O Paradoxo da Hibridização: O Impacto da Combinação de Algoritmos na Justiça e no Risco em Sistemas de Recomendação

João Paulo Prata Costa[1], **Daniel José Chaves Ferreira**[2], **Anderson A. Ferreira**[1,2], **Reinaldo Silva Fortes**[1,2]

[1]Departamento de Computação, Universidade Federal de Ouro Preto.

[2]Programa de Pós-Graduação em Ciência da Computação, Univ. Fed. de Ouro Preto.

{joao.prata, daniel.jcf}@aluno.ufop.edu.br

{anderson.ferreira, reifortes}@ufop.edu.br

Abstract. *This study investigates the impact of weighted hybridization in Recommender Systems from the perspectives of* ***fairness*** *and* ***risk-sensitiveness****. Employing a rigorous evaluation with 20 sliding time windows on the MovieLens dataset, we analyzed how combining models via regression affects the equity and robustness of suggestions. The results reveal a paradox: hybridization acted as a generic leveler, amplifying performance disparities between gender and activity groups while increasing the system's overall risk. These findings demonstrate that optimization focused on mean error minimization sacrifices fine-grained personalization, harming users with complex consumption profiles or those belonging to minority groups.*

Resumo. *Este trabalho investiga o impacto da hibridização ponderada em Sistemas de Recomendação sob as óticas de* ***justiça*** *(fairness) e* ***sensibilidade ao risco****. Empregando uma avaliação rigorosa com 20 janelas temporais deslizantes na base MovieLens, analisou-se como a combinação de modelos via regressão afeta a equidade e a robustez das sugestões. Os resultados evidenciam um fenômeno contraintuitivo: a hibridização atuou como um nivelador genérico, intensificando as disparidades de desempenho entre grupos e elevando o risco global do sistema. Tais achados demonstram que a otimização focada na minimização do erro médio sacrifica a personalização fina, prejudicando usuários com perfis complexos ou pertencentes a grupos minoritários.*

1. Introdução

A onipresença de conteúdos digitais em plataformas de *streaming* e *e-commerces* elevou a importância dos Sistemas de Recomendação (SR) como ferramentas essenciais para a filtragem de informações relevantes. Ao traçar perfis baseados em históricos de consumo e comportamento, esses sistemas buscam personalizar a experiência do usuário [Ma et al. 2011]. No entanto, abordagens isoladas como a Filtragem Colaborativa (FC) e a Filtragem Baseada em Conteúdo (CB) apresentam limitações como o problema de *Cold Start* e a superespecialização. Nesse cenário, a Filtragem Híbrida (FH) surge como uma estratégia para combinar diferentes algoritmos [Burke 2002], visando complementar forças e atenuar fraquezas individuais.

Contudo, além da precisão média global, a evolução dos SRs exige a consideração da ética algorítmica e da estabilidade das predições. O conceito de justiça (*fairness*) refere-se à capacidade do sistema de tratar entidades similares de forma não discriminatória [Pitoura et al. 2021], enquanto a sensibilidade ao risco foca na minimização de resultados severamente insatisfatórios [Wang et al. 2012]. A urgência dessa discussão é evidenciada por casos como o do algoritmo COMPAS, que demonstrou como vieses preditivos podem marginalizar grupos específicos [Angwin et al. 2016].

Diante da hipótese de que a hibridização possa atuar como um mecanismo de estabilização, este trabalho investiga os seguintes questionamentos: (1) *Técnicas de hibridização contribuem para resultados mais justos na recomendação?* (2) *Técnicas de hibridização contribuem para resultados menos sensíveis ao risco?*

Diferente de abordagens que utilizam modelos de linguagem complexos ou redes neurais profundas, este estudo investiga se a arquitetura de hibridização ponderada por regressão, amplamente utilizada para ganho de acurácia na predição, traz benefícios implícitos de justiça e robustez. A principal contribuição desta pesquisa reside em fornecer evidências empíricas de que o ganho de complexidade algorítmica com a hibridização não implica em melhorias éticas. Os resultados obtidos caracterizam o que denominamos **Paradoxo da Hibridização**, em que a busca pelo consenso estatístico na predição de notas acaba por marginalizar perfis comportamentais complexos. Tais achados demonstram que a otimização focada na minimização do erro médio sacrifica a personalização fina em prol de uma redução de erro global enviesada.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 discute os trabalhos relacionados; a Seção 4 descreve a metodologia experimental baseada em janelas temporais; a Seção 5 discute os resultados obtidos e a Seção 6 apresenta as considerações finais.

2. Fundamentação Teórica

Esta seção formaliza os pilares teóricos fundamentais para a análise desenvolvida neste Estudo. A Subseção 2.1 descreve os conceitos de justiça algorítmica de grupo e a métrica de Diferença Absoluta. A Subseção 2.2 apresenta a fundamentação sobre sensibilidade ao risco e a formalização matemática do *GeoRisk*. Por fim, a Subseção 2.3 detalha a arquitetura de hibridização ponderada fundamentada na técnica de *stacking*.

2.1. Justiça em Sistemas de Recomendação

A justiça (*fairness*) em SR pode ser analisada sob as perspectivas individual ou de grupo. Este estudo foca na **Justiça de Grupo**, que avalia se subconjuntos de usuários, definidos por atributos demográficos (como gênero) ou comportamentais (como nível de atividade), recebem recomendações com qualidade equiparável [Pitoura et al. 2021].

