Journal on Interactive Systems, 2025, 16:1, doi: 10.5753/jis.2025.5407

Group Fairness in Recommendation Systems: The Importance of Hierarchical Clustering in Identifying Latent Groups in MovieLens and Amazon Books

Rafael Vargas Mesquita dos Santos [**Federal Institute of Espírito Santo** | ***rafaelv@ifes.edu.br***]
Giovanni Ventorim Comarela [**Federal University of Espírito Santo** | ***gc@inf.ufes.br***]

Institute of Computing, Federal Institute of Espírito Santo, Cachoeiro de Itapemirim, ES, 29322-000, Brazil.

Received: 08 January 2025 • **Accepted:** 01 July 2025 • **Published:** 12 July 2025

Abstract: Fairness in recommendation systems is a critical area of study, particularly when addressing group disparities based on sensitive attributes such as gender, age, activity levels, or user location. This study also explores latent groups identified through hierarchical clustering techniques. The goal is to assess group unfairness across various clustering configurations and collaborative filtering strategies to promote equitable and inclusive recommendation systems. We applied collaborative filtering techniques, including ALS, KNN, and NMF, and evaluated group unfairness using metrics such as R_{grp} for different clustering configurations (e.g., gender, age, activity level, location, and hierarchical clustering) in two datasets: MovieLens and Amazon Books. Hierarchical clustering yielded the highest group unfairness, with ALS and NMF reaching R_{grp} values of 0.0062 and 0.0049 in MovieLens, and NMF and KNN peaking at 0.0972 and 0.0220 in Amazon Books. These results reveal significant fairness disparities across both latent and observable user groups, reinforcing the importance of selecting appropriate filtering strategies and clustering methods to build fair and inclusive recommendation systems.

Keywords: Recommendation System, Group Fairness, Agglomerative Hierarchical Clustering, Latent Groups

1 Introduction

In recent years, digital transformation has revolutionized the way we interact with the world around us. Digital interfaces have rapidly evolved from static systems to dynamic and personalized experiences that accommodate the nuances of individual user interests. In this context, recommendation systems stand out as a crucial innovation, playing a vital role in guiding choices and reinforcing interactions on online platforms.

In a scenario where digital interfaces are becoming increasingly interactive and personalized, recommendation systems emerge as essential tools, significantly shaping the choices and interactions of users on online platforms. The more we know about the user, the better the quality of the items recommended to them [Hazrati and Ricci, 2024; Pereira *et al*., 2018; Cavalcante and Fettermann, 2019; Bernardino and Gonçalves, 2019].

Since their origins in basic suggestions based on collaborative filtering, these systems have evolved to incorporate sophisticated methods utilizing machine learning and artificial intelligence. This progress has allowed companies to offer users increasingly tailored experiences, enhancing customer engagement and satisfaction.

Recommendation systems provide item suggestions to their users, currently being incorporated into e-commerce sites, digital libraries, and social networks [Liu *et al*., 2024; Souza *et al*., 2022]. The proliferation of these systems highlights the urgent need to evaluate and mitigate potential adverse social repercussions that may emerge, especially as they deeply integrate into social networks and digital environments.

The growing complexity of these systems has also raised concerns about their ethical implications. As they become more integrated into everyday life, it is essential to assess how they influence decisions and behaviors, potentially amplifying existing biases and social inequalities.

Recent investigations, such as those conducted by Deldjoo *et al*. [2024], inadvertently reveal that recommendation systems can intensify biases and inequalities, thus creating disparities in the service offered to different segments of the population. This finding underscores the complexity of social interactions mediated by such systems and the importance of promoting equitable and fair practices.

The growing reliance on these systems and their influence on the dynamics of social networks require a detailed examination of their impacts, aimed at developing solutions that ensure an equitable and inclusive digital space. Therefore, the need to transcend the traditional accuracy metric in recommendation systems is highlighted, incorporating fairness as a crucial parameter in evaluating the impact of these technologies on society.

This study proposes an innovative approach, integrating fairness metrics into recommendation algorithms, with the objective of elucidating existing inequalities. Through detailed analysis applied to the MovieLens dataset, we consider sociodemographic and behavioral variables, such as gender, age, and frequency of item ratings, along with the use of hierarchical clustering methods to reveal latent groups. The inclusion of the latter method aims to identify potentially latent groups, whose characteristics are not readily apparent, aiming for a more holistic and inclusive approach in examining the social fairness disparities manifested in the system's recommendations.

2 Related Work

This section begins by presenting the definitions of fairness that underpin this study, providing an overview of the concepts that guide the analysis. It will address the different perspectives on fairness, setting the stage for the discussion on how it is applied and evaluated in recommendation systems. Additionally, the contributions of this article will be described, highlighting the innovative aspects of the research. By clarifying these definitions and contributions, the goal is to provide a framework that guides the rest of the study and emphasizes the importance of fairness in the development and implementation of machine learning models.

Fairness has become a topic of growing interest in the field of machine learning. In this context, a recommendation system is considered fair if it ensures uniformity in the quality of service (i.e., prediction accuracy) for all individuals or user groups [Zafar *et al*., 2017; Rahmani *et al*., 2022; Sonboli *et al*., 2021; Wang *et al*., 2023; Li *et al*., 2023; Santos and Comarela, 2024]. Generally, definitions of fairness can be classified into two categories: individual fairness and group fairness.

Individual fairness pertains to the quality of recommendations made to each user, ensuring they are relevant and fair, based on the interactions and preferences of each individual. Thus, a system that promotes individual fairness ensures that all users receive high-quality recommendations, avoiding algorithmic discrimination that may occur due to personal characteristics unrelated to item selection [Dwork *et al*., 2011; Wu *et al*., 2021; Li *et al*., 2021].

On the other hand, group fairness focuses on the quality of recommendations offered to different user groups, analyzing how each group is treated by the system [Dwork *et al*., 2011; Ekstrand *et al*., 2022; Friedler *et al*., 2016]. The use of clustering techniques, such as hierarchical classification, allows for identifying patterns of unequal treatment that may not be evident at first glance [Alves *et al*., 2024b; Jáñez-Martino *et al*., 2023]. This analysis can reveal inequalities in the quality of recommendations, enabling adjustments in the recommendation process to ensure that all user groups receive equitable recommendations.

We can summarize the contributions of this work as follows:

- We analyzed group fairness using, among others, the agglomerative clustering method (hierarchical method) to group users. This differs from previous studies [Liu *et al*., 2022; Rastegarpanah *et al*., 2019; Santos *et al*., 2024], which focused on clustering users based on a single sensitive attribute, such as gender, race, or age;
- We evaluated three distinct collaborative filtering strategies in recommendation systems. Similar to the study proposed by Leonhardt *et al*. [2018], the intentional choice of varied methods, including model-based and memory-based approaches, aims to comprehensively compare and evaluate different collaborative filtering techniques. However, we introduced an additional contribution by incorporating two model-based strategies for the comparative analysis of different approaches to handling missing data.

