information

MDPI

Article

Bias Assessment Approaches for Addressing User-Centered Fairness in GNN-Based Recommender Systems

Nikzad Chizari [1,*], Keywan Tajfar [2] and María N. Moreno-García [1,*]

[1] Department of Computer Science and Automation, Science Faculty, University of Salamanca, Plaza de los Caídos sn, 37008 Salamanca, Spain
[2] College of Science, School of Mathematics, Statistics, and Computer Science, Department of Statistics, University of Tehran, Tehran 1417935840, Iran
* Correspondence: nikzadchizari@usal.es (N.C.); mmg@usal.es (M.N.M.-G.)

Abstract: In today's technology-driven society, many decisions are made based on the results provided by machine learning algorithms. It is widely known that the models generated by such algorithms may present biases that lead to unfair decisions for some segments of the population, such as minority or marginalized groups. Hence, there is concern about the detection and mitigation of these biases, which may increase the discriminatory treatments of some demographic groups. Recommender systems, used today by millions of users, are not exempt from this drawback. The influence of these systems on so many user decisions, which in turn are taken as the basis for future recommendations, contributes to exacerbating this problem. Furthermore, there is evidence that some of the most recent and successful recommendation methods, such as those based on graphical neural networks (GNNs), are more sensitive to bias. The evaluation approaches of some of these biases, as those involving protected demographic groups, may not be suitable for recommender systems since their results are the preferences of the users and these do not necessarily have to be the same for the different groups. Other assessment metrics are aimed at evaluating biases that have no impact on the user. In this work, the suitability of different user-centered bias metrics in the context of GNN-based recommender systems are analyzed, as well as the response of recommendation methods with respect to the different types of biases to which these measures are addressed.

Keywords: recommender systems; GNN (graph neural network); bias amplification; fairness; sensitive features

check for updates

Citation: Chizari, N.; Tajfar, K.; Moreno-García, M.N. Bias Assessment Approaches for Addressing User-Centered Fairness in GNN-Based Recommender Systems. *Information* **2023**, *14*, 131. https://doi.org/10.3390/info14020131

Academic Editors: Willy Susilo, Jun Hu, Antonio Jiménez-Martín and Zahir M. Hussain

Received: 31 December 2022
Revised: 8 February 2023
Accepted: 15 February 2023
Published: 17 February 2023

1. Introduction

Information overload is an important issue experienced by users when choosing and purchasing products, which prevents them from easily discovering items that match their preferences. The role of recommender system(s) (RS) in providing support to users to overcome this serious problem is unquestionable. These systems provide recommendations relevant to users and create better exposure for items. The usefulness of RS is evidenced by their increasing incorporation in multiple areas including E-commerce platforms, social networks, lifestyle apps, and so on [1–5].

Much effort has been devoted to addressing the problems affecting RS and improving the performance of recommendations. However, researchers are focusing on mitigating different types of biases, which generally result in a decrease in performance. Bias shortcomings, commonly observed in machine learning models, lead to various forms of discrimination. Bias and fairness issues are some of the most critical problems that RS can face. The underlying imbalances and inequalities in data can introduce bias to RS while learning patterns from these historical data [6,7]. This translates into biased recommendations that influence the user's consumption decisions. In addition, the items consumed as a result of these recommendations are incorporated into the data used to generate new models. This makes the data to become increasingly biased and the problem of providing unfair recommendations become worse. However, the data are not the only cause of these inequalities, the design of the algorithms used can result in bias and automatic discrimination in decisions [7,8].

Due to the high implementation of machine learning technologies in society, the negative repercussions of biased models cover a wide range of aspects, including economic, legal, ethical, or security issues, which are detrimental to companies [6,9–12]. Moreover, bias can lead to user dissatisfaction [13]. Apart from these reasons, some international regulations, such as those of the European Union, include the obligation to minimize the risk of biased results in artificial intelligence systems that support decision-making in critical areas [9]. Decisions based on these results often have ethical implications from Aristotle's point of view regarding fairness, which may involve discriminatory treatment of minority or marginalized groups. For instance, some studies indicate that in the music recommendation field, female and older users are provided with a poorer quality of recommendations [14]. Those segments of the population that suffer the negative consequences of bias are the so-called protected groups on which mitigation strategies should focus. All of these facts have driven the research currently being carried out in this field [5,15–17].

Deep learning techniques in the RS area, which are used to improve the performances [18–20], have led to bias amplification in recommendations due to the greater propensity of these methods to magnify this problem. Within this approach, the methods based on graph neural networks (GNNs) perform well in this application domain, although they are still affected by the bias shortcoming. Some studies have shown that graph structures and the used message-passing system inside GNNs promote the amplification of unfairness and other social biases [21,22]. Moreover, in most of the social networks with graph architecture, nodes with similar sensitive attributes are prone to be connected in comparison to other nodes with different sensitive attributes (e.g., young individuals are more likely to start a friendship in social networks). The mentioned phenomenon creates an environment in which nodes with similar sensitive features receive similar representations from the aggregation of neighbors' features in GNN, while different representations are provided for the nodes with different sensitive features. This will lead to significant bias in the decision-making process [22].

There are various types of GNN-based RS, i.e., for conventional recommendations and sequential or session-based recommendations [23]. The most extended are graph convolutional network (GCN), graph attention network (GAT), and graph recurrent network (GRN). These approaches performed better in comparison to other machine learning methods in RS [24,25]. However, due to the aforementioned propensity to bias, one important challenge is to address the treatments of multiple types of biases that can dramatically affect the RS. This requires a prior identification and evaluation process that depends on the objectives pursued and the area of application, among other factors. In the RS field, the characteristics of the algorithms and the objectives may be very different from those of other machine learning fields, so many of the commonly used bias evaluation metrics are not suitable for assessing bias in recommendations. Some fairness metrics evaluate whether predictions are similar for protected and unprotected groups. However, in RS, users' preferences for items are predicted, and these preferences need not be similar for different groups. E.g., the music that young people listen to is different from the music that older people listen to, women may prefer different types of movies than men, and so on. In addition, many of the RS-specific metrics focus on detecting biases in the recommended items rather than unfairness to system users, e.g., they evaluate whether a similar number of songs by male artists are recommended by female artists.

In this study, we focus on user-centered fairness metrics and the behavior of GNN-based RS with respect to them. Our goal is to evaluate which ones are the most appropriate for recommender systems to help select the best strategy according to the objectives. We consider bias metrics at the individual user level, as well as fairness metrics focused on detecting and quantifying discriminatory treatment of groups of users. Groups formed on the basis of two sensitive attributes, gender, and age, were studied. In addition, we examined the extent to which the level of bias was related to the precision of the recommendation models and other quality measures applicable to the item recommendation lists.

The detection of bias in RS is particularly important due to the influence on user decisions and the progressive worsening of the problem caused by this fact. As previously indicated, the recommendations provided by these systems incite users to consume certain items. These consumption data become part of the datasets used to generate new recommendation models, which will be increasingly biased if there is bias in the initial models.

Bias evaluations in the GNN results were widely reported in the literature; however, work in the context of GNN-based recommender systems is very scarce. Moreover, most studies do not take into account the suitability of the assessment metrics in terms of the objective. This aspect is especially important in the field of RS because features are used to detect bias in other application domains; these are features that influence user preferences and produce different results for different groups of users without implying any bias. In addition, the impact of minimizing fairness bias on other dimensions of recommendation quality is omitted in many papers on the subject at hand. Bearing this in mind, the main contributions of this work are as follows:

- Review of the state-of-the-art concerning the evaluation and mitigation of bias in the field of machine learning, especially in the areas of recommender systems and GNN-based methods.
- Study of the adequacy of the metrics for evaluating fairness bias that affects users of recommender systems regarding discrimination between different groups.
- Examination of the behavior of GNN-based recommendation methods against the aforementioned fairness bias and comparison with other recommendation approaches.
- Analysis of the relationship between the values of the fairness metrics and the classic metrics for evaluating the quality of the recommendation lists (precision, recall, mean reciprocal rank, etc.) since the mitigation of biases usually results in a worsening of the recommendation quality evaluated by these measures.

This study is intended to answer the following research questions (RQ) about bias amplification of GNN approaches.

- RQ1: Can the findings reported in the literature in the general context of machine learning be extended to the specific field of recommender systems?
- RQ2: Do the performances of GNN-based recommendation methods against biases depend on dataset characteristics and sensitive attributes?
- RQ3: Are all bias evaluation metrics appropriate for assessing user-centered fairness in recommender systems in all application domains?
- RQ4: Do less bias-prone methods always provide lower-quality recommendations?

The work is organized as follows. After this introductory section, the state-of-the-art is discussed in different levels namely machine learning (ML), GNN algorithms, RS, and GNN-based RS. In the next section, the experimental study is described including the methodology, used datasets, recommendation methods, and evaluation metrics, which are explained in further detail. The following sections are devoted to the presentation and discussion of the results. Finally, the last section provides conclusions and future work.

2. State-of-the-Art

In this section, an overview of previous studies is provided. The section addresses the bias and fairness problems, as well as different forms of assessment and mitigation. The survey was carried out at different levels, including ML, GNN algorithms, RS, and GNN-based RS. It focused on the RS area and paid special attention to GNN-based RS and fairness evaluation metrics.

2.1. Bias and Fairness in Machine Learning (ML)

ML models are designed to work and train with human-generated data that potentially include biases [26,27], which can be generated due to systematic errors during the measurement and sampling procedures [28] among other causes. Such biases can be easily transferred into the ML models, leading to unfair decisions and low quality outcomes [8,10,29]. ML models, moreover, can amplify these kinds of biases and affect the system's decisions [30]. In [31], a systematic and controlled study on the bias amplification problem is conducted by using a simple image classification problem. The results demonstrate various factors involving the bias amplification problem, consisting of dataset bias, model performance, model overconfidence, training time, and training dataset size. Moreover, the difficulty of classification tasks in recognizing class membership can affect bias amplification.

As seen above, bias can occur at each stage of the ML process. They can arise in data collection and pre-processing, algorithm design, model induction, and even in the interpretation and use of the results. In addition, biases may originate outside the ML process, as they may stem from existing inequalities or discrimination in society.

There are different classifications of biases given according to different aspects, such as the phase of the process in which they occur [10] or the source of the bias [29]. According to the mentioned article, biases can come from data, algorithms, or interactions with users. Each group, in turn, encompasses different types of bias.

Training datasets can be biased due to differences between the sampling distribution and the real-world population. In other words, the sample is biased if its distribution does not represent the distribution of data in the real world. [32,33].

Algorithmic bias in ML can come from biased datasets, unreliable models, poor algorithm designs, or historical human biases [34]. This type of bias could appear as a result of under-fitting in the model training due to limitations in both input data and model performance [33].

User interaction biases occur due to the higher probability that users interact with popular items, which may be caused by the fact that the frequency of the interactions with the best-ranked results from search engines, recommendation algorithms, etc. is higher than with other items. This interaction data will subsequently be used by the algorithms making the popularity of those items increase more and more, introducing new biases in the data. It translates into a loop linking the feedback between data, algorithms, and user-interaction biases [29].

Although any bias can be a source of unfairness, and considering that there are many definitions of fairness in the area of machine learning, our work addresses fairness from the perspective of the results produced by the algorithms, which should not be discriminatory towards certain individuals or groups that are considered protected, based on sensitive features. Among these features or attributes, the most common are gender and race, although others, such as nationality, age, etc., can lead to unfair results.

In fact, there are numerous proposals in the literature so far to detect, evaluate, and tackle bias and unfairness issues in the ML area, but most approaches focus on mitigating bias and increasing fairness by considering protected or sensitive features [35]. Inequalities with respect to these features can be detected and quantified by means of different metrics based on statistical parameters [36]. Demographic parity, also called statistical parity [37], is a popular measure of assessing this type of bias. According to this metric, the results are unbiased if they are independent of sensitive attributes. This category includes the Equalized Odds metric [38], used to ensure parity in success/error rates for protected and unprotected groups (e.g., the ′women′ group and ′men′ group when gender is considered sensitive). The study introduced by [26] provides a new technique of ML bias detection and assessment. An alternation function is used to change the values in the potentially biased attributes and detect abrupt changes in the predictions for those that are biased. To assess the magnitude of the bias they propose a metric based on KL divergence. This research mainly focuses on gender and race features.

Moreover, metrics can be classified according to whether they are applied at the individual or group level [35]. The former assesses whether people with the same qualities in relation to some specific aspects receive the same treatments by ML algorithms, while the latter assesses the unequal treatments of different groups.

Some authors associate fairness with model explainability and process transparency. The proposal in [39] includes methods to identify biases at the local and global level, where transparency and explainability are taken as a basis. At the local level, the influence of individual features is examined, while at the global level, statistical analysis of the classification results for the different groups is suggested.

Other papers report different ways of assessing bias according to different perspectives of fairness [29,40], which will be discussed in a later section. As reflected in the extensive literature on the subject, multiple approaches are used to identify and assess fairness in ML. However, most studies agree that not all of them are universally applicable and that the choice depends on the application domain, the objectives pursued, and other factors.

Regarding bias mitigation approaches, there are three different major methods: pre-processing, in-processing, and post-processing [34,35].

Pre-processing methods are focused on increasing the quality of input data for the models. These approaches modify the input data to achieve fairer results. To do so, different techniques can be used including reweighing the input data, decreasing discrimination by learning data transformations, and encoding protected attributes that can lead to learning fair representations of the original data.

In-processing approaches focus on altering the algorithms. This can be considered an optimization task for balancing the fairness and performance of the model. This trade-off point of view usually requires an additional fairness metric for the evaluation phase.

Finally, post-processing approaches are those that do not depend on algorithms and can be implemented after any classification results. These approaches calculate probabilities for each classified observation and adjust the classification threshold for specific groups to acquire higher fairness of the solution.

The literature contains numerous proposals from each of the above groups. However, there are still unresolved issues that need to be addressed [35]. The first challenge is the trade-off between performance and fairness. Many studies show that increasing fairness can result in a decrease in accuracy. The second issue is agreement and incompatibility, which points out that most fairness metrics either emphasize individual or group fairness, but not both of them together. Moreover, many used methods for group fairness, address the problems between groups which can result in magnifying the problems inside the groups. The third one is the issue between context and policy. Many approaches and metrics only focus on optimization and do not provide transparency for the reason behind the unfairness and root of the problem. The fourth and last is the democratization of ML versus the fairness skills gap. Many systems today are based on cloud services or Automated ML. Such types of services help companies to democratize ML by extending its use to non-expert personnel. As a consequence, this can result in a decrease in social sensitivity and fairness.

2.2. Bias and Fairness in GNN-Based Models

Recently, GNN-based models have garnered attention for their strong performance and wide range of applications in graph learning tasks [25,41,42]. Despite their success, many studies have shown that these algorithms suffer from bias and unfairness issues. GNNs are susceptible to discriminating against certain demographic subgroups defined by sensitive features such as age, gender, and race. Furthermore, researchers have given little attention to understanding and mitigating these biases in GNNs [22,27,41,43–45].

Bias problems in GNN algorithms are rooted in various causes, such as the bias of the input network structure. It is believed that the message-passing mechanism in GNNs amplifies this problem, but other aspects of the GNN network structure should not be overlooked. There are some obstacles to comprehending how biases in the structure lead to biased predictions [41]: fairness notion gap, usability gap, and faithfulness gap. The fairness notion gap evaluates bias at the instance level. The usability gap refers to fairness affected by edges in the computational graph and which edges contribute the most (the final edges are not necessarily the ones that contribute the most). Moreover, the faithfulness gap ensures that comprehended bias explanations show the true reason for bias based on the used model.

The study presented in [41] sheds light on these gaps by proposing a bias evaluation metric for the node predictions and a framework for their explanations. The proposed metric evaluates the contribution of each node to the distance of the output distributions of two sensitive subgroups of nodes formed on the basis of the sensitive features.

The literature on this topic includes different approaches to address the problem of biases in GNN-based models, although research in this particular area is still quite limited.

The work presented in [44] proposes a novel bias-aware graph neural network framework by learning invariant node representations in graphs to enhance prediction for unrevealed testing distributions. The two important components in this system are bias identification and invariant prediction. The proposed method is designed specifically for selection bias. The results indicate a strong performance of this method across various datasets.

Graph convolutional networks (GCNs), variants of GNNs, are the targets of a work that focused on bias issues [46]. In this study, a self-supervised-learning degree-specific GCN is introduced to manage the degree-related biases of GCNs regarding models and data. To achieve this goal, a degree-specific GCN layer is designed to model differences and similarities of nodes with various degrees and to decrease the model bias in GCNs. The results on three benchmark datasets illustrate an improvement in accuracy for low-degree nodes in GCN.

The objective of the works reported in [22] and [47] is to eradicate discrimination and increase the fairness of GNNs concerning sensitive attribute information. In [22], the authors propose a method that is capable of reducing bias along with keeping high node classification accuracy. Reference [47] introduces a method to learn a fair adjacency matrix with good graph structural constraints to achieve fair link prediction with minimum effect on the accuracy. The work conducted by [48] proposes two model-agnostic algorithms to execute edge editing by exploiting gradient information of a fairness loss to find edges that enhance fairness.

Other research approaches, such as [49], focus more on training data to eliminate bias in GNNs and enhance the fairness of the system toward sensitive groups. This study introduces a novel unified normalization framework that decreases the bias in GNN-based learning and tackles some learning challenges such as convergence and stability. The main aim is to enhance fairness concerning statistical parity and equal opportunity compared to fairness-aware baselines. Moreover, a sampling method is proposed in [25] to address bias and high variance problems for GNNs. To do so, a novel reward function is provided to avoid unstable and high-variance results. The proposed method is optimized compared to previous bandit-GNN use cases, thus, enhancing the accuracy of outcomes.