Uma métrica central é a **Diferença Absoluta (AD)**, que quantifica a disparidade entre as utilidades observadas para dois grupos distintos [Fu et al. 2020]:

$$AD = |f(G_0) - f(G_1)| \tag{1}$$

sendo que $f(G)$ representa o desempenho médio de um grupo em relação a uma métrica de qualidade (ex: F1 ou NDCG). A justiça do sistema é inversamente proporcional ao valor de AD, sendo o cenário ideal aquele em que $AD = 0$.

2.2. Sensibilidade ao Risco e Robustez

A otimização tradicional de SR foca na melhoria da eficácia média global, o que pode mascarar resultados severamente insatisfatórios para determinadas parcelas de usuários. A sensibilidade ao risco avalia a robustez do modelo, entendida como a capacidade de performar consistentemente em diferentes cenários [Wang et al. 2012].

Para mensurar essa robustez, adota-se o cálculo do **GeoRisk** (G_{RISK}), originalmente proposto por [Wang et al. 2012] e [Dinçer et al. 2014] no contexto de Recuperação de Informação, e adaptado para recomendação por [Fortes 2022]. Ele se baseia na função Z_{RISK}, que pondera ganhos e perdas de desempenho em relação a uma base de risco:

$$Z_{RISK}(x_j) = \sum_{u \in U^+} z(x_j, u) + (1 + \alpha) \sum_{u \in U^-} z(x_j, u) \tag{2}$$

sendo que U^+ representa o subconjunto de usuários para os quais o algoritmo avaliado obteve desempenho superior, enquanto U^- delimita o caso oposto, desempenho inferior; $z(x_j, u)$ representa o desvio de desempenho do algoritmo x_j para o usuário u em relação a um algoritmo de referência. O coeficiente α introduz uma penalidade assimétrica para resultados negativos. No contexto desta pesquisa, α é configurado para ser maior que zero, garantindo que quedas acentuadas de desempenho (resultados que alienam o usuário) tenham um peso maior na métrica do que ganhos incrementais de precisão.

A partir de Z_{RISK}, obtém-se o G_{RISK}:

$$G_{RISK}(x_j) = \sqrt{\frac{qi(x_j)}{|U|} \times \Phi\left(\frac{Z_{RISK}(x_j)}{|U|}\right)} \tag{3}$$

sendo que $qi(x_j)$ é o desempenho total e Φ representa a distribuição normal padrão acumulada. Valores elevados de G_{RISK} indicam sistemas mais estáveis e menos sensíveis ao risco extremo, sendo esta métrica essencial para sistemas em produção que buscam evitar o "pior cenário" para o usuário.

2.3. Hibridização ponderada

A hibridização ponderada é implementada, neste trabalho, por meio de uma arquitetura de *stacking* (empilhamento). Nessa abordagem, o processo é dividido em dois níveis [Burke 2002]: **Nível 0 (Modelos Base):** Algoritmos clássicos (ex: KNN, SVD, NMF) geram predições de notas; **Nível 1 (Meta-modelo):** Um regressor é treinado utilizando as predições do Nível 0 como variáveis de entrada (*features*) e a nota real como alvo (*target*).

O fundamento teórico do *stacking* reside na redução do viés e da variância: ao combinar modelos com diferentes vieses indutivos, o meta-modelo deveria, em tese, aprender os contextos em que cada constituinte falha, promovendo uma recomendação mais robusta e precisa.

3. Trabalhos Relacionados

A evolução dos Sistemas de Recomendação (SR) tem sido marcada por uma transição de métricas puramente focadas em erro para uma visão multidimensional que engloba ética e robustez. Esta seção contextualiza o presente estudo diante do estado da arte em justiça, risco e hibridização no contexto de SRs. Inicialmente, abordam-se pesquisas sobre justiça algorítmica. Em seguida, discute-se o uso de métricas sensíveis ao risco. E,por fim, analisam-se estudos que relacionam a hibridização à qualidade das sugestões.

Justiça Algorítmica em Recomendação. O estudo da justiça (*fairness*) em SRs ganhou relevância acadêmica ao identificar que algoritmos de filtragem colaborativa podem herdar e amplificar vieses presentes nos dados de treinamento [Pitoura et al. 2021]. Pesquisas recentes investigam como equilibrar a utilidade da recomendação com a equidade entre grupos de usuários [Burke et al. 2018, Deldjoo et al. 2023, Wang et al. 2023]. Surveys abrangentes documentam ainda que o *popularity bias* afeta grupos demográficos de forma assimétrica, intensificando disparidades de gênero em sistemas de filtragem colaborativa [Klimashevskaia et al. 2024, Boratto et al. 2025]. Enquanto a maioria desses trabalhos foca em modelos isolados ou propõe arquiteturas híbridas já orientadas à equidade [Farnadi et al. 2018], este estudo diferencia-se ao investigar se a camada de hibridização ponderada por regressão (comumente aplicada para ganho de acurácia, sem objetivos explícitos de *fairness*) atua efetivamente como um redutor de injustiça ou se, paradoxalmente, a agrava.

Sistemas de Recomendação Sensíveis ao Risco. A aplicação de métricas de sensibilidade ao risco em SRs é uma adaptação de conceitos consolidados na Recuperação de Informação (RI), onde se busca minimizar a variabilidade do desempenho [Wang et al. 2012, Dinçer et al. 2014]. No domínio de recomendação, o uso do *GeoRisk* permitiu avaliar a robustez de algoritmos clássicos frente a usuários com perfis atípicos [Fortes 2022]. Esta pesquisa expande tais investigações ao aplicar o escrutínio do risco sobre arquiteturas híbridas de dois níveis, verificando se a combinação de modelos base resulta em um sistema globalmente mais resiliente ou mais suscetível a falhas regionais.