3 Material and Methods

This section provides a detailed explanation of the methodology employed in this study, which includes an in-depth analysis of the MovieLens dataset. We apply various collaborative filtering techniques to explore and evaluate social fairness measures in recommendation systems. The primary objective is to understand how these algorithms perform across different user groups, specifically with respect to fairness, by examining the variation in recommendations given to users from diverse social categories. The methodology incorporates a range of techniques, from data preprocessing to algorithmic evaluation, with a focus on comparing group fairness across different recommendation algorithms, particularly considering latent groups, highlighting their effectiveness and limitations in promoting social fairness.

3.1 Datasets

This case study used two public datasets well-known in the field of recommendation systems.

The first, **MovieLens 1M**[1] Harper and Konstan [2015], contains approximately 1 million ratings for about 4000 movies, assigned by 6000 users on a 1-to-5-star scale.

The second dataset, **Amazon Books**[2] Bagchi [2022], is a subset of book data available on Amazon. The full dataset comprises records of 20000 books, 30000 users, and a total of 35080 ratings.

To ensure a balance between statistical power and computational efficiency, a sampling methodology was applied to both datasets, following the approach proposed by Rastegarpanah *et al*. [2019]. For each dataset, we randomly selected 300 users and the 1000 items (movies or books) with the highest number of ratings, thus ensuring sufficient density in the interaction matrix for the experiments.

3.2 Recommendation Strategies

To estimate unknown ratings, three different collaborative filtering strategies were tested in recommendation systems, each with distinct characteristics. Specifically, methods with different approaches were chosen to evaluate and compare the behavior of these techniques:

- **ALS** *(Alternating Least Squares)*: This model-based method is employed in matrix factorization, minimizing the quadratic error alternately by fixing one factor at a time. It uses boolean masks to directly ignore missing values during optimization;
- **NMF** *(Non-negative Matrix Factorization)*: Also model-based, it uses iterative optimization techniques with a non-negativity constraint. It requires filling in missing values with averages or other imputations before optimization, as it cannot process them directly;
- **KNN** *(K-Nearest Neighbors)*: On the other hand, KNN is a memory-based method centered on similarity between users or items. For a given user or item, the

[1] `https://grouplens.org/datasets/movielens/`

[2] `https://www.kaggle.com/datasets/saurabhbagchi/books-dataset`

system identifies the k nearest neighbors (where k is an integer) and makes recommendations based on these neighbors' preferences;

This review highlights the intentional choice of methods with distinct approaches, incorporating the differentiation between model-based methods (ALS and NMF) and memory-based (KNN). Additionally, it emphasizes the importance of analyzing the handling of missing data, which is particularly relevant for fairness assessments in results.

3.3 Hyperparameter Optimization

The following hyperparameters were optimized in the recommendation strategies used. These parameters were chosen to ensure a balance between accuracy and computational efficiency:

- **ALS** *(Alternating Least Squares)*: The hyperparameter 'rank' was set to 20, representing the number of latent factors, and 'lambda', also at 20, acts as regularization to prevent overfitting.
- **NMF** *(Non-negative Matrix Factorization)*: The 'components' was set to 5, determining the number of latent factors used to approximate the user-item interaction matrix. Additionally, the algorithm was configured with 'max_iter' set to 200, which limits the number of iterations, and 'threshold' set to 10^{-4}, defining the convergence tolerance.
- **KNN** *(K-Nearest Neighbors)*: The optimized hyperparameters include 'k' equal to 5, specifying the number of nearest neighbors considered during recommendations. The algorithm also used 'max_iter' set to 200 to limit the number of iterations and a convergence 'threshold' set to 10^{-5} to determine when the optimization process should stop.

3.4 User Clustering

Users were clustered based on age, number of ratings submitted (activity), and agglomerative clustering, in addition to one demographic attribute specific to each dataset: gender for MovieLens, and location for Amazon Books. Grouping users based on these characteristics helps identify behavioral patterns, as these factors often influence preferences. It is important to note that individuals within the same network tend to form stable social groups, sharing similar behaviors and preferences over time, as suggested by previous studies Coelho *et al*. [2023].

The choice between gender and location as a clustering criterion stems from the information available in each dataset. MovieLens provides user gender data, while Amazon Books does not include this information, but does provide geographic data, allowing segmentation by location (North America, Europe, and other territories). This adaptation ensures that the analyses respect the specificities of each dataset, maintaining methodological consistency.

Agglomerative clustering was adopted to investigate the influence of the correlation between multiple user attributes — such as age, location/gender, and activity — on group unfairness. By combining these variables in an unsupervised manner, this type of clustering enables the identification of latent groups whose composition may result in disparities not evident in traditional segmentation methods.

The imbalance observed in demographic variables — such as gender or location — reflects the asymmetric distributions naturally found in real-world platforms, as documented by Ekstrand *et al*. [2018] and Chen *et al*. [2018]. This approach aligns with methodological recommendations on representativeness in the evaluation of recommender systems Beutel *et al*. [2019], allowing algorithms to be tested under conditions that reflect the challenges faced in real-world applications.

- Gender: male and female;
- Location: users grouped into three geographic regions based on their declared origin — North America (NA), Europe (EU), and Other Territories (OT), which includes Asia, Oceania, South America, and Africa;
- Age: under 18, 18 to 24, 25 to 34, 35 to 44, 45 to 49, 50 to 55, and over 55;
- Activity (95-5): one group containing the top 5% of users with the highest number of ratings, and the remaining 95% in another group. The 5% group represents active users, while the 95% group represents inactive users;
- Agglomerative clustering[3]: the optimal number of groups was determined based on silhouette score analysis[4]. A hierarchical method was used to cluster users into five distinct groups, considering characteristics such as gender, age, and number of ratings. The method employed the Euclidean distance metric (*metric='euclidean'*) and Ward's linkage criterion (*linkage='ward'*) to minimize the variance within each group. The goal was to identify clusters that are not immediately evident. The K-Means clustering method[5] (Lloyd's algorithm) was also tested; however, for the comparison presented in this paper, agglomerative clustering was chosen, as it resulted in higher levels of group unfairness.

Table 1 shows the number of users in each group for the different clustering configurations in the MovieLens dataset. Table 2 presents the corresponding configurations for the Amazon Books dataset. These distributions provide insight into how users are categorized based on location, age, activity level, and agglomerative clustering. The analysis of these clusters helps to understand the diversity of the user base and supports decisions related to recommendations, fairness, and system performance.

