaliadas. O conjunto Q^+ é composto por consultas que possuem diferença positiva entre o método testado e a base de risco, de forma a representar os resultados positivos. Já o conjunto

Q^- possui consultas com diferença negativa, representando as perdas. Por fim, considere que $Si = \sum_{q=1}^{c} x_{iq}$ é o desempenho esperado para as consultas no sistema i, $Qq = \sum_{i=1}^{r} x_{iq}$ é a efetividade do sistema dentro da consulta individual q e $N = r \sum_{i=1}^{c} \sum_{q=1}^{c} x_{iq}$ a soma de todos os elementos.

No mesmo trabalho, Dinçer, Ounis e Macdonald (2014) ressaltam que Z_{RISK} não leva em consideração a média de desempenho dos sistemas, e não permite comparação entre eles. Para contornar tal problema, eles propõem a função G_{RISK}.

$$G_{RISK}(i) = \sqrt{\frac{Si}{c} \times \Phi\left(\frac{Z_{RISK}(i)}{c}\right)} \tag{2.21}$$

tal que Φ representa a distribuição normal padrão acumulada.

Essas bases de cálculo são normalmente utilizadas em trabalhos relacionados à recuperação da informação, como (SOUSA et al., 2019), contudo não se deve desprezar a importância de otimizar medidas sensíveis ao risco no contexto de SRs. Fortes (2022) traz essa abordagem para o campo da recomendação, utilizando as funções Z_{RISK} e G_{RISK} para mensurar sensibilidade ao risco em SRs.

2.4 Trabalhos Relacionados

Wang et al. (2025) propõem uma abordagem inovadora para lidar com a questão de *fairness* em SRs, com foco na disparidade entre usuários ativos e inativos. No trabalho, os autores dividem os usuários em dois grupos com base no número total de interações históricas ($|H_{u_k}|$) registradas no sistema. Usuários cujo número de interações ultrapassa um limiar empírico (τ) são considerados *ativos*, enquanto os demais são classificados como *inativos*.

A principal contribuição do estudo de Wang et al. (2025) é a introdução do método *LLMFda* (*Large Language Model-based Fair Data Augmentation*), que utiliza modelos de linguagem de larga escala (LLMs) para gerar interações sintéticas para usuários inativos. Essa técnica visa reduzir a escassez de dados desse grupo e, consequentemente, diminuir a diferença na qualidade das recomendações entre usuários ativos e inativos. Os experimentos realizados mostram que o uso de LLMs é capaz de melhorar significativamente as métricas de *NDCG@K* e *Precision@K* para o grupo de inativos, contribuindo assim para uma maior equidade no sistema de recomendação.

Em outra vertente, Fortes (2022) apresenta um estudo detalhado que formaliza o conceito de sensibilidade ao risco no contexto de Sistemas de Recomendação. Embora o trabalho do autor também explore a otimização de múltiplos aspectos de qualidade, sua contribuição fundamental para esta pesquisa é a adaptação e aplicação de métricas matemáticas robustas para mensurar a variabilidade e o risco nas recomendações. Para calcular a sensibilidade ao risco, Fortes (2022) adaptou as funções Z_{RISK} e G_{RISK}, aplicando-as da seguinte maneira:

$$Z_{RISK}(x_j) = \sum_{u \in U^+} z(x_j, u) + (1 + \alpha) \sum_{u \in U^-} z(x_j, u) \tag{2.22}$$

$$G_{RISK}(qi(x_j)) = \sqrt{\frac{qi(x_j)}{|U|} \times \Phi\left(\frac{Z_{RISK}(x_j)}{|U|}\right)} \tag{2.23}$$

Nestas formulações, $z(x_j, u) = \frac{qi(x_j,u) - e(x_j,u)}{\sqrt{e(x_j,u)}}$, sendo que $e(x_j, u) = \frac{qi(x_j) \times Tu}{N}$. A variável $qi(x_j, u)$ representa a qualidade da solução (recomendação) x_j para o usuário u em relação a uma métrica de avaliação qi (como NDCG ou F1-score). Por sua vez, $qi(x_j) = \sum_{y=1}^{|U|} qi(x_j, u_y)$ representa o desempenho total da solução para todos os usuários, e $Tu = \sum_{j=1}^{|X|} qi(x_j, u)$ representa a soma de todas as soluções para o usuário u. Os conjuntos U^+ e U^- representam valores de $z(x_j, u)$ positivos e negativos, respectivamente, enquanto N atua como o fator de normalização global. A variável α define o peso (penalidade) que a degradação no desempenho deve receber. Na Equação 2.23, Φ representa a função de distribuição acumulada da normal padrão.

Note que, ao apresentar essas formas rigorosas de mensurar sensibilidade ao risco em SR, Fortes (2022) fornece as ferramentas exatas necessárias para investigar empiricamente a ocorrência de resultados insatisfatórios extremos e sua possível correlação com a falta de justiça (*fairness*) nas predições.

Conectando esses dois conceitos, Rodrigues et al. (2025) propõem a função de perda *RiskLoss*, uma abordagem contínua e diferenciável para otimização sensível ao risco em modelos de *ranking* com redes neurais profundas. A proposta é motivada pelo problema do viés de otimização pela média, que frequentemente mascara resultados muito ruins para usuários ou contextos mais difíceis. Para lidar com isso, os autores desenvolvem uma estratégia de otimização baseada em *multi-dropout*, avaliando e penalizando a variabilidade do desempenho do modelo diretamente durante o treinamento.

Ainda que o trabalho de Rodrigues et al. (2025) não foque exclusivamente em *fairness*, os autores reconhecem a forte relação conceitual entre sensibilidade ao risco e justiça algorítmica. Os experimentos mostram que a *RiskLoss* reduz consistentemente o número de recomendações ruins — especialmente para os usuários mais "difíceis"de serem modelados. Isso sugere fortemente que mitigar riscos extremos de má recomendação é um caminho que também contribui para uma experiência mais justa e equilibrada entre diferentes perfis demográficos ou comportamentais de usuários.

A revisão da literatura evidencia, portanto, que a busca por equidade (*fairness*) e robustez (*risk-sensitivity*) tem sido predominantemente abordada através de modificações complexas nas funções de objetivo matemáticas, alterações nas funções de perda durante o treinamento profundo ou na geração massiva de dados sintéticos.

O presente trabalho se conecta a Wang et al. (2025) ao compartilhar o objetivo de analisar e mitigar disparidades entre usuários altamente ativos e inativos. No entanto, difere

fundamentalmente na abordagem: enquanto Wang foca em *Data Augmentation* via LLMs para criar dados irreais, esta pesquisa investiga se a simples arquitetura de hibridização de algoritmos clássicos é suficiente para reduzir essa disparidade apenas pela combinação de vieses indutivos naturais.

Em relação a Fortes (2022), este trabalho adota diretamente as métricas Z_{RISK} e G_{RISK} utilizadas pelo autor. Contudo, a aplicação aqui se distancia da formulação teórica multiobjetivo; as métricas são utilizadas de forma prática como ferramentas de avaliação para testar a hipótese de que o empilhamento de modelos apresenta maior robustez do que abordagens isoladas.

Por fim, ao contrário de Rodrigues et al. (2025), que propõe alterações nas funções de perda (*Loss Functions*) em redes neurais complexas, este estudo adota uma perspectiva puramente empírica e arquitetural. O objetivo é verificar se o *Stacking* (hibridização ponderada por regressão) — uma técnica amplamente utilizada na indústria apenas para ganho de precisão global — traz, de forma implícita e natural, os mesmos benefícios de justiça e redução de risco buscados por algoritmos mais onerosos, oferecendo assim um diagnóstico fundamental sobre o comportamento real dessas combinações.

3 Metodologia e Desenvolvimento

Neste capítulo, será detalhada a estratégia adotada, bem como os recursos utilizados para alcançar os objetivos estabelecidos neste trabalho.

A Seção 3.1 apresenta uma breve descrição do *dataset* utilizado. A Seção 3.2 detalha os algoritmos de recomendação constituintes e híbridos avaliados, enquanto a Seção 3.3 define as métricas adotadas. A Seção 3.4 descreve o método experimental, contemplando a estratégia de avaliação temporal e o pré-processamento dos dados. Em seguida, a Seção 3.5 explica a dinâmica de execução, separação e treinamento dos modelos. Por fim, a Seção 3.6 detalha os procedimentos práticos adotados para mensurar a equidade (*fairness*) e a sensibilidade ao risco.

3.1 Dataset

Para os experimentos deste trabalho foi utilizado o dataset *MovieLens 1M*[1]. O dataset contém 1.000.209 de avaliações feitas por 6.040 usuários sobre aproximadamente 3.900 filmes. O dataset é dividido em 3 partes, que serão detalhadas nas subseções a seguir.

3.1.1 Ratings

O arquivo de *ratings* armazena o histórico de interações explícitas dos usuários com as obras cinematográficas. Cada registro neste conjunto de dados é composto por quatro atributos principais: o identificador único do usuário (`userId`), o identificador único do filme avaliado (`movieId`), a nota atribuída (`rating`) — que varia em uma escala discreta de 1 a 5 estrelas —, e o carimbo de tempo (`timestamp`), que registra em segundos o momento exato em que a avaliação foi realizada. A Tabela abaixo ilustra o formato tabular em que essas informações estão estruturadas.

userId	**movieId**	**rating**	**timestamp**
1	1193	5	978300760
1	661	3	978302109
1	914	3	978301968
…	…	…	…

Tabela 3.1 – Demonstração tabular do arquivo de *ratings* do *Movielens 1M*.

[1] Disponível em <https://grouplens.org/datasets/movielens/1m/>.

3.1.2 Users

O arquivo de *Users* do *MovieLens 1M* contém informações demográficas dos usuários e está estruturado com os seguintes atributos: identificador único do usuário (*userId*), gênero ("M" para masculino e "F" para feminino), faixa etária, ocupação e código de área (*zip-code*).

3.1.3 Usuários e Itens

O conjunto também disponibiliza informações complementares sobre usuários e filmes, armazenadas nos arquivos `users.dat` e `movies.dat`. O primeiro contém dados demográficos básicos, como gênero, idade e ocupação, enquanto o segundo lista o título dos filmes e seus respectivos gêneros.

3.2 Algoritmos Utilizados

Nos experimentos conduzidos, foram avaliados dois grupos de algoritmos de recomendação: métodos constituintes e métodos híbridos. Essa divisão teve como objetivo permitir uma análise comparativa entre abordagens consolidadas na literatura e modelos derivados de técnicas de aprendizado supervisionado.

3.2.1 Algoritmos Clássicos

O grupo de métodos clássicos foi composto pelos algoritmos *Stochastic itemKNN*, *NMF*, *BiasedSVD* e *BiasedMF*. Esses algoritmos representam abordagens tradicionais de filtragem colaborativa, amplamente utilizadas como referência em trabalhos baseados no conjunto de dados *MovieLens*.

A versão estocástica de *itemKNN* foi empregada neste trabalho, pois introduz um mecanismo de amostragem probabilística na seleção dos vizinhos mais similares, o que gera variabilidade entre execuções, possibilitando o cálculo de intervalos de confiança e análises estatísticas mais consistentes.

Por sua vez, os algoritmos *NMF* (*Non-negative Matrix Factorization*), *BiasedSVD* e *BiasedMF* pertencem à classe de métodos baseados em fatoração de matrizes.

3.2.2 Algoritmos Híbridos

Os resultados híbridos foram gerados a partir de técnicas de regressão supervisionada, empregando os seguintes algoritmos: *BayesianRidge*, *Tweedie*, *Ridge*, *RandomForest*, *Bagging*, *AdaBoost*, *GradientBoosting* e *LinearSVR*.

Nesses modelos, as predições geradas pelos métodos clássicos são utilizadas como variáveis de entrada (*features*) para estimar novas avaliações, permitindo a combinação de diferentes perspectivas de recomendação.

A fim de garantir uma comparação justa e o melhor desempenho possível de cada método, foi realizada uma etapa de otimização dos hiperparâmetros. Isso foi feito através do algoritmo *RandomizedSearchCV* da biblioteca (`scikit-learn`), com validação cruzada interna de 3 dobras (`cv=3`), função objetivo de erro quadrático médio (`scoring = 'neg_mean_squared_error'`) e reprodutibilidade (`random_state = 42`). O número de amostras na busca (`n_iter`) foi fixado em 15, com exceção do *RandomForest* (10 iterações) por custo computacional. Sempre que possível, utilizou-se paralelização (`n_jobs=-1`). Os espaços de busca contemplaram hiperparâmetros centrais de cada família de modelos:

- **BayesianRidge**: `n_iter` $[200, 1000]$, `alpha_1`, `alpha_2`, `lambda_1`, `lambda_2` em distribuições *uniform* de pequena magnitude (ordem de 10^{-7} a 10^{-5}).
- **Ridge**: `alpha` $\in [0.1, 2.0]$, `max_iter` $[500, 2000]$, `solver` $\in \{$`auto, svd, cholesky, lsqr`$\}$.
- **TweedieRegressor**: `power` $\in [1.0, 3.0]$, `alpha` $\in [0.1, 2.0]$, `max_iter` $[500, 1500]$.
- **RandomForestRegressor**: `n_estimators` $[50, 300]$, `max_depth` $[5, 25]$, `min_samples_split` $[2, 10]$, `min_samples_leaf` $[1, 5]$, `max_features` $\in \{$`sqrt, log2, None`$\}$.
- **BaggingRegressor**: `n_estimators` $[10, 100]$, `max_samples` $\in [0.5, 1.5]$, `max_features` $\in [0.5, 1.5]$, `bootstrap` $\in \{$`True, False`$\}$.
- **AdaBoostRegressor**: `n_estimators` $[50, 200]$, `learning_rate` $\in [0.01, 1.01]$, `loss` $\in \{$`linear, square, exponential`$\}$.
- **GradientBoostingRegressor**: `n_estimators` $[50, 300]$, `learning_rate` $\in [0.01, 0.31]$, `max_depth` $[3, 10]$, `subsample` $\in [0.6, 1.6]$, `min_samples_split` $[2, 10]$.
- **LinearSVR**: `epsilon` $\in [0.0, 0.5]$, `C` $\in [0.1, 2.1]$, `max_iter` $[1000, 5000]$, `tol` $\in [10^{-5}, 10^{-3}]$.