Hibridização e o Trade-off de Desempenho. A hibridização ponderada via *stacking* foi formalizada como uma técnica capaz de mitigar problemas de *Cold Start* e esparsidade ao integrar diferentes vieses indutivos [Burke 2002]. Tradicionalmente, a literatura foca na melhoria de métricas de erro, como RMSE e MAE [Bao et al. 2009, Sill et al. 2009]. No entanto, pesquisas que relacionam hibridização a equidade e risco ainda são incipientes. Este trabalho preenche esta lacuna, desafiando a premissa de que a combinação de modelos melhora a qualidade geral ao evidenciar que a otimização focada no erro médio pode marginalizar perfis de usuários com maior variância comportamental.

4. Metodologia Experimental

Esta seção detalha os procedimentos, recursos e protocolos adotados para a condução dos experimentos. A Seção 4.1 descreve as características do conjunto de dados utilizado. A Subseção 4.2 detalha a estratégia de avaliação temporal via janelas deslizantes. A Subseção 4.3 apresenta os algoritmos e meta-modelos avaliados. A Subseção 4.4 define as métricas de qualidade. A Subseção 4.5 descreve os critérios de agrupamento de usuários, enquanto a Subseção 4.6 detalha o protocolo de validação estatística.

4.1. Dataset

Utilizou-se o *dataset MovieLens 1M*, que compreende 1.000.209 avaliações realizadas por 6.040 usuários sobre aproximadamente 3.900 filmes. As estatísticas descritivas estão sumarizadas na Tabela 1.

A escolha do *MovieLens 1M* em detrimento de versões maiores (como 20M ou 25M) foi estratégica e fundamentada na viabilidade da infraestrutura experimental. Dado o delineamento rigoroso desta pesquisa, que envolve 20 janelas temporais, a combinação

Tabela 1. Estatísticas descritivas do dataset MovieLens 1M.

Característica	Valor
Número de Usuários	6.040
Número de Itens (Filmes)	3.900 (aprox.)
Número de Avaliações	1.000.209
Escala de Notas	1 a 5
Densidade	4,24%
Esparsidade	95,76%

de múltiplos algoritmos de Nível 0 e Nível 1, a otimização de hiperparâmetros via *RandomizedSearchCV* e cinco execuções independentes para cada cenário, a utilização de conjuntos de dados massivos tornaria o tempo de experimentação proibitivo. O custo computacional cresceria exponencialmente, dificultando a realização das inúmeras iterações de treino e teste necessárias para garantir a robustez estatística dos resultados de justiça e risco. Além disso, *MovieLens 1M* contém informações demográficas dos usuários, como gênero, o que permitiu a estratificação em diferentes grupos para a mensuração da justiça.

4.2. Estratégia de Avaliação Temporal

Para garantir que a avaliação reflita a natureza dinâmica dos SRs e evite o viés de vazamento de dados (*data leakage*), adotou-se a estratégia de janelas de tempo deslizantes (*sliding windows*). A escolha justifica-se pela obsolescência natural das preferências: um modelo treinado em dados estáticos de anos anteriores falha em capturar tendências sazonais ou mudanças de gosto dos usuários [Quadrana et al. 2018, Ji et al. 2023].

O fluxo experimental foi dividido em **20 janelas temporais**, cada uma com duração total de 15 meses. O avanço entre janelas consecutivas foi fixado em 1 mês. Em cada recorte, os primeiros 12 meses foram destinados ao treinamento e os 3 meses subsequentes ao teste. Especificamente para o treinamento do Nível 1 (meta-modelo híbrido), aplicou-se uma subdivisão interna: as predições dos algoritmos constituintes foram geradas em um subconjunto de 9 meses de treino e 3 meses de validação, servindo como entrada (*features*) para o regressor. Este rigor metodológico assegura que o sistema aprenda a hibridizar com base em comportamentos recentes antes da avaliação final no conjunto de teste. A Figura 1 ilustra essa organização.

4.3. Algoritmos

Os experimentos foram implementados em Python utilizando a biblioteca *Scikit-learn*.O código-fonte e as instruções para reprodução estão publicamente disponíveis em `https://github.com/JoaoPauloPrata/sbbd-fairness-and-risk.git`.

Os algoritmos de Nível 0 (constituintes), *Stochastic itemKNN*, *NMF*, *BiasedSVD* e *BiasedMF*, foram selecionados pela variedade de estratégias. Limitou-se a 4 algoritmos devido à complexidade do processo experimental. Para o Nível 1 (híbridos), os meta-modelos foram selecionados por sua diversidade de abordagens estatísticas, que vão de modelos lineares a métodos de *ensemble*. A otimização de hiperparâmetros foi realizada via *RandomizedSearchCV* (3-*fold CV*) com 15 iterações. A Tabela 2 detalha os espaços de busca definidos para os regressores de Nível 1, garantindo que cada híbrido opere em seu ponto de máxima eficácia para a métrica alvo (MSE).