3.5 Algorithm Module: Fairness Measures

In this subsection, we present the algorithm module developed to calculate social fairness measures for the proposed case study. It is pertinent to mention that all implementations

[3]https://scikit-learn.org/stable/modules/clustering#hierarchical-clustering

[4]https://scikit-learn.org/stable/modules/clustering#silhouette-coefficient

[5]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans

Table 1. User clustering configurations and group composition — using the MovieLens dataset

Clustering	Groups	Quantity
Gender	Male	240
	Female	60
Age	Under 18	5
	18 to 24	53
	25 to 34	142
	35 to 44	58
	45 to 49	24
	50 to 55	12
	Over 55	6
Activity	Active	15
	Inactive	285
Agglomerative	Group 1	146
	Group 2	69
	Group 3	42
	Group 4	25
	Group 5	18

Table 2. User clustering configurations and group composition — using the Amazon Books dataset

Clustering	Groups	Quantity
Location	NA (North America)	258
	EU (Europe)	26
	OT (Other Territories)	16
Age	Group 1	17
	Group 2	47
	Group 3	114
	Group 4	58
	Group 5	20
	Group 6	15
	Group 7	29
Activity	Active	15
	Inactive	285
Agglomerative	Group 1	34
	Group 2	187
	Group 3	37
	Group 4	41

of the fairness measures used in the proposed fairness algorithm were based on the work of Rastegarpanah *et al*. [2019], providing a solid foundation for our approach to addressing social fairness in recommendation systems.

The implementation of all group fairness analysis codes on the MovieLens dataset is available in the repository ravarmes/recsys-rgrp-movielens[6] on GitHub.

Considering all the specifications and discussions from the previous section, we will formally define the metrics that specify the objective functions associated with individual fairness and group fairness.

We will start by presenting the system configuration, notation, and problem definition. Suppose $X \in \mathbb{R}^{n \times m}$ is a partially observed rating matrix of n users and m items, such that the element $\mathbf{x}_{ij}$ denotes the rating given by user i for item j. Let Ω be the set of indices of known ratings in X. Moreover, Ω_i denotes the indices of known item ratings for user i, and Ω_j denotes the indices of known user ratings for item j.

For a matrix A, $P_\Omega(A)$ is a matrix whose elements in $(i,j) \in \Omega$ are a_{ij} and zero elsewhere. Similarly, for a vector a, $P_{\Omega_j}(a)$ is a vector whose elements in $i \in \Omega_j$ are the corresponding elements of a and zero elsewhere. Throughout the article, we denote column j of A by the vector a_j and row i of A by the vector a^i. All vectors are column vectors.

Given a traditional recommendation system, an estimated recommendation matrix $\hat{X} = [\hat{X}_{ij}]_{n \times m}$ is generated. In this recommendation problem, we assume users in a set $\{u_1, u_2, ..., u_n\}$ and items in a set $\{v_1, v_2, ..., v_m\}$.

3.5.1 Individual Fairness

For each user i, we define ℓ_i, the user's loss for i, as the mean squared error estimate over the known ratings of user i. And individual unfairness R_{indv} as the variation of users' losses.

$$\ell_i = \frac{||P_{\Omega^i}(\hat{\mathbf{x}}^i - \mathbf{x}^i)||_2^2}{|\Omega^i|} \tag{1}$$

$$R_{indv}(X, \hat{X}) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l>k} (\ell_k - \ell_l)^2 \tag{2}$$

3.5.2 Group Fairness

For each group g of users, we define the group's loss ℓ_g, calculated as the mean squared error estimate over the known ratings of users in group g. The social fairness measure R_{grp} is defined as the variation of group losses. Let G denote the number of groups, where each group is indexed by g. Then we have:

$$L_i = \frac{||P_{\Omega_{G_i}}(\hat{X} - X)||_2^2}{|\Omega_{G_i}|} \tag{3}$$

$$R_{grp}(X, \hat{X}, G) = \frac{1}{g^2} \sum_{k=1}^{g} \sum_{l>k} (L_k - L_l)^2 \tag{4}$$

3.6 Overview of Methodological Process

Figure 1 presents a flowchart that outlines the general steps of the methodology used for analyzing group fairness in the MovieLens dataset. At the start of the process, we have the partially filled rating matrix ($\mathbf{X}$), which serves as input to traditional recommendation algorithms, such as ALS (*Alternating Least Squares*), NMF (*Non-negative Matrix Factorization*), and KNN (*K-Nearest Neighbors*).

Each of these traditional recommendation algorithms computes an estimated recommendation matrix ($\hat{\mathbf{X}}$). Subsequently, users are studied in four distinct groupings: gender, age, activity level, and through agglomerative clustering.

The developed algorithm is applied by calculating an effectiveness value, Root Mean Square Error (RMSE), as well as values for Group Unfairness R_{grp} for each estimated matrix $\hat{\mathbf{X}}$ within each cluster.

This procedure enables the identification of algorithms and groupings that exhibit greater or lesser fairness from the perspective of groups. Additionally, it provides a detailed analysis of the groups within the clusters, assessing which of

[6] https://github.com/ravarmes/recsys-rgrp-movielens

Table 3. Comparison of group unfairness (R_{grp}) and prediction error (RMSE) across recommendation strategies and clustering criteria — using the MovieLens dataset

Strategy	Grouping	R_{grp}	RMSE
ALS	Activity	0.00129660	0.8751394
	Age	0.00170270	
	Gender	0.00426530	
	Agglomerative	0.00615080	
NMF	Activity	0.00400160	0.8335659
	Age	0.00161200	
	Gender	0.00301780	
	Agglomerative	0.00487470	
KNN	Activity	0.00195270	1.0338668
	Age	0.00767110	
	Gender	0.00053500	
	Agglomerative	0.00304140	

Table 4. Comparison of group unfairness (R_{grp}) and prediction error (RMSE) across recommendation strategies and clustering criteria — using the Amazon Books dataset

Strategy	Grouping	R_{grp}	RMSE
ALS	Activity	0.0025600	0.8751394
	Age	0.0226210	
	Location	0.0536636	
	Agglomerative	0.0221096	
NMF	Activity	0.0500996	0.8335659
	Age	0.0245520	
	Location	0.0861890	
	Agglomerative	0.0972347	
KNN	Activity	0.0158433	1.0338668
	Age	0.0208155	
	Location	0.0178347	
	Agglomerative	0.0219661	

them receive the most or least favorable recommendations within the respective groupings.

4 Results and Discussions

In this section, we present a comprehensive analysis of group fairness metrics across different collaborative filtering strategies and user grouping configurations. The analyses were conducted on two widely used datasets in recommender systems: **MovieLens** and **Amazon Books**. The obtained results are illustrated through visualizations comparing the performance of the strategies according to measures of group unfairness (R_{grp}) and group losses (L_i).

Figures 2 and 3 present the group unfairness (R_{grp}) for the MovieLens and Amazon Books datasets, respectively, separated by filtering strategy and grouping configuration. Figures 4 and 5 show the group losses (L_i) in each domain; Figures 6 and 7 display the distribution of group unfairness; and finally, Figures 8 and 9 provide a direct comparison of unfairness across recommendation strategies. This parallel organization allows us to verify whether the fairness patterns observed in MovieLens remain consistent in the Amazon Books scenario, which presents a distinct domain and a different user profile.