Ao final da busca, o melhor estimador (`best_estimator_`) é selecionado e seus hiperparâmetros são registrados em arquivo com carimbo temporal para reprodutibilidade. Em seguida, os modelos otimizados são empregados para gerar predições sobre o conjunto-alvo da janela, e as saídas são persistidas em arquivos tabulares por método e por janela.

3.3 Métricas Utilizadas

Para a avaliação dos métodos foram utilizadas as métricas *F1-score* (Eq. 2.13), *Mean Absolute Error* – MAE (Eq. 2.8), *Root Mean Squared Error* – RMSE (Eq. 2.7) e *Normalized Discounted Cumulative Gain* – NDCG (Eq. 2.9).

Essas métricas foram empregadas como medidas base de qualidade tanto para a análise de sensibilidade ao risco quanto para a avaliação de *fairness*.

No contexto de sensibilidade ao risco, os valores obtidos por cada métrica foram utilizados na construção das matrizes de desempenho submetidas ao cálculo do *GeoRisk*, conforme definido na Seção 2.3. Valores mais elevados de *GeoRisk* indicam maior robustez do algoritmo, refletindo um melhor equilíbrio entre desempenho e estabilidade ao longo das janelas analisadas.

Para a análise de equidade, as mesmas métricas foram aplicadas no cálculo da diferença absoluta entre grupos de usuários, de acordo com a Equação 2.18. Nesse caso, a qualidade em termos de *fairness* aumenta à medida que a diferença absoluta diminui, indicando que os grupos comparados receberam desempenhos mais semelhantes.

Assim, o *GeoRisk* foi adotado como medida de sensibilidade ao risco dos algoritmos, enquanto a diferença absoluta foi utilizada como indicador de *fairness*, permitindo uma avaliação complementar entre robustez e equidade.

3.4 Método Experimental

Nesta seção, são descritos os procedimentos experimentais adotados para a realização dos testes e das análises dos algoritmos de recomendação. O objetivo principal do experimento é avaliar a sensibilidade ao risco e o *fairness* de diferentes abordagens, comparando métodos clássicos de filtragem colaborativa com modelos híbridos baseados em regressão supervisionada.

3.4.1 Estratégia de Avaliação Temporal

A avaliação dos algoritmos de recomendação foi conduzida por meio de uma estratégia baseada em janelas de tempo deslizantes (*sliding windows*) com interseção entre elas. Esse método permite observar a variação de desempenho dos modelos ao longo do tempo, reduzindo a dependência de um único recorte temporal e tornando a análise mais robusta e representativa do comportamento real de um sistema de recomendação em operação contínua.

Nessa abordagem, o conjunto de dados é segmentado em várias janelas de comprimento semelhante, que avançam progressivamente no tempo, com sobreposição parcial entre si. Em cada janela, o modelo é treinado com interações anteriores e avaliado em interações posteriores, preservando sempre a ordem cronológica dos eventos.

Segundo Quadrana, Cremonesi e Jannach (2018), a utilização de uma única divisão

treino–teste em dados com dependência temporal pode gerar resultados enviesados, uma vez que ignora a evolução natural das preferências dos usuários. Os autores recomendam a adoção de múltiplas divisões temporais consecutivas, o que permite uma avaliação mais estável e confiável dos algoritmos.

De modo análogo, Ji et al. (2023) descrevem a aplicação de janelas de tempo móveis para avaliação sequencial, nas quais o modelo é continuamente atualizado com novos dados e avaliado em períodos futuros. Essa metodologia, também conhecida como avaliação pré-sequencial (*prequential evaluation*), tem sido empregada em cenários de *streaming* e recomendação *online*, constituindo uma forma rigorosa de avaliação temporal.

No presente trabalho, adotou-se uma variação desse método, dividindo o conjunto de dados em vinte janelas temporais de 15 meses, com avanço de 1 mês entre janelas consecutivas. Em cada janela, os primeiros 12 meses foram utilizados para treinamento dos modelos, enquanto os 3 meses restantes compuseram o conjunto de teste, sobre o qual as predições foram realizadas. Dessa forma, garantiu-se que as predições fossem sempre realizadas sobre eventos futuros em relação ao treinamento, evitando qualquer forma de vazamento de informação (*data leakage*).

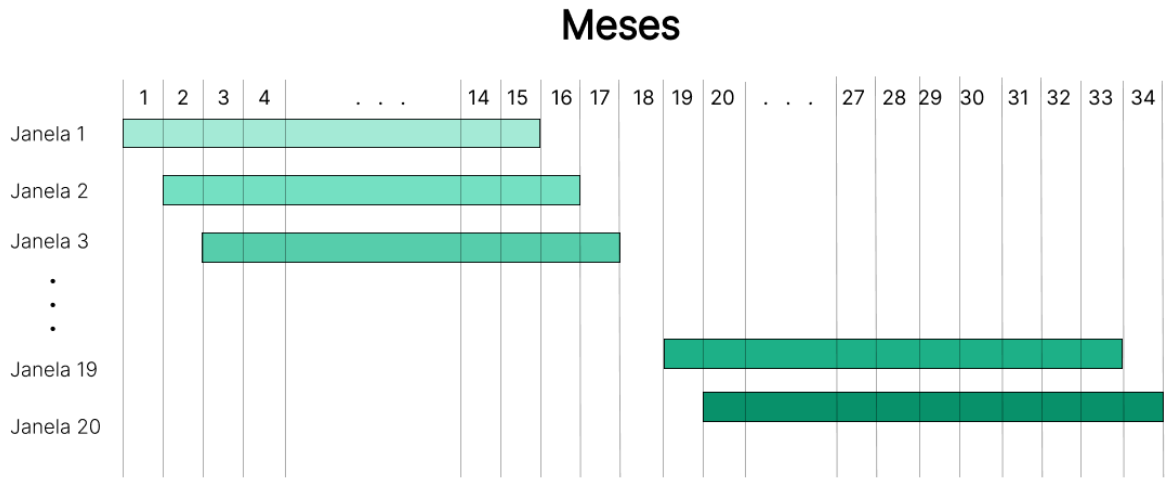

Figura 3.1 – Divisão das janelas ao longo dos meses.

3.4.2 Pré-processamento dos Dados

Durante o pré-processamento, os dados foram padronizados e organizados de modo a permitir a execução de experimentos sob uma perspectiva temporal. O campo `timestamp` foi convertido para o tipo `datetime` e utilizado como referência para a segmentação do conjunto de dados em janelas de tempo. Cada janela temporal, como citado na subseção anterior, representa um intervalo de 15 meses, com avanço de 1 mês entre janelas consecutivas, caracterizando o processo de janelas deslizantes (*sliding windows*).

De forma geral, para cada janela temporal, foram produzidos quatro conjuntos distintos de dados: dois destinados à construção e avaliação dos métodos híbridos; e dois destinados à avaliação dos métodos tradicionais.

O procedimento de divisão seguiu a seguinte lógica:

- **Treino e teste para os métodos híbridos:** Foram extraídos dois subconjuntos menores, correspondentes aos nove primeiros meses (*train*) e aos três meses seguintes (*test*) dentro de cada janela. Esses conjuntos foram utilizados para gerar predições dos métodos tradicionais, que posteriormente serviriam como variáveis de entrada (*features*) para os modelos de regressão empregados nos métodos híbridos.
- **Treino e teste para os métodos tradicionais:** Foram gerados dois subconjuntos adicionais, de maior abrangência temporal, correspondentes aos doze primeiros meses (*train*) e aos três meses seguintes (*test*) de cada janela. Esses conjuntos foram utilizados diretamente para o treinamento e predição dos resultados utilizados para avaliar os métodos citados em 3.2.1.

Figura 3.2 – Divisão de uma janela para gerar predições utilizadas no treino de metodos hibridos.

Figura 3.3 – Divisão de uma janela para gerar predições dos métodos tradicionais.

3.5 Execução dos Algoritmos

Para cada uma das 20 janelas temporais foram realizadas 5 execuções independentes dos algoritmos, totalizando 100 amostras por método. Essa estratégia foi adotada com o objetivo de obter distribuições empíricas dos resultados, permitindo uma análise estatística mais robusta, especialmente no cálculo de intervalos de confiança.

Os algoritmos tradicionais foram executados duas vezes, com finalidades distintas. A primeira execução utilizou o conjunto de treino e teste ilustrado na Seção 3.4.2, sendo destinada à geração dos resultados que serviram como base para o treinamento dos métodos de regressão (modelos híbridos).

A entrada de treino dos métodos híbridos foi composta, para cada instância usuário–item, por:

- Um vetor de predições $(\hat{r}_{BiasedMF}, \hat{r}_{NMF}, \hat{r}_{StochasticitemKNN}, \hat{r}_{BiasedSVD})$;
- A nota real atribuída pelo usuário ao item correspondente (r_{real}).

Dessa forma, os modelos de regressão foram treinados para aprender uma função de combinação das predições geradas pelos métodos tradicionais.

A segunda execução dos métodos tradicionais utilizou o particionamento apresentado na Figura 3.3. Os resultados obtidos nessa etapa foram aqueles efetivamente considerados na avaliação final dos métodos tradicionais. Além disso, as predições geradas nessa execução também foram utilizadas para compor o vetor de *features* empregado pelos métodos híbridos na fase de teste, durante a geração das predições finais.

Dessa forma, o processo experimental garantiu separação adequada entre as fases de treinamento e avaliação, preservando a consistência metodológica da comparação entre métodos tradicionais e híbridos.

3.6 Obtenção de Métricas de Avaliação

Esta seção descreve os procedimentos adotados para a obtenção das métricas utilizadas na avaliação dos métodos de recomendação implementados. O processo de avaliação teve como objetivo analisar o desempenho médio dos algoritmos em termos de equidade (*fairness*) e sensibilidade ao risco, fornecendo o material empírico necessário para responder às questões de pesquisa deste trabalho.

Inicialmente, foram realizados os cálculos de *fairness* com o intuito de avaliar a equidade dos algoritmos em relação a diferentes subgrupos de usuários. Essa análise foi conduzida considerando duas dimensões distintas: (i) gênero dos usuários; e (ii) nível de atividade ao longo do tempo. Em ambos os casos, a *fairness* foi mensurada por meio da diferença absoluta de desempenho entre os grupos, permitindo a identificação de possíveis vieses nos resultados dos algoritmos.

Em seguida, foram obtidas as métricas de sensibilidade ao risco dos métodos de recomendação, com o objetivo de analisar a robustez dos algoritmos diante de variações de desempenho entre usuários e responder às questões centrais deste estudo.

As subseções a seguir detalham a metodologia adotada para o cálculo das métricas de *fairness* e de sensibilidade ao risco. A análise e discussão dos resultados obtidos são apresentadas no capitulo 4, investigando se a hibridização dos métodos contribui para a redução de desigualdades entre diferentes grupos de usuários, sem comprometer o desempenho global do sistema.

3.6.1 Cálculo de *Fairness* por Atividade

Com o objetivo de avaliar a equidade dos métodos de recomendação em relação ao nível de atividade dos usuários, foi conduzida uma análise baseada no comportamento de interação observado ao longo do tempo. Para esse fim, os usuários do conjunto de dados *MovieLens 1M* foram segmentados em subgrupos de acordo com seu grau de atividade em cada janela temporal, utilizando o algoritmo de clusterização não supervisionada *K-Means*.

A clusterização foi realizada de forma independente em cada uma das vinte janelas temporais, considerando exclusivamente informações de atividade dos usuários pertencentes ao conjunto de treino de cada janela, de modo a evitar qualquer forma de vazamento de informação (*data leakage*). O número de clusters foi fixado em $k = 2$, permitindo distinguir usuários menos ativos daqueles mais ativos em cada período analisado.

Essa estratégia permite que o mesmo usuário seja associado a diferentes grupos ao longo do tempo, refletindo variações em seus padrões de interação e garantindo uma avaliação dinâmica da equidade.

Para cada algoritmo e para cada janela temporal, as métricas de desempenho foram calculadas separadamente para os dois grupos de atividade. A medida de *fairness* foi então obtida por meio da diferença absoluta entre os resultados médios dos grupos, conforme definido na Seção 3.3 e formalizado na Equação 2.18.

Foram obtidos quatro resultados de diferença absoluta para cada algoritmo, correspondentes às métricas F1-score, RMSE, NDCG e MAE.

3.6.2 Cálculo de *Fairness* por Gênero

Com o objetivo de avaliar a equidade dos métodos de recomendação em relação ao gênero dos usuários, foi realizada uma análise com base nas informações demográficas disponíveis no conjunto de dados *MovieLens 1M*. O atributo `Gender` permitiu segmentar os usuários em dois subgrupos: (i) usuários do gênero masculino; e (ii) usuários do gênero feminino.

Diferentemente da análise por atividade, a segmentação por gênero é estática ao longo do tempo, pois se trata de uma característica demográfica invariável no conjunto de dados. Ainda assim, a avaliação foi realizada ao longo das vinte janelas temporais, permitindo observar possíveis variações no comportamento dos algoritmos ao longo do tempo.

Para cada algoritmo e para cada janela temporal, as métricas de desempenho foram calculadas separadamente para os dois grupos de gênero. A medida de *fairness* foi então obtida por meio da diferença absoluta entre os resultados médios dos grupos, conforme descrito na Seção 3.3 e formalizado na Equação 2.18.

Assim como na análise por atividade, foram obtidos quatro resultados de diferença absoluta para cada algoritmo, um para cada métrica base considerada: F1-score, RMSE, NDCG

e MAE.

3.6.3 Cálculo de sensibilidade ao risco

Com o objetivo de mensurar a robustez dos métodos em diferentes cenários, a sensibilidade ao risco foi obtida por meio da métrica G_{RISK} (GeoRisk), definida na Equação 2.23.

No contexto deste trabalho, cada *query* q_i corresponde a uma janela temporal, enquanto cada sistema representa um dos algoritmos avaliados. Dessa forma, as matrizes de desempenho foram construídas considerando, em cada linha, os resultados obtidos por todos os métodos em uma determinada janela, e, em cada coluna, o desempenho de um método ao longo das diferentes janelas.

Foram calculados quatro resultados distintos de risco para cada algoritmo, um para cada métrica de qualidade considerada: *F1-score*, RMSE, NDCG e MAE. Em cada caso, os valores da métrica correspondente foram utilizados como base para a construção da matriz submetida ao cálculo do *GeoRisk*.