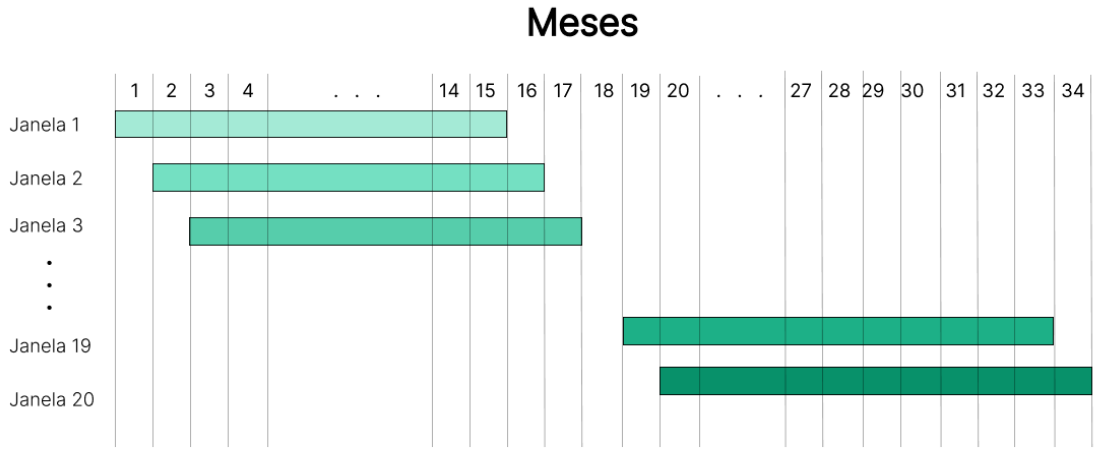


Figura 1. Divisão das janelas temporais ao longo dos meses.

4.4. Métricas de Qualidade

Quatro métricas foram adotadas para mensurar a precisão e o ranqueamento:

- **RMSE e MAE:** Avaliam o erro de predição. O RMSE (*Root Mean Squared Error*) penaliza erros de maior magnitude, enquanto o MAE (*Mean Absolute Error*) fornece uma média linear dos desvios [Herlocker et al. 2001].
- **F1-score:** Média harmônica entre precisão e revocação, utilizada para avaliar a eficácia do sistema em recomendar itens relevantes [Herlocker et al. 2001].
- **NDCG:** O *Normalized Discounted Cumulative Gain* mede a qualidade do ranqueamento, priorizando a presença de itens relevantes no topo da lista sugerida [Burke et al. 2018].

4.5. Segmentação de Grupos para Fairness

A avaliação de *fairness* investigou disparidades em duas dimensões:

1. **Gênero:** Segmentação binária (Masculino e Feminino) derivada dos metadados demográficos, buscando avaliar se a hibridização favorece a maioria demográfica do conjunto de dados.
2. **Atividade:** Aplicação dinâmica do algoritmo **K-Means** ($k = 2$) em cada janela temporal.O atributo utilizado para a clusterização foi a **quantidade total de interações** de cada usuário no conjunto de treino. Esta segmentação é vital para distinguir usuários "Mais Ativos" (perfis densos) de usuários "Menos Ativos" (perfis esparsos), permitindo observar se a complexidade algorítmica penaliza usuários com menos histórico.

Tabela 2. Espaço de busca para otimização de hiperparâmetros (Nível 1).

Algoritmo	Hiperparâmetro	Intervalo / Valores
Ridge / BayesianRidge	Alpha (α)	$[10^{-6}, 10^{1}]$ (escala log)
RandomForest	n_estimators / max_depth	$[50, 200]/[5, 20]$
GradientBoosting	learning_rate / max_depth	$[0.01, 0.2]/[3, 10]$
AdaBoost	n_estimators / learning_rate	$[50, 150]/[0.1, 1.0]$
LinearSVR	C / epsilon (ϵ)	$[0.1, 10]/[0.0, 0.2]$

4.6. Avaliação e Protocolo Estatístico

Para garantir a confiabilidade dos achados, cada experimento foi submetido a um protocolo estatístico rigoroso. A sensibilidade ao risco foi quantificada pelo *GeoRisk* (G_{RISK}) e a equidade pela Diferença Absoluta (AD).

Foram realizadas **5 execuções independentes** para cada uma das 20 janelas, totalizando uma amostra de $n = 100$ por algoritmo. Sobre este conjunto de dados, aplicou-se a Análise de Variância (**ANOVA**) para identificar se existiam diferenças globais significativas entre os modelos. Constatada a significância ($\alpha = 0.05$), utilizou-se o teste *post-hoc* de **Tukey HSD** para realizar comparações par a par. Este teste permite classificar os algoritmos em grupos de desempenho estatisticamente equivalentes, isolando os efeitos da hibridização de variações aleatórias.

5. Avaliação dos Resultados

Esta seção analisa o desempenho dos métodos sob as óticas de equidade e robustez. A Seção 5.1 apresenta a avaliação de *fairness* por gênero e atividade. A Seção 5.2 discute a sensibilidade ao risco via *GeoRisk*. A Seção 5.3 analisa o desempenho isolado e o paradoxo observado, enquanto a Seção 5.4 aprofunda a discussão sobre a redundância algorítmica e o conflito entre métricas.

Os resultados correspondem à média de 100 amostras experimentais, com intervalos de confiança de 95% e validação via teste de Tukey HSD. Embora as métricas de erro (RMSE e MAE) tenham sido processadas, elas não apresentaram variações estatisticamente significativas entre os modelos. No contexto de SR, tal cenário reforça a necessidade de priorizar métricas de ranqueamento (F1 e NDCG), visto que a utilidade prática do sistema reside na ordenação dos itens no topo da lista e não apenas na precisão pontual das predições. Dessa forma, a análise concentra-se em F1 e NDCG, que evidenciaram com maior clareza o impacto da hibridização na personalização e na justiça.