The quantitative data from these analyses are summarized in Tables 3 and 4, which present the group unfairness values (R_{grp}) and RMSE, respectively for the MovieLens and Amazon Books datasets. The highest R_{grp} values for each recommendation algorithm are highlighted in red, facilitating the identification of the most unfavorable configurations in terms of fairness.

These results enable a comparative assessment of the fairness of the strategies, controlling for accuracy through similar RMSEs. This aligns with the main objective of this study, which is to evaluate the impact of recommendation strategies on group fairness, regardless of overall predictive performance.

The comparative analysis of Figure 2 and Figure 3 reveals a recurring and noteworthy pattern. The Agglomerative Clustering approach, which identifies latent groups based on behavioral patterns, consistently points to high levels of unfairness in both datasets. It is observed that, for both MovieLens (with the ALS and NMF strategies) and Amazon Books (with NMF and KNN), the agglomerative clustering configuration resulted in the highest disparity indices in two out of the three evaluated filtering strategies. Additionally, when considering the average unfairness across strategies for each dataset, this clustering method also stands out with the highest values. This observation suggests that significant disparities may be concentrated in latent groups, highlighting the relevance of investigating such hidden structures beyond predefined demographic groups when analyzing fairness in recommendation systems.

Still regarding the results from Figures 2 and 3, it is important to note that we also tested the K-means algorithm, a partitional alternative to the hierarchical agglomerative clustering method used in the main analyses. For the MovieLens dataset, the group unfairness levels obtained with K-means were lower than those observed with the agglomerative method, across the ALS, NMF, and KNN algorithms (K-means: [0.0027223, 0.0025223, 0.0036480]; Agglomerative: [0.0061508, 0.0048747, 0.0030414]). Similarly, in the Amazon Books dataset, unfairness values were also lower with K-means for most algorithms (K-means: [0.0156997, 0.0855969, 0.0015300]; Agglomerative: [0.0221096, 0.0972347, 0.0219661]). Therefore, we chose to emphasize the agglomerative method in our analyses, as it more clearly exposed the disparities between groups, providing a more suitable scenario to investigate the mechanisms of algorithmic unfairness.

Tables 5 and 6 present a sensitivity analysis of the k-Nearest Neighbors (KNN) algorithm for different values of the k parameter, examining its impact on both recommendation accuracy and algorithmic fairness. This analysis is essential to understand how the choice of parameter k simultaneously affects the predictive performance and fairness of the recommendation system.

We chose $k = 5$ as the optimal value for our in-depth analysis of algorithmic unfairness, both for the MovieLens and Amazon Books datasets. In the case of MovieLens, this value yields the highest disparity in the activity-based grouping (R_{grp}, Activity $= 0.0020$), while also maintaining significant disparity levels in the age (0.0077), gender (0.0005), and agglomerative (0.0030) groupings, with an RMSE of 1.0339. Similarly, in Amazon Books, $k = 5$ also stands out with the highest disparity value for age grouping (R_{grp}, Age $=$

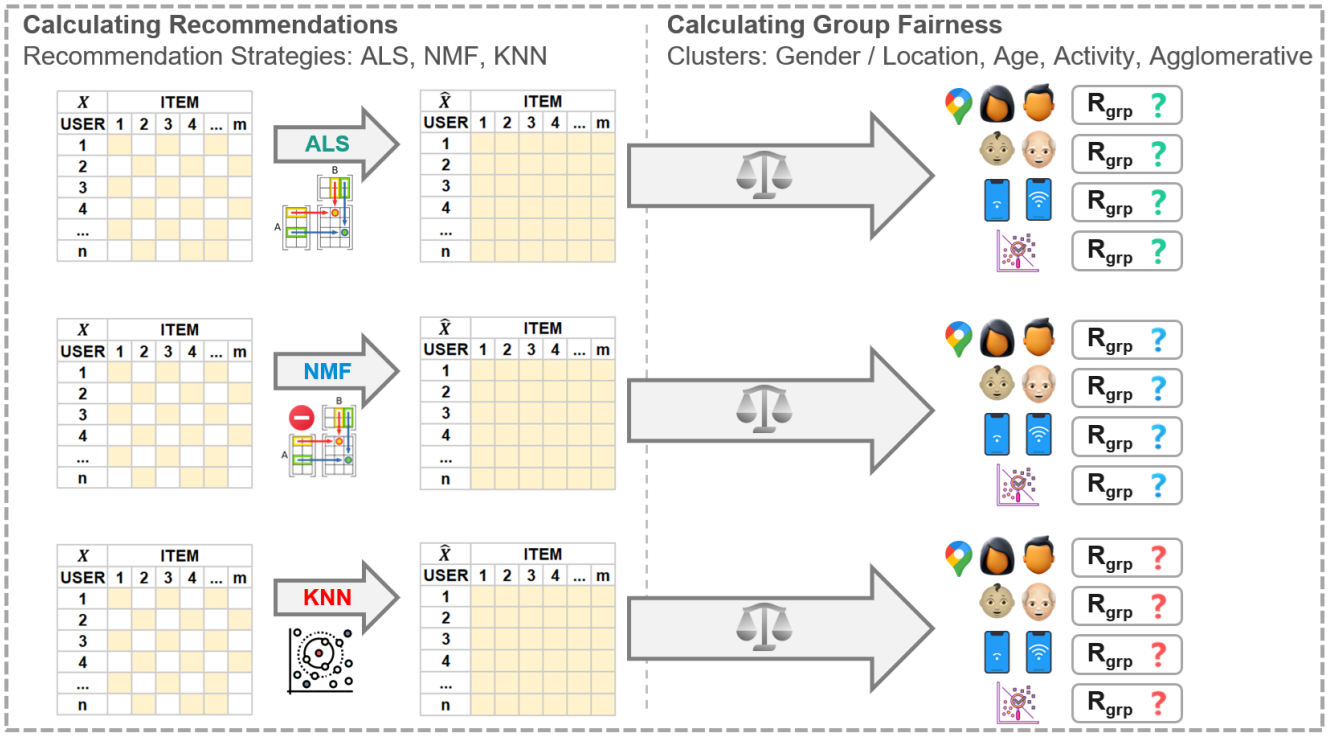

Figure 1. Flowchart of the methodological framework for evaluating group fairness in recommendation systems. The proposed evaluation is applied to the MovieLens and Amazon Books datasets