2.3. Bias and Fairness in Recommender Systems

Bias and fairness problems in the RS area can have different meanings and can be divided into different categories. From a top-down perspective, bias can be categorized into three different classes analogous to those established for ML in [29], namely bias in input data, which refers to the data collection phase from users, bias in the model, or algorithmic bias, which takes place in the learning phase of recommendation models based upon the collected data, and bias in results, which influence subsequent user actions and decisions [5,50].

To expand the bias issue in more depth, The three classes can be further subdivided into different subclasses in a circular form.

Data bias generally happens due to differences between the distribution of test and training data. Data bias itself can have different forms including selection bias, exposure bias, conformity bias, and position bias. Selection bias can happen due to skewed rating distribution, which means that the observed ratings are samples that do not represent all ratings. Exposure bias takes place when users are only exposed to some specific items, and unobserved interactions do not necessarily indicate a negative preference. In addition, bias can limit the users' choices and contaminate users' feedback, which can amplify exposure bias [5,51]. Conformity bias can be the result of skewed interaction labels, which means some users are willing to behave similarly to the other group members. Position bias occurs when users are willing to choose items in better positions (e.g., better ranks) rather than real relevant items [5].

Algorithmic bias could occur at any time throughout the generation of the recommendation model, which covers data pre-processing, training, and evaluation. An important algorithmic bias that is not considered harmful is inductive bias. Inductive bias occurs in model generalization, regarding the model's assumptions for learning the training dataset in a better way and generalizing the decisions on the unseen test data [5].

Bias in outcomes can be classified into two subgroups, namely popularity bias and unfairness. Popularity bias, usually considered the most important type, roots in the long-tail phenomenon in ratings, meaning a small group of popular items receive most of the user interactions. The mentioned issue can result in receiving higher scores for popular items from the model and dismissing the unpopular ones [5,16]. Unfairness can occur because of systematic discrimination against certain groups [5].

Mentioned types of biases can together lead to a circular pattern in which biases in data are transferred to the models and from the models to the results. The circle closes with the transfer of bias from the results back to the data. At each of these stages, new biases can be introduced [5,52]. This circular behavior can complicate the process of bias recognition and mitigation [5,53].

Biases impact the robustness of the recommendation models generated from the data. Often, these models are built and studied under unrealistic noiseless data and independent training and testing data. However, it is essential now to study RS in realistic scenarios where the input data in RS can be influenced by malicious attacks, sparsity, and bias issues that can lead to a decrease in the performance of RS [54]. Precisely the evaluation of robustness in recommendation models is the objective of a recent tool presented in [54]. Regarding bias, this tool is based on the analysis of the performance of subpopulations of the test set with respect to some sensitive attributes such as gender or age. For the particular aspect of bias, the idea is similar to the one presented in our work, in which different performance evaluations are performed for the protected groups.

Conversational recommender systems (CRSs), in which the interaction with the system resembles the interaction between humans, are taken into consideration in [4]. The objective is to systematically investigate the popularity bias issue in recently used CRSs from different perspectives, such as exposure, success, and conversational utility. This work presents a set of metrics focused on popularity bias and aimed at this particular type of system. The work by Abdollahpouri et al. in reference [55] focuses on popularity bias and long-tail problems in RS. This study proposes appropriate metrics to assess the long-tail distribution. These metrics are the average recommendation popularity (ARP), average percentage of long tail (APLT) items, and average coverage of long tail (ACLT) items.

Process fairness focuses on fair allocation in the process. It refers to fairness in the models, features (e.g., race and gender), and the learned representations. On the other hand, outcome fairness, also called distributive justice, turns fair allocation into a fair recommendation outcome [56].

Outcome fairness itself has two sub-categories, i.e., grouped by target and grouped by concept. Furthermore, grouped by the target can be either group-level or individual-level fairness. Fairness concerning groups indicates that the results related to the different groups must be fair. Individual fairness declares that the results are supposed to be fair at the level of each particular individual [56]. 'Grouped by concept' has its own categorization: (1) Fairness at the individual level holds that the same treatment should be given to homogeneous individuals. At the group level, different groups should receive similar treatment. (2) Calibrated fairness, also called merit-based fairness, ensures that the merit of an individual (or group) should correlate with the value of the result. (3) Counterfactual fairness requires that individuals have the same results in both the real and the counterfactual world. (4) Envy-free fairness holds that individuals are not supposed to face envy issues toward the outcomes of others. (5) The Rawlsian 'maximin' fairness implies the maximization of the result values for the weakest individual or group. (6) 'Maximin'-shared fairness, which indicates that all individuals (or groups) should receive better outcomes than their 'maximin' share [56]. Figure 1 summarizes the classifications described above.

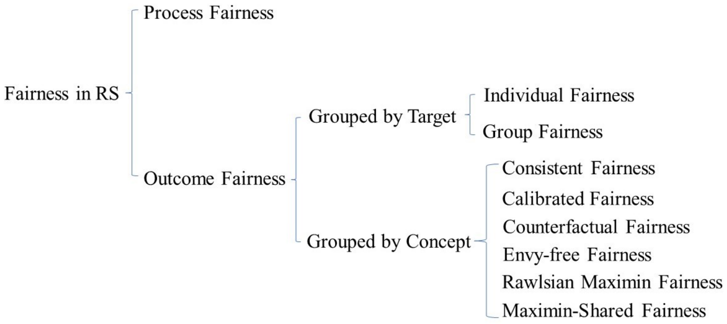


Figure 1. Fairness in RS [56].

The relation between bias and fairness is very important. A systematic analysis was conducted by the authors of [57]; it focused on alleviating processes for consumer unfairness in the rating prediction tasks of two real-world datasets (LastFM and MovieLens). The first analysis assesses the effect of bias mitigation on the accuracy of the models, such as NDCG/RMSE. The second analysis measures the impact of bias mitigation on unfairness. The third analysis, furthermore, checks if disparate impact always damages minority groups based on demographic parity (DP). This study highlights the challenges of this subject matter, hence, adding some solutions and optimization techniques. Moreover, choosing the right metrics for this kind of evaluation is vital.

Fairness in rating prediction is analyzed in [40]. This study delivers and maps different fairness concepts and definitions between ML and RS and spots the similarities and gaps. Finally, a new bias mitigation strategy is proposed to tackle potential unfairness.

Improving fairness and dealing with sensitive attributes are challenges in RS. Some works, such as [58,59], aim to achieve fair results by considering gender (a sensitive feature). Reference [58] introduces an unbiased gender recommendation to equilibrate the accuracy between men and women. By acquiring the preferences of users, a component with adversarial learning is designed to remove gender bias, which uses the user–item graph to learn users and items representations. In addition, a multi-hop mechanism is proposed to boost representations by aggregation of users' higher-order neighbors. To avoid damaging the accuracy of the model, an end-to-end training framework with adversarial learning is used, which removes gender-specific features and keeps common features. Reference [59] focuses more on discrimination and gender bias in the field of music recommendations. To decrease artist gender bias, this work involves the analysis of CF approaches to determine their behavior regarding the enhancement or reduction of this type of bias. Outcomes were analyzed by means of two different metrics, preference ratio (PR) and bias disparity (BD). The work proves that bias can be amplified by CF RS.

In addition to the previously mentioned work, [60] focuses on gender bias in RS. A model-agnostic bias mitigation method is proposed to reduce fairness and maintain accuracy at the same time. The results on two book rating datasets show a great decrease in the bias, with a minor impact on the performance of the used K-nearest neighbors' family.

The study in [7] addresses the problem of bias and fairness in the RS area by proposing a method to recognize and describe the potentially discriminated user groups. In this research, they concentrate on creating descriptions of specific user segments aimed at better understanding the underlying data characteristics and hidden biases but focusing on discrimination against users rather than items.

The work presented in [61] investigates statistically significant contrast in recommendation utility by considering both age and gender. To do so different metrics including NDCG (normalized discounted cumulative gain), MRR (mean reciprocal rank), and RBP (rank-biased precision) are used. In addition, to verify if differences in utility are significant across demographics, Kruskal–Wallis significance tests on mean NDCG values between the demographic groups are used. The Bonferroni correction is implemented for multiple testing. The results show that the recommendation performance steadily drops for older users, and this performance is lower for women compared with men. This study indicates that the total usage and popularity of consumed products are powerful predictors of recommendation performance, and change dramatically across demographic groups.

The research conducted in [62] focuses on movie and book recommendations with the intention of characterizing imbalances in the distribution of the user–item data and regarding where items are produced (geographic imbalance). It is important to evaluate the disparate impact on advantaged and disadvantaged groups. To do so, items are divided into groups based on their continent of production and characterize how represented is each group in the data and then the given visibility and exposure are measured to observe disparities toward the most represented groups. To address this issue, equity with a re-ranking approach is proposed that controls the number of recommendations given to the items produced in a continent (visibility) and the positions in which items are ranked in the recommendation list (exposure). This method considers the loss in effectiveness, hence, regulating the fairness of providers coming from different continents. NDCG is used as a metric to measure the performance of the models. The mentioned method shows more equity for providers with respect to both visibility and exposure.

Typical solutions include proposing a user-centered fairness re-ranking framework that is applied on top of a base ranking model to mitigate its unfair behavior towards a specific user group (i.e., disadvantaged groups). In [63], a user-oriented fairness study is re-produced and extensive experiments are provided to analyze their proposed method's dependency on various fairness and recommendation aspects, including the recommendation domain, nature of the base ranking model, and user grouping method. Additionally, this work evaluates the final recommendations provided by the re-ranking framework from both user (e.g., NDCG and user fairness) and item-side (e.g., novelty and item fairness) metrics. Interesting trends and trade-offs between the model's performance related to various evaluation metrics are discovered. For example, the referred study claims that the definition of the advantaged/disadvantaged user groups plays a crucial role in the effectiveness of the fairness algorithm and how it improves the performance of certain baseline ranking models.

The authors of [64] propose a new approach for fair recommendation with optimized antidote data, which aims to improve the fairness performances of RS by building a small and carefully crafted antidote dataset. To this end, this approach formulates the antidote data generation task as a mathematical optimization problem that minimizes the unfairness of the targeted RS without disturbing the deployed recommendation algorithms. Extensive experiments show that their proposed antidote data generation algorithm significantly improves the fairness of RS with small amounts of antidote data.

Fairness in RS is studied in a previously referenced work [54] in which the analysis of the robustness of recommender models is its main objective. The robustness evaluation toolkit presented in this work helps accelerate the process of a comprehensive robustness evaluation for RS models. The NDCG and area under the curve (AUC) metrics are used to evaluate the results.

Many studies separately address user and item fairness concerns and ignore the fact that RS operate in two-sided marketplaces. In [65], an optimization-based re-ranking approach that seamlessly integrates fairness constraints on both the consumer and producer side into a common goal framework is presented. This method demonstrates through large-scale experiments on 8 datasets that their proposed method is able to improve both consumer and manufacturer fairness without reducing overall recommendation quality and denote the role algorithms play in minimizing data bias.

Some studies focus on specific scenarios. Reference [66] considers the fairness problem in RS under the premium scenario. According to this research, current RS can be unfair toward premium users/items, which should receive better quality services, due to control of confounding factors. To address this issue, a new metric is designed to evaluate fairness when premium users are involved and compare premium and standard group behavior.

Furthermore, a flexible and contextual fairness-aware recommendation framework is introduced which provides a good distribution to fit user or item group scores. The results show that the proposed method achieves better services for premium users/items.

2.4. Bias and Fairness in GNN-Based RS

It is claimed that the use of the GNN-based method for RS can improve the accuracy of the results [18–20]. On the other hand, this enhancement in performance can cause bias and fairness problems [21,22]. As discussed in Section 2.2, the structure of graphs together with the message-passing system inside GNNs can amplify bias problems, leading to unfair outcomes.

Moreover, many RS work in the social network area, which includes graph structure. In these systems, nodes of homogenous sensitive attributes are probable to make connections with each other compared with other nodes of different sensitive attributes (e.g., connections among young individuals in social networks). This phenomenon creates an environment in which nodes of similar sensitive features receive similar representations from neighbors' features aggregation in GNN, while different representations are received by the nodes of different sensitive features. This results in a notable bias problem in decision-making [22].

Certain sensitive attributes can have a strengthening effect on the biases present in the network of GNN-based RS, and this problem has led to measuring fairness in these approaches. Suitable metrics for addressing this problem should take into account the proportion of positive classifications for each group associated with different values of the sensitive attribute [23,67].

There are some recent works in the field of GNN-based RS that copes with fairness issues and sensitive attributes. An article in this area [67] centers on quantifying and addressing fairness problems. This novel study on algorithmic fairness is focused on analyzing group fairness (disparate impact) of the current graph embedding. In addition, the statistical parity measure is extended to evaluate the disparate impact and quantify fairness for groups, taking into account the sensitive attributes of pairs of users. This leads to a new idea of equality of representation to evaluate fairness in friendship RS. Mentioned methods are applied to real-world datasets and then a novel fairness-aware graph embedding algorithm is proposed to mitigate the bias. In the results, the mentioned mitigation method improves the two metrics referred to above on a large scale.

The study conducted in reference [68] aims to make fair recommendations by eliminating sensitive information in representation learning. To achieve this goal, user and item embeddings and a sensitive feature set are used in the proposed model to filter sensitive information by means of a graph-based adversarial training process. To measure the fairness of the model, classification performance is calculated. Due to the dataset imbalance regarding the gender attribute, binary classification performance was measured from the Area Under the Curve (AUC) metric. In addition, micro-averaged F1 is used for the rest of the multivalued attributes. The proposed model was validated with Lastfm-360K and MovieLens datasets.

The lack of sensitive attributes in RS datasets does not mean that there are no fairness issues to be addressed, but this is another challenge that experts are faced with [22]. Although in RS the interaction between users and items does not explicitly contain any sensitive information from users, due to the high correlation between users and their attributes, directly apply of the modern user and item representation learning could lead to leakage of mentioned sensitive information [23]. In the underlying graphical structure of RS, there is no independence between users, but rather those with similar preferences or behavior are correlated in an implicit way. This can lead to unwanted biases and critical issues for CF recommendations [23]. In addition to previously mentioned issues, recent GNN algorithms suffer from the problem of weak generalization due to societal bias in data [22].

To address these issues, many approaches have been provided. Some introduced approaches are based on graph embedding methods, by means of which each user is represented by a lower-dimensional vector that includes structural information (e.g., user's neighborhood, popularity) within the network. These methods are especially useful in online social networks (OSN), which face discrimination against minority groups [67].

Some work as [69] focus on the long-tail problem, related to item popularity, in the context of session-based recommendation models and its two possible approaches, namely RNN-based (recurrent neural network) models and GNN-based models. In this paper, two accuracy metrics are used including MRR and recall for performance evaluation, and three long-tail-based metrics are used including coverage, tail coverage, and tail.

Another study [70] regarding the GNN-based RS area addresses the exposure bias issue by neighbor aggregation. The edge weight for the aggregation process is given by the inverse propensity score with Laplacian normalization computed from the user–item bipartite graph. This highlights the less popular neighbors in the embedding. The mentioned method is able to balance the biased local structure of each target node. The metrics used in this study are based on accuracy, including NDCG and hit ratio.

Graph-based collaborative filtering (CF) approaches in RS are addressed in [71]. The authors analyzed accuracy and novelty, concluding that the vanilla GNN layers in models may lead to recommendations biased to popular items, leading to poor novelty performances. This paper presents a theoretical analysis of the roots of popularity bias in the methods under study, and proposes a way of handling the trade-off between accuracy and novelty based on the graph structure. The results show that the popularity bias is amplified by symmetric neighborhood aggregation in the majority of graph-based CF models. This amplification, moreover, can be enlarged with the graph propagation depth.

Other papers in the literature study other problems of GNN-based recommendation methods, such as cold start or heterogeneous interactions, which may ultimately result in biased or unfair recommendations to users. For instance, many studies on GNN-based RS do not consider the social inconsistency problem, which means social links are not necessarily consistent with the process of rating prediction. This phenomenon can be found in two different levels, including context and relational level. The work reported in [72] considers addressing this social inconsistency issue and reducing the cold start problem for rating prediction in the social media area by combining social links with interactions among items and users. This study strengthens the consistency between sample neighbors by calculating consistency scores between neighbors and relating them to sampling probability.

The research conducted in reference [73] focuses on online shopping baskets, which help users to purchase their products faster. This shopping basket can be created based on users and items collaborative filtering signals and multi-item correlations. This representation of basket intent can be modeled with a basket-item link prediction task in the user–basket–item (UBI) graph. This research aims to investigate the collectivity and heterogeneity characteristics in the mentioned area. 'Collectivity' refers to the semantics of each node, which should be aggregated from both directly and indirectly connected neighbors. Heterogeneity, on the other hand, is comprehended from multi-type interactions and multi-type nodes in the UBI graph. This study introduces a novel framework with three types of aggregators accurately designed for three types of nodes based on GNN. Mentioned aggregators collectively learn node embeddings from both neighborhood and high-order contexts. The interactive layers in the aggregators, moreover, can distinguish different types of interactions. The results indicate a strong performance of the introduced model.