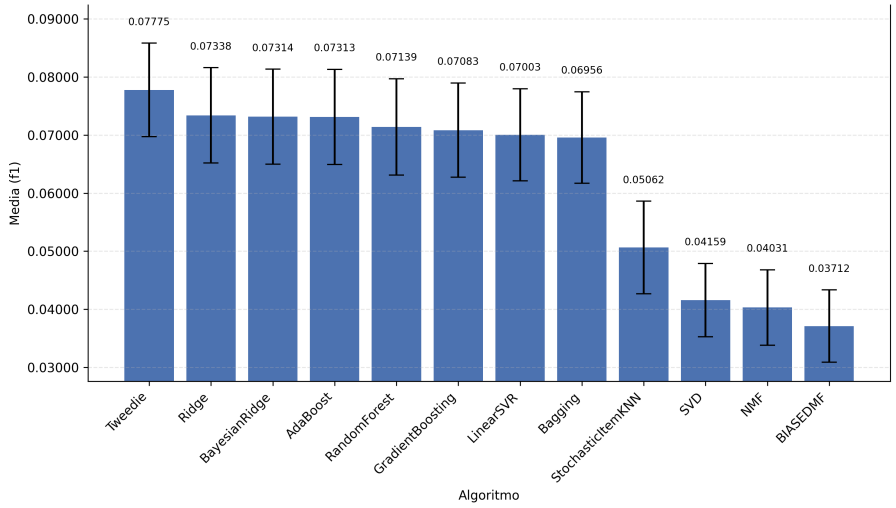
5.1. Justiça por Gênero e Atividade

A avaliação de *fairness* baseou-se na Diferença Absoluta (AD), onde menores valores indicam maior equidade. Na análise por **gênero**, os testes estatísticos revelaram que os algoritmos híbridos obtiveram resultados de *fairness* inferiores aos constituintes sob métricas de ranqueamento, conforme ilustrado na Figura 2. Enquanto os métodos clássicos mantiveram disparidades baixas, a hibridização ponderada ampliou a diferença de desempenho entre usuários masculinos e femininos.

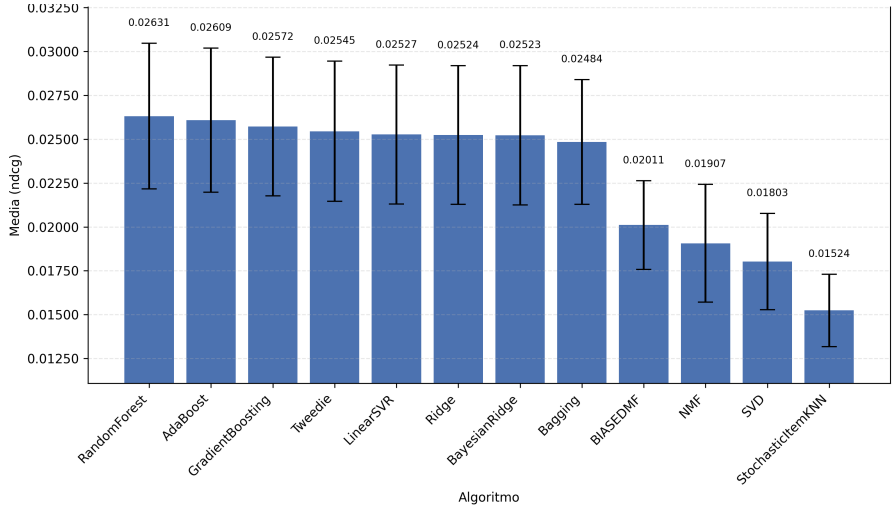
Padrão similar foi observado na análise por **atividade**, conforme ilustrado na Figura 3. Para a métrica NDCG, os métodos constituintes foram significativamente mais justos que todos os modelos híbridos. Os resultados refutam a hipótese de que a hibridização atuaria como mecanismo de calibração para usuários menos ativos, uma vez que os híbridos apresentaram disparidades superiores.

5.2. Sensibilidade ao Risco (GeoRisk)

A robustez dos sistemas foi mensurada pelo G_{RISK}, onde valores maiores indicam menor sensibilidade a quedas drásticas de desempenho. Conforme ilustrado na Figura 4, os algoritmos constituintes (*BiasedMF* e *SVD*) se destacaram como os mais robustos.



(a) *Fairness* Gênero (F1).



(b) *Fairness* Gênero (NDCG).

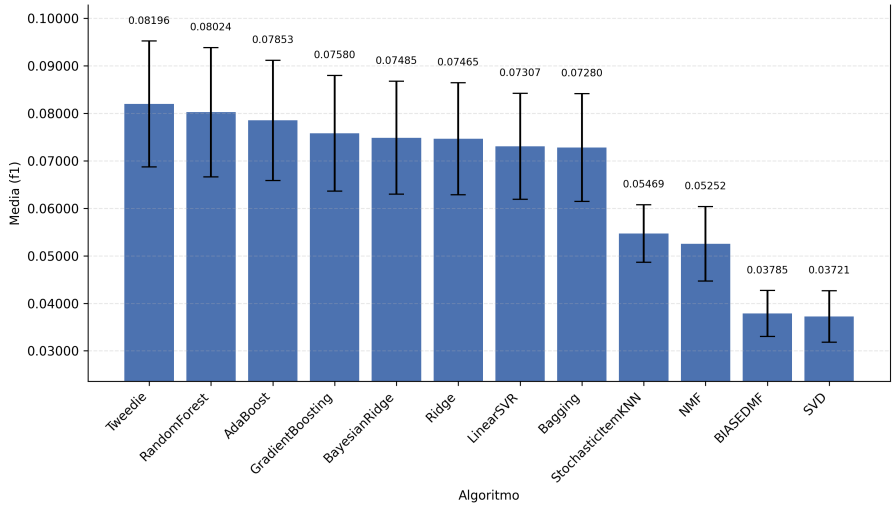
Figura 2. Resultados de *Fairness* por Gênero para métricas de ranqueamento.