Assim, o G_{RISK} permitiu avaliar a robustez relativa dos algoritmos ao longo das vinte janelas temporais, sendo que valores mais elevados indicam métodos mais estáveis e menos sensíveis a variações de desempenho.

3.6.4 Teste estatístico

Para garantir a validade e o rigor científico das comparações entre os algoritmos constituintes e os modelos híbridos, os resultados das métricas de qualidade, justiça e sensibilidade ao risco foram submetidos a testes de significância estatística. Primeiramente, aplicou-se a Análise de Variância (ANOVA) de via única (com nível de significância $\alpha = 0.05$) para verificar a existência de diferenças globais entre as médias dos sistemas avaliados ao longo das 100 amostras experimentais (20 janelas temporais $\times$ 5 execuções).

Como a ANOVA indica apenas se há alguma diferença, mas não especifica entre quais modelos, utilizou-se o teste *post-hoc* de *Tukey HSD* (*Honest Significant Difference*) para realizar as comparações par a par (*pairwise*). Esse teste foi escolhido por controlar rigorosamente a taxa de falsos positivos decorrentes de múltiplas comparações, permitindo atestar com 95% de confiança quais algoritmos apresentaram desempenhos estatisticamente equivalentes ou distintos.

4 Avaliação dos resultados

Nesta seção são analisados os resultados obtidos a partir da execução dos métodos clássicos e híbridos ao longo das vinte janelas temporais. A avaliação tem como objetivo investigar se a hibridização contribui para: (i) reduzir as disparidades entre grupos de usuários, tais quais foram mensuradas pela diferença absoluta; e (ii) aumentar a robustez dos sistemas, mensurada pelo GeoRisk.

Todos os resultados apresentados correspondem à média das 100 execuções realizadas por método (20 janelas $\times$ 5 execuções), acompanhadas de desvio padrão e intervalo de confiança de 95%.

A organização desta avaliação está dividida em três partes: a Seção 4.1 aborda a justiça em relação ao gênero dos usuários; a Seção 4.2 analisa a justiça sob a perspectiva do nível de atividade; a Seção 4.3 discute o impacto da hibridização na sensibilidade ao risco e, por fim, a Seção 4.4 analisa os isoladamente os resultados dos diferentes grupos, buscando explicar os comportamentos observados nas seções anteriores a ela nesse capítulo.

4.1 Avaliando *Fairness* por gênero

Nesta seção, são apresentados os resultados do cálculo de *fairness* em relação ao gênero dos usuários. As Figuras 4.1, 4.2, 4.3 e 4.4 exibem o *fairness* para cada uma das quatro métricas-base (F1, NDCG, RMSE e MAE), cujas formulações matemáticas são detalhadas na Seção 2.1.

Após a validação estatística e a aplicação do teste de *Tukey HSD*, evidenciou-se, nos resultados da Figura 4.1, que os algoritmos constituintes não apresentam diferenças significativas entre si. Em contrapartida, apresentam diferença relevante em relação aos algoritmos híbridos, que obtiveram resultados de *fairness* de gênero inferiores (ou seja, maior disparidade). Quanto aos resultados exibidos na Figura 4.2, observou-se que apenas o algoritmo *StochasticItemKNN* se mostrou significativamente superior aos métodos híbridos. De maneira oposta, para as métricas de erro (Figuras 4.3 e 4.4), o teste ANOVA demonstrou que apenas o algoritmo *StochasticItemKNN* foi inferior aos algoritmos de regressão, ou seja, com valores de erro maiores do que os algoritmos híbridos.

De forma geral, os resultados evidenciam que, nos casos em que há diferença estatística, os métodos clássicos mantiveram resultados consistentemente mais justos (maior *fairness*) entre os grupos masculino e feminino ao longo das janelas temporais. Os métodos híbridos, por sua vez, apresentaram disparidades superiores e maior variação entre as execuções. Esse comportamento indica que, no cenário experimental adotado, a hibridização ponderada não contribuiu para a redução das desigualdades de gênero. Assim, **no que se refere à pergunta de pesquisa sobre a**

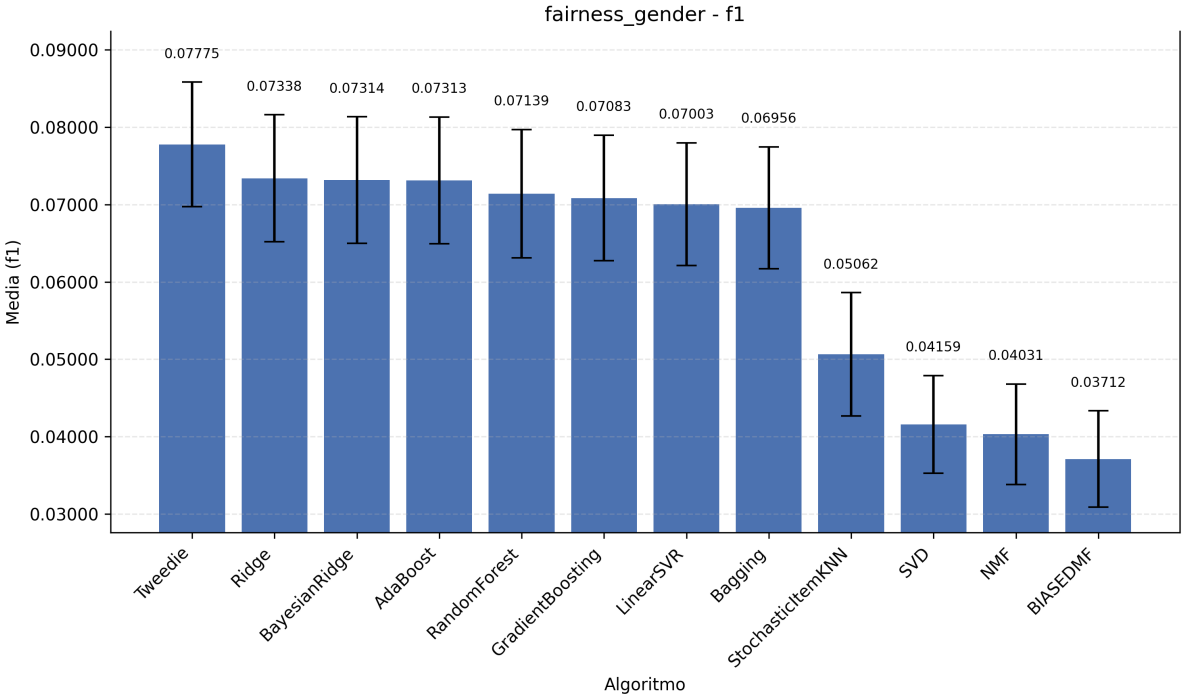

Figura 4.1 – *Fairness* calculado com base em F1, com intervalos de confiança.

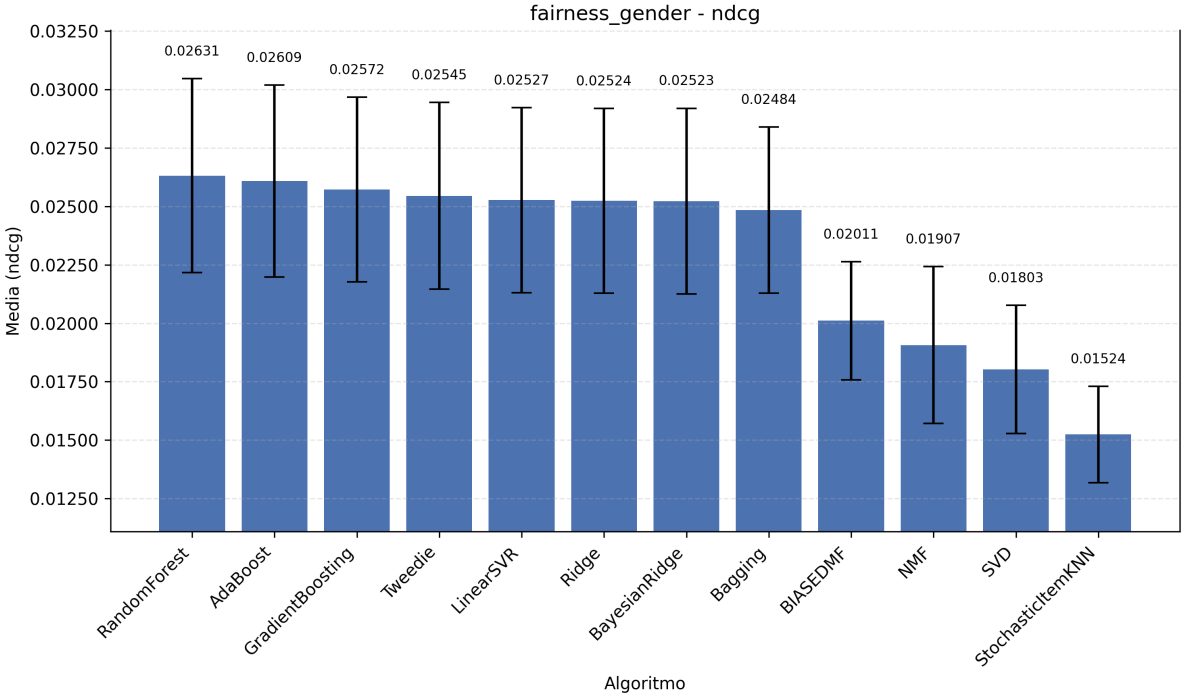

Figura 4.2 – *Fairness* calculado com base em NDCG, com intervalos de confiança.

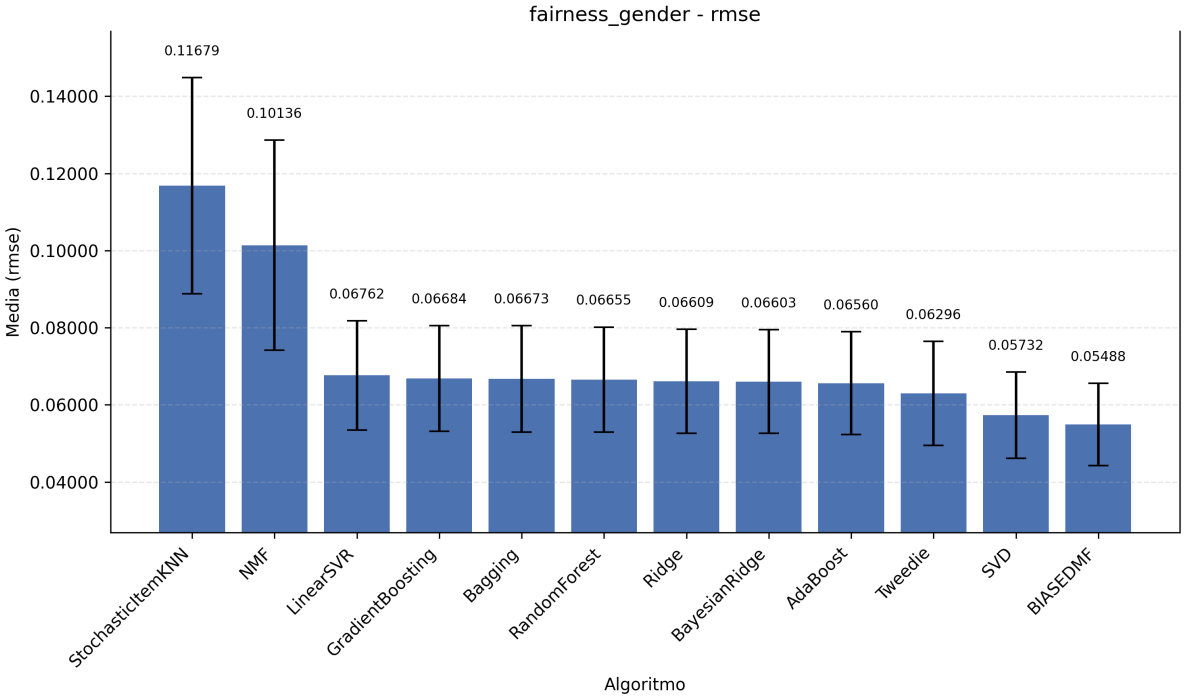

Figura 4.3 – *Fairness* calculado com base em RMSE, com intervalos de confiança.

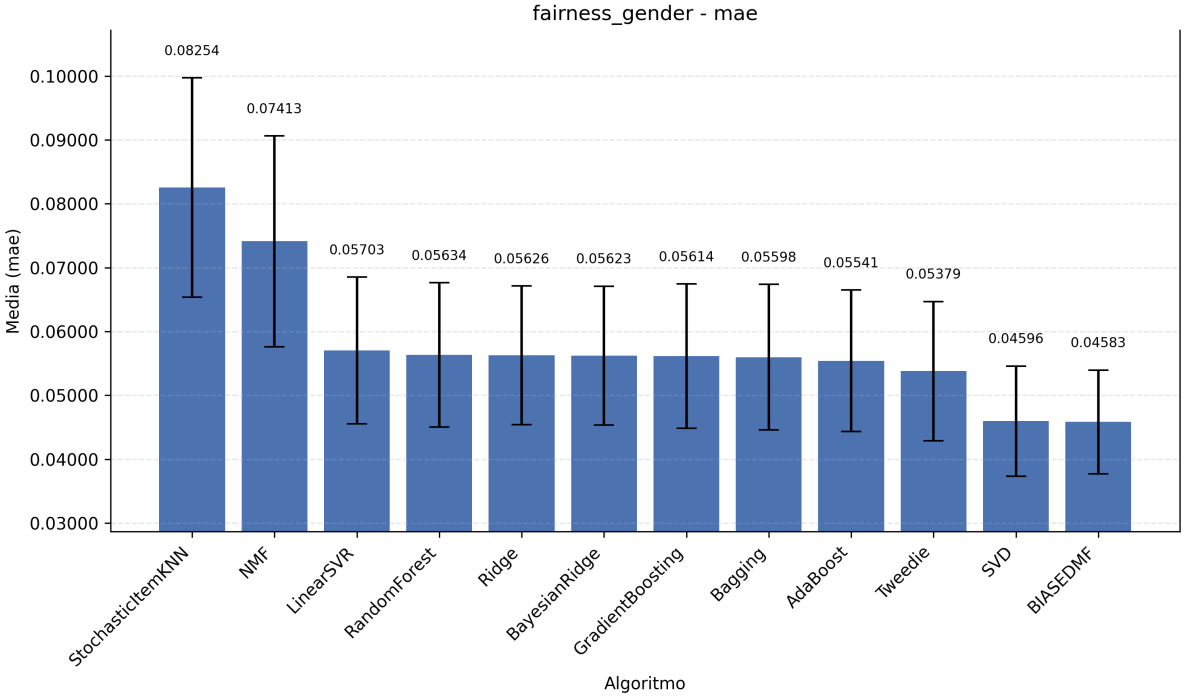

Figura 4.4 – *Fairness* calculado com base em MAE, com intervalos de confiança.

capacidade da hibridização de promover maior justiça nas recomendações, os resultados observados para o critério de gênero não sustentam tal hipótese.