After reviewing the state-of-the-art and examining the challenges in GNN-based RS regarding bias and fairness problems, it appears that although the lack of sensitive features in this area is a critical issue, most of the studies in this field aimed to address the problem of discrimination against minorities and information leakage. These kinds of unfairness and discrimination are clearly against mentioned regulations and anti-discriminatory laws. Finally, the trade-off between accuracy and bias mitigation is relevant in many studies.

Table 1 summarizes the state-of-the-art covered in this research.

Table 1. Literature review.

Area of Research	Focus	Publications
Bias and fairness in ML	Understanding, detection, and evaluation of bias and/or fairness in ML Fairness in information networks Bias management in ML	[26,29–35,38,40] [27] [10]
Bias and fairness in GNNs	Understanding, detection, and evaluation of bias and/or fairness in GNNs Bias and/or fairness mitigation in GNNs	[25,41,43,45] [46–49]
Bias and fairness in RS	Understanding, detection, and evaluation of bias and/or fairness in RS Bias and/or fairness mitigation in RS	[5,16,50,52] [4,7,51,54,55,60–66,69]
Bias and fairness in GNN-based RS	Understanding, detection, and evaluation of bias and/or fairness in GNN-based RS Bias and/or fairness mitigation in GNN-based RS	[18,20,21,23,56,57] [22,53,58,67,68,70–73]

3. Study of Performance against Bias of Recommendation Methods

The main objective of this study is to analyze the behavior of GNN-based recommendation methods in regard to biases, with a special focus on those that impact users and result in unfair treatment of some individuals or groups based on sensitive characteristics, such as gender or age. To achieve this, it was necessary to select appropriate RS datasets containing sensitive attributes from different application domains. Moreover, several recommendation algorithms representative of both GNN-based and other approaches were selected for comparison. Finally, diverse bias assessment metrics were analyzed and the most suitable ones to reach the objectives in the context of RS were applied.

3.1. Methodology

In the experimental study, we attempted to answer the research questions posed in the introduction. This purpose was present in the three fundamental stages of this process: selection of data, selection of the recommendation methods, and selection of the procedures and metrics for evaluating the results.

The choice of datasets was conditioned by the presence of the sensitive attributes and biases under study. Thus, the selected datasets contain the gender and age of users, as well as the unequal distribution of instances for the different values of these attributes, in addition to an important item popularity bias. These three real-world datasets are MovieLens 100K, LastFM 100K, and book recommendation. The descriptions and EDA (exploratory data analysis) of these datasets can be seen in the following subsection.

Secondly, for the study on the biased behavior of GNN-based recommendation methods, the most representative algorithms in this category were selected. To compare the results of these methods with other approaches, it was necessary to search for appropriate algorithms for comparison. In this way, three different types of RS techniques were tested, including collaborative filtering (CF), matrix factorization (MF), and GNN-based approaches. The specific algorithms for each category are described in the 'recommendation methods' section.

The third step was related to the evaluation of results and involved making decisions on both the validation procedure and the assessment metrics. Regarding the former, a cross-validation technique was used to separate the data into train, test, and validation sets. Moreover, 80% of each dataset was used in the training set, 10% in the test set, and 10% in the validation set. The selection of the metrics required a detailed process of analysis and classification to identify those that met the objectives of the research and were valid in the field of RS. Some evaluation metrics were used to assess bias in recommendations that affect users at the individual level, such as average popularity or the Gini index. Moreover, fairness metrics were applied to evaluate discrimination against certain groups, such as absolute unfairness or non-parity among others. In addition, to study the relationship between accuracy and bias in the recommendations, different metrics related to the precision of the lists of recommended items were applied.

3.2. Benchmark Datasets

The three real-world datasets used in this study include sensitive information and suffer from popularity bias. The description and EDA of these three mentioned datasets are provided below. First, Table 2 with the three dataset information is shown. In it, we can see that the MovieLens and LastFM datasets contain the sensitive attributes age and gender, while the book recommendation dataset contains only the sensitive attribute age.

Table 2. Dataset information.

Dataset	Features	Description	Data Type
MovieLens	Age	Age of users	int
	Rating	Rating on movies provided by users	float
	User id	IDs of the users	int
	Movie id	IDds of the movies	int
	Gender	Gender of users	String
	Occupation	Users' job	String
	Movie title	The title of rated movies	String
LastFM	Weight	Listening count per artist and user	float
	Age	Age of users	float
	User id	IDs of the users	int
	Item id	IDs of the artists	int
	Gender	Gender of users	String
	Country	Users' territory of living	String
	Name	Name of the artists	String
Book Recommendation	Rating	User score on books	float
	Age	Age of users	float
	User id	IDs of the users	int
	Item id	IDs of the books	int
	Location	Users' location	String

3.2.1. MovieLens 100k

The first dataset chosen for this research is MovieLens [74]. This dataset has been used in many studies in the RS area. It was collected gradually and selected randomly from the MovieLens website, which is a non-commercial web-based movie recommender system. The dataset contains users' ratings of movies on a star scale in the interval [1, 5]. In this research, a subset of the real MovieLens dataset called ml-100k is used, which includes 100,000 records of ratings. This dataset contains user information, such as "Gender" and "Age" attributes, which are sensitive features according to the capAI guidance [75]. Furthermore, in this dataset, the long-tail phenomenon occurs, which is the main cause of popularity bias toward more popular movies. The combination of mentioned bias and user information is the main reason behind choosing this dataset.

For a better understanding of the MovieLens dataset, some graphical representations are shown below. These allow us to carry out the Exploratory Data Analysis (EDA). Figure 2 shows that most of the ratings are given to a small portion of items, the most popular ones. This effect, known as long-tail, is evidence of the popularity bias suffered by the dataset. Figure 3 illustrates the distribution of age and gender. It shows that a majority number of users are young individuals and there is a majority of men in all age ranges. Figure 4 shows the age and gender distributions of MovieLens dataset.

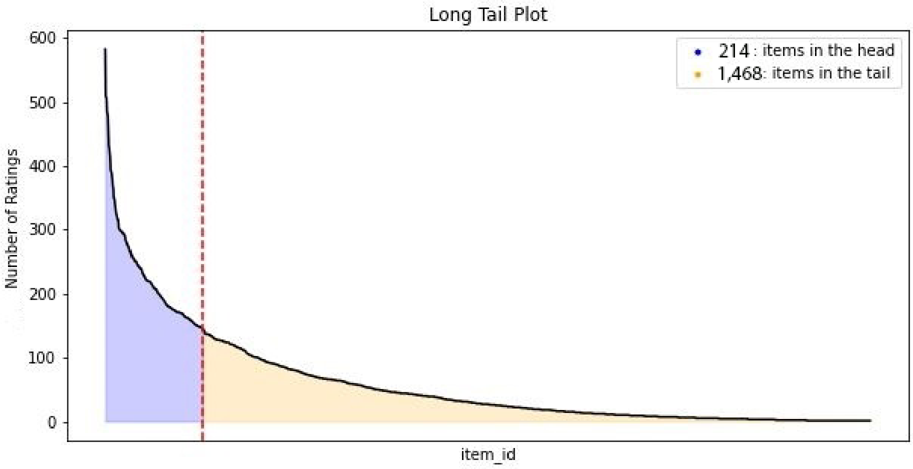


Figure 2. Number of ratings per item in the MovieLens dataset.

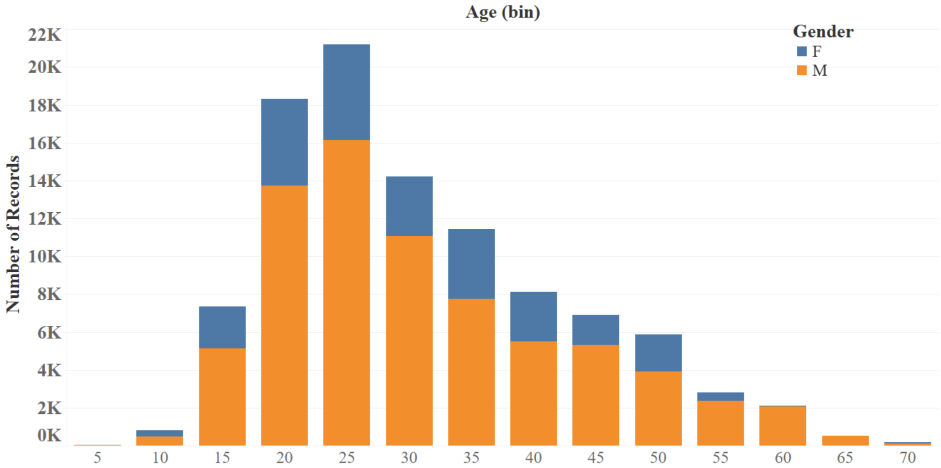
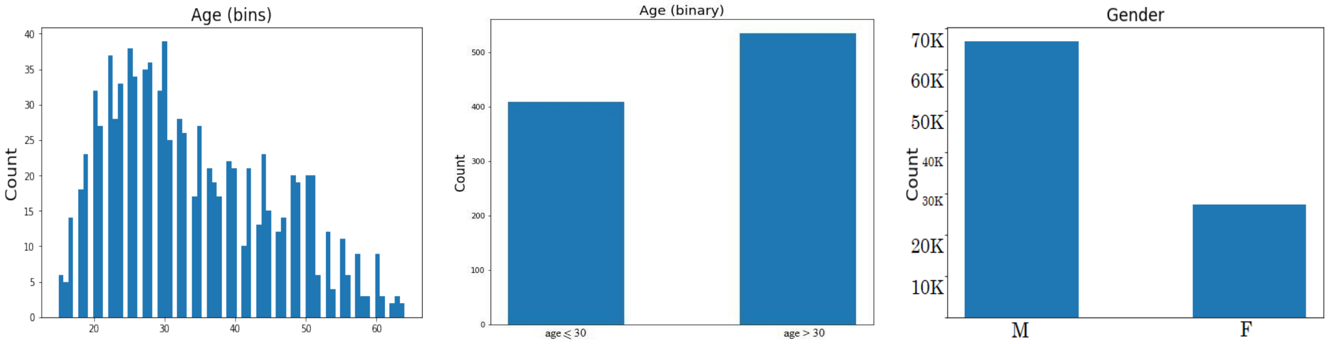


Figure 3. Age/gender distribution in the MovieLens dataset.



(**a**) Age distribution. (**b**) Gender distribution. (**c**) Binary age distribution.

Figure 4. Age and gender distributions of MovieLens dataset.

3.2.2. LastFM 100k

The LastFM dataset [76] was popular in RS for the music recommendations collected by Òscar Celma. This dataset contains user and artist information from different parts of the world. Instead of a rating option, this dataset shows the times that each user has listened to each artist (weight). The dataset used in this research is a subset of the LastFM 360K, which was pre-processed for RS implementation. In this subset, we used 100,000 interactions in the dataset. According to capAI guidance [75], gender and age are considered sensitive attributes in this dataset. In this dataset, the number of times that users listened to certain music is presented as the weight. This feature is normalized into the scale of [1–5] to acquire better accuracy.

The EDA for the LastFM dataset is provided below.

Figure 5 shows the number of ratings per artist in the LastFM dataset. Similar to MovieLens, this dataset has a significant popularity bias. Items in the long tail account for 90.3% out of the 41,269 items in the dataset.

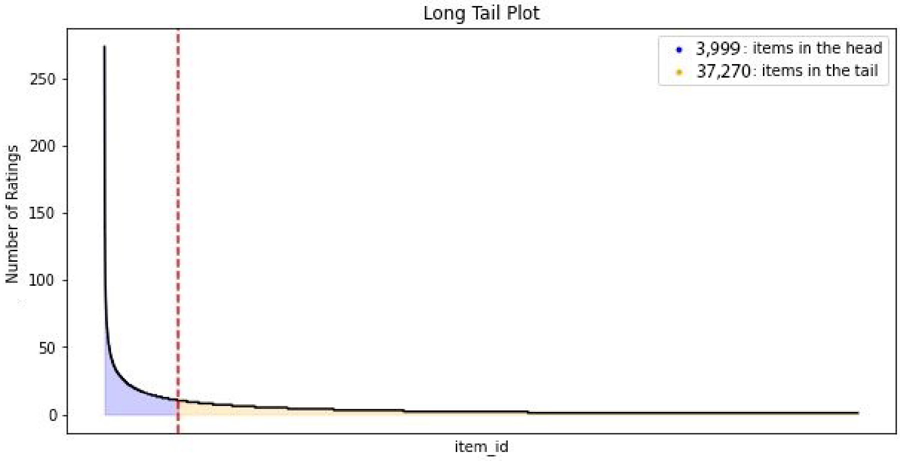


Figure 5. Long-tail distribution in the LastFM dataset.

Figure 6 shows age/gender distribution for the LastFM dataset in a stack histogram chart. Figure 7 shows gender and binary age distributions in the LastFM dataset. Based on the gender distribution in this dataset, the male gender dominates over the female gender.

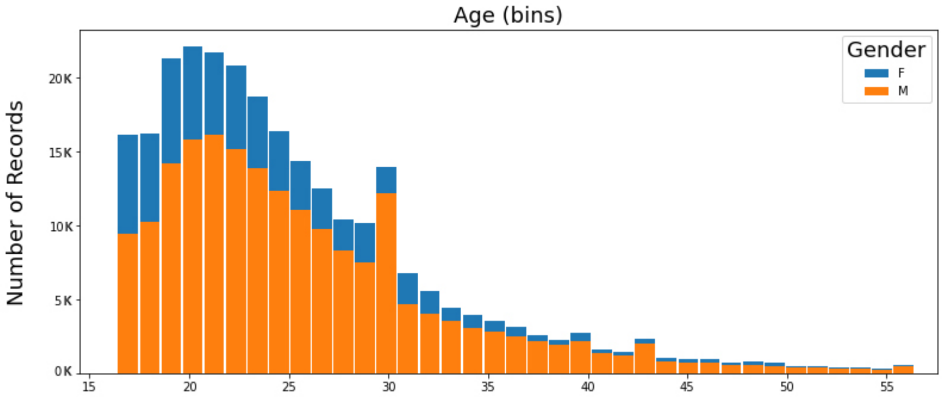
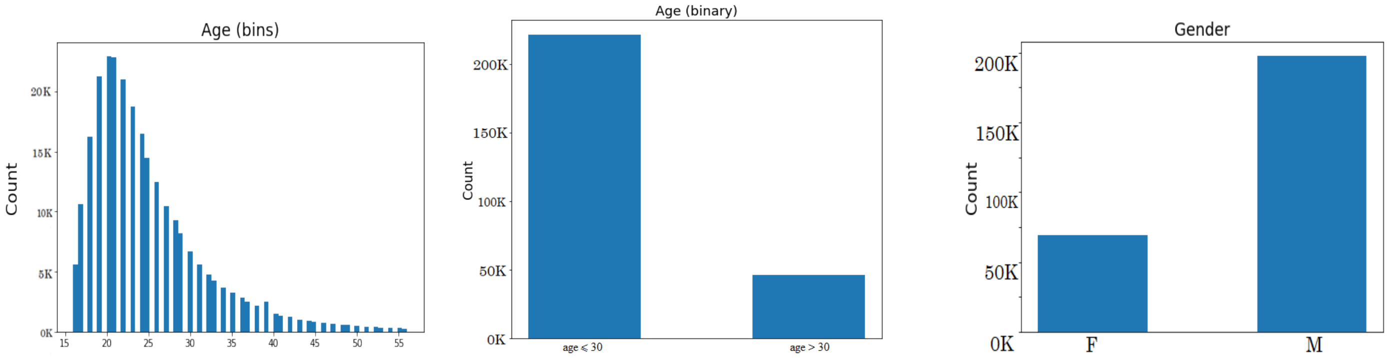


Figure 6. Age/gender distribution in the LastFM dataset.



(**a**) Age distribution.

(**b**) Gender distribution.

(**c**) Binary age distribution.

Figure 7. The age and gender distributions of the LastFM dataset.

3.2.3. Book Recommendation 100k

This dataset [77] includes ratings from users of different books. In this study, a 100,000-record sample of this dataset was chosen for the experiment. This sample completely follows the real dataset distributions.

The following is the EDA of the book recommendation subset.

Figure 8 shows the long-tail phenomenon in the book recommendation dataset. Figure 9 shows the age distribution histogram and binary age distribution of the book recommendation dataset.