A hibridização ponderada resultou em valores de G_{RISK} estatisticamente inferiores em quase todas as janelas, sugerindo que a camada de regressão, ao ser otimizada para minimizar o MSE global, torna-se excessivamente sensível a variações nas predições dos modelos base. Em cenários de incerteza (janelas com mudanças bruscas de catálogo ou comportamento), o híbrido falha de forma mais severa que os modelos isolados, aumentando o risco de oferecer recomendações irrelevantes para subgrupos de usuários.

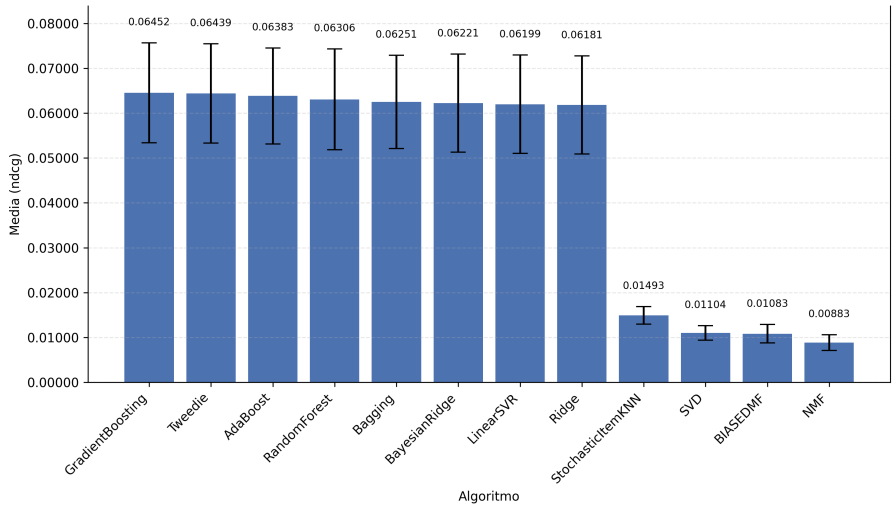
5.3. Análise do Paradoxo: O Nivelador Genérico

Para compreender a degradação na justiça e na robustez, avaliou-se o desempenho isolado dos grupos, conforme ilustrado na Figura 5. Observou-se que o aumento da injustiça nos híbridos decorreu de uma degradação generalizada que afetou os perfis de forma assimétrica. Na análise de atividade, o impacto negativo foi severamente mais acentuado no grupo de usuários “Mais Ativos”.

Este comportamento pode ser explicado pela natureza estatística da hibridização



(a) *Fairness* Atividade (F1).

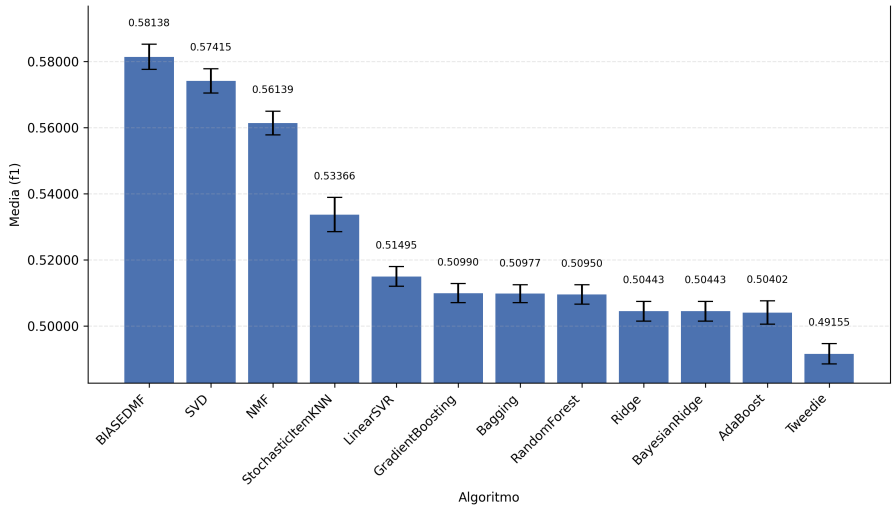


(b) *Fairness* Atividade (NDCG).