4.2 Avaliando *Fairness* por atividade

Esta seção apresenta os resultados do cálculo de *fairness* em relação ao nível de atividade dos usuários (frequência de interações). As Figuras 4.5, 4.6, 4.7 e 4.8 apresentam os *Fairness* calculados para cada métrica entre os grupos de usuários mais e menos ativos.

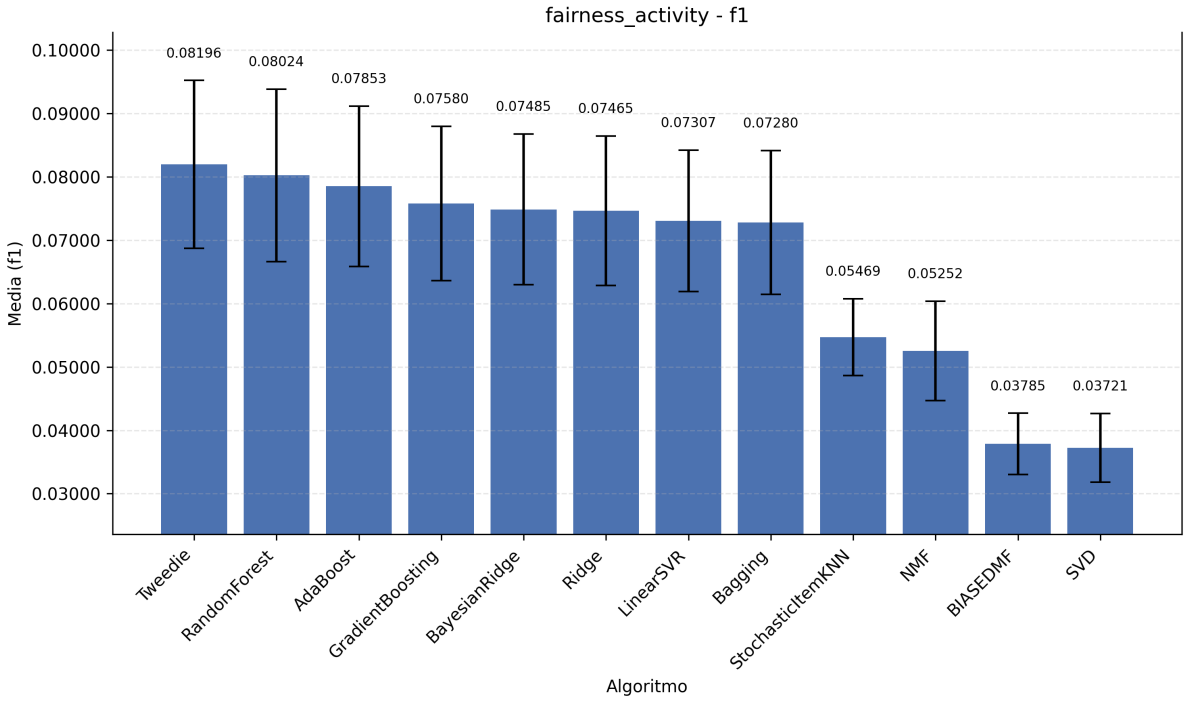

Figura 4.5 – *Fairness* calculado com base em F1, com intervalos de confiança.

Os testes estatísticos revelaram padrões de comportamento distintos para este critério. Na métrica F1 (Figura 4.5), os algoritmos de fatoração (BIASEDMF, NMF e SVD) apresentaram diferenças significativas em relação à quase totalidade dos métodos híbridos. O algoritmo *StochasticItemKNN* apresentou um comportamento atípico nesta métrica, assemelhando-se estatisticamente aos métodos de regressão. Já para a métrica NDCG (Figura 4.6), observou-se um isolamento claro dos métodos constituintes, que se mostraram significativamente superiores (menor disparidade) a todos os modelos híbridos.

Quanto às métricas de erro (Figuras 4.7 e 4.8), o teste de *Tukey HSD* evidenciou que, embora o NMF e o SVD mantenham perfis de *fairness* singulares, algoritmos como o BIASEDMF apresentam maior proximidade estatística com os resultados dos híbridos. De forma geral, os resultados confirmam que os métodos clássicos mantêm, consistentemente, menores valores de diferença absoluta entre os grupos de usuários ao longo das janelas temporais.

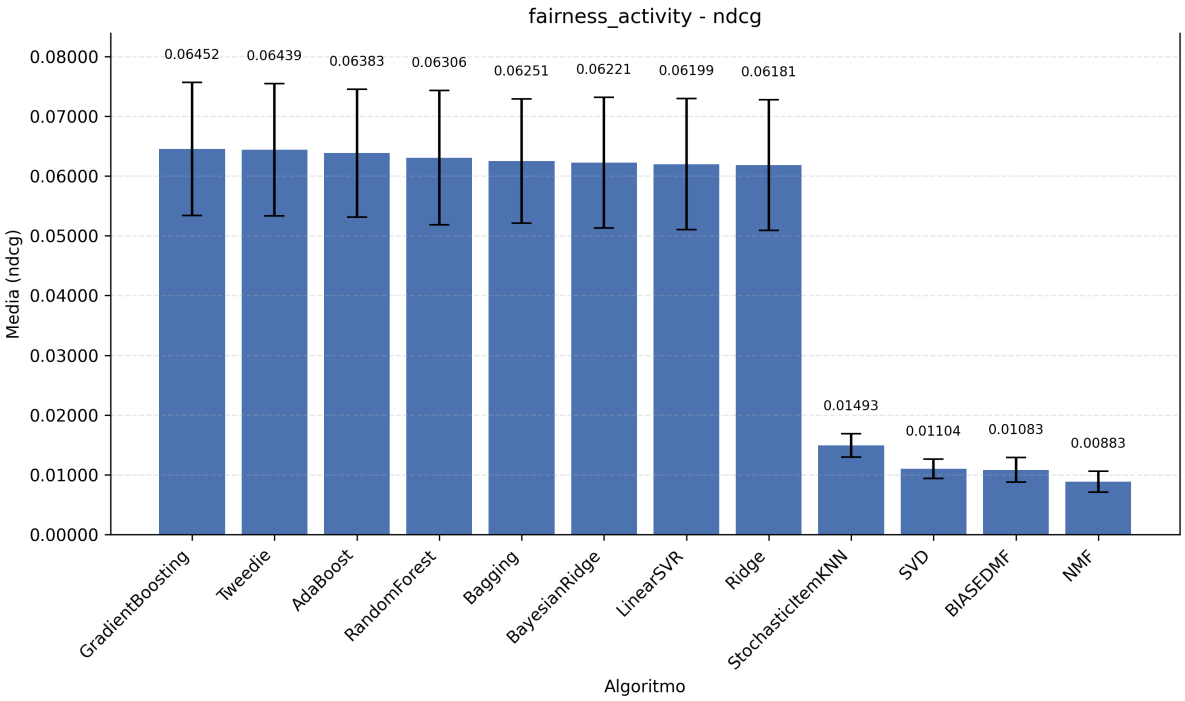

Figura 4.6 – *Fairness* calculado com base em NDCG, com intervalos de confiança.

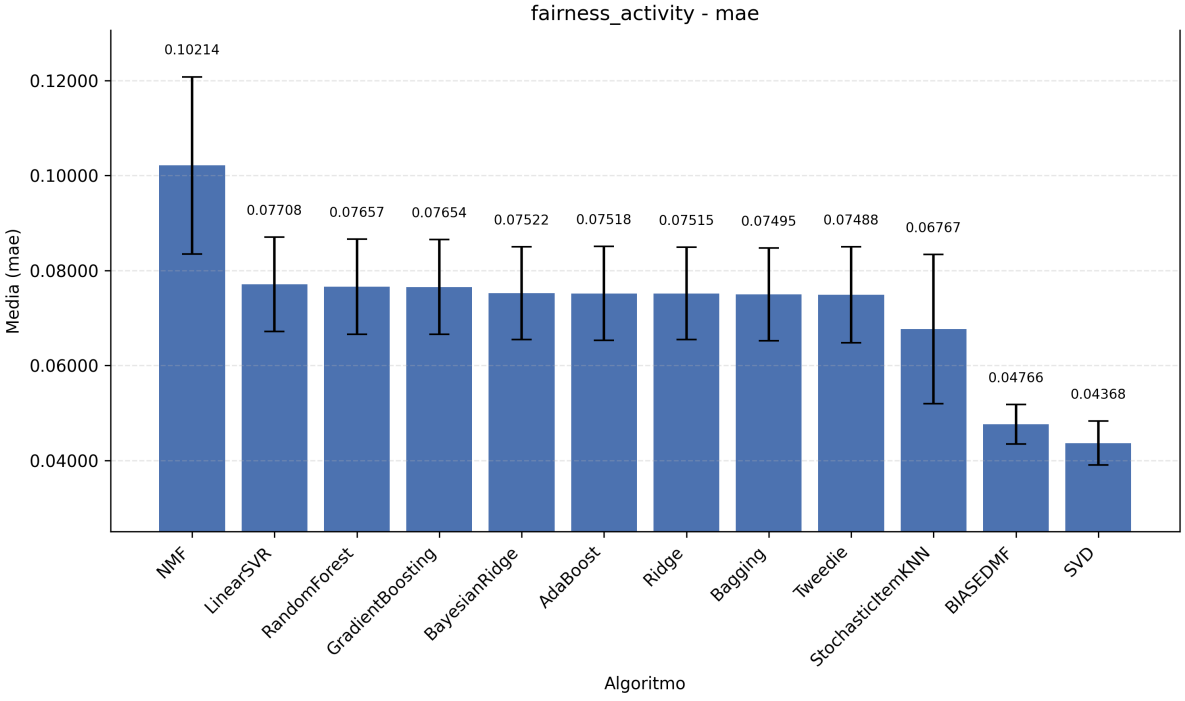

Figura 4.7 – *Fairness* calculado com base em MAE, com intervalos de confiança.

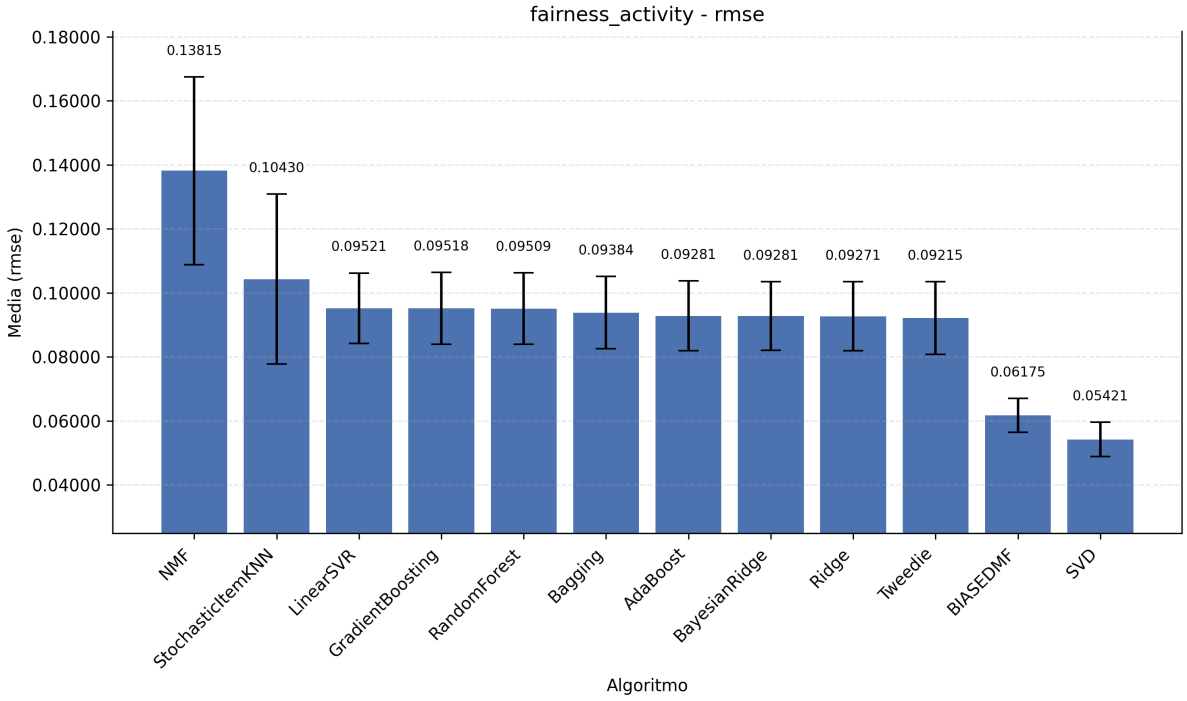

Figura 4.8 – *Fairness* calculado com base em RMSE, com intervalos de confiança.

Em contrapartida, os métodos híbridos apresentaram disparidades superiores e maior variabilidade. Portanto, **os resultados indicam que a hibridização ponderada também não foi capaz de promover maior justiça nas recomendações sob a perspectiva do nível de atividade dos usuários, refutando a hipótese de que a combinação de modelos reduziria o viés algorítmico neste cenário experimental**.

4.3 Avaliando Sensibilidade ao Risco (*GeoRisk*)

Nesta seção, são apresentados os resultados da sensibilidade ao risco, mensurada pela métrica *GeoRisk*. Conforme detalhado na Seção 3.3, diferentemente da diferença absoluta utilizada para avaliar *fairness*, no *GeoRisk* valores maiores indicam sistemas mais robustos, com menor variação na qualidade das recomendações entre os usuários e, consequentemente, menor sensibilidade ao risco. As Figuras 4.9, 4.10, 4.11 e 4.12 apresentam os resultados calculados para cada uma das quatro métricas base.