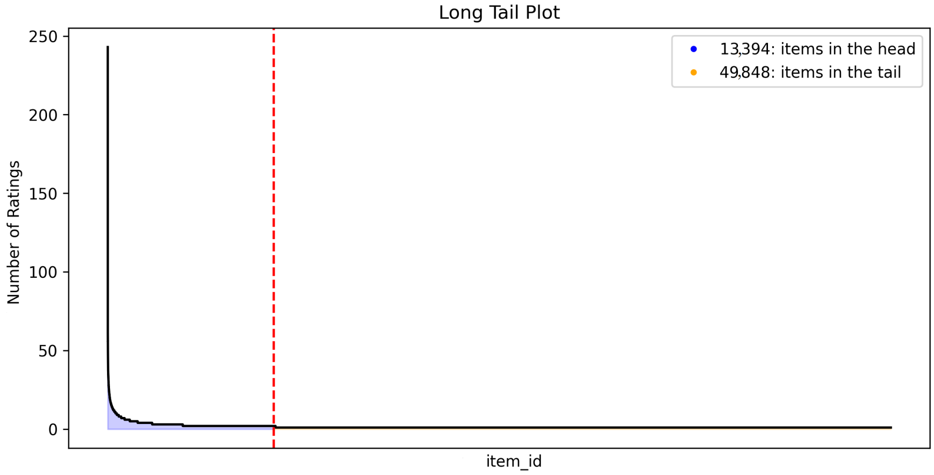
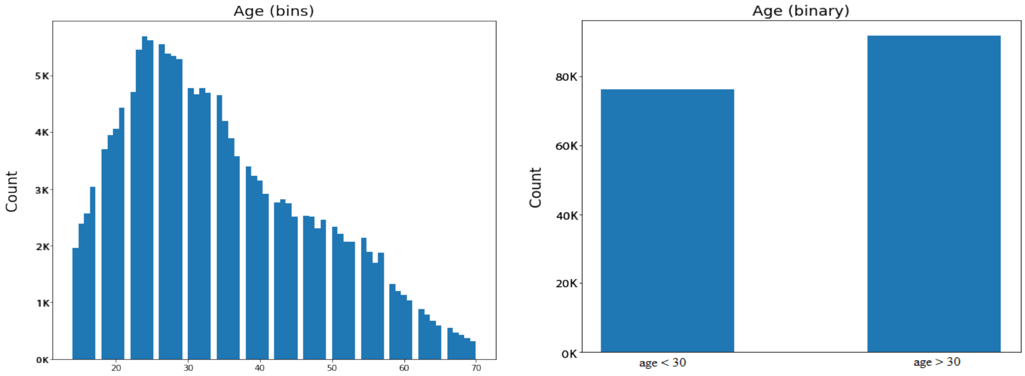


Figure 8. Long-tail distribution in the book recommendation dataset.



(**a**) Age distribution.

(**b**) Binary age distribution.

Figure 9. Age distribution histogram and binary age distribution in the book recommendation dataset.

3.3. Recommendation Methods

In this research, three recommendation approaches were used to compare the effect of different types of algorithms on the bias of the results: collaborative filtering (CF), matrix factorization (MF), and GNN-based approaches. The most representative algorithms for each of these groups were tested. Implementation of these methods can provide a wide range of comparable results that can accelerate the process of bias and fairness analysis. The following section includes a brief description of the methods of each approach that were used in this study.

1. Collaborative filtering (CF).
   CF approaches are implemented based on ratings given to items by users, which enclose the user preferences. The recommendation algorithms predict the ratings that users would give to items not rated by them by calculating user or item similarities. These similarities indicate that similar ratings given to a certain item by two different users can also happen for new items. Items can be recommended to a user based

on the previous items consumed or rated by the user. In CF approaches, ratings are represented in a user–item rating matrix that is used to find similarities among users and items. A drawback related to these approaches is that user and item features are often not available since recommendations in CF are provided by only using the feedback of other users. CF techniques can be implemented in two different ways, as can be seen below:

- User-based: This technique is used to predict the items that a user might desire on the basis of ratings given to that item by other users who share similar preferences with the target user [1].
- Item-based: This technique predicts the rating of a user for an item on the basis of the ratings given by the user to similar items. The similarity of the items is calculated from the values of the ratings they receive from users. Item-based approaches are usually more reliable, faster, and do not need to be updated frequently, but the results are sometimes worse than those of user-based methods [1].

Figure 10 shows the differences between user-based and item-based approaches.

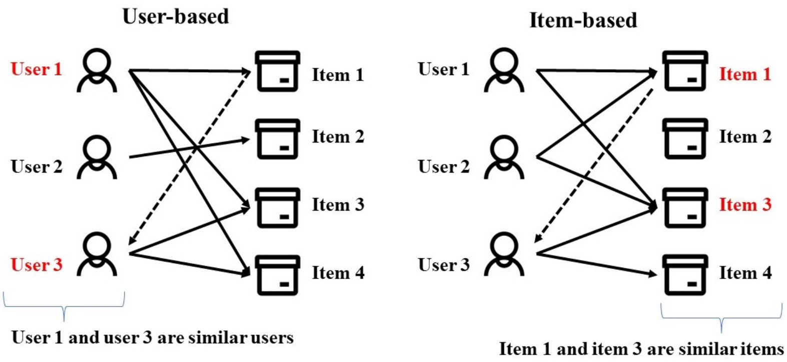


Figure 10. Differences between user-based and item-based approaches.

The CF methods used:

- ItemKNN: It is a widely used algorithm belonging to the item-based group, where the similarity between items is computed based on the ratings given by the users. Customers usually are more likely to consume items with the same characteristics as those previously consumed by them, which is the main idea behind this method. It follows a model-based approach involving two important components. The first is in charge of inducing a model which captures the relations between different items, and the second component uses the model to provide recommendations for a user. The response time to the user of this method is quite short because the model is already created at the recommendation time. In addition, it provides good accuracy compared with other similar CF methods [78–81].
- Neural collaborative filtering model with the interaction-based neighborhood (NNCF): This is a CF method where complex interactions between users and items are modeled by means of deep learning, although neighborhood information is used to improve the performance. The results of traditional methods, such as simple linear factorization, can be enhanced by NNCF, giving the best support to the complex interactions among users and items. Another advantage of this method is the possibility to obtain high-quality user–item embeddings. [82–84].

2. Matrix Factorization:
Recently the usage of MF methods in RS has increased significantly thanks to their advantages and the capability of reducing the storage size [1,85]. The purpose of MF models is to address the sparsity problem of the rating matrix through its transformation into two more compact matrices of latent factors of users and items. The inner product of these matrices encloses user preferences for items [1,86].

Used MF approaches:

- Deep matrix factorization (DMF): This method benefits from neural network architecture by constructing a user–item matrix containing explicit ratings and non-preference implicit feedback. The matrix is the input for learning common low-dimensional space for the deep structure learning architecture. To optimize this method, a new loss function based on binary cross entropy is introduced, which considers explicit and implicit feedback. Compared with other conventional models, DMF shows better accuracy in top-K recommendations by using implicit feedback, hence, reconstructing the ratings of users through learning hidden structures from explicit ratings. In addition, DMF supports two-channel structures for combining side information from both users and items. Several studies show that DMF methods provide good accuracy and high efficiency [87–89].
- Neural collaborative filtering (NeuMF): In this method, a neural network architecture replaces the inner product of user–item interaction in CF models. Neural network-based collaboration (NCF) is an approach that generalizes matrix factorization and can be improved by using non-linear kernels. To do so, a multi-layer perceptron is used to learn the user–item interaction function. The use of general NCF in NeuMF can be useful for the combination of different models and using the side information. Studies indicate that acceptable accuracy of deep neural networks comes from their good capacity and nonlinearity [85,90].

3. GNN-based.
Graph learning (GL), machine learning applied to graph structure data, is a rapidly developing technology with good capabilities [23]. Graph Learning-based recommender system(s) (GLRS) are introduced by using relational data in this structure [8]. In real-world systems and applications, objects are often connected either implicitly or explicitly, forming a graph structure. In the field of RS, users, items, attributes, and context are the objects of the structure. They are strongly connected and influence each other via various relations. Using graph techniques can significantly improve the quality of RS. In addition, GL has both a great capacity to learn complex relations and a strong potential for seizing knowledge encapsulated in different types of graphs [56]. Diverse relations in RS can be comprehended with entities, including users, items, and attributes. A wide variety of graphs can be used to represent these entities. To implement a conventional RS, user, item, and interaction between them are sufficient, but to evaluate various metrics and enhance the fairness and accuracy of the model, other information regarding users and/or items can be used. This information can be categorized into two main groups, namely user–item interaction data (user ratings, clicks, purchases. . .), and side information data (user and item attributes). The first group can be further categorized according to whether the interactions are sequential or general [56]. Additional sub-classifications of each group are shown in Table 3.

Table 3. A summary of data representations in RS and the representing graph [56].

Data Class	Data Subclass	Representing Graph
General interaction	Explicit interaction, implicit interaction	Weighted bipartite graph unweighted bipartite graph
Sequential interaction	Single-type interactions Multi-type interactions	Directed homogeneous graph Directed heterogeneous graph
Side information	Attribute information, social information, external knowledge	Heterogeneous graph homogeneous graph tree or heterogeneous graph

Information regarding the type of interaction that happens between users and items represented in the user–item matrix. Data from the mentioned interaction can be either explicit or implicit. The first type occurs when users express their opinions about items directly (e.g., ratings on items). Implicit information is induced from the actions of the user in the interaction with the system (e.g., click, view) [56,91].

The GNN Methods used.

- LightGCN: LightGCN is a simplified variant of the graph convolution network (GCN) containing the most essential components of GCN for recommendation tasks. This method involves the linear propagation of the user and item embeddings on the user–item interaction graph. Moreover, the calculation of the final embedding is performed by using the weighted sum of the embeddings that are learned throughout all layers [92]. LightGCN adopts the same symmetric normalization used in standard GCN for controlling the expansion size of embeddings by means of graph convolution operations. Several studies have indicated the good performance of LightGCN in comparison to traditional approaches [93,94].
- Neural graph collaborative filtering (NGCF): The method represents user–item interactions as a graph structure, which is used to create high-order connectivity and generate embeddings on it. In this structure, the collaborative signal is explicitly transferred through the process of embedding [56]. In addition, multiple embedding propagation layers with concatenated outputs are used to finally predict items to be recommended. NGCF is known for its good performance with respect to model optimization [95,96].
- Self-supervised graph learning (SGL): SGL is a version of GCN optimized for improving accuracy and robustness. This method is very noise tolerant and outperforms other models in this matter. An enhanced classical supervised recommendation task with support for the self-supervised task is used in this method to reinforce the learning of node representation via self-discrimination. Various views of a node can be generated in this structure, which maximizes the agreement between the different views of the same node compared to the views of other nodes. Many strides show that the SGL method performs very well in RS tasks with respect to the accuracy of the results [21,97–99].
- Disentangled graph collaborative filtering (DGCF): This method is designed to focus more on user–item relationships by extracting factors and producing disentangled representations. DGCF models distribution for each user–item interaction and iteratively filters the intent-aware interaction graphs and representations while maintaining the independence of different intents. Although executing DGCF is very time-consuming, this method outperforms many state-of-the-art models [100–102].

3.4. Evaluation Metrics

There are a large number of metrics in the literature that can be used for bias assessment in recommender systems. The choice depends largely on the objectives and the type of discrimination to be detected. In this section, some of them will be analyzed with a focus on user-centric fairness metrics. Those used in the experimental study will be described in detail, aiming at reaching a comprehensive understanding of the behavior of the models regarding bias and fairness. In the case of rank metrics aimed at recommendation lists, different values of K were used to select the top-K-ranked items of the list, where K represents the size of the list.

Before describing the metrics, we will introduce the notation used (Table 4).

Since one of our objectives is to test whether the reduction of biases has a negative impact on the accuracy and precision of the recommendations, metrics for their evaluation were used in this study. Table 5 shows the main metrics in this category. These are rank-based metrics, since the evaluation will be performed on top-K item recommendation lists.

Table 4. Table of notations.

Notation	Definition
U	A set of users
I	A set of items
u	A user
i, j	An item
$R(u)$	A ground-truth set of items that user u interacted with
$\hat{R}(u)$	A ranked list of items that a model produces
K	The length of the recommendation list
$M(x)$	Algorithmic mechanism for RS with input x and output y
θ	Distribution which generates x
Θ	A set of distributions of θ which generate each instance x
g_i	a variable indicating which group the ith user belongs to
$E_g[y]_j$	The average predicted score for the jth item from disadvantaged users
$E_{\neg g}[y]_j$	The average predicted score for the jth item from advantaged users
$E_g[r]_j$	The average ratings for the disadvantaged users
$E_{\neg g}[r]_j$	The average ratings for the advantaged users

Table 5. Metrics for performance evaluation.

Metric Name	Description
Mean Reciprocal Rank (MRR)	It is calculated for the first relevant element found in the top-K item list. Let $Rank_u^*$ be the position of that element in the list provided by a given algorithm for the user u. $$MRR@K = \frac{1}{\|U\|}\sum_{u\in U}\frac{1}{Rank_u^*}$$
Normalized Discounted Cumulative Gain (NDCG)	Discounted cumulative gain (DCG) is a metric applicable to relevant items in the list, where the graded relevance of the items is penalized logarithmically as their position descends in the list. The cumulative gain is calculated up to a rank K. NDCG is the ratio between DCG and the maximum possible DCG. $\delta(0)$ is an indicator function. $$NDCG@K = \frac{1}{\|U\|}\sum_{u\in U}\frac{1}{\sum_{i=1}^{min(\|R(u)\|,K)}\frac{1}{log_2(i+1)}}\sum_{i=1}^{K}\delta(i\in R(u))\frac{1}{\log_2(i+1)}$$
Precision	A well-known measure for ranked lists, which represents the fraction of relevant items out of all the recommended items. Usually expressed as the average of the metric values for each user. $\|\hat{R}(u)\|$ represents the item count of $\hat{R}(u)$ $$Precision@K = \frac{1}{\|U\|}\sum_{u\in U}\frac{\|\hat{R}(u)\cap R(u)\|}{\|\hat{R}(u)\|}$$
Recall	It is a measure similar to precision, but in this case, its value is the ratio between relevant items in the top-K recommendation list and all relevant items. $\|R(u)\|$ represents the item count of $R(u)$ $$Recall@K = \frac{1}{\|U\|}\sum_{u\in U}\frac{\|\hat{R}(u)\cap R(u)\|}{\|R(u)\|}$$
Hit Ratio (HR)	This metric evaluates how many 'hits' were included in a top-K item list. A hit is an item that appears in the ground-truth set. $\delta(0)$ is an indicator function. $\delta(b) = 1$ if b is true, otherwise it would be 0. $\emptyset$ denotes the empty set. $$HR@K = \frac{1}{\|U\|}\sum_{u\in U}\delta(\hat{R}(u)\cap R(u) \neq \emptyset)$$

As mentioned earlier, the main focus of this research is on analyzing bias and unfairness measurement in the GNN-based RS field. To assess the general popularity bias three different metrics (Gini index, item coverage, and average popularity) are used, as can be seen in Table 6. However, our study is more focused on quantifying unfairness that has an impact on the user, especially discrimination toward disadvantaged groups identified through certain sensitive attributes (gender and age). To this end, specific metrics were introduced for this purpose. These are differential fairness (DF), value unfairness, absolute unfairness, and non-parity. Their definition can be found in Table 6.

Average popularity evaluates the effect on recommendations of the long-tail distribution. It takes into account the popularity of the recommended items. In addition, The Gini index is a very appropriate metric in this context because its purpose is to measure inequalities. In this study, it is applied to assess the diversity of the recommended items. Its values lie between 0 and 1, where 0 represents the equality of the distribution while 1 represents the maximum inequality. Another metric related to the popularity of the items is Item Coverage, which depicts the percentage of recommended items in relation to the total number of items.

Differential Fairness is used to check for differences in the recommendations for the groups with different values of the sensitive attributes (gender and age in our study). Both intersectionality and behavior toward minorities are considered in DF [103,104]. The first refers to fairness for each of the protected attributes individually, and the second affects the aforementioned anti-discrimination laws.

Value unfairness quantifies the deviations in the predicted scores above or below the true values for different groups of users. If in all groups the deviation is similar or the overestimation and underestimation are in balance with each other, Value unfairness turns small. If there is a repeated overestimation for one group and an underestimation for another, the metric values are small too [105,106].

Table 6. Metrics for bias measurement.

Metric Name	Description
Average Popularity	Measures the average popularity of the items included in the recommendations. It is calculated as follows, where $\phi(i)$ is the number of interactions or ratings on item i in the training data [107]. $$AveragePopularity@K = \frac{1}{\vert U\vert}\sum_{u\in U}\frac{\sum_{i\in R_u}\phi(i)}{\vert R_u\vert}$$
Gini index [51]	Measures the diversity of recommended items. The calculation is shown below, where $P(i)$ indicates the number of occurrences of the item i and the recommendation list, which is indexed in non-decreasing order [108]. $$GiniIndex@K = \frac{\sum_{i=1}^{\vert I\vert}(2i-\vert I\vert-1)P(i)}{\vert I\vert\sum_{i=1}^{\vert I\vert}P(i)}$$
Item Coverage	Represents the percentage of recommended items over all items [109] $$ItemCoverage@K = \frac{\vert\cup_{u\in U}\hat{R}(u)\vert}{\vert I\vert}$$
Differential fairness (DF) for gender (a sensitive attribute) [103,110]	Measures the bias in the recommendations received by the protected groups. An algorithm or mechanism $M(x)$ is ϵ-differentially fair with respect to (A,θ) if for all $\theta\in\Theta$ with $x\sim\theta$, and $y\in Range(M)$. For all $(s_i,s_j)\in A\times A$ where $P(s_i)>0$, $P(s_j)>0$. $s_i,s_j\in A$ are tuples of all protected attribute values. $$e^{-\epsilon}\leq\frac{P_{M,\theta}(M(x)=y\vert s_i,\theta)}{P_{M,\theta}(M(x)=y\vert s_j,\theta)}\leq e^{\epsilon}$$

Table 6. *Cont.*

Metric Name	Description
Value Unfairness [105,106,111]	Computes instability in signed estimation error over user types $$U_{val} = \frac{1}{I}\sum_{j=1}^{I} \lvert (E_g[y]_j - E_g[r]_j) - (E_{\neg g}[y]_j - E_{\neg g}[r]_j) \rvert$$ where $E_g[y]_j$ is the average predicted score fir the jth item for protected groups. $$E_g[y]_j := \frac{1}{\lvert i : ((i,j) \in X) \wedge g_i \rvert} \sum_{i:((i,j)\in X)\wedge g_i} y_{ij}$$
Absolute Unfairness [105,106]	Measures differences in absolute estimation error over groups of users $$U_{abs} = \frac{1}{I}\sum_{j=1}^{I} \lvert \lvert E_g[y]_j - E_g[r]_j \rvert - \lvert E_{\neg g}[y]_j - E_{\neg g}[r]_j \rvert \rvert$$
Non-Parity Unfairness [105,106]	It is the absolute difference between the rating average of protected and unprotected groups $$U_{par} = \lvert E_g[y] - E_{\neg g}[y] \rvert$$

Another used metric for sensitive features is absolute unfairness, which only reflects the magnitude of the deviations of the predicted scores for items regardless of whether they are positive or negative [105,106]. This avoids compensating for overestimates and underestimates. If the estimation error is much greater for one group than for another, there will be discriminatory or unfair treatment for users who receive poorer recommendations.