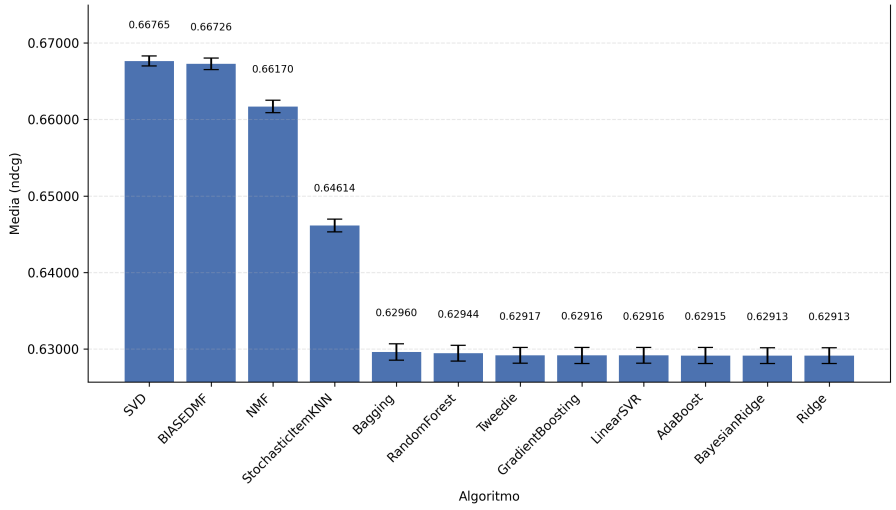
Figura 3. Resultados de *Fairness* por Atividade para métricas de ranqueamento.

ponderada. Ao tentar minimizar o erro global (MSE), o meta-modelo de Nível 1 atua como um **nivelador genérico**, ajustando as predições para uma média de consenso (regressão à média). Usuários altamente ativos tendem a possuir perfis ecléticos, cujas preferências geram predições com maior variância nos modelos base do Nível 0.

O regressor de hibridização, ao buscar a redução do erro médio, acaba por tratar essa variância necessária para a personalização fina como "ruído estatístico". Consequentemente, o modelo híbrido sacrifica o atendimento a esses perfis complexos em prol de uma maior estabilidade no grupo majoritário (usuários menos ativos e mais previsíveis). Este efeito revela um *trade-off* crítico: a busca pela redução do erro global em modelos híbridos pode, paradoxalmente, reforçar o *status quo* demográfico e comportamental dos dados, atuando como um mecanismo que silencia a diversidade de interesses em favor de um padrão médio de consumo.



(a) *GeoRisk* (F1).



(b) *GeoRisk* (NDCG).

Figura 4. Resultados de *GeoRisk* para as métricas de ranqueamento.

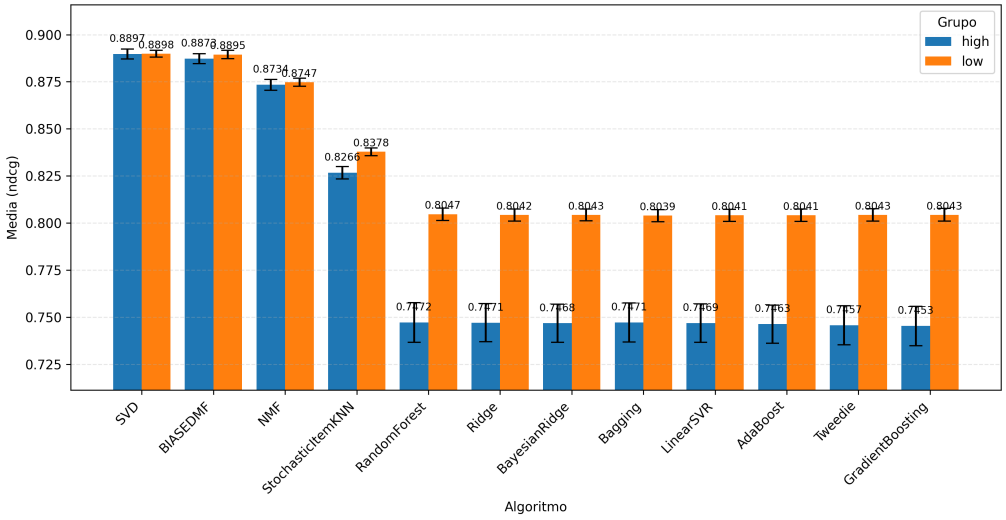


Figura 5. Desempenho isolado por nível de atividade (NDCG).

5.4. Discussão: Redundância e o Conflito Erro vs. Ranqueamento

Um fator determinante para o insucesso da hibridização ponderada neste cenário é a **redundância algorítmica**. A análise dos pesos atribuídos pelos regressores revelou que o meta-modelo frequentemente prioriza algoritmos cujos erros já são correlacionados. Como resultado, a hibridização não consegue extrair informações complementares dos modelos de Nível 0, o que justifica o **empate estatístico** observado nas métricas de erro absoluto (RMSE e MAE). A complexidade adicional da hibridização ponderada não se traduziu em ganho de precisão, mantendo o desempenho médio no mesmo patamar dos constituintes isolados.

Além disso, identifica-se um conflito fundamental entre o objetivo de otimização do Nível 1 e a qualidade final da recomendação por ranqueamento. Mesmo quando a hibridização atinge um erro médio estatisticamente equivalente ao dos modelos base, o processo de minimização do erro médio quadrático (MSE) altera a distribuição das predições. Para "tentar" aproximar-se da nota real, o regressor acaba por suavizar os valores preditos, o que frequentemente subverte a ordem relativa dos itens na lista final.

Esse fenômeno é crítico: para o SR, a precisão da nota absoluta é menos importante do que a integridade do ranqueamento. Ao buscar reduzir um erro pontual que não resulta em ganho estatístico, a hibridização degrada o NDCG e o F1-score, especialmente para usuários com perfis ecléticos. Assim, o sistema híbrido falha duplamente: não supera os modelos base em eficácia média e os prejudica sistematicamente em termos de equidade e robustez.