A análise dos resultados, amparada pelo teste de *Tukey HSD*, revela um padrão de comportamento consistente com os achados de *fairness* discutidos anteriormente. Para as métricas de ranking (*F1* e *NDCG*), os métodos constituintes apresentaram, de forma sistemática, valores de *GeoRisk* superiores e estatisticamente distintos dos valores obtidos pelos algoritmos híbridos. Este cenário indica que abordagens clássicas, como o *BIASEDMF* e o *SVD*, oferecem maior robustez e menor variabilidade na experiência dos usuários sob a ótica de ranking.

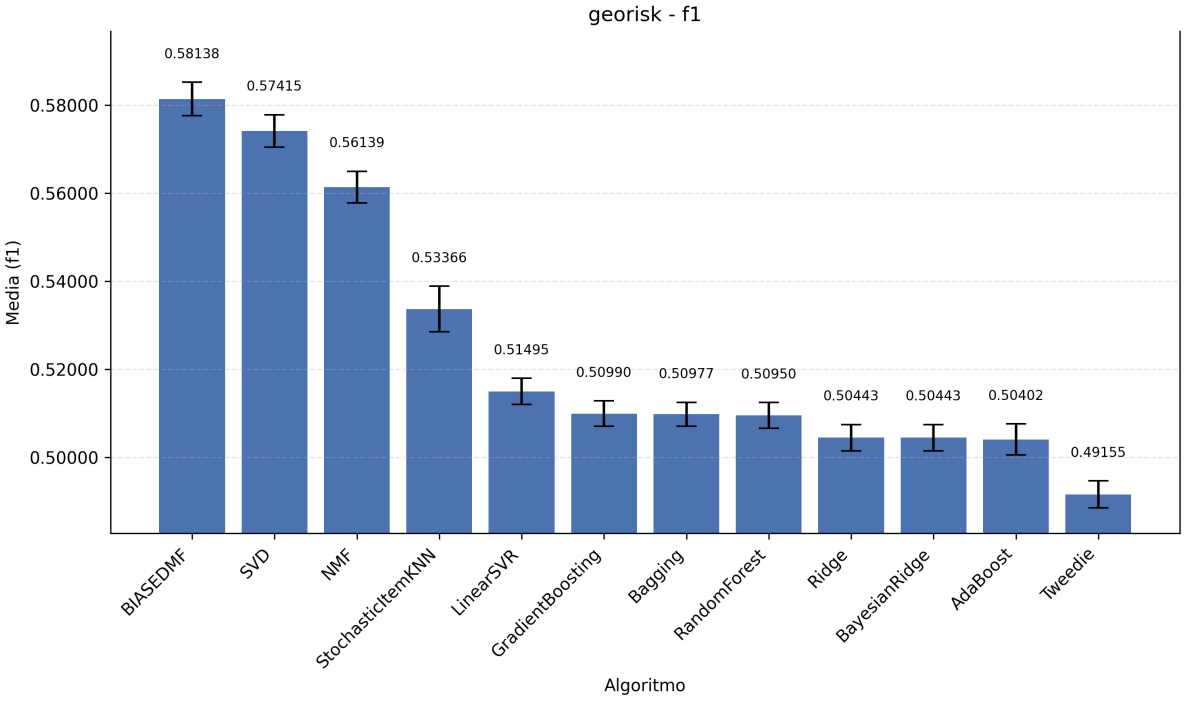

Figura 4.9 – *GeoRisk* médio calculado com base em F1, com intervalos de confiança.

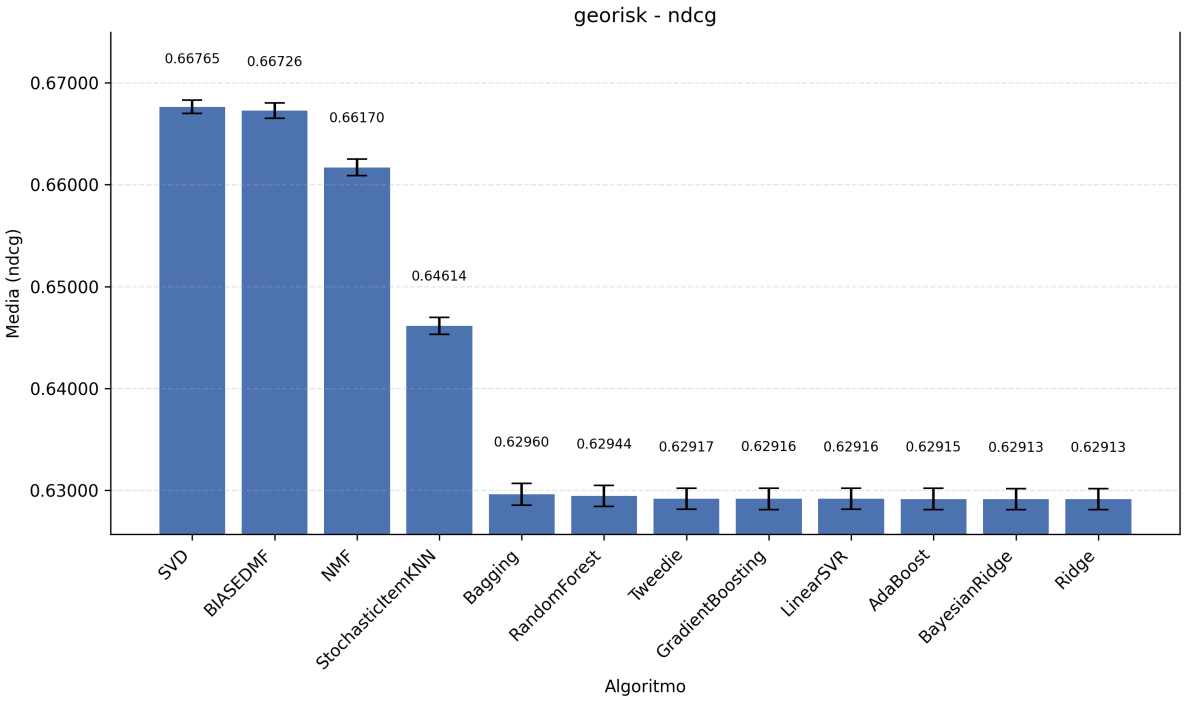

Figura 4.10 – *GeoRisk* médio calculado com base em NDCG, com intervalos de confiança.

Por outro lado, para as métricas de erro (*MAE* e *RMSE*), embora os gráficos sugiram uma vantagem visual para os métodos base, o teste estatístico não evidenciou diferenças significativas

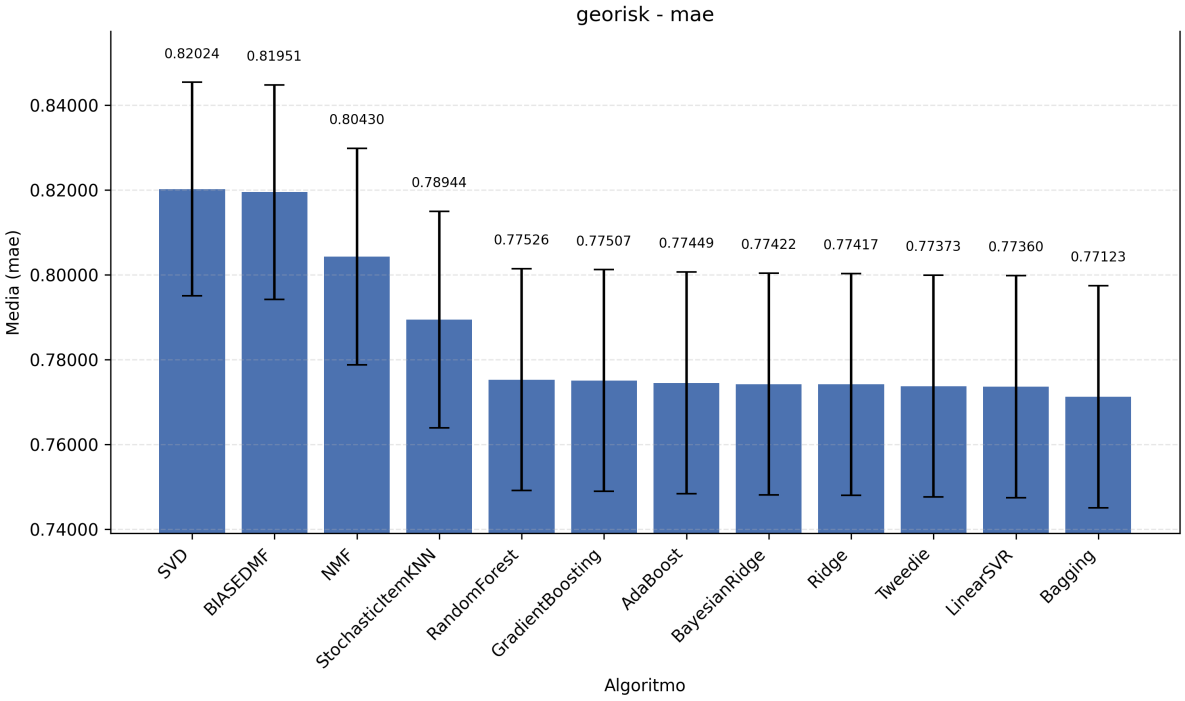

Figura 4.11 – *GeoRisk* médio calculado com base em MAE, com intervalos de confiança.

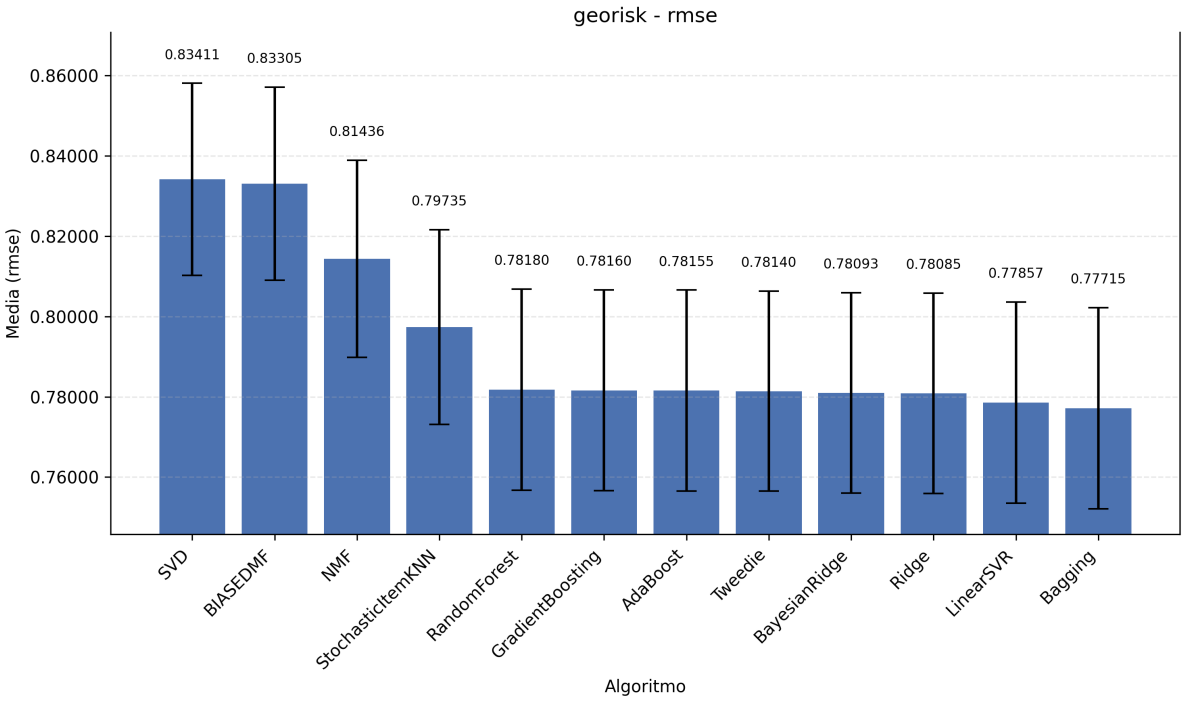

Figura 4.12 – *GeoRisk* médio calculado com base em RMSE, com intervalos de confiança.

entre os algoritmos avaliados. Este “empate técnico” decorre da alta variabilidade observada nos intervalos de confiança dos métodos constituintes dessas métricas específicas, o que impede a

afirmação de superioridade estatística.

Em suma, os experimentos indicam que a hibridização ponderada, no cenário experimental adotado, tende a aumentar a sensibilidade ao risco ou, no melhor dos casos, manter o perfil de instabilidade dos modelos base. Esse comportamento sugere que a etapa de regressão, focada na minimização do erro médio global, pode sacrificar a estabilidade regional das recomendações. Assim, **os resultados observados não sustentam a hipótese de que as técnicas de hibridização avaliadas contribuem para gerar recomendações menos sensíveis ao risco**.

4.4 Análise do Desempenho Isolado por Grupos

Para compreender os motivos pelos quais a hibridização não resultou em melhorias nas métricas analisadas (*fairness* e *GeoRisk*), é fundamental avaliar o desempenho absoluto dos algoritmos, isoladamente, para cada subgrupo (ativos *versus* inativos e masculino *versus* feminino). A observação das médias isoladas revela que o aumento da diferença absoluta (injustiça) nos métodos híbridos não decorreu de uma melhora desproporcional em um grupo privilegiado, mas sim de uma degradação generalizada da qualidade da recomendação, que afetou os grupos de forma assimétrica.

No cenário de **atividade dos usuários**, cujos desempenhos médios estão ilustrados nas Figuras 4.13, 4.14, 4.15 e 4.16, os algoritmos constituintes apresentaram erros de predição (RMSE e MAE) consideravelmente menores no grupo de usuários altamente ativos. Isso evidencia que a maior disponibilidade de dados históricos facilita o aprendizado do perfil do usuário. No entanto, observa-se nas Figuras 4.13 e 4.14 que esses métodos conseguiram manter a qualidade do ranqueamento sem diferença estatisticamente significativa entre os grupos.