To achieve more accurate results of unfairness for sensitive features, Underestimation unfairness is added to the setting where not recommending items the user likes is more important than recommending items the user dislikes. In contrast, overestimation unfairness is added in the setting where users are overwhelmed by recommendations, considering too many recommendations can be harmful.

Another important measure of fairness is non-parity unfairness. This metric is calculated as the absolute difference between the average of the ratings of the different groups [105,106].

Table 6 shows the metrics used in this study to evaluate the sensitivity of the models to the most relevant types of biases in recommender systems.

4. Results of the Experimental Study

This section presents the values of the bias metrics obtained by applying the recommendation methods described in the previous section on the three datasets with sensitive attributes, whose exploratory study was presented previously. These real-world datasets are MovieLens 100K, LastFM 100K, and book recommendation. The detailed results allow us to test not only whether GNN-based recommendation methods amplify biases in the data compared to more traditional approaches, but whether their behavior is similar across all types of biases. Since this work is especially focused on user-centered fairness, the main objective is to check the performance of GNN-based methods with respect to appropriate metrics for this type of bias. An important aspect of the analysis is to find out which models provide a better balance between accuracy and bias sensitivity since the improvement of one of them usually has a negative influence on the other.

First, the results of the metrics applicable to top-K item recommendation lists are shown. They were obtained for different values of K (5, 10, and 15). Subsequently, the results of the fairness metrics applicable to rating prediction are presented.

Tables 7–9 contain the results of the list metrics obtained from the application of the eight recommendation methods on MovieLens, LastFM, and book recommendation datasets, respectively.

Table 7. Results of the MovieLens dataset.

Approach	Method	Top K	Recall	Precision	MRR	NDCG	HIT	Item Coverage	Gini Index	Average Popularity
CF	ItemKNN	K = 5	0.15	0.23	0.44	0.28	0.63	0.19	0.93	231.96
CF	ItemKNN	K = 10	0.22	0.18	0.46	0.27	0.75	0.24	0.93	249.74
CF	ItemKNN	K = 15	0.31	0.16	0.46	0.29	0.84	0.29	0.89	208.12
CF	NNCF	K = 5	0.15	0.24	0.47	0.29	0.64	0.17	0.95	284.47
CF	NNCF	K = 10	0.24	0.19	0.46	0.22	0.78	0.25	0.91	217.70
CF	NNCF	K = 15	0.28	0.15	0.47	0.27	0.81	0.30	0.91	231.28
MF	DMF	K = 5	0.14	0.22	0.43	0.26	0.62	0.18	0.94	256.29
MF	DMF	K = 10	0.21	0.17	0.42	0.25	0.73	0.20	0.93	252.25
MF	DMF	K = 15	0.29	0.16	0.45	0.28	0.83	0.28	0.90	219.49
MF	NeuMF	K = 5	0.15	0.23	0.45	0.27	0.65	0.25	0.91	228.52
MF	NeuMF	K = 10	0.23	0.18	0.46	0.27	0.78	0.36	0.89	212.41
MF	NeuMF	K = 15	0.30	0.16	0.46	0.28	0.83	0.40	0.86	196.89
GNN	NGCF	K = 5	0.15	0.24	0.48	0.29	0.66	0.15	0.95	277.85
GNN	NGCF	K = 10	0.25	0.20	0.49	0.30	0.77	0.25	0.93	255.49
GNN	NGCF	K = 15	0.32	0.17	0.49	0.31	0.86	0.32	0.89	219.13
GNN	LightGCN	K = 5	0.11	0.17	0.36	0.21	0.55	0.05	0.98	245.13
GNN	LightGCN	K = 10	0.18	0.14	0.37	0.21	0.67	0.07	0.97	312.47
GNN	LightGCN	K = 15	0.23	0.12	0.38	0.21	0.76	0.10	0.96	292.8
GNN	SGL	K = 5	0.15	0.25	0.47	0.29	0.66	0.24	0.91	229.24
GNN	SGL	K = 10	0.25	0.20	0.49	0.29	0.80	0.31	0.89	209.39
GNN	SGL	K = 15	0.31	0.17	0.49	0.30	0.85	0.34	0.88	200.63
GNN	DGCF	K = 5	0.12	0.15	0.30	0.18	0.55	0.27	0.94	269.46
GNN	DGCF	K = 10	0.22	0.14	0.40	0.23	0.74	0.17	0.96	278.14
GNN	DGCF	K = 15	0.26	0.11	0.36	0.22	0.82	0.46	0.90	232.08

Table 8. Results of the LastFM dataset.

Approach	Method	Top K	Recall	Precision	MRR	NDCG	HIT	Item Coverage	Gini Index	Average Popularity
CF	ItemKNN	K = 5	0.10	0.06	0.13	0.17	0.30	0.10	0.94	25.90
CF	ItemKNN	K = 10	0.14	0.04	0.12	0.19	0.45	0.17	0.90	19.19
CF	ItemKNN	K = 15	0.21	0.03	0.46	0.29	0.84	0.23	0.85	15.33
CF	NNCF	K = 5	0.10	0.09	0.29	0.33	0.45	0.10	0.94	39.79
CF	NNCF	K = 10	0.15	0.03	0.30	0.37	0.57	0.17	0.90	25.46
CF	NNCF	K = 15	0.17	0.03	0.47	0.38	0.64	0.23	0.85	18.91
MF	DMF	K = 5	0.12	0.09	0.29	0.33	0.46	0.10	0.94	40.21
MF	DMF	K = 10	0.16	0.05	0.31	0.37	0.58	0.17	0.90	20.09
MF	DMF	K = 15	0.17	0.04	0.45	0.39	0.65	0.23	0.85	18.10
MF	NeuMF	K = 5	0.11	0.09	0.29	0.33	0.46	0.10	0.94	39.89
MF	NeuMF	K = 10	0.17	0.05	0.30	0.37	0.58	0.17	0.90	25.49
MF	NeuMF	K = 15	0.20	0.04	0.46	0.39	0.64	0.23	0.85	18.77
GNN	NGCF	K = 5	0.08	0.05	0.16	0.19	0.28	0.77	0.66	22.23
GNN	NGCF	K = 10	0.11	0.03	0.17	0.22	0.37	0.95	0.56	14.87
GNN	NGCF	K = 15	0.14	0.02	0.49	0.23	0.42	0.99	0.48	11.44
GNN	LightGCN	K = 5	0.09	0.05	0.15	0.18	0.25	0.33	0.85	22.18
GNN	LightGCN	K = 10	0.12	0.03	0.16	0.19	0.30	0.70	0.97	13.66
GNN	LightGCN	K = 15	0.17	0.02	0.16	0.20	0.35	0.89	0.53	10.04
GNN	SGL	K = 5	0.07	0.03	0.09	0.11	0.16	0.83	0.49	13.16
GNN	SGL	K = 10	0.09	0.02	0.49	0.13	0.28	0.31	0.88	11.46
GNN	SGL	K = 15	0.13	0.03	0.49	0.15	0.34	0.91	0.40	9.56
GNN	DGCF	K = 5	0.07	0.03	0.13	0.14	0.19	0.22	0.83	18.64
GNN	DGCF	K = 10	0.11	0.01	0.08	0.12	0.20	0.45	0.65	9.16
GNN	DGCF	K = 15	0.13	0.01	0.13	0.16	0.27	0.59	0.57	8.12

Table 9. Results of the book recommendation dataset.

Approach	Method	Top K	Recall	Precision	MRR	NDCG	HIT	Item Coverage	Gini Index	Average Popularity
CF	ItemKNN	K =5	0.19	0.01	0.05	0.08	0.19	0.13	0.93	2.44
CF	ItemKNN	K = 10	0.19	0.01	0.05	0.08	0.2	0.21	0.89	2.23
CF	ItemKNN	K = 15	0.17	0.03	0.45	0.39	0.65	0.23	0.85	18.1
CF	NNCF	K = 5	0.15	0.03	0.08	0.1	0.16	0.08	0.95	5.08
CF	NNCF	K = 10	0.22	0.02	0.09	0.12	0.24	0.15	0.9	4.04
CF	NNCF	K = 15	0.26	0.04	0.1	0.13	0.29	0.21	0.87	3.4
MF	DMF	K = 5	0.17	0.03	0.12	0.12	0.18	0.08	0.95	5.85
MF	DMF	K = 10	0.19	0.02	0.13	0.13	0.22	0.17	0.91	4.12
MF	DMF	K = 15	0.21	0.03	0.13	0.14	0.24	0.19	0.87	3.26
MF	NeuMF	K = 5	0.18	0.03	0.13	0.14	0.2	0.08	0.95	6.13
MF	NeuMF	K = 10	0.24	0.02	0.13	0.15	0.26	0.14	0.9	4.35
MF	NeuMF	K = 15	0.27	0.04	0.14	0.16	0.29	0.2	0.86	3.49
GNN	NGCF	K = 5	0.04	0.01	0.02	0.02	0.04	0.34	0.71	1.28
GNN	NGCF	K =10	0.08	0.01	0.02	0.03	0.09	0.57	0.56	1.28
GNN	NGCF	K = 15	0.13	0.02	0.03	0.05	0.12	0.72	0.48	1.27
GNN	LightGCN	K = 5	0.04	0.01	0.02	0.02	0.04	0.27	0.79	2.08
GNN	LightGCN	K = 10	0.05	0.01	0.02	0.02	0.05	0.49	0.64	1.62
GNN	LightGCN	K = 15	0.08	0.02	0.03	0.04	0.08	0.63	0.54	1.54
GNN	SGL	K = 5	0.03	0.01	0.01	0.02	0.03	0.27	0.83	1.62
GNN	SGL	K = 10	0.07	0.02	0.02	0.04	0.11	0.39	0.74	1.63
GNN	SGL	K = 15	0.1	0.03	0.02	0.11	0.34	0.5	0.68	1.59
GNN	DGCF	K = 5	0.03	0.01	0.01	0.02	0.04	0.29	0.78	1.57
GNN	DGCF	K = 10	0.01	0.01	0.02	0.03	0.08	0.51	0.63	1.4
GNN	DGCF	K = 15	0.09	0.01	0.02	0.04	0.11	0.67	0.53	1.36

The results of the fairness metrics based on sensitive attributes are shown below. These metrics are differential fairness (DF), value unfairness (VU), absolute unfairness (AV), and non-parity unfairness. The gender attribute was studied in the MovieLens and LastFM datasets (Tables 10 and 11) and the age attribute in the three datasets (Tables 12–14). The values considered for gender were male and female, and for age—under and equal to 30 years old and over 30 years old.

Table 10. Results of fairness metrics for gender in the MovieLens dataset.

Approach	Method	Avg. DF	VU	AU	Non-Parity Unfairness
CF	ItemKNN	3.2702	2.0264	2.0195	1.8335
CF	NNCF	1.5023	0.4967	0.4523	0.0613
MF	DMF	2.3341	0.233	0.1685	0.0211
MF	NeuMF	2.188	0.2113	0.1413	0.0072
GNN	NGCF	1.5023	0.4967	0.4523	0.0613
GNN	LightGCN	1.4138	0.2159	0.1851	0.0233
GNN	SGL	0.6419	0.2152	0.2148	0.0002
GNN	DGCF	0.6855	0.2116	0.2103	0.0015

Table 11. Results of fairness metrics for gender in the LastFM dataset.

Approach	Method	Avg. DF	VU	AU	Non-Parity Unfairness
CF	ItemKNN	0.7052	0.3101	0.3101	0
CF	NNCF	5.9771	0.1507	0.1491	0.0016
MF	DMF	6.0327	0.1486	0.1476	0.0032
MF	NeuMF	5.0174	0.2228	0.222	0.0003
GNN	NGCF	2.4747	0.351	0.3422	0.002
GNN	LightGCN	0.6764	0.3094	0.3094	0.0001
GNN	SGL	0.7263	0.3032	0.3032	0
GNN	DGCF	0.7287	0.4249	0.4249	0.0004

Table 12. Results of fairness metrics for age range in the MovieLens dataset.

Approach	Method	Avg. DF	VU	AU	Non-Parity Unfairness
CF	ItemKNN	2.7436	1.6859	1.6808	0.5607
CF	NNCF	1.9234	0.1935	0.1233	0.5607
MF	DMF	2.1169	0.2116	0.1434	0.0022
MF	NeuMF	2.188	0.2113	0.1413	0.0049
GNN	NGCF	1.2998	0.3895	0.3456	0.032
GNN	LightGCN	1.0907	0.2051	0.1752	0.003
GNN	SGL	0.6419	0.2152	0.2148	0.0002
GNN	DGCF	0.6509	0.2011	0.2008	0

Table 13. Results of fairness metrics for age range in the LastFM dataset.

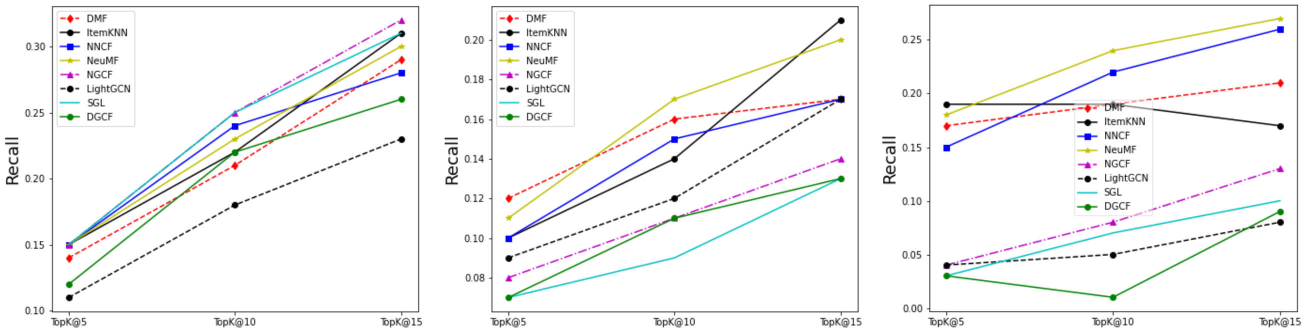
Approach	Method	Avg. DF	VU	AU	Non-Parity Unfairness
CF	ItemKNN	0.8071	0.2865	0.2865	0
CF	NNCF	6.4780	0.1405	0.1397	0.0138
MF	DMF	0.7182	0.4294	0.4294	0.0164
MF	NeuMF	5.4643	0.2033	0.2029	0.0185
GNN	NGCF	2.8171	0.3309	0.3262	0.0597
GNN	LightGCN	0.7237	0.2943	0.2943	0.0003
GNN	SGL	0.7504	0.2903	0.2903	0
GNN	DGCF	0.6960	0.431	0.4323	0.0004

Table 14. Results of fairness metrics for age range in the book recommendation dataset.

Approach	Method	Avg. DF	VU	AU	Non-Parity Unfairness
CF	ItemKNN	0.7875	0.4928	0.4927	0.0034
CF	NNCF	7.2301	0.4794	0.4417	0.0001
MF	DMF	7.3435	0.4814	0.4426	0.0058
MF	NeuMF	4.3972	0.5739	0.565	0.0093
GNN	NGCF	5.1748	0.5609	0.541	0.0004
GNN	LightGCN	0.8551	0.4965	0.4964	0.0002
GNN	SGL	0.5654	0.4931	0.4931	0.00015
GNN	DGCF	0.6408	0.494	0.494	0.0004

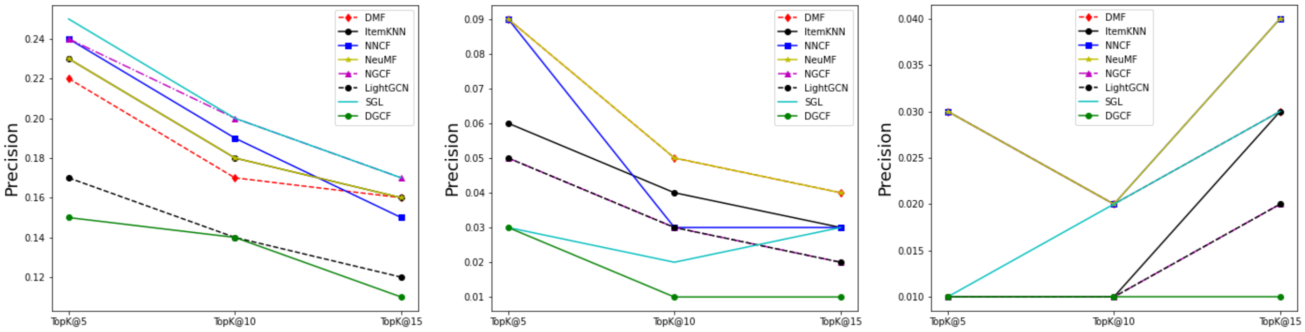
To facilitate the comparative analysis of the results and to obtain a better insight into the behavior of the algorithms, the data in the tables are graphically represented in the following figures.