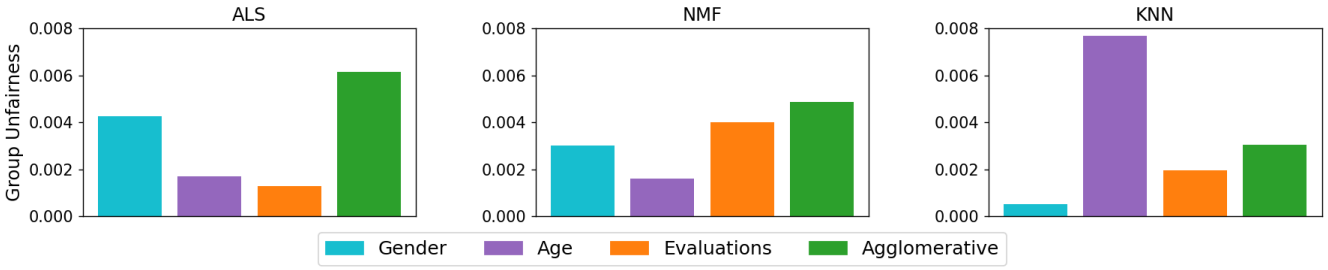

Figure 2. Analysis of group unfairness (R_{grp}) by filtering strategy and user grouping — using the MovieLens dataset

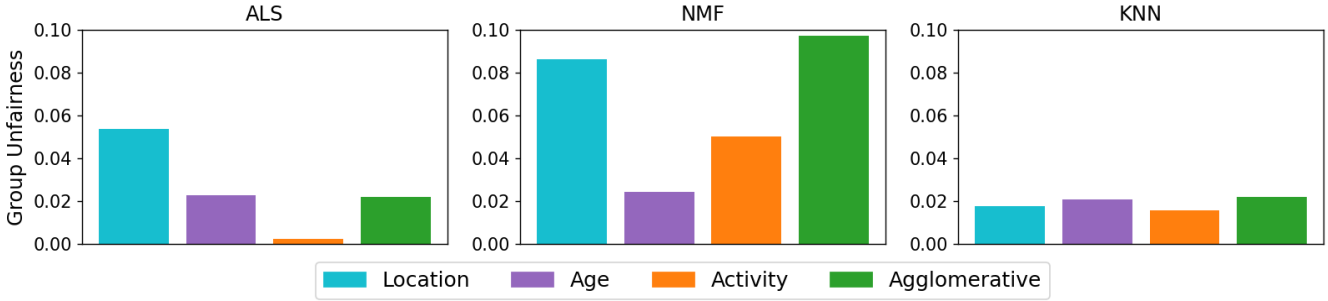

Figure 3. Analysis of group unfairness (R_{grp}) by filtering strategy and user grouping — using the Amazon Books dataset

0.0208), while maintaining high disparity levels for activity (0.0158) and location (0.0178), and a competitive RMSE of 0.5730. In both datasets, $k = 5$ offers a good balance between accuracy and sensitivity to group variations, avoiding the overfitting observed at $k = 3$ and the performance degradation associated with higher k values. Therefore, this configuration proves to be appropriate for investigating algorithmic discrimination mechanisms and testing mitigation strategies across different data contexts.

Tables 7 and 8 present the group unfairness values (R_{grp}) for different clustering configurations, ranging from 3 to 7 clusters, in the ALS, NMF, and KNN algorithms, considering the MovieLens and Amazon Books datasets. The results indicate that the number of clusters significantly affects the fairness of recommendations, with considerable variations in R_{grp} values across the evaluated configurations. In some cases, configurations with fewer clusters exhibited lower levels of unfairness, as observed in the NMF algorithm with 3 clusters in the MovieLens dataset (0.0028). Conversely, configurations with 5 or more clusters showed higher levels of unfairness, such as the values of 0.0062 for ALS on MovieLens and 0.0371 for KNN on Books, both with 5 or 7 clusters.

Despite these variations, the final clustering configurations — 5 clusters for MovieLens and 4 clusters for Ama-

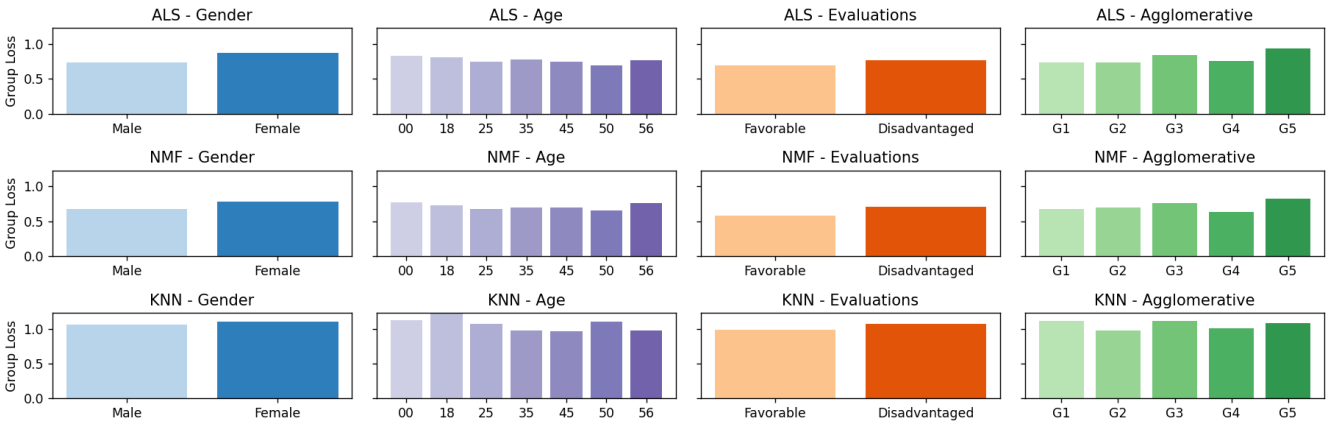

Figure 4. Group loss values (L_i) by filtering strategy and user grouping — using the MovieLens dataset

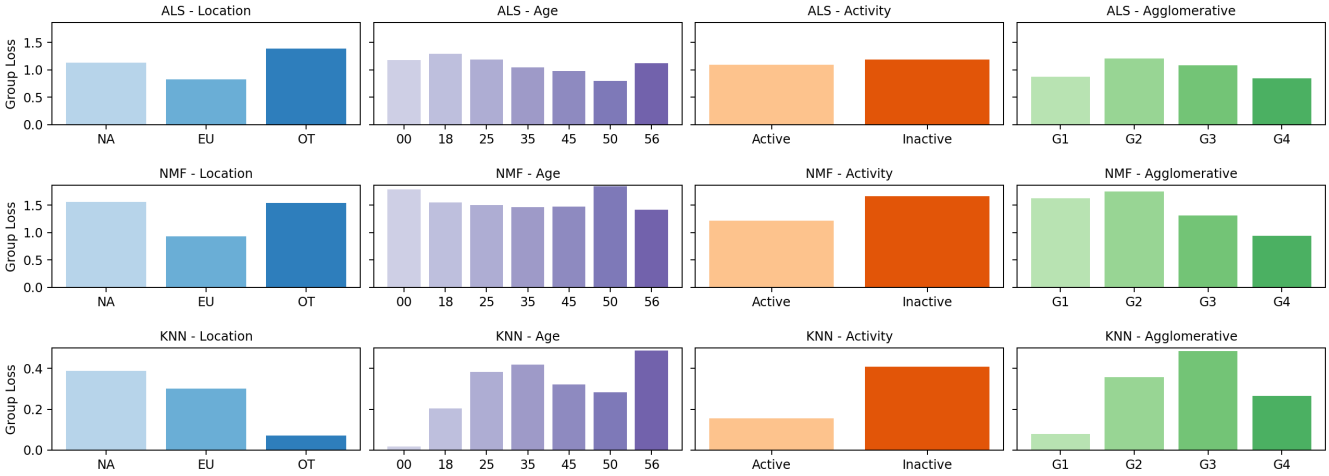

Figure 5. Group loss values (L_i) by filtering strategy and user grouping — using the Amazon Books dataset