Por outro lado, os métodos híbridos baseados em regressão apresentaram uma perda expressiva na qualidade global. Conforme demonstram as Figuras 4.13 e 4.14, o impacto negativo no ranqueamento (NDCG) e na precisão (*F1-score*) foi ainda mais severo no grupo de usuários altamente ativos do que nos inativos. Uma hipótese para explicar tal comportamento é a de que usuários muito ativos tendem a possuir perfis de consumo mais ecléticos e complexos. Ao tentar minimizar o erro global, o regressor utilizado na hibridização pode atuar como um suavizador, levando as predições para uma média mais genérica. Esse efeito de "regressão à média" acabaria prejudicando a personalização fina necessária para ranquear adequadamente os itens de usuários com histórico denso, adicionando ruído às predições sólidas geradas pelos métodos base.

A seguir, as Figuras 4.17, 4.18, 4.19 e 4.20 exibem as disparidades no desempenho médio entre os usuários com base em seu gênero.

Em relação à análise por **gênero**, as médias isoladas confirmam a presença de um viés natural no conjunto de dados *MovieLens 1M*, predominantemente composto por avaliações de usuários do sexo masculino. Os métodos clássicos refletiram esse desbalanceamento demográfico,

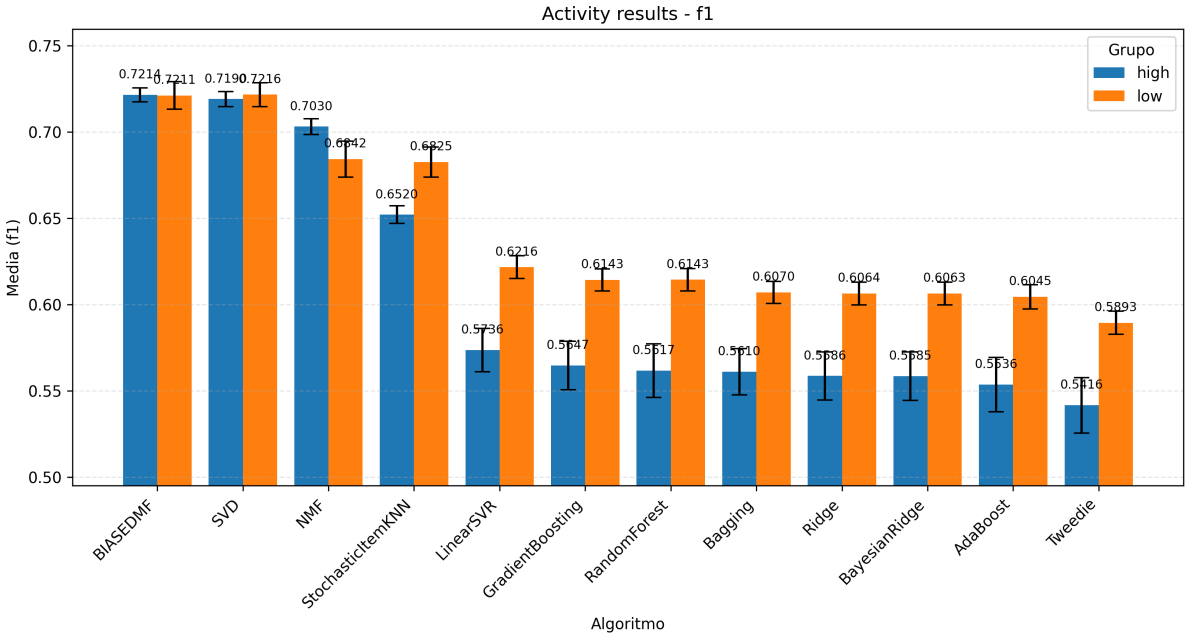

Figura 4.13 – Resultado médio do cálculo de F1, com base na atividade dos usuários.

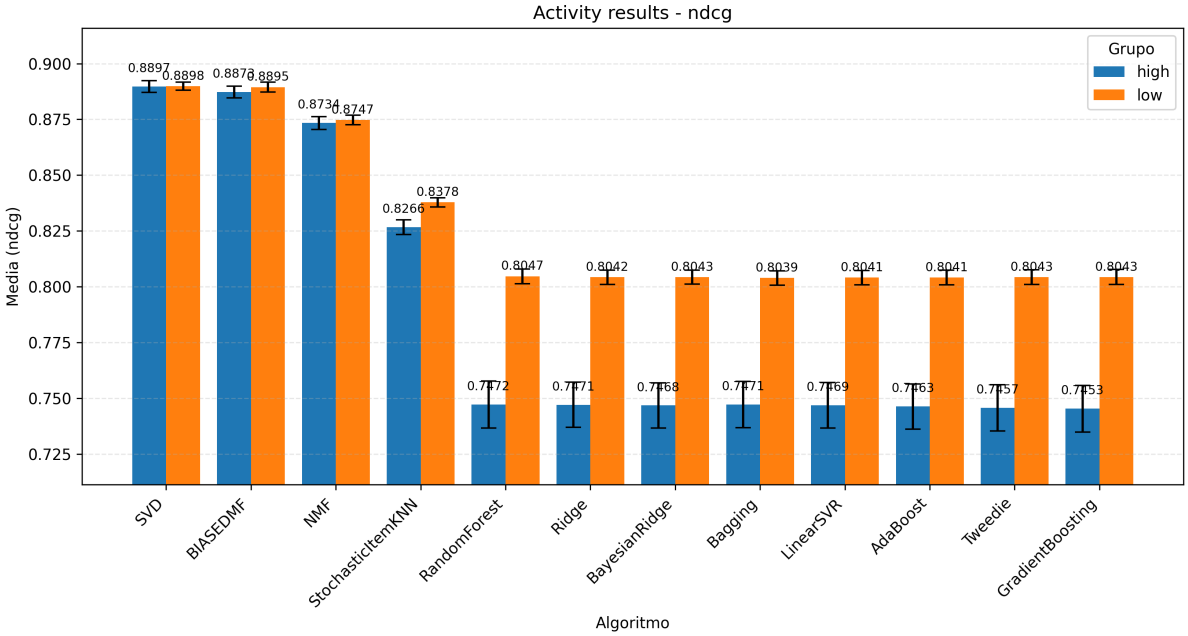

Figura 4.14 – Resultado médio do cálculo de NDCG, com base na atividade dos usuários.

apresentando resultados ligeiramente superiores para homens, como observado nas métricas de erro das Figuras 4.19 e 4.20. Ao aplicar as técnicas de hibridização, os dados mostram que os modelos preservaram a vantagem de desempenho no grupo masculino, ao mesmo tempo em que elevaram a taxa de erro global em ambos os sexos.

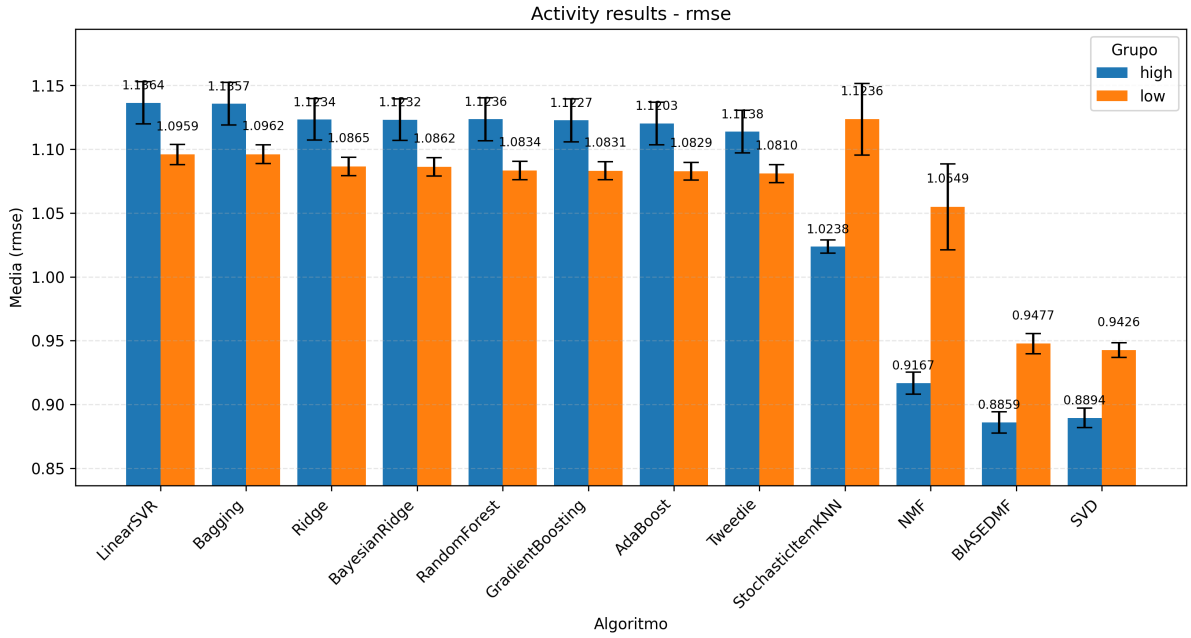

Figura 4.15 – Resultado médio do cálculo de RMSE, com base na atividade dos usuários.

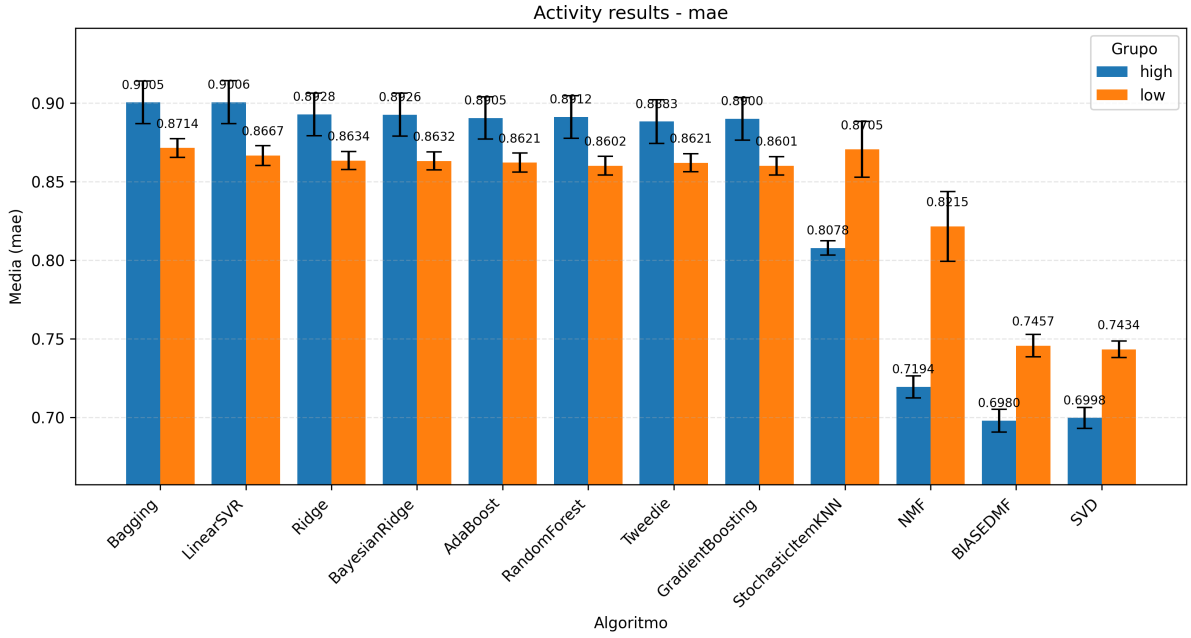

Figura 4.16 – Resultado médio do cálculo de MAE, com base na atividade dos usuários.

Esse comportamento corrobora a hipótese levantada na análise de atividade: ao aplicar uma camada de regressão que suaviza as predições em direção a um consenso global, a fim de minimizar o erro de treinamento, o sistema passa a favorecer o grupo demográfico majoritário (homens), cujos padrões ditam a "média" do *dataset*. Consequentemente, o modelo penaliza subgrupos cujas preferências desviam desse padrão majoritário, como o público feminino.

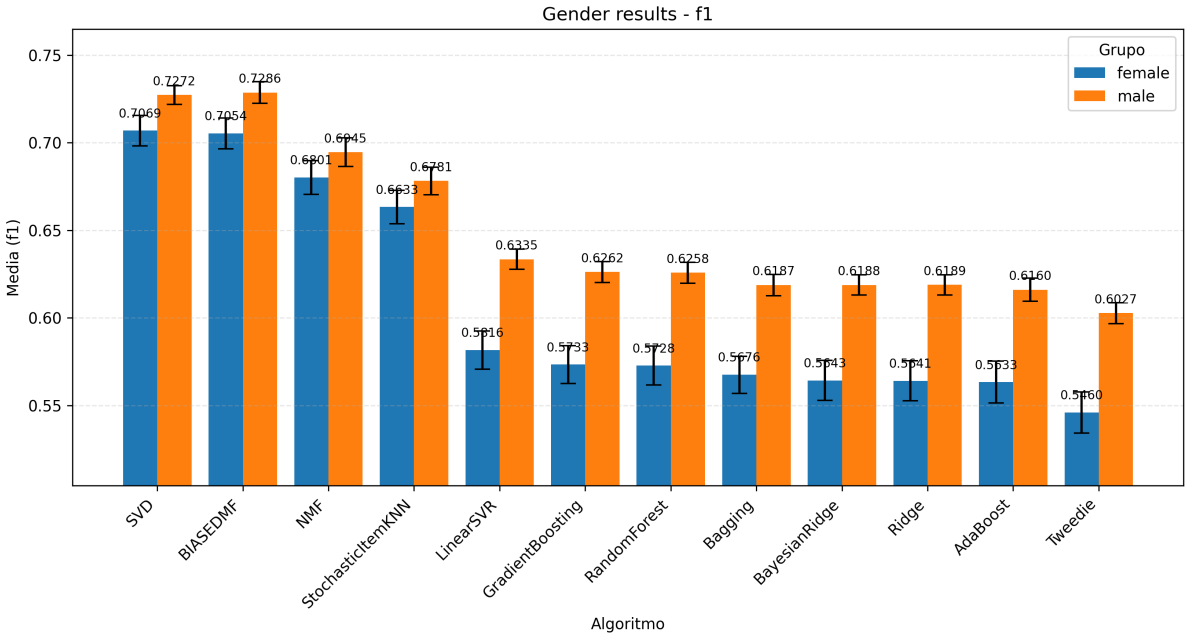

Figura 4.17 – Resultado médio do cálculo de F1, com base no gênero dos usuários.