First, the metrics related to the quality of the recommendation lists are shown. Figure 11 shows recall values of all algorithms for item recommendation lists with three different sizes (values of K). In the same way, the precision results are displayed in Figure 12. Figure 13 illustrates the values of the MRR metric, Figure 14 shows those of the HIT measure, and Figure 15 shows NDCG.



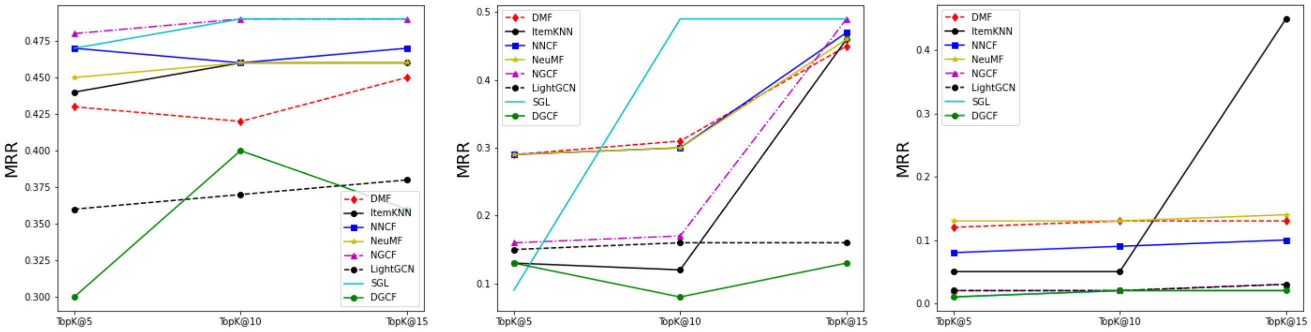
(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

Figure 11. Recall results for the three datasets.



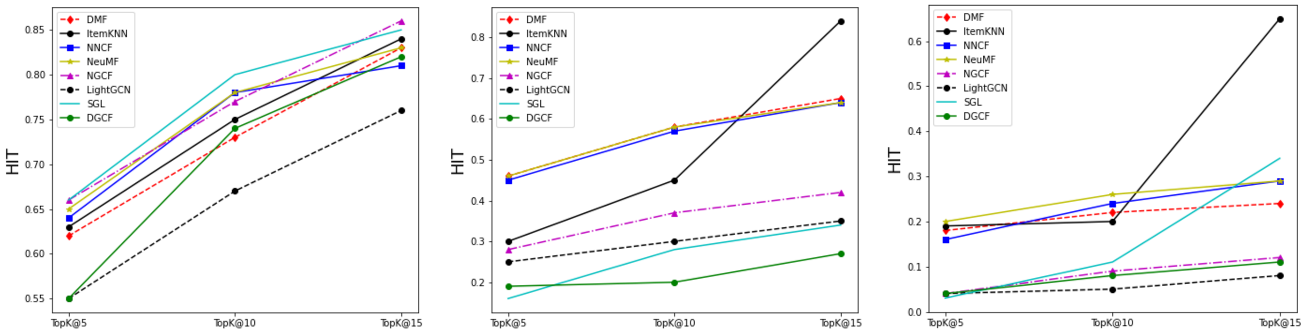
(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

Figure 12. Precision results for the three datasets.



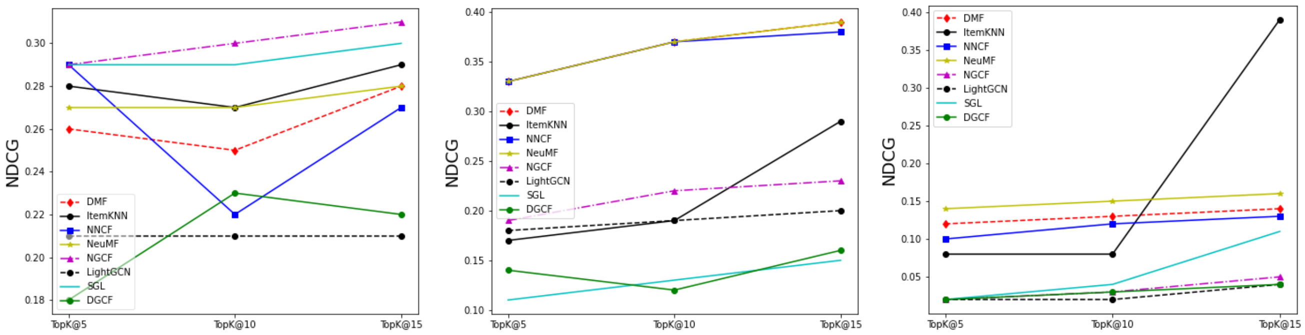
(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

Figure 13. MRR results for the three datasets.



(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

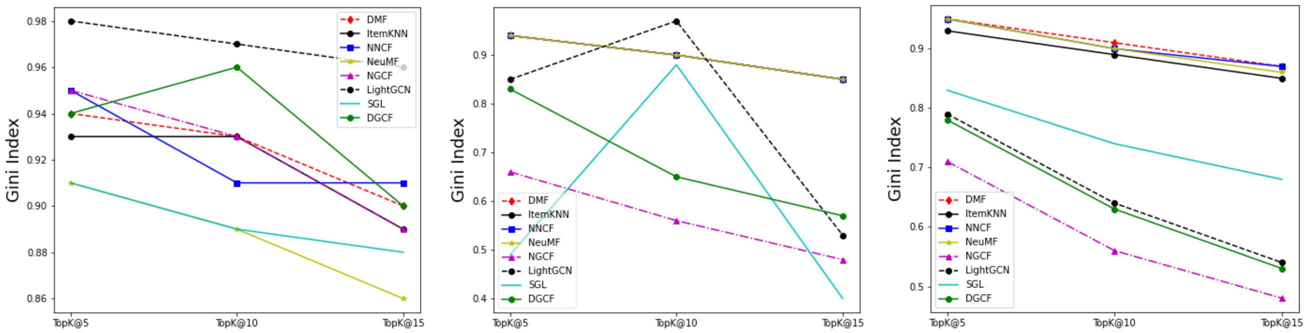
Figure 14. Results of HIT for the three datasets.



(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

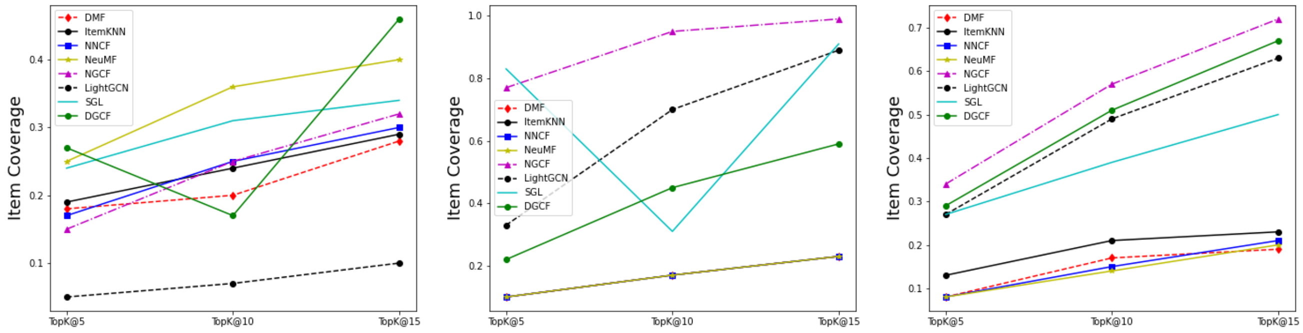
Figure 15. NDCG results for the three datasets.

Although this study is more oriented to measuring biases that may have impacts on the unfair treatment of users, we analyze biases related to the popularity and diversity of recommendations since they affect users in individual ways. Within this category, we include the Gini index, which measures the diversity of the distribution of items in the recommendation lists. Likewise, the classic metrics of coverage and average popularity are considered in this group. The Gini index is shown in Figure 16, item coverage in Figure 17, and average popularity in Figure 18.



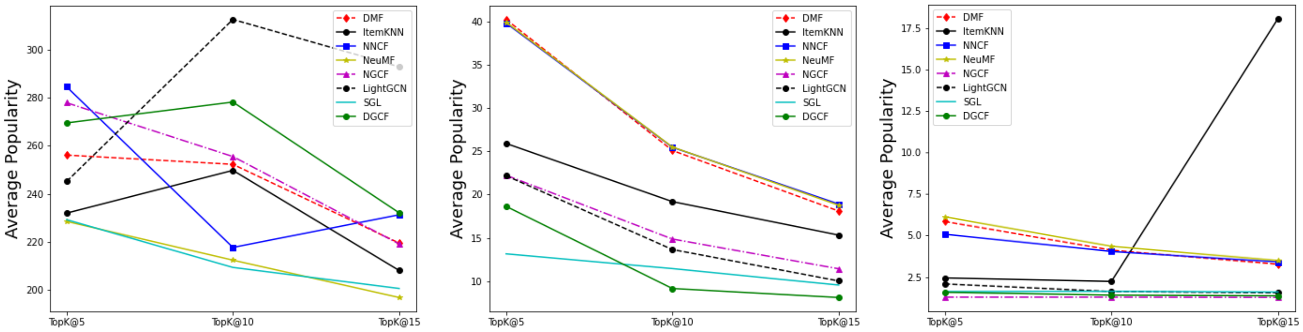
(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

Figure 16. Results of the Gini index for the three datasets.



(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

Figure 17. Item coverage results for the three datasets.



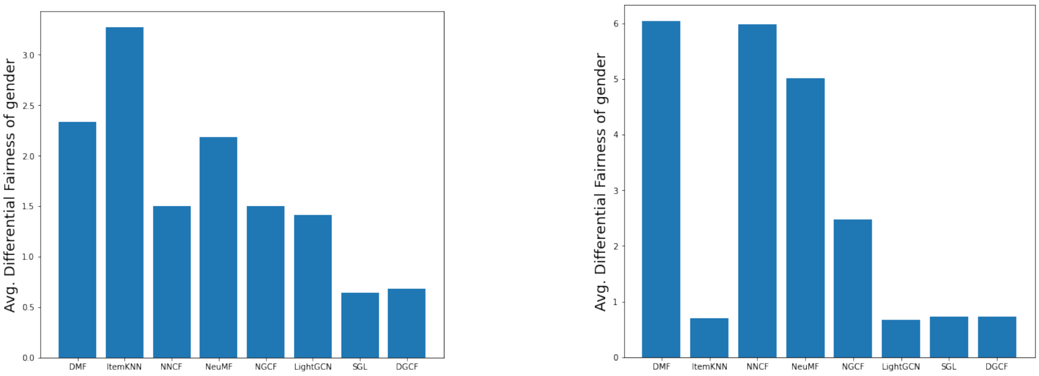
(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

Figure 18. Average popularity results for the three datasets.

Finally, we present the graphs corresponding to the metrics aimed at assessing the fairness of recommendations in user groups. In our study, the groups are based on gender and age-sensitive attributes. All these metrics are computed from the ratings predicted by the models; therefore, they do not apply to recommendation lists.

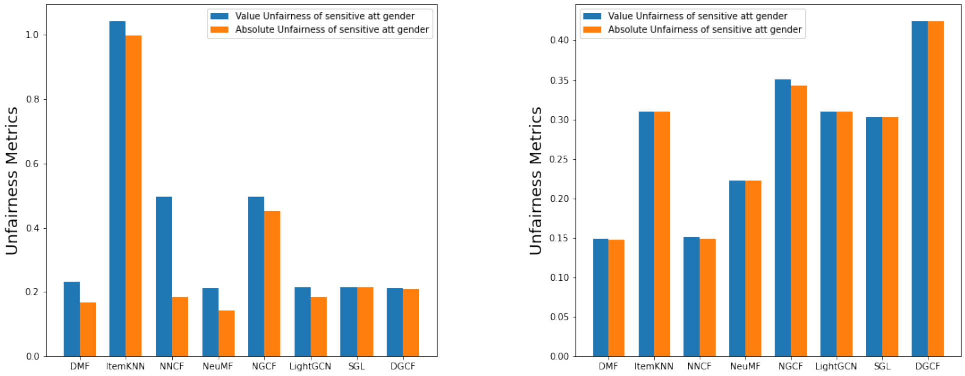
Below, the results of fairness metrics for the gender attribute are provided. These results are the outcome of this experiment on MovieLens and LastFM datasets, which are the two datasets that have this attribute.

We start with the visualization of the Differential Fairness values in Figure 19. Next, Figure 20 illustrates the absolute unfairness and value unfairness results. Last, Figure 21 shows non-parity unfairness.



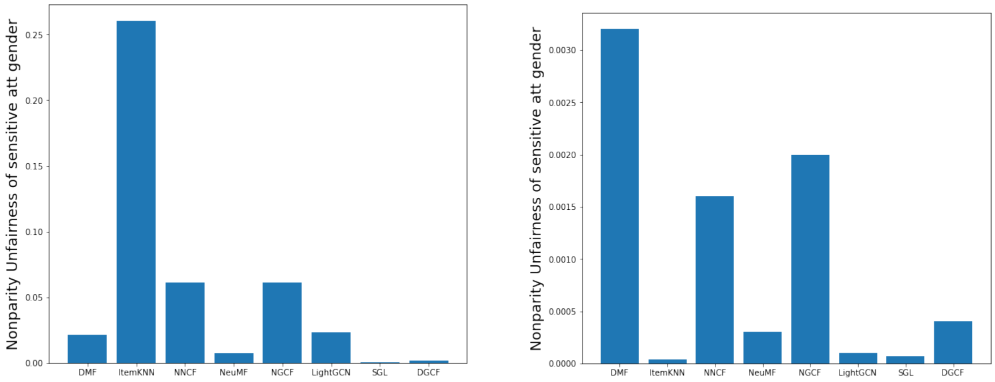
(**a**) MovieLens. (**b**) LastFM.

Figure 19. Results of the differential fairness of the sensitive attribute gender for two datasets.



(**a**) MovieLens. (**b**) LastFM.

Figure 20. Value and absolute unfairness results of gender for MovieLens and LastFM.

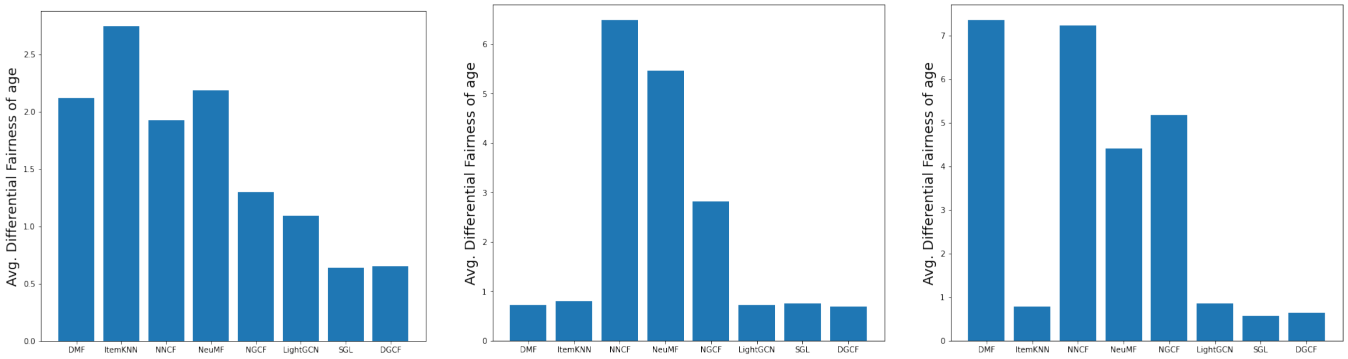


(**a**) MovieLens. (**b**) LastFM.

Figure 21. Non-parity Unfairness results of gender for MovieLens and LastFM.

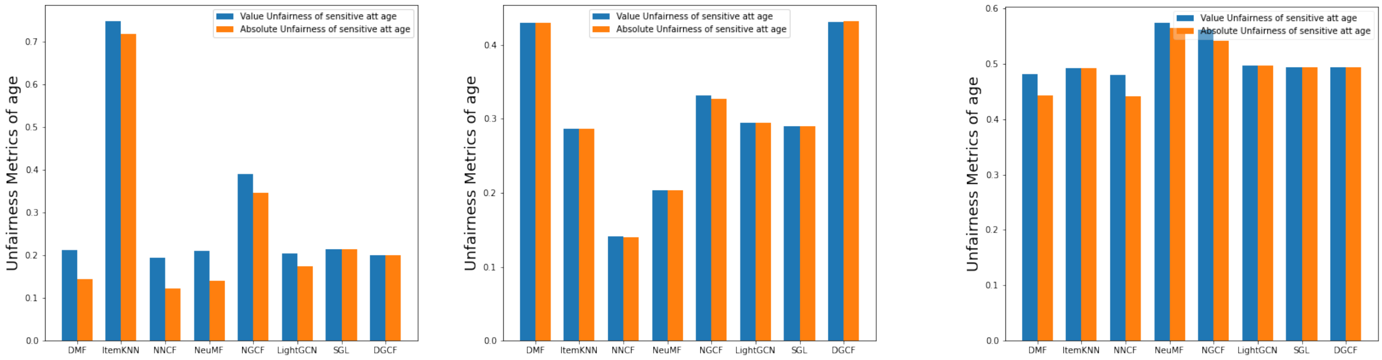
After presenting the values of the fairness metrics, considering the gender-sensitive attribute, we move on to the visualization of the results corresponding to the age attribute, whose values were divided into two intervals. This last attribute is present in the three datasets studied.