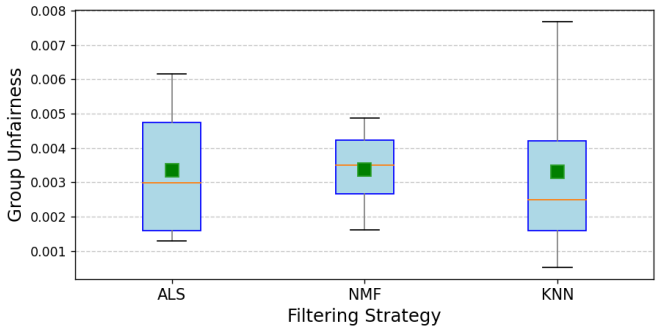

Figure 6. Distribution of group unfairness across filtering strategies — using the MovieLens dataset

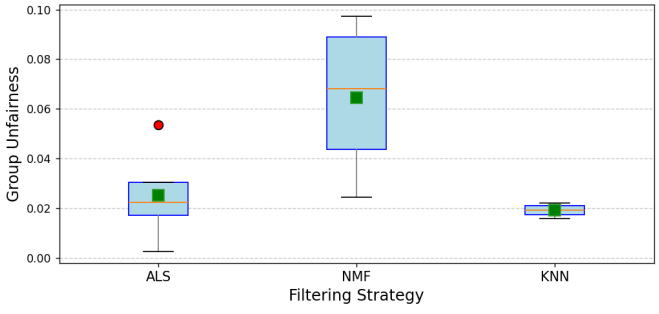

Figure 7. Distribution of group unfairness across filtering strategies — using the Amazon Books dataset

zon Books — were primarily based on the analysis of the Silhouette Index (Section 3.4), which indicated these config-

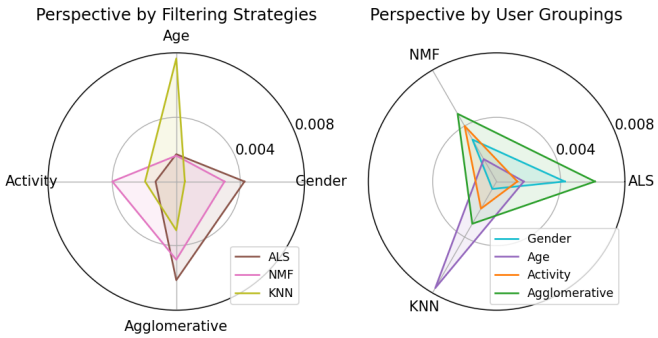

Figure 8. Comparative analysis of group unfairness among filtering strategies — using the MovieLens dataset

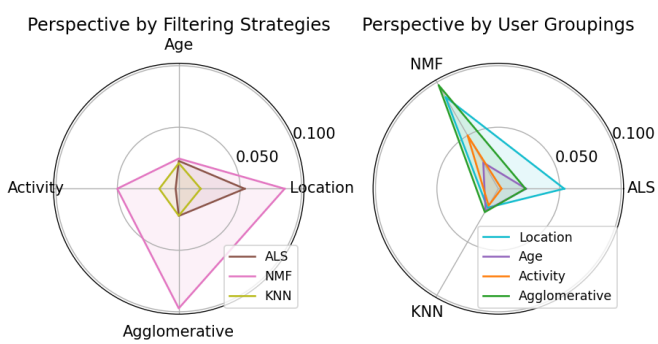

Figure 9. Comparative analysis of group unfairness among filtering strategies — using the Amazon Books dataset

Table 5. Sensitivity analysis of the neighborhood size parameter (k) in the KNN algorithm on group fairness metrics across different user clustering strategies — using the MovieLens dataset

k	KNN R_{grp}				RMSE
	Activity	Age	Gender	Agglom.	
3	0.0013	0.0082	0.0001	0.0021	0.9542
5	**0.0020**	**0.0077**	**0.0005**	**0.0030**	**1.0339**
7	0.0006	0.0082	0.0012	0.0035	1.0832
10	0.0000	0.0114	0.0010	0.0035	1.1320
15	0.0002	0.0095	0.0011	0.0031	1.1862

Table 6. Sensitivity analysis of the neighborhood size parameter (k) in the KNN algorithm on group fairness metrics across different user clustering strategies — using the Amazon Books dataset

k	KNN R_{grp}				RMSE
	Activity	Age	Location	Agglom.	
3	0.0195	0.0192	0.0172	0.0091	0.5628
5	**0.0158**	**0.0208**	**0.0178**	**0.0110**	**0.5730**
7	0.0155	0.0234	0.0187	0.0144	0.5813
10	0.0109	0.0289	0.0193	0.0222	0.5922
15	0.0070	0.0306	0.0161	0.0353	0.6031

urations as the most suitable in terms of internal cohesion and group separation. In the case of MovieLens, although the 5-cluster configuration showed higher unfairness in some algorithms, it also provided a clearer segmentation of user profiles, which can make existing inequalities among groups more evident. For the Books dataset, besides satisfactory performance in terms of unfairness, the 4-cluster configuration showed the most balanced distribution of users per group among all tested configurations, favoring a more representative and robust analysis. Thus, the chosen configurations reflect a compromise between segmentation quality, fairness in recommendations, and stability in result interpretation.

Figures 4 and 5 provide a detailed view of group unfairness, presenting the group losses L_i for each group within the clustering configurations for the MovieLens and Amazon Books datasets, respectively. In particular, for the MovieLens dataset, the groups 'Women' and 'Inactive' presented the highest losses across the ALS, NMF, and KNN methods. Similarly, in the Amazon Books dataset, the 'Inactive' group also exhibited high losses, as did user groups from North America ('NA') and the age group '56+', indicating a concerning disparity in the equitable allocation of recommendations in both scenarios. Additionally, the Agglomerative Clustering scenario in both datasets reveals that, although the KNN algorithm presented greater group fairness (more uniform loss distribution), it often displayed lower accuracy. This observation highlights the complexity and critical importance of finding an appropriate balance between fairness and accuracy in recommender systems.