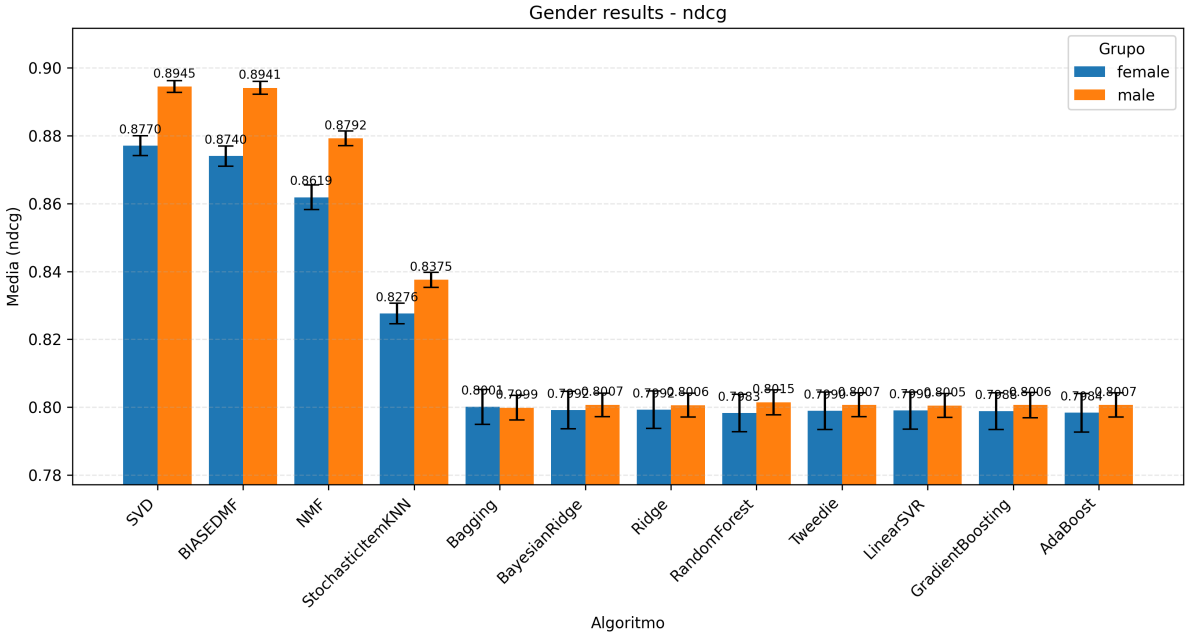

Figura 4.18 – Resultado médio do cálculo de NDCG, com base no gênero dos usuários.

Em síntese, **a análise das médias isoladas sugere que a hibridização com modelos de regressão não atuou como mecanismo de calibração ou de correção de viés**. Em vez disso, os indícios apontam para uma perda de personalização. Ao nivelar as recomendações em função de tendências globais, a hibridização reduziu a eficácia geral do sistema e desestabilizou as predições de forma desigual, prejudicando perfis mais complexos (usuários ativos) e mantendo

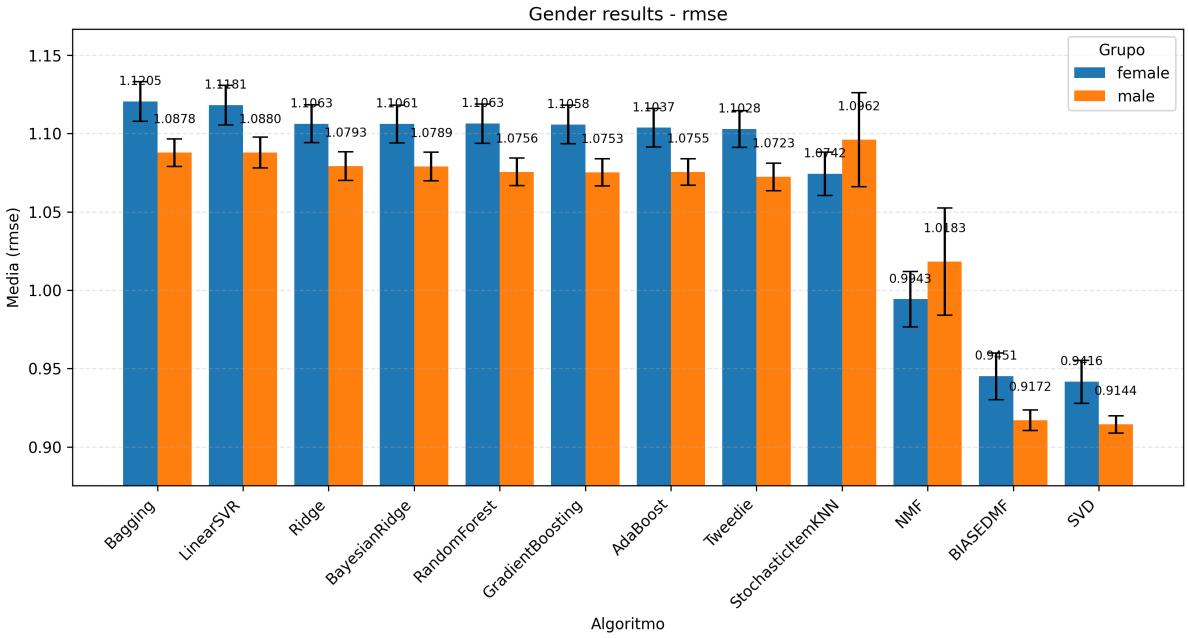

Figura 4.19 – Resultado médio do cálculo de RMSE, com base no gênero dos usuários.

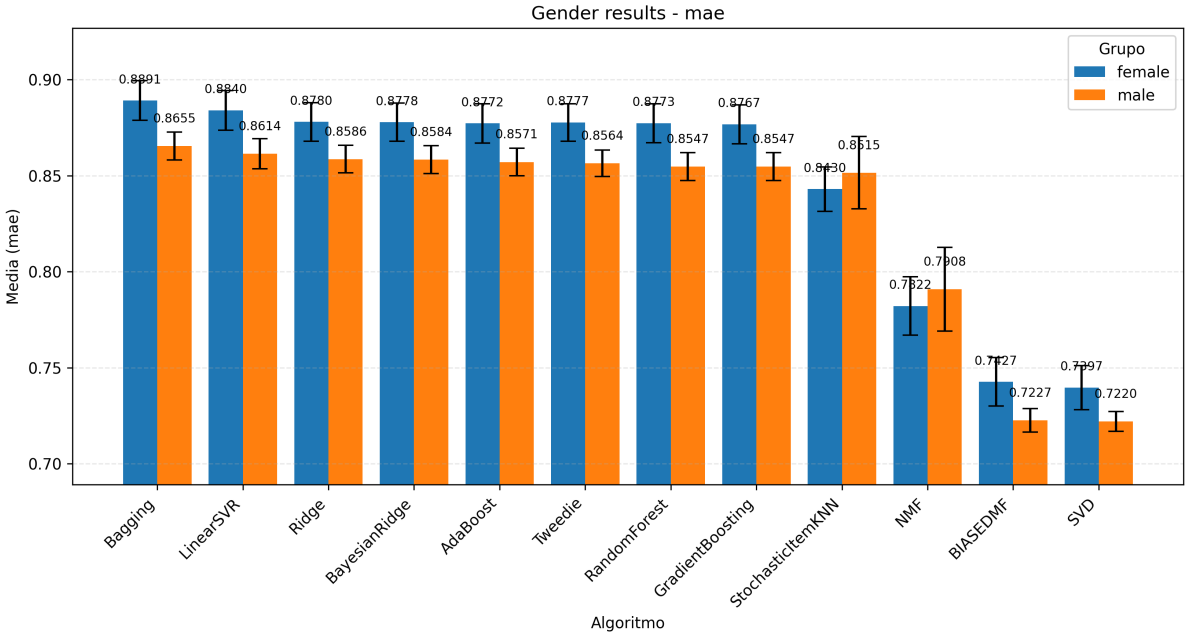

Figura 4.20 – Resultado médio do cálculo de MAE, com base no gênero dos usuários.

as desvantagens de grupos minoritários (mulheres). Isso se refletiu estatisticamente no aumento das métricas de injustiça e na diminuição da robustez global (sensibilidade ao risco) discutidas nas seções anteriores.

5 Considerações Finais

Neste capítulo, são apresentadas as considerações finais sobre a pesquisa desenvolvida. Inicialmente, a Seção 5.1 expõe as conclusões derivadas dos experimentos empíricos, respondendo diretamente às questões de pesquisa levantadas no início deste trabalho. Em seguida, a Seção 5.2 destaca as principais contribuições deste estudo para a literatura de Sistemas de Recomendação. A Seção 5.3 reconhece as limitações inerentes à metodologia e aos dados adotados. Por fim, a Seção 5.4 sugere possíveis caminhos e melhorias arquiteturais para pesquisas futuras.

5.1 Conclusões

Este trabalho teve como objetivo principal investigar o impacto das técnicas de hibridização em Sistemas de Recomendação (SRs) sob a ótica de métricas que vão além da precisão tradicional. O foco central consistiu em avaliar empírica e estatisticamente se a combinação de métodos, através de modelos de regressão, seria capaz de promover resultados mais justos (*fairness* por gênero e atividade) e menos sensíveis ao risco (*GeoRisk*).

Para garantir a robustez das análises, os experimentos foram conduzidos utilizando uma abordagem de validação temporal com 20 janelas deslizantes sobre o conjunto de dados *MovieLens 1M*, mitigando o risco de vieses gerados por um único recorte estático de tempo.

Com base nos resultados obtidos, é possível responder diretamente às duas indagações que guiaram este estudo:

- **Técnicas de hibridização contribuem para resultados mais justos na recomendação?** Os resultados indicam que **não**. A hibridização ponderada por regressão ampliou a diferença absoluta entre os grupos avaliados. Na análise por atividade, o ranqueamento piorou mais severamente entre os usuários altamente ativos do que entre os inativos. Na análise por gênero, os modelos híbridos elevaram a taxa de erro global, mas mantiveram a vantagem estatisticamente significativa em relação ao público masculino.
- **Técnicas de hibridização contribuem para resultados menos sensíveis ao risco?** Os dados demonstram que **não**. Os valores de *GeoRisk* revelaram que os algoritmos híbridos foram mais suscetíveis a quedas acentuadas de desempenho para subconjuntos de usuários, apresentando menor robustez e maior variabilidade do que métodos constituintes como *userKNN* ou fatoração de matrizes isoladas.

O comportamento contraintuitivo observado, em que a combinação de múltiplos algoritmos piorou a equidade e a robustez, pode ser explicado pelo funcionamento matemático dos

modelos de regressão empregados. Ao tentar minimizar o erro global (como o MSE) durante o treinamento, a camada híbrida tende a atuar como um "suavizador", puxando as predições para uma média de consenso.

Nesse processo de regressão à média, o modelo passa a favorecer os padrões estatísticos dominantes no *dataset*. No caso do *MovieLens 1M*, esse padrão é ditado por usuários do gênero masculino (a esmagadora maioria demográfica) e por usuários com padrões de consumo previsíveis e *mainstream* (geralmente usuários menos ativos). Consequentemente, o modelo destrói a personalização fina necessária para atender a usuários com perfis complexos (os muito ativos) e pune subgrupos cujas preferências desviam da norma (como o público feminino). A hibridização, portanto, atuou como um nivelador genérico, sacrificando a precisão individual em prol de uma redução de erro de treinamento global enviesada.

5.2 Contribuições do Trabalho

A principal contribuição desta pesquisa é fornecer evidências empíricas de que o ganho de complexidade algorítmica, através da combinação de modelos, não se traduz automaticamente em recomendações mais éticas ou robustas. O trabalho demonstra que a arquitetura híbrida padrão, por ser "cega" a atributos demográficos e níveis de atividade, pode na verdade amplificar vieses já existentes nos dados de origem.

5.3 Limitações

É importante reconhecer as limitações empíricas deste estudo. Em primeiro lugar, os experimentos foram restritos ao *dataset MovieLens 1M*. Embora seja um padrão na literatura de SRs, apresenta um forte desbalanceamento demográfico e se concentra exclusivamente no domínio de filmes. Além disso, não foi avaliada a correlação entre as predições dos algoritmos constituintes. Em abordagens de hibridização, a eficácia da combinação depende dos modelos base capturarem sinais diferentes e cometerem erros distintos. A ausência dessa análise impede verificar se a falha em melhorar as métricas do sistema foi agravada por uma possível alta redundância (correlação) entre as predições geradas pelos métodos clássicos utilizados como entrada para os regressores.

5.4 Trabalhos Futuros

A partir das conclusões e limitações mapeadas, abrem-se diversas frentes para trabalhos futuros:

1. **Exploração de Outras Arquiteturas Híbridas:** Investigar se modelos híbridos em cascata ou mistos preservam melhor a personalização do que o método ponderado (regressão)

empregado neste trabalho.

2. **Diversificação de Domínios:** Replicar a mesma metodologia de janelas deslizantes em conjuntos de dados de domínios variados (como *e-commerce*, música ou leitura) e com distribuições demográficas mais balanceadas, a fim de verificar se o fenômeno da regressão à média persiste em diferentes contextos de recomendação.

3. **Análise de Correlação e Diversidade:** Incorporar etapas de seleção de *features* baseadas na correlação de Pearson entre os métodos constituintes, garantindo que apenas os algoritmos com alta diversidade preditiva sejam selecionados para compor o modelo híbrido.

4. **Exploração de estratégias de otimização multi-objetivo**: Uma vez identificada a dificuldade de modelos híbridos em obter resultados mais justos e menos sensíveis ao risco, a exploração de métodos multi-objetivo, que considerem explicitamente estes fatores de qualidade, pode configurar um caminho promissor para garantir resultados híbridos mais justos e menos sensíveis ao risco.

Referências

ABDOLLAHPOURI, H.; MANSOURY, M.; BURKE, R.; MOBASHER, B. The connection between popularity bias, calibration, and fairness in recommendation. *arXiv preprint arXiv:2101.00790*, 2021.

ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L. How we analyzed the COMPAS recidivism algorithm. *ProPublica*, 2016. Disponível em: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

BURKE, R. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, Springer, v. 12, n. 4, p. 331–370, Nov 2002. ISSN 0924-1868, 1573-1391.