Figure 22 shows the results of differential fairness, Figure 23 shows those corresponding to value unfairness and absolute unfairness, and Figure 24 shows non-parity unfairness values.



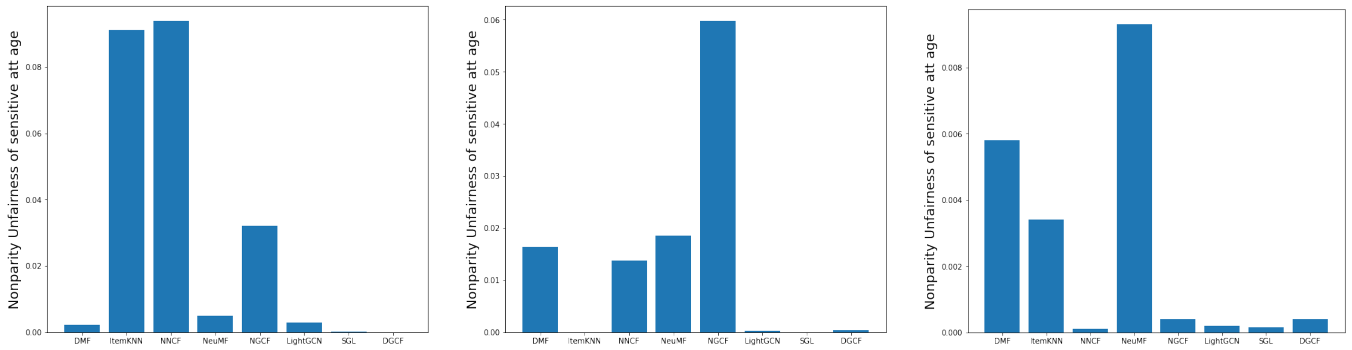
(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

Figure 22. Differential fairness results of age for three datasets.



(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

Figure 23. Value and absolute unfairness results of binary age for three datasets.



(**a**) MovieLens. (**b**) LastFM. (**c**) Book recommendation.

Figure 24. Non-parity Unfairness results of age for three datasets.

5. Discussion of Results

Recent literature studies have addressed the problem of bias and fairness in recommender systems, although the focus on GNN-based methods is more limited. Some studies on GNN application in other domains have concluded that these approaches increase performance (in terms of prediction quality) but accentuate biases compared to other methods. However, it is not known whether these conclusions can be extended to the field of RS and all types of biases specific to this area. In this section, we intend to answer these and other research questions formulated in the introduction to this work, after analyzing the results of the extensive study presented in the previous sections. This will allow gaining insight into an issue of such relevance in the context of recommender systems.

In this section, we will analyze the performance against different types of biases of GNN-based recommendation methods in comparison with other classical methods and analyze the impact on the quality (precision, recall, etc.) of the recommendation lists since one of the goals of bias mitigation is to keep the accuracy as high as possible.

Since our study is mainly aimed at user-centered fairness, we will first differentiate between individual-level metrics and group-level metrics. In the former category are the Gini index, item coverage, and average popularity, and in the latter—differential fairness (DF), value unfairness, absolute unfairness, and non-parity. These will be discussed in terms of their appropriateness for the RS field.

One observation of this study is the disparate behavior of the algorithms with the different datasets in terms of quality metrics for item recommendation lists (precision, recall, NDCG). This is to be expected because it has been shown that the accuracies of GNN-based methods largely depend on the characteristics of the datasets. In our case, the number of records of all the datasets is similar, but they vary in terms of the number of users and items to be recommended.

Regarding the bias metrics at the individual level, the results between datasets vary, although not as much as in the previous metrics.

In the context of our work, the Gini index should have high values, which express high inequality and therefore great diversity in the recommendations. The results of our study show that an increase in the values of this metric has an impact on a decrease in precision and other quality metrics of the recommendations. LightGCN is the model that provides the highest Gini values in the MovieLens and LastFM datasets and median values in the Books dataset. However, with this modelm the lowest values of NDCG, HIT, and MRR are obtained in the MovieLens and Book datasets while medium-low values of these metrics are obtained in the LastFM dataset. The behavior in relation to precision and recall is similar. In contrast, NGCF, which is the model that yields the best accuracy and recall results in MovieLens, has medium-high Gini values in this dataset. In the other two datasets, NGCF presents medium-low values of the Gini index and very low precision and recall values. The results for MRR, HIT, and NDCG are similar to those for precision and recall. DGCF gives better Gini results than NGCF but its precision, recall, MRR, HIT, and NDCG values are among the worst. The consequence that can be drawn from this is that the degree of the negative impact of a high Gini value on the accuracy and analogous metric can be very different depending on the algorithm.

Item coverage is another bias metric that affects users individually since the lower the coverage, the lower the probability that the user will receive recommendations of items that they might like. The NGCF models give the highest values of this metric with the LastFM and Book datasets where precision and recall are rather low. Moreover, coverage is quite high with the MovieLens dataset, although with the latter it is surpassed by NeuMF and SGL. With this dataset, the highest precision and recall were achieved. DGCF achieves very good coverage in the Book datasets and medium coverage in the remaining datasets, while, as has already been seen, it is the worst in terms of the quality of the recommendation lists. LightGCN presents the lowest value with the MovieLens dataset and medium–high values with the other two datasets. Therefore, the change in the performance of the models depending on the objective of the evaluation is confirmed here, which can be either the bias or the accuracy of the recommendations.

Regarding the average popularity metric, the goal is to achieve low values so that unpopular items are recommended. The SGL models are the ones with the most uniform behavior in the three datasets, providing very low values for this metric. NGCF achieves the lowest value among all the models with the Book dataset, while its value is medium with MovieLens and LastFM. DGCF models provide good results with the book and LastFM datasets but one of the worst values with MovieLens. In this way, it is confirmed once again that the gain in precision amplifies the biases, but not always to the same degree.

In general, we can say that LightGCN is the algorithm belonging to GNN-based approach that exhibits the most irregular behavior across datasets and list sizes. This uneven performance is found in the classical approaches of collaborative filtering and matrix factorization. Within these classical methods, the one that almost always achieves good precision, recall, MRR, HIT, and NDCG values, is NeuMF, but its results in terms of bias metrics are very irregular.

Regarding the biases at the level of individual users, we can conclude that although the ranking of the models changes depending on whether the quality of the lists or the biases are evaluated, the algorithms based on GNN are not the worst positioned in relation to the biases but some of them reach good Gini index values, coverage, and Average Popularity.

We next examine group-level fairness metrics based on sensitive attributes, such as gender and age. The metrics used in this study are differential fairness (DF), value unfairness (VU), absolute unfairness (AU), and non-parity unfairness. Among them, it is important to differentiate between two different approaches. Within the first are DF and non-parity, whose objective is to find differences in predicted ratings for different user groups. The objectives of VU and AU are different since they are focused on measuring and comparing the quality of recommendations for each group. We consider the latter to be the most appropriate in the domain of recommender systems since the attributes used to create the groups (in our case gender and age) have a proven influence on user preferences. Therefore, the fact that they receive different recommendations need not be unfair or discriminatory.

When analyzing the differential fairness results, all of the GNN-based models generally present lower values than the classical methods with all the datasets and for both the gender and age attributes. There are only a few exceptions where a classical method reaches a value similar to or lower than the GNN-based method. Therefore, the performance of GNN models in relation to this fairness metric is quite good, with NGCF being the worst performer in this category.

Although the results of the Non-Parity metric are somewhat more irregular, the techniques based on GNN, except for NGCF, provide low values of unfairness for both sensitive attributes and almost always lower than the classical algorithms. The results of this metric for the methods in the latter category differ greatly from one dataset to another.

Finally, the results obtained for value and absolute unfairness are quite similar. Unlike what happens with the previous metrics, the results of these metrics with the GNN models are generally worse than with the classical models, except for some occasional exceptions in which some HF method gives the worst result. This behavior occurs for both the gender and age attributes, although the results are better with the MovieLens dataset, especially for the gender attribute.

In relation to the fairness metrics at the group level, we can conclude that precisely the worst behavior of the GNN-based methods occurs with the most appropriate metrics for recommender systems, which are those that are based on the quality of the recommendations. On the contrary, the results of these methods are better for the other type of metrics, which evaluate the similarity in the ratings for the different groups.

Considering all of the results provided, it seems that GNN-based methods have great potential in providing accurate recommendations. It can be concluded that the performances of these methods outperformed the other used models. In addition, among GNN-based approaches, SGL provided higher results on the MovieLens dataset. However, some types of biases may be amplified based on the target dataset and chosen model. In most cases, a higher performance of the model resulted in bias amplification, and unfairness toward disadvantaged groups. This can show the trade-off between accuracy and bias.

Once the results were analyzed, we are in a position to answer the research questions that are the subject of this study. The questions below are followed by findings related to each one.

- RQ1: Can the findings reported in the literature in the general context of machine learning be extended to the specific field of recommender systems? Although the literature states that GNN methods are more prone to bias than other classical techniques, the same cannot be said in the area of recommender systems, since some of these methods perform well against bias while maintaining the accuracy of the recommendations.
- RQ2: Does the performance of GNN-based recommendation methods against biases depend on dataset characteristics and sensitive attributes? The study showed that the fairness metrics present irregular results for the different datasets. For example, the tested algorithms yield totally different values for the gender-related unfairness metrics in the MovieLens and LastFM datasets, with the gender imbalance being very similar in both datasets. This reveals that the bias in the results is highly dependent on other characteristics of the data. This irregular behavior occurs with different sensitivity attributes.
- RQ3: Are all bias evaluation metrics appropriate for assessing user-centered fairness in recommender systems in all application domains? The literature review has allowed us to compile the most commonly used bias metrics, some of which do not assess the quality of the results but rather the similarity of the results themselves for different groups. Because these groups are formed on the basis of sensitive attributes such as gender or age, which were shown to influence preferences, these metrics are not appropriate in the field of recommender systems whose objective is to predict user preferences. In most of the application domains of RS, such as recommendations of movies, music, etc., preferences change according to these attributes. In fact, some methods use them to generate better recommendations.
- RQ4: Do less bias-prone methods always provide lower-quality recommendations? Although the decrease in biases generally results in low values of quality metrics such as precision, NDCG, etc., this does not always happen. Some of the GNN-based recommendation methods present good values both for these last metrics and for the bias metrics, presenting in some cases better behavior against bias than classical methods.

6. Conclusions and Future Work

Bias and fairness problems are some of the most vital issues in RS, which can lead to discrimination and result in heavy different costs for companies and people. In this study, bias and fairness issues in GNN-based RS were taken into consideration. The identification and quantification of these problems can be done through different types of metrics, which can consider unfair treatments for both items that are recommended (for example, discrimination of artists, tourist places, product providers, etc.) and the users that receive the recommendation (discrimination of users based on age, race, gender, etc.). This work focuses on the analysis of biases and unfairness in this last group. Thus, the results of different user-centric bias metrics provided by different recommendation methods were analyzed. Our main objective was to analyze the behaviors of the methods based on GNN against the biases that affect the user and to compare them with the behaviors of collaborative filtering and matrix factorization methods. To obtain a wide range of comparative results, several algorithms of each type were implemented and tested on three real-world datasets.

The chosen datasets are from different types of RS, which can help advance research on this topic in different scales. These datasets contain sensitive attributes that can be sources of unfairness and suffer from the long-tail phenomenon, which points to popularity item bias.

In this study, the quality of recommendations (via lists of items) was evaluated using rank metrics, and biases were evaluated at the individual and group levels. The analysis reveals that GNN-based methods have great potential in providing accurate recommendations, although the results vary depending on the characteristics of the dataset used. It has been shown that these methods do not always behave poorly in the face of biases, and some can produce good values of the bias metrics while maintaining fairly good precision in the recommendations. The most negative conclusion obtained in relation to recommender systems is that the most appropriate fairness metrics for this type of system present poor values when GNN-based recommender methods are applied. This can lead to more ways to improve these methods to mitigate biases.

In future work, we will investigate the main roots of bias amplification and unfairness in GNN-based recommendation algorithms and conduct bias mitigation solutions with respect to the performance and reliability of the models. We will consider more sensitive attributes and non-binary attributes.

Author Contributions: Conceptualization, N.C., K.T. and M.N.M.-G.; methodology, N.C. and M.N.M.-G.; software, N.C. and K.T.; validation, N.C. and K.T.; formal analysis, N.C. and K.T.; investigation, N.C. and K.T.; resources, M.N.M.-G.; data curation, N.C.; writing—original draft preparation, N.C.; writing—review and editing, N.C., K.T. and M.N.M.-G.; visualization, N.C. and K.T.; supervision, M.N.M.-G.; project administration, M.N.M.-G.; funding acquisition, M.N.M.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were used. Details are provided in Section 3.1.

Conflicts of Interest: There are no conflict of interest.

References

1. Ricci, F.; Rokach, L.; Shapira, B. Recommender Systems: Techniques, Applications, and Challenges. In *Recommender Systems Handbook*; Springer: New York, NY, USA, 2022; pp. 1–35.
2. Zheng, Y.; Wang, D.X. A survey of recommender systems with multi-objective optimization. *Neurocomputing* **2022**, *474*, 141–153. [CrossRef]
3. Pérez-Marcos, J.; Martín-Gómez, L.; Jiménez-Bravo, D.M.; López, V.F.; Moreno-García, M.N. Hybrid system for video game recommendation based on implicit ratings and social networks. *J. Ambient Intell. Humanized Comput.* **2020**, *11*, 4525–4535. [CrossRef]
4. Lin, S.; Wang, J.; Zhu, Z.; Caverlee, J. Quantifying and Mitigating Popularity Bias in Conversational Recommender Systems. *arXiv* **2022**, arXiv:2208.03298.
5. Chen, J.; Dong, H.; Wang, X.; Feng, F.; Wang, M.; He, X. Bias and debias in recommender system: A survey and future directions. *arXiv* **2020**, arXiv:2010.03240.
6. Boratto, L.; Marras, M. Advances in Bias-aware Recommendation on the Web. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, 8–12 March 2021; pp. 1147–1149.
7. Misztal-Radecka, J.; Indurkhya, B. Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems. *Inf. Process. Manag.* **2021**, *58*, 102519. [CrossRef]
8. Gao, C.; Wang, X.; He, X.; Li, Y. Graph neural networks for recommender system. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Tempe, AZ, USA, 21–25 February 2022; pp. 1623–1625.

9. Di Noia, T.; Tintarev, N.; Fatourou, P.; Schedl, M. Recommender systems under European AI regulations. *Commun. ACM* **2022**, *65*, 69–73. [CrossRef]
10. Fahse, T.; Huber, V.; Giffen, B.V. Managing bias in machine learning projects. In Proceedings of the International Conference on Wirtschaftsinformatik, online, 9–11 March 2021; pp. 94–109.
11. Kordzadeh, N.; Ghasemaghaei, M. Algorithmic bias: Review, synthesis, and future research directions. *Eur. J. Inf. Syst.* **2022**, *31*, 388–409. [CrossRef]
12. Boratto, L.; Fenu, G.; Marras, M. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Inf. Process. Manag.* **2021**, *58*, 102387. [CrossRef]
13. Gao, C.; Lei, W.; Chen, J.; Wang, S.; He, X.; Li, S.; Li, B.; Zhang, Y.; Jiang, P. CIRS: Bursting Filter Bubbles by Counterfactual Interactive Recommender System. *arXiv* **2022**, arXiv:2204.01266.
14. Wang, Y.; Ma, W.; Zhang, M.; Liu, Y.; Ma, S. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.* **2022**, *41*, 52. [CrossRef]
15. Yalcin, E.; Bilge, A. Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Inf. Process. Manag.* **2022**, *59*, 103100. [CrossRef]
16. Ahanger, A.B.; Aalam, S.W.; Bhat, M.R.; Assad, A. Popularity Bias in Recommender Systems-A Review. In Proceedings of the International Conference on Emerging Technologies in Computer Engineering, Jaipur, India, 4–5 February 2022; pp. 431–444.
17. Tran, T.N.T.; Felfernig, A.; Tintarev, N. Humanized recommender systems: State-of-the-art and research issues. *ACM Trans. Interact. Intell. Syst.* **2021**, *11*, 9. [CrossRef]
18. Steck, H.; Baltrunas, L.; Elahi, E.; Liang, D.; Raimond, Y.; Basilico, J. Deep learning for recommender systems: A Netflix case study. *AI Mag.* **2021**, *42*, 7–18. [CrossRef]
19. Khan, Z.Y.; Niu, Z.; Sandiwarno, S.; Prince, R. Deep learning techniques for rating prediction: A survey of the state-of-the-art. *Artif. Intell. Rev.* **2021**, *54*, 95–135. [CrossRef]
20. Mu, R. A survey of recommender systems based on deep learning. *IEEE Access* **2018**, *6*, 69009–69022. [CrossRef]
21. Chizari, N.; Shoeibi, N.; Moreno-García, M.N. A Comparative Analysis of Bias Amplification in Graph Neural Network Approaches for Recommender Systems. *Electronics* **2022**, *11*, 3301. [CrossRef]
22. Dai, E.; Wang, S. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, 8–12 March 2021; pp. 680–688.
23. Wu, S.; Sun, F.; Zhang, W.; Xie, X.; Cui, B. Graph neural networks in recommender systems: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 97. [CrossRef]
24. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81. [CrossRef]
25. Zhang, Q.; Wipf, D.; Gan, Q.; Song, L. A biased graph neural network sampler with near-optimal regret. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8833–8844.
26. Alelyani, S. Detection and Evaluation of Machine Learning Bias. *Appl. Sci.* **2021**, *11*, 6271. [CrossRef]
27. Zeng, Z.; Islam, R.; Keya, K.N.; Foulds, J.; Song, Y.; Pan, S. Fair representation learning for heterogeneous information networks. In Proceedings of the International AAAI Conference on Weblogs and Social Media, Atlanta, GA, USA, 6–9 June 2022; Volume 15.
28. Bruce, P.; Bruce, A.; Gedeck, P. *Practical statistics for Data Scientists*, 2nd ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2020; pp. 50–51.
29. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**. [CrossRef]
30. Bernhardt, M.; Jones, C.; Glocker, B. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nat. Med.* **2022**, *28*, 1157–1158. [CrossRef] [PubMed]
31. Hall, M.; van der Maaten, L.; Gustafson, L.; Adcock, A. A Systematic Study of Bias Amplification. *arXiv* **2022**, arXiv:2201.11706.
32. Gu, J.; Oelke, D. Understanding bias in machine learning. *arXiv* **2019**, arXiv:1909.01866.
33. Blanzeisky, W.; Cunningham, P. Algorithmic factors influencing bias in machine learning. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Bilbao, Spain, 13–17 September 2021; pp. 559–574.
34. Akter, S.; Dwivedi, Y.K.; Sajib, S.; Biswas, K.; Bandara, R.J.; Michael, K. Algorithmic bias in machine learning-based marketing models. *J. Bus. Res.* **2022**, *144*, 201–216. [CrossRef]
35. Caton, S.; Haas, C. Fairness in machine learning: A survey. *arXiv* **2020**, arXiv:2010.04053.
36. Verma, S.; Rubin, J. Fairness definitions explained. In Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (Fairware), Gothenburg, Sweden, 29 May 2018; pp. 1–7.
37. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 259–268.
38. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. *arXiv* **2016**, arXiv:1610.02413.
39. Seymour, W. Detecting Bias: Does an Algorithm Have to Be Transparent in Order to Be Fair? *BIAS 2018*, **2018**, *4*, 543–555.
40. Ashokan, A.; Haas, C. Fairness metrics and bias mitigation strategies for rating predictions. *Inf. Process. Manag.* **2021**, *58*, 102646. [CrossRef]