Figures 6 and 7 offer a detailed view of the distribution of group unfairness for the ALS, KNN, and NMF collaborative filtering strategies using different clustering configurations. Through boxplots, one can observe variations, medians, and the presence of outliers in the unfairness measurements, providing a comparative analysis across strategies. This visual representation highlights differences in each filtering strategy's performance in terms of fairness, enabling a more precise evaluation of each approach's effectiveness in promoting fairer and more inclusive recommender systems.

Table 7. Sensitivity analysis on the influence of the number of user clusters over group fairness metrics across recommendation algorithms — using the MovieLens dataset

Clusters	Agglomerative R_{grp}		
	ALS	NMF	KNN
3	0.0038	0.0028	0.0031
4	0.0030	0.0031	0.0034
5	**0.0062**	**0.0049**	**0.0030**
6	0.0058	0.0042	0.0029
7	0.0053	0.0036	0.0040

Table 8. Sensitivity analysis on the influence of the number of user clusters over group fairness metrics across recommendation algorithms — using the Amazon Books dataset

Clusters	Agglomerative R_{grp}		
	ALS	NMF	KNN
3	0.0181	0.1045	0.0087
4	**0.0221**	**0.0972**	**0.0220**
5	0.0232	0.0897	0.0110
6	0.0243	0.1050	0.0117
7	0.0298	0.0900	0.0371

Figures 6 and 7 reveal notable variations in group unfairness across strategies. In MovieLens, KNN exhibits the greatest variability, which aligns with its reliance on local data. However, for the Amazon Books dataset, the NMF strategy demonstrates drastically greater dispersion, with a higher median and a wider interquartile range, suggesting significant instability in fairness for this scenario. In contrast, KNN on Amazon Books presents the lowest variability, being the most consistent. The presence of an outlier in the ALS strategy for Amazon Books highlights the need to investigate specific configurations that may lead to atypical levels of unfairness, while ALS and NMF in MovieLens appear more stable due to their ability to capture global patterns and apply regularization.

Figures 8 and 9 present a comparative analysis of group unfairness through two side-by-side radar charts, offering distinct perspectives on collaborative filtering strategies (ALS, KNN, NMF) and their interactions with different user clusterings. The first chart illustrates how each filtering strategy impacts unfairness levels across different clusterings, providing a direct view of variations across strategies. On the other hand, the second chart highlights the influence of each user clustering on the filtering strategies, showing each strategy's sensitivity to different clusterings. Together, these charts provide a holistic and detailed view of the dynamics between strategies and clusterings, emphasizing the complexity of promoting fairness in recommender systems.

In Figures 8 and 9, we expand the analysis to a dual visual perspective. In MovieLens, the KNN strategy shows great variability, with a peak in unfairness for the 'Age' clustering. In Amazon Books, NMF is the most unbalanced strategy, exhibiting extremely high unfairness for the 'Agglomerative' clustering. Shifting focus, the second chart in each figure explores the influence of the clusterings. For both datasets, Agglomerative Clustering stands out as the most challenging, generating the highest levels of unfairness, especially for NMF (Amazon Books) and ALS (MovieLens). While in

MovieLens the 'Age' clustering was the most sensitive, in Amazon Books it was the 'Location' clustering that caused the greatest disparity among filtering strategies, demonstrating how the demographic characteristics of each dataset directly influence fairness outcomes.

Given the complexity of the addressed scenario, it is essential to emphasize the importance of carefully choosing the recommendation algorithm, considering how it interacts with different user clusterings to ensure fair and equitable recommendations. This perspective aligns with the work of Abdollahpouri *et al*. [2020], who demonstrated how high-accuracy algorithms can amplify inequalities among user groups. The chosen approach significantly affects users' perception of fairness, impacting acceptance and satisfaction with the system, as evidenced by Alves *et al*. [2024a].

User clustering analysis is fundamental when addressing group unfairness, as it reveals correlations between variables by identifying latent groups. Our study advances significantly in applying intersectional analysis to recommender systems, originally proposed by Crenshaw [1989] and adapted to algorithmic contexts by Burke *et al*. [2018] and Ekstrand *et al*. [2018]. Unlike these works, which consider intersectional groups as direct combinations of known attributes, our approach innovates by employing hierarchical clustering to uncover hidden grouping patterns. To illustrate hypothetically, we may consider scenarios where significant disparities emerge in latent groups discovered through clustering, transcending traditional demographic categories. This type of phenomenon would be consistent with Sonboli *et al*. [2020], but would extend their findings by revealing injustices invisible to predefined intersectional analyses.

Our methodological contribution lies in integrating unsupervised clustering techniques with fairness metrics and in systematically comparing predefined and latent groups. This approach is relevant considering Zehlike *et al*. [2022], who demonstrated how conventional bias mitigation methods often fail by not considering complex interactions between attributes. Our empirical results show that groups identified by hierarchical clustering often present higher unfairness than groups defined by traditional attributes, with techniques like ALS and NMF exhibiting significantly higher R_{grp} values in these groups. Our methodology provides a more robust framework for identifying subtle aspects of unfairness invisible in conventional analyses, promoting significant advances in implementing more equitable recommender systems.

5 Conclusion

The analysis of the results provides a deep understanding of the interaction between collaborative filtering strategies and clustering configurations. Firstly, the variability in group unfairness among the filtering strategies (ALS, KNN, and NMF) emphasizes the importance of careful selection of the strategy to promote fairness in recommendations. It is observed that different strategies exhibit varied performances concerning group unfairness, with KNN displaying distinct behavior in certain clusters, such as Gender and Age, standing out in the comparative analysis.

Significantly, the clustering approach using the Agglomerative method revealed the highest levels of group unfairness in most of the filtering strategies evaluated. This outcome highlights the inherent complexity of recommendation systems and the need for a comprehensive evaluation that considers the interactions among all involved variables.

The results indicate the necessity for a more tailored approach in the design of recommendation systems. Incorporating social fairness considerations from conception through implementation of such systems is essential. It is recommended that developers and researchers investigate filtering strategies and clustering methods aligned with fairness objectives, aiming to create recommendation systems that are not only efficient but also fair and inclusive.

These conclusions underscore the intrinsic complexity in the pursuit of fair recommendation systems and point to promising areas for future research. One direction to explore is conducting multivariate analyses to examine the interactions among the various variables that characterize users, thereby contributing to the enhancement of fairness and inclusion on digital platforms.

Declarations

Acknowledgements

We would like to thank Federal Institute of Espírito Santo (IFES) and Federal University of Espírito Santo (Ufes) for their academic and financial support.

Authors' Contributions

Rafael V. M. S. was responsible for writing the original draft, research, methodology application, data collection and curation, as well as conducting the formal analysis and translating this work into English. Giovanni V. C. supervised the project, contributing to data analysis and providing insights into the results, as well as editing, correcting, and suggesting improvements for this document. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets used, as well as all implementations and results, are available in the repository: `https://github.com/ravarmes/recsys-rgrp-movielens-jis`. Last access on 12 July 2025.