6. Considerações Finais

Este estudo investigou o impacto das técnicas de hibridização ponderada em Sistemas de Recomendação sob a ótica de métricas de justiça (*fairness*) e sensibilidade ao risco (*GeoRisk*). Através de uma avaliação experimental rigorosa com 20 janelas temporais deslizantes, foi possível observar o comportamento de modelos de regressão ao combinarem algoritmos clássicos de filtragem colaborativa.

As evidências obtidas permitem responder às perguntas de pesquisa levantadas: (1) ***A hibridização contribui para resultados mais justos?*** Os resultados indicam que **não**. A hibridização ponderada por regressão ampliou a Diferença Absoluta (AD) entre grupos. Na análise por atividade, o ranqueamento degradou-se mais severamente entre usuários altamente ativos, enquanto na análise demográfica, os modelos híbridos elevaram o erro global e mantiveram vieses estatísticos contra o público feminino; (2) ***A hibridização contribui para resultados menos sensíveis ao risco?*** Os dados demonstram que **não**. Os valores de G_{RISK} revelaram que algoritmos híbridos são mais suscetíveis a quedas acentuadas de desempenho para subconjuntos de usuários, apresentando menor robustez do que métodos isolados como *BiasedMF* ou *SVD*.

A principal contribuição deste trabalho é a demonstração teórica e empírica de que a hibridização por *stacking* pode atuar de forma deletéria sobre a personalização. Ao formalizarmos o fenômeno como um efeito de nivelamento genérico, alertamos a comunidade de SR para o fato de que a busca pelo consenso estatístico sacrifica a diversidade de interesses, resultando em um sistema estruturalmente menos justo e mais suscetível a variações de risco.

Como principais limitações, reconhecem-se o uso de apenas um conjunto de dados, a ausência de análise de correlação entre os algoritmos constituintes, o que pode ter mascarado redundâncias preditivas, e a ausência de outros modelos híbridos. Para trabalhos futuros, sugere-se: (1) Ampliação dos conjuntos de dados, dos algoritmos constituintes, dos *baselines* híbridos e das análises comparativas de resultados; (2) A exploração de arquiteturas híbridas em cascata; (3) A inclusão de critérios de diversidade na seleção de modelos base; (4) O desenvolvimento de estratégias de otimização multiobjetivo que incorporem explicitamente termos de justiça e risco na função de perda dos regressores.

Agradecimentos

Os autores agradecem a Universidade Federal de Ouro Preto (UFOP) e o Departamento de Computação (DECOM/UFOP), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), e o INCT-TILDIAR (grant # 408490/2024-1).

Uso de ferramentas de Inteligência Artificial

Os autores declaram o uso de ferramentas de inteligência artificial generativa para a revisão gramatical e o aprimoramento da fluidez textual deste artigo.

Referências

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica*.

Bao, X., Bergman, L., and Thompson, R. (2009). Stacking Recommendation Engines with Additional Meta-features. In *ACM RecSys*, pages 109–116, New York, NY, USA.

Boratto, L. et al. (2025). Popularity bias in recommender systems: The search for fairness in the long tail. *Information*, 16(2).

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Burke, R., Sonboli, N., and Ordonez-Gauger, A. (2018). Balanced neighborhoods for multi-sided fairness in recommendation. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214. PMLR.

Deldjoo, Y., Jannach, D., Bellogin, A., Difonzo, A., and Zanzonelli, D. (2023). Fairness in recommender systems: Research landscape and future directions. *User Modeling and User-Adapted Interaction*.

Dinçer, B. T., Ounis, I., and Macdonald, C. (2014). Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In *Proceedings of the 36th European Conference on Information Retrieval*, pages 26–38. Springer.

Farnadi, G., Kouki, P., Thompson, S. K., Srinivasan, S., and Getoor, L. (2018). A fairness-aware hybrid recommender system. *CoRR*, abs/1809.09030.

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

Fu, Z., Xian, Y., Gao, R., Zhao, J., Huang, Q., Ge, Y., Xu, S., Geng, S., Shah, C., Zhang, Y., and de Melo, G. (2020). Fairness-aware explainable recommendation over knowledge graphs.

Herlocker, J., Konstan, J. A., Borchers, A., and Riedl, J. (2001). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 19(1):5–53.

Ji, Y., Sun, A., Zhang, J., and Li, C. (2023). A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems (TOIS)*, 41(3).

Klimashevskaia, A., Jannach, D., Elahi, M., and Trattner, C. (2024). A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*.

Ma, H., Zhou, D., Liu, C., Lyu, M. R., and King, I. (2011). Recommender systems with social regularization. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 287–296, New York, NY, USA. Association for Computing Machinery.

Pitoura, E., Stefanidis, K., and Koutrika, G. (2021). Fairness in rankings and recommendations: an overview. *The VLDB Journal*, pages 651–654.

Quadrana, M., Cremonesi, P., and Jannach, D. (2018). Sequence-aware recommender systems. *ACM Computing Surveys (CSUR)*, 51(4):1–36.

Sill, J., Takacs, G., Mackey, L., and Lin, D. (2009). Feature-Weighted Linear Stacking. *arXiv:0911.0460 [cs]*.

Wang, L., Bennett, P. N., and Collins-Thompson, K. (2012). Robust ranking models via risk-sensitive optimization.

Wang, Y., Chen, J., et al. (2023). A survey on fairness-aware recommender systems. *Information Fusion*.