41. Dong, Y.; Wang, S.; Wang, Y.; Derr, T.; Li, J. On Structural Explanation of Bias in Graph Neural Networks. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 316–326.
42. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef]
43. Dong, Y.; Liu, N.; Jalaian, B.; Li, J. Edits: Modeling and mitigating data bias for graph neural networks. In Proceedings of the ACM Web Conference 2022, Lyon, France, 25–29 April 2022; pp. 1259–1269.
44. Chen, Z.; Xiao, T.; Kuang, K. BA-GNN: On Learning Bias-Aware Graph Neural Network. In Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE), Virtual Event, 9–12 May 2022; pp. 3012–3024.
45. Xu, B.; Shen, H.; Sun, B.; An, R.; Cao, Q.; Cheng, X. Towards consumer loan fraud detection: Graph neural networks with role-constrained conditional random field. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 4537–4545.
46. Tang, X.; Yao, H.; Sun, Y.; Wang, Y.; Tang, J.; Aggarwal, C.; Mitra, P.; Wang, S. Graph convolutional networks against degree-related biases. *arXiv* **2020**, arXiv:2006.15643.
47. Li, P.; Wang, Y.; Zhao, H.; Hong, P.; Liu, H. On dyadic fairness: Exploring and mitigating bias in graph connections. In Proceedings of the International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
48. Loveland, D.; Pan, J.; Bhathena, A.F.; Lu, Y. FairEdit: Preserving Fairness in Graph Neural Networks through Greedy Graph Editing. *arXiv* **2022**, arXiv:2201.03681.
49. Kose, O.D.; Shen, Y. FairNorm: Fair and Fast Graph Neural Network Training. *arXiv* **2022**, arXiv:2205.09977.
50. Baeza-Yates, R. Data and algorithmic bias in the web. In Proceedings of the 8th ACM Conference on Web Science, Hannover, Germany, 22–25 May 2016.
51. Sun, W.; Khenissi, S.; Nasraoui, O.; Shafto, P. Debiasing the human-recommender system feedback loop in collaborative filtering. In Proceedings of the 2019 World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 645–651.
52. Fabbri, F.; Croci, M.L.; Bonchi, F.; Castillo, C. Exposure Inequality in People Recommender Systems: The Long-Term Effects. In Proceedings of the International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 6–9 June 2022; Volume 16, pp. 194–204.
53. Mansoury, M.; Abdollahpouri, H.; Pechenizkiy, M.; Mobasher, B.; Burke, R. A graph-based approach for mitigating multi-sided exposure bias in recommender systems. *ACM Trans. Inf. Syst. (TOIS)* **2021**, *40*. [CrossRef]
54. Ovaisi, Z.; Heinecke, S.; Li, J.; Zhang, Y.; Zheleva, E.; Xiong, C. RGRecSys: A Toolkit for Robustness Evaluation of Recommender Systems. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Tempe, AZ, USA, 21–25 February 2022; pp. 1597–1600.
55. Abdollahpouri, H.; Burke, R.; Mobasher, B. Managing popularity bias in recommender systems with personalized re-ranking. In Proceedings of the Thirty-Second International Flairs Conference, Sarasota, FL, USA, 19–22 May 2019.
56. Wang, S.; Hu, L.; Wang, Y.; He, X.; Sheng, Q.Z.; Orgun, M.A.; Cao, L.; Ricci, F.; Yu, P.S. Graph learning based recommender systems: A review. *arXiv* **2021**, arXiv:2105.06339.
57. Boratto, L.; Fenu, G.; Marras, M.; Medda, G. Consumer fairness in recommender systems: Contextualizing definitions and mitigations. In Proceedings of the European Conference on Information Retrieval, Stavanger, Norway, 10–14 April 2022; pp. 552–566.
58. Liu, H.; Wang, Y.; Lin, H.; Xu, B.; Zhao, N. Mitigating sensitive data exposure with adversarial learning for fairness recommendation systems. *Neural Comput. Appl.* **2022**, *34*, 18097–18111. [CrossRef]
59. Shakespeare, D.; Porcaro, L.; Gómez, E.; Castillo, C. Exploring artist gender bias in music recommendation. *arXiv* **2020**, arXiv:2009.01715.
60. Saxena, S.; Jain, S. Exploring and Mitigating Gender Bias in Recommender Systems with Explicit Feedback. *arXiv* **2021**, arXiv:2112.02530.
61. Neophytou, N.; Mitra, B.; Stinson, C. Revisiting popularity and demographic biases in recommender evaluation and effectiveness. In Proceedings of the European Conference on Information Retrieval, Stavanger, Norway, 10–14 April 2022; pp. 641–654.
62. Gómez, E.; Boratto, L.; Salamó, M. Provider fairness across continents in collaborative recommender systems. *Inf. Process. Manag.* **2022**, *59*, 102719. [CrossRef]
63. Rahmani, H.A.; Naghiaei, M.; Dehghan, M.; Aliannejadi, M. Experiments on Generalizability of User-Oriented Fairness in Recommender Systems. *arXiv* **2022**, arXiv:2205.08289.
64. Fang, M.; Liu, J.; Momma, M.; Sun, Y. FairRoad: Achieving Fairness for Recommender Systems with Optimized Antidote Data. In Proceedings of the 27th ACM on Symposium on Access Control Models and Technologies, New York, NY, USA, 8–10 June 2022; pp. 173–184.
65. Naghiaei, M.; Rahmani, H.A.; Deldjoo, Y. Cpfair: Personalized consumer and producer fairness re-ranking for recommender systems. *arXiv* **2022**, arXiv:2204.08085.
66. Li, Y.; Hedia, M.L.; Ma, W.; Lu, H.; Zhang, M.; Liu, Y.; Ma, S. Contextualized Fairness for Recommender Systems in Premium Scenarios. *Big Data Res.* **2022**, *27*, 100300. [CrossRef]
67. Rahman, T.; Surma, B.; Backes, M.; Zhang, Y. Fairwalk: Towards fair graph embedding. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 1–7 January 2019.
68. Wu, L.; Chen, L.; Shao, P.; Hong, R.; Wang, X.; Wang, M. Learning fair representations for recommendation: A graph-based perspective. In Proceedings of the Web Conference 2021, Virtual, 19–23 April 2021; pp. 2198–2208.

69. Chen, J.; Wu, W.; Shi, L.; Zheng, W.; He, L. Long-tail session-based recommendation from calibration. *Appl. Intell.* **2022**, *53*, 4685–4702. [CrossRef]
70. Kim, M.; Oh, J.; Do, J.; Lee, S. Debiasing Neighbor Aggregation for Graph Neural Network in Recommender Systems. *arXiv* **2022**, arXiv:2208.08847.
71. Zhao, M.; Wu, L.; Liang, Y.; Chen, L.; Zhang, J.; Deng, Q.; Wang, K.; Shen, X.; Lv, T.; Wu, R. Investigating Accuracy-Novelty Performance for Graph-based Collaborative Filtering. *arXiv* **2022**, arXiv:2204.12326.
72. Yang, L.; Liu, Z.; Dou, Y.; Ma, J.; Yu, P.S. Consisrec: Enhancing gnn for social recommendation via consistent neighbor aggregation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Online, 11–15 July 2021; pp. 2141–2145.
73. Liu, Z.; Wan, M.; Guo, S.; Achan, K.; Yu, P.S. Basconv: Aggregating heterogeneous interactions for basket recommendation with graph convolutional neural network. In Proceedings of the 2020 SIAM International Conference on Data Mining, Cincinnati, OH, USA, 7–9 May 2020; pp. 64–72.
74. Harper, M.; Konstan, J.A. The MovieLens Datasets: Distributed by GroupLens at the University of Minnesota. 2021. Available online: https://grouplens.org/datasets/movielens/ (accessed on 14 February 2023).
75. Floridi, L.; Holweg, M.; Taddeo, M.; Amaya Silva, J.; Mökander, J.; Wen, Y. capAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. University of Oxford, 2022. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091 (accessed on 14 February 2023).
76. Celma, O. *Music Recommendation and Discovery in the Long Tail*; Springer: Berlin, Germany, 2010.
77. Mobius, A. Book Recommendation Dataset. 2020. Available online: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset (accessed on 14 February 2023)
78. Al-Ghamdi, M.; Elazhary, H.; Mojahed, A. Evaluation of Collaborative Filtering for Recommender Systems. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 5–53. [CrossRef]
79. Airen, S.; Agrawal, J. Movie recommender system using k-nearest neighbors variants. *Natl. Acad. Sci. Lett.* **2022**, *45*, 75–82. [CrossRef]
80. Deshpande, M.; Karypis, G. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst. (TOIS)* **2004**, *22*, 143–177. [CrossRef]
81. Bahadorpour, M.; Neysiani, B.S.; Shahraki, M.N. Determining optimal number of neighbors in item-based kNN collaborative filtering algorithm for learning preferences of new users. *J. Telecommun. Electron. Comput. Eng. (JTEC)* **2017**, *9*, 163–167.
82. Sang, L.; Xu, M.; Qian, S.; Wu, X. Knowledge graph enhanced neural collaborative filtering with residual recurrent network. *Neurocomputing* **2021**, *454*, 417–429. [CrossRef]
83. Girsang, A.S.; Wibowo, A.; Jason; Roslynlia. Neural collaborative for music recommendation system. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2021; Volume 1071, p. 012021.
84. Bai, T.; Wen, J.R.; Zhang, J.; Zhao, W.X. A neural collaborative filtering model with interaction-based neighborhood. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 1979–1982.
85. Kuang, H.; Xia, W.; Ma, X.; Liu, X. Deep matrix factorization for cross-domain recommendation. In Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 13–14 March 2021; Volume 5, pp. 2171–2175.
86. Himabindu, T.V.; Padmanabhan, V.; Pujari, A.K. Conformal matrix factorization based recommender system. *Inf. Sci.* **2018**, *467*, 685–707. [CrossRef]
87. Xue, H.J.; Dai, X.; Zhang, J.; Huang, S.; Chen, J. Deep matrix factorization models for recommender systems. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017; Volume 17, pp. 3203–3209.
88. Yi, B.; Shen, X.; Liu, H.; Zhang, Z.; Zhang, W.; Liu, S.; Xiong, N. Deep matrix factorization with implicit feedback embedding for recommendation system. *IEEE Trans. Ind. Inform.* **2019**, *15*, 4591–4601. [CrossRef]
89. Liang, G.; Sun, C.; Zhou, J.; Luo, F.; Wen, J.; Li, X. A General Matrix Factorization Framework for Recommender Systems in Multi-access Edge Computing Network. *Mobile Netw. Appl.* **2022**, *27*, 1629–1641. [CrossRef]
90. Zhang, Z.; Liu, Y.; Xu, G.; Luo, G. Recommendation using DMF-based fine tuning method. *J. Intell. Inf. Syst.* **2016**, *47*, 233–246. [CrossRef]
91. Schafer, J.B.; Frankowski, D.; Herlocker, J.; Sen, S. Collaborative filtering recommender systems. In *The Adaptive Web*; Springer: Berlin, Germany, 2007; pp. 291–324.
92. He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 25–30 July 2020; pp. 639–648.
93. Broman, N. Comparasion of Recommender Systems for Stock Inspiration. Bachelor's Thesis, Stockholm School of Economics, Stockholm, Sweden, 2021.
94. Ding, S.; Feng, F.; He, X.; Liao, Y.; Shi, J.; Zhang, Y. Causal incremental graph convolution for recommender system retraining. *arXiv* **2022**, arXiv:2108.06889.
95. Sun, W.; Chang, K.; Zhang, L.; Meng, K. INGCF: An Improved Recommendation Algorithm Based on NGCF. In Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing, Xiamen, China, 3–5 December 2021; pp. 116–129.

96. Wang, X.; He, X.; Wang, M.; Feng, F.; Chua, T.S. Neural graph collaborative filtering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 165–174.
97. Wu, J.; Wang, X.; Feng, F.; He, X.; Chen, L.; Lian, J.; Xie, X. Self-supervised graph learning for recommendation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Online, 11–15 July 2021; pp. 726–735.
98. Yang, C. Supervised Contrastive Learning for Recommendation. *arXiv* **2022**, arXiv:2201.03144.
99. Tang, H.; Zhao, G.; Wu, Y.; Qian, X. Multisample-based Contrastive Loss for Top-k Recommendation. *IEEE Trans. Multimed.* **2021**, *25*, 339–351. [CrossRef]
100. Wang, X.; Jin, H.; Zhang, A.; He, X.; Xu, T.; Chua, T.S. Disentangled graph collaborative filtering. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 25–30 July 2020; pp. 1001–1010.
101. Bourhim, S.; Benhiba, L.; Idrissi, M.J. A Community-Driven Deep Collaborative Approach for Recommender systems. In Proceedings of the 2019 IEEE International Conference on Web Services (ICWS), Milan, Italy, 8–13 July 2019.
102. Sha, X.; Sun, Z.; Zhang, J. Disentangling Multi-Facet Social Relations for Recommendation. *IEEE Trans. Comput. Soc. Syst.* **2021**, *9*, 867–878. [CrossRef]
103. Foulds, J.R.; Islam, R.; Keya, K.N.; Pan, S. Differential Fairness. NeurIPS 2019 Workshop on Machine Learning with Guarantees, Vancouver, Canada. UMBC Faculty Collection. 2019. Available online: https://www.semanticscholar.org/paper/Differential-Fairness-Foulds-Islam/cf3081d5fa83750a89898ae1adcef7925ed8af81 (accessed on 14 February 2023).
104. Foulds, J.R.; Islam, R.; Keya, K.N.; Pan, S. An intersectional definition of fairness. In Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, 20–24 April 2020; pp. 1918–1921.
105. Yao, S.; Huang, B. Beyond parity: Fairness objectives for collaborative filtering. *arXiv* **2017**, arXiv:1705.08804.
106. Farnadi, G.; Kouki, P.; Thompson, S.K.; Srinivasan, S.; Getoor, L. A fairness-aware hybrid recommender system. *arXiv* **2018**, arXiv:1809.09030.
107. Naghiaei, M.; Rahmani, H.A.; Dehghan, M. The Unfairness of Popularity Bias in the Book Recommendation. *arXiv* **2022**, arXiv:2202.13446.
108. Lazovich, T.; Belli, L.; Gonzales, A.; Bower, A.; Tantipongpipat, U.; Lum, K.; Huszar, F.; Chowdhury, R. Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *arXiv* **2022**, arXiv:2202.01615.
109. Wang, X.; Wang, W.H. Providing Item-side Individual Fairness for Deep Recommender Systems. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 117–127.
110. Islam, R.; Keya, K.N.; Zeng, Z.; Pan, S.; Foulds, J. Debiasing career recommendations with neural fair collaborative filtering. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 3779–3790.
111. Aalam, S.W.; Ahanger, A.B.; Bhat, M.R.; Assad, A. Evaluation of Fairness in Recommender Systems: A Review. In Proceedings of the International Conference on Emerging Technologies in Computer Engineering, Jaipur, India, 4–5 February 2022; pp. 456–465.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.