BURKE, R. Hybrid web recommender systems. In: *The Adaptive Web*. Springer Berlin Heidelberg, 2007. v. 4321, p. 377–408. ISBN 978-3-540-72078-2. Disponível em: <https://doi.org/10.1007/978-3-540-72079-9_12>.

BURKE, R.; SONBOLI, N.; ORDONEZ-GAUGER, A. Balanced neighborhoods for multi-sided fairness in recommendation. In: FRIEDLER, S. A.; WILSON, C. (Ed.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 2018. (Proceedings of Machine Learning Research, v. 81), p. 202–214. Disponível em: <https://proceedings.mlr.press/v81/burke18a.html>.

CHEN, L.; AL. et. Innovative stacking model with bayesian ridge regression meta-learner. *PMC*, 2025. Demonstra a eficácia do Bayesian Ridge como meta-modelo em arquiteturas stacking.

DINÇER, B. T.; OUNIS, I.; MACDONALD, C. Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In: SPRINGER. *Proceedings of the 36th European Conference on Information Retrieval*. [S.l.], 2014. p. 26–38.

FORTES, R. S. Enhancing the multi-objective recommendation from three new perspectives: data characterization, risk-sensitiveness, and prioritization of the objectives. 2022.

FU, Z.; XIAN, Y.; GAO, R.; ZHAO, J.; HUANG, Q.; GE, Y.; XU, S.; GENG, S.; SHAH, C.; ZHANG, Y.; MELO, G. de. *Fairness-Aware Explainable Recommendation over Knowledge Graphs*. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2006.02046>.

GOLDBERG, D.; NICHOLS, D.; OKI, B. M.; TERRY, D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, ACM New York, v. 35, n. 12, p. 61–70, 1992.

HARDT, M.; PRICE, E.; SREBRO, N. Equality of opportunity in supervised learning. 2016.

HERLOCKER, J.; KONSTAN, J. A.; BORCHERS, A.; RIEDL, J. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, ACM, v. 19, n. 1, p. 5–53, 2001.

JI, Y.; SUN, A.; ZHANG, J.; LI, C. A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems (TOIS)*, Association for Computing Machinery, New York, NY, USA, v. 41, n. 3, 2023.

KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix factorization techniques for recommender systems. *Computer*, IEEE, v. 42, n. 8, p. 30–37, 2009. Base teórica para SVD e BiasedSVD [3, 2].

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999. Referência fundamental para algoritmos NMF em sistemas de recomendação [1, 2].

LI, Y.; CHEN, H.; FU, Z.; GE, Y.; ZHANG, Y. User-oriented fairness in recommendation. *CoRR*, abs/2104.10671, 2021. Disponível em: <https://arxiv.org/abs/2104.10671>.

LIMA, R.; COMARELLA, G.; BELéM, F.; REIS, J. Justiça em sistemas de recomendação: Uma análise de técnicas de regularização. In: *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*. Porto Alegre, RS, Brasil: SBC, 2022. p. 177–189. ISSN 2763-8979. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/21805>.

MA, H.; ZHOU, D.; LIU, C.; LYU, M. R.; KING, I. Recommender systems with social regularization. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2011. (WSDM '11), p. 287–296. ISBN 9781450304931. Disponível em: <https://doi.org/10.1145/1935826.1935877>.

PITOURA, E.; STEFANIDIS, K.; KOUTRIKA, G. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, Springer, p. 651–654, Oct 2021.

QUADRANA, M.; CREMONESI, P.; JANNACH, D. Sequence-aware recommender systems. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 4, p. 1–36, 2018.

RODRIGUES, P. H. S.; SOUSA, D. X. de; FRANçA, C.; RABBI, G.; ROSA, T. C.; GONçALVES, M. A. Risk-sensitive optimization of neural deep learning ranking models with applications in ad-hoc retrieval and recommender systems. *Information Processing & Management*, Elsevier, v. 62, p. 104126, 2025. Disponível em: <https://doi.org/10.1016/j.ipm.2025.104126>.

SOUSA, D. X.; CANUTO, S.; GONçALVES, M. A.; ROSA, T. C.; MARTINS, W. S. Risk-sensitive learning to rank with evolutionary multi-objective feature selection. *Information Processing Management*, 2019.

STIJGER, E. *Active learning in recommender systems for predicting vulnerabilities in software*. Master's Thesis — Utrecht University, 2025. Pesquisa sobre Item-based KNN e métricas de ranking como nDCG [7].

TOMBUS, A. C. Impact of recommender systems in e-commerce - a worldwide empirical analysis. *Journal of Business Research*, 2025. Análise de mercado e eficácia industrial de sistemas de recomendação até 2026 [5, 6].

WANG, L.; BENNETT, P. N.; COLLINS-THOMPSON, K. Robust ranking models via risk-sensitive optimization. 2012.

WANG, X.; WANG, W. H. Providing item-side individual fairness for deep recommender systems. In: . [S.l.: s.n.], 2020.

WANG, Y.; ZHOU, H.; LU, G.-F.; GAO, C.; MENG, S. Improving user-oriented fairness in recommendation via data augmentation: Don't worry about inactive users. *Journal of Systems and Software*, v. 225, p. 112387, 2025. Disponível em: <https://doi.org/10.1016/j.jss.2025.112387>.

YANG, J.; EMURA, T. Shared parameter alternating tweedie (sa-tweedie) for recommender systems. *Preprints*, 2025. Modelo especializado para dados esparsos com inflação de zeros.

ZHANG, Z.-K.; LIU, C.; ZHANG, Y.-C.; ZHOU, T. Solving the cold start problem in recommender systems with social tags. 2010.

Apêndices

APÊNDICE A – Resultados detalhados

Tabela A.1 – Resultados de fairness por atividade (F1) ao longo das janelas temporais.

Método	Média	IC 95% (Inf.)	IC 95% (Sup.)
Tweedie	0.08196	0.06871	0.09522
RandomForest	0.08024	0.06662	0.09387
AdaBoost	0.07853	0.06588	0.09117
GradientBoosting	0.07580	0.06363	0.08798
BayesianRidge	0.07485	0.06295	0.08675
Ridge	0.07465	0.06283	0.08647
LinearSVR	0.07307	0.06191	0.08424
Bagging	0.07280	0.06143	0.08416
StochasticItemKNN	0.05469	0.04861	0.06076
NMF	0.05252	0.04467	0.06037
BIASEDMF	0.03785	0.03301	0.04269
SVD	0.03721	0.03178	0.04264

Tabela A.2 – Resultados de fairness por atividade (MAE) ao longo das janelas temporais.

Método	Média	IC 95% (Inf.)	IC 95% (Sup.)
NMF	0.10214	0.08349	0.12078
LinearSVR	0.07708	0.06714	0.08702
RandomForest	0.07657	0.06656	0.08658
GradientBoosting	0.07654	0.06657	0.08652
BayesianRidge	0.07522	0.06549	0.08496
AdaBoost	0.07518	0.06528	0.08508
Ridge	0.07515	0.06542	0.08489
Bagging	0.07495	0.06521	0.08470
Tweedie	0.07488	0.06477	0.08498
StochasticItemKNN	0.06767	0.05192	0.08341
BIASEDMF	0.04766	0.04350	0.05181
SVD	0.04368	0.03904	0.04833

Tabela A.3 – Resultados de fairness por atividade (NDCG) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
GradientBoosting	0.06452	0.05338	0.07566
Tweedie	0.06439	0.05333	0.07546
AdaBoost	0.06383	0.05313	0.07454
RandomForest	0.06306	0.05183	0.07429
Bagging	0.06251	0.05213	0.07289
BayesianRidge	0.06221	0.05126	0.07316
LinearSVR	0.06199	0.05103	0.07294
Ridge	0.06181	0.05086	0.07275
StochasticItemKNN	0.01493	0.01299	0.01688
SVD	0.01104	0.00941	0.01266
BIASEDMF	0.01083	0.00878	0.01288
NMF	0.00883	0.00706	0.01059

Tabela A.4 – Resultados de fairness por atividade (RMSE) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
NMF	0.13815	0.10882	0.16748
StochasticItemKNN	0.10430	0.07770	0.13090
LinearSVR	0.09521	0.08422	0.10620
GradientBoosting	0.09518	0.08395	0.10641
RandomForest	0.09509	0.08397	0.10621
Bagging	0.09384	0.08260	0.10508
AdaBoost	0.09281	0.08186	0.10376
BayesianRidge	0.09281	0.08207	0.10355
Ridge	0.09271	0.08198	0.10343
Tweedie	0.09215	0.08084	0.10347
BIASEDMF	0.06175	0.05646	0.06704
SVD	0.05421	0.04881	0.05962

Tabela A.5 – Resultados de fairness por gênero (F1) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
Tweedie	0.07775	0.06970	0.08580
Ridge	0.07338	0.06519	0.08158
BayesianRidge	0.07314	0.06496	0.08132
AdaBoost	0.07313	0.06495	0.08131
RandomForest	0.07139	0.06310	0.07969
GradientBoosting	0.07083	0.06273	0.07894
LinearSVR	0.07003	0.06210	0.07796
Bagging	0.06956	0.06167	0.07746
StochasticItemKNN	0.05062	0.04265	0.05860
SVD	0.04159	0.03529	0.04789
NMF	0.04031	0.03382	0.04679
BIASEDMF	0.03712	0.03090	0.04333

Tabela A.6 – Resultados de fairness por gênero (MAE) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
StochasticItemKNN	0.08254	0.06538	0.09970
NMF	0.07413	0.05763	0.09063
LinearSVR	0.05703	0.04556	0.06850
RandomForest	0.05634	0.04505	0.06763
Ridge	0.05626	0.04538	0.06715
BayesianRidge	0.05623	0.04534	0.06711
GradientBoosting	0.05614	0.04483	0.06744
Bagging	0.05598	0.04460	0.06736
AdaBoost	0.05541	0.04431	0.06651
Tweedie	0.05379	0.04288	0.06469
SVD	0.04596	0.03734	0.05459
BIASEDMF	0.04583	0.03772	0.05395

Tabela A.7 – Resultados de fairness por gênero (NDCG) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
RandomForest	0.02631	0.02216	0.03047
AdaBoost	0.02609	0.02199	0.03020
GradientBoosting	0.02572	0.02178	0.02967
Tweedie	0.02545	0.02146	0.02944
LinearSVR	0.02527	0.02131	0.02922
Ridge	0.02524	0.02128	0.02919
BayesianRidge	0.02523	0.02125	0.02920
Bagging	0.02484	0.02129	0.02840
BIASEDMF	0.02011	0.01758	0.02263
NMF	0.01907	0.01571	0.02242
SVD	0.01803	0.01528	0.02077
StochasticItemKNN	0.01524	0.01318	0.01730

Tabela A.8 – Resultados de fairness por gênero (RMSE) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
StochasticItemKNN	0.11679	0.08878	0.14480
NMF	0.10136	0.07410	0.12863
LinearSVR	0.06762	0.05346	0.08179
GradientBoosting	0.06684	0.05317	0.08052
Bagging	0.06673	0.05294	0.08052
RandomForest	0.06655	0.05294	0.08016
Ridge	0.06609	0.05263	0.07956
BayesianRidge	0.06603	0.05257	0.07950
AdaBoost	0.06560	0.05224	0.07897
Tweedie	0.06296	0.04946	0.07647
SVD	0.05732	0.04614	0.06850
BIASEDMF	0.05488	0.04419	0.06556

Tabela A.9 – Resultados de GeoRisk (F1) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
BIASEDMF	0.58138	0.57757	0.58520
SVD	0.57415	0.57049	0.57780
NMF	0.56139	0.55782	0.56496
StochasticItemKNN	0.53366	0.52849	0.53884
LinearSVR	0.51495	0.51197	0.51793
GradientBoosting	0.50990	0.50702	0.51278
Bagging	0.50977	0.50706	0.51248
RandomForest	0.50950	0.50659	0.51241
Ridge	0.50443	0.50144	0.50741
BayesianRidge	0.50443	0.50145	0.50741
AdaBoost	0.50402	0.50050	0.50753
Tweedie	0.49155	0.48849	0.49461

Tabela A.10 – Resultados de GeoRisk (MAE) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
SVD	0.82024	0.79506	0.84542
BIASEDMF	0.81951	0.79422	0.84480
NMF	0.80430	0.77877	0.82982
StochasticItemKNN	0.78944	0.76390	0.81498
RandomForest	0.77526	0.74908	0.80144
GradientBoosting	0.77507	0.74893	0.80121
AdaBoost	0.77449	0.74833	0.80066
BayesianRidge	0.77422	0.74805	0.80038
Ridge	0.77417	0.74800	0.80033
Tweedie	0.77373	0.74756	0.79991
LinearSVR	0.77360	0.74741	0.79980
Bagging	0.77123	0.74503	0.79744

Tabela A.11 – Resultados de GeoRisk (NDCG) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
SVD	0.66765	0.66701	0.66829
BIASEDMF	0.66726	0.66652	0.66801
NMF	0.66170	0.66088	0.66252
StochasticItemKNN	0.64614	0.64531	0.64698
Bagging	0.62960	0.62853	0.63067
RandomForest	0.62944	0.62841	0.63048
Tweedie	0.62917	0.62815	0.63020
GradientBoosting	0.62916	0.62811	0.63022
LinearSVR	0.62916	0.62813	0.63019
AdaBoost	0.62915	0.62811	0.63020
BayesianRidge	0.62913	0.62810	0.63015
Ridge	0.62913	0.62809	0.63016

Tabela A.12 – Resultados de GeoRisk (RMSE) ao longo das janelas temporais.

Método	**Média**	**IC 95% (Inf.)**	**IC 95% (Sup.)**
SVD	0.83411	0.81018	0.85805
BIASEDMF	0.83305	0.80897	0.85713
NMF	0.81436	0.78984	0.83888
StochasticItemKNN	0.79735	0.77308	0.82162
RandomForest	0.78180	0.75675	0.80685
GradientBoosting	0.78160	0.75659	0.80660
AdaBoost	0.78155	0.75651	0.80658
Tweedie	0.78140	0.75649	0.80631
BayesianRidge	0.78093	0.75598	0.80588
Ridge	0.78085	0.75590	0.80580
LinearSVR	0.77857	0.75352	0.80362
Bagging	0.77715	0.75211	0.80218