Citation Diversity Statement

In this work, our references were selected based solely on the quality and relevance of the publications. The identity of the authors, including their sex, country, or any other characteristics, was not considered in the selection process. By focusing on the merit of the work, we aim to promote a fair and unbiased scholarly environment.

References

Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., and Pizzato, L. (2020). Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30. DOI: https://doi.org/10.1007/s11257-019-09256-1.

Alves, G., Jannach, D., Ferrari De Souza, R., and Manzato, M. G. (2024a). User perception of fairness-calibrated recommendations. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '24, page 78–88, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3627043.3659558.

Alves, P., Martins, A., Negrão, F., Novais, P., Almeida, A., and Marreiros, G. (2024b). Are heterogeinity and conflicting preferences no longer a problem? personality-based dynamic clustering for group recommender systems. *Expert Systems with Applications*, 255:124812. DOI: https://doi.org/10.1016/j.eswa.2024.124812.

Bagchi, S. (2022). Books Dataset: Books, Users and Ratings. Kaggle.

Bernardino, G. S. and Gonçalves, A. L. (2019). A education profile model applied in the context of recommender systems. *IEEE Latin America Transactions*, 17(3):505–512. DOI: https://doi.org/10.1109/TLA.2019.8863321.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., and Goodrow, C. (2019). Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2212–2220. DOI: https://doi.org/10.1145/3292500.3330745.

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.

Cavalcante, C. G. S. and Fettermann, D. C. (2019). Recommendations for product development of intelligent products. *IEEE Latin America Transactions*, 17(10):1645–1652. DOI: https://doi.org/10.1109/TLA.2019.8986442.

Chen, J., Feng, N., Bernstein, M. S., and Zhu, H. (2018). Investigating the impacts of gender representation in recommendation. In *Workshop on Recommendation in Multi-stakeholder Environments at the ACM RecSys*.

Coelho, N. L. L., Figueiredo, T. F., and Figueiredo, R. (2023). Uso de programação linear inteira para geração e análise de agrupamentos de políticos da câmara dos deputados. In *BraSNAM 2023 - XII Brazilian Workshop on Social Network Analysis and Mining*, João Pessoa, PB, Brasil. Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/brasnam.2023.230858.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(1):139–167. DOI: https://doi.org/10.4324/9780429499142-5.

Deldjoo, Y., Jannach, D., Bellogin, A., Difonzo, A., and Zanzonelli, D. (2024). Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, 34(1):59–108. DOI: https://doi.org/10.1007/s11257-023-09364-z.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2011). Fairness through awareness.

Ekstrand, M. D., Das, A., Burke, R., and Diaz, F. (2022). Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16(1–2):1–177. DOI: https://doi.org/10.1561/1500000079.

Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., and Pera, M. S. (2018). All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. 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 172–186. PMLR.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016). On the (im)possibility of fairness.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):Article 19. DOI: https://doi.org/10.1145/2827872.

Hazrati, N. and Ricci, F. (2024). Choice models and recommender systems effects on users' choices. *User Model User-Adap Inter*, 34(1):109–145. DOI: https://doi.org/10.1007/s11257-023-09366-x.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., and Alegre, E. (2023). Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. *Applied Soft Computing*, 139:110226. DOI: https://doi.org/10.1016/j.asoc.2023.110226.

Leonhardt, J., Anand, A., and Khosla, M. (2018). User fairness in recommender systems. *CoRR*, abs/1807.06349. DOI: https://doi.org/10.1145/3184558.3186949.

Li, Y., Chen, H., Xu, S., Ge, Y., Tan, J., Liu, S., and Zhang, Y. (2023). Fairness in recommendation: Foundations, methods and applications.

Li, Y., Chen, H., Xu, S., Ge, Y., and Zhang, Y. (2021). Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1054–1063, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3404835.3462966.

Liu, S., Ge, Y., Xu, S., Zhang, Y., and Marian, A. (2022). Fairness-aware federated matrix factorization. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 168–178, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3523227.3546771.

Liu, S., Sun, J., Deng, X., and et al. (2024). Towards platform profit-aware fairness in personalized recommendation. *Int. J. Mach. Learn. & Cyber.*. DOI: https://doi.org/10.1007/s13042-024-02149-9.

Pereira, F. S. F., Linhares, C. D. G., Ponciano, J. R., Gama, J., de Amo, S., and Oliveira, G. M. B. (2018). That's my jam! uma análise temporal sobre a evolução das preferências dos usuários em uma rede social de músicas. In *BraSNAM 2018 - VII Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, RS, Brasil. Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/brasnam.2018.3587.

Rahmani, H. A., Deldjoo, Y., Tourani, A., and Naghiaei, M. (2022). The unfairness of active users and popularity bias in point-of-interest recommendation.

Rastegarpanah, B., Gummadi, K. P., and Crovella, M. (2019). Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19. ACM. DOI: https://doi.org/10.1145/3289600.3291002.

Santos, R. and Comarela, G. (2024). Development of an equity strategy for recommendation systems. In *Anais do V Workshop sobre as Implicações da Computação na Sociedade*, pages 24–35, Porto Alegre, RS, Brasil. SBC. DOI: https://doi.org/10.5753/wics.2024.1975.

Santos, R. V. M., Comarela, G. V., and Junior, M. C. G. L. (2024). Implementation of fairnessmeasures: A case study in the cultural context for different strategies in recommendation systems. *iSys- Revista Brasileira de Sistemas de Informação*, 17:1–16. DOI: https://doi.org/10.5753/isys.2024.4214.

Sonboli, N., Eskandanian, F., Burke, R., Liu, W., and Mobasher, B. (2020). Opportunistic multi-aspect fairness through personalized re-ranking. UMAP '20, page 239–247, New York, NY, USA. Association for Computing Machinery. DOI: https://doi.org/10.1145/3340631.3394846.

Sonboli, N., Smith, J. J., Cabral Berenfus, F., Burke, R., and Fiesler, C. (2021). Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 274–279. ACM. DOI: https://doi.org/10.1145/3450613.3456835.

Souza, E., Lichtnow, D., and Gasparini, I. (2022). Estratégia de pós-processamento aplicada a um sistema de recomendação de artigos para a melhora da diversidade. In *BraSNAM 2022 - XI Brazilian Workshop on Social Network Analysis and Mining*, pages 216–221, Porto Alegre, RS, Brasil. Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/brasnam.2022.222805.

Wang, Y., Ma, W., Zhang, M., Liu, Y., and Ma, S. (2023). A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.*, 41(3). DOI: https://doi.org/10.1145/3547333.

Wu, Y., Cao, J., Xu, G., and Tan, Y. (2021). Tfrom: A two-sided fairness-aware recommendation model for both customers and providers.

Zafar, M., Valera, I., Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW '17: Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. DOI: https://doi.org/10.1145/3038912.3052660.

Zehlike, M., Yang, K., and Stoyanovich, J. (2022). Fairness in ranking: A survey. *ACM Computing Surveys*, 55(6):1–34. DOI: https://doi.org/10.1145/3533379.